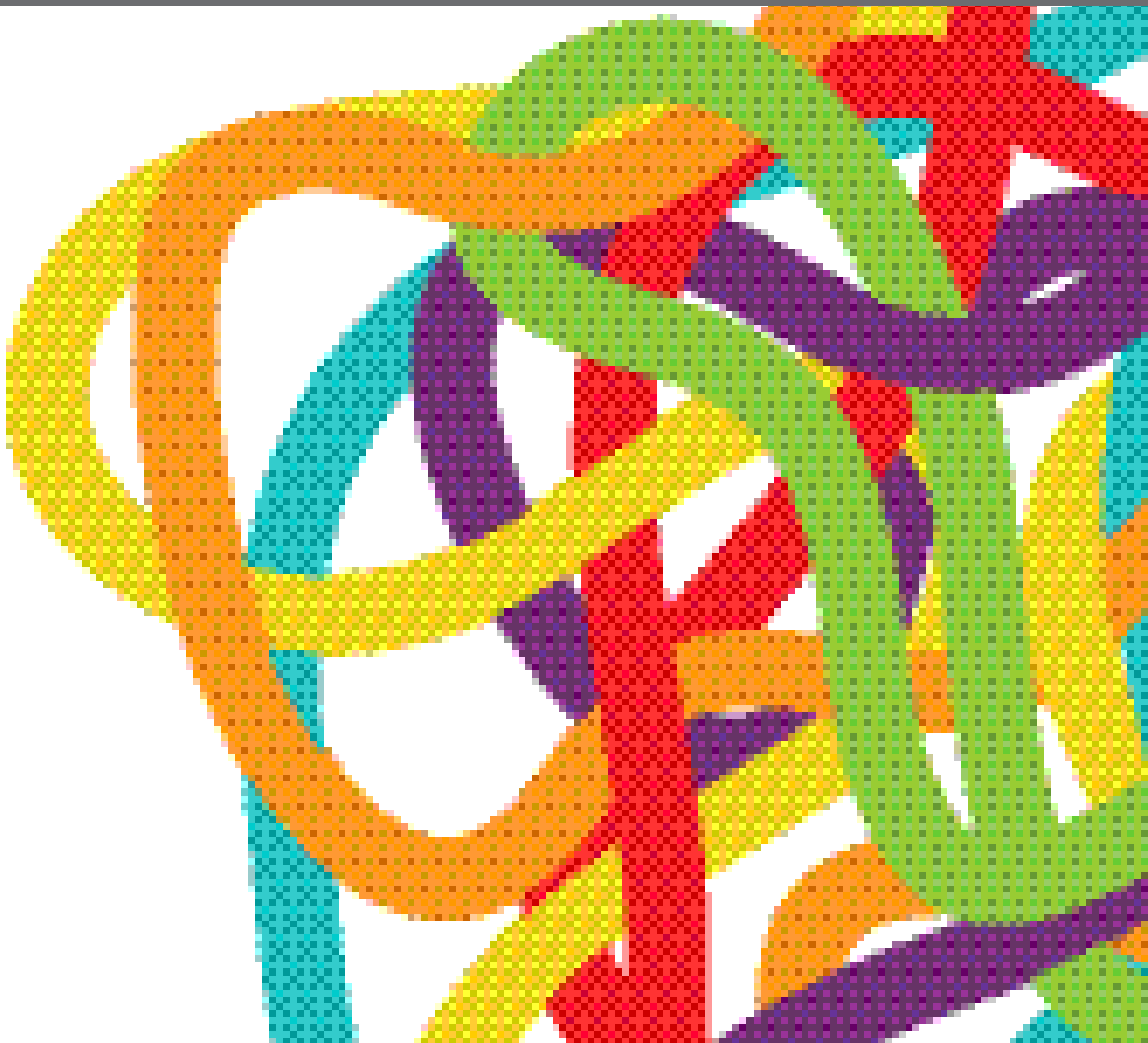
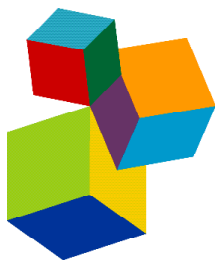


MULTI-OMIC DATA INTEGRATION IN ONCOLOGY

EDITED BY: Chiara Romualdi, Enrica Calura, Davide Risso,
Sampsa Hautaniemi and Francesca Finotello
PUBLISHED IN: Frontiers in Oncology and Frontiers in Genetics





frontiers

Frontiers eBook Copyright Statement

The copyright in the text of individual articles in this eBook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this eBook is the property of Frontiers.

Each article within this eBook, and the eBook itself, are published under the most recent version of the Creative Commons CC-BY licence.

The version current at the date of publication of this eBook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or eBook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714

ISBN 978-2-88966-151-0

DOI 10.3389/978-2-88966-151-0

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.

Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

MULTI-OMIC DATA INTEGRATION IN ONCOLOGY

Topic Editors:

Chiara Romualdi, University of Padua, Italy

Enrica Calura, University of Padua, Italy

Davide Risso, University of Padua, Italy

Sampsa Hautaniemi, University of Helsinki, Finland

Francesca Finotello, Innsbruck Medical University, Austria

Citation: Romualdi, C., Calura, E., Risso, D., Hautaniemi, S., Finotello, F., eds.
(2020). Multi-omic Data Integration in Oncology. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88966-151-0

Table of Contents

- 05 Editorial: Multi-omic Data Integration in Oncology**
Francesca Finotello, Enrica Calura, Davide Risso, Sampsa Hautaniemi and Chiara Romualdi
- 09 Identification of Specific Long Non-Coding Ribonucleic Acid Signatures and Regulatory Networks in Prostate Cancer in Fine-Needle Aspiration Biopsies**
Zehuan Li, Jianghua Zheng, Qianlin Xia, Xiaomeng He, Juan Bao, Zhanghan Chen, Hiroshi Katayama, Die Yu, Xiaoyan Zhang, Jianqing Xu, Tongyu Zhu and Jin Wang
- 21 Computational Methods for the Integrative Analysis of Genomics and Pharmacological Data**
Jimmy Caroli, Martina Dori and Silvio Bicciato
- 27 Genomic and Transcriptomic Landscape of Tumor Clonal Evolution in Cholangiocarcinoma**
Geng Chen, Zhixiong Cai, Xiuqing Dong, Jing Zhao, Song Lin, Xi Hu, Fang-E Liu, Xiaolong Liu and Huqing Zhang
- 38 Computational Oncology in the Multi-Omics Era: State of the Art**
Guillermo de Anda-Jáuregui and Enrique Hernández-Lemus
- 59 Multi-Omic Regulation of the PAM50 Gene Signature in Breast Cancer Molecular Subtypes**
Soledad Ochoa, Guillermo de Anda-Jáuregui and Enrique Hernández-Lemus
- 73 Big Data-Based Identification of Multi-Gene Prognostic Signatures in Liver Cancer**
Meiliang Liu, Xia Liu, Shun Liu, Feifei Xiao, Erna Guo, Xiaoling Qin, Liuyu Wu, Qiuli Liang, Zerui Liang, Kehua Li, Di Zhang, Yu Yang, Xingxi Luo, Lei Lei, Jennifer Hui Juan Tan, Fuqiang Yin and Xiaoyun Zeng
- 96 Impact of Data Preprocessing on Integrative Matrix Factorization of Single Cell Data**
Lauren L. Hsu and Aedin C. Culhane
- 105 Toward Systems Biomarkers of Response to Immune Checkpoint Blockers**
Óscar Lapuente-Santana and Federica Eduati
- 114 Integrative Network Fusion: A Multi-Omics Approach in Molecular Profiling**
Marco Chierici, Nicole Bussola, Alessia Marcolini, Margherita Francescato, Alessandro Zandonà, Lucia Trastulla, Claudio Agostinelli, Giuseppe Jurman and Cesare Furlanello
- 128 Integrated Multi-Omics Analyses in Oncology: A Review of Machine Learning Methods and Tools**
Giovanna Nicora, Francesca Vitali, Arianna Dagliati, Nophar Geifman and Riccardo Bellazzi
- 139 Integrated Transcriptome Analysis of Human Visceral Adipocytes Unravels Dysregulated microRNA-Long Non-coding RNA-mRNA Networks in Obesity and Colorectal Cancer**
Sabrina Tait, Antonella Baldassarre, Andrea Masotti, Enrica Calura, Paolo Martini, Rosaria Vari, Beatrice Scazzocchio, Sandra Gessani and Manuela Del Cornò

**157 *Unraveling the Complexity of the Cancer Microenvironment With
Multidimensional Genomic and Cytometric Technologies***

Natasja L. de Vries, Ahmed Mahfouz, Frits Koning and Noel F. C. C. de Miranda

**172 *Multi-Omics Characterization of the 4T1 Murine Mammary Gland Tumor
Model***

Barbara Schrörs, Sebastian Boegel, Christian Albrecht, Thomas Bukur,
Valesca Bukur, Christoph Holtsträter, Christoph Ritzel, Katja Manninen,
Arbel D. Tadmor, Mathias Vormehr, Ugur Sahin and Martin Löwer



Editorial: Multi-omic Data Integration in Oncology

Francesca Finotello¹, Enrica Calura², Davide Risso³, Sampsa Hautaniemi⁴ and Chiara Romualdi^{2*}

¹ Biocenter, Institute of Bioinformatics, Medical University of Innsbruck, Innsbruck, Austria, ² Department of Biology, University of Padua, Padua, Italy, ³ Department of Statistical Sciences, University of Padua, Padua, Italy, ⁴ Research Program in Systems Oncology, Faculty of Medicine, University of Helsinki, Helsinki, Finland

Keywords: multi-omic, single-cell, transcriptomics, pathways, cancer, data integration

Editorial on the Research Topic

Multi-omic Data Integration in Oncology

In the next few years, we are going to witness changes in the treatment of cancer patients due to molecular and personalized medicine. Indeed, many hospitals are already starting routine genome-wide screening to complement and inform diagnosis and treatment choices. However, the majority of molecular aberrations identified in cancers have synergic interactions in many aspects of cell signaling beyond the genome. The complexity of cancers cross cell boundaries especially studying the tumor microenvironment as a heterogeneous and dynamic network of interacting cells (1), one of the new hot topics for anticancer treatment development. In this scenario, multi-omic technologies and single-cell data can shed light on these interactions by generating high-throughput datasets portraying the genomes, transcriptomes, proteomes, metabolomes, and epigenomes of tumors.

Large-scale cancer genomic projects, such as The Cancer Genome Atlas (TCGA) (2), have generated petabytes of multi-omic data portraying this heterogeneity. Importantly, these data have been made available to the scientific community, shifting the main challenge from data collection to data analysis and integration, and allowing for development of novel data analysis methods. However, while computational and statistical analyses of single-omics datasets are well-established—excluding the still challenging single-cell data analyses—the integration of multi-omic data is still far from being standardized. As the number of datasets grows and the biological knowledge increases, existing methods should be extended or generalized, and new computational tools need to be proposed to cope with the complexity and multi-level structure of the available information. In this special issue, de Anda-Jáuregui and Hernández-Lemus presented a comprehensive review of the state of the art of multi-omic data analysis in oncology, encompassing a wide range of tasks, such as data acquisition and processing, data management, identification of therapeutic targets, as well as patient classification, diagnosis, and prognosis.

One of the major challenges in the analysis of multi-omic data is how to integrate the different data modalities. Nicora et al. reviewed a selection of recent tools for the computational integration of multi-omic data sets based on: deep learning, network integration, data clustering or factorization, and feature extraction or transformation. This emerging field has already contributed a rich catalog of freely available tools: the most widely used approaches are network-based methods,

OPEN ACCESS

Edited and reviewed by:

Claudio Sette,
Catholic University of the Sacred
Heart, Rome, Italy

*Correspondence:

Chiara Romualdi
chiara.romualdi@unipd.it

Specialty section:

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

Received: 05 August 2020

Accepted: 07 August 2020

Published: 15 September 2020

Citation:

Finotello F, Calura E, Risso D,
Hautaniemi S and Romualdi C (2020)
Editorial: Multi-omic Data Integration
in Oncology. *Front. Oncol.* 10:1768.
doi: 10.3389/fonc.2020.01768

but deep learning strategies are becoming increasingly popular. In this context, Chierici et al. proposed a computational framework for high-throughput data integration (called Integrative Network Fusion, INF), which leverages network structures and machine learning models to extract multi-omic predictive biomarkers for cancer subtype identification. By integrating gene expression, protein expression, and copy-number data across three TCGA cancer types, INF showed a higher predictive performance with respect to simple juxtaposition of single-omics analyses and enabled the extraction of more biologically meaningful biomarkers. INF was designed to integrate an arbitrary number of omic layers, allowing to extend the framework to other types of data, such as histopathological and radiological images.

The main goal of most integrative methods is the identification of multi-omic signatures that can be diagnostic (healthy vs. disease), prognostic (good vs. poor patient outcome), or predictive (good vs. poor response to therapeutic interventions). The selection of the optimal signature size, that is the number of molecular features needed to stratify patients, is not trivial. In general, the smaller the signature size, the easier its clinical applicability, but the lower its accuracy, due to patients heterogeneity. In this perspective meta-analysis studies that exploit data from previously published studies can increase the signature robustness and reliability. Liu et al. combined extensive text mining and transcriptomic data to identify and validate a small prognostic signature in liver cancer. By selecting more than thousand genes known to be involved in liver cancer initiation and progression, they identified a triplet of genes associated with survival. Using three independent cohorts and specific experimental assays to confirm transcript and protein expression levels, they found that low expression of *F2*, *GOT2*, and *TRPV1* is associated with poor prognosis in liver cancer. In a parallel study, Li et al. identified a small diagnostic signature composed of long non-coding RNAs (RP11-33A14.1, RP11-423H2.3, and LAMTOR5-AS1) that, combined with clinical and previously-published molecular biomarkers, is able to predict prostate cancer from fine needle aspiration biopsies with high sensitivity and specificity. Looking for potential molecular functions of the signature elements, the authors suggested and validated a sponge mechanism, that sees miR-7, miR-24-3p, and miR-30 as the three main miRNAs sequestered by the long non-coding RNAs, which in turn interact with the RNA binding protein FUS.

While the identification of precise molecular signatures is fundamental for clinical practice, the understanding of the actual mechanisms driving these alterations in specific cancers or cancer subtypes is crucial to design new pharmacological treatments. Ochoa et al. investigated the regulatory elements that drive the various expression behaviors of the PAM50 signature (3) in different breast cancer subtypes. The authors integrated coding and non-coding gene expression, methylation levels, and information on transcription factors (TF)-target interaction data via a generalized elastic-net model. Using breast tumors and normal adjacent tissues from the TCGA, they identified both subtype-specific regulators and regulators acting across subtypes, such as miR-21 and miR-10b. With a

similar aim, Tait et al. combine transcriptomic data to study the expression patterns of non-coding elements (miRNAs and long non-coding RNAs, ncRNA) underlying dysfunctional adipocyte phenotype in obesity and colorectal cancer. The authors inferred lncRNA-miRNA-mRNA modules, highlighting several ncRNA modulations and dysregulated pathways that are common to both obesity and colorectal cancer. Chen et al., using whole exome and transcriptome sequencing, studied the genomic and transcriptomic landscape of cholangiocarcinoma. The authors investigated subnetworks that were greatly influenced by tumor clonal or subclonal mutations impacting gene expression.

Immunotherapy with checkpoint blockers has drastically advanced treatment of different types of cancer over the past years, improving overall patient survival compared to standard therapy. However, response to treatment remains hard to predict due to the large intra- and inter-patient heterogeneity. Lapuente-Santana and Eduarti reviewed the benefit of multi-omic approaches for biomarker discovery in the immuno-oncology field. They present multi-omic approaches that could help understand how different immune cell types can influence the efficacy of immunotherapy with checkpoint blockers and how the cells interact in the tumor microenvironment, shaping the immune response, and resistance to immunotherapy. The authors suggest that a combination of dynamic mathematical models and longitudinal data could further improve our understanding of the tumor microenvironment role in the response to immunotherapy and provide the rationale for alternative personalized treatments.

Another field that recently had a boost from multi-omic integration strategies is pharmacogenomics. The term pharmacogenomics is generally used to define the variability of drug response due to the patients' genomic landscape. In this context, cancer cell lines have been the most widely used models to explore the molecular basis of drug sensitivity. Starting from the first NCI-60 project (4), several other studies investigating the link between the genomic makeup and drug response in cancer cell lines have been carried out (5–7). Caroli et al. reviewed the databases and computational tools that have been developed to integrate cancer cell lines genomic profiles and sensitivity to small molecule perturbations obtained from different screenings.

Multimodal omics can be integrated *in silico* to respond to complex biological questions that require a systems biology approach. One of such examples is the prediction of tumor neoantigens, namely mutated peptides that are bound to the major histocompatibility complex molecules of cancer cells and can elicit anticancer immune responses. Schrörs et al. derived an integrated map of the genome, transcriptome, and neoantigen landscape of one of the most widely used breast cancer models: the 4T1 murine mammary cancer cell line. They found that 4T1 cells share molecular features with triple-negative breast cancer and, thus, represent a promising model for preclinical studies. Moreover, the authors confirmed experimentally the antigenic potential of 23 mutated peptides selected from the pool of neoantigens predicted *in silico* using IFN γ -ELISpot assays.

Despite their recognized value to advancing and informing immuno-oncology and precision medicine, standard “bulk”

technologies are intrinsically limited by the sequencing of heterogeneous cell mixtures, which renders a blended average portrayal of the tumor microenvironment. Rapidly-emerging single-cell technologies allow to disentangle the phenotypes of individual cells, providing unprecedented insights into the cellular and spatial diversity of the tumor microenvironment. However, the sparsity, noise, and high-dimensionality of single-cell data pose unique challenges to data analysis. Hsu and Culhane provide a guide to dimensionality reduction techniques that are vital to extract the major sources of variations from single-cell RNA-sequencing data prior to performing downstream data integration, clustering and analysis. The authors focused on principal component analysis (PCA), a matrix factorization method that can easily scale to large datasets when used with sparse-matrix representations; they described its relationship with singular value decomposition, the differences between using correlation or covariance matrices, the impact of data scaling, log-transformation, and standardization, and how to recognize artifacts in PCA plots. Moreover, they described how canonical correlation analysis (CCA), another popular matrix factorization approach, can be used to integrate single-cell data from different platforms or studies.

Despite their promise, single-cell technologies, such as flow cytometry, mass cytometry, or single-cell RNA sequencing, are still limited by the lack of information on spatial context and multicellular interactions. de Vries et al. show how multimodal and spatially-resolved single-cell data can advance our understanding of the inter-cellular organization and communication in the tumor microenvironment. They present recent developments in spatial, tissue-based techniques, such as multiparameter fluorescence, imaging mass cytometry, and *in situ* transcriptomics, as well as, multidimensional single-cell technologies and studies that integrate multiple single-cell modalities to disentangle complex cell interactions in the tumor microenvironment. These approaches hold the promise to uncover the sources of intra-tumor heterogeneity that hamper cancer treatment but require the development of dedicated bioinformatic tools for the data analysis and interpretation and tight collaboration between oncologists, immunologists, pathologists, and bioinformaticians for the extraction of mechanistic rationales and actionable targets.

Overall, our collection of original research articles and reviews covers a wide range of multi-omic applications in oncology. The scenario that emerges is that transcriptomics, methylomics, and genomics are the three most frequently analyzed and integrated data, both in bulk and single-cell studies. To fully understand the complex interactions of the molecular processes underlying cellular mechanisms a fine temporal and spatial resolution is required. Spatial transcriptomics (8), a set of techniques that allow the (sub-) cellular characterization of gene expression, has the potential to unveil the complex interplay between cell types but gives rise to new computational and statistical challenges, also in terms of data integration. In addition, important information can be exploited by integrating omics data and biomedical images (9), a field that is experiencing new advances in terms of sensitivity and resolution. Multi-modal integrative analysis will soon become the standard to study complex systems, and we look forward to exciting new computational developments to tackle data heterogeneity, computational efficiency and results interpretation, and can ultimately push the oncology field forward.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

FF was supported by the Austrian Science Fund (FWF) (project no. T 974-B30). DR was supported by the Programma per Giovani Ricercatori Rita Levi Montalcini granted by the Italian Ministry of Education, University, and Research, by the National Cancer Institute of the National Institutes of Health (2U24CA180996), and by the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation (CZF2019-002443). SH was supported by the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 667403 for HERCULES. CR and EC were supported by Italian Association for Cancer Research (IG 21837 to CR and MFAG 2019 23522 to EC).

REFERENCES

1. Finotello F, Eduati F. Multi-omics profiling of the tumor microenvironment: paving the way to precision immuno-oncology. *Front Oncol.* (2018) 8:430. doi: 10.3389/fonc.2018.00430
2. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet.* (2013) 45:1113–20. doi: 10.1038/ng.2764
3. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol.* (2009) 27:1160–7. doi: 10.1200/JCO.2008.18.1370
4. Shoemaker RH. The NCI60 human tumour cell line anticancer drug screen. *Nat Rev Cancer.* (2006) 6:813–23. doi: 10.1038/nrc1951
5. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature.* (2012) 483:603–7. doi: 10.1038/nature11003
6. Basu A, Bodycombe NE, Cheah JH, Price EV, Liu K, Schaefer GI, et al. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell.* (2013) 154:1151–61. doi: 10.1016/j.cell.2013.08.003
7. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic

- biomarker discovery in cancer cells. *Nucleic Acids Res.* (2013) 41:D955–61. doi: 10.1093/nar/gks1111
8. Burgess DJ. Spatial transcriptomics coming of age. *Nat Rev Genet.* (2019) 20:317. doi: 10.1038/s41576-019-0129-z
9. Sun R, Limkin EJ, Vakalopoulou M, Derclé L, Champiat S, Han SR, et al. A radiomics approach to assess tumour-infiltrating CD8 cells and response to anti-PD-1 or anti-PD-L1 immunotherapy: an imaging biomarker, retrospective multicohort study. *Lancet Oncol.* (2018) 19:1180–91. doi: 10.1016/S1470-2045(18)30413-3

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Finotello, Calura, Risso, Hautaniemi and Romualdi. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Identification of Specific Long Non-Coding Ribonucleic Acid Signatures and Regulatory Networks in Prostate Cancer in Fine-Needle Aspiration Biopsies

Zehuan Li^{1,2†}, Jianghua Zheng^{3†}, Qianlin Xia^{4†}, Xiaomeng He^{1†}, Juan Bao¹, Zhanghan Chen², Hiroshi Katayama⁵, Die Yu¹, Xiaoyan Zhang¹, Jianqing Xu¹, Tongyu Zhu^{1,6} and Jin Wang^{1*}

OPEN ACCESS

Edited by:

Chiara Romualdi,
University of Padova, Italy

Reviewed by:

Smrithi Rajendran,
University of California,
Santa Cruz, United States
Chin-Yo Lin,
University of Houston,
United States

*Correspondence:

Jin Wang
wjincityu@yahoo.com

[†]These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Genetics

Received: 04 June 2019

Accepted: 17 January 2020

Published: 14 February 2020

Citation:

Li Z, Zheng J, Xia Q, He X, Bao J,
Chen Z, Katayama H, Yu D, Zhang X,
Xu J, Zhu T and Wang J (2020)
Identification of Specific Long Non-
Coding Ribonucleic Acid Signatures and
Regulatory Networks in Prostate Cancer
in Fine-Needle Aspiration Biopsies.
Front. Genet. 11:62.
doi: 10.3389/fgene.2020.00062

¹ Scientific Research Center, Shanghai Public Health Clinical Center, Fudan University, Shanghai, China, ² Department of General Surgery, Zhongshan Hospital, Fudan University, Shanghai, China, ³ Department of Laboratory Medicine, Zhoupu Hospital Affiliated to Shanghai University of Medicine & Health Sciences, Shanghai, China, ⁴ Department of Laboratory Medicine, Shanghai Jiao Tong University Affiliated Sixth People's Hospital, Shanghai, China, ⁵ Department of Molecular Oncology, Okayama University Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama, Japan, ⁶ Department of Urology, Shanghai Key Laboratory of Organ Transplantation, Zhongshan Hospital, Fudan University, Shanghai, China

Prostate cancer (PCa) is one of the most common tumors in men and can be lethal, especially if left untreated. A substantial majority of PCa patients not only are diagnosed based on fine needle aspiration (FNA) biopsies, but their treatment choices are also largely driven by the pathological findings obtained with these FNA specimens. It is widely believed that lncRNAs have strong biological significance, but their specific functions and regulatory networks have not been elucidated. lncRNAs may serve as key players and regulators of PCa carcinogenesis and could be novel biomarkers of this cancer. To identify potential markers for early detection of PCa, in this study, we employed a competing endogenous RNA (ceRNA) microarray to identify differentially expressed lncRNAs (DelncRNAs) in PCa tissue and quantitative real-time PCR (qRT-PCR) analysis to validate these DelncRNAs in FNA biopsies. We demonstrated that a total of 451 lncRNAs were differentially expressed in four pairs of PCa/adjacent tissues, and upregulation of the lncRNAs RP11-33A14.1, RP11-423H2.3, and LAMTOR5-AS1 was confirmed in FNA biopsies of PCa by qRT-PCR and was consistent with the ceRNA array data. The association between the expression of the lncRNA LAMTOR5-AS1 and aggressive cancer was also investigated. Regulatory network analysis of DelncRNAs showed that the lncRNAs RP11-33A14.1 and RP11-423H2.3 targeted miR-7, miR-24-3p, and miR-30 and interacted with the RNA binding protein FUS. Knockdown of these DelncRNAs in PCa cells also demonstrated the effects of RP11-423H2.3 on miR-7/miR-24/miR-30 or LAMTOR5-AS1 on miR-942-5p/miR-542-3p via direct interaction. The results of these studies indicate that these three specific lncRNA signatures and regulatory

networks might serve as risk prediction and diagnostic biomarkers for prostate cancer, even in biopsies obtained by FNA.

Keywords: prostate cancer, long non-coding ribonucleic acid, regulatory networks, fine needle aspiration biopsies, microribonucleic acid, ribonucleic acid binding proteins, biomarker

INTRODUCTION

Prostate cancer is the second most common tumor among men worldwide, leading to the highest morbidity and mortality along with lung and bronchial cancer. In 2018, the incidence of prostate cancer (PCa) among all new cancer cases was 19%, and in the USA, ~29,000 men died from prostate cancer (Siegel et al., 2017; Siegel et al., 2018), which is usually diagnosed at a localized stage by the combination of prostate-specific antigens (PSAs), magnetic resonance imaging (MRI), digital rectal examination (DRE), and transrectal ultrasound (TRUS)-guided biopsy (Carroll et al., 2018); most panel members favor informed testing beginning at the age of 45 years. Despite these detection methods and systemic therapies, including radiation therapy, prostatectomy, androgen deprivation therapy, immunotherapy, and chemotherapy (Mohler et al., 2018), several patients are still diagnosed at a late stage of development (Siegel et al., 2018). Moreover, while PCa remains indolent in most individuals, in a minority of patients, PCa behaves aggressively. PSA, which is the most common prostatic marker, has a high specificity for prostate cancer, but its expression cannot be detected in ~5% of patients with high-grade PCa (Epstein, 1993; Van Der Toom et al., 2019) or, conversely, leads to the overdiagnosis of clinically insignificant cancer (Tan et al., 2019). Thus, biomarkers that accurately diagnose prostate cancer and, more importantly, differentiate indolent from life-threatening prostate cancer are urgently required.

Noncoding RNAs (ncRNAs) play key roles in cancer progression and could be used to develop novel biomarkers of prostate cancer (Shan et al., 2017; Xia et al., 2018). Answering the many unknown questions regarding ncRNAs' participation in prostate cancer progression, such as how ncRNAs participate in many pathological processes leading to the development of prostate cancer, how they significantly interact with proteins, and the degree of their specificity and ease of detection in tissues, serum, plasma, and urine could lead to the development of novel biomarkers of this aggressive cancer. In our previous studies, we

demonstrated that four differentially expressed genes (TGBL1, HOXA7, KRT15, and TGM4) in FNA biopsies could facilitate the diagnosis of prostate cancer, which was significantly improved over PSA (Shan et al., 2017), and we found that differentially expressed circular RNAs (circRNAs) (circ_0062019 and circ_0057558) and the host gene SLC19A1 of circ_0062019 could be used as potential novel biomarkers of PCa (Xia et al., 2018). Long noncoding RNAs (lncRNAs) are currently defined as RNA transcripts longer than 200 nucleotides that do not appear to code proteins but control cell fate during development through complex mechanisms, and their dysregulation underlies some human disorders caused by chromosomal deletions and translocations (Batista and Chang, 2013). lncRNAs include several types of RNA transcripts, such as antisense, intronic, and intergenic transcripts, pseudogenes, and retrotransposons (Lee, 2012), which are more cell type-specific than protein-coding genes, and their aberrant expression has been documented in various cancers, including PCa (Hon et al., 2017; Misawa et al., 2017). lncRNAs were found to be involved in prostate carcinogenesis by mediating enhancer-promoter looping, alternative splicing, and antisense gene silencing, antagonizing transcription regulators and repressing DNA repair (Walsh et al., 2014). For example, the lncRNA SChLAP1 promotes aggressive PCa mechanistically by impairing the SWI/SNF axis-mediated regulation of their gene expression and genomic binding (Prensner et al., 2013). The lncRNA NEAT1, which is regulated by estrogen receptor alpha (ER α), drives an oncogenic cascade in PCa and is associated with therapeutic resistance (Chakravarty et al., 2014). The lncRNA HOTAIR increases the androgen receptor-mediated transcriptional program and promotes the growth of castration-resistant prostate cancer (Zhang et al., 2015). Other lncRNAs, such as lncRNA ZEB1-AS1 (Su et al., 2017) and lncRNA HOXD-AS1 (Gu et al., 2017), can also regulate cell proliferation and chemoresistance as oncogenes. However, some lncRNAs, such as lncRNA TUG1 and lncRNA CTB-89H12.4, can mediate sponge regulatory networks as tumor suppressors (Du et al., 2016). Preclinically, the interfering lncRNA MALAT1 can suppress enzalutamide-resistant PCa progression (Wang et al., 2017b). Therefore, lncRNAs play multifaceted roles in PCa and may serve as risk prediction, diagnostic, prognostic, and predictive biomarkers of PCa.

In this study, we applied a competing endogenous RNA (ceRNA) microarray to identify differentially expressed lncRNAs in PCa tissue. Through further validation of the most differentially expressed lncRNAs in prostate biopsy tissues, we found that three lncRNAs, i.e., RP11-33A14.1, RP11-423H2.3, and LAMTOR5-AS1, and their regulatory networks may serve as novel diagnostic biomarkers of PCa.

Abbreviations: AUC, area under the curve; BPH, benign prostatic hyperplasia; ceRNA, competing endogenous RNA; DEGs, differentially expressed genes; DelncRNAs, differentially expressed lncRNAs; DRE, digital rectal examination; ER α , estrogen receptor alpha; FNA, Fine-Needle Aspiration; FPG, fasting plasma glucose; FUS/TLS, fused in sarcoma/translocated in liposarcoma; GEO, gene expression omnibus; lncRNAs, long noncoding RNAs; MRI, magnetic resonance imaging; ncRNAs, noncoding RNAs; NEAT1, nuclear paraspeckle assembly transcript 1; NMD, nonsense mediated RNA decay; PCa, prostate cancer; PSA, prostate specific antigen; PTBP1, polypyrimidine tract-binding protein 1; qRT-PCR, quantitative real-time polymerase chain reaction; ROC, receiver operating characteristic; RBPs, RNA binding proteins; TC, total cholesterol; TG, total triglyceride; TRUS, transrectal ultrasound.

MATERIALS AND METHODS

Cell Lines and Cell Culture

The prostate cancer cell lines 22Rv1 (ATCC No. CRL-2505), DU145 (ATCC No. HTB-81), LNCaP (ATCC No. CRL-1740), and PC3 (ATCC No. CRL-1435) were purchased from the Culture Collection of the Chinese Academy of Sciences, Shanghai, China (<http://www.cellbank.org.cn/>). DU145 and PC3 were cultured in MEM (Cat#: 41500034, Life Technologies) and F-12 (GIBCO, 21700075, Life Technologies), respectively; LNCaP and 22Rv1 were maintained in RPMI-1640 (Cat#: 31800022, Life Technologies) supplemented with 10% fetal bovine serum (FBS) (Thermo Fisher Scientific, Waltham, MA, US) at 37°C in 5% CO₂. The human prostatic epithelial cell lines (HPEpic) were purchased from Shanghai Xinyu Biological Technology Co., Ltd. All cells were cultured according to the ATCC standard procedure.

Prostate Tumor and Benign Prostatic Hyperplasia Tissue Samples

Four pairs of fresh prostate tumor and paracancerous tissues and 105 cases of prostate tissues on fine needle biopsies (FNA), including 48 cases of PCa tissues and 57 cases of benign prostatic hyperplasia (BPH) tissues, were acquired from Zhongshan Hospital Affiliated with Fudan University. This research was approved by the Ethics Committee of Zhongshan Hospital Affiliated with Fudan University and Shanghai Public Health Clinical Center. Written informed consent was obtained from all patients for the use of their tissue samples and clinical records. Each tissue was confirmed by a pathologist specializing in prostate cancer, and a Gleason score was provided for the risk stratification. All samples were stored at -80°C after surgical resection.

Ribonucleic Acid Purification, Competing Endogenous Ribonucleic Acid Microarray, and Data Analysis

Total RNA was extracted and purified using TRIzol reagent (Invitrogen, Carlsbad, CA, US) and an RNeasy Mini Kit (QIAGEN, GmbH, Germany) following the manufacturer's instructions. The total RNA was quantified by a NanoDrop 2000 spectrophotometer (NanoDrop, US) and selected by limiting the 260/280 nm absorbance ratio of the samples to 1.8–2.0. The selected RNA samples were assessed by an Agilent Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA, US) to inspect the RNA integrity. Four pairs of prostate tumor and paracancerous tissues were used for the microarray assay to investigate the differentially expressed lncRNAs between the cancer tissues and paracancerous tissues (Xia et al., 2018). The total RNA was amplified and labeled by a Low Input Quick Amp WT Labeling Kit (Santa Clara, CA, US) and labeled by Cy3-labeled CTP with T7 RNA polymerase. The labeled cRNAs were purified by an RNeasy Mini Kit (QIAGEN, GmbH, Germany) and loaded onto SBC human (4*180 K) ceRNA microarrays, including 68,423 ncRNAs, 88,371 circRNAs, and 18,853

messenger RNAs (mRNAs) (Shanghai Biotech Co., Ltd., Shanghai, China). The microarray hybridization was performed following the manufacturer's standard protocols using a Gene Expression Hybridization Kit (Santa Clara, CA, US) in a hybridization oven (Santa Clara, CA, US). The hybridized slides were washed, fixed, and finally scanned to obtain images using an Agilent Microarray Scanner (Agilent Technologies, Santa Clara, CA, US). The data were extracted with Feature Extraction software 10.7 (Agilent Technologies, Santa Clara, CA, US), and the raw data were normalized by the quantile algorithm in the limma package in R. The significantly differentially expressed lncRNAs between the prostate cancer and paracancerous tissues were identified and retained by screening for fold change > 2.0 at $p < 0.05$. The prostate cancer microarray datasets were deposited in the Gene Expression Omnibus (GEO) database under accession number GSE140927.

Regulatory Network Analysis of Differential Long Non-Coding Ribonucleic Acids and Microribonucleic Acids in Prostate Cancer

For an integrative analysis of prostate cancer-specific differentially expressed lncRNAs and miRNAs, we searched the GEO database for miRNA expression profiling studies related to prostate cancer. The two miRNA expression datasets were downloaded from the National Center for Biotechnology Information GEO database (GSE76260 and GSE36802). All patients' records/information were anonymized and deidentified prior to the analysis. In total, 106 prostate clinical specimens (53 cancer and 53 non-neoplastic tissues/matched benign prostate tissues) were collected from GEO to create the data downloaded from 47 patients with prostate cancer in two different platforms, including an Affymetrix Multispecies miRNA-1 Array and Illumina Human v2 MicroRNA Expression BeadChip. We applied unpaired Student's t-tests to determine the expression differences between the groups. The differential expression values are displayed as a log of the fold-change. All analyses were performed with R statistical software. We predicted the candidate genes targeted by these miRNAs based on TargetScan (Whitehead Institute for Biomedical Research, Cambridge, MA, US) (Lewis et al., 2003) or miRecords (LC Sciences, Houston, TX, US) (Xiao et al., 2009). We also applied GEO2R to determine the involvement of dysregulated miRNAs in PCa and used the microRNA.org databases and the hypergeometric method to calculate the p-values in the miRNA target analysis. Furthermore, we analyzed the potential target microRNAs (miRNAs) of the differential lncRNAs online (<http://www.mircode.org/>). To understand the protein-lncRNA interactions of the differentially expressed lncRNAs, we constructed a lncRNA-mRNA network based on the transcripts. By analyzing the possible combination of lncRNAs and mRNAs, we predicted the target mRNAs of the differentially expressed lncRNAs (<http://starbase.sysu.edu.cn/starbase2/>) (Li et al., 2014) and generated a lncRNA-mRNA regulatory network map by Cytoscape3.5.1 software (Shannon et al., 2003).

Knockdown of Differentially Expressed Long Non-Coding Ribonucleic Acids in Prostate Cancer Cells

We applied si-RP11-423H2.3 and si-LAMTOR5-AS1 to knock down the expression of RP11-423H2.3 and LAMTOR5-AS1 in the prostate PC3 and DU145 cancer cells (the target sequence of RP11-423H2.3 was AAGGACAGCTTGCCTGACT; the target sequence of LAMTOR5-AS1 was CTGGTCTACTGTCACAACA; and siRNA-GFP was the control). All siRNAs were designed and synthesized by Ribobio (Guangzhou, China). qRT-PCR was applied to validate the transfection efficiency and expression level of relevant lncRNAs and target miRNAs. The siRNA with the best transfection efficiency was selected for subsequent experiments. Prostate PC3 and DU145 cancer cells were transfected with siRNAs at a concentration of 50 nM using 5 μ l of Lipofectamine 3000 (Invitrogen, CA, US) according to the manufacturer's protocol.

Quantitative Real-Time Polymerase Chain Reaction Analysis

Total RNA was isolated from 105 clinical specimens and prostate cells using TRIzol reagent (Invitrogen, Carlsbad, CA, USA). In total, 600 ng of total RNA per sample was used for complementary DNA (cDNA) synthesis using a PrimeScript™ RT Reagent Kit with gDNA Eraser (Takara, Cat#: RR047A, Japan). Real-time quantitative reverse transcription PCR (qRT-PCR) was performed with SYBR Premix Ex Taq™ II (Takara, Cat#: RR820A, Japan) using the LightCycler 480 II Instrument (Roche Molecular Systems, Inc). We performed qRT-PCR in a total reaction volume of 10 μ l, including 5 μ l of 2 $\times$ SYBR Green PCR buffer, 0.4 μ l of forward primer (10 μ M), 0.4 μ l of reverse primer (10 μ M), 0.2 μ l of ROX Reference Dye II, 3.5 μ l of ddH₂O, and 15 ng of cDNA. The reaction was initiated at 95°C for 1 min followed by 95°C (5 s) and 60°C (30 s) for 40 cycles. The expression of the lncRNAs was normalized to the level of 18S. The specific primers of the lncRNAs, miRNAs, and 18S are presented in **Table S1**. The data were collected and analyzed using the $2^{-\Delta\Delta Ct}$ method.

Immunoblots

Prostate PC3 and DU145 cancer cells were transfected with si-RP11-423H2.3, si-LAMTOR5-AS1, or siRNA-GFP (si-Control) using Lipofectamine 3000 (Invitrogen, CA, US) according to the manufacturer's protocol. After 72 h, protein samples were lysed in radioimmunoprecipitation assay (RIPA) buffer supplemented with protease inhibitors. Thirty micrograms of total protein were loaded per lane separated on a 10% sodium dodecyl sulfate (SDS)-polyacrylamide gel by electrophoresis, and proteins transferred onto nitrocellulose membranes. The membranes were blocked with 5% milk in phosphate buffered saline with tween 20 (PBST) and then incubated with a rabbit anti-UPF1 (Cat#: D161327, BBI Solutions) or rabbit anti-FUS (Cat#: D223360, BBI Solutions), or β -actin (N-21) rabbit polyclonal antibody (Cat#: sc-130656, Santa Cruz Biotechnology, Inc) at 4°C overnight. After washing with PBST, the blots were treated with a horseradish peroxidase (HRP)

conjugated anti-rabbit IgG. Detection of blots was performed using Meilunbio® fg super sensitive ECL luminescence reagent (Dalian Meilun Biotechnology Co., Ltd.) (Zhang et al., 2019).

Statistical Analyses

We collected clinical data from 105 prostate tissues, and a Student's t-test was used to analyze the differences in lncRNA expression between the prostate cancer group and BPH group. A Pearson correlation analysis was used to investigate the relationship between the differential lncRNAs and clinical parameters. The results were regarded as statistically significant at $p < 0.05$. All graphs were generated using GraphPad Prism 7.0 software (GraphPad Software Inc., La Jolla, CA, USA). The statistical analysis was performed using SPSS 22.0 (IBM-SPSS Inc., Chicago, IL, USA). Receiver operating characteristic (ROC) curves were applied to evaluate the clinical diagnostic value of the differential lncRNAs and the combination of PSA and lncRNAs.

RESULTS

Differential Profiling of Long Non-Coding Ribonucleic Acids in Prostate Cancer

To identify potential biomarkers of PCa, we first performed ceRNA microarray profiling of PCa patients and detected many transcripts in the PCa and adjacent normal tissues. We collected four pairs of tumor/adjacent normal tissue paraffin specimens and applied a ceRNA microarray to detect the transcripts in the PCa and adjacent normal tissues (Xia et al., 2018). A heatmap (**Figure 1A**) and scatter plots (**Figure 1B**) of the differential lncRNAs between the PCa tissues and normal tissues are shown in **Figure 1**. The heatmap indicates that 451 lncRNAs (**Figure 1A**) were differentially expressed with a fold change > 2.0 at $p < 0.05$. Among these lncRNAs, 217 lncRNAs were upregulated, and 234 lncRNAs were downregulated, in four pairs of PCa/adjacent tissues (**Table 1**). Among the differentially expressed lncRNAs, the most upregulated lncRNA is LINC00675, and the most downregulated lncRNA is RP11-864N7.4.

Validation of Key Differentially Expressed Long Non-Coding Ribonucleic Acids (RP11-33A14.1, RP11-423H2.3, and LAMTOR5-AS1) Using Fine Needle Aspiration Samples

We further carried out a qRT-PCR analysis of the related differential lncRNAs, including RP11-33A14.1, RP11-423H2.3, LAMTOR5-AS1, LINC00675, RP11-118K6.2, and RP11-423H2.3, in the normal prostate cell line HPEpic, PCa cells (22Rv1, DU145, LNCaP, and PC3 cells), and 105 FNA prostate tissues (48 PCa tissues and 57 BPH tissues) (**Figure S1**). The results revealed that the lncRNAs RP11-33A14.1 (**Figure 2A**), RP11-423H2.3 (**Figure 2B**), and LAMTOR5-AS1 (**Figure 2C**) were upregulated in the four PCa cells. We further validated these lncRNAs in 48 PCa tissues and 57 BPH tissues. The results showed that in the PCa tissues, the lncRNAs RP11-33A14.1, RP11-423H2.3, and LAMTOR5-AS1 were upregulated by

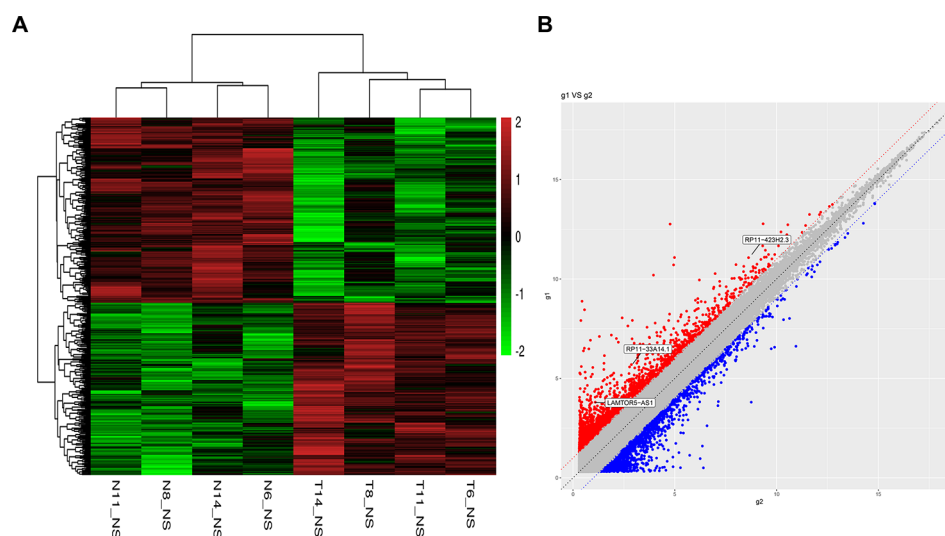


FIGURE 1 | Heatmap and scatter plots of differential long non-coding RNAs (lncRNAs) in prostate tumor tissues and normal tissues. **(A)** Heatmap of differential lncRNAs; **(B)** scatter plots of differential lncRNAs.

TABLE 1 | Top 10 of the differentially expressed long non-coding ribonucleic acids (lncRNA) in prostate cancer (PCa) (cancer/paracancerous tissue).

Accession	Gene symbol	Relation	Fold change	P values	FDR
NR_036581	LINC00675	intergenic	7.58	0.045	0.59
ENST00000439575	RP11-118K6.2	intergenic	7.41	0.023	0.55
ENST00000609245	LAMTOR5-AS1	intergenic	7.30	0.001	0.47
ENST00000414475	RP11-33A14.1	intergenic	6.96	0.049	0.59
lnc-NTM-4:1	—	intergenic	6.27	0.030	0.55
ENST00000503263	RP11-423H2.3	intergenic	5.46	0.024	0.55
ENST00000366189	RP11-423H2.3	intergenic	5.13	0.047	0.59
lnc-KAZALD1-1:1	—	intronic_sense	5.11	0.044	0.59
ENST00000605909	RP11-16D22.2	intergenic	4.71	0.012	0.51
ENST00000623288	RP11-423H2.4	intergenic	4.51	0.032	0.56
ENST00000371162	MIR4435-1HG	intergenic	-3.76	0.018	0.53
lnc-PRICKLE2-6:1	—	exonic_sense	-3.79	0.041	0.58
lnc-AC079135.1.1-8:1	—	exonic_sense	-3.92	0.032	0.56
lnc-ZDHHC13-5:1	—	exonic_sense	-4.19	0.027	0.55
ENST00000451884	MIR4435-1HG	intergenic	-4.19	0.003	0.47
lnc-JPH2-1:1	—	exonic_sense	-4.54	0.009	0.50
lnc-C15orf54-4:2	—	exonic_sense	-4.59	0.010	0.50
lnc-TPD52L3-1:1	—	exonic_sense	-5.31	0.013	0.51
NONHSAT018709	—	exonic_sense	-6.45	0.015	0.51
ENST00000624759	RP11-864N7.4	intronic_sense	-15.56	0.025	0.55

11.12 ± 3.66-fold (**Figure 2D**), 4.44 ± 1.87-fold (**Figure 2E**), and 1.89 ± 0.76-fold (**Figure 2F**), respectively ($p < 0.05$), further confirming the results of our ceRNA microarray.

Differentially Expressed Long Non-Coding Ribonucleic Acids as Novel Biomarkers of Prostate Cancer Associated With Prostate-Specific Antigens Levels and the Progression of Prostate Cancer

We assessed the diagnostic effectiveness of the differential lncRNAs in differentiating between PCa and BPH tissues by an

ROC curve (**Figure 3**). The areas under the curve (AUCs) of lncRNAs RP11-33A14.1, RP11-423H2.3, and LAMTOR5-AS1 were 0.697, 0.620, and 0.641, respectively (**Figure 3** and **Table 2**). When the three differential lncRNAs were combined, the AUC was 0.754 (**Figure 3D**). We further analyzed the PSA level using the results of the 3 differential lncRNAs, and the AUC was 0.984. The sensitivity was 97.9%, and the specificity was 84.2% (**Figure 3E**). To clarify the characteristics of these differential lncRNAs in PCa, we applied a Pearson correlation analysis to analyze the correlation between these lncRNAs and the corresponding clinical parameters. As shown in **Table 3**,

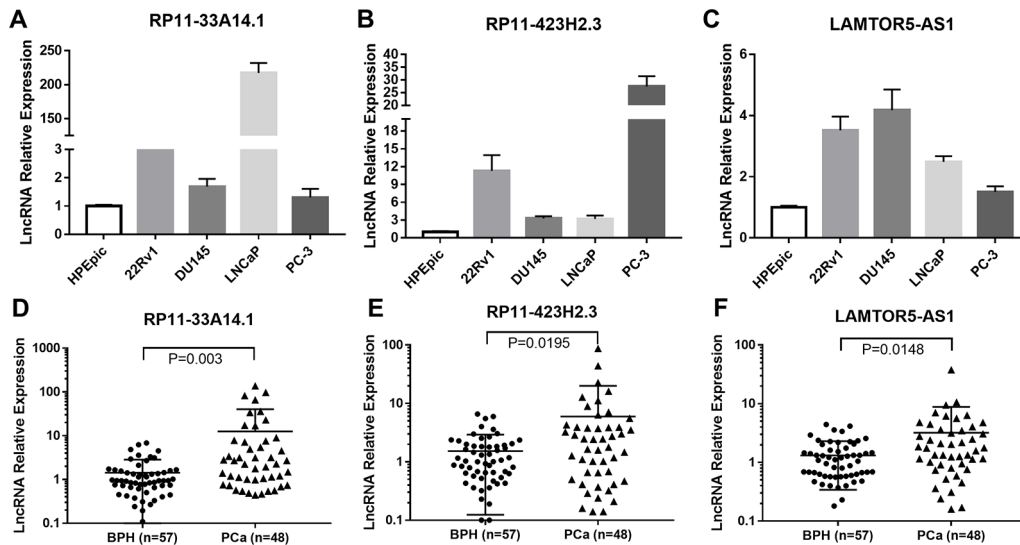


FIGURE 2 | Quantitative real-time (qRT)-PCR analysis of the gene expression levels of lncRNAs (RP11-33A14.1, RP11-423H2.3, and LAMTOR5-AS1) in prostate cells and tumor tissue fine needle aspiration (FNA) samples. RP11-33A14.1 (**A**, **D**), RP11-423H2.3 (**B**, **E**), and LAMTOR5-AS1 (**C**, **F**) in prostate cells (**A–C**), and tumor tissue samples compared to benign prostatic hyperplasia (BPH) tissue samples (**D–F**).

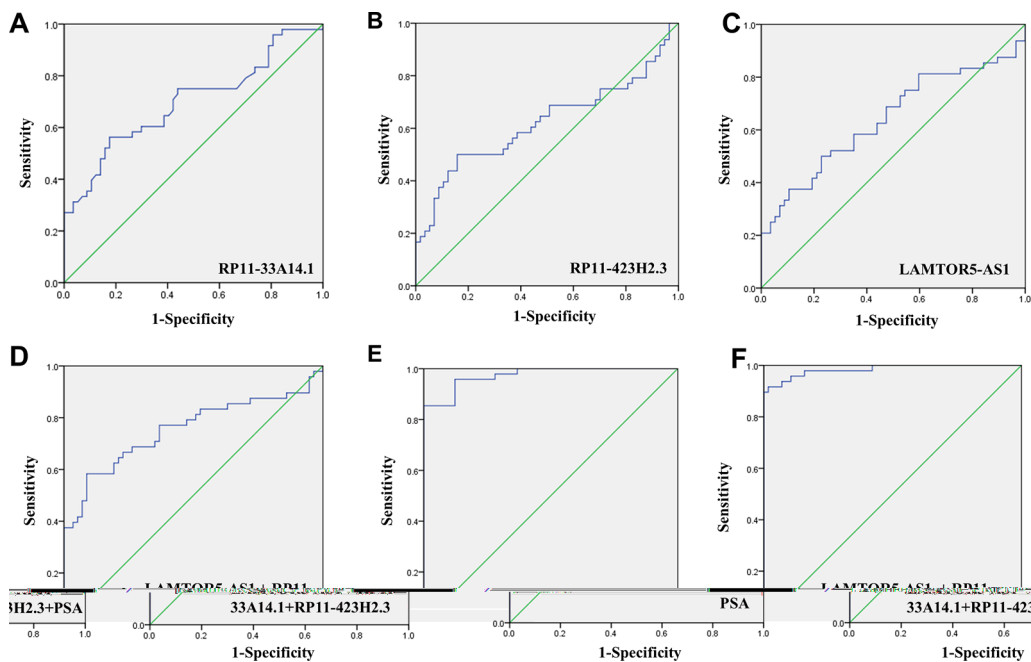


FIGURE 3 | ROC curve showing expression levels of differentially expressed long non-coding RNAs (lncRNAs). LncRNA RP11-33A14.1 (**A**), RP11-423H2.3 (**B**), and LAMTOR5-AS1 (**C**) in prostate cancer (PCa) patients and benign prostatic hyperplasia (BPH) controls; the three lncRNAs combination (**D**); prostate-specific antigen (PSA) only (**E**); and the three lncRNAs and PSA combination (**F**). The receiver operating characteristic (ROC) curves were analyzed using univariate (log-rank) analysis.

TABLE 2 | ROC analysis of the diagnostic efficiency of differential long non-coding ribonucleic acids (lncRNAs) (RP11-33A14.1, RP11-423H2.3, and LAMTOR5-AS1) and serum prostate-specific antigen (PSA) in prostate cancer (PCa) patients and benign prostatic hyperplasia (BPH) controls.

Biomarker	Sensitivity (%)	Specificity (%)	AUC (95% CI)	P-value
RP11-33A14.1	60.4	70.2	0.697 (0.506–0.734)	0.001
RP11-423H2.3	56.2	61.4	0.620 (0.506–0.734)	0.035
LAMTOR5-AS1	58.3	64.9	0.641 (0.531–0.751)	0.013
LAMTOR5-AS1 + RP11-33A14.1	77.1	63.2	0.754 (0.655–0.854)	<0.001
RP11-33A14.1 + RP11-423H2.3				
PSA	95.8	84.2	0.974 (0.946–0.998)	<0.001
LAMTOR5-AS1 + RP11-33A14.1	97.9	84.2	0.984 (0.964–1.004)	<0.001
RP11-33A14.1 + RP11-423H2.3 + PSA				

TABLE 3 | Association between the differential long non-coding ribonucleic acids (lncRNAs) and clinical parameter in prostate cancer (PCa) patients.

Clinical parameter	RP11-33A14.1, r (P)*	RP11-423H2.3, r (P)*	LAMTOR5-AS1, r (P)*
Age	0.165 (0.094)	0.324 (0.001)	0.258 (0.008)
PSA	0.025 (0.799)	0.347 (0.000)	0.803 (0.000)
Cholesterol (TC)	−0.158 (0.281)	−0.196 (0.370)	−0.299 (0.166)
Triglyceride (TG)	−0.161 (0.463)	0.021 (0.924)	−0.215 (0.324)
Fasting plasma glucose (FPG)	0.003 (0.990)	0.455 (0.077)	−0.179 (0.506)
Gleason score	0.020 (0.941)	−0.247 (0.091)	0.243 (0.096)

*Bold values denote statistical significance at the $p < 0.05$ level.

the lncRNA LAMTOR5-AS1 is positively correlated with the PSA level of the patients ($p < 0.001$). A combined Gleason score of 6 or 7 indicates that PCa is likely to grow but may not spread quickly. A score of 8–10 is suggestive of aggressive prostate cancer that is potentially lethal [24]. In this study, we investigated the association between the expression of lncRNA LAMTOR5-AS1 and aggressive cancer (Gleason score 8–10, $p < 0.05$) (Table 4) and found that lncRNA LAMTOR5-AS1 expression was higher in the less aggressive PCa (Gleason score 6–7; GS6–7) than in the aggressive PCa (Gleason score 8–10; GS8–10), yet its expression in GS8–10 was higher than in

TABLE 4 | Association between the differential long non-coding ribonucleic acids (lncRNAs) and aggressive prostate cancer (PCa).

Histologic diagnosis	lncRNA (mean $\pm$ SD)		
	RP11-33A14.1	RP11-423H2.3	LAMTOR5-AS1
Aggressive cancer (Gleason score 8–10)	1.19 $\pm$ 1.48	3.58 $\pm$ 6.23	1.92 $\pm$ 2.10
Less aggressive cancer (Gleason score 6–7)	1.55 $\pm$ 1.46	4.61 $\pm$ 8.12	3.42 $\pm$ 3.01
p-value	0.414	0.653	0.038

*Bold values denote statistical significance at the $p < 0.05$ level.

non-cancer tissues ($p = 0.023$) (Figure S2), which indicated that LAMTOR5-AS1 might be useful in the early diagnosis of PCa.

Regulatory Network Analysis of Differentially Expressed Long Non-Coding Ribonucleic Acids, Their Target Microribonucleic Acids, and Their Interaction With Ribonucleic Acids Binding Protein in Prostate Cancer

Subsequently, we predicted the miRNAs likely to be targeted by these three lncRNAs. In total, 100 miRNAs with binding sites for lncRNA RP11-33A14.1 and 47 miRNAs with binding sites for lncRNA RP11-423H2.3 were selected for subsequent analysis (Figures 4A, B). We also analyzed the miRNA expression profiles of GSE76260 and GSE36802 from the GEO databases. The microarray dataset GSE76260 included 32 prostate cancer and 32 non-neoplastic tissue samples; GSE36802 included 21 pairs of prostate cancer samples and matched benign prostate tissues. We identified 53 miRNAs that were differentially expressed between the prostate cancer tumor tissue and the normal controls. We found that compared with the normal controls, 28 miRNAs were upregulated (Figure 4A), and 25 miRNAs were repressed in the prostate cancer tissue samples (Figure 4B) in the two GEO datasets. Taken together, these results indicate that miR-7 predicted from lncRNAs RP11-33A14.1 and RP11-423H2.3 was upregulated in the prostate cancer tumor tissue in the two GEO datasets (Figure 4A). In contrast, two miRNAs (miR-24 and miR-30c) predicted from the two lncRNAs were repressed in the prostate cancer tumor tissue in the two GEO datasets (Figure 4B). Furthermore, we found that lncRNAs RP11-33A14.1 and RP11-423H2.3 both target miR-7, miR-24-3p, and miR-30 (Figure 4C). However, we only obtained two predicted miRNAs (miR-542-3p and miR-30c) for LAMTOR5-AS1 if we combined these two GEO datasets and utilized the miRDB database to identify target miRNAs. Next, we applied three reference datasets, DIANA-TarBase (<http://www.microrna.gr/tarbase>) (Karagkouni et al., 2018), lncRNASNP2 (<http://bioinfo.life.hust.edu.cn/lncRNASNP/#!/mirna/>), and miRDB (<http://www.mirdb.org/>), to predict the targeted miRNAs of LAMTOR5-AS1 and overlapped the three predicted results. Furthermore, we selected the top miRNAs (miR-550b-3p, miR-942-5p, miR-542-3p, miR-7162-3p, miR-4653, miR-3921, and miR-181b-3p) (Table S3) with the highest context scores (score > 70 in two predicted datasets) to establish a lncRNA-miRNA network for LAMTOR5-AS1 (Figure 4C). Finally, we analyzed the regulatory networks of lncRNAs RP11-33A14.1, RP11-423H2.3, and LAMTOR5-AS1 and predicted their potential RNA binding proteins (RBPs) using the starBase database. We found that lncRNAs RP11-423H2.3 and LAMTOR5-AS1 shared common RBPs, including eIF4AIII, U2AF65, and UPF1. More intriguingly, lncRNAs RP11-33A14.1, RP11-423H2.3, and LAMTOR5-AS1 interact with the same RBP FUS (Figure 4D).

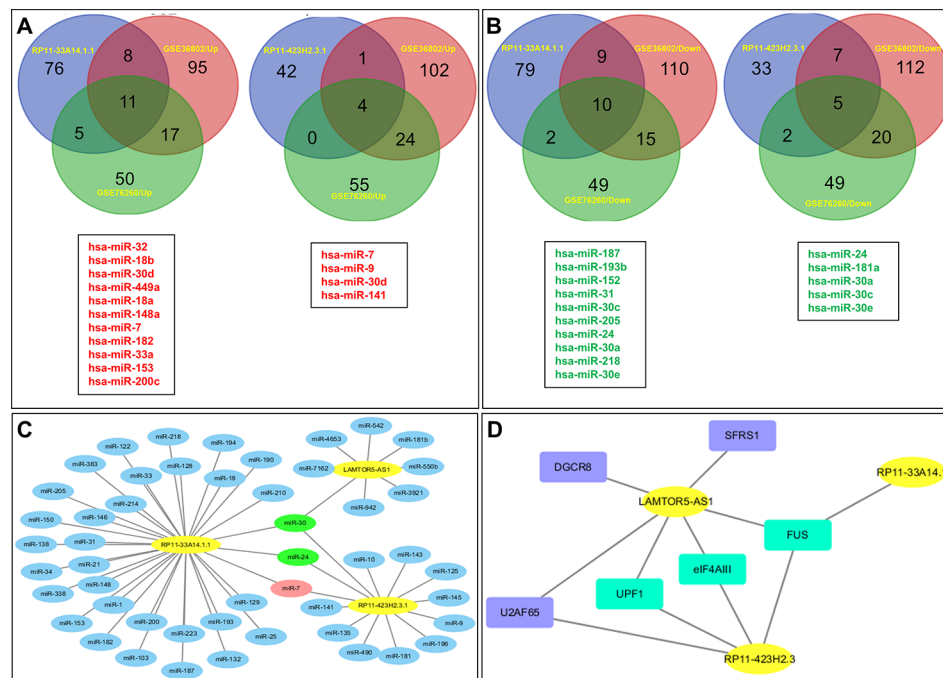


FIGURE 4 | Targeted microRNAs (miRNAs) of differentially expressed long non-coding RNAs (lncRNAs) in prostate cancer (PCa) and their regulatory network analysis. The Venn diagram demonstrates that the dysregulated miRNAs in PCa from the expression profiles of GSE76260 and GSE36802 in the GEO databases are the targeted miRNAs of lncRNAs in PCa (**A, B**), with upregulation in PCa (**A**) and downregulation in PCa (**B**); regulatory network analysis of differential lncRNAs, their targeted miRNAs (**C**); lncRNA RP11-423H2.3 and LAMTOR5-AS1 shared common RNA-binding proteins (**D**).

DISCUSSION

Prostate cancer is one of the most common cancers in men and ranges from low risk states amenable to active surveillance to high-risk states that can be lethal, especially if left untreated (Eskra et al., 2019). Although the diagnosis cornerstone of PCa has been prostate-specific antigen levels and numerous biomarkers have been introduced over the past decade, there is still a critical need for the development of relatively noninvasive and clinically useful methods for the screening, detection, prognosis, disease monitoring, and prediction of treatment efficacy of PCa.

Noncoding RNAs (ncRNAs) are typically classified into small and lncRNAs based on their size ranges of <200 or >200 nucleotides, and these RNAs are actively transcribed to a versatile group of RNA transcripts without protein-coding potential (over 80% of the genome) (Kapranov et al., 2007; Djebali et al., 2012). The dysregulation of lncRNAs has been implicated in the development and progression of a variety of cancers (Das et al., 2019). However, notably few lncRNAs have been functionally characterized and experimentally validated in PCa. In this study, the lncRNAs RP11-33A14.1, RP11-423H2.3, and LAMTOR5-AS1 were found to be upregulated in FNA biopsies of PCa. Several members of the lncRNA RP11 family are related to malignancies, including glioblastoma, renal cell carcinoma, and colorectal cancer. The lncRNA RP11-838N2.4

enhances the cytotoxic effects of temozolomide by inhibiting the functions of miR-10a in glioblastoma cell lines (Liu et al., 2016). The lncRNA RP11-436H11.5 functions as a ceRNA to upregulate BCL-W expression by sponging miR-335-5p, thereby promoting proliferation and invasion in renal cell carcinoma (Wang et al., 2017a). The downregulation of long noncoding RNA RP11-708H21.4 is associated with a poor prognosis in colorectal cancer and promotes tumorigenesis by regulating the AKT/mTOR pathway (Sun et al., 2017). RP11-380D23.2 drives the distal-proximal patterning of the lung by regulating PITX2 expression (Banerjee et al., 2018). The lncRNA LAMTOR5-AS1, which is known as late endosomal/lysosomal adaptor-2C MAPK and MTOR activator 5 (LAMTOR5) antisense RNA 1, was first shown to be associated with PCa in this report. Subsequently, we assessed the diagnostic effectiveness of differential lncRNAs in differentiating between PCa and BPH tissues. When the PSA level was combined with the three differential lncRNAs, the AUC was 0.984, the sensitivity was 97.9% and the specificity was 84.2%, which are better than the values obtained using PSA only. We previously demonstrated that different levels of two circRNAs (circ_0057558 and circ_0062019) and four genes DEGs (ITGBL1, TGM4, KRT15, and HOXA7) could help to distinguish PCa patients from non-PCa patients (Shan et al., 2017; Xia et al., 2018); thus, we proposed that combining these biomarkers might improve the diagnostic efficiency of PCa. We

demonstrated that when the expression of two circRNAs (circ_0057558 and circ_0062019) or 4 differentially expressed genes (DEGs) (ITGBL1, TGM4, KRT15, and HOXA7) were considered along with the three differentially expressed lncRNAs (DelncRNAs), the AUC was 0.935 (**Figure S3A**) and 0.968 (**Figure S3B**), the sensitivity was 85.0% and 93.8%, and the specificity was 89.2 and 92.7%, respectively. We also attempted to include only one gene (ITGBL1) and one circRNA (circ_0062019), which were the best biomarkers for the diagnosis of PCa in our previous publications, and found that when the expression of ITGBL1 and circ_0062019 was considered along with the three DelncRNAs, the AUC was 0.957 (**Figure S3C**), the sensitivity was 93.3%, and the specificity was 92.3% (**Table S2**), which were significantly improved compared to three lncRNAs. We also demonstrated that the lncRNA LAMTOR5-AS1 is positively correlated with the PSA level in patients and is more closely related to less aggressive PCa than to aggressive PCa, indicating that LAMTOR5-AS1 may be useful in the early diagnosis of PCa and that these differentially expressed lncRNAs might be novel biomarkers of PCa.

We further performed a regulatory network analysis of the differentially expressed lncRNAs and predicted that miR-7, miR-24, and miR-30 were target miRNAs of lncRNAs RP11-33A14.1 and RP11-423H2.3. Among these miRNAs, two miRNAs (miR-7 and 30d) were upregulated (**Figure 4A**), but four miRNAs (miR-24, miR-30a, miR-30c, and miR-30e) were repressed in the prostate cancer tumor tissue (**Figure 4B**) in the two GEO datasets. To determine if possible mechanisms of action that target miRNA expression were affected by these DelncRNAs, we knocked down RP11-423H2.3 and

LAMTOR5-AS1 in PCa cells. Our results revealed that knockdown of RP11-423H2.3 reduced the expression levels of miR-24-3p, miR-30a, miR-30d, and miR-30e and upregulated miR-7-1-3p in both PC3 and DU145 cells (**Figures 5A–C**). We also found that when LAMTOR5-AS1 was knocked down (**Figure 5D**), miR-942-5p, and miR-542-3p were repressed in PC3 cells (**Figure 5E**) but upregulated in DU145 cells (**Figure 5F**). In keeping with the ceRNA regulatory mechanism, lncRNAs can function as molecular decoys or sponges of microRNAs (Salmena et al., 2011), which might cause increased expression of miR-7-1-3p following knockdown of RP11-423H2.3. On other hand, some lncRNAs could also be processed to generate miRNAs or activate miRNA expression (Yoon et al., 2014), so that several miRNAs were deregulated after knockdown of RP11-423H2.3 or LAMTOR5-AS1, which supported the effects of RP11-423H2.3 on miR-7/miR-24/miR-30 or LAMTOR5-AS1 on miR-942-5p/miR-542-3p *via* direct interaction. miR-7 can inhibit the stemness of prostate cancer stem-like cells and tumorigenesis by repressing the KLF4/PI3K/Akt/p21 pathway (Chang et al., 2015). miR-24 serves as a tumor suppressor role in PCa and was repressed in prostate cancer cell lines and tumor tissue, which was correlated with high PSA serum levels and related to prostate cancer progression (Lynch et al., 2016). miR-30 was also downregulated in prostate cancer cells compared to that in the prostate immortalized normal epithelial-derived cell line RWPE-1, which may be associated with tumor suppressor functions in prostate cancer (Kao et al., 2014), and miR-30 has been identified as a direct regulator of androgen receptor signaling in prostate cancer by complementary functional microRNA library screening

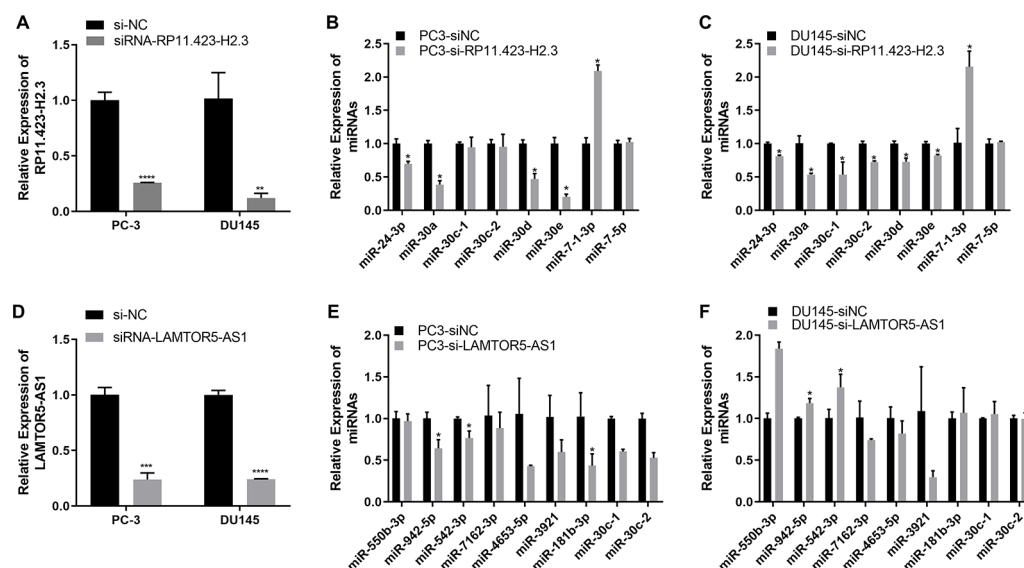


FIGURE 5 | Quantitative real-time (qRT-PCR) analysis of the gene expression levels of long non-coding RNAs (lncRNAs) and target microRNAs (miRNAs) in prostate cancer (PCa) cells with knockdown of RP11-423H2.3 or LAMTOR5-AS1. Knockdown of RP11-423H2.3 (**A–C**) and LAMTOR5-AS1 (**D–F**) in prostate cells; the expression levels of target miRNAs in PC3 cells (**B, E**) and DU145 cells (**C, F**). * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

(Kumar et al., 2016). miR-30a-5p and miR-30b were not only found to be lower in PCa tumors than in benign tissues but significantly increased when VCaP and PC3 cells were treated with saracatinib and PP2. However, miR-30c was different (Kao et al., 2014). miR-30b-3p and miR-30d-5p can be direct regulators of androgen receptor signaling in prostate cancer, and inhibition of miR-30b-3p and miR-30d-5p can increase androgen receptor (AR) expression and promote androgen-independent cell growth (Kumar et al., 2016). Finally, we determined that the lncRNAs RP11-423H2.3 and LAMTOR5-AS1 shared common RBPs, including eIF4AIII, U2AF65, and UPF1. Some lncRNAs can recruit regulatory compounds and affect gene expression by interacting with RBPs (Jia et al., 2017). The lncRNA MEG3 interacts with the RBP polypyrimidine tract-binding protein 1 (PTBP1) and induces cholestatic liver injury (Zhang et al., 2017). LncRNAs might affect the expression level of neighboring genes by a cis-regulated function. We found that all three lncRNAs, i.e., RP11-33A14.1, RP11-423H2.3, and LAMTOR5-AS1, interacted with FUS, while the loss of FUS expression may contribute to cancer progression (Brooke et al., 2011). The DNA and RNA helicase UPF1 played key roles in nonsense mediated RNA decay (NMD) that could selectively degrade aberrant RNA transcripts (Azzalin and Lingner, 2006). FUS was a multifunctional protein and participated in many RNA metabolism pathways, and mutant FUS suppressed protein biosynthesis and disrupted NMD regulation (Kamelgarn et al., 2018). FUS expression was also inversely correlated with Gleason grade of prostate cancer (Ghanbarpanah et al., 2018). We demonstrated that deregulation of FUS and UPF1 was in both PC3 and DU145 cells following knockdown of RP11-423H2.3 or LAMTOR5-AS1 (**Figure S4**), which implied that RBP FUS and UPF1 with lncRNA RP11-423H2.3 or LAMTOR5-AS1 interactions might affect prostate cancer progression. Deregulation of the RNA-binding protein fused in sarcoma/translocated in liposarcoma (FUS/TLS) in breast cancer cells by interacting with the lncRNA nuclear paraspeckle assembly transcript 1 (NEAT1) and miR-548ar could induce cell apoptosis (Wang et al., 2016). As FUS is a member of the TET protein family, this protein was found to be inversely regulated by miR-141 in human neuroblastoma (Wang et al., 2016) and can be activated by lncRNA XIST, which also served as a ceRNA in cervical cancer progression while competitively binding with miR-200a (Zhu et al., 2018). FUS promoted conditions that favored cell-cycle arrest by reducing proliferator factors and was a key link between androgen receptor signaling and the progression of the cell cycle in prostate cancer (Brooke et al., 2011; Ghanbarpanah et al., 2018).

CONCLUSIONS

While we continue to search for smarter and more reliable, precise, and cost-effective screening methods, we continue to

advocate shared decision-making in prostate cancer screening to serve our patients' best interests. The differentially expressed lncRNAs and their specific regulatory networks may serve as potential biomarkers for the clinical diagnosis and treatment of PCa, which could guide decisions regarding whom to biopsy and whom to re-biopsy after an initial negative biopsy with continued suspicion of PCa and might support an individual oncological approach in the future.

DATA AVAILABILITY STATEMENT

The prostate cancer microarray datasets were deposited in the Gene Expression Omnibus (GEO) database under accession number GSE140927.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of Zhongshan Hospital Affiliated with Fudan University and Shanghai Public Health Clinical Center. Written informed consent was obtained from all patients for the use of their tissue samples and clinical records.

AUTHOR CONTRIBUTIONS

JW and JZ planned overall concepts and designed the experiments. ZL, QX, XH, ZC, and DY performed the experiments. QX, ZL, HK and JW interpreted the data. XZ, TZ, JB and JX supported the study. ZL participated in drafting the manuscript. JW wrote and revised the manuscript.

ACKNOWLEDGMENTS

This research was supported by a grant from the National Natural Science Foundation of China (81672383, 81372318), a grant (2018ZX10302103-003) from the National Special Research Program of China for Important Infectious Diseases, China, and a grant (PWRL2017-07) supported by Pudong New District Commission of Health and Family Planning Leading Talent Program, Shanghai, China. The authors also want to thank Ms. Xiaoxiao Sun (Sinotech Genomics Co., Ltd., Shanghai, China) for microarray data analysis of our manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.00062/full#supplementary-material>

REFERENCES

- Azzalin, C. M., and Lingner, J. (2006). The double life of UPF1 in RNA and DNA stability pathways. *Cell Cycle* 5, 1496–1498. doi: 10.4161/cc.5.14.3093
- Banerjee, P., Surendran, H., Bharti, K., Morishita, K., Varshney, A., and Pal, R. (2018). Long noncoding RNA RP11-380D23.2 drives distal-proximal patterning of the lung by regulating PITX2 expression. *Stem Cells* 36, 218–229. doi: 10.1002/stem.2740
- Batista, P. J., and Chang, H. Y. (2013). Long noncoding RNAs: cellular address codes in development and disease. *Cell* 152, 1298–1307. doi: 10.1016/j.cell.2013.02.012
- Brooke, G. N., Culley, R. L., Dart, D. A., Mann, D. J., Gaughan, L., McCracken, S. R., et al. (2011). FUS/TLS is a novel mediator of androgen-dependent cell-cycle progression and prostate cancer growth. *Cancer Res.* 71, 914–924. doi: 10.1158/0008-5472.CAN-10-0874
- Carroll, P. R., Kellogg Parsons, J., Bahnson, R. R., Castle, E. P., Catalona, W. J., and Lurie, R. H. (2018). Clinical practice guidelines in oncology: prostate cancer early detection. *NCCN*. Version 2.2018.
- Chakravarty, D., Shoner, A., Nair, S. S., Giannopoulou, E., Li, R., Hennig, S., et al. (2014). The oestrogen receptor alpha-regulated lncRNA NEAT1 is a critical modulator of prostate cancer. *Nat. Commun.* 5, 5383. doi: 10.1038/ncomms5383
- Chang, Y. L., Zhou, P. J., Wei, L., Li, W., Ji, Z., Fang, Y. X., et al. (2015). MicroRNA-7 inhibits the stemness of prostate cancer stem-like cells and tumorigenesis by repressing KLF4/PI3K/Akt/p21 pathway. *Oncotarget* 6, 24017–24031. doi: 10.18632/oncotarget.4447
- Das, R., Feng, F. Y., and Selth, L. A. (2019). Long non-coding RNAs in prostate cancer: biological and clinical implications. *Mol. Cell Endocrinol.* 480, 142–152. doi: 10.1016/j.mce.2018.10.023
- Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., et al. (2012). Landscape of transcription in human cells. *Nature* 489, 101–108. doi: 10.1038/nature11233
- Du, Z., Sun, T., Hacisuleyman, E., Fei, T., Wang, X., Brown, M., et al. (2016). Integrative analyses reveal a long noncoding RNA-mediated sponge regulatory network in prostate cancer. *Nat. Commun.* 7, 10982. doi: 10.1038/ncomms10982
- Epstein, J. I. (1993). PSA and PAP as immunohistochemical markers in prostate cancer. *Urol. Clin. North Am.* 20, 757–770.
- Esakra, J. N., Rabizadeh, D., Pavlovich, C. P., Catalona, W. J., and Luo, J. (2019). Approaches to urinary detection of prostate cancer. *Prostate Cancer Prostatic Dis.* 22 (3), 362–381. doi: 10.1038/s41391-019-0127-4
- Ghanbarpanah, E., Kohanpour, M. A., Hosseini-Beheshti, F., Yari, L., and Keshvari, M. (2018). Structure and function of FUS gene in prostate cancer. *Bratislav Lek Listy* 119, 660–663. doi: 10.4149/BLL_2018_118
- Gu, P., Chen, X., Xie, R., Han, J., Xie, W., Wang, B., et al. (2017). lncRNA HOXD-AS1 regulates proliferation and chemo-resistance of castration-resistant prostate cancer via recruiting WDR5. *Mol. Ther.* 25, 1959–1973. doi: 10.1016/j.ymthe.2017.04.016
- Hon, C. C., Ramilowski, J. A., Harshbarger, J., Bertin, N., Rackham, O. J., Gough, J., et al. (2017). An atlas of human long non-coding RNAs with accurate 5' ends. *Nature* 543, 199–204. doi: 10.1038/nature21374
- Jia, L., Xi, Q., Wang, H., Zhang, Z., Liu, H., Cheng, Y., et al. (2017). miR-142-5p regulates tumor cell PD-L1 expression and enhances anti-tumor immunity. *Biochem. Biophys. Res. Commun.* 488, 425–431. doi: 10.1016/j.bbrc.2017.05.074
- Kamelgarn, M., Chen, J., Kuang, L., Jin, H., Kasarskis, E. J., and Zhu, H. (2018). ALS mutations of FUS suppress protein translation and disrupt the regulation of nonsense-mediated decay. *Proc. Natl. Acad. Sci. U. S. A.* 115, E11904–E11913. doi: 10.1073/pnas.1810413115
- Kao, C. J., Martiniez, A., Shi, X. B., Yang, J., Evans, C. P., Dobi, A., et al. (2014). miR-30 as a tumor suppressor connects EGF/Src signal to ERG and EMT. *Oncogene* 33, 2495–2503. doi: 10.1038/onc.2013.200
- Kapranov, P., Willingham, A. T., and Gingeras, T. R. (2007). Genome-wide transcription and the implications for genomic organization. *Nat. Rev. Genet.* 8, 413–423. doi: 10.1038/nrg2083
- Karakouni, D., Paraskevopoulou, M. D., Chatzopoulos, S., Vlachos, I. S., Tastsoglou, S., Kanellos, I., et al. (2018). DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA-gene interactions. *Nucleic Acids Res.* 46, D239–d245. doi: 10.1093/nar/gkx1141
- Kumar, B., Khaleghzadegan, S., Mears, B., Hatano, K., Kudrolli, T. A., Chowdhury, W. H., et al. (2016). Identification of miR-30b-3p and miR-30d-5p as direct regulators of androgen receptor signaling in prostate cancer by complementary functional microRNA library screening. *Oncotarget* 7, 72593–72607. doi: 10.18632/oncotarget.12241
- Lee, J. T. (2012). Epigenetic regulation by long noncoding RNAs. *Science* 338, 1435–1439. doi: 10.1126/science.1231776
- Lewis, B. P., Shih, I. H., Jones-Rhoades, M. W., Bartel, D. P., and Burge, C. B. (2003). Prediction of mammalian microRNA targets. *Cell* 115, 787–798. doi: 10.1016/S0092-8674(03)01018-3
- Li, J. H., Liu, S., Zhou, H., Qu, L. H., and Yang, J. H. (2014). starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.* 42, D92–D97. doi: 10.1093/nar/gkt1248
- Liu, Y., Xu, N., Liu, B., Huang, Y., Zeng, H., Yang, Z., et al. (2016). Long noncoding RNA RP11-838N2.4 enhances the cytotoxic effects of temozolomide by inhibiting the functions of miR-10a in glioblastoma cell lines. *Oncotarget* 7, 43835–43851. doi: 10.18632/oncotarget.9699
- Lynch, S. M., McKenna, M. M., Walsh, C. P., and McKenna, D. J. (2016). miR-24 regulates CDKN1B/p27 expression in prostate cancer. *Prostate* 76, 637–648. doi: 10.1002/pros.23156
- Misawa, A., Takayama, K. I., and Inoue, S. (2017). Long non-coding RNAs and prostate cancer. *Cancer Sci.* 108, 2107–2114. doi: 10.1111/cas.13352
- Mohler, J. L., Lee, R. J., Antonarakis, E. S., Armstrong, A. J., D'Amico, A. V., and Davis, B. J. (2018). Clinical practice guidelines in oncology: prostate cancer. *NCCN Version 4.2018*.
- Prensner, J. R., Iyer, M. K., Sahu, A., Asangani, I. A., Cao, Q., Patel, L., et al. (2013). The long noncoding RNA SChLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex. *Nat. Genet.* 45, 1392–1398. doi: 10.1038/ng.2771
- Salmena, L., Poliseno, L., Tay, Y., Kats, L., and Pandolfi, P. P. (2011). A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* 146, 353–358. doi: 10.1016/j.cell.2011.07.014
- Shan, M., Xia, Q., Yan, D., Zhu, Y., Zhang, X., Zhang, G., et al. (2017). Molecular analyses of prostate tumors for diagnosis of malignancy on fine-needle aspiration biopsies. *Oncotarget* 8, 104761–104771. doi: 10.18632/oncotarget.22289
- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303
- Siegel, R. L., Miller, K. D., and Jemal, A. (2017). Cancer Statistics 2017. *CA Cancer J. Clin.* 67, 7–30. doi: 10.3322/caac.21387
- Siegel, R. L., Miller, K. D., and Jemal, A. (2018). Cancer statistics 2018. *CA Cancer J. Clin.* 68, 7–30. doi: 10.3322/caac.21442
- Su, W., Xu, M., Chen, X., Chen, N., Gong, J., Nie, L., et al. (2017). Long noncoding RNA ZEB1-AS1 epigenetically regulates the expressions of ZEB1 and downstream molecules in prostate cancer. *Mol. Cancer* 16, 142. doi: 10.1186/s12943-017-0711-y
- Sun, L., Jiang, C., Xu, C., Xue, H., Zhou, H., Gu, L., et al. (2017). Down-regulation of long non-coding RNA RP11-708H21.4 is associated with poor prognosis for colorectal cancer and promotes tumorigenesis through regulating AKT/mTOR pathway. *Oncotarget* 8, 27929–27942. doi: 10.18632/oncotarget.15846
- Tan, G. H., Nason, G., Ajib, K., Woon, D. T. S., Herrera-Caceres, J., Alhunaiddi, O., et al. (2019). Smarter screening for prostate cancer. *World J. Urol.* doi: 10.1007/s00345-019-02719-5
- Van Der Toom, E. E., Axelrod, H. D., De La Rosette, J. J., De Reijke, T. M., Pienta, K. J., and Valkenburg, K. C. (2019). Prostate-specific markers to identify rare prostate cancer cells in liquid biopsies. *Nat. Rev. Urol.* 16, 7–22. doi: 10.1038/s41585-018-0119-5
- Walsh, A. L., Tuzova, A. V., Bolton, E. M., Lynch, T. H., and Perry, A. S. (2014). Long noncoding RNAs and prostate carcinogenesis: the missing 'linc'? *Trends Mol. Med.* 20, 428–436. doi: 10.1016/j.molmed.2014.03.005
- Wang, Z., Lei, H., and Sun, Q. (2016). MicroRNA-141 and its associated gene FUS modulate proliferation, migration and cisplatin chemosensitivity in neuroblastoma cell lines. *Oncol. Rep.* 35, 2943–2951. doi: 10.3892/or.2016.4640

- Wang, K., Jin, W., Song, Y., and Fei, X. (2017a). LncRNA RP11-436H11.5, functioning as a competitive endogenous RNA, upregulates BCL-W expression by sponging miR-335-5p and promotes proliferation and invasion in renal cell carcinoma. *Mol. Cancer* 16, 166. doi: 10.1186/s12943-017-0735-3
- Wang, R., Sun, Y., Li, L., Niu, Y., Lin, W., Lin, C., et al. (2017b). Preclinical study using malat1 small interfering RNA or androgen receptor splicing variant 7 Degradation enhancer ASC-J9((R)) to suppress enzalutamide-resistant prostate cancer progression. *Eur. Urol.* 72, 835–844. doi: 10.1016/j.eururo.2017.04.005
- Xia, Q., Ding, T., Zhang, G., Li, Z., Zeng, L., Zhu, Y., et al. (2018). Circular RNA expression profiling identifies prostate cancer-specific circRNAs in prostate cancer. *Cell Physiol. Biochem.* 50, 1903–1915. doi: 10.1159/000494870
- Xiao, F., Zuo, Z., Cai, G., Kang, S., Gao, X., and Li, T. (2009). miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res.* 37, D105–D110. doi: 10.1093/nar/gkn851
- Yoon, J. H., Abdelmohsen, K., and Gorospe, M. (2014). Functional interactions among microRNAs and long noncoding RNAs. *Semin. Cell Dev. Biol.* 34, 9–14. doi: 10.1016/j.semcdb.2014.05.015
- Zhang, A., Zhao, J. C., Kim, J., Fong, K. W., Yang, Y. A., Chakravarti, D., et al. (2015). LncRNA HOTAIR enhances the androgen-receptor-mediated transcriptional program and drives castration-resistant prostate cancer. *Cell Rep.* 13, 209–221. doi: 10.1016/j.celrep.2015.08.069
- Zhang, L., Yang, Z., Trottier, J., Barbier, O., and Wang, L. (2017). Long noncoding RNA MEG3 induces cholestatic liver injury by interaction with PTBP1 to facilitate shp mRNA decay. *Hepatology* 65, 604–615. doi: 10.1002/hep.28882
- Zhang, X., Liu, T., Li, Z., Feng, Y., Corpe, C., Liu, S., et al. (2019). Hepatomas are exquisitely sensitive to pharmacologic ascorbate (P-AscH⁻). *Theranostics* 9, 8109–8126. doi: 10.7150/thno.35378
- Zhu, H., Zheng, T., Yu, J., Zhou, L., and Wang, L. (2018). LncRNA XIST accelerates cervical cancer progression via upregulating Fus through competitively binding with miR-200a. *BioMed. Pharmacother.* 105, 789–797. doi: 10.1016/j.biopha.2018.05.053

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Li, Zheng, Xia, He, Bao, Chen, Katayama, Yu, Zhang, Xu, Zhu and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Computational Methods for the Integrative Analysis of Genomics and Pharmacological Data

Jimmy Caroli, Martina Dori and Silvio Bicciato *

Department of Life Sciences, University of Modena and Reggio Emilia, Modena, Italy

OPEN ACCESS

Edited by:

Davide Rizzo,
University of Padova, Italy

Reviewed by:

Nehme El-Hachem,
McGill University, Canada
Jun Zhong,
National Cancer Institute (NCI),
United States

*Correspondence:

Silvio Bicciato
silvio.bicciato@unimore.it

Specialty section:

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

Received: 05 December 2019

Accepted: 03 February 2020

Published: 27 February 2020

Citation:

Caroli J, Dori M and Bicciato S (2020)
Computational Methods for the
Integrative Analysis of Genomics and
Pharmacological Data.
Front. Oncol. 10:185.
doi: 10.3389/fonc.2020.00185

Since the pioneering NCI-60 panel of the late '80's, several major screenings of genetic profiling and drug testing in cancer cell lines have been conducted to investigate how genetic backgrounds and transcriptional patterns shape cancer's response to therapy and to identify disease-specific genes associated with drug response. Historically, pharmacogenomics screenings have been largely heterogeneous in terms of investigated cell lines, assay technologies, number of compounds, type and quality of genomic data, and methods for their computational analysis. The analysis of this enormous and heterogeneous amount of data required the development of computational methods for the integration of genomic profiles with drug responses across multiple screenings. Here, we will review the computational tools that have been developed to integrate cancer cell lines' genomic profiles and sensitivity to small molecule perturbations obtained from different screenings.

Keywords: genomics, pharmacogenomics, integration, bioinformatics, online databases

INTRODUCTION

Clinical responses to cancer treatment are strongly influenced by the patient's genomic landscape, pushing modern therapeutics toward a more personalized approach (1). To this end, despite their inability to reflect many aspects of a drug's behavior in the human body, cancer cell lines have been the most widely used models to explore the molecular basis of drug activity. Indeed, since the NCI-60 project, several major screenings of unite genetic profiling and drug testing have been created to investigate how genomic portraits can shape cancer response to therapy. These efforts required the definition of integrated frameworks that, leveraging on high-throughput technologies and computational methods, addressed the identification of genomic factors of cancer vulnerability associated with drug sensitivity. The NCI-60 project (https://ntp.cancer.gov/discovery_development/nci-60/) has been the first extensive screening of a massive number of chemical compounds (>50,000) on a well-defined set of cancer cell lines (60 across nine different tumoral tissues) (2, 3). Building on the NCI-60 approach, several other projects investigated the interplay between genomic backgrounds and responses to drug treatment in cancer cell lines (Figure 1A). All cancer cell line screenings basically adopt two approaches. In the first strategy, the molecular profiles of untreated cells and their response to various compounds are investigated in parallel to assess or predict how the molecular portraits determine intrinsic cell sensitivity and resistance to drugs or potential drugs. In the second, cell lines are profiled both before and after treatment to assess how their expression profiles respond to perturbation by the various agents tested. In particular, the Cancer Cell Line Encyclopedia (CCLE, <https://portals.broadinstitute.org/ccle>) project fully characterized the molecular profiles of more than 1,000 untreated cancer cell lines

along with their response to a panel of 24 Food and Drug Administration (FDA)-approved drugs (4–6). Similarly, the Genomics of Drug Sensitivity in Cancer (GDSC, <https://www.cancerrxgene.org>) and the Cancer Therapeutics Response Portal (CTRP, <http://portals.broadinstitute.org/ctrp/>) linked genomic features of more than 800 cancer cell lines to their sensitivity to hundreds of chemical compounds comprising FDA-approved drugs, clinical candidates, and small molecules (7–11). Conversely, the Connectivity Map (CMap) and its recent development, L1000 (CLUE, <https://clue.io>), profiled cancer cell lines before and after the treatment with several chemical compounds and genomic perturbagens, retrieving gene signatures directly associated to their administration (12–14). Although these screenings share a similar experimental pipeline, most of the produced data are heterogeneous and lack concordance in terms of investigated cell lines, tested compounds, and genomic information. In this review, we will describe some computational tools for the integrative analysis of data from different pharmacogenomics resources.

INTEGRATIVE ANALYSIS OF GENOMICS AND PHARMACOLOGICAL DATA

Inspired by the NCI-60 project, several collaborative efforts scaled up the number of cancer cell lines investigated in pharmacogenomics studies from the original 60 to more than 1,400, planning to reach over 10,000 publicly available cancer models in the near future (15). The massive amount of genomic and drug response data generated by these screenings are commonly collected in databases that, through dedicated web portals, provide direct insights into potential interactions between the analyzed cancer cell lines and the tested drugs. These databases are commonly equipped with computational resources specifically designed for the navigation and the analysis of the pharmacogenomics data, as for instance GDSCTools (16), CellMiner (17), Enrichr (18), L1000 Viewer (19), PharmacoGx, and PharmacoDB (20, 21), and the recently deployed RING (22). However, most of these tools are database specific and have limited capabilities in integrating data obtained from different screenings. This limitation is mostly due to the heterogeneity of data provided by the various studies, with drug tests not standardized across projects and genomic profiling not always available for the entire panel of cell lines. In addition, data are often unbalanced, with experiments comprising a high number of cell lines screened on few drugs (e.g., CCLE and GDSC) and, vice versa, screenings of large pools of chemical compounds performed on small cohorts of cancer cell lines (as in the NCI-60). Finally, while genomic data are rather homogeneous and can be easily integrated across studies after removing batch effects, pharmacological data derived from distinct experimental designs must be kept separate as they are profoundly different in terms of analytical assays, tested drug concentration, and retrieved inhibitory potential (23, 24). Despite these intrinsic limitations, several approaches have been proposed for the integrative analysis of genomics and pharmacological data collected from different screenings (Figure 1B). In particular, CellMinerCDB

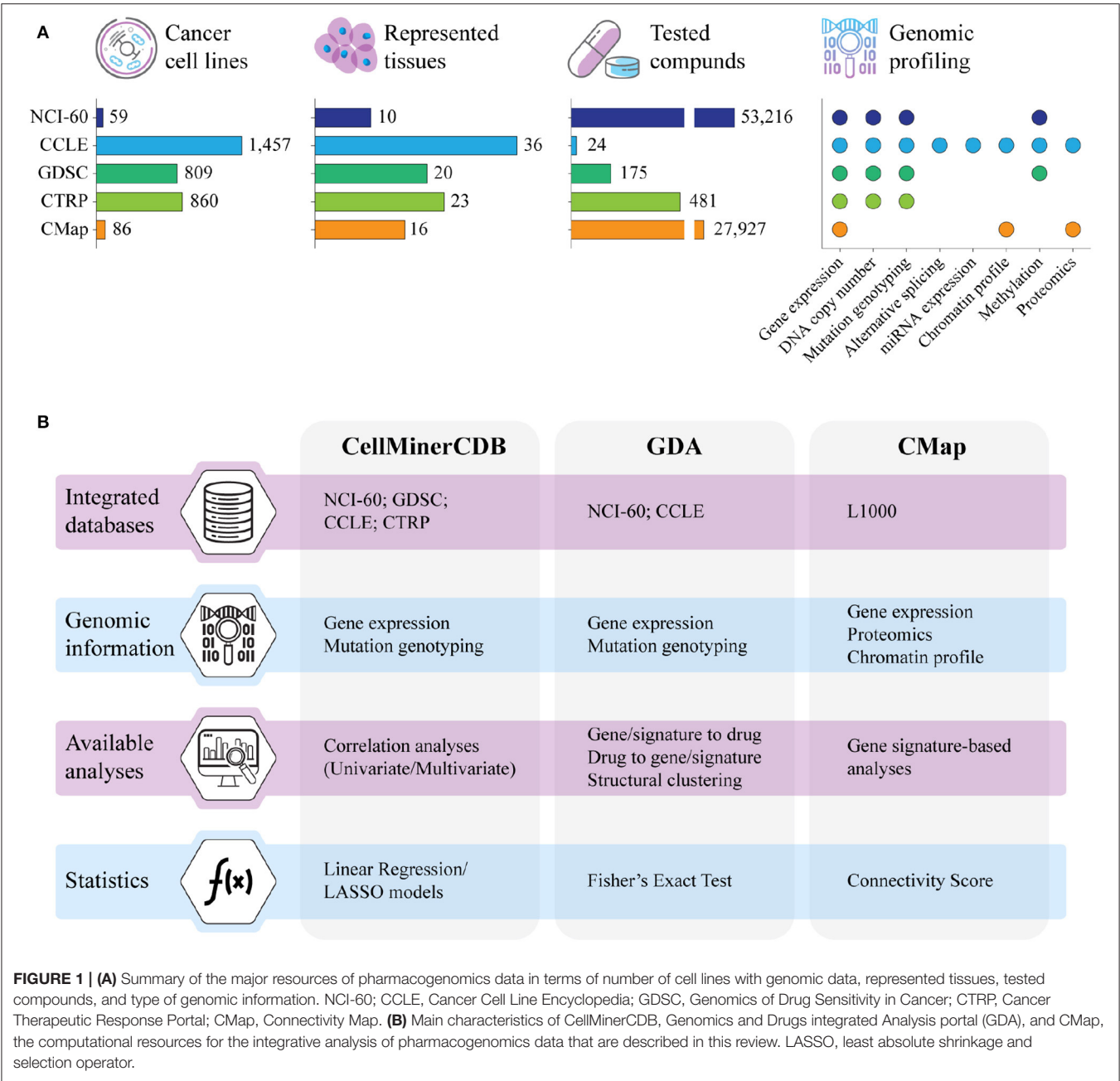
combines genomic profiles from NCI-60, CCLE, GDSC, and CTRP with the pharmacological data provided by the NCI-60 screening (25); the Genomics and Drugs integrated Analysis portal (GDA) integrates pharmacological data derived from the NCI-60 with the genomic information of NCI-60 and CCLE (26); and the CMap enables the investigation of the L1000 data through the correlation of gene lists and transcriptional signatures modulated by the drug treatment (12, 14, 27).

CellMinerCDB: Integrative Cross-Database Genomics and Pharmacogenomics Analyses

CellMinerCDB (<https://discover.nci.nih.gov/cellminerfdb/>) expands the analysis power of CellMiner, the original NCI-60 analysis tool, with the integration of the cancer cell line data from the Sanger/Massachusetts General Hospital GDSC, the Broad/Novartis CCLE, and the Broad CTRP (25, 28). The integrated database comprises all molecular profiles of almost 1,400 different cancer cell lines, together with drug activity for more than 20,000 compounds. The guiding element, used to link pharmacological information to genomic data from different sources, is the set of common cancer cell lines between the NCI-60 and the other resources, with 55 NCI-60 lines shared with GDSC, 44 with CCLE, and 671 in common between CCLE and GDSC. CellMinerCDB performs correlation analyses to investigate and visualize relationships between the drug activity of a compound and the specific profile of a selected molecular feature across all the available cell lines (univariate analysis). In addition, linear regression methods are implemented for the integrative analysis of multiple identifiers (multivariate analysis). The confidence of the associations is assessed by statistical analyses conducted through a basic linear regression model or using least absolute shrinkage and selection operator (LASSO). An interesting feature of CellMinerCDB is the possibility to compare patterns associated to either drug activity or molecular data via the *Compare Pattern* function of the univariate analysis search. This analysis allows the identification of genomic determinants of drug response, as exemplified by the connection found between the expression of Schlafen 11 (SLFN11) and the response to several DNA-targeted anticancer drugs as platinum derivatives, topoisomerase inhibitors, and poly (ADP-ribose) polymerase (PARP) inhibitors (25).

Genomics and Drugs Integrated Analysis

GDA (gda.unimore.it/) is a web-based tool designed for the integrative analysis of drug response, mutations, and gene expression profiles derived from the NCI-60 consortium and the CCLE (26, 29). GDA comprises 73 cancer cell lines shared by NCI-60 and CCLE and treated with 50,816 compounds and integrates the drug response data from the NCI-60 screening with the mutations and genomic information derived from both CCLE and NCI-60. GDA allows four different types of analyses, namely, *from drug to gene*, *from gene to drug*, *from signature to drug*, and *from drug to signature*. Pharmacological and genomic data can be queried to identify drugs correlated to gene mutations (from gene to drug), gene mutations associated



to drug responses (from drug to gene), and drugs associated to active gene signatures (from signature to drug). Starting from a drug correlated to gene mutations, gene expression profiles can be used to identify genes differentially expressed in cell lines sensitive to the selected compound. The statistics behind GDA is based on drug response data. Basically, all pairs of cell lines and drugs are defined as responsive if the relative sensitivity is smaller than two standard deviations of the left tail of the distribution of all relative sensitivities, and non-responsive otherwise. Based on genomic data, cell lines are classified as mutant if treated with the compound and carrying the selected set of mutations and as wild type if treated with the compound but without the specific set of mutations. Given these classifications, compounds are ranked using a score defined by

the fraction of responsive in mutant multiplied by the fraction of non-responders in wild type. This score ranks each drug based on the enrichment of responsive in the mutant group. The statistical significance of this ranking is computed using a one-tailed Fisher's exact test for the enrichment of responsive in mutant as compared to non-responsive in wild type, given the number of non-responsive in mutant and responsive in wild type. Results are accessible through interactive graphical representations and tables and can be directly fed to external tools as Enrichr for functional annotation (18). When used to identify compounds able to inhibit the proliferation potential of cancer cell lines with aberrant nuclear YAP/TAZ activation, GDA retrieved imatinib analogs and statins as potentially active drugs. Following GDA indications, *in vitro* studies demonstrated that

the combination of statins with dasatinib, an imatinib analog enhances YAP/TAZ nuclear exclusion, is able to block YAP/TAZ transcriptional activity, and is much more active in inducing apoptosis in different tissues (29).

Connectivity Map and the CMap Linked User Environment

CMap (<https://www.broadinstitute.org/connectivity-map-cmap>) was one of the first computational resources developed for the investigation of connections between transcriptomics and drug-induced perturbations (12). As extensively reviewed in Musa et al. (30), the goal of CMap is to identify drug or disease-associated gene signatures correlating with transcriptomics changes induced by the administration of drugs or chemical compounds (31, 32). The original project comprised the gene expression profiling of three cancer cell lines before and after the treatment with 164 different small molecules, obtaining drug-associated gene signatures for each cell line. This initial version has been recently scaled up through the L1000 Assay Platform, a method to analyze the expression levels of 978 selected landmark transcripts (assayed with 1,058 probes, including 80 controls) that have been shown to be sufficient to recover more than 80% of the information relative to the full transcriptome (14). This new approach translated into the screening of 86 different cancer cell lines using 27,927 unique perturbagens, including 19,811 small molecules and 7,494 genetic perturbations (consisting of overexpression or knockdown of different genes associated with human diseases or biological pathways). This large-scale screening finally resulted in a collection of 476,251 gene expression signatures that can be analyzed through the CMap Linked User Environment (CLUE, <https://clue.io>). In CLUE, the Query tool allows to input a gene signature (i.e., a list of genes upregulated and downregulated) and search for perturbagens (chemical and/or genetic) that induce a similar (or opposite) expression profile in the treated cells. The statistical significance of the association is assessed through a connectivity score that takes into account the strength of the similarity between the query and the induced signature as compared to the enrichment of all other signatures in the database (14). This approach proved its efficacy in the identification of a novel inhibitor for the serine-threonine kinase CSNK1A, an enzyme essential in specific subtypes of myelodysplastic syndrome and acute myeloid leukemia. Starting from the loss of function signature of CSNK1A1, authors searched CMap for compounds mimicking the loss of this kinase and identified one compound (BRD-1868) with a high connectivity score relative to this signature. Further enzymatic assays confirmed both the binding between BRD-1868 and CSNK1A1 and its inhibitory effect on enzymatic activity (14). From its first publication, CLUE has been expanded to include also proteomics analysis ranging from expression arrays to histone modification signatures.

CONCLUDING REMARKS

Efforts to decipher the molecular mechanisms of cancer stimulated scientists to explore the interconnection between

the genomic landscape of cancer models and their response to drug treatments. This resulted in large pharmacogenomics screenings that, with the advent of high-throughput technologies, generated large amounts of genomics and pharmacological data. However, the integration of these precious information is still challenging due to the variable type and number of drugs and cancer cell lines that have been screened by the various projects and the heterogeneous assays used for drug testing in the different studies (23, 24, 33–35). Despite these intrinsic difficulties, several computational approaches have been developed for the integrative analysis of genomics and pharmacological data. Their application allowed to discover several new connections between drug sensitivity and genomic backgrounds, enabling the potential repurposing of commercially available drugs to cancer treatment (36–38). However, these computational resources, although proven effective, still suffer the limitations of the original studies as the sparsity of the drug and cell interaction matrices, the effective impossibility to merge drug response data across different screenings, and the criticalities of cancer cell lines as a reliable cancer model (39–41). To this end, the project for a Patient-Derived Model Database (PDMB) launched in 2012 by the NCI might represent a potential breakthrough as genomic and drug response data directly collected from patients and patient-derived xenografts (PDXs) will reproduce more accurately the cancer disease and its environment than any cell line model (42). Furthermore, while novel experimental models are generating more accurate data, advanced computational methods are under development to enhance the analytical potential of existing algorithms. As recently discussed (43–45), artificial intelligence approaches as network-based models, deep-learning frameworks, and machine-learning techniques are increasingly applied to investigate pharmacogenomics connections and drug repositioning. These methods can be effective not only for data integration but also to predict new interactions and applications of already approved drugs (46–48). In summary, computational approaches for the integration of genomic and pharmacological data have the potential to become crucial for the systematic identification of new biomarkers of drug sensitivity and the discovery of novel anticancer drugs on the basis of specific genetic abnormalities, as long as reliable cellular models and highly curated data become available.

AUTHOR CONTRIBUTIONS

JC and SB conceived the project. JC, MD, and SB wrote and revised the manuscript.

FUNDING

This work was supported by funds from the Italian Association for Cancer Research (AIRC) Special Program Molecular Clinical Oncology 5 per mille (grant no. 10016) and from the Italian Epigenomics Flagship Project (Epigen) of the Italian Ministry of Education, University and Research.

REFERENCES

- Roden DM, George AL Jr. The genetic basis of variability in drug responses. *Nat Rev Drug Discov.* (2002) 1:37–44. doi: 10.1038/nrd705
- Shoemaker RH. The NCI60 human tumour cell line anticancer drug screen. *Nat Rev Cancer.* (2006) 6:813–23. doi: 10.1038/nrc1951
- Abaan OD, Polley EC, Davis SR, Zhu YJ, Bilke S, Walker RL, et al. The exomes of the NCI-60 panel: a genomic resource for cancer biology and systems pharmacology. *Cancer Res.* (2013) 73:4372–82. doi: 10.1158/0008-5472.CAN-12-3342
- Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature.* (2012) 483:603–7. doi: 10.1038/nature11003
- Ghandi M, Huang FW, Jane-Valbuena J, Kryukov GV, Lo CC, McDonald ER 3rd, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature.* (2019) 569:503–8. doi: 10.1038/s41586-019-1186-3
- Cancer Cell Line Encyclopedia Consortium and Genomics of Drugs Sensitivity in Cancer Consortium. Pharmacogenomic agreement between two cancer cell line data sets. *Nature.* (2015) 528:84–7. doi: 10.1038/nature15736
- Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, Lau KW, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature.* (2012) 483:570–5. doi: 10.1038/nature11005
- Basu A, Bodycombe NE, Cheah JH, Price EV, Liu K, Schaefer GI, et al. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell.* (2013) 154:1151–61. doi: 10.1016/j.cell.2013.08.003
- Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* (2013) 41:D955–61. doi: 10.1093/nar/gks1111
- Seashore-Ludlow B, Rees MG, Cheah JH, Cokol M, Price EV, Coletti ME, et al. Harnessing connectivity in a large-scale small-molecule sensitivity dataset. *Cancer Discov.* (2015) 5:1210–23. doi: 10.1158/2159-8290.CD-15-0235
- Rees MG, Seashore-Ludlow B, Cheah JH, Adams DJ, Price EV, Gill S, et al. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nat Chem Biol.* (2016) 12:109–16. doi: 10.1038/nchembio.1986
- Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science.* (2006) 313:1929–35. doi: 10.1126/science.1132939
- Lamb J. The Connectivity Map: a new tool for biomedical research. *Nat Rev Cancer.* (2007) 7:54–60. doi: 10.1038/nrc2044
- Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell.* (2017) 171:1437–52 e1417. doi: 10.1016/j.cell.2017.10.049
- Boehm JS, Golub TR. An ecosystem of cancer cell line factories to support a cancer dependency map. *Nat Rev Genet.* (2015) 16:373–4. doi: 10.1038/nrg3967
- Cokelaer T, Chen E, Iorio F, Menden MP, Lightfoot H, Saez-Rodriguez J, et al. GDSCTools for mining pharmacogenomic interactions in cancer. *Bioinformatics.* (2018) 34:1226–8. doi: 10.1093/bioinformatics/btx744
- Reinhold WC, Sunshine M, Liu H, Varma S, Kohn KW, Morris J, et al. CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. *Cancer Res.* (2012) 72:3499–511. doi: 10.1158/0008-5472.CAN-12-1370
- Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinform.* (2013) 14:128. doi: 10.1186/1471-2105-14-128
- Musa A, Tripathi S, Dehmer M, and Emmert-Streib F. L1000 viewer: a search engine and web interface for the LINCS data repository. *Front Genet.* (2019) 10:557. doi: 10.3389/fgene.2019.00557
- Smirnov P, Safikhani Z, El-Hachem N, Wang D, She A, Olsen C, et al. PharmacGx: an R package for analysis of large pharmacogenomic datasets. *Bioinformatics.* (2016) 32:1244–6. doi: 10.1093/bioinformatics/btv723
- Smirnov P, Kofia V, Maru A, Freeman M, Ho C, El-Hachem N, et al. PharmacDB: an integrative database for mining *in vitro* anticancer drug screening studies. *Nucleic Acids Res.* (2018) 46:D994–1002. doi: 10.1093/nar/gkx911
- Politano G, Di Carlo S, Benso A. “One DB to rule them all”—the RING: a Regulatory INteraction Graph combining TFs, genes/proteins, SNPs, diseases and drugs. *Database.* (2019) 2019:108. doi: 10.1093/database/baz108
- Haibe-Kains B, El-Hachem N, Birkbak NJ, Jin AC, Beck AH, Aerts HJ, et al. Inconsistency in large pharmacogenomic studies. *Nature.* (2013) 504:389–93. doi: 10.1038/nature12831
- Weinstein JN, Lorenzi PL. Cancer: discrepancies in drug sensitivity. *Nature.* (2013) 504:381–3. doi: 10.1038/nature12839
- Rajapakse VN, Luna A, Yamade M, Loman L, Varma S, Sunshine M, et al. CellMinerCDB for integrative cross-database genomics and pharmacogenomics analyses of cancer cell lines. *Science.* (2018) 10:247–64. doi: 10.1016/j.jsci.2018.11.029
- Caroli J, Sorrentino G, Forcato M, Del Sal G, Bicciato, S. GDA, a web-based tool for Genomics and Drugs integrated analysis. *Nucleic Acids Res.* (2018) 46:W148–56. doi: 10.1093/nar/gky434
- Duan Q, Reid SP, Clark NR, Wang Z, Fernandez NF, Rouillard AD, et al. L1000CDS(2): LINCS L1000 characteristic direction signatures search engine. *NPJ Syst Biol Appl.* (2016) 2:15. doi: 10.1038/npsba.2016.15
- Polley E, Kunkel M, Evans D, Silvers T, Delosh R, Laudeman J, et al. Small cell lung cancer screen of oncology drugs, investigational agents, and gene and microRNA expression. *J Natl Cancer Inst.* (2016) 108. doi: 10.1093/jnci/djw122
- Taccioli C, Sorrentino G, Zannini A, Caroli J, Beneventano D, Anderlucci L, et al. MDP, a database linking drug response data to genomic information, identifies dasatinib and statins as a combinatorial strategy to inhibit YAP/TAZ in cancer cells. *Oncotarget.* (2015) 6:38854–65. doi: 10.18632/oncotarget.5749
- Musa A, Ghorai LS, Zhang SD, Glazko G, Yli-Harja O, Dehmer M, et al. A review of connectivity map and computational approaches in pharmacogenomics. *Brief Bioinform.* (2018) 19:506–23. doi: 10.1093/bib/bbw112
- Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, et al. Functional discovery via a compendium of expression profiles. *Cell.* (2000) 102:109–26. doi: 10.1016/S0092-8674(00)00015-5
- Dudley JT, Sirota M, Shenoy M, Pai RK, Roedder S, Chiang AP, et al. Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. *Sci Transl Med.* (2011) 3:96ra76. doi: 10.1126/scitranslmed.3002648
- Papillon-Cavanagh S, De Jay N, Hachem N, Olsen C, Bontempi G, Aerts HJ, et al. Comparison and validation of genomic predictors for anticancer drug sensitivity. *J Am Med Inform Assoc.* (2013) 20:597–602. doi: 10.1136/amiajnl-2012-001442
- Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, et al. A landscape of pharmacogenomic interactions in cancer. *Cell.* (2016) 166:740–54. doi: 10.1016/j.cell.2016.06.017
- Safikhani Z, Smirnov P, Freeman M, El-Hachem N, She A, Rene Q, et al. Revisiting inconsistency in large pharmacogenomic studies. *F1000Res.* (2016) 5:2333. doi: 10.12688/f1000research.9611.1
- Kim JH, Scialli AR. Thalidomide: the tragedy of birth defects and the effective treatment of disease. *Toxicol Sci.* (2011) 122:1–6. doi: 10.1093/toxsci/kfr088
- Pritchard JLE, O'Mara TA, Glubb DM. Enhancing the promise of drug repositioning through genetics. *Front Pharmacol.* (2017) 8:896. doi: 10.3389/fphar.2017.00896
- Sabatine MS, Giugliano RP, Keech AC, Honarpour N, Wiviott SD, Murphy SA, et al. Evolocumab and clinical outcomes in patients with cardiovascular disease. *N Engl J Med.* (2017) 376:1713–22. doi: 10.1056/NEJMoa1615664
- Mullard A. Reliability of 'new drug target' claims called into question. *Nat Rev Drug Discov.* (2011) 10:643–4. doi: 10.1038/nrd3545
- Ben-David U, Siranosian B, Ha G, Tang H, Oren Y, Hinohara K, et al. Genetic and transcriptional evolution alters cancer cell line drug response. *Nature.* (2018) 560:325–30. doi: 10.1038/s41586-018-0409-3
- Mullard A. Can you trust your cancer cell lines? *Nat Rev Drug Discov.* (2018) 17:613. doi: 10.1038/nrd.2018.154
- Mer AS, Ba-Alawi W, Smirnov P, Wang YX, Brew B, Ortmann J, et al. Integrative pharmacogenomics analysis of patient-derived xenografts. *Cancer Res.* (2019) 79:4539–50. doi: 10.1158/0008-5472.CAN-19-0349

43. Kalinin AA, Higgins GA, Reamaroon N, Soroushmehr S, Allyn-Feuer A, Dinov ID, et al. Deep learning in pharmacogenomics: from gene regulation to patient stratification. *Pharmacogenomics*. (2018) 19:629–50. doi: 10.2217/pgs-2018-0008
44. Chiu YC, Chen HHH, Zhang T, Zhang S, Gorthi A, Wang LJ, et al. Predicting drug response of tumors from integrated genomic profiles by deep neural networks. *BMC Med Genom*. (2019) 12(Suppl.1):18–18. doi: 10.1186/s12920-018-0460-9
45. Sakellaropoulos T, Vougas K, Narang S, Koinis F, Kotsinas A, Polyzos A, et al. A deep learning framework for predicting response to therapy in cancer. *Cell Rep*. (2019) 29:3367–73.e3364. doi: 10.1016/j.celrep.2019.11.017
46. Menden MP, Iorio F, Garnett M, McDermott U, Benes CH, Ballester PJ, et al. Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS ONE*. (2013) 8:e61318. doi: 10.1371/journal.pone.0061318
47. Costello JC, Heiser LM, Georgii E, Gönen M, Menden MP, Wang NJ, et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nat Biotechnol*. (2014) 32:1202–12. doi: 10.1038/nbt.2877
48. Kalamara A, Tobalina L, Saez-Rodriguez J. How to find the right drug for each patient? *Adv Challenges Pharmacogenom*. (2018) 10:53–62. doi: 10.1016/j.coisb.2018.07.001

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Caroli, Dori and Biciato. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Genomic and Transcriptomic Landscape of Tumor Clonal Evolution in Cholangiocarcinoma

Geng Chen¹, Zhixiong Cai¹, Xiuqing Dong², Jing Zhao¹, Song Lin¹, Xi Hu¹, Fang-E Liu³, Xiaolong Liu² and Huqing Zhang^{1*}

¹ School of Life Sciences and Technology, Xi'an Jiaotong University, Xi'an, China, ² The United Innovation of Mengchao Hepatobiliary Technology Key Laboratory of Fujian Province, Mengchao Hepatobiliary Hospital of Fujian Medical University, Fuzhou, China, ³ Department of Nursing, School of Medicine, Xi'an Peihua University, Xi'an, China

OPEN ACCESS

Edited by:

Enrica Calura,
University of Padova, Italy

Reviewed by:

Xueqiu Lin,
Stanford University, United States
Jun Zhong,
National Cancer Institute (NCI),
United States

*Correspondence:

Huqing Zhang
huqzhang@mail.xjtu.edu.cn

Specialty section:

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Genetics

Received: 11 June 2019

Accepted: 19 February 2020

Published: 13 March 2020

Citation:

Chen G, Cai Z, Dong X, Zhao J,
Lin S, Hu X, Liu F-E, Liu X and
Zhang H (2020) Genomic
and Transcriptomic Landscape
of Tumor Clonal Evolution
in Cholangiocarcinoma.
Front. Genet. 11:195.
doi: 10.3389/fgene.2020.00195

Cholangiocarcinoma remained a severe threat to human health. Deciphering the genomic and/or transcriptomic profiles of tumor has been proved to be a promising strategy for exploring the mechanism of tumorigenesis and development, which could also provide valuable insights into Cholangiocarcinoma. However, little knowledge has been obtained regarding to how the alteration among different omics levels is connected. Here, using whole exome sequencing and transcriptome sequencing, we performed a thorough evaluation for the landscape of genome and transcriptome in cholangiocarcinoma and illustrate the alteration of tumor on different biological levels. Meanwhile, we also identified the clonal structure of each included tumor sample and discovered different clonal evolution patterns related to patients' survival. Furthermore, we extracted subnetworks that were greatly influenced by tumor clonal/subclonal mutations or transcriptome change. The topology relationship between genes affected by genomic/transcriptomic changes in biological interaction networks revealed that alteration of genome and transcriptome was highly correlated, and somatic mutations located on important genes might affect the expression of numerous genes in close range.

Keywords: cholangiocarcinoma, clonal evolution, sequencing, transcriptome, genome

INTRODUCTION

Cholangiocarcinoma (CCA), a heterogeneous malignant tumor currently acknowledged as the second most common primary liver cancer, showed increasing incidents worldwide during past decades. Although CCA is considered as a rare cancer in most countries due to its relative low incidents (lower than 6 cases per 100,000 people), the situations are different in several countries including China and Thailand, where CCA incident reaches an exceptionally high level. Among all CCA cases, intrahepatic cholangiocarcinoma takes up only 10%, while a minority (15%) of these patients were diagnosed with resectable disease status (Cardinale et al., 2018; Rizvi et al., 2018). While the most promising therapeutic strategy for CCA is surgical operation combined with chemo-/radio-therapy, this approach was considered only suitable for early stage CCA and later stage CCA patients often face the difficulty of lacking effective treatment options. Thus, most CCA patients usually suffered from poor prognosis (5-year survival rate less than 10%). Meanwhile,

the heterogeneity of tumor on multiple levels (e.g., genomic, transcriptional) often resulted in resistance to therapy, which further intensifies the challenge of CCA treatments. Thus, a thorough evaluation of the landscape on CCA genome and transcriptome could provide clinically related insights into the genesis and progression of CCA.

Just like other tumors, CCA is developed on the basis of acquiring tumor somatic mutations and clonal evolution. When tumor arises and progresses, the acquisition of somatic mutations randomly happened, resulting in different groups of tumor cells with distinct genetic features. The tumor clone, built up with the complicated constitution of groups of tumor cells (which could be referred as subclones), evolves during its development, dynamically changing its structure to better fit the micro-environment (Greaves and Maley, 2012; McGranahan and Swanton, 2017). During this entirely evolutionary process, certain somatic mutations could give tumor cells survival advantage and subpopulation carrying these genomic alterations expanded, while subclones with mutations reducing survival capacity diminished. Thus, deciphering the clonal evolution in CCA could provide valuable information regarding crucial genetic events in tumorigenesis and progression and how different biological pathways might be affected by these genetic events, which in turn could help further understand the intrinsic mechanisms of tumor progression. Indeed, such efforts have been made in other types of cancer including leukemia (Ferrando and López-Otín, 2017) and solid tumors such as hepatocellular carcinoma (Chen et al., 2018) and breast cancer (Hoadley et al., 2016), and different clonal evolution patterns have been discovered with high correlation with patients' clinical course.

However, the evolutionary process in CCA still requires further investigation. What more, although the importance of clonal evolution is widely acknowledged, how tumor clonal structure affects tumor transcriptome remained poorly explored. Understanding how somatic mutation interacted with such transcriptome change could further provide valuable insights into the evolutionary mechanism of CCA development. To explore the genetic and transcriptional landscape of intrahepatic CCA, we performed whole exome sequencing and transcriptome sequencing on tumor and corresponding peritumor tissue of 9 CCA patients. The differences on genetic and transcriptional levels were investigated and tumor clonal evolution was deciphered to discover the molecular pathways taking part in the deregulation of tumor cells. These findings will be of great value in understanding the mechanism of CCA development and how transcriptome interact with genetic alterations.

MATERIALS AND METHODS

Sample Collection

Tumor and corresponding peritumor tissue samples were collected from 9 patients diagnosed with intrahepatic cholangiocarcinoma during their surgical operation for tumor removal. The detailed clinical information is provided in **Table 1**. All human tissue sample collection procedures and usage of these samples were approved by the Institution Review Board of

TABLE 1 | Clinical characteristics of 9 enrolled CCA patients.

Clinicopathological variables	Patient number (<i>n</i> = 9)	Percentage (100%)
Sex		
Male	5	55.6
Female	4	44.4
Age at first enrolled year, Mean ± SD	62.44 ± 11.78	
HBV infection		
Negative	6	66.7
Positive	3	33.3
HBV DNA		
≤10 ³	7	77.8
10 ³ –10 ⁴	1	11.1
10 ⁴ –10 ⁵	1	11.1
Maximal tumor size, cm		
0–2.5	2	22.2
2.5–5.0	2	22.2
5.0–10	5	55.6
Tumor number		
Single	8	88.9
Multiple	1	11.1
Liver cirrhosis		
Absent	5	55.6
Present	4	44.4
Microvascular invasion		
Yes	3	33.3
No	6	66.7
PVTT		
Yes	1	11.1
No	8	88.9
Microsatellite lesion		
Absent	8	88.9
Present	1	11.1
TNM		
I	5	55.6
II	1	11.1
IV	3	33.3
BCLC		
0	1	11.1
A	4	44.5
B	1	11.1
C	3	33.3

PVTT, Portal vein tumor thrombosis; TNM, The TNM Classification of Malignant Tumors; BCLC, the Barcelona Clinic Liver Cancer staging system.

Mengchao Hepatobiliary Hospital of Fujian Medical University and written consents were obtained from all participated patients included in this study.

Whole Exome/Transcriptome Sequencing

Whole-exome and transcriptome sequencing were performed to capture the genetic and transcriptional features for the acquired tumor and corresponding peritumor tissue on Illumina HiSeq 3000 system.

Whole Exome Sequencing Data Processing

Somatic single nucleotide variants (SNV) and copy number alterations (CNA) were detected for the whole exome sequencing data of tumor tissue samples using the corresponding peritumor as control. To identify SNVs, SomaticSniper (version 1.0.5.0) (Larson et al., 2012) were applied using default parameters provided in the algorithm manual and only SNVs with somatic score ≥ 40 were accepted for downstream analysis. The identified SNVs were further filtered with such criteria to rule out possible false discovery: (1) read depth ≥ 50 in both tumor and peritumor tissues; (2) variant allele frequency $\geq 10\%$ in tumor tissue; (3) variant allele frequency $< 10\%$ in normal peritumor tissues. The detected SNVs were then annotated using wANNOVAR to obtain related gene and functional information. For CNVs, TitanCNA (version 1.17.1) (Ha et al., 2014) was applied on the tumor tissue's whole exome sequencing data using the corresponding peritumor as control using the workflow script provided by the algorithm.

Transcriptome Sequencing Data Processing

All acquired Transcriptome sequencing reads were first aligned to ribosomal rRNA sequences to remove ribosomal RNA sequence. The unmapped reads were then aligned to human genome reference (GRCH37) using star with GENCODE gene annotation. The gene expression was quantified with fragments per kilobase of exon per million mapped fragments (FPKM) and genes with no read counts in $> 50\%$ samples were not included in downstream analysis. Differentially expressed genes were identified using limma package. Genes with adjusted p value < 0.05 (Benjamini-Hochberg correction) and fold-change > 2 or < 0.5 were then considered as significantly differentially expressed between CCA tumor and peritumor.

Clonal Evolution in CCA

For each CCA tumor sample, inference of subclonal population was conducted using Sclust (Cun et al., 2018). Sclust provided a copy-number analysis method incorporated with mutational clustering to accurately determines copy-number states and subclonal populations. In brief, whole exome sequencing data of the paired tumor and peritumor samples were first processed using command bam process to extract the read ratio and SNP information. Then, the copy number analysis is conducted with command cn for each patient, using the obtained read ratio and SNP information together with SomaticSniper mutation calling results. Finally, the mutational clustering was performed using command cluster based on above results to identify tumor clonal structure.

Discovery of Altered Subnetworks Influenced by Somatic Mutations and Transcriptome Change

HotNet2 was applied to discover altered subnetworks in the large gene interaction networks. HotNet2 required two input files for subnetwork identification: Heat scores and Interaction

network. For somatic mutations, Heat scores for HotNet2 were generated based on mutation distribution across all patients; For transcriptome, Heat scores were generated based on the adjusted p -value produced by DESeq2 package. Network hint + hi2012 and irefindex9 provided by HotNet2 was used as the Interaction network for this analysis. The algorithm was run using all recommended parameters provided by algorithm authors and the identified subnetworks were visualized using Cytoscape (version 3.4.0) (Shannon et al., 2003).

RESULTS

Case Summary

In total, 9 patients that were diagnosed with CCA and received surgical operation in Mengchao Hepatobiliary Hospital were included in this study. According to previous reports regarding inflammatory context of liver tumors (Bishayee, 2014; Banales et al., 2016), we chose peritumor tissue as sequencing control to better capture the CCA characteristics. During their surgery, cholangiocarcinoma tumor tissues along with corresponding peritumor tissues were collected and the tumor existence for all patients was histologically confirmed. Then, whole-exome and transcriptome sequencing were performed for acquired tissue samples. Among all included patients, 77.8% (7/9) were diagnosed with TNM staging I-II and the other 22.2% were diagnosed with TNM staging III. The average diameter of tumor in each patient was 5.1 cm (range, 2.0–9.5 cm), while Vascular tumor thrombus was seen in 44.4% (4/9) of all patients. Detailed clinical information for all included patients before they received surgical operation is presented in **Table 1** and the corresponding clinical courses were demonstrated in **Figure 1A**.

Landscape of CCA Genome and Transcriptome

Whole-exome sequencing achieved a mean average depth of $194.67 \times$ cross all collected tissue samples. To identify tumor somatic mutations, SomaticSniper was applied on all tumor tissue samples using corresponding peritumor as control. Meanwhile, copy number variation was identified using TitanCNA. In total, an average of 378 somatic SNVs (range, 260–529) were detected in tumor tissues, and the distribution of SNVs across human Genome was visualized in **Figure 1C**. Annotation of acquired SNVs revealed a number of common mutated genes across tumor samples, containing several known cancer-related genes (**Figure 1B**). Several members of mucin (MUC16, MUC3A, MUC6, and MUC4) were among the most frequently mutated genes, which is consistence with previous reports (Chang et al., 2006; Pereira et al., 2016; Liu et al., 2018; Pareja et al., 2019). Other noteworthy genes included DSPP, PER3, MTCH2, and KRT18, all have been reported with important roles in tumor formation and development. On the other hand, a number of copy number of variations was also identified in tumor samples, showing a wide-spread instability of cancer genome (**Figure 1D**).

Meanwhile, transcriptome sequencing revealed a significant change on transcriptional level, with a total of 2366 differentially expressed genes identified between CCA tumor and peritumor

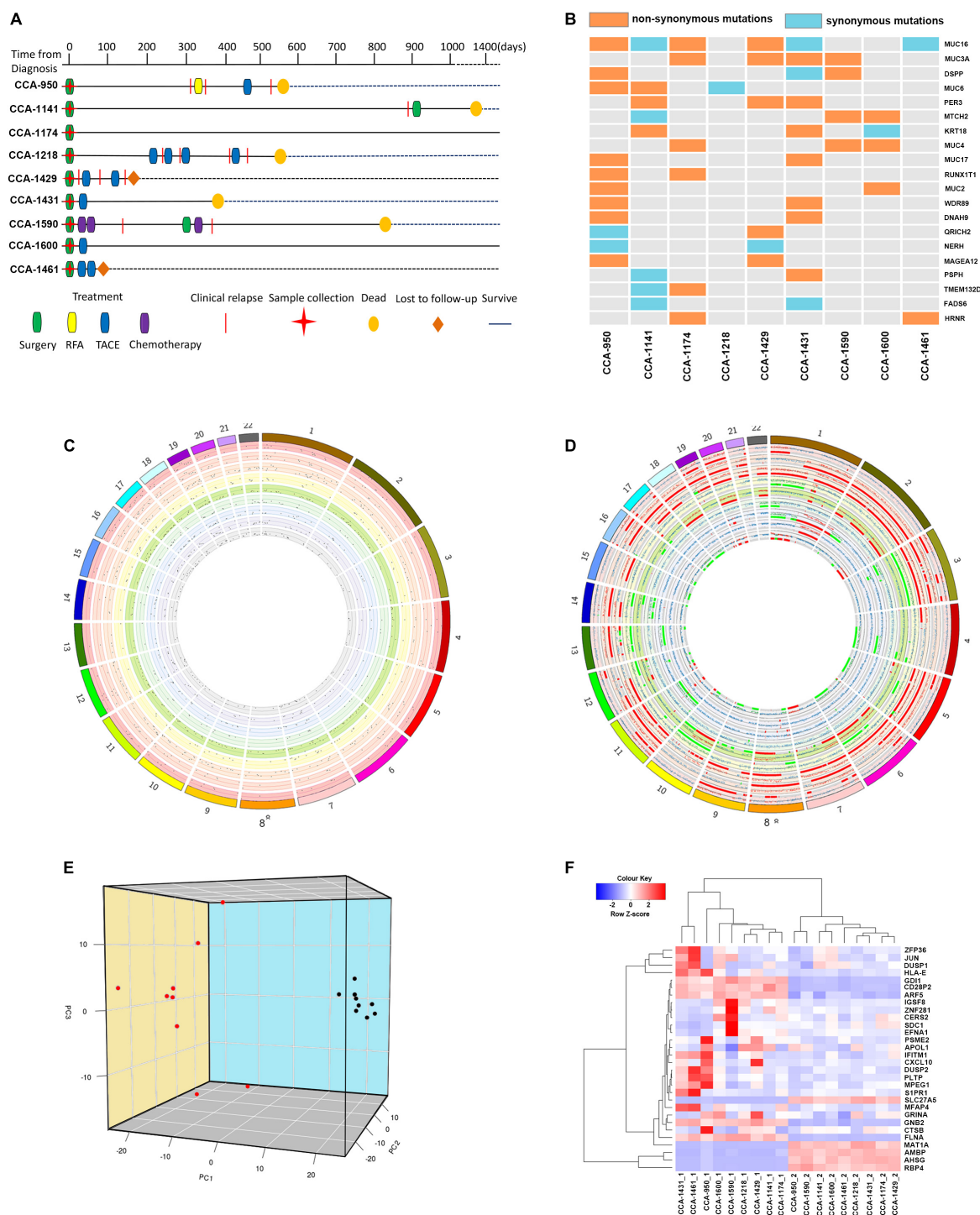


FIGURE 1 | The Clinical courses and the genome and transcriptome landscape of CCA. **(A)** The clinical course of 9 included CCA patients. RFA, Radiofrequency ablation; TACE, Transarterial chemoembolization. **(B)** The common mutated genes with somatic SNVs identified in included CCA patients. Different color indicated the functional type of somatic SNVs in these genes (orange: non-synonymous mutation; light blue: synonymous mutation; gray: not mutated). **(C)** The genomic distribution of somatic SNVs for included CCA patients. Each circle represented a single patient. Dots in the dot plot represented identified somatic SNVs and their heights indicated corresponding variant allele frequencies. **(D)** The genomic distribution of somatic CNVs for included CCA patients. Each circle represented a single patient. The scatter plot showed the logR value for each segment, and regions with different color indicated their copy number status (red: copy number gain; gray normal; green: copy number loss). **(E)** Principal component analysis of CCA transcriptome. The image showed the three-dimension distribution of each sample on the first three principal components. Red dots represented peritumor samples and black dots represented tumor samples. **(F)** Clustering of included tissue samples using top genes correlated with the first three principal components. Genes names and sample names were provided.

samples. To provide a clear classification based on samples' transcriptional features, principal component analysis was conducted to better characterize these samples. Not surprisingly, tumor samples and peritumor samples were well divided by the first three principal components, which explained 21.96%, 10.60%, and 8.68% of variation in samples' transcriptome, respectively (**Figure 1E**).

The results showed that the top genes positively associated with PC1 included RBP4, SLC27A5, and PCK2, all of which were known tumor-related genes and correlated with cancer patients' survival (Anderson and Stahl, 2013; Leithner et al., 2014, 2015; Balsa-Martinez and Puigserver, 2015). Meanwhile, PC1 negatively associated genes included FLNA, ARF5, and SLC25A6, suggesting its connection to cancer development (Savoy and Ghosh, 2013; Casalou et al., 2016; Shao et al., 2016; Cho et al., 2019). For PC2, top positively correlated genes included IFITM1 and GPX1, both have been reported to be associated with risk of numerous cancers (Ravn-Haren et al., 2006; Arsova-Sarafinowska et al., 2009; Lee et al., 2012; Ogoni et al., 2016), while most negatively PC2 correlated genes included common-known tumor over-expressed genes such as EFNA1 (Nakamura et al., 2005; Xiang-Dan et al., 2010).

As in PC3, most noteworthy genes positively correlated with this principal component are ZFP36 and DUSP1, both are known for their function of regulation in cancer progression

(Montorsi et al., 2016; Nagahashi et al., 2018). Other important correlated genes included t CXCL9 and CXCL10, and they served as important regulators of immune activation in tumor microenvironment (Bronger et al., 2016; Ding et al., 2016; Tokunaga et al., 2018).

Using top genes correlated with the first three principal components, transcriptome clustering revealed that tumor sample and peritumor samples could be indeed well separated (**Figure 1F**), suggesting that CCA tumors indeed have distinct gene expression patterns compared to peritumor tissues.

Clonal Evolution in CCA

To explore the evolutionary process driving tumorigenesis and development, ScIust algorithm was applied to infer subclonal populations in cancer genomes. Combining copy-number analysis and mutation clustering approach, ScIust could accurately determine copy-number states as well as cellular prevalence of mutations. As shown in **Figure 2A** and **Supplementary Figure S1**, different types of clonal structure were revealed. For 7 of the included patients (CCA-1218, CCA-1431, CCA-1461, CCA-950, CCA-1429, CCA-1590, and CCA-1600), no subclonal mutations were identified since all mutations within each sample could be clustered into one single cluster according to their allele frequencies. These results showed that during the tumor clonal evolution of these patients, the

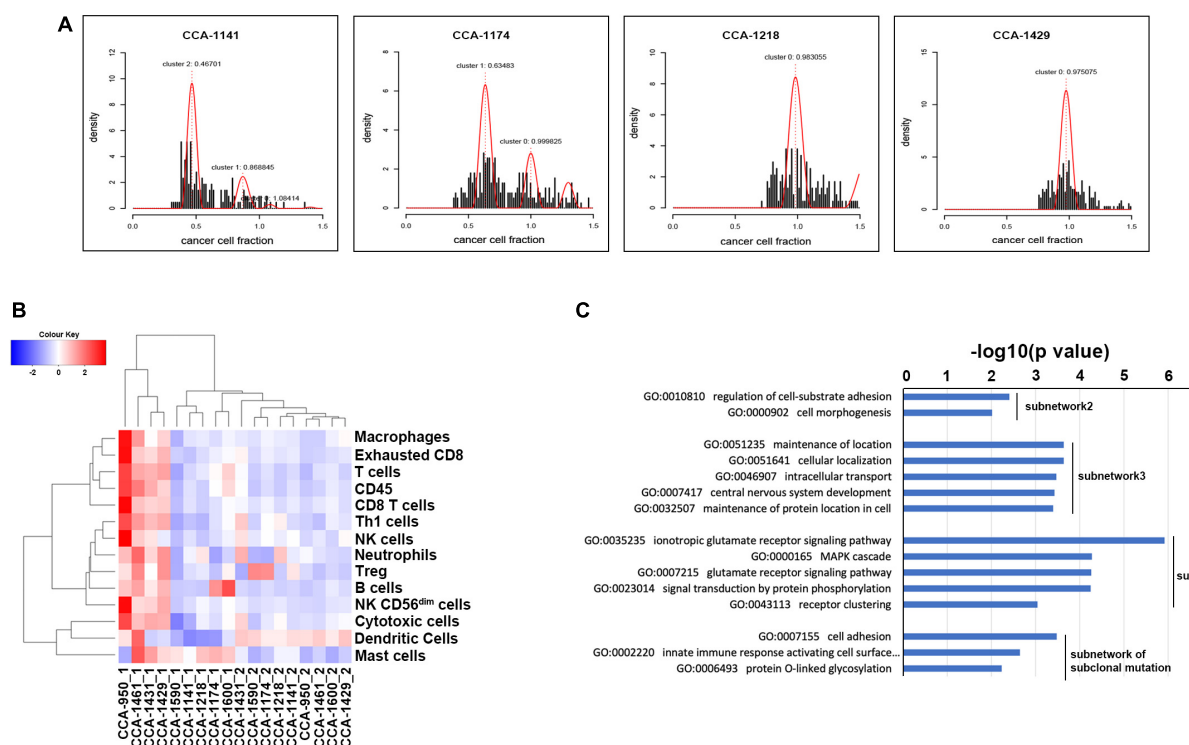
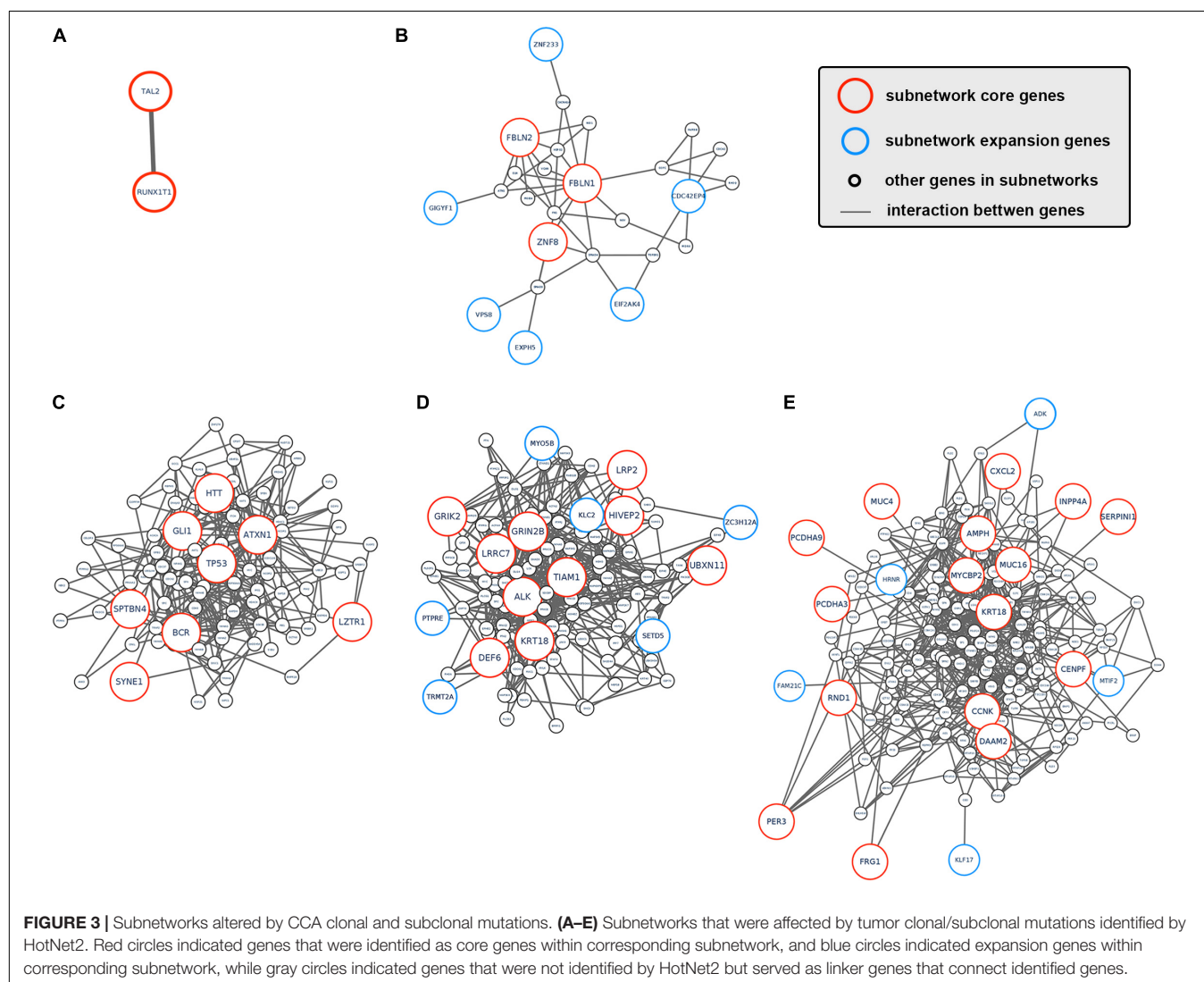


FIGURE 2 | Clonal evolution of CCA and its influence on biological interaction network. **(A)** Mutation clusters identified by ScIust in 4 of 9 included CCA patients. Additional cluster(s) other than cluster 0 were subclonal mutation clusters. Patient identifiers were provided above each plot. **(B)** Clustering of included tissue sample using known immune signatures. **(C)** Go term enrichment results in biological pathways for each identified subnetwork. Subnetwork 2–4 indicated subnetwork altered by clonal mutations and subnetwork of subclonal mutation indicated the subnetwork altered by subclonal mutations. Subnetwork 1 only contained two genes and did not show significant enrichment in Gene Ontology of biological pathways.

randomly accumulated mutations might not create subclones with significant survival advantage. The other 2 patients (CCA-1141 and CCA-1174), on the other hand, presented considerable portion of subclonal mutations. In patient CCA-1141, two large subclonal mutation clusters were observed, with cellular frequency of 46.70% and 86.88%, respectively. The other patient, CCA-1174, also showed one considerable subclonal mutation clusters, accounting for 63.48% of all tumor cells. The existence of a large number of subclonal mutations might suggest that the emerge of these tumor subclones took place in the later stage of tumor development, while a high cellular frequency further indicated that they possessed notable survival advantage. Surprisingly, these two patients with subclonal mutations identified showed better prognostic outcome compared to other patients, with relapse-free survival and over-all survival both longer than 20 months. One possible explanation is that in this kind of patients, some critical mutations that might greatly benefit tumors' growth took place in the later period of tumor development (which explained the expanding tumor

subclones), while other tumor acquired these genetic alterations in the early stage, and thus resulted in the differences in patients' prognosis. Evaluation of known immune signature based on gene expression further revealed that CCA-1141 and CCA-1174 could be categorized into cold tumor with relatively low level of cells correlated with immune response (**Figure 2B**). This result suggested that the clonal evolution of CCA might be closely related to its immune microenvironment, and high level of infiltration might suppress the evolutionary process of tumor cells.

To better understand how tumor clonal evolution affected different biological pathways/processes in tumor cells, we first divided patients' somatic mutations into clonal mutations and subclonal mutations, and then HotNet2 algorithm was used to scan gene interaction networks for altered subnetworks affected by different categories of mutations. For clonal mutations, four subnetworks were identified (**Figures 3A–E**). The first subnetwork contained only 2 core genes: RUNX1T1 and TAL2 (**Figure 3A**). These two genes were both related to



gene transcription and their dysregulation has been reported to promote tumorigenesis in various cancer. The second subnetwork (**Figure 3B**) contained three core genes (FBLN1, FBLN2, and ZNF8, label with red) and six expansion genes (CDC42EP4, EIF2AK4, EXPH5, GIGYF1, VPS8, and ZNF233, labeled with blue). Gene Ontology (GO) term enrichment analysis revealed that this subnetwork is closely related with extracellular matrix structure, cell-substrate adhesion and cell morphogenesis (**Figure 2C**), suggesting that tumor clonal mutation would show a tendency to affect biological pathways related to cells' interaction with microenvironment, which is critical for tumor development. The third subnetwork was made up of eight highly interacted genes, namely ATXN1, BCR, GLI1, HTT, LZTR1, SPTBN4, SYNE1, and TP53 (**Figure 3C**). All these genes were known as oncogenes, including a well-known driver gene in various cancer, TP53. The last and biggest subnetwork (**Figure 3D**) including 10 core genes (ALK, DEF6, GRIK2, GRIN2B, HIVEP2, KRT18, LRP2, LRRC7, TIAM1, UBXN11) and 6 expansion genes (KLC2, MYO5B, PTPRE, SETD5, TRMT2A, and ZC3H12A), most of which served as important components of multiple signaling pathways and involved in regulation of cancer cell.

Interestingly, several subnetworks altered by tumor clonal mutations were closely related to major metabolism pathways. It's within expectation since one well-known intrinsic character for tumor cells is its abnormal metabolism.

On the other hand, we also analyzed the subnetwork affected by tumor subclonal mutations. Considered that two out of nine patients were identified with subclonal mutations, HotNet2 identified only one subnetwork that was altered by subclonal mutations (**Figure 3E**). GO analysis revealed that the mutated genes were most relevant to cell adhesion. This result suggested that subclonal mutations benefiting tumor metastasis might bring survival advantage for corresponding tumor subclones.

Transcriptome Analysis Revealed Alteration in Pathways Enriched in CCA Clonal Evolutionary Process

we next explored the transcriptome landscape to evaluate the change in gene expression during CCA development. Using limma algorithm, a total of 2366 differentially expressed genes [$|\log(\text{fold-change})| \geq 1$ and $P_{\text{adjusted}} < 0.05$] were identified in CCA tumor comparing to peritumor samples (**Figure 4A**). Among these genes, 1833 were significantly upregulated in CCA and 533 were downregulated. Transcriptome clustering using the top 20 differentially expressed genes also showed an excellent separation between tumor and peritumor samples (**Figure 4B**). GO-term enrichment analysis revealed that the up-regulated genes (**Figure 4C**) were mostly enriched in the regulation of biological process (GO:0048519, GO:0048522 and GO:0048523), while down-regulated genes (**Figure 4D**) were mostly enriched in metabolism related biological processes including carboxylic acid metabolic process (GO:0019752) and oxoacid metabolic process (GO:0043436).

Next, HotNet2 was once again applied to identify the altered subnetworks affected by transcriptome aberration.

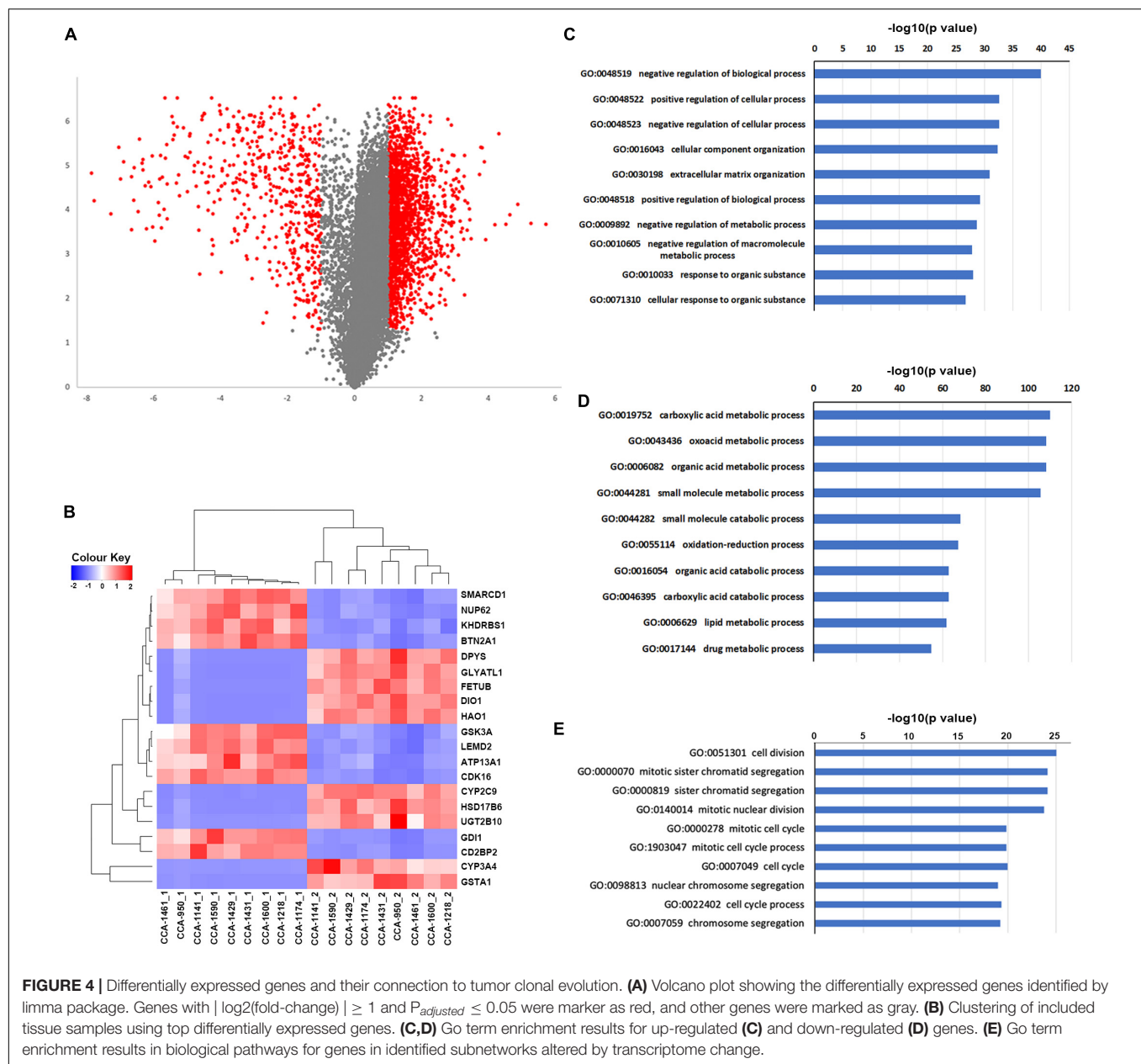
Surprisingly, genes identified in subnetworks affected by somatic mutations (clonal or subclonal) rarely appeared in subnetworks affected by transcriptome change. However, mapping genes affected by transcriptome change back to biological interaction networks revealed that many of these genes were in close range of the altered subnetworks affected by tumor somatic mutations (**Figures 5A–E**). It appeared that tumor genomic alterations created a spreading aberration across the biological interaction network and thus a number of genes were under their influence, resulting in a wide-range change of tumor transcriptome. Meanwhile, Gene Ontology enrichment analysis revealed that subnetworks altered by transcriptome change were dominantly enriched in biological processes related to cell division and cell cycle (**Figure 4E**), including cell division (GO:0051301), cell cycle (GO:0004857), protein localization (GO:0008104) and cellular component organization (GO:0016043), indicating notable change of proliferation capacity happened during tumor clonal evolution. It's not surprising that cell morphogenesis (GO:0000902), cellular localization (GO:0051641), intracellular transport (GO:0046907) and maintenance of protein location in cell (GO:0032507), four biological pathways that had been reported to be significantly enriched for mutation-affected subnetworks, were also enriched for these transcriptome-change-affected genes.

Furthermore, we also found that these multi-omics-altered subnetworks were significantly overlapped with pathways presented in kegg database (**Supplementary Figures S2–S21**). Noteworthy, all hot subnetworks were significantly overlapped with pathways in cancer (hsa05200), while other enriched pathways included cell-cycle (Vermeulen et al., 2003), ECM-receptor interaction (Lu et al., 2012) and VEGF signaling pathway (Roskoski, 2007), all have been reported to be related with tumor progression.

To further investigate if the altered pathways could be clinically related, we obtain the gene expression profile from TCGA-CHOL dataset and use Cox regression analysis to identify potential biomarkers for CCA patients' overall survival. Univariate cox regression analysis revealed that 14 genes within the hot subnetworks showed expression pattern significantly correlated with patients' overall survival (**Supplementary Figure S22**), including PTN and EGFR, two major players in tumor progression. Then these genes were utilized to generate the multivariate Cox regression model using stepwise forward selection. The acquired model consisted of 4 genes (PTPRZ1, CFH, RCN2 and VPS4B) and corresponding model parameters were summarized in **Supplementary Table S1**. The prognostic value was then calculated from the model score as follows:

$$\text{prognostic value} = \frac{e^{\text{score}}}{1 + e^{\text{score}}}$$

Applying the 50-percentage cutoff of prognostic value, the TCGA-CHOL dataset could be divided into two risk groups with distinct prognostic patterns (Kaplan-Meier survival analysis, $p = 0.00015$, **Supplementary Figure S23**).

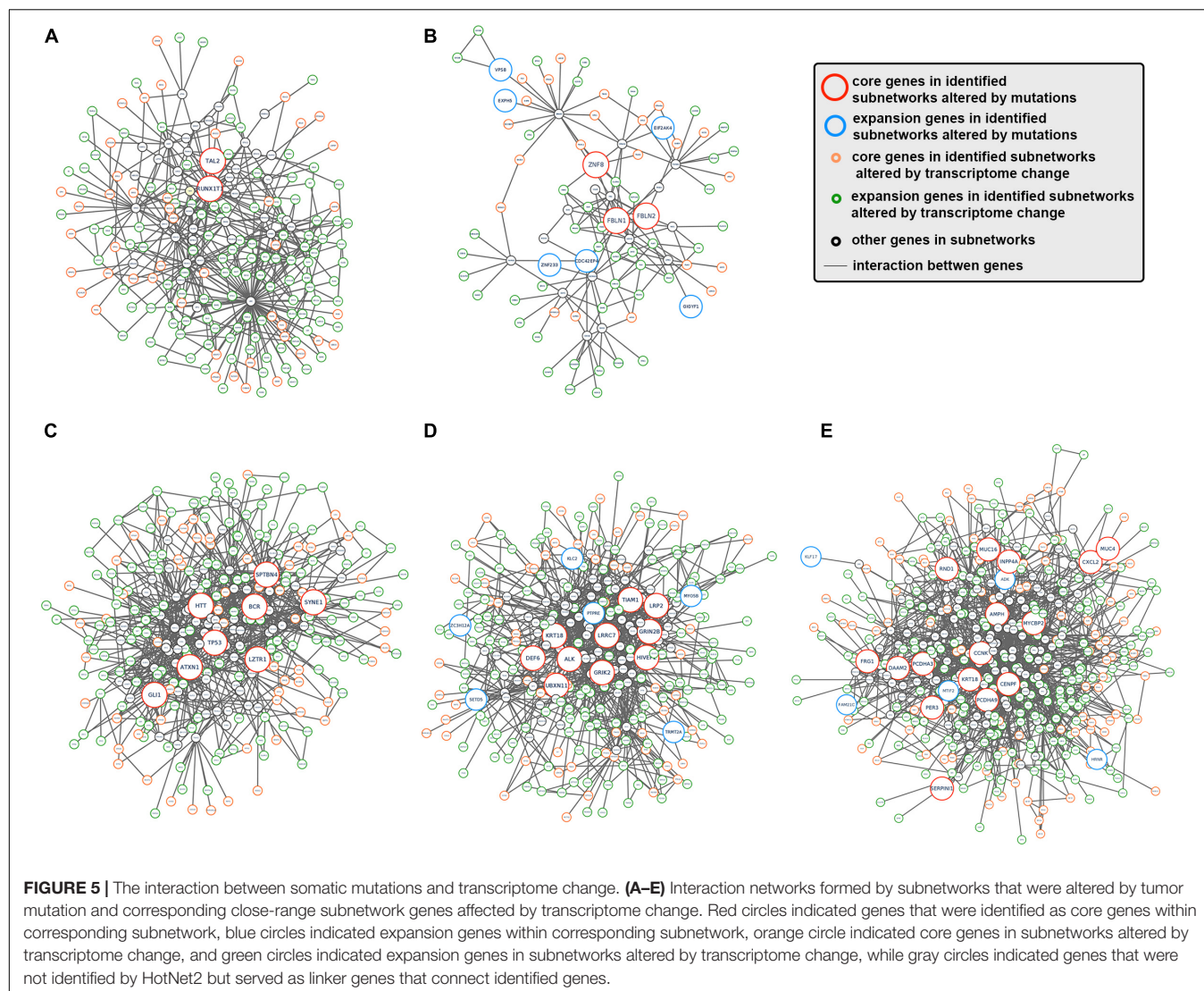


All these results suggested that the alteration of tumor genome and transcriptome were closely related, and the influence of driver gene mutations might spread to faraway downstream.

DISCUSSION

Clonal evolution has been proved to be one of the most important concepts in tumor genesis and development. Currently, a lot of researches have been conducted in variable kinds of tumors and revealed different clonal evolution patterns along with cancer development, providing insights into better understanding of their evolutionary mechanism. These valuable knowledges were of great value in prognosis evaluation and treatment selection.

In our analysis including 9 cholangiocarcinoma patients, we discovered that a major portion (7/9) of CCA cases did not show visible subclones within the primary tumors, indicating the existence of mature clonal structure after tumorigenesis. Interestingly, the other two CCA patients with considerable subclones demonstrated significantly longer RFS and OS compared to these patients without visible subclones. Above phenomena might suggest that the forming of a stable and lasting clonal structure at early stage might lead to worse clinical outcome for CCA cases. Another intriguing finding is that the expanding subclones in tumor were connected to relatively low immune signatures (as we showed before), showing a close interaction between tumor and its immune microenvironment. Meanwhile, identification of subnetworks



affected by CCA clonal/subclonal mutations revealed that clonal mutations' influence spread across a number of different biological pathways, while subclonal mutations influence mainly focused on pathways that benefiting tumor metastasis. This result indicated that most mutations with survival advantage were acquired during early stage of CCA development and acquisition of mutations on key regulator genes could affect how tumor evolved.

Cancer development involved biological alteration/dysregulation on multiple biological levels, including genomic, epigenomic and transcriptomic. Although a lot of studies have been conducted on every single omics level, discovering a variety of patterns and mechanism for how these alterations contribute to tumorigenesis, one major question still remained largely unanswered: how the alteration on multiple biological levels interact? In our analysis, we identified key subnetworks that were greatly affected by genomic and transcriptomic changes. Interestingly, although genes in subnetworks greatly affected by genomic change rarely

overlapped with those under the influence of transcriptome alteration, it appeared that these two groups of genes were in close range within biological interaction networks, suggesting that dysregulation of genome and transcriptome were closely related. One possible explanation might be that genes that were mutated served as sources of disturbance and affected the expression of their neighbor genes. This disturbance could further spread, creating a large-scale change of tumor transcriptome.

CONCLUSION

In conclusion, integrating whole exome and transcriptome sequencing technology, our analysis demonstrated the landscape of CCA genome as well as transcriptome and discovered the different clonal evolution patterns in these patients. We also identified biological pathways significantly altered by tumor somatic mutations and transcriptome change and reveal the connection among the alteration on different omics levels, which

could bring insight for better understanding the mechanism of CCA development and help future prognosis evaluation.

DATA AVAILABILITY STATEMENT

The raw sequence data reported in this paper have been deposited in the Genome Sequence Archive (Wang et al., 2017) in BIG Data Center (Big Data Center Members, 2018), Beijing Institute of Genomics (BIG), Chinese Academy of Sciences, under accession numbers HRA000085, which can be accessed at <https://bigd.big.ac.cn/gsa-human>.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Institution Review Board of Mengchao Hepatobiliary Hospital of Fujian Medical University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

GC, ZC, and HZ contributed the conception and design of the study. GC, JZ, and XH performed the bioinformatic analysis. XD,

SL, and ZC performed the sample collection and clinical data collection. HZ, GC, F-EL, and XL interpreted the analysis results. GC and HZ wrote the manuscript. ZC, F-EL, and XL wrote the sections of the manuscript. All authors contributed to manuscript revision, read and approved the submitted version.

FUNDING

This work was supported by the National Science and Technology Major Project of China (Grant No. 2018ZX10302205) and National Natural Science Foundation of China (Grant No. 61372151).

ACKNOWLEDGMENTS

The results shown here are part based upon data generated by the TCGA Research Network: <https://www.cancer.gov/tcga>.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.00195/full#supplementary-material>

REFERENCES

- Anderson, C. M., and Stahl, A. (2013). SLC27 fatty acid transport proteins. *Mol. Asp. Med.* 34, 516–528. doi: 10.1016/j.mam.2012.07.010
- Arsova-Sarafinovska, Z., Matevska, N., Eken, A., Petrovski, D., Banev, S., Dzokova, S., et al. (2009). Glutathione peroxidase 1 (GPX1) genetic polymorphism, erythrocyte GPX activity, and prostate cancer risk. *Int. Urol. Nephrol.* 41, 63–70. doi: 10.1007/s11255-008-9407-y
- Balsa-Martinez, E., and Puigserver, P. (2015). Cancer cells hijack gluconeogenic enzymes to fuel cell growth. *Mol. Cell* 60, 509–511. doi: 10.1016/j.molcel.2015.11.005
- Banales, J. M., Cardinale, V., Carpino, G., Marziani, M., Andersen, J. B., Invernizzi, P., et al. (2016). Expert consensus document: cholangiocarcinoma: current knowledge and future perspectives consensus statement from the European Network for the Study of Cholangiocarcinoma (ENS-CCA). *Nat. Rev. Gastroenterol. Hepatol.* 13, 261–280. doi: 10.1038/nrgastro.2016.51
- Big Data Center Members (2018). Database resources of the BIG data center in. *Nucleic Acids Res.* 46, D14–D20. doi: 10.1093/nar/gkx897
- Bishayee, A. (2014). The role of inflammation and liver cancer. *Adv. Exp. Med. Biol.* 816, 401–435. doi: 10.1007/978-3-0348-0837-8_16
- Bronger, H., Singer, J., Windmüller, C., Reuning, U., Zech, D., Delbridge, C., et al. (2016). CXCL9 and CXCL10 predict survival and are regulated by cyclooxygenase inhibition in advanced serous ovarian cancer. *Br. J. Cancer* 115, 553–563. doi: 10.1038/bjc.2016.172
- Cardinale, V., Bragazzi, M. C., Carpino, G., Di Matteo, S., Overi, D., Nevi, L., et al. (2018). Intrahepatic cholangiocarcinoma: review and update. *Hepatoma Res.* 4:20.
- Casalou, C., Faustino, A., and Barral, D. C. (2016). Arf proteins in cancer cell migration. *Small GTPases* 7, 270–282. doi: 10.1080/21541248.2016.1228792
- Chang, S. C., Lin, J. K., Yang, S. H., Wang, H. S., Li, A. F. Y., and Chin-Wen, C. (2006). Relationship between genetic alterations and prognosis in sporadic colorectal cancer. *Int. J. Cancer* 118, 1721–1727. doi: 10.1002/ijc.21563
- Chen, G., Cai, Z., Li, Z., Dong, X., Xu, H., Lin, J., et al. (2018). Clonal evolution in long-term follow-up patients with hepatocellular carcinoma. *Int. J. Cancer* 143, 2862–2870. doi: 10.1002/ijc.31844
- Cho, S. H., Pak, K., Jeong, D. C., Han, M. E., Oh, S. O., and Kim, Y. H. (2019). The AP2M1 gene expression is a promising biomarker for predicting survival of patients with hepatocellular carcinoma. *J. Cell. Biochem.* 120, 4140–4146. doi: 10.1002/jcb.27699
- Cun, Y., Yang, T.-P., Achter, V., Lang, U., and Peifer, M. (2018). Copy-number analysis and inference of subclonal populations in cancer genomes using ScLust. *Nat. Protoc.* 13, 1488–1501. doi: 10.1038/nprot.2018.033
- Ding, Q., Lu, P., Xia, Y., Ding, S., Fan, Y., Li, X., et al. (2016). CXCL9: evidence and contradictions for its role in tumor progression. *Cancer Med.* 5, 3246–3259. doi: 10.1002/cam4.934
- Ferrando, A. A., and López-Otín, C. (2017). Clonal evolution in leukemia. *Nat. Med.* 23, 1135–1145.
- Greaves, M., and Maley, C. C. (2012). Clonal evolution in cancer. *Nature* 481, 306–313.
- Ha, G., Roth, A., Khattra, J., Ho, J., Yap, D., Prentice, L. M., et al. (2014). TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res.* 24, 1881–1893. doi: 10.1101/gr.180281.114
- Hoadley, K. A., Siegel, M. B., Kanchi, K. L., Miller, C. A., Ding, L., Zhao, W., et al. (2016). Tumor evolution in two patients with basal-like breast cancer: a retrospective genomics study of multiple metastases. *PLoS Med.* 13:e1002174. doi: 10.1371/journal.pmed.1002174
- Larson, D. E., Harris, C. C., Chen, K., Koboldt, D. C., Abbott, T. E., Dooling, D. J., et al. (2012). SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* 28, 311–317. doi: 10.1093/bioinformatics/btr665
- Lee, J., Goh, S. H., Song, N., Hwang, J. A., Nam, S., JuChoi, I., et al. (2012). Overexpression of IFITM1 has clinicopathologic effects on gastric cancer and is regulated by an epigenetic mechanism. *Am. J. Pathol.* 181, 43–52. doi: 10.1016/j.ajpath.2012.03.027
- Leithner, K., Hrzenjak, A., and Olschewski, H. (2014). Gluconeogenesis in cancer: door wide open. *Proc. Natl. Acad. Sci. U.S.A.* 111, E4394–E4394.
- Leithner, K., Hrzenjak, A., Trötz Müller, M., Moustafa, T., Köfeler, H. C., Wohlkoeing, C., et al. (2015). PCK2 activation mediates an adaptive response

- to glucose depletion in lung cancer. *Oncogene* 34, 1044–1050. doi: 10.1038/onc.2014.47
- Liu, B., Hu, F. F., Zhang, Q., Hu, H., Ye, Z., Tang, Q., et al. (2018). Genomic landscape and mutational impacts of recurrently mutated genes in cancers. *Mol. Genet. Genomic Med.* 6, 910–923. doi: 10.1002/mgg3.458
- Lu, P., Weaver, V. M., and Werb, Z. (2012). The extracellular matrix: a dynamic niche in cancer progression. *J. Cell Biol.* 196, 395–406. doi: 10.1083/jcb.201102147
- McGranahan, N., and Swanton, C. (2017). Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell* 168, 613–628. doi: 10.1016/j.cell.2017.01.018
- Montorsi, L., Filippo, G., Claudia, A., Caporali, A., Martello, A., Atene, C. G., et al. (2016). Loss of ZFP36 expression in colorectal cancer correlates to wnt/ β -catenin activity and enhances epithelial-to-mesenchymal transition through upregulation of ZEB1, SOX9 and MACC1. *Oncotarget* 7, 59144–59157. doi: 10.18632/oncotarget.10828
- Nagahashi, M., Yamada, A., Katsuta, E., Aoyagi, T., Wei-Ching, H., Terracina, K. P., et al. (2018). Targeting the SphK1/S1P/S1PR1 axis that links obesity, chronic inflammation, and breast cancer metastasis. *Cancer Res.* 78, 1713–1725. doi: 10.1158/0008-5472.CAN-17-1423
- Nakamura, R., Kataoka, H., Sato, N., Kanamori, M., Ihara, M., Igarashi, H., et al. (2005). EPHA2/EFNA1 expression in human gastric cancer. *Cancer Sci.* 96, 42–47. doi: 10.1111/j.1349-7006.2005.00007.x
- Ogony, J., Choi, H. J., Lui, A., Cristofanilli, M., and Lewis-Wambi, J. (2016). Interferon-induced transmembrane protein 1 (IFITM1) overexpression enhances the aggressive phenotype of SUM149 inflammatory breast cancer cells in a signal transducer and activator of transcription 2 (STAT2)-dependent manner. *Breast Cancer Res.* 18:25. doi: 10.1186/s13058-016-0683-7
- Pareja, F., Lee, J. Y., Brown, D. N., Piscuoglio, S., Gualarte-Mérida, R., Selenica, P., et al. (2019). The genomic landscape of mucinous breast cancer. *J. Natl. Cancer Inst.* 111, 737–741.
- Pereira, B., Suet-Feung, C., Rueda, O. M., Hans-Kristian, M. V., Provenzano, E., Bardwell, H. A., et al. (2016). The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nat. Commun.* 7, 11479. doi: 10.1038/ncomms11479
- Ravn-Haren, G., Olsen, A., Tjønneland, A., Dragsted, L. O., Nexø, B. A., Wallin, H., et al. (2006). Associations between GPX1 Pro198Leu polymorphism, erythrocyte GPX activity, alcohol consumption and breast cancer risk in a prospective cohort study. *Carcinogenesis* 27, 820–825. doi: 10.1093/carcin/bgi267
- Rizvi, S., Khan, S. A., Hallemeier, C. L., Kelley, R. K., and Gores, G. J. (2018). Cholangiocarcinoma - evolving concepts and therapeutic strategies. *Nat. Rev. Clin. Oncol.* 15, 95–111. doi: 10.1038/nrclinonc.2017.157
- Roskoski, R. Jr. (2007). Vascular endothelial growth factor (VEGF) signaling in tumor progression. *Crit. Rev. Oncol. Hematol.* 62, 179–213. doi: 10.1016/j.critrevonc.2007.01.006
- Savoy, R. M., and Ghosh, P. M. (2013). The dual role of filamin A in cancer: can't live with (too much of) it, can't live without it. *Endocr. -Relat. Cancer* 20, R341–R356. doi: 10.1530/ERC-13-0364
- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303
- Shao, Q.-Q., Zhang, T.-P., Zhao, W.-J., Liu, Z.-W., You, L., Zhou, L., et al. (2016). Filamin A: insights into its exact role in cancers. *Pathol. Oncol. Res.* 22, 245–252. doi: 10.1007/s12253-015-9980-1
- Tokunaga, R., Zhang, W., Naseem, M., Puccini, A., Berger, M. D., Soni, S., et al. (2018). CXCL9, CXCL10, CXCL11/CXCR3 axis for immune activation - A target for novel cancer therapy. *Cancer Treat. Rev.* 63, 40–47. doi: 10.1016/j.ctrv.2017.11.007
- Vermeulen, K., Van Bockstaele, D. R., and Berneman, Z. N. (2003). The cell cycle: a review of regulation, deregulation and therapeutic targets in cancer. *Cell Prolif.* 36, 131–149. doi: 10.1046/j.1365-2184.2003.00266.x
- Wang, Y., Song, F., Zhu, J., Zhang, S., Yang, Y., Chen, T., et al. (2017). GSA: genome sequence archive. *Genom. Proteom. Bioinform.* 15, 14–18. doi: 10.1016/j.gpb.2017.01.001
- Xiang-Dan, C., Mi-Jin, L., Goung-Ran, Y., In-Hee, K., Hee-Chul, Y., Eun-Young, S., et al. (2010). EFNA1 ligand and its receptor EphA2: potential biomarkers for hepatocellular carcinoma. *Int. J. Cancer* 126, 940–949. doi: 10.1002/ijc.24798

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Chen, Cai, Dong, Zhao, Lin, Hu, Liu, Liu and Zhang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Computational Oncology in the Multi-Omics Era: State of the Art

Guillermo de Anda-Jáuregui^{1,2*} and Enrique Hernández-Lemus^{1,3*}

¹ Computational Genomics Division, National Institute of Genomic Medicine, Mexico City, Mexico, ² Cátedras Conacyt Para Jóvenes Investigadores, National Council on Science and Technology, Mexico City, Mexico, ³ Center for Complexity Sciences, Universidad Nacional Autónoma de México, Mexico City, Mexico

OPEN ACCESS

Edited by:

Francesca Finotello,
Innsbruck Medical University, Austria

Reviewed by:

Raoul Jean Pierre Bonnal,
Istituto Nazionale Genetica Molecolare
(INGM), Italy
Barbara Di Camillo,
University of Padova, Italy
Dietmar Rieder,
Innsbruck Medical University, Austria

*Correspondence:

Guillermo de Anda-Jáuregui
gdeanda@inmegen.edu.mx
Enrique Hernández-Lemus
ehernandez@inmegen.gob.mx

Specialty section:

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

Received: 01 December 2019

Accepted: 10 March 2020

Published: 07 April 2020

Citation:

de Anda-Jáuregui G and
Hernández-Lemus E (2020)
Computational Oncology in the
Multi-Omics Era: State of the Art.
Front. Oncol. 10:423.
doi: 10.3389/fonc.2020.00423

Cancer is the quintessential complex disease. As technologies evolve faster each day, we are able to quantify the different layers of biological elements that contribute to the emergence and development of malignancies. In this multi-omics context, the use of integrative approaches is mandatory in order to gain further insights on oncological phenomena, and to move forward toward the precision medicine paradigm. In this review, we will focus on computational oncology as an integrative discipline that incorporates knowledge from the mathematical, physical, and computational fields to further the biomedical understanding of cancer. We will discuss the current roles of computation in oncology in the context of multi-omic technologies, which include: data acquisition and processing; data management in the clinical and research settings; classification, diagnosis, and prognosis; and the development of models in the research setting, including their use for therapeutic target identification. We will discuss the machine learning and network approaches as two of the most promising emerging paradigms, in computational oncology. These approaches provide a foundation on how to integrate different layers of biological description into coherent frameworks that allow advances both in the basic and clinical settings.

Keywords: multi-omics analysis, computational oncology, data integration, cancer complexity, machine learning, network science

1. CANCER: THE COMPLEX DISEASE

Cancer is by now widely accepted to be the quintessential complex disease: a proper description of the pathological phenotype can only be achieved by properly integrating the myriad of interconnected biological elements and their relationships with their environment (1). As a complex system, cancer exhibits features, such as: emergent patterns, adaptive and collective behaviors, self-organization, non-linear dynamics, and interactions forming complex networks (2). Examples of these can be found in the *Hallmarks of Cancer* (3, 4), as seen in **Figure 1**.

On a system-wide fashion, every tumor is involved in interactions with non-cancer elements: such as gene-environment interactions (GxE) (5), micro-environmental interactions (6), and those with the immune system (7); intercellular interactions within the tumor environment (8); and intracellular interactions, such as transcriptional regulation and gene co-expression (9, 10), signaling (11, 12) and metabolic pathways (13, 14), as well as protein interactions (15). These are exemplified in **Figure 2**. It soon becomes evident that a major source of cancer complexity lies on the many layers of interacting elements involved in the phenomenon.

2. THE MULTI-OMICS PARADIGM

2.1. Multi-Omics in a Nutshell

Multimomics is the name given to the modelization approach in biology that makes use of more than one of the current high-throughput biomolecular experimental techniques (a.k.a.

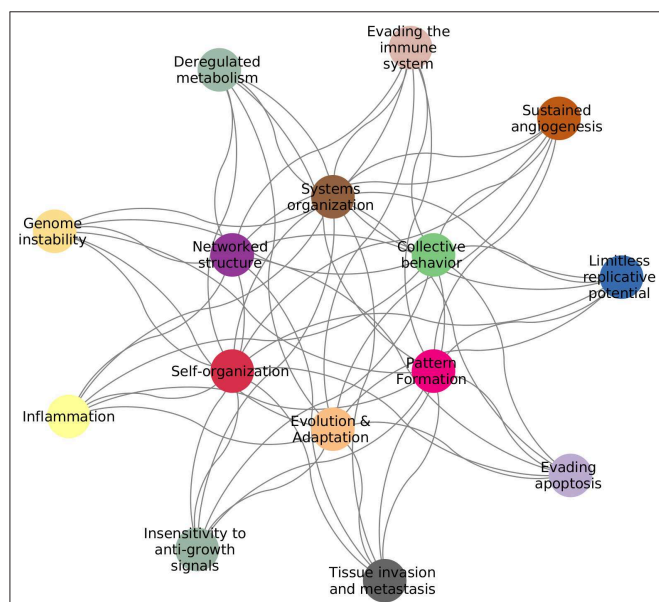


FIGURE 1 | Hallmarks of cancer complexity. The defining features of cancer (3, 4) are intrinsically connected to the defining features of complex systems (2).

omics) in order to characterize biological systems at the phenomenological level. It is understood that every omic contributes on a specific fashion to shape the actual biological phenotype under study. For this reason, it has become evident that there is a need for integrating frameworks to gather and organize the knowledge gained with each experimental approach into mechanistic or semi-mechanistic descriptions of the biological phenomenon. This issue has been deemed particularly relevant for the study of complex phenotypes, such as cancer tumors (16).

The rapid development of sequencing strategies as well as genotyping and expression microarrays led to the development of gene models to account for the molecular aspects of biology at the whole cellular level (and even at the organ and organism scales). The coming of age and popularization (driven by an almost exponential lowering of the costs) of next gen sequencing techniques leads to an explosion of new approaches to understand complex phenotypes that in turn have sped up the rise of high throughput proteomics, metabolomics catching up. Single cell technologies and a number of arising sequence based approaches (ChIP-seq, ATAC-seq) are becoming usual tools of biomedical and in particular cancer research (see **Figure 3**, for an account of the fastly increasing number of PubMed publications based on these omic tools).

In spite of this, the integrative approach to multi-omic modeling is far from trivial due to the broad diversity of data types, dynamic ranges and sources of experimental and analytical errors characteristic of each omic. In spite of these facts, a number of approaches to multi-omic integration have been proposed [see, for instance, discussions in Hernández-Lemus (17, 18)]. Said approaches make use of tools from statistics, probability, machine

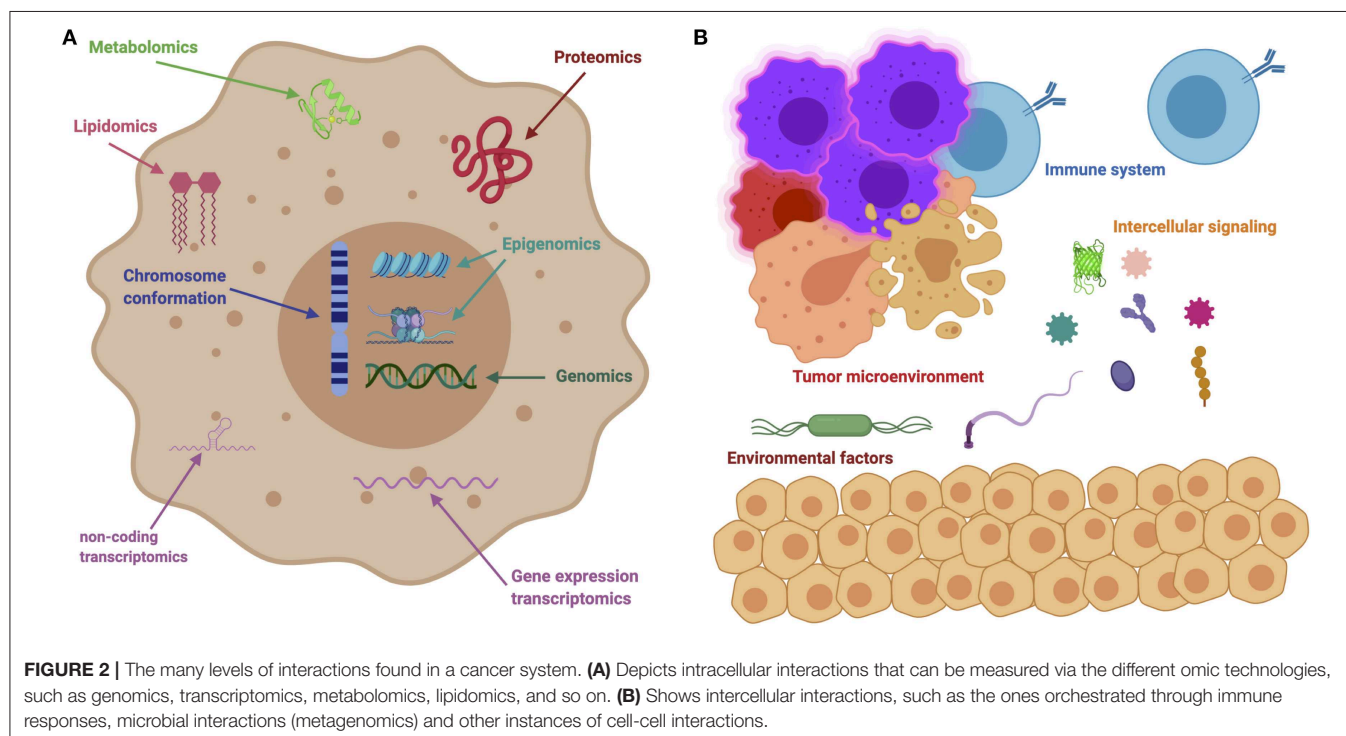


FIGURE 2 | The many levels of interactions found in a cancer system. **(A)** Depicts intracellular interactions that can be measured via the different omic technologies, such as genomics, transcriptomics, metabolomics, lipidomics, and so on. **(B)** Shows intercellular interactions, such as the ones orchestrated through immune responses, microbial interactions (metagenomics) and other instances of cell-cell interactions.

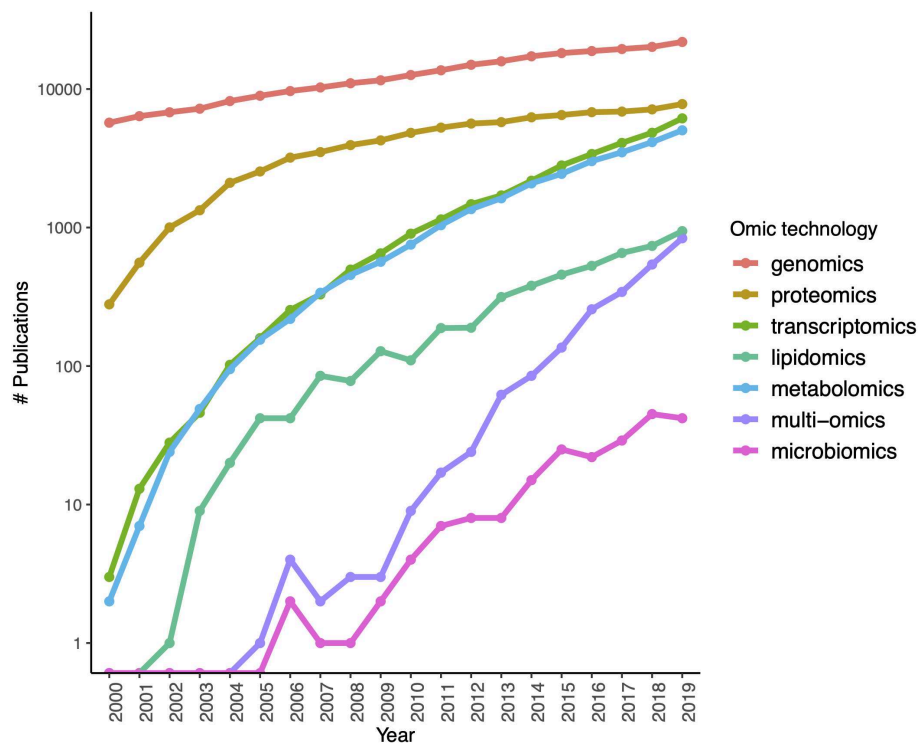


FIGURE 3 | Growth of interest in omics technologies in the twenty-first century: the number of Pubmed publications mentioning each omic technology in its title or abstract measured yearly since the year 2000.

learning and network science to classify, explore and provide guidelines for feature selection and their application is very much rooted in the tenets of systems biology.

The systematic study of cancer given by multi-omics is founded on the acknowledgment of a contribution of many different factors in the development and maintenance of the malignant state, including genetic aberrations, epigenetic alterations, changes in the response to cellular signaling, metabolic alterations, and beyond (19). Hence, by analyzing cancer as a complex pathology, the systems biology paradigm tries to gain insight into the molecular origins of the disease by looking at the diverse contributions, from DNA mutations (both germline and somatic), to deregulation of the gene expression programmes, the phenomenon of hormone disruption, that may or not be supplemented by metabolic abnormalities, and aberrant pathway signaling.

Cancer is also a multiscale pathology, aside from the biomolecular events just mentioned there is the influence of the environment and lifestyle that is known to be able to modify the onset, development, and outcome of tumors and their metastases. Multiomic analysis under a systems biology framework makes possible to use the unprecedented power of current high-throughput molecular and computational tools to draw a more complete figure of the different players in tumorigenesis and tumor establishment. At the same time, it may provide us with new instruments and strategies useful in basic and clinical research laboratories, but also in translational medicine and therapeutic endeavors.

These different levels of description have been independently studied for years. However, even if the advent of high-throughput technologies has permitted the development of systems biology, system-level models (conforming the theoretical foundations of these multiomic studies) are still under development.

2.2. The Systems Biology Framework

In essence, the foundational basis of systems biology is that of considering biological phenomena as systems, i.e., constructs formed by a large number of complex molecular and environmental components interacting at different levels to shape the functional features of said system. Tumor behavior, for instance, is determined by a combination of changes in genomic information that may (or may not) be associated with abnormal gene expression profiles; affecting protein abundance, but also modifying protein structure and folding, as well as supramolecular assembly. Changes in the regulatory patterns may also affect cell signaling mechanisms; and their responses. Hence, the complex interaction of nucleic acids and proteins in replication, transcription, metabolic, and signaling networks are considered the ultimate causes for the functioning (or malfunctioning, if preferred) of the tumor cell. We can notice that these are interdependent phenomena that cannot be treated separately, hence the need for integrative methodologies.

Another pivotal challenge in contemporary studies undertaken following a systems biology view is hence data integration. Data integration allows for the understanding of the enormous datasets generated by experimental multi-omics.

This is indeed a highly non-trivial task, since just the data management of such large amounts of information represents a challenge that has been called the big data paradigm.

3. THE ROLES OF COMPUTATION IN THE AGE OF CANCER MULTI-OMICS

We have identified four main roles that computation plays in the analysis of high-throughput data. These are the raw data acquisition from high-throughput instruments; the processing of raw data to quantitative data; the storage and management of massive omics data, for instance in remote repositories; and finally the deployment of data analysis models. These roles are illustrated in **Figure 4**. In this section, we will discuss select aspects of each of these roles.

3.1. Data Acquisition and Processing

The acquisition, processing, and manipulation of omic data generated in high throughput experiments requires, due to the very nature of these experiments (see **Figure 5**), the use of specialized bioinformatics pipelines. As the complexity of these datasets increases due to the natural

evolution of these technologies, so do the associated challenges evolve (20). Bioinformatics workflow management systems can be used to develop, maintain, and foster reproducibility of a give pipeline or workflow. Examples of these systems include Galaxy (21), Snakemake (22), Nextflow (23), and the general purpose Common Workflow Language (24).

It should be noted that a large number of tools for omic data analysis are available as packages for the R language contained in the *Bioconductor* project (25), a repository of bioinformatics open source software. It is important, however, to acknowledge the existence of other software ecosystems, such as the *Biopython* project (26). Although the number of packages in Bioconductor is greater than that found in Biopython [see for instance (27)], the main takeaway should be that there is a large number of tools available to researchers that can be used in any combination suitable for their research question.

3.1.1. Genomics

The oldest of the omic technologies, genomic analyses focus on the genomic sequence and its variations: insertions, deletions (INDELs), single nucleotide variations (SNVs), copy number

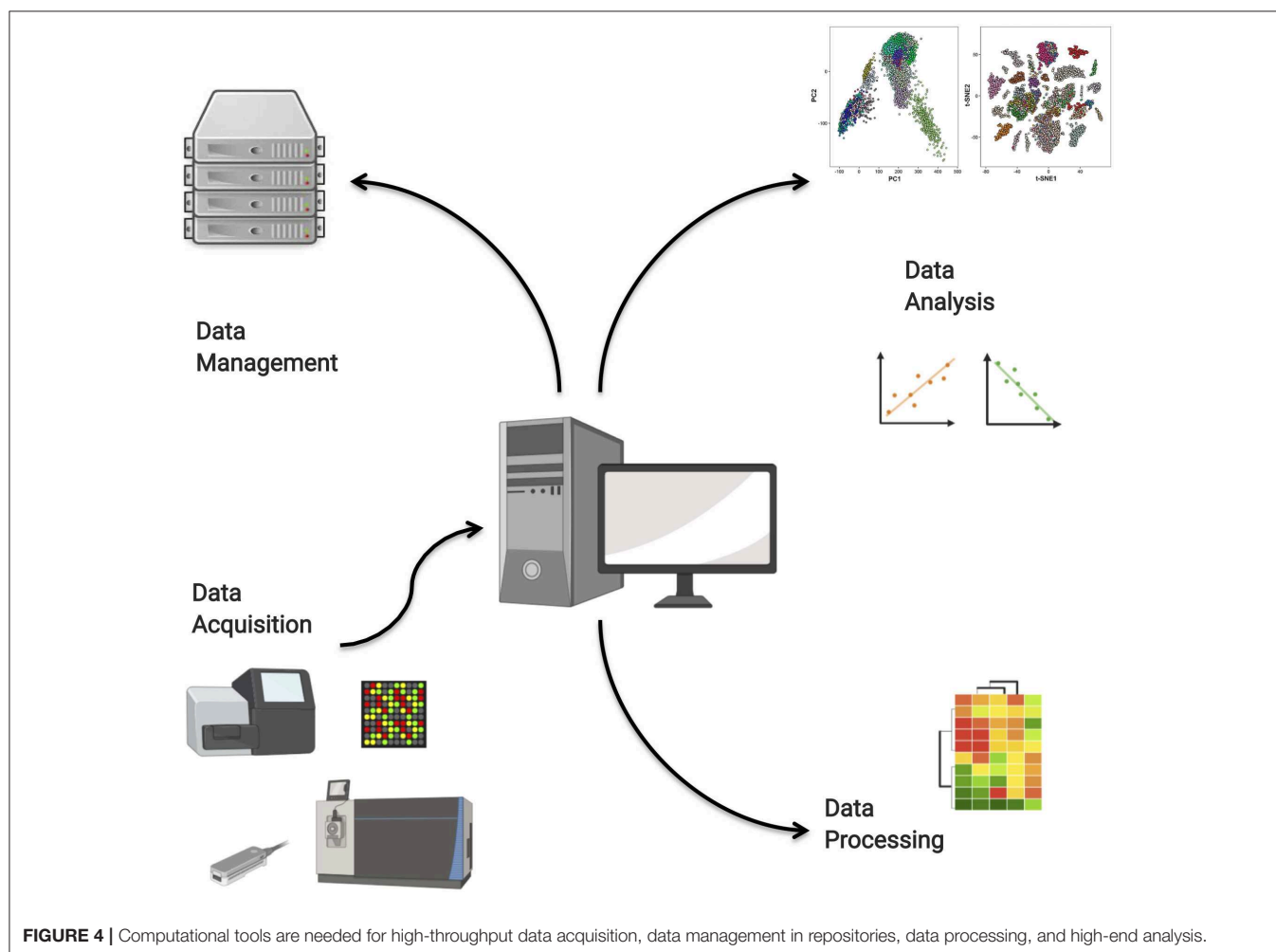


FIGURE 4 | Computational tools are needed for high-throughput data acquisition, data management in repositories, data processing, and high-end analysis.

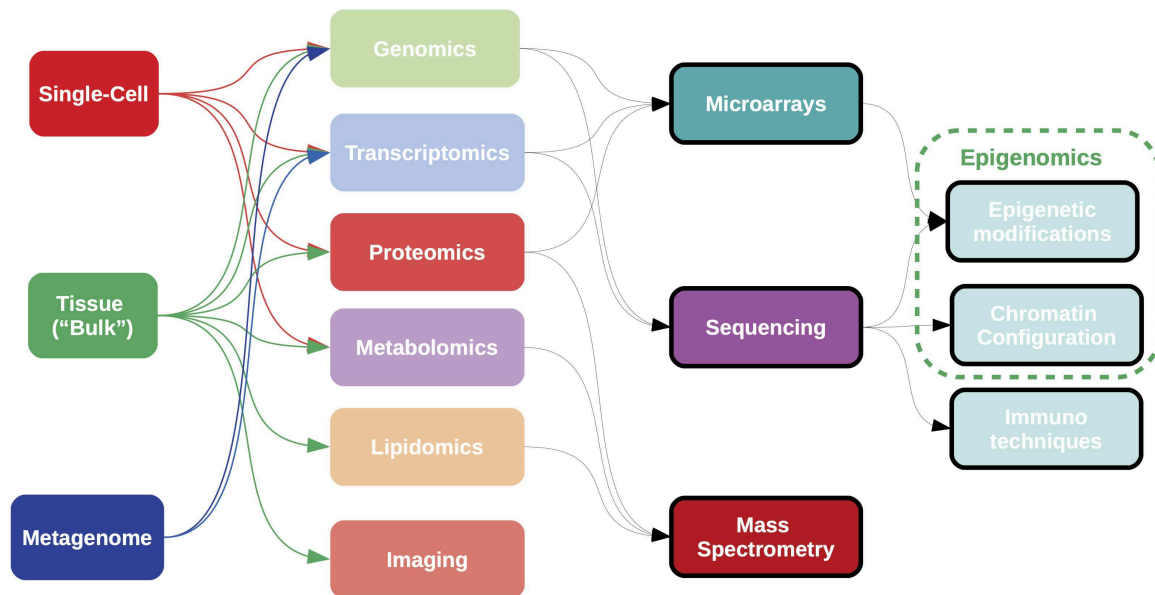


FIGURE 5 | Samples for omics analyses can be obtained from “bulk” tissue, single cell data, or heterogeneous populations, such as metagenomes. Most current omics data are generated using technologies either array-based, sequence-based, or mass spectrometry-based; although high-throughput imaging data is becoming important in the clinical setting. Complementary techniques exist for the analysis of epigenetic states. Each combination of sample type, omic measurement and analytical technology requires a specific bioinformatic pipeline for data acquisition and processing.

variations (CNVs), and so forth. The relationship between genomic alterations and cancer is well-known (28).

Microarrays have long been used for genotyping. Although specifics of microarray technology may vary across manufacturers, most modern DNA microarrays can be analyzed using well-established tools available in the *Arrays* (29). Such tools can handle arrays for different genotyping tasks, including SNP and copy number assays [for instance, copy number detection from exome sequencing using *CODEX* (30)].

Although DNA microarrays remain in use, next generation sequencing (NGS) technologies are quickly becoming commonplace. The analysis of NGS data entails a workflow that involves sequence acquisition and alignment to a reference genome. A number of downstream analysis pipelines can follow; for instance, a variant discovery workflow would involve variant calling, filtering, annotation, and prioritization (31). The first step to analyze NGS data is to use a sequence aligner tool on the sequence data (stored in FASTQ format). Some popular aligners are the stand-alone *BWA* (32), *Bowtie* (33), *Bowtie2* (34), and *SNAP* (35), with aligned sequences being stored in SAM (Sequence Alignment Map, text-based) or BAM (Binary Alignment Map) files. These aligned sequences are the input for downstream genotyping analyses (36, 37).

Such *standards* are indeed a matter of state-of-the-trade in the academic research community indeed. Regarding pipelines approved by regulatory instances, there is in fact an official FDA guideline document to this end: “Considerations for Design, Development, and Analytical Validation of Next Generation Sequencing (NGS)—Based *in vitro* Diagnostics (IVDs) Intended to Aid in the Diagnosis of Suspected Germline Diseases”

available for download at <https://www.fda.gov/media/99208/download>. The Guideline document (99208) actually refers to a Software Documentation Guideline: “General Principles of Software Validation; Final Guidance for Industry and FDA Staff” which is however quite outdated (last revised January, 11, 2002) (<https://www.fda.gov/media/73141/download>). Some NGS tools however are actually available as a web service at <https://precision.fda.gov/>. For a review on these guidelines and tools see (38).

3.1.2. Epigenomics

With the recent advent of high-throughput omic technologies to probe chemical modifications in the tumor genomes it has become more and more evident that such epigenomic modifications are present and likely play relevant roles in many cancers. These variations include DNA methylation and histone modifications, both in oncogenes and in other cancer-associated genes. Mutations in genes involved in epigenetic regulation have also been found in several tumor types. The computational analysis of epigenomic data may provide us new insights about cancer initiation and progression. More relevant perhaps, such studies will pave the way for a more efficient identification of genetic and epigenetic biomarkers for diagnosis, prognosis or response to therapy. These in turn, may accelerate the development of novel therapeutic approaches.

Epigenomics often presents another view of functional processes complementary to that of genomics. Sometimes epigenomic techniques even allow for a better understanding of genome-associated phenomena. Such is the case of high-throughput immunoprecipitation assays, such as ChIP-Seq.

ChIP-Seq and other experiments based on the analysis of short reads show the effects of multi-reads, i.e., reads that map to more than one genomic region. Determination of the origin of such multi-reads indeed results critical for the accurate mapping of reads to repetitive regions, such as copy number variants (39, 40). Current computational approaches have been refined to cover up for this phenomenon even at the single-cell level (41).

The epigenome contains the set of potentially inheritable chemical modifications of DNA and histone proteins that can control gene expression activity (42). There are several mechanisms which are contained within the epigenomics concept, each requiring a different high throughput molecular technique for its measurement. Each of these techniques, in turn, requires the use of a dedicated set of computational tools. These include:

- **DNA methylation:** The methylation state of a DNA region can alter its transcriptional activity. This state can be measured using either array-based methods or sequencing methods, such as the popular whole-genome bisulfite sequencing (WGBS) (43). Data from array based methods can be processed using the aforementioned array packages, along with dedicated packages, such as *methylationArrayAnalysis* (44). Similarly, those obtained using sequence-based methods can make use of dedicated tools, such as the *bsseq* (45) or *methyAnalysis* (46) packages.
- **Chromatin remodeling:** Regions where nucleosomes are sparse and physical access to the DNA sequence is enabled are identified as open chromatin. Chromatin accessibility is a dynamical and complex framework modulated by diverse elements, including nucleosome occupancy and turnover rate, histone modifications, ATP-dependent chromatin remodeling complexes and even TF binding (47, 48). Open chromatin has emerged as indicative of transcriptional regulatory potential or activity across the human genome because most of the TFs analyzed to date bind within open regions (49). Chromatin architecture is modified by changing its accessibility affecting gene expression rates. This remodeling can be controlled by histone modifications, which include acetylation, methylation, ubiquitination, and SUMOylation, among others. Overall chromatin accessibility can be also measured by techniques, such as ATAC-seq (50), a high throughput NGS technique to assess genome-wide chromatin accessibility. Due to the characteristic biochemical design of the assay ATAC-seq is a faster and more sensitive analysis of the chromatin accessibility than other alternatives, such as DNase-seq.

ChIP-seq (51) data is used to identify genomic locations with an overabundance of proteins of interest; such identification uses the so-called *peak callers* (52, 53). These include *SICER2* (54), *PeakRanger* (55), *GEM* (56) *MUSIC* (57), *PePr* (58), *DFilter* (59), and *MACS* (60); benchmarks for these algorithms can be found at <https://github.com/skchronicles/PeakCalling>.

MACS is a popular peak caller that uses dynamic Poisson distribution; its successor, *MACS2* (61), improves the algorithm to, amongst other things, make it more suitable for calling differential regions. Differential binding

analysis (that is, identifying sites in which exhibit a different binding behavior between biological conditions) can be useful to identify relevant regions that may be driving cancer phenotypes, using ChIP-seq data. Tools for this task include *DiffBind* (62), a package that provides functions to handle the results of peak set callers, such as *MACS*. Another tool for this task is *csaw* (63), useful for de novo detection of differentially bound regions using a sliding window approach. In-depth comparison of differential ChIP-seq analysis tools can be found in (64).

- **Chromosome conformation:** The three-dimensional organization of the genome allows for interactions between regions that are distant in terms of sequence, even belonging to other chromosomes. These higher-order chromosome structures are a current area of research in oncology (65). Chromosome configuration capture techniques are able to quantify interactions between genomic loci. These *C-techs* are based on the original 3C, *Chromosome configuration capture* (66); able to quantify interactions between a single pair of loci. It was followed by: 4C (*Chromosome configuration capture-on-chip*) (67), which captures interactions between one locus and all others; 5C (chromosome conformation capture carbon copy) (68), which captures all interactions between two sets of loci; and Hi-C (high-resolution chromosome conformation capture) (69, 70) to detect interactions between all possible loci pairs. Development of computational analysis tools for chromosome conformation capture data is ongoing, although there are available packages for the detection of significant interactions for all these technologies (71–73).

It has been known for some time that higher order chromatin arrangements are associated with chromosomal alterations in cancer. For instance, it has been argued that spatial chromosome conformation and negative selection may be powerful driving forces behind somatic copy number alterations (74). More recently, chromatin conformation capture has allowed the identification of putative pharmacological targets in breast cancer (75). Genomic loci interactions may even affect the expression of biomarkers related to hallmarks of cancer, such as hypoxia (76).

Packages, such as *methyPipe* and *compEpiTools* provide an integral platform for the comprehensive and integrative analysis of the first two classes of epigenomic data (77), whereas *ATACseqQC* (78) is a package offering quality control tools for ATAC-seq data, while *esATAC* (79) offers a whole analysis pipeline and the *GenomicInteractions* package (80) offers a complete framework for the analysis of chromosome conformation data.

3.1.3. Transcriptomics

Transcriptomic analyses are used to measure the presence and abundance of RNA in a given physiological context (81). Perhaps the most common application of transcriptomic technologies is to measure gene expression. The gene expression profile of a phenotype can be used as a barcode of its biological state. Such barcodes can be compared, through differential expression analyses, to pinpoint cellular changes in cancers (82). The expression profile is the product of the gene regulatory program

encoded in the genome and the epigenome. By measuring gene expression, we are indirectly capturing the regulatory changes that are at the core of the disease.

The development of gene expression microarray technology (83) has made gene expression measurement more technically and economically viable than the measurement of protein abundance. Therefore, methods for the measurement of biological activity (i.e., pathways) have been developed with transcriptomic data in mind (84). Studying the molecular phenotype of cells via transcriptomics has become an invaluable tool providing a proxy to the functional state of cells and its regulatory interactions, both in cancer (85, 86), and in healthy phenotypes (87). Nevertheless, it should be noted that the correspondence between gene and protein abundance is far from perfect (88), which highlights the need for multi-omics.

Beyond gene expression, whole transcriptomic analyses involve the measurement of non-coding (nc) RNA, such as micro-RNA (miR), long non-coding RNAs (lnc-RNA), small nucleolar, Piwi-interacting, enhancer RNAs, among others (89, 90). The role of these transcripts, particularly in terms of their contribution to the regulatory program, remains an active area of study.

As previously mentioned, transcriptomic technologies are one of the most developed omics, second only to genomics itself. Measurement of transcript abundance can be done using either expression microarrays or RNA-sequencing (91, 92). Each methodology has technical considerations, but the general steps for their analyses are similar: acquire and preprocess data, removing technical artifacts; quality control; and data normalization. The resulting data can be represented as an expression matrix: an NxM matrix where rows represent transcripts, and columns represent samples (or observations). It should be noted that most expression pipelines are oriented toward differential expression analyses [see for instance (93)]; this should be taken into account in case that is not the intended use-case.

Starting points for RNA-seq data analysis include either alignment based methods, such as *Bowtie* (33), and *STAR* (94), or alignment-free methods, such as *kallisto* (95) and *Salmon* (96).

Cancer-related omic experiments often rely on specific, tailor-made analytics. One instance of this is provided by alignment-free RNA-Seq analysis methods, such as the ones performed by *kallisto*, *Salmon*, etc. Alignment-free methods (AFMs) are particularly well-suited to study cancer transcriptomics to look up at the role and abundance of fusion transcripts that may give rise to chimeric proteins (97, 98). Another reason behind the use of AFMs is that it is known that different RNASeq pipelines present differences that may be important when analyzing cancer genomes and transcriptomes (99, 100).

Further require different tools for quantification, quality control, and normalization of expression data. For instance, a popular pipeline is composed of the aforementioned *Bowtie* as a short read aligner, *TopHat* (101) for the identification of splice junctions, *Cufflinks* (102) for transcriptome assembly and differential expression analysis, and *CummeRbund* (103) for result exploration; it should be noted that, while this pipeline is still widely used and maintained (e.g., *Bowtie2* latest release was

02/28/20), other approaches are being gradually embraced by the community (104); for instance, the *HiSat2* (105), *StringTie* (106), and *Ballgown* (107).

In the case of tools like *STAR*, we need to be aware that fusion detection using *STAR-fusion* is mainly limited by the length of single-end reads. The *STAR-fusion* wiki (<https://github.com/STAR-Fusion/STAR-Fusion/wiki>) indicates the need for at least 100 base length. In the case of other approaches, such as *FusionHunter* (108) the authors recommend to align to a pseudo-reference and discard junction spanning reads with <6 bp matches on either gene. *Arriba* is a relevant tool to call for gene fusions, based also in the *STAR*-alignment (<https://github.com/suhrig/arriba/>). *Arriba* was the winner of the DREAM SMC-RNA Challenge (<https://www.synapse.org/#!/Synapse:syn2813589/wiki/401435>) (109).

An advantage of the modular design of these pipelines is that it is possible to combine tools from different workframes, depending on experimental and analytical needs: For instance, *Salmon* provides tools to connect with differential expression tools, such as *DESeq2* (110), *edgeR* (111), *limma* (112), or *sleuth* (113). A detailed discussion of these methods is beyond the scope of this article; please see Conesa et al. (114) for an in-depth review.

3.1.4. Proteomics

Proteomic analyses are used to identify and quantify the set of proteins present within a biological system of interest (115). The study of cancer proteomes is promising as a way of identifying biomarkers and therapeutic targets (116). This is not surprising: proteins are the molecular unit from which cellular structure and function arises.

Historically, high throughput proteomics technologies have developed at a slower pace than genomics and transcriptomics technologies. Microarray approaches to proteomics have been developed, with varied levels of success and applications (117, 118). However, the bigger breakthroughs have come through the use of mass spectrometry (119).

Various steps of proteomics analysis involve data analysis (120). During data acquisition, the detected molecular fragments must be identified. This is often done by comparing fragments to databases in real-time (121, 122). Later, the assembly of proteins from identified peptide fragments requires another set of computational methods (123). The development of such methods remains an active area of research (124, 125). The *Bioconductor* offers a streamlined set of tools for the management of proteomics data, from data processing to functional analysis (126). Another alternative for protein quantification is the *maxquant* toolset (127).

3.1.5. Metabolomics and Lipidomics

Metabolic alterations are important contributors to cancer development (128). Cancer metabolomics has become an important research topic in oncology (129), with the promise of providing novel insights on cancer development and potential therapeutic options. Lipidomics is actually a subset of metabolomics (130). The study of cancer lipidomics may lead to

the identification of biomedical important findings, such as novel biomarkers (131).

Like proteomics before, metabolomics and lipidomics studies have been possible thanks to the use of mass spectrometry. The analytical considerations for the extraction and quantification of these types of compounds have some differences to those used for proteomics. This is expected, as the chemical nature of metabolites and lipids are fundamentally different (132, 133). In turn, bioinformatic and chemoinformatic approaches to high-throughput metabolite profiling exhibit some modifications (134).

Analysis frameworks for metabolomic and lipidomic data are currently available. The *metab* package (135) provides an analysis pipeline for metabolomics derived from gas chromatography—mass spectrometry data. The *metaRbolomics* package (136) is a general toolbox that goes from data processing to functional analysis. Finally, the *lipidr* package (137) is a similar framework focused on lipidomics data.

3.1.6. Unraveling the Complexity Within Samples: Single Cell, Imaging, Microbiome

The aforementioned technologies were all developed for the detection and quantification of analytes extracted from a complex biological matrix, obtained from tissue, plasma, or a similar fluid. As such, the data from these omics is an aggregate of the different cellular contexts present in the sample. The environment within and surrounding cancer tumors is notably heterogeneous (138, 139). There is knowledge to be gained by recovering the omics diversity within samples.

Cancer is an extremely heterogeneous disease at the cellular and molecular level. Tumor heterogeneity caused by the concurrence of multiple cell lineages and differentiation stages, determined to an extent by the processes of clonal evolution. This has led to an early adoption of single cell analysis techniques. The case of single cell sequencing to study the genomic and epigenomic features of the different cell populations within a tumor by considering the characteristics of individual cells has revealed as an appealing approach to deal with said cell-to-cell variability (140–142).

Cancer cell heterogeneity also exists beyond the genome. Tumor evolution under complex environmental scenarios often leads to variability in epigenetic modifications. Single cell sequencing and imaging techniques have proven to be quite effective to characterize cellular plasticity induced by epigenomic phenomena (143). Aside from scMethSeq, and scDNase Seq, other techniques, such as single-cell chromatin accessibility assays are starting to shed light to how epigenomic subpopulations in cancer may have the potential to impact tumor features, such as drug sensitivity and clonal dynamics (144).

Single-cell omics analyses rely on experimental techniques for the isolation of single cells from a sample, using microfluidics or fluorescence-activated cell sorting methods (145). Single-cell RNA-seq (scRNA-seq) is currently the most developed high-throughput omics technology for individual cell analysis (146).

Data from scRNA-seq experiments can be thought to be very similar to so-called “bulk” data. Data from scRNA-seq is, in fact, sparser, more variable, and with more complex expression values

distributions. As such, data analyses techniques may need to account for different assumptions than their “bulk” counterparts (147). Again, the development of these novel bioinformatics tools is an active area of research (148). The *Bioconductor* ecosystem has a complete framework for the analysis of scRNA-seq from low-level (149) to functional analyses (150). *Scanpy* (151) provides a toolkit for single-cell gene expression analysis in a Python environment. Another single-cell genomics toolkit is *Seurat* (152) for R.

Integration of single-cell RNA-seq with other profiling tools is an important research area (153); as along with *single-cell*, there are other technologies that can provide a more complete picture of the cancer heterogeneity. High throughput imaging techniques (154) can be generated and computationally analyzed (155, 156). Imaging techniques can be used along with omics to recover the spatial distribution of molecules within cells and throughout tissues. Tools, such as *CellProfiler* (157) allow for a high-throughput analysis of data. Imaging techniques can be combined with single-cell methods: for instance, *MERFISH* can simultaneously measure copy number and distribution of RNA in single cells (158); *Slide-seq* (159) can measure transcriptomes at a high spatial resolution.

Space-resolved transcriptomics or spatial transcriptomics (ST) is a set of *in situ* transcript capturing methodologies aiming at quantification and visualization of gene expression patterns in individual tissue sections or regions. ST methods have indeed revealed relevant tissular phenomena linked to tumor evolution and in some cases have been able to allow the prediction of clinical outcomes in, for instance, breast cancer subtypes (160).

ST mapping of prostate tumors, on the other hand, have resulted key in the identification of gene expression gradients in stroma adjacent to tumor regions. This in turn has resulted in patient re-stratification based of tumor microenvironment features (161). A similar approach has been taken to trace tumor advance in malignant melanoma (162). A combination of ST with scRNASeq has led some researchers to propose the concept of a “tumor atlas,” a roadmap to navigate tumor spatial and cellular heterogeneity (163).

Multi-omic analysis is not devoid of technical and logistic conundrums. Perhaps the most obvious is the availability of the different sample types from a single source in the same experiments. Cell cultures may provide a way out to this problem, however *in vitro* conditions are often not resembling some aspects of interest in complex phenotypes, such as cancer. In recent times, three dimensional cell culture techniques have allowed the design and development of more *realistic* models, such as the case of organoids and tumoroids. These models may represent a good compromise between cell line studies and biopsy-captured tissue experiments (164). Multi-omic approaches are starting to be applied on lab-grown organoids with relative success (165, 166). In order to analyze such data some novel computational tools are being developed and adapted (167).

The role of the immune system in cancer response is another area of active research. CITE-Seq is an RNASeq method that incorporates epitope analysis thus leading to semiquantitative information regarding surface protein abundance via antibody

assays, even at the single cell level (168). This novel technique is starting to be applied to provide the answer to fundamental questions in oncology, such is the case of tumorigenesis (169)

Finally, the role of the microbiome in cancer is being recognized (170); the integration of metagenomic, and perhaps *meta-omics* data (171), could provide key insights into cancer pathogenesis and therapeutics.

3.2. Data Management

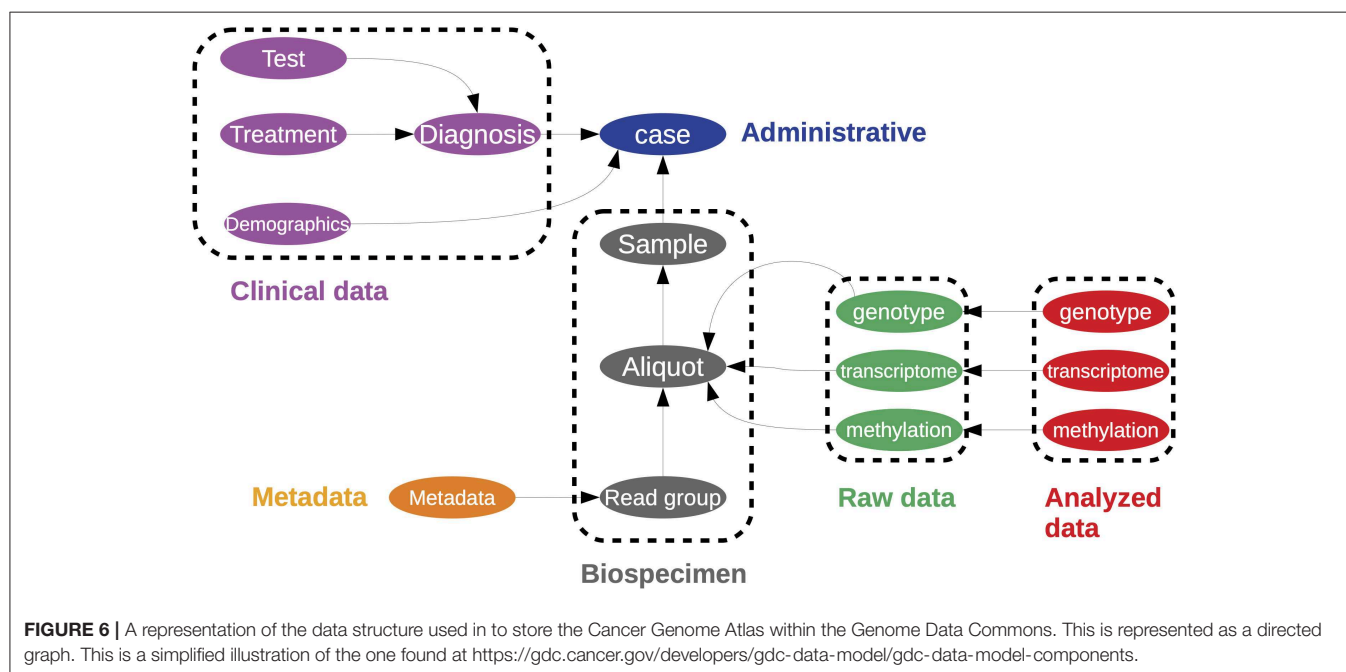
The push for open data in the field of biomedical genomics since the gestation of the Human Genome Project has led to the emergence of a rich Genomic Commons (172). Making data available in public repositories makes for faster scientific discovery, although there are challenges to be overcome, both ethical/legal (173), and technological.

Challenges of data management include defining the type of data to be stored and how to store it; the policies for data access, sharing, and re-use; and long term archiving policies (174). Arguably, the most successful repository of cancer multiomics is NIH's Genome Data Commons (GDC) (175). The Genome Data Commons contains all data generated by the Cancer Genome Atlas (TCGA) project (176); although it should be noted that not all data is publicly accessible. The data is organized as a directed graph comprised of interconnected entities (**Figure 6**), with each entity having an associated set of properties and links. Data is publicly accessible either through the *gdc-client* command line tool, the REST API for programmatic access to the database, or through dedicated packages, such as *rtcg* (177). A recent account by *The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (PCAWG)* of these resources and analyses is presented in (178). Furthermore, a larger collection of datasets can be accessed through the Broad Institute's *Firehose* (<http://gdac.broadinstitute.org/>); cloud computing enabled data

access is provided through the Cancer Genome Collaboratory (<https://cancercollaboratory.org/>).

The impact of TCGA at the forefront of multiomics research is inarguable. As a publicly available resource, it provides data for method development and validation. This is used by a lot of current projects. However, there are other datasets with either single layer or multiomic datasets that can also be integrated. And wetlab researchers still carry out their projects, contributing to the cancer multiomics community. Integrating data from both, local experimental projects and large collaborative endeavors, such as TCGA is indeed a common practice in many places, such as our institution, the National Institute of Genomic Medicine in Mexico. Doing so allows to contrast specific hypothesis for the different research groups with the statistical power obtained via the much larger datasets generated by international multicentric collaborative projects.

As mentioned, it is possible to extract a lot of knowledge from the systematic re-analysis of data available in large public datasets. Perhaps, the more comprehensive of these databases is the one by the TCGA/Genome Data Commons/International Cancer Genome Consortium, TCGA. Retrieving the data via their Application Programming Interface (API) (<https://gdc.cancer.gov/developers/gdc-application-programming-interface-api>) demands some familiarity with command line tools and coding that may be beyond of most non-bioinformaticians. The project's data portal (<https://portal.gdc.cancer.gov/>) provides easy to use interfaces, but may be limited on its application to broader analyses. To date there is a number of commercially available platforms that provide a gentler access to the TCGA data. Such is the case of Qiagen's OncoLand database (<https://digitalinsights.qiagen.com/products-overview/discovery-insights-portfolio/content-exploration-and-databases/qiagen-oncoland/>) and the cloud-based analytics solution Seven Bridges



(<https://docs.sevenbridges.com/docs/tcga-data>). A limitation, aside from being subscription based alternatives that require a payment is that they are not customizable, which means that not all possible (nor desired) analysis may be performed.

There are, however a number of resources not only to access the data but to actually perform different levels of downstream analysis. Such is the case of imputation approaches to missing data in the TCGA database (179) (<https://github.com/mrendleman/MachineLearningTCGAHNSC-BINF/>).

Perhaps, the best combination of usability and versatility is present in the TCGA Workflow suite available as an R/Bioconductor package (180) (<https://www.bioconductor.org/packages/release/workflows/vignettes/TCGAWorkflow/inst/doc/TCGAWorkflow.html>).

4. COMPUTATIONAL TOOLS FOR MULTI-OMICS DATA INTEGRATION

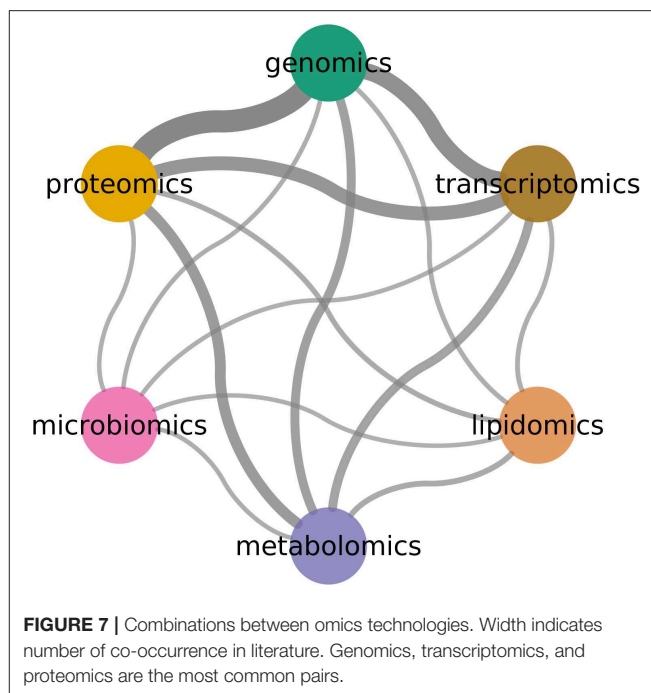
An often-asked question is why try to integrate multiple omics technologies using complex models. Perhaps the simplest argument is that the biological phenomena is not comprised of independent layers of biological features: integrative models will be, due to this simple fact, closer to the system of study. As omics technologies become available, researchers have used them together to try and capture a better description of the phenomena (see **Figure 7**).

Improving our current cancer diagnostic capabilities is a major goal of biomedical research: the role of molecular technologies in the development of these tools has long been recognized (181). It is expected that multi-omic integration is able to provide better predictive tools than single molecular technologies, due to the fact that each technology is capturing just a slice of the whole complex pathological system; multi-omics data are expected to be of value for both basic and clinical research, as long as they are able to recover biological insights beyond those obtainable from the simple addition of each analysis layer (182, 183).

It may soon become evident that the formalisms that can lead to such level of description are, by necessity, complex (184). A remaining question is what multiomic combinations are able to achieve better diagnostic results. Selecting this optimal omics combination is not trivial, since there are practical constraints (such as economic and technical limitations) in the clinical setting in which such diagnostic tools are to be deployed (185). Computational tools and bioinformatic approaches play an important role in the design of such studies. A list of such tools is presented in Supplementary Materials as **Table 1**.

4.1. Multi-Omics Data Representation and Preparation

The success of a computational method could arguably be influenced by the design principles implemented in its data representation. The *MultiAssayExperiment* package (186) provides an eponymous data class to contain multi-omics experiments. Like other *Bioconductor* classes, *MultiAssayExperiment* is object-oriented. It can contain the



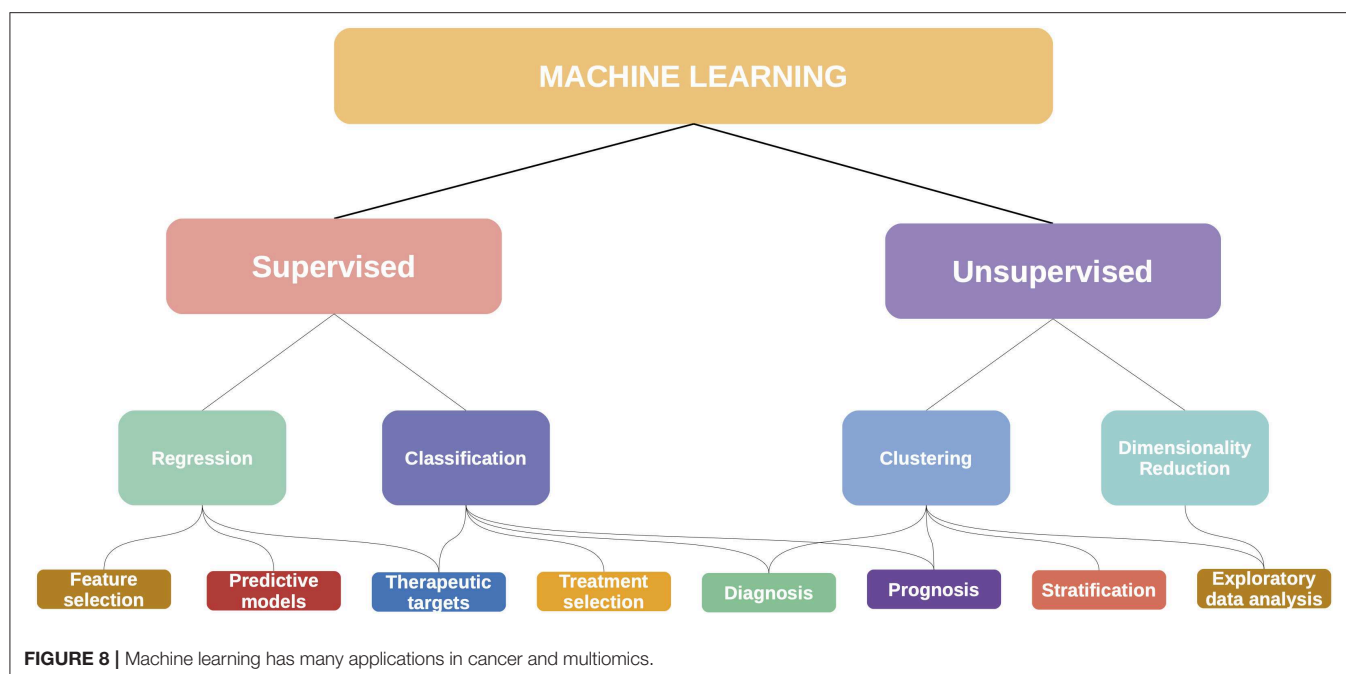
information of different (multi-omics) experiments, linking features, patients, and experiments. Furthermore, by sharing design principles with the rest of the *S4-Bioconductor* classes, it is highly interoperable.

An important issue with large scale multi-omics studies is the problem of missing and mislabeled samples. Whether by technical limitations or human error, the samples associated with a given patient may not have all measurements; or samples from two different patients may get mixed-up. There are packages available to handle these problems. The *missRow* package (187) can be used to handle missing data, combining multiple imputation with multiple factor analysis. The *omicsPrint* package (188), in turn, can be used to evaluate data linkage through the use of linear discriminant analysis.

The *STATegRa* (189) project provides a framework for multi-omics data analysis and integration: these are *MixOmics* (190), descended from the *integrOmics* project (191); and just like the *Bioconductor* project, the major advantage of such projects is the increased interoperability due to the sharing of design principles. For instance, within the *STATegRa* project, there is an Experiment Manager System (192); *MOSim* (193) a tool that provides methods for the generation of synthetic multi-omics datasets. These datasets can be used for the benchmarking and validating of other integration tools; and an experimental multi-omics dataset (194).

4.2. Multi-Omics Data Integration as a Data Science Problem

For this review, we approached these methods from a *data science* perspective, considering that each method is in essence solving a machine learning task (or set of tasks). In **Figure 8** we show



some of these mappings, although it should be noted that these categories may be fluid: an unsupervised clustering analysis can become the basis for a supervised classifier, with diagnostic and prognostic applications. This is the story of the PAM50 algorithm for breast cancer (195).

4.3. Exploratory Data Analysis

Exploratory data analysis (EDA) is a vital first step in omics analyses (196). Through EDA the nature of the data can be understood, allowing for better decisions at a further modeling step.

Unsupervised learning approaches can provide a hypothesis-free understanding of the data behavior. This will reflect the nature of the underlying biological phenomenon. *Unsupervised clustering analyses* attempt to group samples based on the similarity of their measured features. The assumption is that this unsupervised classification will recover relevant biological differences. Multi-omics can increase the efficiency of such approaches (197).

Multi Omic data analysis is often performed with the aim of unveiling non-trivial molecular and systemic interactions that are difficult or impossible to see if one relies on a single omic approach. However, since we are tacitly assuming that the different omic levels of description may have synergistic effects that are key to develop more accurate models of tumor biology. Since multi omic approaches may generate a plethora of interdependent data it is useful to design analytical strategies for dimensionality reduction, feature selection and integration of all this information.

Aside from intelligibility, there are additional reasons to make dimensionality reduction schemes, one of these is that a multi omic study combines different information sources, hence dramatically increasing the number of features, often keeping

the number of samples constant, in order to preserve statistical power we need to rely only on the most informative variables (198–200).

Computational tools to this end have been developed, such as the following: <https://www.bioconductor.org/packages/release/bioc/html/mixOmics.html> <https://bioconductor.org/packages/release/bioc/html/STATegRa.html> For an extensive list of computational tools in the context of cancer biology, see (186).

One can make use of *dimensionality reduction techniques* in order to embed multi-omic data observations into a lower-dimensional space that can be used for either manual (i.e., visual) inspections or as the input for unsupervised clustering (or other analysis tools). Popular dimensionality reduction methods:

- Principal Component Analysis (PCA) is a classical (201) method based on an orthogonal transformation of the set of observations.
- T-distributed stochastic neighbor embedding (t-SNE) (202) is a method based on the minimization of the Kullback-Leibler divergence between the probability distribution of pairs of high-dimensional objects.
- The Uniform Manifold Approximation and Projection (UMAP) (203) is a non-linear technique in which data are projected into a Riemannian manifold.

Data visualization is an important part of EDA: the graphical representation of data can be sufficient for the identification of complex patterns (204). Visualizing high-dimensional biological data can be helpful from a purely data-driven point of view: for instance, to understand the variability within a phenomenon. Combinations of dimensionality reduction, data clustering, and visual inspection can be effective to identify subpopulations within a dataset. The most common visualization for these tasks is perhaps the scatterplot, but it is far from the only: for instance,

hexbins (205) can be used to explore sc-RNAseq data, which can be useful to overcome overplotting problems related to the order in which points are drawn in the canvas.

Visualization can also be coupled with other biological information, for instance locating the genomic regions in which epigenomic features are found. Visualizations, such as the *Circos* plot (206) can be used for the detailed representation of multi-omics data and their location in specific genomic regions; The *omicCircos* (207) implementation is compatible with the standard data classes used in *Bioconductor*. The multiOmicsViz *multiOmicsViz* package is useful to visualize the effects of one omics layer to another, visualized in within the spatial chromosome context. The *Gviz* package (208) provides a full R graphics system solution for genome browser-style visualizations. Such representation is useful to represent the behavior of different experimental layers (as tracks) in a sequence context. For ChIP-seq data visualization, tools like *PAVIS* (209) may be used. Single Cell RNA-seq data visualization suites, such as *SingleCell Signature Explorer* (210) can be useful for exploratory analysis of such datasets. In the case of chromatin capture data, visualization toolboxes, such as *HiBrowse* (211), the *Epigenome Browser* (212), and *Juicebox* (213). For a thorough review of Hi-C visualization consult (214).

Common exploratory data analysis tools are implemented either in base R or as packages from CRAN (since their use is not necessarily limited to biological data). However, there are packages providing integrated EDA tools for multi-omics and oncology. The *OMICsPCA* package (215) provides omics-oriented tools for PCA analysis. The *CancerSubtypes* package (216) contains several data preprocessing, quality control, and clustering methods, focused on the identification of cancer subpopulations from multi-omics data. *Biocancer* (217) provides an interactive multi-omics data exploratory toolkit. The *omicade4* package (218) provides an implementation of multiple co-inertia analysis (MCIA), another dimensionality reduction technique; these tools were used for the integration of transcriptome and proteome data from the NCI-60 cancer cell line panel. The Multi-omics Autoencoder Integration (*maui*) is a tool for multi-omics data analysis for Python. It allows for latent factor model coupled with artificial neural networks for multiomics data integration. *iClusterPlus* is a Bioconductor package based on the original *iCluster* (219) algorithm for integrative cluster analysis combining different types of genomic data.

4.4. Statistical Models: Classifiers, Predictors, and Feature Selection

Exploratory methods provide a useful description of biological phenomena. Nevertheless, in the oncology context, the identification of actionable elements is most desired, to generate translational value. The generation of models and feature selection strategies can lead to such results.

In this context, *statistical models* are computational (and thus mathematical) representations of the relationships between observed variables. These models can be useful to solve a given

task based on some input data (220). Examples of these tasks include the *classification* of samples and the *prediction* of the state of a feature of interest.

Classification models have important biomedical applications (185). If a classification is able to discriminate between physiological states it can have translational use: A model that discriminates between health and disease has *diagnostic* utility; A model that discriminates between different disease outcomes has *prognostic* utility, which can be used for *stratification* purposes. Molecular classifiers have been quite successful in oncology: perhaps the best example being breast cancer (221). Classification models can be developed using *supervised* methods (that is, the model is trained with class information); but *unsupervised* methods, such as the previously discussed clustering, may be able to recover groupings that capture biological and clinical differences.

Predictive models can provide insights into the molecular mechanisms driving physiological states. These can reveal the interactions between different omics, as well as between individual biomolecules. Furthermore, predictive models can have translational applications, including their use in prognostic tools (222).

Feature selection consists in the selection of a subset of measured variables that are most informative: that is, they contribute the most for the model to accomplish its task. Proper feature selection is important for biomedical models (223), as (1) removing uninformative (“irrelevant” or “redundant”) features simplifies the model and increases its performance; and (2) a smaller set of features is less expensive to measure, increasing the translational potential of a given model.

Common applications of statistical models in the clinical context of cancer are the prediction of susceptibility, recurrence, and survival (223). Additionally, classification and association models are regularly used for the interpretation of molecular studies of cancer. For instance, biomarker discovery (224) is an often sought target for modeling based on biochemical and multi-omics analyses. This is an important area of study, since actionable biomarkers are not particularly common (225).

4.4.1. Implementations and Use-Cases

Novel tools for the implementation of oncology models using model data are being released constantly. Many of these packages combine exploratory, supervised, and unsupervised tools, providing a wide range of analysis tools. *mixOmics* (190) is a self-described omics data integration project; it includes an eponymous package that provides different exploratory and integrative multivariate methods, including (independent) PCA, Canonical Correlation Analysis, Partial Least Squares regression (PLS), and PLS-Discriminant Analysis (DA). Part of the larger project is the *Data Integration Analysis for Biomarker discovery using Latent Variable approaches for Omics studies* (DIABLO) framework, which has been used for the identification of a multi-omics signature of breast cancer molecular subtypes (226).

Other tools also follow this combined design principle. The *ropls* package (227), for instance, incorporates the tools for PCA, as well as (Orthogonal) PLS. Multi-Omics Factor Analysis (MOFA) is implemented in the eponymous package (228).

This factor analysis model has been used for the unsupervised detection of groups in a leukemia dataset, and the selection of informative multi-omic features associated with oxidative stress. *OmicsMarkeR* (229) also provides a variety of classification and feature selection tools; originally developed for metabolomics, this tool has been used for the study skin cancer progression (230). Some packages include different classifier methods to generate an ensemble model; such is the case of *Biosigner* (231) which combines PLS-DA, Random Forests, and Support Vector Machines to select discriminant features across omics.

We agree with the assumption that multi-omics specific tools can improve workflows by adhering to a single design philosophy. However, we also agree that this is convenient, but not necessary. For instance, a diagnostic panel for pancreatic cancer was recently identified with a Random Forest implementation (232) using genomics, transcriptomics, and immunohistochemistry data. In another study, biomarker candidates for pancreatic cancer are identified using a Support Vector Machine on miRNA and gene transcriptomics (233).

Predictive models can be used to identify the contribution of one omics layer to the activity of another. For instance, *epigenomix* (234) uses Bayesian mixture models to integrate ChIP-seq and gene transcription data. The *Integrative analysis of Multi-omics data for Alternative Splicing* (235) package integrates expression, sQTLs, and methylation to provide mechanistic insights behind the manifestation of alternative splicing.

Predictive methods have been used to integrate multi-omics with other sources of big data, with publicly available implementations. The packages *rexposome* and *omicRexposome* (236) have been used to study the *exposome*, defined as the set of environmental exposures. Using multi-canonical correlation analyses and multiple co-inertia analysis, exposome-wide associations have been made to multi-omic data. The *OmicsLonDA* package (237) offers a method that uses linear mixed-effect models and smoothing spline regression models to identify time periods with differential omics levels. A highlight of this package is the consideration for the use of physiological measurements from wearable sensors, which may provide applications for *nowcasting*, the prediction of near-future states.

4.4.2. Functional Aggregation

One could argue that analysis methods can be more informative if there is a way of associating the findings to the wider body of biomedical knowledge. Mapping omics data to functional features, such as pathways and functional genesets, is a strategy that can provide such readily interpretable results. *Functional enrichment* approaches, such as *over-representation analysis* (ORA) and *gene-set enrichment analysis* (GSEA), are effectively *feature extraction* methods that can be used as biologically relevant dimensionality reduction methods. The results of such methods can serve as starting points for more complex models, such as interactions among functions (238). For a detailed discussion of functional analysis, see (84).

The development of methods for effective functional enrichment based on multi-omics data is ongoing. *Multi-omics gene-set analysis* (MOGSA) (239) approaches the problem by using multivariate analysis, and using projections of data and

genesets to lower dimensional spaces, to generate an enrichment score. *Massive integrative gene set analysis* (MIGSA) (240) takes a different approach, making independent functional associations for each omics layer (using ORA and Functional Class Scoring). Instead of providing an aggregated measurement, the functional associations of each layer are stored in a special data structure, allowing flexible analyses. This method has been used to functionally characterize breast cancer molecular subtypes from a multi-omics perspective.

Functional aggregation can be used as the basis for other data analysis tasks. In *pathwayPCA* (241), exploratory data analysis is done by analyzing the functional enrichment of each omics set separately, and aggregating them via consensus. This method was used to study heterogeneity in an ovarian cancer dataset. In the original work for the *Divergence analysis* (242) method for high-dimensional omics data analysis, the authors evaluate the effect of using functional aggregation for their data classification task. Functional aggregation methods are an important part of high-throughput drug initiatives, as can be seen by their prominence in the iLINCS platform (243).

4.5. The Network Paradigm

As we have stated throughout this work, biological phenomena are complex, interconnected systems. The data that we recover from high-throughput multi-omics is not isolated. Any biological system is not just the sum of its parts, but the sum of its biological elements *and their relationships*. With this in mind, the integration of high-throughput data within a network paradigm becomes appealing. Some advantages of a network approach to multi-omics integration are:

- A network representation of multi-omics data can be studied using all the foundations and tools of network science (244). Network topological parameters can be associated with important biological features; furthermore, dynamical processes can be modeled over networks.
- As previously noted, the functional level of biological description is fundamentally composed of molecular interactions. In other words, measurable functions can be thought to emerge from biological networks. Functional analyses can benefit from considering the way in which the participating molecules interact.
- The integration of interaction information can lead to more informative models (245).

A network perspective can enhance every aspect of the multi-omics analysis. For instance, mapping omics data to pathway networks can provide an opportunity to biologically contextualize the data. A classic tool for this is the *pathview* (246) package. The *Graphite* (247) package is a more flexible alternative, as it allows the visualization of pathways from different data sources, and provides proper graph objects that can be manipulated using network visualization tools. Recently, the *metaGraphite* package provided a major update to the original tool, effectively incorporating multi-omics through the addition of a metabolomics layer.

Network approaches can be used for classification and prognosis. For instance, the *micrographite* (248) package provides

a method to integrate micro-RNA and mRNA data through their association to canonical pathways. This approach has been useful in identifying key micro-RNAs in myeloma (249), primary myelofibrosis (250), and ovarian cancer (251). *Mergeomics* (252) integrates data from genomic, epigenetic, and transcriptional association studies through a functional enrichment method, the results of which are used as the basis for a network construction; however, this tool has not been used in a cancer context. *pwOmics* (253) is another tool that leverages biological network knowledge to integrate multi-omics data. In particular, this tool is well-suited for the study of time series analyses.

While mapping data to predefined networks can be useful to gain a much-needed biological context, high-throughput technologies offer the opportunity to actually *infer* networks from the data itself. With such approach, data analysis problems can be transformed into network analysis problems. For instance, feature clustering becomes network module detection, which can be then used as the basis for a functional enrichment analysis (254).

While network reconstruction from omics data can be a powerful tool, it should be stated that every network reconstructed from data has an underlying hypothesis, which defines what the links between elements represent. This hypothesis should be at the center of any interpretation of the topological or functional associations recovered from a network. Furthermore, one must remember that comparison between reconstructed networks of different biological conditions will yield information about biological differences only if the method for network reconstruction does not deviate for each condition. For a discussion on this subject, see (255). This point is particularly relevant when discussing multi-omics data integration, as many of the network reconstruction methods available were developed for gene expression data. Proper validation of a method should be conducted before using it with other types of data.

There are some recent implementations of network reconstruction methods that have been developed with multi-omics data in mind. *MAGIA*² (256) is a tool for the reconstruction of micro-RNA and transcription factor regulatory circuits; it has been used for the analysis of expression regulation in the NCI60 cell panel. The *Discordant* method (257) uses a mixture model to identify differential correlation: that is, statistical dependencies between feature pairs that are lost or gained from one biological state or another. This method has been evaluated for its use with different types of omics data. The *Netboost* (258) is a network reconstruction method infers statistical dependency based on multi-omics data, and uses a modularity approach to reduce dimensionality; the method has been used for the classification and survival analysis of acute myeloid leukemia data. *AMARETTO* (259) identifies pairwise relationships between different omic layers to select cancer driver genes. A module detection approach is used to construct a dimensionally reduced module network, which is further analyzed to identify molecular signatures.

Probabilistic network reconstruction is a powerful data analysis technique. In such a model, features are connected based on an information-theoretical similarity measure, such as mutual information, between their expression profiles.

Unlike correlation metrics (260), mutual information can capture non-linear relationships between features, which makes it suitable for the analysis of transcriptomics (261). We have applied these methods for the reconstruction of micro-RNA and gene co-expression bipartite networks with minor adjustments; the analysis of such networks has yielded interesting insights on the nature of functional control by micro-RNAs (262). A current research interest the authors of this work is the extension of probabilistic network reconstruction for multi-omics reconstruction, in order to construct *probabilistic multilayer networks* (263) that can be studied using the recent tensorial formalism of multilayer networks (264).

4.6. Data Science in Biology—A Word of Warning

An important aspect of any data science project is the crucial role of both technical and domain specific expertise. The analysis of biological networks in particular can pose some complication for biological scientists not familiar with the field of network science; a network visualization may be presented as result, without an adequate evaluation of network topology or other structural and dynamic parameters. Similar behaviors can be found with other applications of data science tools.

A data-driven analysis without the participation of a domain expert risks the pursuit of non-relevant questions. On the other hand, even though a bioinformatics tool may be developed with an increased usability in mind, the level of complexity of both the computational method may require a deeper understanding of the algorithm's assumptions and limitations in order to reach valid results. With this in mind, it is evident that proper computational approaches to biological questions require a fundamental understanding of both in order to reach scientifically solid conclusions. In many cases, the key to achieve this is to strive for multidisciplinary approaches.

5. CONCLUSION

Cancer is the paradigmatic complex phenotype. We have been able to capture some of this complexity via experimental measurements with the different high throughput biomolecular technologies generically termed *omics*. Each single-technology derived data type has its own set of caveats and complexities. An additional challenge lies in the fact that each data type is able to account for a fraction of the large set of cancer aspects or features. Recent times have witnessed the development of new ways to gather and analyze these partial information layers together, under the name of multi-omics.

There are, however, multiple approaches to multi-omic computational modeling and integration, some of the most relevant have been described and discussed here. Our aim has been that of presenting the current state of the art of computational oncology tools for multiomic studies of complex cancer phenotypes. Novel developments in the multiomic computational analysis come from different fields, ranging from purely mathematical developments (263, 264), to machine learning and computational intelligence applications (179, 223), to single-cell sequencing and imaging studies (139, 145) and

more. However, in our view, the development of methods to integrate all these different analytical approaches into intelligible and statistically robust frameworks will provide the field with unprecedented advances both in our understanding of cancer biology and in our impact in the clinical settings. The field is fast-growing and currently under development, with novel algorithmic approaches being constantly released, but we believe that the present account is a good starting point.

AUTHOR CONTRIBUTIONS

GA-J and EH-L contributed to reviewing and classifying the literature, structured the review, prepared the figures, wrote, and revised the manuscript. EH-L contributed to funding and general oversight of the project.

FUNDING

This work was supported by the Consejo Nacional de Ciencia y Tecnología [SEP-CONACYT-2016-285544 and FRONTERAS-2017-2115], and the National Institute of Genomic Medicine,

México. Additional support has been granted by the Laboratorio Nacional de Ciencias de la Complejidad, from the Universidad Nacional Autónoma de México. EH-L was recipient of the 2016 Marcos Moshinsky Fellowship in the Physical Sciences.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Laura Lucila Gómez Romero (INMEGEN) for a recent discussion on current sequence-based methods. **Figures 2, 4** were generated using Biorender (<https://biorender.com/>). **Figure 4** includes images from Wikipedia, released under a Creative Commons Attribution-Share Alike 3.0.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fonc.2020.00423/full#supplementary-material>

REFERENCES

- Knox SS. From “omics” to complex disease: a systems biology approach to gene-environment interactions in cancer. *Cancer Cell Int.* (2010) 10:11. doi: 10.1186/1475-2867-10-11
- Sayama H. *Introduction to the Modeling and Analysis of Complex Systems*. Geneseo, NY: Open SUNY Textbooks (2015). Available online at: <http://textbooks.opensuny.org/introduction-to-the-modeling-and-analysis-of-complex-systems/>
- Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell.* (2000) 100:57–70. doi: 10.1016/S0092-8674(00)81683-9
- Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell.* (2011) 144:646–74. doi: 10.1016/j.cell.2011.02.013
- McAllister K, Mechanic LE, Amos C, Aschard H, Blair IA, Chatterjee N, et al. Current challenges and new opportunities for gene-environment interaction studies of complex diseases. *Am J Epidemiol.* (2017) 186:753–61. doi: 10.1093/aje/kwx227
- Barriga V, Kuol N, Nurgali K, Apostolopoulos V. The complex interaction between the tumor micro-environment and immune checkpoints in breast cancer. *Cancers.* (2019) 11:1205. doi: 10.3390/cancers11081205
- Gonzalez H, Hagerling C, Werb Z. Roles of the immune system in cancer: from tumor initiation to metastatic progression. *Genes Dev.* (2018) 32:1267–84. doi: 10.1101/gad.314617.118
- Zhou JX, Taramelli R, Pedrini E, Knijnenburg T, Huang S. Extracting intercellular signaling network of cancer tissues using ligand-receptor expression patterns from whole-tumor and single-cell transcriptomes. *Sci Rep.* (2017) 7:8815. doi: 10.1038/s41598-017-09307-w
- Brabletz T, Jung A, Reu S, Porzner M, Hlubek F, Kunz-Schughart LA, et al. Variable β -catenin expression in colorectal cancers indicates tumor progression driven by the tumor environment. *Proc Natl Acad Sci USA.* (2001) 98:10356–61. doi: 10.1073/pnas.171610498
- Kammula US, Kuntz EJ, Francone TD, Zeng Z, Shia J, Landmann RG, et al. Molecular co-expression of the c-Met oncogene and hepatocyte growth factor in primary colon cancer predicts tumor stage and clinical outcome. *Cancer Lett.* (2007) 248:219–28. doi: 10.1016/j.canlet.2006.07.007
- Van Gool B, Dedieu S, Emonard H, Roebroek AJ. The matricellular receptor LRP1 forms an interface for signaling and endocytosis in modulation of the extracellular tumor environment. *Front Pharmacol.* (2015) 6:271. doi: 10.3389/fphar.2015.00271
- Terra M, Oberkamp M, Fayolle C, Rosenbaum P, Guillerey C, Dadaglio G, et al. Tumor-derived TGF β alters the ability of plasmacytoid dendritic cells to respond to innate immune signaling. *Cancer Res.* (2018) 78:3014–26. doi: 10.1158/0008-5472.CAN-17-2719
- Mayers JR, Vander Heiden MG. Nature and nurture: what determines tumor metabolic phenotypes? *Cancer Res.* (2017) 77:3131–4. doi: 10.1158/0008-5472.CAN-17-0165
- Davidson SM, Papagiannakopoulos T, Olenchock BA, Heyman JE, Keibler MA, Luengo A, et al. Environment impacts the metabolic dependencies of Ras-driven non-small cell lung cancer. *Cell Metab.* (2016) 23:517–28. doi: 10.1016/j.cmet.2016.01.007
- Serrels A, Lund T, Serrels B, Byron A, McPherson RC, von Kriegsheim A, et al. Nuclear FAK controls chemokine transcription, Tregs, and evasion of anti-tumor immunity. *Cell.* (2015) 163:160–73. doi: 10.1016/j.cell.2015.09.001
- Hernández-Lemus E, Reyes-Gopar H, Espinal-Enríquez J, Ochoa S. The many faces of gene regulation in cancer: a computational oncogenomics outlook. *Genes.* (2019) 10:865. doi: 10.3390/genes10110865
- Hernández-Lemus E. Systems biology and integrative omics in breast cancer. In: Barh D, editor. *Omics Approaches in Breast Cancer*. New Delhi: Springer (2014). p. 333–52. doi: 10.1007/978-81-322-0843-3_17
- Hernández-Lemus E. Further steps toward functional systems biology of cancer. *Front Physiol.* (2013) 4:256. doi: 10.3389/fphys.2013.00256
- Du W, Elemento O. Cancer systems biology: embracing complexity to develop better anticancer therapeutic strategies. *Oncogene.* (2015) 34:3215–25. doi: 10.1038/onc.2014.291
- Davis-Turak J, Courtney SM, Hazard ES, Glen WB, da Silveira WA, Wesselman T, et al. Genomics pipelines and data integration: challenges and opportunities in the research setting. *Expert Rev Mol Diagn.* (2017) 17:225–37. doi: 10.1080/14737159.2017.1282822
- Goecks J, Nekrutenko A, Taylor J, Galaxy Team T. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* (2010) 11:R86. doi: 10.1186/gb-2010-11-8-r86
- Koster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics.* (2012) 28:2520–2. doi: 10.1093/bioinformatics/bts480
- Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. *Nat Biotechnol.* (2017) 35:316–9. doi: 10.1038/nbt.3820

24. Amstutz P, Crusoe MR, Tijanić N, Chapman B, Chilton J, Heuer M, et al. *Common Workflow Language, v1.0*. (2016). Available online at: <https://www.commonwl.org/>
25. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with bioconductor. *Nat Methods*. (2015) 12:115–21. doi: 10.1038/nmeth.3252
26. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. (2009) 25:1422–3. doi: 10.1093/bioinformatics/btp163
27. Grüning B, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, et al. Bioconda: A sustainable and comprehensive software distribution for the life sciences. *Nat Methods*. (2018) 15:475–6. doi: 10.1038/s41592-018-0046-7
28. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. *Nature*. (2009) 458:719–24. doi: 10.1038/nature07943
29. Maintainer BP. *arrays: Using Bioconductor for Microarray Analysis*. Washington, DC (2019).
30. Jiang Y, Oldridge DA, Diskin SJ, Zhang NR. CODEX: a normalization and copy number variation detection method for whole exome sequencing. *Nucleic Acids Res*. (2015) 43:e39. doi: 10.1093/nar/gku1363
31. Roy S, Coldren C, Karunamurthy A, Kip NS, Klee EW, Lincoln SE, et al. Standards and guidelines for validating next-generation sequencing bioinformatics pipelines. *J Mol Diagn*. (2018) 20:4–27. doi: 10.1016/j.jmoldx.2017.11.003
32. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. (2009) 25:1754–60. doi: 10.1093/bioinformatics/btp324
33. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. (2009) 10:R25. doi: 10.1186/gb-2009-10-3-r25
34. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. (2012) 9:357–9. doi: 10.1038/nmeth.1923
35. Zaharia M, Bolosky WJ, Curtis K, Fox A, Patterson D, Shenker S, et al. *Faster and More Accurate Sequence Alignment with SNAP*. (2011). Available online at: <http://arxiv.org/abs/1111.5572v1>; <http://arxiv.org/pdf/1111.5572v1>
36. Magis AT, Funk CC, Price ND. SNAPR: a bioinformatics pipeline for efficient and accurate RNA-seq alignment and analysis. *IEEE Life Sci Lett*. (2015) 1:22–5. doi: 10.1109/LLS.2015.2465870
37. Arora S, Morgan M. *Sequencing: Introduction to Bioconductor for Sequence Data*. Washington, DC (2019).
38. Luh F, Yen Y. FDA guidance for next generation sequencing-based testing: balancing regulation and innovation in precision medicine. *NPJ Genom Med*. (2018) 3:1–3. doi: 10.1038/s41525-018-0067-2
39. Zhang Q, Keleş S. CNV-guided multi-read allocation for ChIP-seq. *Bioinformatics*. (2014) 30:2860–7. doi: 10.1093/bioinformatics/btu402
40. Shen L, Shao NY, Liu X, Maze I, Feng J, Nestler EJ. diffReps: detecting differential chromatin modification sites from ChIP-seq data with biological replicates. *PLoS ONE*. (2013) 8:e65598. doi: 10.1371/journal.pone.0065598
41. Grosselin K, Durand A, Marsolier J, Poitou A, Marangoni E, Nemati F, et al. High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. *Nat Genet*. (2019) 51:1060–6. doi: 10.1038/s41588-019-0424-9
42. Bernstein BE, Meissner A, Lander ES. The mammalian epigenome. *Cell*. (2007) 128:669–81. doi: 10.1016/j.cell.2007.01.033
43. Fan S, Chi W. Methods for genome-wide DNA methylation analysis in human cancer. *Brief Funct Genomics*. (2016) 15:432–42. doi: 10.1093/bfpg/elw010
44. Maksimovic J. *methylationArrayAnalysis: A Cross-Package Bioconductor Workflow for Analysing Methylation Array*. Washington, DC (2019).
45. Hansen KD, Langmead B, Irizarry RA. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol*. (2012) 13:R83. doi: 10.1186/gb-2012-13-10-r83
46. Du P, Bourgon R. *methyAnalysis: DNA Methylation Data Analysis and Visualization*. Washington, DC (2019).
47. Bell O, Tiwari VK, Thomä NH, Schübeler D. Determinants and dynamics of genome accessibility. *Nat Rev Genet*. (2011) 12:554. doi: 10.1038/nrg3017
48. Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet*. (2019) 20:207–20. doi: 10.1038/s41576-018-0089-8
49. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. *Nature*. (2012) 489:75. doi: 10.1038/nature11232
50. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr Protoc Mol Biol*. (2015) 109:21–9. doi: 10.1002/0471142727.mb2129s109
51. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of *in vivo* protein-DNA interactions. *Science*. (2007) 316:1497–502. doi: 10.1126/science.1141319
52. Pepke S, Wold B, Mortazavi A. Computation for ChIP-seq and RNA-seq studies. *Nat Methods*. (2009) 6:S22–32. doi: 10.1038/nmeth.1371
53. Sarkar D, Gentleman R, Lawrence M, Yao Z. *chipseq: A Package for Analyzing ChIPseq Data*. Washington, DC (2019).
54. Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics*. (2009) 25:1952–8. doi: 10.1093/bioinformatics/btp340
55. Feng X, Grossman R, Stein L. PeakRanger: a cloud-enabled peak caller for ChIP-seq data. *BMC Bioinformatics*. (2011) 12:139. doi: 10.1186/1471-2105-12-139
56. Guo Y, Mahony S, Gifford DK. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput Biol*. (2012) 8:e1002638. doi: 10.1371/journal.pcbi.1002638
57. Harmanci A, Rozowsky J, Gerstein M. MUSIC: identification of enriched regions in ChIP-Seq experiments using a mappability-corrected multiscale signal processing framework. *Genome Biol*. (2014) 15:474. doi: 10.1186/s13059-014-0474-3
58. Zhang Y, Lin YH, Johnson TD, Rozek LS, Sartor MA. PePr: a peak-calling prioritization pipeline to identify consistent or differential peaks from replicated ChIP-Seq data. *Bioinformatics*. (2014) 30:2568–75. doi: 10.1093/bioinformatics/btu372
59. Kumar V, Muratani M, Rayan NA, Kraus P, Lufkin T, Ng HH, et al. Uniform, optimal signal processing of mapped deep-sequencing data. *Nat Biotechnol*. (2013) 31:615–22. doi: 10.1038/nbt.2596
60. Zhang Y, Liu T, Meyer CA, Eeckhoutte J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*. (2008) 9:R137. doi: 10.1186/gb-2008-9-9-r137
61. Feng J, Liu T, Qin B, Zhang Y, Liu XS. Identifying ChIP-seq enrichment using MACS. *Nat Protoc*. (2012) 7:1728–40. doi: 10.1038/nprot.2012.101
62. Ross-Innes CS, Stark R, Teschendorff AE, Holmes KA, Ali HR, Dunning MJ, et al. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*. (2012) 481:389–93. doi: 10.1038/nature10730
63. Lun ATL, Smyth GK. *De novo* detection of differentially bound regions for ChIP-seq data using peaks and windows: controlling error rates correctly. *Nucleic Acids Res*. (2014) 42:e95. doi: 10.1093/nar/gku351
64. Steinhauser S, Kurzawa N, Eils R, Herrmann C. A comprehensive comparison of tools for differential ChIP-seq analysis. *Brief Bioinform*. (2016) 17:953–66. doi: 10.1093/bib/bbv110
65. Jia R, Chai P, Zhang H, Fan X. Novel insights into chromosomal conformations in cancer. *Mol Cancer*. (2017) 16:173. doi: 10.1186/s12943-017-0741-5
66. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. *Science*. (2002) 295:1306–11. doi: 10.1126/science.1067799
67. Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, De Wit E, et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat Genet*. (2006) 38:1348. doi: 10.1038/ng1896
68. Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, et al. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res*. (2006) 16:1299–309. doi: 10.1101/gr.5571506
69. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions

- reveals folding principles of the human genome. *Science*. (2009) 326:289–93. doi: 10.1126/science.1181369
70. Van Berkum NL, Lieberman-Aiden E, Williams L, Imaev M, Gnirke A, Mirny LA, et al. Hi-C: a method to study the three-dimensional architecture of genomes. *J Vis Exp*. (2010) 39:e1869. doi: 10.3791/1869
 71. Thongjuea S, Stadhouders R, Grosveld FG, Soler E, Lenhard B. r3Cseq: an R/bioconductor package for the discovery of long-range genomic interactions from chromosome conformation capture and next-generation sequencing data. *Nucleic Acids Res*. (2013) 41:e132. doi: 10.1093/nar/gkt373
 72. Klein FA, Pakozdi T, Anders S, Ghavi-Helm Y, Furlong EEM, Huber W. FourCSeq: analysis of 4C sequencing data. *Bioinformatics*. (2015) 31:3085–91. doi: 10.1093/bioinformatics/btv335
 73. Servant N, Lajoie BR, Nora EP, Giorgetti L, Chen CJ, Heard E, et al. HiTC: exploration of high-throughput 'C' experiments. *Bioinformatics*. (2012) 28:2843–4. doi: 10.1093/bioinformatics/bts521
 74. Fudenberg G, Getz G, Meyerson M, Mirny LA. High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. *Nat Biotechnol*. (2011) 29:1109. doi: 10.1038/nbt.2049
 75. Baxter JS, Leavy OC, Dryden NH, Maguire S, Johnson N, Fedele V, et al. Capture Hi-C identifies putative target genes at 33 breast cancer risk loci. *Nat Commun*. (2018) 9:1–13. doi: 10.1038/s41467-018-03411-9
 76. Stone JK, Kim JH, Vukadin L, Richard A, Giannini HK, Lim STS, et al. Hypoxia induces cancer cell-specific chromatin interactions and increases MALAT1 expression in breast cancer cells. *J Biol Chem*. (2019) 294:11213–24. doi: 10.1074/jbc.RA118.006889
 77. Kishore K, de Pretis S, Lister R, Morelli MJ, Bianchi V, Amati B, et al. methylPipe and compEpiTools: a suite of R packages for the integrative analysis of epigenomics data. *BMC Bioinformatics*. (2015) 16:313. doi: 10.1186/s12859-015-0742-6
 78. Ou J, Liu H, Yu J, Kelliher MA, Castilla LH, Lawson ND, et al. ATACseqQC: a Bioconductor package for post-alignment quality assessment of ATAC-seq data. *BMC Genomics*. (2018) 19:3. doi: 10.1186/s12864-018-4559-3
 79. Wei Z, Zhang W, Fang H, Li Y, Wang X. esATAC: an easy-to-use systematic pipeline for ATAC-seq data analysis. *Bioinformatics*. (2018) 34:2664–5. doi: 10.1093/bioinformatics/bty141
 80. Harmston N, Ing-Simmons E, Perry M, Barešić A, Lenhard B. Genomic Interactions: an R/bioconductor package for manipulating and investigating chromatin interaction data. *BMC Genomics*. (2015) 16:963. doi: 10.1186/s12864-015-2140-x
 81. Zhang H, He L, Cai L. Transcriptome sequencing: RNA-seq. In: Huang T, editor. *Computational Systems Biology*. New York, NY: Humana Press (2018). p. 15–27.
 82. Jeong E, Moon SU, Song M, Yoon S. Transcriptome modeling and phenotypic assays for cancer precision medicine. *Arch Pharm Res*. (2017) 40:906–14. doi: 10.1007/s12272-017-0940-z
 83. Babu MM. Introduction to microarray data analysis. *Comput Genom Theory Appl*. (2004) 225:249. doi: 10.1007/0-306-47815-3_1
 84. García-Compos MA, Espinal-Enríquez J, Hernández-Lemus E. Pathway analysis: state of the art. *Front Physiol*. (2015) 6:383. doi: 10.3389/fphys.2015.00383
 85. Li JR, Sun CH, Li W, Chao RF, Huang CC, Zhou XJ, et al. Cancer RNA-Seq Nexus: a database of phenotype-specific transcriptome profiling in cancer cells. *Nucleic Acids Res*. (2015) 44:D944–51. doi: 10.1093/nar/gkv1282
 86. Tomczak K, Czerwińska P, Wizniewski M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol*. (2015) 19:A68. doi: 10.5114/wo.2014.47136
 87. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The genotype-tissue expression (GTEx) project. *Nat Genet*. (2013) 45:580. doi: 10.1038/ng.2653
 88. Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet*. (2012) 13:227–32. doi: 10.1038/nrg3185
 89. Jiang Z, Zhou X, Li R, Michal JJ, Zhang S, Dodson MV, et al. Whole transcriptome analysis with sequencing: methods, challenges and potential solutions. *Cell Mol Life Sci*. (2015) 72:3425–39. doi: 10.1007/s00018-015-1934-y
 90. Kaikkonen MU, Adelman K. Emerging roles of non-coding RNA transcription. *Trends Biochem Sci*. (2018) 43:654–67. doi: 10.1016/j.tibs.2018.06.002
 91. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, et al. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*. (2008) 321:956–60. doi: 10.1126/science.1160342
 92. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*. (2008) 5:621. doi: 10.1038/nmeth.1226
 93. Slonim DK, Yanai I. Getting started in gene expression microarray analysis. *PLoS Comput Biol*. (2009) 5:e1000543. doi: 10.1371/journal.pcbi.1000543
 94. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. (2013) 29:15–21. doi: 10.1093/bioinformatics/bts635
 95. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*. (2016) 34:525–7. doi: 10.1038/nbt.3519
 96. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods*. (2017) 14:417–9. doi: 10.1038/nmeth.4197
 97. Li Y, Heavican TB, Vellichirammal NN, Iqbal J, Guda C. ChimeRScope: a novel alignment-free algorithm for fusion transcript prediction using paired-end RNA-Seq data. *Nucleic Acids Res*. (2017) 45:e120. doi: 10.1093/nar/gkx315
 98. Yang X, Saito Y, Rao A, Kim HJ, Singh P, Scott E, et al. Alignment-free filtering for cfRNA fusion fragments. *Bioinformatics*. (2019) 35:i225–32. doi: 10.1093/bioinformatics/btz346
 99. Raplee ID, Evsikov AV, Marin de Esvikova C. Aligning the aligners: comparison of RNA sequencing data alignment and gene expression quantification tools for clinical breast cancer research. *J Personal Med*. (2019) 9:18. doi: 10.3390/jpm9020018
 100. Babarinde IA, Li Y, Hutchins AP. Computational methods for mapping, assembly and quantification for coding and non-coding transcripts. *Comput Struct Biotechnol J*. (2019) 17:628–37. doi: 10.1016/j.csbj.2019.04.012
 101. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*. (2009) 25:1105–11. doi: 10.1093/bioinformatics/btp120
 102. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*. (2010) 28:511–5. doi: 10.1038/nbt.1621
 103. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*. (2012) 7:562–78. doi: 10.1038/nprot.2012.016
 104. Baruzzo G, Hayer KE, Kim EJ, Camillo BD, FitzGerald GA, Grant GR. Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nat Methods*. (2017) 14:135–9. doi: 10.1038/nmeth.4106
 105. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol*. (2019) 37:907–15. doi: 10.1038/s41587-019-0201-4
 106. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol*. (2019) 20:278. doi: 10.1186/s13059-019-1910-1
 107. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc*. (2016) 11:1650–67. doi: 10.1038/nprot.2016.095
 108. Li Y, Chien J, Smith DI, Ma J. FusionHunter: identifying fusion transcripts in cancer using paired-end RNA-seq. *Bioinformatics*. (2011) 27:1708–10. doi: 10.1093/bioinformatics/btr265
 109. Ellrott K, Buchanan A, Creason A, Mason M, Schaffter T, Hoff B, et al. Reproducible biomedical benchmarking in the cloud: lessons from crowd-sourced data challenges. *Genome Biol*. (2019) 20:1–9. doi: 10.1186/s13059-019-1794-0
 110. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. (2014) 15:550. doi: 10.1186/s13059-014-0550-8

111. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* (2012) 40:4288–97. doi: 10.1093/nar/gks042
112. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-seq and microarray studies. *Nucleic Acids Res.* (2015) 43:e47. doi: 10.1093/nar/gkv007
113. Pimentel H, Bray NL, Puente S, Melsted P, Pachter L. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nat Methods.* (2017) 14:687–90. doi: 10.1038/nmeth.4324
114. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. *Genome Biol.* (2016) 17:13. doi: 10.1186/s13059-016-0881-8
115. Aslam B, Basit M, Nisar MA, Khurshid M, Rasool MH. Proteomics: technologies and their applications. *J Chromatogr Sci.* (2017) 55:182–96. doi: 10.1093/chromsci/bmw167
116. Yakkoui Y, Temel Y, Chevet E, Negroni L. Integrated and quantitative proteomics of human tumors. *Methods Enzymol.* (2017) 586:229–46. doi: 10.1016/b.s.mie.2016.09.034
117. Sutandy FXR, Qian J, Chen CS, Zhu H. Overview of protein microarrays. *Curr Protoc Protein Sci.* (2013) Chapter 27:Unit 27.1. doi: 10.1002/0471140864.ps2701s72
118. Atak A, Mukherjee S, Jain R, Gupta S, Singh VA, Gahoi N, et al. Protein microarray applications: autoantibody detection and posttranslational modification. *Proteomics.* (2016) 16:2557–69. doi: 10.1002/pmic.201600104
119. Cho WC. Mass spectrometry-based proteomics in cancer research. *Expert Rev Proteomics.* (2017) 14:725–7. doi: 10.1080/14789450.2017.1365604
120. Schmidt A, Forne I, Imhof A. Bioinformatic analysis of proteomics data. *BMC Syst Biol.* (2014) 8:S3. doi: 10.1186/1752-0509-8-S2-S3
121. Graumann J, Scheltema RA, Zhang Y, Cox J, Mann M. A framework for intelligent data acquisition and real-time database searching for shotgun proteomics. *Mol Cell Proteomics.* (2012) 11:M111.013185. doi: 10.1074/mcp.M111.013185
122. Hoopmann MR, Moritz RL. Current algorithmic solutions for peptide-based proteomics data generation and identification. *Curr Opin Biotechnol.* (2013) 24:31–8. doi: 10.1016/j.copbio.2012.10.013
123. Nesvizhskii AI. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J Proteomics.* (2010) 73:2092–123. doi: 10.1016/j.jprot.2010.08.009
124. Kopczynski D, Sickmann A, Ahrends R. Computational proteomics tools for identification and quality control. *J Biotechnol.* (2017) 261:126–30. doi: 10.1016/j.jbiotec.2017.06.1199
125. Mihășan M, Wormwood KL, Sokolowska I, Roy U, Woods AG, Darie CC. Mass spectrometry-and computational structural biology-based investigation of proteins and peptides. In: *Advancements of Mass Spectrometry in Biomedical Research*. Cham: Springer (2019). p. 265–287.
126. Gatto L, Christoforou A. Using R and bioconductor for proteomics data analysis. *Biochim Biophys Acta.* (2014) 1844:42–51. doi: 10.1016/j.bbapap.2013.04.032
127. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol.* (2008) 26:1367–72. doi: 10.1038/nbt.1511
128. Vazquez A, Kamphorst JJ, Markert EK, Schug ZT, Tardito S, Gottlieb E. Cancer metabolism at a glance. *J Cell Sci.* (2016) 129:3367–73. doi: 10.1242/jcs.181016
129. Armitage EG, Ciborowski M. Applications of metabolomics in cancer studies. *Adv Exp Med Biol.* (2017) 965:209–34. doi: 10.1007/978-3-319-47656-8_9
130. Yang K, Han X. Lipidomics: techniques, applications, and outcomes related to biomedical sciences. *Trends Biochem Sci.* (2016) 41:954–69. doi: 10.1016/j.tibs.2016.08.010
131. Perrotti F, Rosa C, Cicalini I, Sacchetta P, Del Boccio P, Genovesi D, et al. Advances in lipidomics for cancer biomarkers discovery. *Int J Mol Sci.* (2016) 17:1992. doi: 10.3390/ijms17121992
132. Zhang A, Sun H, Yan G, Wang P, Wang X. Mass spectrometry-based metabolomics: applications to biomarker and metabolic pathway research. *Biomed Chromatogr.* (2016) 30:7–12. doi: 10.1002/bmc.3453
133. Hu T, Zhang JL. Mass-spectrometry-based lipidomics. *J Sep Sci.* (2018) 41:351–72. doi: 10.1002/jssc.201700709
134. Meier R, Ruttkies C, Treutler H, Neumann S. Bioinformatics can boost metabolomics research. *J Biotechnol.* (2017) 261:137–41. doi: 10.1016/j.jbiotec.2017.05.018
135. Aggio R, Villas-Boas SG, Ruggiero K. Metab: an R package for high-throughput analysis of metabolomics data generated by GC-MS. *Bioinformatics.* (2011) 27:2316–8. doi: 10.1093/bioinformatics/btr379
136. Stanstrup J, Broeckling CD, Helmus R, Hoffmann N, Mathé E, Naake T, et al. The metaRbolomics toolbox in bioconductor and beyond. *Metabolites.* (2019) 9:E200. doi: 10.3390/metabo9100200
137. Mohamed A, Molendijk J. *lipidr: Data Mining and Analysis of Lipidomics Datasets*. R package version 200. Washington, DC (2019).
138. Yuan Y. Spatial heterogeneity in the tumor microenvironment. *Cold Spring Harb Perspect Med.* (2016) 6:a026583. doi: 10.1101/cshperspect.a026583
139. Prasetyanti PR, Medema JP. Intra-tumor heterogeneity from a cancer stem cell perspective. *Mol Cancer.* (2017) 16:41. doi: 10.1186/s12943-017-0600-4
140. Sierant MC, Choi J. Single-cell sequencing in cancer: recent applications to immunogenomics and multi-omics tools. *Genomics Inform.* (2018) 16:e17. doi: 10.5808/GI.2018.16.4.e17
141. Ren X, Kang B, Zhang Z. Understanding tumor ecosystems by single-cell sequencing: promises and limitations. *Genome Biol.* (2018) 19:1–14. doi: 10.1186/s13059-018-1593-z
142. Navin NE. The first five years of single-cell cancer genomics and beyond. *Genome Res.* (2015) 25:1499–507. doi: 10.1101/gr.191098.115
143. Lo PK, Zhou Q. Emerging techniques in single-cell epigenomics and their applications to cancer research. *J Clin Genom.* (2018) 1:1–16. doi: 10.4172/JCG.1000103
144. Litzénburger UM, Buenrostro JD, Wu B, Shen Y, Sheffield NC, Kathiria A, et al. Single-cell epigenomic variability reveals functional cancer heterogeneity. *Genome Biol.* (2017) 18:15. doi: 10.1186/s13059-016-1133-7
145. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet.* (2015) 16:133–45. doi: 10.1038/nrg3833
146. Haque A, Engel J, Teichmann SA, Lönnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med.* (2017) 9:75. doi: 10.1186/s13073-017-0467-4
147. Gao S. Data analysis in single-cell transcriptome sequencing. *Methods Mol Biol.* (2018) 1754:311–26. doi: 10.1007/978-1-4939-7717-8_18
148. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med.* (2018) 50:96. doi: 10.1038/s12276-018-0071-8
149. Lun ATL, McCarthy DJ, Marioni JC. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res.* (2016) 5:2122. doi: 10.12688/f1000research.9501.2
150. Amezcua RA, Carey VJ, Carpp LN, Geistlinger L, Lun ATL, Marini F, et al. *Orchestrating Single-Cell Analysis With Bioconductor*. Washington, DC (2019).
151. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* (2018) 19:15. doi: 10.1186/s13059-017-1382-0
152. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol.* (2018) 36:411–20. doi: 10.1038/nbt.4096
153. Stuart T, Satija R. Integrative single-cell analysis. *Nat Rev Genet.* (2019) 20:257–72. doi: 10.1038/s41576-019-0093-7
154. Pegoraro G, Misteli T. High-throughput imaging for the discovery of cellular mechanisms of disease. *Trends Genet.* (2017) 33:604–15. doi: 10.1016/j.tig.2017.06.005
155. Pau G, Fuchs F, Sklyar O, Boutros M, Huber W. EBIImage-an R package for image processing with applications to cellular phenotypes. *Bioinformatics.* (2010) 26:979–81. doi: 10.1093/bioinformatics/btq046
156. Pau G, Zhang X, Boutros M, Huber W. *imageHTS: Analysis of High-Throughput Microscopy-Based Screens*. Washington, DC (2019).
157. McQuinn C, Goodman A, Chernyshev V, Kamentsky L, Cimini BA, Karhohs KW, et al. CellProfiler 3.0: next-generation image processing for biology. *PLoS Biol.* (2018) 16:e2005970. doi: 10.1371/journal.pbio.2005970
158. Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science.* (2015) 348:aaa6090. doi: 10.1126/science.aaa6090

159. Rodriques SG, Stickels RR, Goeva A, Martin CA, Murray E, Vanderburg CR, et al. Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. *Science*. (2019) 363:1463–7. doi: 10.1126/science.aaw1219
160. Yoosuf N, Navarro JF, Salmén F, Ståhl PL, Daub CO. Identification and transfer of spatial transcriptomics signatures for cancer diagnosis. *Breast Cancer Res.* (2020) 22:1–10. doi: 10.1186/s13058-019-1242-9
161. Berglund E, Maaskola J, Schultz N, Friedrich S, Marklund M, Bergenstråhle J, et al. Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nat Commun.* (2018) 9:1–13. doi: 10.1038/s41467-018-04724-5
162. Thrane K, Eriksson H, Maaskola J, Hansson J, Lundberg J. Spatially resolved transcriptomics enables dissection of genetic heterogeneity in stage III cutaneous malignant melanoma. *Cancer Res.* (2018) 78:5970–9. doi: 10.1158/0008-5472.CAN-18-0747
163. Moncada R, Wagner F, Chiodin M, Devlin JC, Baron M, Hajdu CH, et al. Building a tumor atlas: integrating single-cell RNA-Seq data with spatial transcriptomics in pancreatic ductal adenocarcinoma. *bioRxiv*. (2018) 254375. doi: 10.1101/254375
164. Xu H, Lyu X, Yi M, Zhao W, Song Y, Wu K. Organoid technology and applications in cancer research. *J Hematol Oncol.* (2018) 11:116. doi: 10.1186/s13045-018-0662-9
165. Lindeboom RG, van Voorthuysen L, Oost KC, Rodríguez-Colman MJ, Luna-Velez MV, Furlan C, et al. Integrative multi-omics analysis of intestinal organoid differentiation. *Mol Syst Biol.* (2018) 14:e8227. doi: 10.15252/msb.20188227
166. Finotello F, Eduati F. Multi-omics profiling of the tumor microenvironment: paving the way to precision immuno-oncology. *Front Oncol.* (2018) 8:430. doi: 10.3389/fonc.2018.00430
167. Finotello F, Rieder D, Hackl H, Trajanoski Z. Next-generation computational tools for interrogating cancer immunity. *Nat Rev Genet.* (2019) 20:724–46. doi: 10.1038/s41576-019-0166-7
168. Mimitou EP, Cheng A, Montalbano A, Hao S, Stoeckius M, Legut M, et al. Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. *Nat Methods.* (2019) 16:409–12. doi: 10.1038/s41592-019-0392-0
169. Praktijn SD, Obermayer B, Zhu Q, Fang L, Liu H, Quinn H, et al. Tracing tumorigenesis in a solid tumor model at single-cell resolution. *Nat Commun.* (2020) 11:1–12. doi: 10.1038/s41467-020-14777-0
170. Rajagopala SV, Vashee S, Oldfield LM, Suzuki Y, Venter JC, Telenti A, et al. The human microbiome and cancer. *Cancer Prev Res (Phila).* (2017) 10:226–34. doi: 10.1158/1940-6207.CAPR-16-0249
171. Mallick H, Ma S, Franzosa EA, Vatanen T, Morgan XC, Huttenhower C. Experimental design and quantitative analysis of microbial community multiomics. *Genome Biol.* (2017) 18:228. doi: 10.1186/s13059-017-1359-z
172. Contreras JL, Knoppers BM. The genomic commons. *Annu Rev Genomics Hum Genet.* (2018) 19:429–53. doi: 10.1146/annurev-genom-083117-021552
173. Cook-Deegan R, McGuire AL. Moving beyond Bermuda: sharing data to build a medical information commons. *Genome Res.* (2017) 27:897–901. doi: 10.1101/gr.216911.116
174. Jansen P, van den Berg L, van Overveld P, Boiten JW. Research data stewardship for healthcare professionals. In: Kubben P, Dumontier M, Dekker A, editors. *Fundamentals of Clinical Data Science*. Cham: Springer (2018) p. 37–53.
175. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, et al. Toward a shared vision for cancer genomic data. *N Engl J Med.* (2016) 375:1109–12. doi: 10.1056/NEJMp1607591
176. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, et al. The cancer genome atlas pan-cancer analysis project. *Nat Genet.* (2013) 45:1113–20. doi: 10.1038/ng.2764
177. Kosinski M, Biecek P. *RTCGA: The Cancer Genome Atlas Data Integration*. R package version 1.14.0 (2019). Available online at: <https://rtcga.github.io/RTCGA>
178. Campbell PJ, Räscher G, Kahles A, Lehmann KV, Davidson NR, Stark SG, et al. Pan-cancer analysis of whole genomes. *Nature*. (2020) 578:82–93. doi: 10.1038/s41586-020-1969-6
179. Rendleman MC, Buatti JM, Braun TA, Smith BJ, Nwakama C, Beichel RR, et al. Machine learning with the TCGA-HNSC dataset: improving usability by addressing inconsistency, sparsity, and high-dimensionality. *BMC Bioinformatics.* (2019) 20:339. doi: 10.1186/s12859-019-2929-8
180. Silva TC, Colaprico A, Olsen C, D'Angelo F, Bontempi G, Ceccarelli M, et al. TCGA workflow: analyze cancer genomics and epigenomics data using bioconductor packages. *F1000Res.* (2016) 5:1542. doi: 10.12688/f1000research.8923.1
181. Parkinson DR, Johnson BE, Sledge GW. Making personalized cancer medicine a reality: challenges and opportunities in the development of biomarkers and companion diagnostics. *Clin Cancer Res.* (2012) 18:619–24. doi: 10.1158/1078-0432.CCR-11-2017
182. Huang S, Chaudhary K, Garmire LX. More is better: recent progress in multi-omics data integration methods. *Front Genet.* (2017) 8:84. doi: 10.3389/fgene.2017.00084
183. Chakraborty S, Hosen MI, Ahmed M, Shekhar HU. Onco-multi-OMICS approach: a new frontier in cancer research. *Biomed Res Int.* (2018) 2018:9836256. doi: 10.1155/2018/9836256
184. Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, et al. Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics.* (2016) 17:15. doi: 10.1186/s12859-015-0857-9
185. Yoo BC, Kim KH, Woo SM, Myung JK. Clinical multi-omics strategies for the effective cancer management. *J Proteomics.* (2018) 188:97–106. doi: 10.1016/j.jprot.2017.08.010
186. Ramos M, Schiffer L, Re A, Azhar R, Basunia A, Rodriguez C, et al. Software for the integration of multiomics experiments in bioconductor. *Cancer Res.* (2017) 77:e39–42. doi: 10.1158/0008-5472.CAN-17-0344
187. Voillet V, Besse P, Liaubet L, Cristobal MS, González I. Handling missing rows in multi-omics data integration: multiple imputation in multiple factor analysis framework. *BMC Bioinformatics.* (2016) 17:402. doi: 10.1186/s12859-016-1273-5
188. van Itersen M, Cats D, Hop P, Heijmans BT. omicsPrint: detection of data linkage errors in multiple omics studies. *Bioinformatics.* (2018) 34:2142–3. doi: 10.1093/bioinformatics/bty062
189. Consortia S. *STATegRa: Classes and Methods for Multi-Omics Data Integration*. R package version 1.20.0 (2019).
190. Rohart F, Gautier B, Singh A, Lê Cao KA. mixOmics: an R package for 'omics feature selection and multiple data integration. *PLoS Comput Biol.* (2017) 13:e1005752. doi: 10.1371/journal.pcbi.1005752
191. Cao KAL, González I, Déjean S. integrOmics: an R package to unravel relationships between two omics datasets. *Bioinformatics.* (2009) 25:2855–6. doi: 10.1093/bioinformatics/btp515
192. Hernández-de Diego R, Boix-Chova N, Gómez-Cabrero D, Tegner J, Abugessaisa I, Conesa A. STATegra EMS: an experiment management system for complex next-generation omics experiments. *BMC Syst Biol.* (2014) 8:S9. doi: 10.1186/1752-0509-8-S2-S9
193. Martínez-Mira C, Conesa A, Tarazona S. *MOSim: Multi-Omics Simulation in R*. Washington, DC (2018).
194. Gomez-Cabrero D, Tarazona S, Ferreirós-Vidal I, Ramirez RN, Company C, Schmidt A, et al. STATegra, a comprehensive multi-omics dataset of B-cell differentiation in mouse. *Sci Data.* (2019) 6:256. doi: 10.1038/s41597-019-0202-7
195. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol.* (2009) 27:1160–7. doi: 10.1200/JCO.2008.18.1370
196. Meng C, Zeleznik OA, Thallinger GG, Kuster B, Gholami AM, Culhane AC. Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief Bioinform.* (2016) 17:628–41. doi: 10.1093/bib/bbv108
197. Tini G, Marchetti L, Priami C, Scott-Boyer MP. Multi-omics integration—a comparison of unsupervised clustering methodologies. *Brief Bioinform.* (2017) 20:1269–79. doi: 10.1093/bib/bbx167
198. Yugi K, Kubota H, Hatano A, Kuroda S. Trans-omics: how to reconstruct biochemical networks across multiple 'omic' layers. *Trends Biotechnol.* (2016) 34:276–90. doi: 10.1016/j.tibtech.2015.12.013
199. Singh A, Shannon CP, Gautier B, Rohart F, Vacher M, Tebbutt SJ, et al. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics.* (2019) 35:3055–62. doi: 10.1093/bioinformatics/bty1054

200. Angermueller C, Pärnamaa T, Parts L, Stegle O. Deep learning for computational biology. *Mol Syst Biol.* (2016) 12:878. doi: 10.15252/msb.20156651
201. Pearson K. LIII. On lines and planes of closest fit to systems of points in space. *Lond Edinb Dublin Philos Mag J Sci.* (1901) 2:559–72. doi: 10.1080/14786440109462720
202. van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res.* (2008) 9:2579–605.
203. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. (2018). Available online at: <http://arxiv.org/abs/1802.03426v2>; <http://arxiv.org/pdf/1802.03426v2>
204. Tufte E. *The Visual Display of Quantitative Information*. Graphics Press (1983).
205. Freytag S. *schex: Hexbin Plots for Single Cell Omics Data*. Washington, DC: R package version 1.0.0 (2019).
206. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. *Genome Res.* (2009) 19:1639–45. doi: 10.1101/gr.092759.109
207. Hu Y, Yan C, Hsu CH, Chen QR, Niu K, Komatsoulis GA, et al. OmicCircos: a simple-to-use R package for the circular visualization of multidimensional omics data. *Cancer Inform.* (2014) 13:13–20. doi: 10.4137/CIN.S13495
208. Hahne F, Ivanek R. Visualizing genomic data using Gviz and bioconductor. *Stat Genom.* (2016) 1418:335–51. doi: 10.1007/978-1-4939-3578-9_16
209. Huang W, Loganantharaj R, Schroeder B, Fargo D, Li L. PAVIS: a tool for peak annotation and visualization. *Bioinformatics.* (2013) 29:3097–9. doi: 10.1093/bioinformatics/btt520
210. Pont F, Tosolini M, Fournié JJ. Single-cell signature explorer for comprehensive visualization of single cell signatures across scRNA-seq datasets. *Nucleic Acids Res.* (2019) 47:e133. doi: 10.1093/nar/gkz601
211. Paulsen J, Sandve GK, Gundersen S, Lien TG, Trengereid K, Hovig E. HiBrowse: multi-purpose statistical analysis of genome-wide chromatin 3D organization. *Bioinformatics.* (2014) 30:1620–2. doi: 10.1093/bioinformatics/btu082
212. Li D, Hsu S, Purushotham D, Sears RL, Wang T. WashU epigenome browser update 2019. *Nucleic Acids Res.* (2019) 47:W158–65. doi: 10.1093/nar/gkz348
213. Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* (2016) 3:99–101. doi: 10.1016/j.cels.2015.07.012
214. Yardımcı GG, Noble WS. Software tools for visualizing Hi-C data. *Genome Biol.* (2017) 18:26. doi: 10.1186/s13059-017-1161-y
215. Das S, Tripathy DS. OMICsPCA: An R Package for Quantitative Integration and Analysis of Multiple Omics Assays From Heterogeneous Samples. R package version 1.2.0 (2019).
216. Xu T, Le TD, Liu L, Su N, Wang R, Sun B, et al. CancerSubtypes: an R/bioconductor package for molecular cancer subtype identification, validation and visualization. *Bioinformatics.* (2017) 33:3131–3. doi: 10.1093/bioinformatics/btx378
217. Mezhdoud K. *bioCancer: Interactive Multi-Omics Cancers Data Visualization and Analysis*. R package version 1.12.0 (2019). Available online at: <http://kmezhdoud.github.io/bioCancer>
218. Meng C, Kuster B, Culhane AC, Gholami AM. A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics.* (2014) 15:162. doi: 10.1186/1471-2105-15-162
219. Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics.* (2009) 25:2906–12. doi: 10.1093/bioinformatics/btp543
220. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY: Springer (2009).
221. Provenzano E, Ulaner GA, Chin SF. Molecular classification of breast cancer. *PET Clinics.* (2018) 13:325–38. doi: 10.1016/j.cpet.2018.02.004
222. Syed-Abdul S, Iqbal U, Li YC. Predictive analytics through machine learning in the clinical settings. *Comput Methods Prog Biomed.* (2017) 144:A1–2. doi: 10.1016/S0169-2607(17)30552-7
223. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J.* (2015) 13:8–17. doi: 10.1016/j.csbj.2014.11.005
224. Streeter OE, Beron PJ, Iyer PN. Precision medicine: genomic profiles to individualize therapy. *Otolaryngol Clin North Am.* (2017) 50:765–73. doi: 10.1016/j.otc.2017.03.012
225. Schwaederle M, Daniels GA, Piccioni DE, Fanta PT, Schwab RB, Shimabukuro KA, et al. On the road to precision cancer medicine: analysis of genomic biomarker actionability in 439 patients. *Mol Cancer Ther.* (2015) 14:1488–94. doi: 10.1158/1535-7163.MCT-14-1061
226. Singh A, Shannon CP, Gautier B, Rohart F, Vacher M, Tebbutt SJ, et al. DIABLO: from multi-omics assays to biomarker discovery, an integrative approach. *bioRxiv.* (2016). doi: 10.1101/067611
227. Thévenot EA, Roux A, Xu Y, Ezan E, Junot C. Analysis of the human adult urinary metabolome variations with age, body mass index, and gender by implementing a comprehensive workflow for univariate and OPLS statistical analyses. *J Proteome Res.* (2015) 14:3322–35. doi: 10.1021/acs.jproteome.5b00354
228. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, et al. MultiOmics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology.* (2018) 14. doi: 10.15252/msb.20178124
229. Determan C. Optimal algorithm for metabolomics classification and feature selection varies by dataset. *Int J Biol.* (2015) 7. doi: 10.5539/ijb.v7n1p100
230. Bhalla S, Kaur H, Dhali A, Raghava GPS. Prediction and analysis of skin cancer progression using genomics profiles of patients. *Sci Rep.* (2019) 9:15790. doi: 10.1038/s41598-019-52134-4
231. Rinaudo P, Boudah S, Junot C, Thévenot EA. Biosigner: a new method for the discovery of significant molecular signatures from omics data. *Front Mol Biosci.* (2016) 3:26. doi: 10.3389/fmolb.2016.00026
232. Long NP, Jung KH, Anh NH, Yan HH, Nghi TD, Park S, et al. An integrative data mining and omics-based translational model for the identification and validation of oncogenic biomarkers of pancreatic cancer. *Cancers.* (2019) 11:155. doi: 10.3390/cancers11020155
233. Kwon MS, Kim Y, Lee S, Namkung J, Yun T, Yi SG, et al. Integrative analysis of multi-omics data for identifying multi-markers for diagnosing pancreatic cancer. *BMC Genomics.* (2015) 16:S4. doi: 10.1186/1471-2164-16-S9-S4
234. Klein HU, Schäfer M, Porse BT, Hasemann MS, Ickstadt K, Dugas M. Integrative analysis of histone ChIP-seq and transcription data using Bayesian mixture models. *Bioinformatics.* (2014) 30:1154–62. doi: 10.1093/bioinformatics/btu003
235. Han S, Lee Y. IMAS: Integrative Analysis of Multi-Omics Data for Alternative Splicing. R package version 1.8.0 (2019).
236. Hernandez-Ferrer C, Wellenius GA, Tamayo I, Basagaña X, Sunyer J, Vrijheid M, et al. Comprehensive study of the exposome and omic data using rexpome bioconductor packages. *Bioinformatics.* (2019) 35:5344–5. doi: 10.1093/bioinformatics/btz526
237. Metwally AA, Zhang T, Snyder M. OmicsLonDA: Omics Longitudinal Differential Analysis. R package version 1.0.0 (2019). Available online at: <https://github.com/aametwally/OmicsLonDA>
238. de Anda-Jáuregui G, Guo K, McGregor BA, Feldman EL, Hur J. Pathway crosstalk perturbation network modeling for identification of connectivity changes induced by diabetic neuropathy and pioglitazone. *BMC Syst Biol.* (2019) 13:1. doi: 10.1186/s12918-018-0674-7
239. Meng C, Basunia A, Peters B, Gholami AM, Kuster B, Culhane AC. MOGSA: integrative single sample gene-set analysis of multiple omics data. *Mol Cell Proteomics.* (2019) 18:S153–68. doi: 10.1074/mcp.TIR118.001251
240. Rodriguez JC, Merino GA, Llera AS, Fernández EA. Massive integrative gene set analysis enables functional characterization of breast cancer subtypes. *J Biomed Inform.* (2019) 93:103157. doi: 10.1016/j.jbi.2019.103157
241. Odom GJ, Ban Y, Liu L, Sun X, Pico AR, Zhang B, et al. pathwayPCA: an R package for integrative pathway analysis with modern PCA methodology and gene selection. *bioRxiv.* (2019). doi: 10.1101/615435
242. Dinalankara W, Ke Q, Xu Y, Ji L, Pagane N, Lien A, et al. Digitizing omics profiles by divergence from a baseline. *Proc Natl Acad Sci USA.* (2018) 115:4545–52. doi: 10.1073/pnas.1721628115
243. Pilarczyk M, Najafabadi MF, Kouril M, Vasiliauskas J, Niu W, Shamsaei B, et al. Connecting omics signatures of diseases, drugs, and mechanisms of actions with iLINC. *bioRxiv.* (2019). doi: 10.1101/826271

244. Albert R, Barabási AL. Statistical mechanics of complex networks. *Rev Mod Phys.* (2002) 74:47–97. doi: 10.1103/RevModPhys.74.47
245. Quesada D, Cruz-Monteagudo M, Fletcher T, Duardo-Sanchez A, González-Díaz H. Complex networks and machine learning: from molecular to social sciences. *Appl Sci.* (2019) 9:4493. doi: 10.3390/app9214493
246. Luo W, Brouwer C. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics.* (2013) 29:1830–1. doi: 10.1093/bioinformatics/btt285
247. Sales G, Calura E, Cavalieri D, Romualdi C. Graphite—a bioconductor package to convert pathway topology to gene network. *BMC Bioinformatics.* (2012) 13:20. doi: 10.1186/1471-2105-13-20
248. Calura E, Martini P, Sales G, Beltrame L, Chiorino G, D’Incalci M, et al. Wiring miRNAs to pathways: a topological approach to integrate miRNA and mRNA expression profiles. *Nucleic Acids Res.* (2014) 42:e96. doi: 10.1093/nar/gku354
249. Calura E, Bisognin A, Manzoni M, Todoerti K, Taiana E, Sales G, et al. Disentangling the microRNA regulatory milieu in multiple myeloma: integrative genomics analysis outlines mixed miRNA-TF circuits and pathway-derived networks modulated in t(4;14) patients. *Oncotarget.* (2015) 7:2367–78. doi: 10.18632/oncotarget.6151
250. Calura E, Pizzini S, Bisognin A, Coppe A, Sales G, Gaffo E, et al. A data-driven network model of primary myelofibrosis: transcriptional and post-transcriptional alterations in CD34⁺ cells. *Blood Cancer J.* (2016) 6:e439. doi: 10.1038/bcj.2016.47
251. Calura E, Paracchini L, Fruscio R, DiFeo A, Ravaggi A, Peronne J, et al. A prognostic regulatory pathway in stage I epithelial ovarian cancer: new hints for the poor prognosis assessment. *Ann Oncol.* (2016) 27:1511–9. doi: 10.1093/annonc/mdw210
252. Shu L, Zhao Y, Kurt Z, Byars SG, Tukiainen T, Kettunen J, et al. Mergeomics: multidimensional data integration to identify pathogenic perturbations to biological systems. *BMC Genomics.* (2016) 17:874. doi: 10.1186/s12864-016-3198-9
253. Wachter A, Beißbarth T. pwOmics: an R package for pathway-based integration of time-series omics data using public database knowledge. *Bioinformatics.* (2015) 31:3072–4. doi: 10.1093/bioinformatics/btv323
254. Alcalá-Corona SA, de Anda-Jáuregui G, Espinal-Enriquez J, Tovar H, Hernández-Lemus E. Network modularity and hierarchical structure in breast cancer molecular subtypes. In: *Springer Proceedings in Complexity.* (2018) p. 352–8.
255. de Anda-Jáuregui G. Guideline for comparing functional enrichment of biological network modular structures. *Appl Netw Sci.* (2019) 4:13. doi: 10.1007/s41109-019-0128-1
256. Bisognin A, Sales G, Coppe A, Bortoluzzi S, Romualdi C. MAGIA2: from miRNA and genes expression data integrative analysis to microRNA-transcription factor mixed regulatory circuits (2012 update). *Nucleic Acids Res.* (2012) 40:W13–21. doi: 10.1093/nar/gks460
257. Siska C, Bowler R, Kechris K. The discordant method: a novel approach for differential correlation. *Bioinformatics.* (2015) 32:690–6. doi: 10.1093/bioinformatics/btv633
258. Schlosser P, Knaus J, Schmutz M, Döhner K, Plass C, Bullinger L, et al. Netboost: Boosting-Supported Network Analysis Improves High-Dimensional Omics Prediction in Acute Myeloid Leukemia and Huntington’s Disease. (2019). Available from: <http://arxiv.org/abs/1909.12551v1>; <http://arxiv.org/pdf/1909.12551v1>
259. Champion M, Brennan K, Croonenborghs T, Gentles AJ, Pochet N, Gevaert O. Module analysis captures pancancer genetically and epigenetically deregulated cancer driver genes for smoking and antiviral response. *EBioMedicine.* (2018) 27:156–66. doi: 10.1016/j.ebiom.2017.11.028
260. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics.* (2008) 9:559. doi: 10.1186/1471-2105-9-559
261. Khatamian A, Paull EO, Califano A, Yu J. SJARACNe: a scalable software tool for gene network reverse engineering from big data. *Bioinformatics.* (2018) 35:2165–6. doi: 10.1093/bioinformatics/bty907
262. de Anda-Jáuregui G, Espinal-Enriquez J, Drago-García D, Hernández-Lemus E. Nonredundant, highly connected micrornas control functionality in breast cancer networks. *Int J Genom.* (2018) 2018:1–10. doi: 10.1155/2018/9585383
263. Hernández-Lemus E, Espinal-Enriquez J, de Anda-Jáuregui G. *Probabilistic Multilayer Networks.* (2018). Available online at: <http://arxiv.org/abs/1808.07857v1>; <http://arxiv.org/pdf/1808.07857v1>
264. De Domenico M, Solé-Ribalta A, Cozzo E, Kivelä M, Moreno Y, Porter MA, et al. Mathematical formulation of multilayer networks. *Phys Rev X.* (2013) 3:041022. doi: 10.1103/PhysRevX.3.041022

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 de Anda-Jáuregui and Hernández-Lemus. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Multi-Omic Regulation of the PAM50 Gene Signature in Breast Cancer Molecular Subtypes

Soledad Ochoa^{1,2}, Guillermo de Anda-Jáuregui^{1,3*} and Enrique Hernández-Lemus^{1,4*}

¹ Computational Genomics Division, National Institute of Genomic Medicine, Mexico City, Mexico, ² Graduate Program in Biomedical Sciences, Universidad Nacional Autónoma de México, Mexico City, Mexico, ³ Cátedras Conacyt para Jóvenes Investigadores, National Council on Science and Technology, Mexico City, Mexico, ⁴ Center for Complexity Sciences, Universidad Nacional Autónoma de México, Mexico City, Mexico

OPEN ACCESS

Edited by:

Chiara Romualdi,
University of Padova, Italy

Reviewed by:

Tanja Kunej,
University of Ljubljana, Slovenia
Valentina Silvestri,
Sapienza University of Rome, Italy

*Correspondence:

Guillermo de Anda-Jáuregui
gdeanda@inmegen.edu.mx
Enrique Hernández-Lemus
ehernandez@inmegen.gob.mx

Specialty section:

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

Received: 01 December 2019

Accepted: 29 April 2020

Published: 22 May 2020

Citation:

Ochoa S, de Anda-Jáuregui G and
Hernández-Lemus E (2020)
Multi-Omic Regulation of the PAM50
Gene Signature in Breast Cancer
Molecular Subtypes.
Front. Oncol. 10:845.
doi: 10.3389/fonc.2020.00845

Breast cancer is a disease that exhibits heterogeneity that goes from the genomic to the clinical levels. This heterogeneity is thought to be captured (at least partially) by the so-called breast cancer molecular subtypes. These molecular subtypes were initially defined based on the unsupervised clustering of gene expression and its correlate with histological, morphological, phenotypic and clinical features already known. Later, a 50-gene signature, PAM50, was defined in order to identify the biological subtype of a given sample within the clinical setting. The PAM50 signature was obtained by the use of unsupervised statistical methods, and therefore no limitation was set on the biological relevance (or lack of) of the selected genes beyond its predictive capacity. An open question that remains is what are the regulatory elements that drive the various expression behaviors of this set of genes in the different molecular subtypes. This question becomes more relevant as the measurement of more biological layers of regulation becomes accessible. In this work, we analyzed the gene expression regulation of the 50 genes in the PAM50 signature, in terms of (a) gene co-expression, (b) transcription factors, (c) micro-RNAs, and (d) methylation. Using data from the Cancer Genome Atlas (TCGA) for the Luminal A and B, Basal, and HER2-enriched molecular subtypes as well as normal tumor adjacent tissue, we identified predictors for gene expression through the use of an elastic net model. We compare and contrast the sets of identified regulators for the gene signature in each molecular subtype, and systematically compare them to current literature. We also identified a unique set of predictors for the expression of genes in the PAM50 signature associated with each of the molecular subtypes. Most selected predictors are exclusive for a PAM50 gene and predictors are not shared across subtypes. There are only 13 coding transcripts and 2 miRNAs selected for the four subtypes. *MIR-21* and *miR-10b* connect almost all the PAM50 genes in all the subtypes and normal tissue, but do it in an exclusive manner, suggesting a cancer switch from *miR-10b* coordination in normal tissue to *miR-21*. The PAM50 gene sets of selected predictors that enrich for a function across subtypes, support that different regulatory molecular mechanisms are taking place. With this study we aim to a wider understanding of the regulatory mechanisms that differentiate the expression of the PAM50 signature, which in turn could perhaps help understand the molecular basis of the differences between the molecular subtypes.

Keywords: multi-omic approaches, breast cancer subtypes, PAM50, elastic net, data integration

1. INTRODUCTION

Breast cancer is the most common cause of cancer death among females (1). Breast tumors have been classified in molecular subtypes with distinctive clinical characteristics and a recognizable gene expression signature (2). Such signature has been reduced to 50 genes that achieve the best separation of subtypes, attaining the PAM50 classifier (3). However, the physiological implications of the difference in gene expression, if any, are not well-understood.

Given that gene expression is regulated by several interconnected mechanisms (4–7), differences across subtypes are expected for these mechanisms. Evidence of this was found in the form of distinguishable patterns of DNA methylation, mutation and miRNA expression that shape groups partially equivalent to the molecular subtypes (8). These patterns imply a link between the different omics and PAM50 gene expression, but do not clarify which genomic, epigenetic or post transcriptional changes drive the expression signature of such molecular subtypes. To advance in the identification of such drivers of molecular subtypes expression, we propose the use of a sparse model of PAM50 gene expression.

Sparse models achieve the selection of the best predictors of an independent variable by fitting penalized linear models. The penalization of the regression coefficients aim is to shrink them toward zero in such a way that predictors contributing lowly to prediction i.e., poorly associated with the independent variable, end up with null coefficient values and get filtered out of the model (9). Ridge Regression, Least Absolute Shrinkage and Selection Operator, and Elastic Network methods apply different penalizations. The elastic network approach selects groups of pairwise correlated variables instead of choosing a single predictor from the group (10, 11), augmenting the space of predictors of interest but also incrementing false positive rates (12).

Sparse models have been proposed for multi-omic sample classification (13, 14) and biomarker identification (15–17); but their capacity to simplify multi-omics co-interpretation has only been tested in the evaluation of the extent of different omics effects over a phenotype (18, 19). Here, the predictor selection capability of the elastic network approach is exploited to identify the CpGs, coding transcripts, and miRNAs most associated with the expression of the PAM50 genes in order to outline molecular differences behind the gene expression patterns characterizing breast cancer subtypes within a true multi-omic framework. The hypothesis is that PAM50 gene expression patterns are accompanied by distinctive regulatory elements, reflecting the way gene expression is controlled in the different breast cancer subtypes.

2. METHODS

2.1. Data Acquisition

Concurrent experimental samples of DNA methylation, transcript and miRNA expression were downloaded from the GDC (<https://portal.gdc.cancer.gov/repository>) at May 2019. Only samples with Illumina Human Methylation 450, RNA-seq

and miRNA-seq measures were kept; filtering out samples quantified with the Illumina Human Methylation 27 BeadChip, which covers a smaller portion of the genome than the one we wanted to target. Subtype classification was also downloaded from the GDC through TCGABiolinks R package (20).

After preprocessing them according to Aryee et al. (21), Tarazona et al. (22), and Tam et al. (23), and biomaRt v95, values of methylation for 384,575 probes and expression for 16,475 coding transcripts and 433 miRNA precursors were obtained for 45 unique samples of Her2, 395 LumA, 128 LumB, and 125 Basal subtypes, plus 75 samples of non-tumor (normal adjacent) tissue.

2.2. Elastic Network Implementation

The three different data types were concatenated and normalized to have mean = 0 and standard deviation = 1. Eighty percent of the samples for each subtype were used for training, leaving the rest for testing as in Liu et al. (13). Using the R package glmnet (24), elastic network models were fitted per subtype for each gene in the PAM50 classifier with the linked script <https://github.com/CSB-IG/PAM50multiomics/blob/master/enetGLMNET.R>. The mixing parameter was held fixed at 0.5 because such value has shown a good performance (10), but shrinkage parameter (λ) was optimized between values from 0.001 and 1,000 through repeated cross-validation.

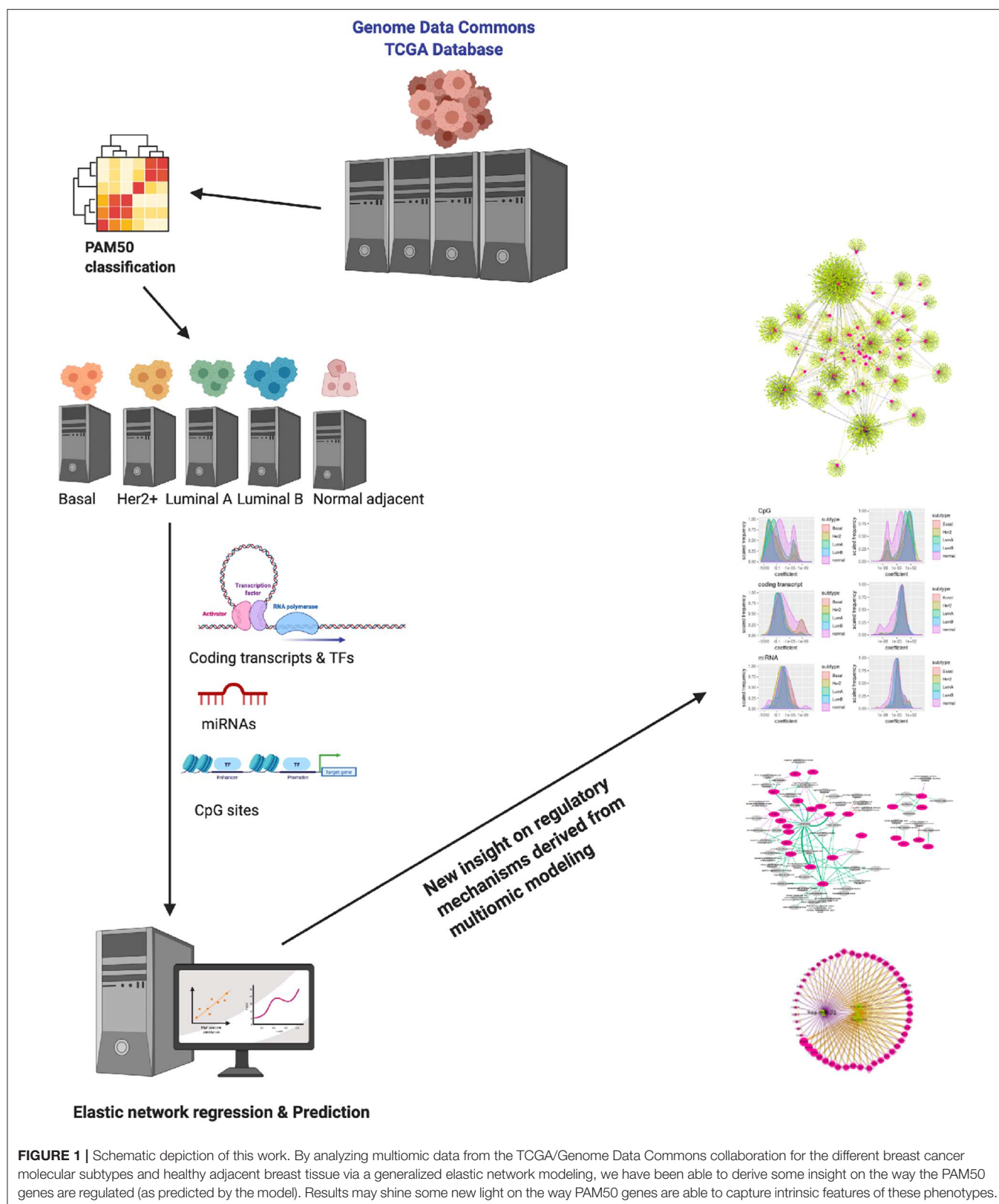
Cross-validation was repeated 100 times with $k = 3$ -folds for the subtypes with <100 training samples (Her2+ subtype and normal tissue) and $k = 5$ for the more represented subtypes (Luminal A, Luminal B, and Basal). Chosen λ parameters were used to predict testing data and root mean squared error (RMSE) was calculated per model. Fitting was repeated with the same specifications, for only 40 samples per subtype to verify the effect of data set size.

2.3. Omics Comparison

For each PAM50 gene model, RMSE was calculated for the testing data either with (1) the complete set of selected predictors, (2) only with selected CpGs, (3) just with selected coding transcripts, or (4) solely with selected miRNAs. Omic's specific RMSE were evaluated by zeroing all coefficients not associated to the omic of interest in the already fitted models with the linked script <https://github.com/CSB-IG/PAM50multiomics/blob/master/RMSEperOmics.R>, in an approach similar to the one used by Setty et al. (25) to search for key regulators. Obtained values shape RMSE distributions per omic which were compared via Kolmogorov–Smirnov test. This was done both per subtype per omic and mixing all the subtypes in a distribution per omic. *P*-values obtained were corrected for multiple testing with the FDR method.

2.4. Test vs. Reported Links Between Predictors and PAM50 Genes

Enrichment for previously reported regulatory links between PAM50 genes and CpGs, TFs, and miRNAs were tested by simple Fisher's Exact Test. Tests repeated by subtypes had *p*-values adjusted by FDR. Regulatory targets were taken from Illumina's annotation in the case of CpGs and from databases accessible through R packages in the case of TFs



and miRNAs, with the linked script <https://github.com/CSB-IG/PAM50multiomics/blob/master/validateInteractions.R>. tftargets <https://github.com/slowkow/tftargets> is the package used to

retrieve TF targets. It queries both predicted and validated data from TRED(2007), ITFP(2008), ENCODE(2012), and TRRUST(2015) databases at the date specified in parentheses

next to each resource, plus the lists curated by Neph et al. (26) and Marbach et al. (27).

The package used to retrieve miRNA targets is multiMiR v2.2 (28), it queries DIANA-microT-CDS, ElMMo, MicroCosm, miRanda, miRDB, PicTar, PITA, TargetScan, miRecords, miRTarBase, and TarBase, also reporting both experimentally validated and predicted results. Universe size for enrichment tests were taken from these databases, constrained to regulators measured in the input datasets. The hypothesis is that models selected reported associations between a PAM50 gene and a regulator measured in the input dataset more than expected.

2.5. Analysis of the Selected Predictors

Selected predictors and associated coefficient values were loaded to Cytoscape to construct a network of PAM50 gene predictors per subtype. PAM50 genes are taken as targets while predictors are sources, this makes a directed network were out and indegree are estimated. Predictors with the largest outdegree were submitted to an analysis of differential expression and their coefficient value distributions were compared to the global miRNA distribution via Kolmogorov–Smirnov tests. The differential analysis of miRNA expression was done per subtype by limma's package *treat* function in order to control for both fold change and significance (29). A minimum fold change of 1.1 was used.

2.6. Gene Enrichment Analysis

Every set of predictors selected for a PAM50 gene was submitted to functional enrichment analysis with the R package *HTSanalyzeR* v2.13.1 (30) versus the GO-BP with the linked script <https://github.com/CSB-IG/PAM50multiomics/blob/master/enrichment.R>. Sets enriched across subtypes were further tested via Fisher's Exact Test with the alternative hypothesis that selection in one subtype is exclusive with regards to selection another subtype.

The code to perform all previous analyses (see **Figure 1**) can be found at the following GitHub repository: <https://github.com/CSB-IG/PAM50multiomics>

3. RESULTS

Elastic network models were fitted per gene, regressing PAM50 gene expression to DNA methylation, miRNA and coding transcript expression. Elastic networks model shrink the regression coefficients toward 0, filtering predictors by its strength of association with the variable of interest. This ability for feature selection was exploited versus unfiltered omic data to identify the CpGs, coding transcripts and miRNAs most related to the PAM50 genes in cancer subtypes and normal tissue.

We fitted five models for each PAM50 gene, one per subtype and one for the normal tissue, since differences are expected for each of the 5 phenotypes. Descriptors of models per subtype and omic are reported in **Table 1**.

The output of the model are lists of associations between PAM50 genes and the selected predictors. Each selected predictor has a coefficient of regression whose value reflects the extent of association with the PAM50 gene. Coefficients are never zero,

TABLE 1 | Size of input and output of the models per subtype: Basal, Her2+, Luminal A, Luminal B as well as normal (i.e. tumor-adjacent healthy tissue).

	Basal	Her2+	LumA	LumB	Normal
Samples	125	45	395	128	75
Selected CpGs	3,090	2,514	7,173	1,485	5,373
Known CpGs selected	9	0	21	12	0
Selected coding transcripts	1,525	591	3,115	888	2,340
Selected TFs	207	91	465	133	327
Selected TFs predicted by another software	15	2	49	11	19
Selected TFs experimentally observed	4	3	25	7	9
miRNAs	101	85	174	116	123
Selected miRNAs predicted by another software	7	3	7	2	4
Selected miRNAs experimentally observed	8	5	8	12	5

since this value means predictors can be filtered out of the prediction; but can be both negative and positive indicating an opposite effect over the predicted value. Lists of associations shape networks like the one represented in **Figure 2**. Networks for the other subtypes and the normal tissue can be found at **Figures S1–S4**.

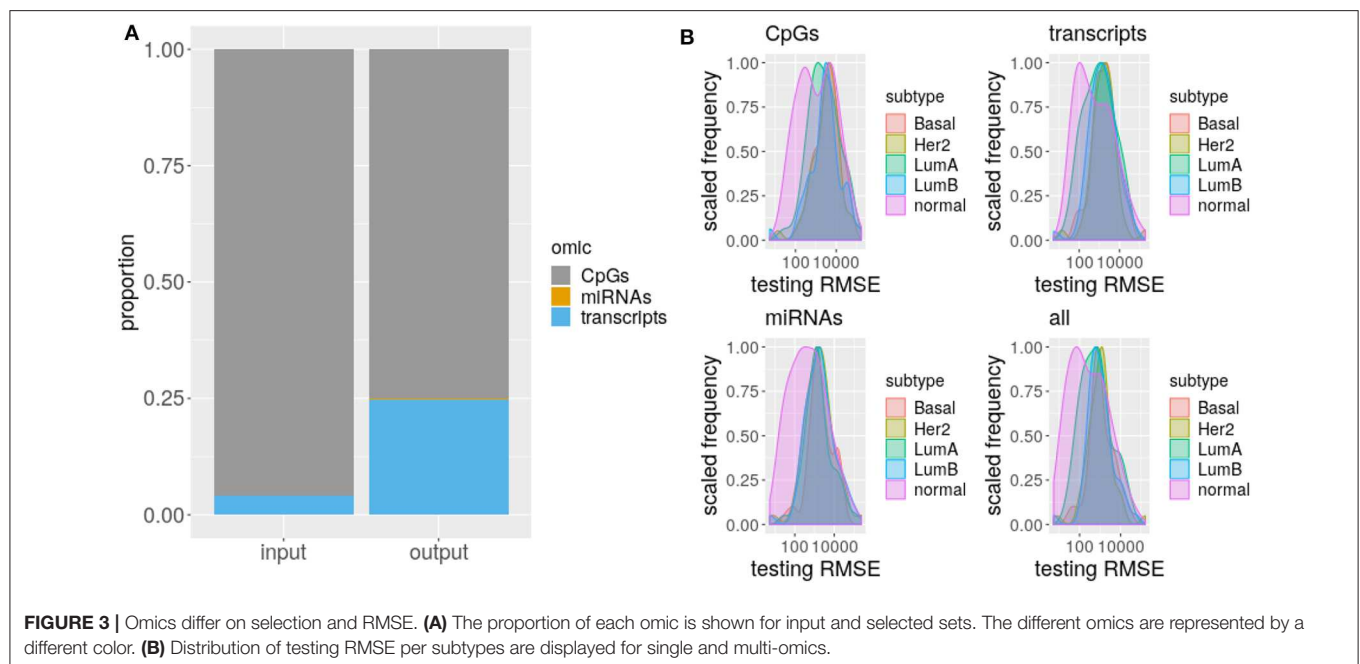
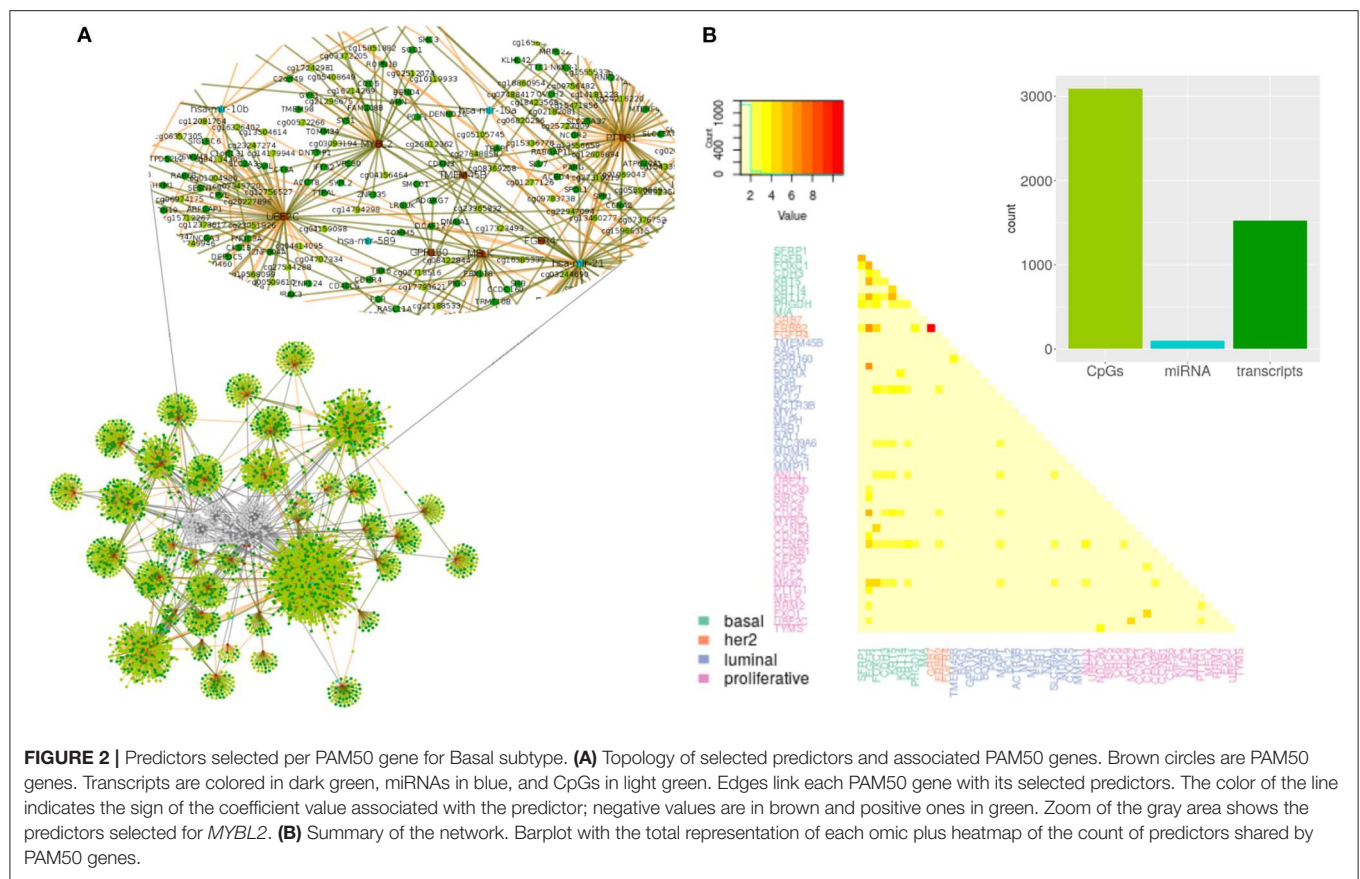
From observation of networks of selected predictors to PAM50 genes, it is evident that CpGs are the most selected predictors, followed by transcripts and with only a few miRNAs selected. It can also be seen that most predictors are exclusive of a PAM50 gene but all the PAM50 genes share predictors whose pattern of expression or methylation links one gene to another. This suggests the complete set of PAM50 expression is coordinated, independently of the gene being of luminal expression, basal, or any other signature.

3.1. Omics Contribute Differently to PAM50 Gene Expression Prediction in Normal Tissue and Cancer

In order to test the reliability of the fitted models, we checked the prediction error and the selection of previously reported associations. Regulation through DNA methylation, miRNA, or TF targeting is hence regarded as true positive and compared to model's results.

The proportion of selected predictors can not be explained solely by the size of the omics taken as input (χ^2 , p -value < $2.2\text{e-}16$, **Figure 3**), specifically, coding transcripts and miRNAs are overrepresented in the models (Fisher's Exact Test, p -value < $2.2\text{e-}16$). Concordantly, there are more true TF (Fisher's Exact Test, p -value $\leq 1.942846\text{e-}05$) and miRNA (Fisher's Exact Test, p -value $\leq 7.573200\text{e-}11$) relations than expected but less CpGs (Fisher's Exact Test, p -value $\leq 4.311267\text{e-}03$). The exception is LumB subtype which has as many true positive CpGs as expected.

Given the difference between input and selected proportion of omics, we hypothesized a discrepant prediction power of



CpGs, coding transcripts, and miRNAs. To test this, we evaluated models carrying the complete set of selected predictors or just the predictors from each omic.

As RMSE is a standard measure to compare regression models that measures how far is the model prediction from the observed data in response variable units, then, the lower its value the better.

Normally, the error decreases the more independent predictors are included in the model, so we choose not to fit again with the selected predictor per omic, but to test the exact same model with the jointly fitted coefficient values, just zeroing predictor's coefficients from other than the omic of interest. This way, the RMSE distribution of a model containing only predictors of a given omic, represents how much of the total prediction is contributed by the predictors from that omic.

As suggested by the difference with the input proportions, DNA methylation is the less predictive omic for all the subtypes, thought this difference is not always significant (CpGs vs. coding transcripts ks. test p -value ≤ 0.03192 for LumB, Her2+, and Basal and CpGs vs. miRNAs ks. test p -value ≤ 0.02222 for Her2+ and Basal). This disagrees with the great prediction improvement reported by Huang et al. (16) for methylation data, a fact that could be driven by the much larger and heterogenous input data used here, that we believe captures better the heterogeneity of breast cancer subtypes. Meanwhile, coding transcript and miRNAs contribute the same, with no significant difference between their distributions for all the subtypes.

Remarkably, the error distribution obtained with the complete set of predictors significantly outperforms CpGs and some subtype miRNAs (ks.test p -value ≤ 0.02222 for LumA and Basal) but never outweighs coding transcripts. Single omics can not beat multi-omics error due to the design of the test, thus the outperforming of CpGs and miRNAs is unsurprising, what is startling is the complete statistical agreement between multi-omics prediction power and coding transcripts prediction power, which supports gene expression as the current best biomarker of molecular subtypes. We must note however that this may be related to (1) more info on RNA and (2) PAM50 was derived from expression signatures.

Finally, there is no significant difference across subtypes RMSE distributions for both single-omics and multi-omics, but CpGs (ks.test p -value ≤ 0.01601952), miRNAs (ks.test p -value ≤ 0.002834981), and multi-omics (ks.test p -value ≤ 0.03919459) distributions of normal tissue differ from the distribution of each subtype, suggesting these omics represent a distinct amount of PAM50 gene expression in normal tissue than in cancer, that is, the association of DNA methylation and miRNA expression with PAM50 gene expression is altered in cancer.

3.2. The Association Strength Distributions of Predictors Are Different for Each Subtype

The difference between omics extends to coefficient values, shown in **Figure 4**. Since coefficients represent the strength of association between predictors and PAM50 expression (16), coefficient values suggest that each omic has a specific association with PAM50 gene expression. Coefficient value distributions are significantly different between subtypes (ks.test p -value $\leq 2.82E-02$) and omics (ks.test p -value ≤ 0.01535) with few exceptions for coding transcripts and miRNAs. Basal, Her2+, and LumB coding transcripts coefficients are not significantly different. Neither are miRNA coefficients of pairs LumA and normal tissue, LumB and Basal subtype, and Basal and Her2.

According to these distributions, DNA methylation has a strong but noisy association with PAM50 gene expression while miRNA (Fisher test p -values ≤ 0.001403597) and coding transcript (Fisher test p -values $\leq 1.086031e-29$) association tends to be positive (**Figure S3**) and more stable. The elevated association between DNA methylation and PAM50 genes expression explains why so many CpGs get selected in spite of its low prediction power. A stronger association between DNA methylation and gene expression than between gene and miRNA expression had previously been found for ovarian cancer by Sohn et al. (18) using a different penalization modeling.

3.3. *miR-21* and *miR-10b* Are the Only Relevant Predictors Selected Across Subtypes

Next, we wanted to see how variable is actually the association between one predictor and the predicted PAM50 gene, that is, the specific coefficient values, not their distributions. For this, we wanted to focus on the predictors selected for a PAM50 gene across subtypes, shown in **Figure 5**. However, as noted before, models selected a great quantity of predictors exclusive for each gene, 93.45% of the selected CpGs, 74.24% of the coding transcript, and 81.37% the miRNAs are not shared between any two genes. In consequence, there are no CpGs associated with any gene for all the subtypes but there are 14 relations with coding transcripts and 51 with miRNAs satisfying this.

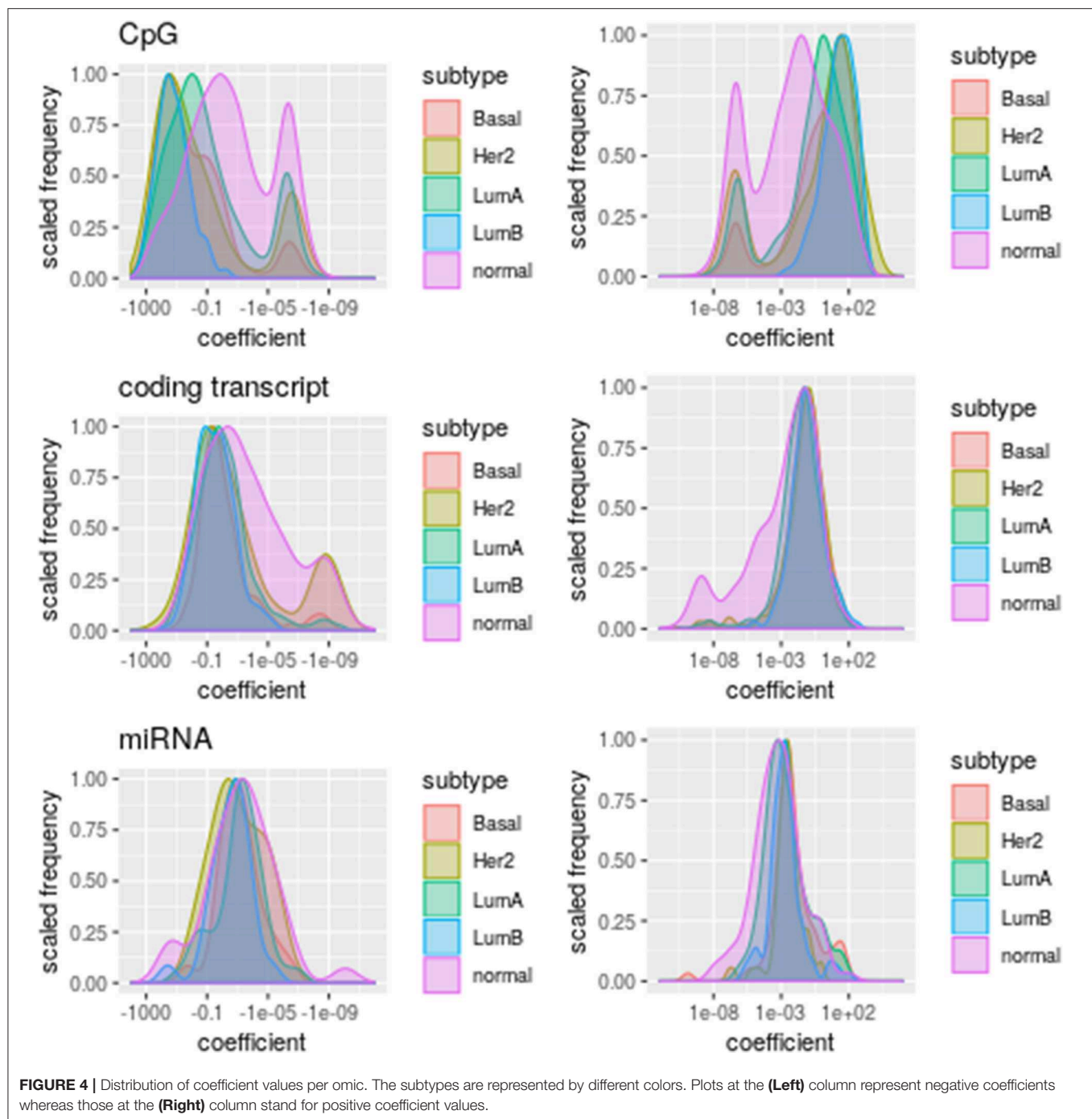
The 13 coding transcripts selected across subtypes as predictors of a specific PAM50 gene are trivial, since they just portray physical linkage. *ELP2* and *SLC39A6* are coded in opposite strands of the same locus while the rest of pairs are contiguous. Most of the associations, 84.77%, connect a PAM50 gene with a coding transcript in another chromosome, but these are not repeatedly selected across subtypes. It is worth mentioning that although all coefficients values are positive, even close predictors, like *YEATS4* and *SLC35E3* carry distinct coefficients.

Regarding miRNAs, there are only two miRNAs repeatedly selected among subtypes, *miR-10b* and *miR-21*. These are known breast cancer markers targeting some PAM50 genes (31). *Mir-21* has been experimentally linked with *BCL2*, *MYC*, *EGFR*, and *ERBB2* expression (32–35) and predicted to target *ESR1* and *FOXA1* (36, 37). On the other hand, *miR-10b* has been linked to *CDC6*, *EGFR*, and *SFRP1* (38, 39). There is no particular pattern among validated associations or coefficients, other than *miR-21* carrying mostly positive coefficient values and *miR-10b* selection extending up to normal tissue (for the full set of validated interactions please see **Supplementary Table S1**).

3.4. Micro-RNA *miR-21* and *miR-10b* Are Universal PAM50 Predictors in Cancer and Health

Next we wanted to check the role of *miR-21* and *miR-10b* per subtype. With this in mind, we revisited the models derived networks, that link PAM50 genes and predictors per subtype.

The networks show that genes overexpressed in each subtype get larger models. About 30% of the luminal genes have models

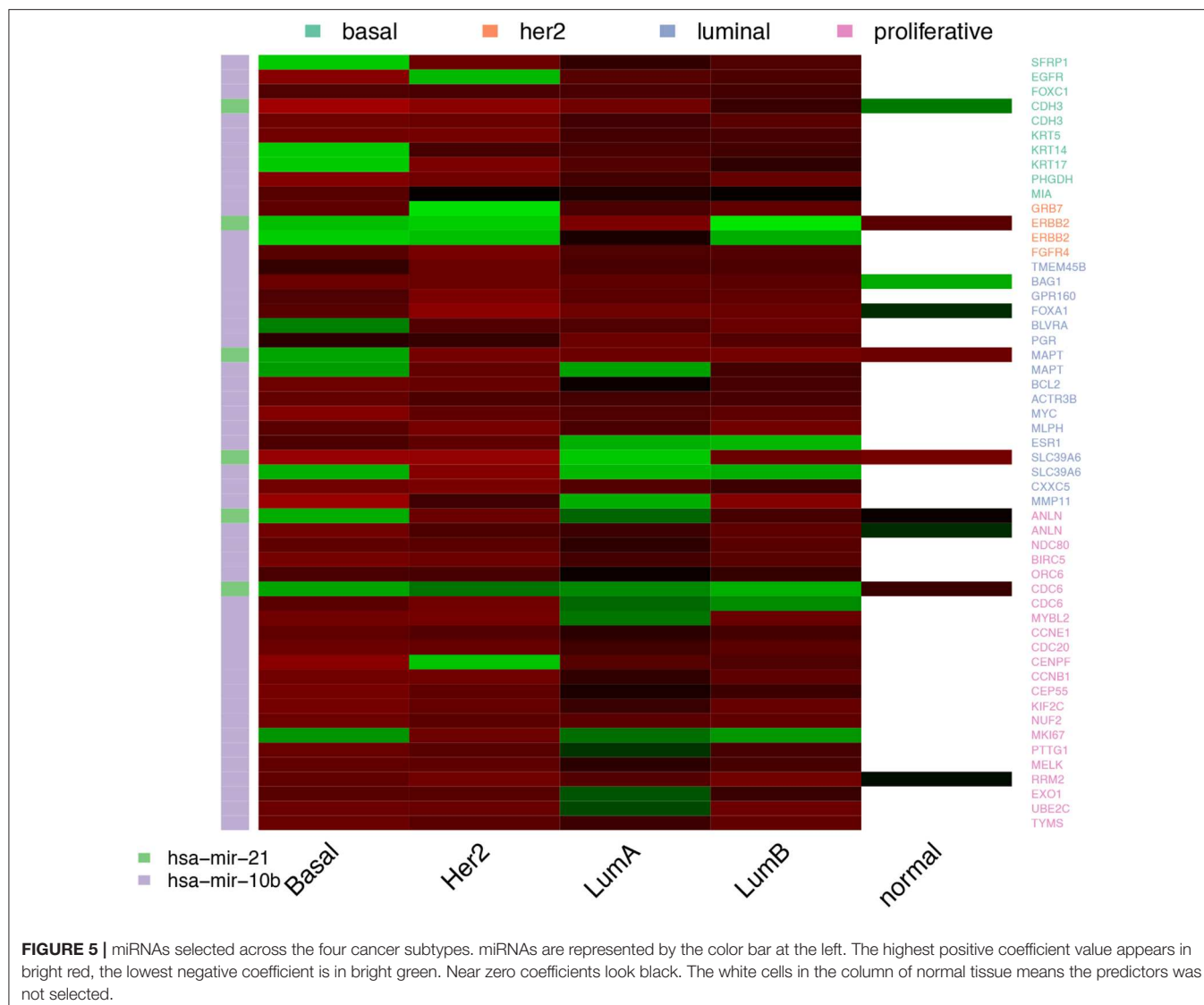


larger than average for LumA subtype, while almost 90% of basal genes have the equivalent for Basal subtype. Her2+ subtype and normal tissue have no clear pattern, but for LumB subtype, half the luminal genes and 28% of the proliferative ones have increased size models.

Predictors that bridge between PAM50 genes can proceed from any omic, but CpGs are significantly underrepresented (Fisher test p -values $\leq 1.81\text{E-}88$). CpGs are at most, selected for two subtypes as predictors of a specific PAM50 gene. There are just 24 CpGs in this situation, of which 15 are shared between

Her2+ and another subtype or the normal tissue, including nine CpGs associated with *ERBB2* but placed in other loci than chromosome 17.

Meanwhile, coding transcripts and miRNAs fulfill this role more often (Fisher test p -values $\leq 5.84\text{E-}03$) than solely input proportions would explain. This is no surprise since both pertain to the same level of molecular features, that of transcripts, as the PAM50 gene expression signature; as such, coding transcript and miRNA may be subject to the same biomolecular pressures. The stunning observation is that one miRNA can link almost all of



the PAM50 genes for all the cases (**Figure 6**). The outstanding miRNAs are again *miR-21* and *miR-10b*.

For normal tissue *miR-10b* was selected as predictor of all PAM50 genes while *miR-21* is linked to only four genes. On the contrary, *miR-21* is connected to most genes in all the breast cancer subtypes, while *miR-10b* is poorly linked. For LumA subtype, shown in **Figure 6B**, both *miR-10b* and *miR-10a* are highly connected, but still can not reach genes like *FOXC1*, which is connected instead with *miR-21*.

Both *miR-10a* and *miR-10b* are members of the miR-10 family encoded within the Hox genes genomic clusters; *miR-10a* resides upstream from *HOXB4* and *miR-10b* upstream from *HOXD4* (40). Due to their relatedness they will be referred as *miR-10a/b*.

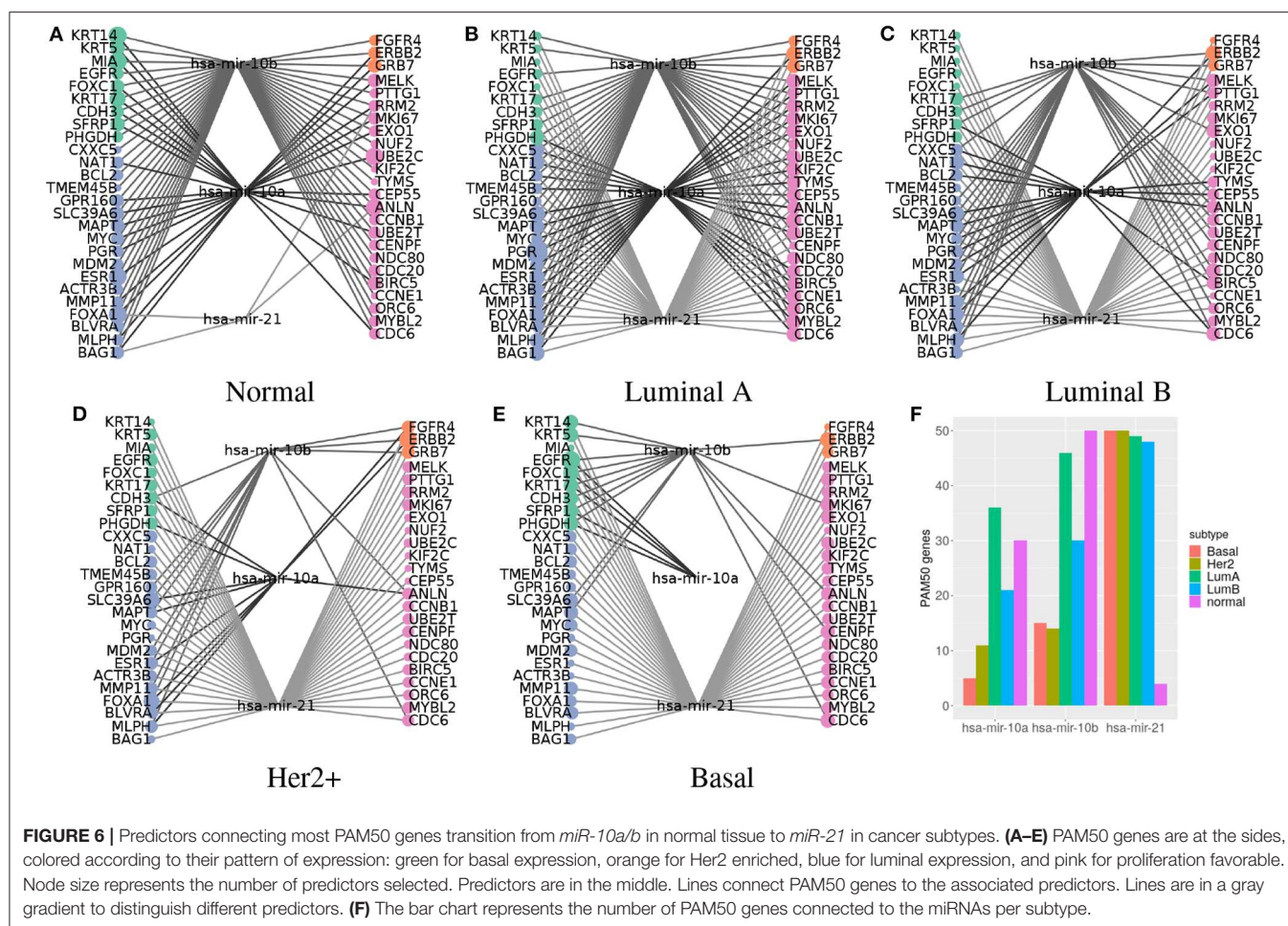
The hub-like behavior of these miRNAs agrees with previous observations of our group of highly connected miRNAs per subtype (41), which are important for network cohesion (42). Although the coefficients networks maintain a large connected component when removing *miR-10a/b* and *miR-21*, tens to

hundreds of predictors are needed to link all the PAM50 genes; when only one of these miRNAs is required to achieve the same.

Given that each miRNA has the potential to target hundreds of genes (43), *miR-10a/b* and *miR-21* are not so exceptional in this regard. However, as explained earlier, only a fraction of PAM50 genes have a regulatory relation with these miRNAs, suggesting most of the detected associations are indirect. Indirectness is consistent with the low values of the coefficients, which range from -0.2938690 to 0.4333184 , when miRNAs coefficient values range within two orders of magnitude higher. Coefficient value distributions of *miR-10a/b* and *miR-21* are also significantly different than the rest of miRNA coefficients (ks.test p -value $\leq 9.068e-05$).

3.5. PAM50 Genes Enrich for Different Functions per Subtype

The selection of predictors we have presented is based on a statistical association with the pattern of expression of a



PAM50 gene. The covariation sustaining such an association may respond to how a specific group of predictors is able to attain some biological function. To test this, functional enrichment was done with the set of selected predictors per gene per subtype, versus Gene Ontology Biological Processes categories (GO-BP) (Figure 7).

Only two PAM50 genes are enriched for some process in the Basal subtype, *FOXC1* (basal cluster) and *ANLN* (proliferative cluster). Neither the *ANLN* enrichment for telomere protection nor the *FOXC1* linkage to transforming growth factor response are within these genes immediate annotated processes. Though *FOXC1* is actually related with *TGFβ* since both are able to regulate EMT (44).

In the case of Her2+, just *ORC6* (proliferative cluster) is enriched for the totally unexpected process of synapse assembly, but, despite the significant *p*-value, we must notice that this is based on only two genes.

LumA is the most enriched subtype. This is not surprising since it has the largest number of selected coding transcripts, which is the starting material for enrichment. The 20 enriched genes are mostly linked to distinct cellular division aspects. The exception are the three keratins, genes with basal expression, which are connected through their normal processes, suggesting

selected predictors respond to the normal gene's function. *MYC* and *UBE2T* are linked to rather wide categories (45) while *MLPH* associates with other than its normal processes. The remaining 14 genes are connected through categories consistent with their proliferative expression, which again alludes to a selection that followed the normal function of the genes. This is again consistent with the available evidence.

For LumB subtype, *MELK* and *CCNB1* enrich for cell division as would be normally expected; while *MYBL2* is unintuitively linked to negative regulation of epithelial cell proliferation, which however, has been reported (46). Finally, the normal tissue shows different cell division aspects coherent with the proliferative expression of its enriched genes.

Altogether, few genes have predictors with significant enrichment extended across subtypes. Eight genes enriched in two subtypes, including *CCNB1*, *MKI67*, and *UBE2C*, that connect with the same processes, the expected ones, for the two subtypes. *MELK* also connects with its normal process for two subtypes but in LumA and LumB subtypes plus normal tissue. *ANLN*, *CEP55*, *KRT17*, *MYBL2*, and *ORC6*, enrich for different processes across subtypes, that is, a fifth of the genes with any kind of enrichment, but five of the nine genes enriched for more than one subtype.

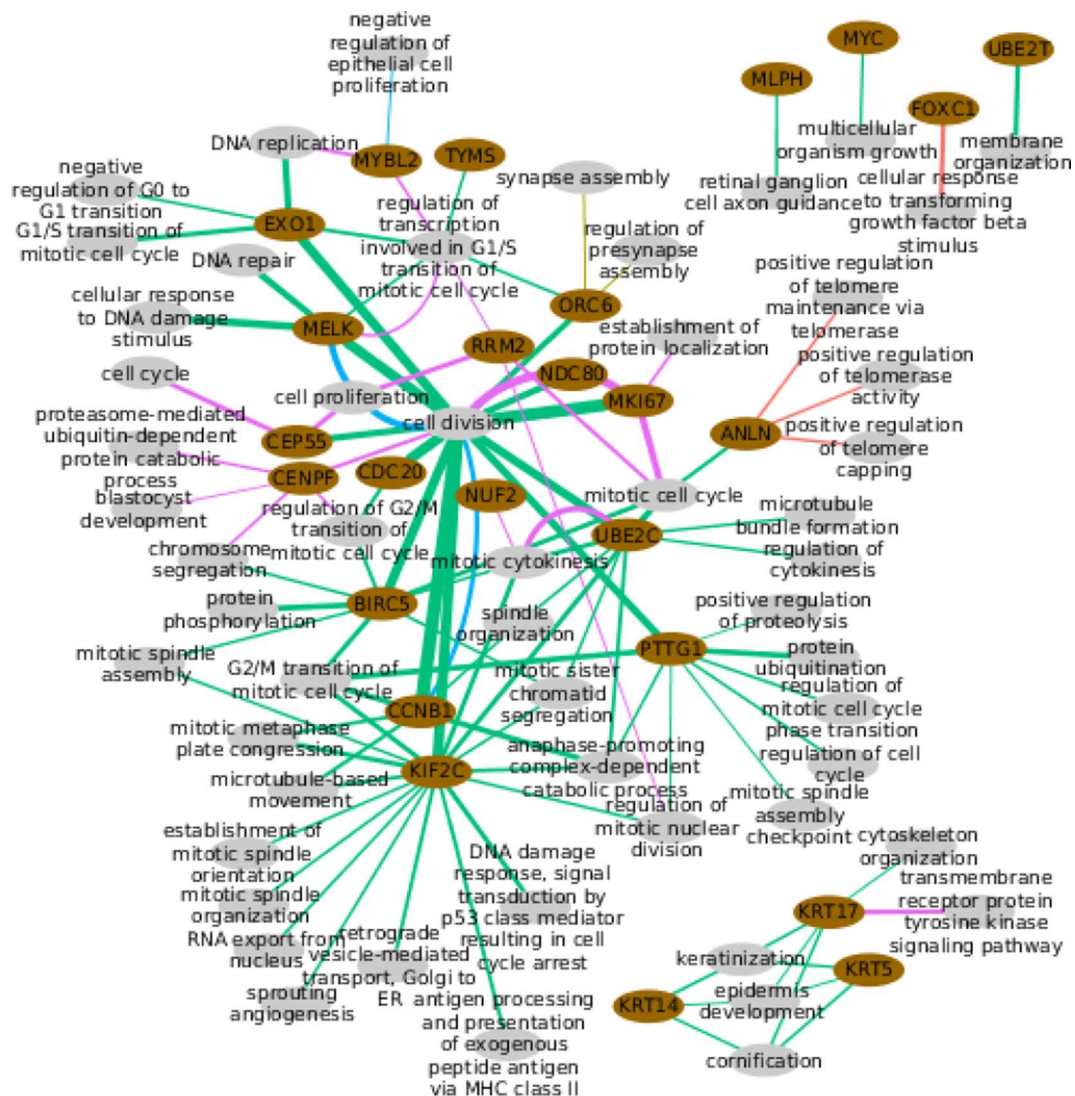


FIGURE 7 | Functional enrichment of the predictors selected per PAM50 gene. Pink ellipses represent PAM50 genes while the gray ellipses represent biological processes. Colored lines link genes with the processes they are significantly enriched to the corresponding subtype. Wider lines indicate a higher number of PAM50 gene predictors in the process.

To further test the functional enrichment per subtype, we compared the sets of predictors selected per subtype for each one of the 9 genes that enrich for several subtypes. Genes enriched for cell division across subtypes, *CCNB1*, *MKI67*, and *MELK* connect to the process via distinct sets of selected predictors. From the beginning, these genes bear different predictors (Fisher's Exact Test H1: less, $p\text{-value} \leq 1.281\text{e-}09$), with a small intersection whose removal does not change the significant enrichment for cell division. This reflects the robustness of the process, which is so important that distinct subsets of the 603 genes annotated in the category are enough to call it.

The other two genes enriched for the same process across subtypes, *UBE2C* for mitotic cytokinesis and, *MELK* for regulation of transcription involved in G1/S transition of mitotic

cell cycle, lost the functional enrichment when the predictors selected in both LumA and normal tissue (the intersection) were removed. This implies LumA mitotic cytokinesis and regulation of transcription may be involved in G1/S transition of mitotic cell cycle relying on the normal tissue mechanism.

The quantity of shared predictors between the sets selected for *CEP55*, indicates that predictor selection in the LumA subtype is exclusive for normal tissue selection (Fisher's Exact Test H1: less, $p\text{-value} = 1.141\text{e-}10$). This means that the differential enrichment between LumA and normal tissue is sustained by different predictors, suggesting *CEP55* fulfills divergent roles in these phenotypes. This matches differences observed between cancer and normal tissue (47) but, to our knowledge, not reported for LumA subtype.

The same reasoning supports *KRT17* and *ORC6* divergent roles across subtypes. It is odd that *KRT17* is linked to kinase signaling for normal tissue and not for a breast cancer subtype, when this has been described for another cancer (48) but this may be associated to tumor incidence over adjacent tissue (49). For *ANLN* and *MYBL2*, selection exclusion between subtypes is not significant, meaning that differential enrichment of these genes could settle on the same predictors, suggesting functional diversity.

4. DISCUSSION

Sparse penalized models have already proven useful to discover molecular mechanisms, cluster samples, and predict outcomes such as survival (50). Penalization permits the fitting of models otherwise unattainable given the relatively small sample sizes and huge number of variables measured by the omics. Here, the elastic network approach was used for integrated interpretation of different omics measuring DNA methylation and expression of both coding transcripts and miRNAs.

However, a large training set is always preferable, and not all breast cancer subtypes have been extensively sampled, which is reflected in the models. For Luminal A, the most frequent and sampled subtype, the highest number of predictors were selected by the models; while Her2+, with only 45 samples, got the lowest number of selected predictors. To assure comparability across subtypes we trained the models again, but now using the same number of samples, 40 samples, for all the subtypes. Patterns found with this subset persist in the analysis of the whole set of data, supporting comparability (Figures S5–S8). Nevertheless, the absence of predictors found for LumA in the smaller subtype's models due to a lack of representation can not be ruled out. This could specifically affect the functional enrichment of PAM50 neighborhoods of predictors and so, the functional divergence between subtypes is not definitive but should be experimentally tested.

Multi-omic modeling of PAM50 gene expression is no better than the sole use of coding transcripts, supporting gene expression as the best biomarker of molecular subtypes. However, our point in using the sparse model was not to predict PAM50 but to identify the molecular differences associated with PAM50 signatures that may lead to functional differences.

At the global level, a reduced prediction power of DNA methylation and miRNAs containing models was observed for all subtypes vs. normal tissue, indicating that the influence of this omics on PAM50 gene expression is reduced for cancer. Although this may be born out of incomplete knowledge or incipient technology, an alteration of these omics has been effectively reported; specifically, a generalized hypomethylation has been observed for breast and other cancers (51).

Different predictors were expected per cancer subtype, but the exclusivity of predictors from all the omics was surprisingly high. Only 13 coding transcripts and 2 miRNAs were selected for the four subtypes. The lack of CpGs selected across subtypes is consistent with the high strength of association it has with

PAM50 gene expression. If the pattern of expression is different between subtypes, the highly associated CpGs should be different.

The ubiquitous selection of *miR-10b* and *miR-21* across subtypes suggests a central role for these miRNAs in breast cancer, which is actually supported by the literature. Proliferation, cell migration, and *in vivo* tumor growth of MCF7 and MDA-MB-231 cell lines implanted in nude mice is inhibited through antagomiR-21 (52) demonstrating the relevance of this miRNA, at least for luminal A and triple negative subtypes. In turn, both sub and overexpression of *miR-10* are oncogenic. *MIR-10b* overexpression enhances cell migration and invasion by targeting *HOXD10*; while subexpression of *miR-10b-3p*, coded in the same *miR-10b* locus, participates in breast cancer onset by upregulating the cell cycle regulators *BUB1*, *PLK1*, and *CCNA2* (53).

Coherent with the ubiquitous selection of *miR-21* breast cancer subtypes and its replacement by *miR-10a/b* in normal tissue. *MIR-21* is significantly overexpressed for all cancer subtypes while *miR-10b* is underexpressed, as previous reports say (31). *MIR-10a* is significantly underexpressed in Basal and Her2+ subtypes and slightly overexpressed in luminal subtypes, but this is not significant in LumB case. The proposal is that when *miR-10b* coordinates PAM50 genes, normal tissue expression is predicted; when *miR-10b* is sub expressed and *miR-21* is overexpressed, this second miRNA gains *miR-10b* place, coordinating cancer expression of the PAM50 genes. Since *miR-10b* has a known role in metastasis (31), it would be interesting to observe the dynamics of the networks throughout the evolution of the disease.

Additionally, the small coefficients associated with these miRNAs are consistent with indirect associations. Considering all these pieces, the transition from hub *miR-10a/b* in normal tissue to *miR-21* in breast cancer through the luminal subtypes, evokes a switch between two master regulators. Master regulators are genes needed for the specification of a lineage by its capacity to regulate downstream genes either directly or not, whose misexpression can re-specify the fate of cells (54).

Nonetheless, sparse models can not select regulators naively, they need to feed on known regulators (16, 25, 55). Then, the regulatory capacity of selected predictor can not be stated, leaving *miR-10a/b* and *miR-21* just as universal predictors of PAM50 genes.

Another limitation of the study is the absence of an estimator of significance or accuracy intrinsic to the methodology (56). Regression models quality is described in terms of RMSE, without an indication of how well the selected predictors describe PAM50 expression. A ROC curve is not feasible, since models would have to be turned into the classification setting, and even this is unreachable, because true negative regulators can not be ascertained, as non regulators could simply be regulators yet to discover.

Finally, it is important to mention that applying the same shrinkage to inherently different molecular levels, like CpG methylation and transcript expression, could shrink to zero all the coefficients of subtler effect predictors (13). Thus, the next implementation of sparse multiomic models on PAM50 expression should adopt multiple penalizations, which could

even ameliorate the bias on subtype representation (57). Distinct values for the mixing parameter should also be probed, as well as data decomposition into latent variables (58).

Future Directions

Apart from exploration of alternative frameworks, the immediate follow up should be the experimental assessment of the observations described here. Specifically, silencing and expression of *miR-10a/b* and *miR-21* need to be tested for each breast cancer subtype. Dissection of interaction between the miRNAs and the PAM50 genes is required too.

Then, more omics could be included in the models. Copy number variation is the first candidate to be incorporated since it is already available in the databases and has a proven effect on Her2+ subtype, in particular regarding the effect of the *Her2* amplicon since it has been associated to regulation of growth and survival processes. But single nucleotide variation and chromatin accessibility are also available for some samples.

Other phenotypes with discriminant patterns of expression could benefit from sparse modeling. There could be significant predictors linked to the glioblastoma subtypes as was observed for breast cancer. Predictors represent potential regulators of the mechanisms behind subtype heterogeneity and, as such, are interesting markers of cancer. In this sense, predictor selection across stages, not subtypes, could illuminate the driving forces behind disease development. Alternative methods like A-JIVE (59) and sPLS (60) would have also exciting outcomes in this settings.

A relevant mid to long term future direction will be the implementation of experimental assays to test for multi-omic synergistic or cooperative phenomena, aiming at providing some mechanistic clues of the biological functions behind. There is however a strong challenge on this given the combinatorial mixture of effects that may be complex to disentangle. Some promissory (yet preliminary) advances are starting to arise.

5. CONCLUSION

Holistic studies of cancer are needed to dissect its complexity. Initiatives like The Cancer Genome Atlas have delivered the distinct molecular perspectives that need to be interpreted as a whole. The elastic net models subject of this work, approach such an integration in a rather simplistic linear form. Yet, the methodology is powerful enough to prove the intuition that PAM50 gene expression patterns are accompanied by distinctive potentially regulatory elements. Predictors are selected in an almost exclusive manner, heavily dictated by the omic of origin, with CpGs strongly associated to PAM50 expression not selected across subtypes. The way *miR-10a/b* and *miR-21*, the only relevant predictors selected for all subtypes,

are connected and differentially expressed, suggest an specific regulatory difference between breast cancer and normal tissue that merits further research.

DATA AVAILABILITY STATEMENT

The datasets analyzed for this study can be found in the Genome Data Commons site <https://bit.ly/2It0i2e>. The code to perform all previous analyses can be found at the following GitHub repository: <https://github.com/CSB-IG/PAM50multiomics>.

AUTHOR CONTRIBUTIONS

SO organized the database, performed the statistical analysis, and wrote the first draft of the manuscript. GA-J contributed to design of the study, generated programming code, and contributed to the writing of the manuscript. EH-L conceived the study, contributed to design of the study, provided funding, discussed findings, and reviewed the writing of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This work was supported by the Consejo Nacional de Ciencia y Tecnología [SEP-CONACYT-2016-285544 and FRONTERAS-2017-2115], and the National Institute of Genomic Medicine, México. Additional support has been granted by the Laboratorio Nacional de Ciencias de la Complejidad, from the Universidad Nacional Autónoma de México. EH-L is recipient of the 2016 Marcos Moshinsky Fellowship in the Physical Sciences.

ACKNOWLEDGMENTS

This paper constitutes a partial fulfilment of the Graduate Program in Biomedical Sciences of the National Autonomous University of México (UNAM) requirements of SO (María de la Soledad Ochoa-Méndez). She acknowledges the scholarship and support provided by the National Council of Science and Technology (CONACyT) and UNAM. **Figure 1** was generated using Biorender (<https://biorender.com/>).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fonc.2020.00845/full#supplementary-material>

Figures S1–S4 depict the topology of the networks for the non-basal subtypes that were not shown. **Table S1** contains a list of all validated interactions.

REFERENCES

- Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *Cancer J Clin.* (2018) 68:394–424. doi: 10.3322/caac.21492
- Prat A, Pineda E, Adamo B, Galván P, Fernández A, Gaba L, et al. Clinical implications of the intrinsic molecular subtypes of

- breast cancer. *Breast*. (2015) 24:S26–35. doi: 10.1016/j.breast.2015.07.008
3. Perou CM, Sørlie T, Eisen MB, Van De Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. *Nature*. (2000) 406:747. doi: 10.1038/35021093
4. Cheng C, Alexander R, Min R, Leng J, Yip KY, Rozowsky J, et al. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res*. (2012) 22:1658–67. doi: 10.1101/gr.136838.111
5. Vimalraj S, Miranda P, Ramyakrishna B, Selvamurugan N. Regulation of breast cancer and bone metastasis by microRNAs. *Dis Mark*. (2013) 35:369–87. doi: 10.1155/2013/451248
6. Cao J, Luo Z, Cheng Q, Xu Q, Zhang Y, Wang F, et al. Three-dimensional regulation of transcription. *Protein Cell*. (2015) 6:241–53. doi: 10.1007/s13238-015-0135-7
7. Liu X, Chen X, Yu X, Tao Y, Bode AM, Dong Z, et al. Regulation of microRNAs by epigenetics and their interplay involved in cancer. *J Exp Clin Cancer Res*. (2013) 32:96. doi: 10.1186/1756-9966-32-96
8. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*. (2012) 490:61–70. doi: 10.1038/nature11412
9. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning*. Vol. 112. New York, NY: Springer (2013). doi: 10.1007/978-1-4614-7138-7
10. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B*. (2005) 67:301–20. doi: 10.1111/j.1467-9868.2005.00503.x
11. Neto EC, Bare JC, Margolin AA. Simulation studies as designed experiments: the comparison of penalized regression models in the “large p, small n” setting. *PLoS ONE*. (2014) 9:e107957. doi: 10.1371/journal.pone.0107957
12. Kirpich A, Ainsworth EA, Wedow JM, Newman JR, Michailidis G, McIntyre LM. Variable selection in omics data: a practical evaluation of small sample sizes. *PLoS ONE*. (2018) 13:e0197910. doi: 10.1371/journal.pone.0197910
13. Liu J, Liang G, Siegmund KD, Lewinger JP. Data integration by multi-tuning parameter elastic net regression. *BMC Bioinformatics*. (2018) 19:369. doi: 10.1186/s12859-018-2401-1
14. Tini G, Marchetti L, Priami C, Scott-Boyer MP. Multi-omics integration-a comparison of unsupervised clustering methodologies. *Brief Bioinformatics*. (2019) 20:1269–79. doi: 10.1093/bib/bbx167
15. Bravo-Merodio L, Williams JA, Gkoutos GV, Acharjee A. -Omics biomarker identification pipeline for translational medicine. *J Transl Med*. (2019) 17:155. doi: 10.1186/s12967-019-1912-5
16. Huang S, Xu W, Hu P, Lakowski TM. Integrative analysis reveals subtype-specific regulatory determinants in triple negative breast cancer. *Cancers*. (2019) 11:507. doi: 10.3390/cancers11040507
17. Singh A, Shannon CP, Gautier B, Rohart F, Vacher M, Tebbutt SJ, et al. DIABLO: an integrative approach for identifying key molecular drivers from multi-omic assays. *Bioinformatics*. (2019) 35:3055–62. doi: 10.1093/bioinformatics/bty1054
18. Sohn KA, Kim D, Lim J, Kim JH. Relative impact of multi-layered genomic data on gene expression phenotypes in serous ovarian tumors. *BMC Syst Biol*. (2013) 7:S9. doi: 10.1186/1752-0509-7-S6-S9
19. Lock EF, Hoadley KA, Marron JS, Nobel AB. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann Appl Stat*. (2013) 7:523. doi: 10.1214/12-AOAS597
20. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Carolini D, et al. TCGAAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res*. (2016) 44:e71. doi: 10.1093/nar/gkv1507
21. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*. (2014) 30:1363–9. doi: 10.1093/bioinformatics/btu049
22. Tarazona S, Furió-Tarí P, Turrà D, Pietro AD, Nueda MJ, Ferrer A, et al. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Res*. (2015) 43:e140. doi: 10.1093/nar/gkv711
23. Tam S, Tsao MS, McPherson JD. Optimization of miRNA-seq data preprocessing. *Brief Bioinformatics*. (2015) 16:950–63. doi: 10.1093/bib/bbv019
24. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. (2010) 33:1–22. doi: 10.18637/jss.v033.i01
25. Setty M, Helmy K, Khan AA, Silber J, Arvey A, Neezen F, et al. Inferring transcriptional and microRNA-mediated regulatory programs in glioblastoma. *Mol Syst Biol*. (2012) 8:605. doi: 10.1038/msb.2012.37
26. Neph S, Stergachis AB, Reynolds A, Sandstrom R, Borenstein E, Stamatoiyannopoulos JA. Circuitry and dynamics of human transcription factor regulatory networks. *Cell*. (2012) 150:1274–86. doi: 10.1016/j.cell.2012.04.040
27. Marbach D, Lamparter D, Quon G, Kellis M, Kutalik Z, Bergmann S. Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nat Methods*. (2016) 13:366–70. doi: 10.1038/nmeth.3799
28. Ru Y, Kechris KJ, Tabakoff B, Hoffman P, Radcliffe RA, Bowler R, et al. The multiMiR R package and database: integration of microRNA-target interactions along with their disease and drug associations. *Nucleic Acids Res*. (2014) 42:e133. doi: 10.1093/nar/gku631
29. McCarthy DJ, Smyth GK. Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics*. (2009) 25:765–71. doi: 10.1093/bioinformatics/btp053
30. Wang X, Terfve C, Rose JC, Markowitz F. HTSanalyzeR: an R/Bioconductor package for integrated network analysis of high-throughput screens. *Bioinformatics*. (2011) 27:879–80. doi: 10.1093/bioinformatics/btr028
31. O’Day E, Lal A. MicroRNAs and their target gene networks in breast cancer. *Breast Cancer Res*. (2010) 12:201. doi: 10.1186/bcr2484
32. Si ML, Zhu S, Wu H, Lu Z, Wu F, Mo YY. miR-21-mediated tumor growth. *Oncogene*. (2007) 26:2799–803. doi: 10.1038/sj.onc.1210083
33. Bhat-Nakshatri P, Wang G, Collins NR, Thomson MJ, Geistlinger TR, Carroll JS, et al. Estradiol-regulated microRNAs control estradiol response in breast cancer cells. *Nucleic Acids Res*. (2009) 37:4850–61. doi: 10.1093/nar/gkp500
34. Barker A, Giles KM, Epis MR, Zhang PM, Kalinowski F, Leedman PJ. Regulation of ErbB receptor signalling in cancer cells by microRNA. *Curr Opin Pharmacol*. (2010) 10:655–61. doi: 10.1016/j.coph.2010.08.011
35. Huang TH, Wu F, Loeb GB, Hsu R, Heidersbach A, Brincat A, et al. Up-regulation of miR-21 by HER2/neu signaling promotes cell invasion. *J Biol Chem*. (2009) 284:18515–24. doi: 10.1074/jbc.M109.006676
36. Gaidatzis D, van Nimwegen E, Hausser J, Zavolan M. Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics*. (2007) 8:69. doi: 10.1186/1471-2105-8-69
37. Maragkakis M, Reczko M, Simossis VA, Alexiou P, Papadopoulos GL, Dalamagas T, et al. DIANA-microT web server: elucidating microRNA functions through target prediction. *Nucleic Acids Res*. (2009) 37:W273–6. doi: 10.1093/nar/gkp292
38. Kishore S, Jaskiewicz L, Burger L, Hausser J, Khorshid M, Zavolan M. A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat Methods*. (2011) 8:559–64. doi: 10.1038/nmeth.1608
39. Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS. MicroRNA targets in Drosophila. *Genome Biol*. (2003) 5:R1. doi: 10.1186/gb-2003-5-1-r1
40. Lund AH. miR-10 in development and cancer. *Cell Death Differ*. (2010) 17:209–14. doi: 10.1038/cdd.2009.58
41. de Anda-Jáuregui G, Espinal-Enríquez J, Drago-García D, Hernández-Lemus E. Nonredundant, highly connected microRNAs control functionality in breast cancer networks. *Int J Genomics*. (2018) 2018:9585383. doi: 10.1155/2018/9585383
42. Drago-García D, Espinal-Enríquez J, Hernández-Lemus E. Network analysis of EMT and MET micro-RNA regulation in breast cancer. *Sci Rep*. (2017) 7:13534. doi: 10.1038/s41598-017-13903-1
43. Helwak A, Kudla G, Dudnakova T, Tollervey D. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell*. (2013) 153:654–65. doi: 10.1016/j.cell.2013.03.043
44. Yu M, Bardia A, Wittner BS, Stott SL, Smas ME, Ting DT, et al. Circulating breast tumor cells exhibit dynamic changes in epithelial and mesenchymal composition. *Science*. (2013) 339:580–4. doi: 10.1126/science.1228522
45. Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res*. (2019) 47:D419–26. doi: 10.1093/nar/gky1038

46. Martin FT, Dwyer RM, Kelly J, Khan S, Murphy JM, Curran C, et al. Potential role of mesenchymal stem cells (MSCs) in the breast tumour microenvironment: stimulation of epithelial to mesenchymal transition (EMT). *Breast Cancer Res Treat.* (2010) 124:317–26. doi: 10.1007/s10549-010-0734-1
47. Jeffery J, Sinha D, Srihari S, Kalimutho M, Khanna KK. Beyond cytokinesis: the emerging roles of CEP55 in tumorigenesis. *Oncogene.* (2016) 35:683–90. doi: 10.1038/onc.2015.128
48. Sankar S, Tanner JM, Bell R, Chaturvedi A, Randall RL, Beckerle MC, et al. A novel role for keratin 17 in coordinating oncogenic transformation and cellular adhesion in Ewing sarcoma. *Mol Cell Biol.* (2013) 33:4448–60. doi: 10.1128/MCB.00241-13
49. Aran D, Camarda R, Odegaard J, Paik H, Oskotsky B, Krings G, et al. Comprehensive analysis of normal adjacent to tumor transcriptomes. *Nat Commun.* (2017) 8:1077. doi: 10.1038/s41467-017-01027-z
50. Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, et al. Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics.* (2016) 17(Suppl. 2):15. doi: 10.1186/s12859-015-0857-9
51. Vidal Ochoa E, Sayols S, Moran S, Guillaumet-Adkins A, Schroeder MP, Royo R, et al. A DNA methylation map of human cancer at single base-pair resolution. *Oncogene.* (2017) 36:5648–57. (2017). doi: 10.1038/onc.2017.176
52. Wang SE, Lin RJ. MicroRNA and HER2-overexpressing cancer. *MicroRNA.* (2013) 2:137–47. doi: 10.2174/22115366113029990011
53. Biagioni F, Bossel Ben-Moshe N, Fontemaggi G, Yarden Y, Domany E, Blandino G. The locus of microRNA-10b: a critical target for breast cancer insurgence and dissemination. *Cell Cycle.* (2013) 12:2371–5. doi: 10.4161/cc.25380
54. Chan SSK, Kyba M. What is a master regulator? *J Stem Cell Res Ther.* (2013) 3:114. doi: 10.4172/2157-7633.1000e114
55. Li W, Zhang S, Liu CC, Zhou XJ. Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics.* (2012) 28:2458–66. doi: 10.1093/bioinformatics/bts476
56. Pineda S, Real FX, Kogevinas M, Carrato A, Chanock SJ, Malats N, et al. Integration analysis of three omics data using penalized regression methods: an application to bladder cancer. *PLoS Genet.* (2015) 11:e1005689. doi: 10.1371/journal.pgen.1005689
57. Lee G, Bang L, Kim SY, Kim D, Sohn KA. Identifying subtype-specific associations between gene expression and DNA methylation profiles in breast cancer. *BMC Med Genomics.* (2017) 10:28. doi: 10.1186/s12920-017-0268-z
58. Lê Cao KA, Martin PGP, Robert-Granié C, Besse P. Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics.* (2009) 10:34. doi: 10.1186/1471-2105-10-34
59. Feng Q, Jiang M, Hannig J, Marron J. Angle-based joint and individual variation explained. *J Multivar Anal.* (2018) 166:241–65. doi: 10.1016/j.jmva.2018.03.008
60. Rohart F, Gautier B, Singh A, Le Cao KA. mixOmics: An R package for-omics feature selection and multiple data integration. *PLoS Comput Biol.* (2017) 13:e1005752. doi: 10.1371/journal.pcbi.1005752

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Ochoa, de Anda-Jáuregui and Hernández-Lemus. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Big Data-Based Identification of Multi-Gene Prognostic Signatures in Liver Cancer

Meiliang Liu^{1†}, Xia Liu^{2†}, Shun Liu¹, Feifei Xiao³, Erna Guo^{1,4}, Xiaoling Qin¹, Liuyu Wu¹, Qiuli Liang¹, Zerui Liang¹, Kehua Li¹, Di Zhang¹, Yu Yang¹, Xingxi Luo¹, Lei Lei¹, Jennifer Hui Juan Tan⁵, Fuqiang Yin^{6,7*} and Xiaoyun Zeng^{1,7*}

¹ School of Public Health, Guangxi Medical University, Nanning, China, ² Key Laboratory of Longevity and Ageing-Related Disease of Chinese Ministry of Education, Centre for Translational Medicine and School of Preclinical Medicine, Guangxi Medical University, Nanning, China, ³ Department of Epidemiology and Biostatistics, University of South Carolina, Columbia, SC, United States, ⁴ School of International Education, Guangxi Medical University, Nanning, China, ⁵ School of Life Sciences and Chemical Technology, Ngee Ann Polytechnic, Singapore, Singapore, ⁶ Life Sciences Institute, Guangxi Medical University, Nanning, China, ⁷ Key Laboratory of High-Incidence-Tumor Prevention and Treatment, Guangxi Medical University, Ministry of Education, Nanning, China

OPEN ACCESS

Edited by:

Chiara Romualdi,
University of Padova, Italy

Reviewed by:

Yanqiang Li,
Houston Methodist Research Institute,
United States
Fang Wang,
University of Texas MD Anderson
Cancer Center, United States

*Correspondence:

Fuqiang Yin
yinfq@mail2.sysu.edu.cn
Xiaoyun Zeng
zengxiaoyun@gxmu.edu.cn

[†]These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

Received: 02 January 2020

Accepted: 29 April 2020

Published: 28 May 2020

Citation:

Liu M, Liu X, Liu S, Xiao F, Guo E,
Qin X, Wu L, Liang Q, Liang Z, Li K,
Zhang D, Yang Y, Luo X, Lei L,
Tan JHJ, Yin F and Zeng X (2020) Big
Data-Based Identification of
Multi-Gene Prognostic Signatures in
Liver Cancer. *Front. Oncol.* 10:847.
doi: 10.3389/fonc.2020.00847

Simultaneous identification of multiple single genes and multi-gene prognostic signatures with higher efficacy in liver cancer has rarely been reported. Here, 1,173 genes potentially related to the liver cancer prognosis were mined with Coremine, and the gene expression and survival data in 370 samples for overall survival (OS) and 319 samples for disease-free survival (DFS) were retrieved from The Cancer Genome Atlas. Numerous survival analyses results revealed that 39 genes and 28 genes significantly associated with DFS and OS in liver cancer, including 18 and 12 novel genes that have not been systematically reported in relation to the liver cancer prognosis, respectively. Next, totally 9,139 three-gene combinations (including 816 constructed by 18 novel genes) for predicting DFS and 3,276 three-gene combinations (including 220 constructed by 12 novel genes) for predicting OS were constructed based on the above genes, and the top 15 of these four parts three-gene combinations were selected and shown. Moreover, a huge difference between high and low expression group of these three-gene combination was detected, with median survival difference of DFS up to 65.01 months, and of OS up to 83.57 months. The high or low expression group of these three-gene combinations can predict the longest prognosis of DFS and OS is 71.91 months and 102.66 months, and the shortest is 6.24 months and 13.96 months. Quantitative real-time polymerase chain reaction and immunohistochemistry reconfirmed that three genes *F2*, *GOT2*, and *TRPV1* contained in one of the above combinations, are significantly dysregulated in liver cancer tissues, low expression of *F2*, *GOT2*, and *TRPV1* is associated with poor prognosis in liver cancer. Overall, we discovered a few novel single genes and multi-gene combinations biomarkers that are closely related to the long-term prognosis of liver cancer, and they can be potential therapeutic targets for liver cancer.

Keywords: liver cancer, gene combinations, data mining, disease-free survival (DFS), overall survival (OS)

INTRODUCTION

Liver cancer is the sixth most common cancer and the fourth leading cause of cancer-related deaths (1). Specifically, hepatocellular carcinoma (HCC) accounts for more than 90% of liver cancer cases from a histopathological perspective. According to the GLOBOCAN 2018 database, there are about 841,000 new HCC cases and 782,000 related deaths worldwide each year, with China accounting for nearly half of the total number of global HCC cases and deaths (2, 3). In China, the Guangxi province has higher morbidity and mortality rates than the national average (4). The high mortality and poor prognosis of HCC poses a global challenge. Despite the slight increase in the 5-year survival rate of liver cancer in China from 10.1 to 12.1% over the periods of 2003–2015, it still remains at a low level (5). A survival analysis of 2,887 liver cancer patients in 14 years showed that the 1-year, 3-year, and 5-year survival rates were 49.3, 26.6, and 19.5%, respectively (6).

Although there are many existing therapies for HCC including surgical resection, transplantation, ablation, and transcatheter chemoembolization, etc., the long-term survival of HCC patients remains poor due to their limited indications and different effects on prognosis (7–10). A 20-year prospective cohort analysis reported that the 5-year survival rates of TNM stage I, II, IIIA, and IVA patients after hepatectomy were 81.7, 77.2, 44, and 28.2%, respectively (11). Therefore, it is of crucial importance to explore new prognostic biomarkers and investigate treatment strategies to improve the overall prognosis of HCC patients.

Currently, the research on prognostic molecular markers of HCC is still ongoing, and many single-gene or multi-gene combination molecular markers related to HCC invasion, metastasis and prognosis are being gradually discovered. For example, the expression of *HMGAI* in HCC is associated with poor prognosis and is found to promote tumor growth and migration *in vitro* (12). The overexpression of *SYPL1* is associated with epithelial-mesenchymal transition (EMT) of HCC cells and can predict the prognosis of HCC (13). *RBM8A* and *SIRT5* promote the migration and invasion of HCC cells by activating the EMT signaling pathway and targeting *E2F1* (14, 15), respectively (16, 17). The *EpCAM* (18), a liver X receptor (*LXR*) (19), *SPAG5* (20), and *KOR* (21) have been shown to be strongly correlated with HCC metastasis, invasion, or prognosis. Arginase-1, *FTCD*, and *MOC-31* have a good performance in the diagnosis of HCC (22). *TMEM88*, *CCL14*, and *CLEC3B* can serve as potential prognostic markers of HCC (23). At the same time, some multi-gene combined prognostic studies on HCC have also been reported. For example, three genes (*UPB1*, *SOCS2*, *RTN3*) combination markers (24) and four genes (*CENPA*, *SPP1*, *MAGEB6*, *HOXD9*) combination models can predict the overall survival in patients with HCC prognosis (25).

However, due to the sample size limitation and the heterogeneity of the samples in different studies, the efficiency of the identified prognostic markers for liver cancer still has ample space to improve. In addition, because of the myriad of gene interaction capabilities and the possibility of synergistic promotion of disease progression, it is of great significance to find some multi-gene combinations that may have better prognostic

efficacy than single genes for prognostic targets of liver cancer. Therefore, the leverage of the large sample sizes of the public data platforms, integrating new and effective mining and screening methods, as well as reliable experimental verification is a very promising direction for the discovery of multiple effective single genes and multi-gene combination prognostic markers of liver cancer.

High-throughput profiling technologies and bioinformatics methods are now being applied to all fields of biomedical research. A mass of cancer data, such as the mRNA expression, copy number variation, single nucleotide polymorphism (SNP), and microRNA expression generated by those tools are collected in public archives such as The Cancer Genome Atlas (TCGA) (<http://cancergenome.nih.gov/>), Coremine (<http://www.coremine.com/medical/>), Oncomine (<https://www.oncomine.org/resource/login.html>), Gene Expression Omnibus database (GEO, <https://www.ncbi.nlm.nih.gov/geo/>), etc. Making full use of the public data from these databases is meaningful for exploring and discovering effective HCC prognostic biomarkers. For instance, Li et al. (24) developed a three-gene prognostic signature composing of three genes *UPB1*, *SOCS2*, and *RTN3*, which was revealed to have prognostic value for HCC patients based on TCGA data. Our previous study used data retrieved from the Coremine, TCGA, and GEO database and discovered that high-expressed E2F transcription factor 3 is associated with poor prognosis of HCC (26).

In this study, we used text mining approach to find the medial related candidate gene list for liver cancer prognosis, and a total of 1,173 genes that might be related to the prognosis of liver cancer were finally obtained. The association of the 1,173 genes with overall survival (OS) and disease-free survival (DFS) was accessed in a large sample of TCGA cohort, in which the subgroups of 319 patients with DFS and 370 with OS were available. The survival analyses are carried out for each of these genes to identify single prognostic markers. Moreover, we performed survival analyses of the gene combinations and performed multiple screening for these HCC prognostic molecular markers, revealing the association between the expression of numerous genes or gene combinations and the survival in HCC patients. We then compared the ability of single genes and multiple gene combinations to predict the prognosis of HCC. Moreover, a huge difference between high and low expression group of these three-gene combinations was detected, with median survival difference of DFS up to 65.01 months, and of OS up to 83.57 months. The high or low expression group of these three-gene combinations can predict the longest prognosis of DFS and OS is 71.91 months and 102.66 months, and the shortest is 6.24 months and 13.96 months. Among the above genes that may be strongly correlated with the prognosis of HCC identified in large sample data, it was found that the combination of the three genes *F2*, *GOT2*, and *TRPV1* that have not been systematically reported has a strong ability to predict the prognosis of HCC. We further verified *F2*, *GOT2*, and *TRPV1* by three independent expression profile microarray data for liver cancer acquired from the Oncomine database, and conducted the quantitative real-time polymerase chain reaction (qRT-PCR) in 20 pairs of HCC and adjacent tissues, and immunohistochemistry

(IHC) staining in 90 pairs of HCC and its precancerous tissues. These results validated that the low expression of *F2*, *GOT2*, and *TRPV1* in liver cancer was associated with the poor prognosis of liver cancer.

MATERIALS AND METHODS

Data Sources

We combined 3 corresponding concepts of the key word “liver cancer” with 2 concepts of the key word “prognosis” and 10 concepts of the key word “outcome,” respectively, (**Supplementary Table S1**), and searched for their corresponding genes or proteins in the Coremine database (<http://www.coremine.com/medical/>). After deleting duplicates, we selected 1,173 gene entries with *p*-values < 0.05 that might be associated with the prognosis of liver cancer for further analyses (**Supplementary Table S2**).

The above genes mined in the Coremine database include some genes obtained from other gene-mining reports; however, the number of samples and data standards in each report is different. Therefore, we selected the cohort of The Cancer Genome Atlas (TCGA) (<http://cancergenome.nih.gov/>), a database with consistent sample size and data standards, to conduct unified batch verification of these genes and conduct three-gene combinations survival analyses.

We studied the relationship between each of the selected 1,173 genes and the prognosis of liver cancer patients in TCGA cohort which downloaded from cBioPortal for Cancer Genomics (<https://www.cbioportal.org/>) in September 2018 (27, 28), and a subgroup of 319 liver cancer samples with HCC DFS corresponding follow-up data and a subgroup of 370 liver cancer samples with HCC OS corresponding follow-up data were chosen.

Survival Analysis and Gene Selection

Kaplan-Meier estimation of survival functions and Log-rank tests were used to evaluate effect of genes on DFS and OS. The Cox proportional hazard model was performed for multivariate analyses of HCC prognosis. Survival analyses were performed using the R survival package in R (version 3.3.1). The Kaplan-Meier survival curves and Cox proportional hazards regression model for DFS and OS were generated by IBM SPSS (version 23.0). The median expression level of a gene was used as a cutoff value for the classification of patients into high and low expression groups (29).

Human Tissue Samples

For the validation studies, we used 20 patients who underwent primary and curative hepatectomy from Apr 2016 to Apr 2018 at the First Affiliated Hospital of Guangxi Medical University. Those patients who have distinctive pathologic diagnosis of HCC without preoperative anticancer treatment were eligible for inclusion in this study. The paraffin-embedded pathologic specimens were collected during surgery and stored in a liquid nitrogen tank until the step of mRNA isolation. All patients received an explanation for the purpose of the study and signed informed consent. The Ethics Committee of Guangxi

Medical University granted approval for this study. For IHC, a commercial biological tissue microarray containing 90 pairs of HCC and adjacent normal liver tissues was constructed by the Biological sample library of Shanghai Outdo Biotech Company, and the survival information of each case was usable. (Microarray: HLivH180Su14).

Quantitative Real-Time Polymerase Chain Reaction (qRT-PCR)

QRT-PCR was performed to evaluate the mRNA expression of selected genes in 20 HCC and their matched precancerous tissues. Total RNA was isolated using Trizol reagent (Life Technologies, Inc., NY, USA) according to the manufacturer's instructions. The concentration and purity of the total RNA were detected using Microplate reader (Biotech Instruments, Inc., VT, USA). RNA reverse transcription was then performed with the PrimeScript™ RT reagent Kit (Takara Biomedical Technology (Beijing) Co., Ltd.) with gDNA Eraser (Perfect Real Time), and qRT-PCR was performed using the TB Green™ Premix Ex Taq™ II (Tli RNaseH Plus) kit (Takara Biomedical Technology (Beijing) Co., Ltd.) protocol in a StepOnePlus system (Applied Biosystems. Life Technologies Holdings Pte Ltd, Singapore).

The sequences of the primers are as follows: *F2*: forward primer, 5'-CTGAGGGTCTGGGTACGAAC-3', reverse primer, 5'-TGGGTAGCGACTCCTCCATAG-3'; *GOT2*: forward primer, 5'-AAGAGTGGCCGGTTTGTTCAC-3', reverse primer, 5'-AGAAAGACATCTCGGCTGAACT-3'; *TRPV1*: forward primer, 5'-TGCACGACGGACAGAACAC-3', reverse primer, 5'-GCGTTGACAAGCTCCTTCAG-3'. The cycle conditions are as follows: after an initial incubation at 95°C for 30 s, the samples were cycled 40 times at 95°C for 5 s and 60°C for 30 s. The relative expression level of each gene in the individual samples was calculated using the $2^{-\Delta\Delta Ct}$ method and normalized using GAPDH as an endogenous control.

Immunohistochemistry (IHC)

EnVision™ FLEX+, Mouse, High pH, (Link) (K8002, Dako) was used for the immunohistochemistry. After the tissue chips were baked and placed in LEICAST5010 (LEICA), PT Link (Dako North America, Inc.) was used for antigen retrieval. Primary antibodies were diluted (*F2*, 1:3000; *GOT2*, 1:80000; *TRPV1*, 1:1500) and incubated overnight at 4°C. The secondary antibody reactions were carried out using the Autostainer Link 48 (Dako North America, Inc.), the sections were subjected to color development with the DAB chromogenic kit, and finally counterstained with Hematoxylin (SLBT4555, Sigma Aldrich). The following antibodies were used: *F2*, 1: Anti-Thrombin (ab83981; Abcam), *GOT2*, 1: Anti-FABP-1 (ab171739; Abcam), *TRPV1*, 1: Anti-VR1 (ab3487; Abcam). All slides were evaluated by two independent pathologists who were blind about the clinicopathologic data.

The expression levels were scored as the staining intensity (0, negative; 1+, weak; 2+, moderate; 3+, strong) multiplied by the proportion of immunopositive staining area (0, < 25%; 1+, 25–50%; 2+, 50–75%; 3+, > 75%) intensity of staining. Expression scores < 5 were considered as “low expression,” and scores ≥ 5 were considered as “high expression.”

Statistics

Statistical analyses were conducted using R 3.3.1 (Auckland, NZ) and IBM SPSS 23.0 (Chicago, USA). McNemar test was used to test the paired 4-fold table experimental data of IHC. The paired *t*-test was used to analyze the qRT-PCR experimental data. Except for single-gene survival analyses and three-gene prognosis survival analyses with *p*-value < 0.01 as statistically significant, other statistical analyses were considered statistically significant with two-sided *p*-value < 0.05.

RESULTS

Selection of Genes Related to Liver Cancer Prognosis and Liver Cancer Samples

We combined 3 corresponding concepts of the key word “liver cancer” [Liver neoplasms (alias Liver Cancer) (disease) (60,666 connections); Liver carcinoma (alias liver cell cancer) (disease) (55,739 connections); Carcinoma, Hepatocellular (alias Adult Liver Cancer) (mesh) (57,034 connections)] with 2 corresponding concepts of the key word “prognosis” [Prognosis (mesh) (77,312 connections); Prognostic Marker (alias Prognosis Marker) (chemical) (22,056 connections)] and 10 corresponding concepts of the key word “outcome” [Fatal Outcome (mesh) (34,016 connections); Outcome Assessment (Health Care) (alias Outcome Study) (mesh) (48,296 connections); Outcome studies (procedure) (9,545 connections); Treatment Outcome (mesh) (77,246 connections); Outcomes research (procedure) (5,540 connections); Outcome monitoring (procedure) (2,030 connections); Patient-focused outcomes (procedure) (3,830 connections); Treatment outcome in HSR (procedure) (998 connections); Patient Reported Outcome Measures (alias Patient Reported Outcome) (mesh) (2,301 connections); Patient Outcome Assessment (mesh) (9,066 connections)], respectively, (**Supplementary Table S1**), and searched for their corresponding genes or proteins in the Coremine database (<http://www.coremine.com/medical/>). With *p*-values < 0.05 as the criteria, a total of 1,173 genes that might be related to the prognosis of liver cancer were finally obtained after screening and elimination of duplicates. As the samples of liver cancer in the Coremine database were not uniform enough, we selected 319 samples for DFS and 370 samples for OS of liver cancer from the TCGA database and obtained the corresponding survival data as well as the expression information of the above 1,173 genes in these samples. This was necessary to carry out the subsequent survival analyses of these genes for liver cancer.

The Single Genes Prognostic Analyses

To clearly describe our process of screening genes, a flowchart of the analysis procedure was developed (**Figure 1**). First, we performed the Kaplan-Meier analysis of each of the 1,173 genes. It was found that the mRNA expression of 276 genes and 283 genes was significantly associated with DFS in 319 patients (*p* < 0.05) and OS in 370 patients (*p* < 0.05), respectively. Additionally, the mRNA expression of 166 of these genes was significantly associated with both DFS and OS (*p* < 0.05).

To further investigate the value of the genes in the prognosis of liver cancer, we chose 135 genes and 149 genes with *p*-values

< 0.01 for DFS and OS, respectively. Next, we used the Cox proportional hazards regression model to employ multivariate analyses on the above genes, respectively to determine the DFS and OS prediction potential of these genes.

The DFS-related multivariate analysis results showed that the expression of 39 genes (*ALDOB*, *APOB*, *AURKB*, *C5*, *CCNF*, *CD4*, *CENPJ*, *CETP*, *COL18A1*, *CPT2*, *DAND5*, *DNASE1*, *EBPL*, *F7*, *FLT3*, *G6PD*, *GNMT*, *ITGB2*, *KLRK1*, *KNG1*, *LMOD1*, *NEK2*, *PCLAF*, *PER1*, *PKM*, *POU2F1*, *PPAT*, *PPIA*, *PRF1*, *PTPN6*, *RUNX3*, *SELP*, *SLCO1B1*, *SPPL2A*, *STAT5A*, *TCF21*, *TRPV1*, *TUSC1*, and *TYMS*) was significantly associated with DFS in HCC patients (*p* < 0.05, **Table 1**). The highly significant results of both the DFS-related single-gene survival analyses for each of these 39 genes and multivariate analysis confirmed that the above 39 genes have a strong association with the DFS of liver cancer, especially the 5-year disease free survival rate of liver cancer.

The OS-related multivariate analysis results showed that the expression of 28 genes (*ABCC1*, *ANXA7*, *APOB*, *ATG7*, *BAK1*, *CA9*, *CCNA2*, *CHD1L*, *CYP3A4*, *E2F1*, *EZH2*, *F2*, *G6PC*, *GMPS*, *GOT2*, *HDAC2*, *HPX*, *KPNA2*, *LAPTM4B*, *MAGEB3*, *MAPT*, *MPV17*, *NTF3*, *PPAT*, *SLC2A1*, *SLC38A1*, *SPP1*, and *TRPV1*) was significantly associated with OS in HCC patients. (*p* < 0.05, **Table 1**). The strongly significant results of both the OS-related single-gene survival analyses and multivariate analysis confirmed that these 28 genes are significantly associated with the OS of liver cancer, especially the 5-year survival rate of liver cancer.

Additionally, among the above-mentioned genes selected after single-gene survival analyses and multivariate analyses, 3 genes (*APOB*, *PPAT*, and *TRPV1*) were significantly associated with both DFS and OS in HCC patients.

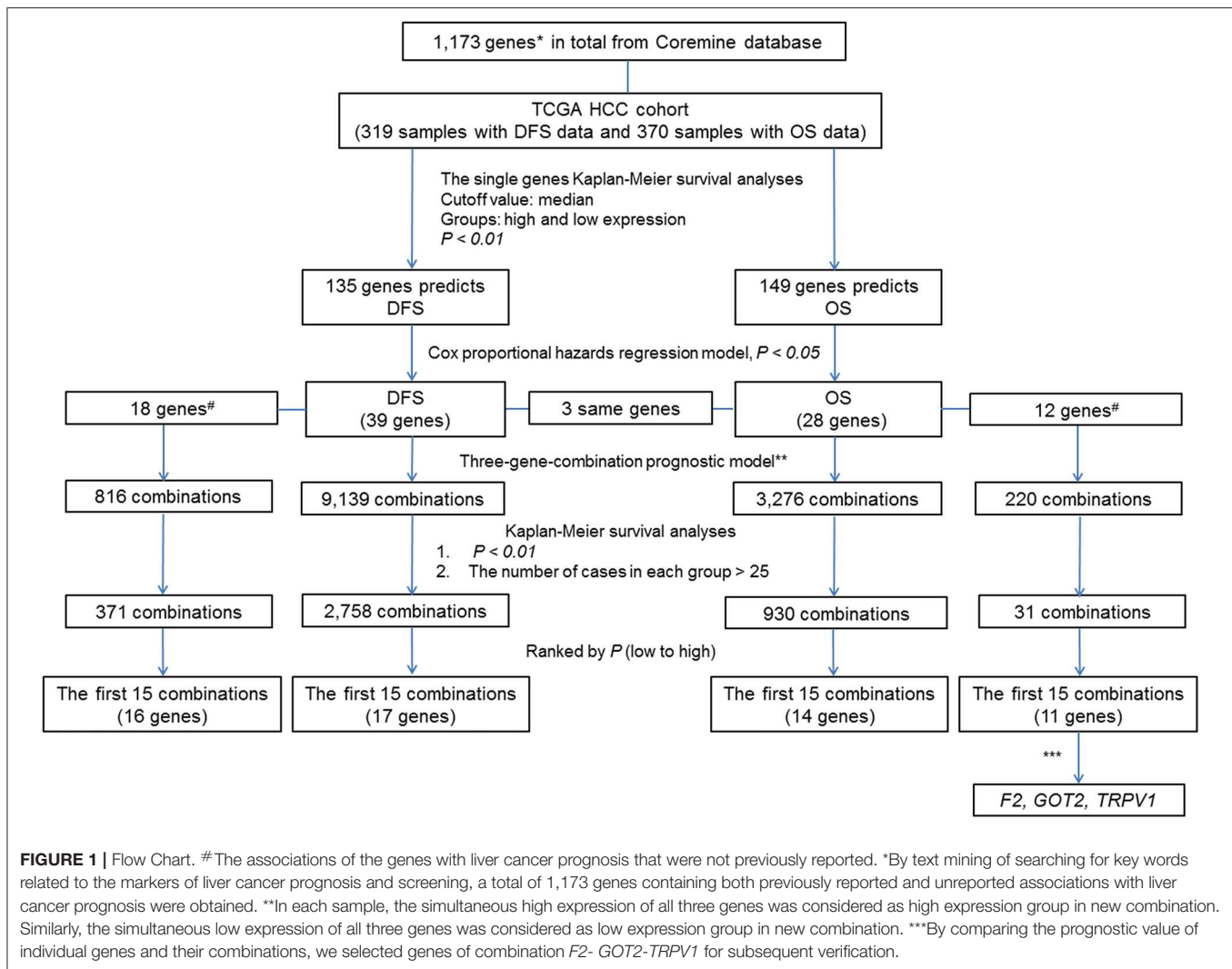
Heat maps of the expression of the above 39 DFS-related genes and 28 OS-related genes in 1173 TCGA liver cancer samples, respectively, which grouped by prognosis status, were shown in **Supplementary Figure S1**.

Three-Gene-Combination Prognostic Model

To reflect the association of the expression of the combined genes with the prognosis of HCC, three-gene-combinations of the above 39 and 28 single genes that are significantly associated with DFS and OS, respectively, were formed, resulting in 9,139 and 3,276 three-gene-combinations for DFS and OS, respectively. In each combination, simultaneous high expression of the three genes in the same case was defined as the co-high expression group. Similarly, simultaneous low expression of the three genes in the same case was considered to be the co-low expression group. In order to ensure the comparability between the high and the low expression group, we deleted combinations which had < 25 cases in the co-high or co-low expression group.

Three-Gene-Combination of Prediction for DFS in Liver Cancer

K-M survival analysis of each of the above 9,139 combinations constituted by 39 DFS-related single genes was first performed. Then, we selected a total of 2,758 combinations with *p*-values < 0.01, excluding the combinations with no more than 25 cases in the co-high expression or co-low expression groups. Apparently,



these selected 2,758 combinations have significant prognostic implications for DFS in liver cancer.

In addition, 18 of the above 39 single genes have not yet been systematically reported to be associated with HCC prognosis, and these 18 genes can combine into 816 three-gene-combinations. The results of the K-M survival analyses showed that 317 combinations had significant association with DFS of liver cancer ($p < 0.01$).

The top 15 combinations of the above 2,758 and 317 combinations with the smallest p -values were chosen. The DFS-related survival analyses diagrams and tables of these combinations and the single genes they contain are as follows (Figures 2, 3; Tables 2, 3).

Three-Gene-Combination of Prediction for OS in Liver Cancer

Similarly, three-gene-combinations of the 28 single genes significantly associated with OS confirmed by the single gene survival analyses and the multivariate analysis were formed, resulting in 3,276 three-gene-combinations. 930 of these 3,276

combinations were screened out on the conditions that the number of cases in both the co-high and co-low expression groups was > 25 , and the p -values were < 0.01 according to the OS-related K-M analyses results.

Furthermore, 12 of the above 28 single genes that were noted to have an unknown association with liver cancer prognosis formed 220 three-gene-combinations. Out of the 220 combinations, there were 31 combinations in which the number of cases in both the co-high and co-low expression groups was > 25 and the OS-related survival analyses results showed $p < 0.01$.

We found 930 of above 3,276 combinations and 31 of above 220 unreported-gene combinations were significant association with OS related survival of liver cancer patients. Among the 930 combinations and 31 combinations mentioned above, the diagrams and tables of the OS-related survival analyses of the top 15 combinations with the smallest p -values and the single genes they contain are as follows (Figures 4, 5; Tables 3, 4) Among the 12 genes that have an unknown association with HCC prognosis, *F2*, *GOT2*, *TRPV1*, and their combination *F2-GOT2-TRPV1* were all significantly associated

TABLE 1 | Multivariate analyses of prognosis of DFS of 319 HCC patients and OS of 370 HCC patients in a TCGA cohort.

Items	Genes	B	SE	Wald	Sig.	Exp (B)	95.0% CI	
							Lower	Upper
DFS associated	<i>ALDOB</i>	−0.580	0.186	9.750	0.002	0.560	0.389	0.806
	<i>APOB</i>	−0.436	0.217	4.023	0.045	0.647	0.423	0.990
	<i>AURKB</i>	0.527	0.211	6.208	0.013	1.694	1.119	2.564
	<i>C5*</i>	−0.420	0.170	6.093	0.014	0.657	0.471	0.917
	<i>CCNF</i>	0.694	0.334	4.310	0.038	2.002	1.040	3.857
	<i>CD4*</i>	−0.774	0.316	6.007	0.014	0.461	0.248	0.856
	<i>CENPJ</i>	1.053	0.243	18.794	0.000	2.867	1.781	4.615
	<i>CETP*</i>	0.829	0.423	3.851	0.050	2.291	1.001	5.245
	<i>COL18A1*</i>	0.417	0.207	4.064	0.044	1.518	1.012	2.278
	<i>CPT2</i>	0.558	0.247	5.114	0.024	1.747	1.077	2.834
	<i>DAND5*</i>	−0.427	0.183	5.466	0.019	0.652	0.456	0.933
	<i>DNASE1*</i>	0.382	0.136	7.927	0.005	1.465	1.123	1.910
	<i>EBPL*</i>	−0.766	0.280	7.463	0.006	0.465	0.268	0.805
	<i>F7*</i>	−0.496	0.175	8.034	0.005	0.609	0.432	0.858
	<i>FLT3*</i>	−0.700	0.240	8.512	0.004	0.497	0.310	0.795
	<i>G6PD</i>	0.477	0.188	6.438	0.011	1.611	1.115	2.328
	<i>GNMT</i>	0.427	0.160	7.118	0.008	1.533	1.120	2.097
	<i>ITGB2*</i>	1.112	0.301	13.662	0.000	3.042	1.686	5.486
	<i>KLRK1</i>	0.932	0.384	5.883	0.015	2.539	1.196	5.390
	<i>KNG1*</i>	0.645	0.277	5.412	0.020	1.906	1.107	3.282
	<i>LMOD1*</i>	−0.873	0.410	4.524	0.033	0.418	0.187	0.934
	<i>NEK2</i>	−0.546	0.263	4.299	0.038	0.579	0.346	0.971
	<i>PCLAF</i>	0.526	0.243	4.700	0.030	1.693	1.052	2.724
	<i>PER1</i>	−0.670	0.221	9.169	0.002	0.512	0.332	0.790
	<i>PKM</i>	−0.645	0.282	5.210	0.022	0.525	0.302	0.913
	<i>POU2F1</i>	0.455	0.142	10.236	0.001	1.577	1.193	2.084
	<i>PPAT*</i>	0.966	0.210	21.121	0.000	2.628	1.741	3.969
	<i>PPIA*</i>	0.626	0.183	11.661	0.001	1.870	1.306	2.679
	<i>PRF1*</i>	−1.676	0.370	20.505	0.000	0.187	0.091	0.386
	<i>PTPN6</i>	−0.610	0.227	7.203	0.007	0.543	0.348	0.848
	<i>RUNX3</i>	0.967	0.375	6.659	0.010	2.629	1.262	5.479
	<i>SELP*</i>	0.790	0.270	8.587	0.003	2.203	1.299	3.736
	<i>SLCO1B1</i>	−0.524	0.213	6.029	0.014	0.592	0.390	0.900
	<i>SPPL2A*</i>	−0.669	0.217	9.528	0.002	0.512	0.335	0.783
	<i>STAT5A</i>	−1.704	0.489	12.149	0.000	0.182	0.070	0.474
	<i>TCF21</i>	−0.979	0.401	5.961	0.015	0.376	0.171	0.824
	<i>TRPV1*</i>	−0.520	0.189	7.604	0.006	0.595	0.411	0.860
	<i>TUSC1</i>	0.423	0.188	5.044	0.025	1.526	1.055	2.207
	<i>TYMS</i>	0.523	0.245	4.558	0.033	1.687	1.044	2.727
OS associated	<i>ABCC1</i>	1.097	0.369	8.841	0.003	2.994	1.453	6.168
	<i>ANXA7*</i>	−0.554	0.201	7.618	0.006	0.575	0.388	0.852
	<i>APOB</i>	−0.791	0.311	6.461	0.011	0.453	0.246	0.834
	<i>ATG7</i>	0.613	0.312	3.876	0.049	1.847	1.003	3.400
	<i>BAK1</i>	−0.490	0.231	4.497	0.034	0.613	0.390	0.964
	<i>CA9</i>	0.761	0.363	4.399	0.036	2.140	1.051	4.356
	<i>CCNA2</i>	0.502	0.203	6.094	0.014	1.652	1.109	2.461
	<i>CHD1L</i>	0.491	0.181	7.377	0.007	1.634	1.147	2.330
	<i>CYP3A4</i>	0.999	0.364	7.539	0.006	2.717	1.331	5.544
	<i>E2F1</i>	0.360	0.172	4.371	0.037	1.433	1.023	2.008

(Continued)

TABLE 1 | Continued

Items	Genes	B	SE	Wald	Sig.	Exp (B)	95.0% CI	
							Lower	Upper
	<i>EZH2</i>	0.985	0.399	6.103	0.013	2.678	1.226	5.852
	<i>F2*</i>	0.711	0.313	5.174	0.023	2.036	1.103	3.757
	<i>G6PC</i>	−0.677	0.341	3.937	0.047	0.508	0.260	0.992
	<i>GMPS</i>	0.733	0.291	6.345	0.012	2.081	1.177	3.681
	<i>GOT2*</i>	−1.509	0.484	9.723	0.002	0.221	0.086	0.571
	<i>HDAC2</i>	0.813	0.316	6.628	0.010	2.255	1.214	4.187
	<i>HPX*</i>	0.930	0.384	5.882	0.015	2.535	1.195	5.378
	<i>KPNA2</i>	0.835	0.284	8.664	0.003	2.305	1.322	4.018
	<i>LAPTM4B</i>	−0.492	0.168	8.616	0.003	0.611	0.440	0.849
	<i>MAGEB3*</i>	0.393	0.179	4.824	0.028	1.482	1.043	2.105
	<i>MAPT*</i>	0.660	0.243	7.349	0.007	1.934	1.201	3.117
	<i>MPV17*</i>	1.141	0.488	5.468	0.019	3.129	1.203	8.141
	<i>NTF3*</i>	1.089	0.357	9.318	0.002	2.973	1.477	5.983
	<i>PPAT*</i>	0.752	0.286	6.897	0.009	2.122	1.210	3.719
	<i>SLC2A1*</i>	−0.921	0.440	4.383	0.036	0.398	0.168	0.943
	<i>SLC38A1*</i>	−0.768	0.289	7.063	0.008	0.464	0.263	0.817
	<i>SPP1</i>	0.604	0.264	5.219	0.022	1.830	1.090	3.073
	<i>TRPV1*</i>	0.453	0.201	5.044	0.025	1.572	1.059	2.334

*The gene has not been systematically reported to be associated with HCC prognosis.

Cox proportional hazard model was used to analyze the impact of 135 genes on DFS and the impact of 149 genes on OS, respectively, $P < 0.05$ were considered to be significant. 39 genes and 28 genes were significantly associated with liver cancer DFS and OS, respectively.

with OS in 370 liver cancer samples from the TCGA data (*F2*: $p = 0.005$; *GOT2*: $p < 0.001$; *TRPV1*: $p = 0.002$; *F2-GOT2-TRPV1*: $p < 0.001$). The overall survival rate in HCC patients with low expression of *F2*, *GOT2*, *TRPV1*, and the three-gene-combination *F2-GOT2-TRPV1* were all significantly lower than that in liver cancer patients with high expression. In addition, the median survival time difference between the high expression group and the low expression group of *F2*, *GOT2*, *TRPV1*, and the three-gene combination *F2-GOT2-TRPV1* was 23.62, 32.26, 35.61, and 55.68 months, respectively. The median survival time difference of this combination was greater than that of a single gene, which was one of the main reasons why we selected these three genes for qRT-PCR and immunohistochemically validation.

Low Expression of *F2*, *GOT2*, and *TRPV1* Predicts Poor Prognosis

Based on the above results of the OS-related survival analyses and multivariate analyses on 28 genes, as well as the results of survival analyses on their three-gene-combinations, we selected three genes *F2*, *GOT2*, and *TRPV1* with strong liver cancer prognostic potential for subsequent validation.

F2, *GOT2*, and *TRPV1* Were Downregulated in HCC Tissues

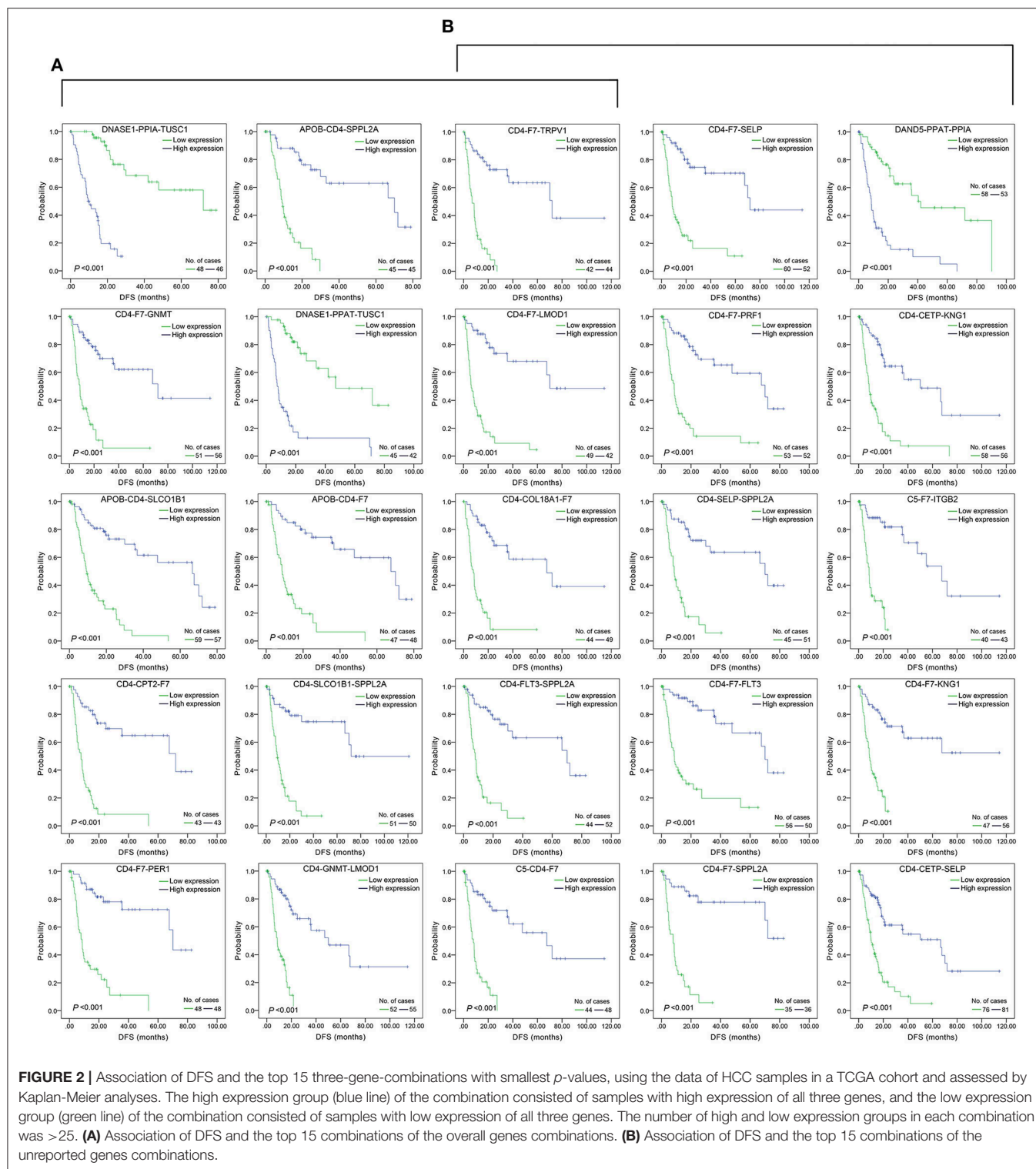
The gene expression in HCC was determined based on three independent microarrays which are all collected in Oncomine database (<https://www.oncomine.org/resource/login.html>). As shown in Roessler Liver 2 Statistics (225 HCC

tissues vs. 220 liver tissues), the expression of *F2*, *GOT2*, and *TRPV1* in HCC tissues were all significantly down-regulated compared with that in normal liver tissues. ($p < 0.001$; **Figure 6**) In addition, based on the Mas Liver Statistics (38 HCC tissue vs. 19 liver tissue), both *F2* and *TRPV1* were significantly down-regulated in HCC tissues. Based on the Chen Liver Statistics (104 HCC tissues vs. 76 liver tissues), both *F2* and *GOT2* were significantly down-regulated in HCC tissues.

The qRT-PCR results of *F2*, *GOT2* and *TRPV1* showed that 20/20, 19/20, and 16/19 of the HCC tissues exhibited significantly lower expression of *F2* ($p < 0.001$; **Figure 7A**), *GOT2* ($p < 0.001$; **Figure 7B**), and *TRPV1* ($p = 0.006$; **Figure 7C**), respectively, when compared with their corresponding non-tumorous tissues.

The protein expression of *F2*, *GOT2*, and *TRPV1* in HCC tissues was evaluated using IHC. Positive staining of *F2*, *GOT2*, and *TRPV1* was mainly localized in the cytoplasm of HCC cells. The representative staining of *F2*, *GOT2*, and *TRPV1* negative and positive protein expression in HCC are shown in **Figure 8A**.

Among 90 HCC tissues and adjacent non-malignant liver tissues, IHC was employed to measure the protein expression of *F2*, *GOT2*, and *TRPV1*, respectively. Low *F2* expression was observed in 62/89 (69.66%) of the HCC tissues, compared to 33/89 (37.08%) in adjacent normal liver tissues ($p < 0.001$); low *GOT2* expression was noted in 72/89 (80.90%) of the HCC tissues, compared to 32/89 (35.96%) in adjacent normal liver tissues ($p < 0.001$); low *TRPV1* expression was also observed in 59/89 (66.29%) of the HCC tissues, compared to 38/89 (42.70%) in adjacent normal liver tissues ($p = 0.002$).



Expression of *F2*, *GOT2*, and *TRPV1* and Their Combination *F2-GOT2-TRPV1* With OS

Based on the above results of single-genes and three-gene combinations survival analyses of TCGA HCC samples, the low expression of *F2*, *GOT2*, *TRPV1* and their

combination *F2-GOT2-TRPV1* was significantly associated with poor OS in HCC. (*F2*: $p = 0.005$; *GOT2*: $p < 0.001$; *TRPV1*: $p = 0.002$; *F2-GOT2-TRPV1*: $p < 0.001$). In addition, the median survival time difference between the high expression group and the low expression group of

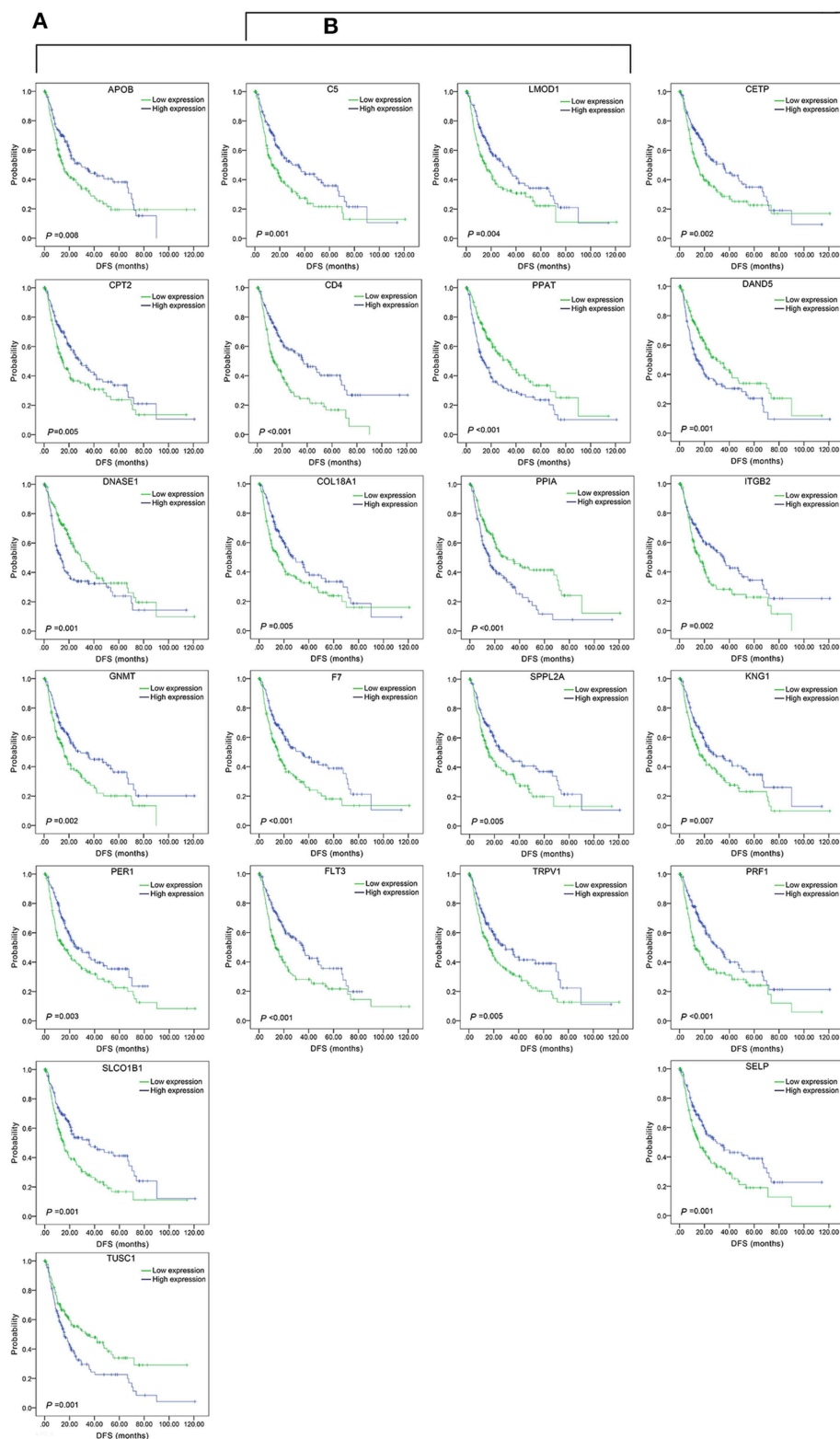


FIGURE 3 | Association of DFS and the individual genes contained in the top 15 combinations with the lowest P -values, using the data of HCC samples in a TCGA cohort and assessed by Kaplan-Meier analyses. **(A)** Association of DFS and the 17 single genes contained in the first 15 total-gene combinations. **(B)** Association of DFS and the 16 single genes contained in the first 15 unreported-gene combinations.

TABLE 2 | The associations of three-gene combinations with disease-free survival (DFS) of HCC patients in a TCGA cohort, analyzed by Kaplan-Meier method.

DFS (Median) of combinations of 39 genes with HCC prognosis								DFS (Median) of combinations of 18 genes have unknown association with HCC prognosis							
		Estimate	Std. Error	95% confidence interval		P	Median survival time difference (H-L)			Estimate	Std. Error	95% confidence interval		P	Median survival time difference (H-L)
				Lower boundary	Upper boundary							Lower boundary	Upper boundary		
<i>DNASE1-PP1A-TUSC1</i>	H	9.490	1.597	6.360	12.620	0.000	−62.420	<i>CD4-F7-TRPV1</i>	H	71.910	20.619	31.498	112.322	0.000	65.010
	L	71.910	24.365	24.154	119.666				L	6.900	1.657	3.652	10.148		
	Overall	21.620	4.848	12.119	31.121				Overall	15.740	5.309	5.334	26.146		
<i>CD4-F7-TRPV1</i>	H	71.910	20.619	31.498	112.322	0.000	65.010	<i>CD4-F7-LMOD1</i>	H	70.070	–	–	–	0.000	63.830
	L	6.900	1.657	3.652	10.148				L	6.240	1.408	3.480	9.000		
	Overall	15.740	5.309	5.334	26.146				Overall	17.640	3.833	10.127	25.153		
<i>CD4-F7-GNMT</i>	H	71.910	22.303	28.196	115.624	0.000	63.370	<i>CD4-COL18A1-F7</i>	H	67.580	21.110	26.205	108.955	0.000	59.660
	L	8.540	1.241	6.108	10.972				L	7.920	1.658	4.670	11.170		
	Overall	21.160	4.039	13.244	29.076				Overall	19.190	3.616	12.104	26.276		
<i>CD4-F7-LMOD1</i>	H	70.070	–	–	–	0.000	63.830	<i>CD4-FLT3-SPPL2A</i>	H	70.070	18.005	34.779	105.361	0.000	62.220
	L	6.240	1.408	3.480	9.000				L	7.850	1.486	4.937	10.763		
	Overall	17.640	3.833	10.127	25.153				Overall	19.650	7.275	5.391	33.909		
<i>CD4-COL18A1-F7</i>	H	67.580	21.110	26.205	108.955	0.000	59.660	<i>C5-CD4-F7</i>	H	67.580	15.374	37.447	97.713	0.000	59.660
	L	7.920	1.658	4.670	11.170				L	7.920	1.414	5.149	10.691		
	Overall	19.190	3.616	12.104	26.276				Overall	21.160	5.704	9.981	32.339		
<i>APOB-CD4-SLCO1B1</i>	H	66.620	13.239	40.672	92.568	0.000	57.130	<i>CD4-F7-SELP</i>	H	71.910	3.184	65.669	78.151	0.000	63.200
	L	9.490	0.918	7.691	11.289				L	8.710	0.783	7.176	10.244		
	Overall	19.650	4.976	9.897	29.403				Overall	21.550	8.496	4.898	38.202		
<i>CD4-CPT2-F7</i>	H	71.910	21.206	30.347	113.473	0.000	64.060	<i>CD4-F7-PRF1</i>	H	70.070	15.899	38.908	101.232	0.000	61.500
	L	7.850	2.024	3.883	11.817				L	8.570	1.055	6.502	10.638		
	Overall	15.700	2.776	10.259	21.141				Overall	21.160	3.455	14.389	27.931		
<i>CD4-F7-PER1</i>	H	70.070	2.855	64.475	75.665	0.000	61.500	<i>CD4-SELP-SPPL2A</i>	H	70.070	3.849	62.525	77.615	0.000	61.500
	L	8.570	1.225	6.170	10.970				L	8.570	0.819	6.964	10.176		
	Overall	25.300	8.227	9.175	41.425				Overall	18.590	5.837	7.149	30.031		
<i>APOB-CD4-SPPL2A</i>	H	70.070	23.928	23.171	116.969	0.000	60.940	<i>CD4-F7-FLT3</i>	H	70.070	3.048	64.097	76.043	0.000	61.360
	L	9.130	0.855	7.455	10.805				L	8.710	1.311	6.140	11.280		

(Continued)

TABLE 2 | Continued

DFS (Median) of combinations of 39 genes with HCC prognosis								DFS (Median) of combinations of 18 genes have unknown association with HCC prognosis							
		Estimate	Std. Error	95% confidence interval		P	Median survival time difference (H-L)			Estimate	Std. Error	95% confidence interval		P	Median survival time difference (H-L)
				Lower boundary	Upper boundary							Lower boundary	Upper boundary		
<i>CD4-FLT3-SPPL2A</i>	Overall	19.190	4.689	10.000	28.380			<i>CD4-F7-SPPL2A</i>	Overall	35.580	12.142	11.781	59.379		
	H	70.070	18.005	34.779	105.361	0.000	62.220		H	–	–	–	–	0.000	–
	L	7.850	1.486	4.937	10.763				L	7.920	1.864	4.266	11.574		
<i>DNASE1-PPAT-TUSC1</i>	Overall	19.650	7.275	5.391	33.909			<i>DAND5-PPAT-PPIA</i>	Overall	24.770	19.276	0.000	62.551		
	H	7.420	1.115	5.235	9.605	0.000	–39.620		H	8.540	0.797	6.978	10.102	0.000	–33.480
	L	47.040	17.350	13.035	81.045				L	42.020	15.014	12.592	71.448		
<i>APOB-CD4-F7</i>	Overall	21.160	5.101	11.162	31.158			<i>CD4-CETP-KNG1</i>	Overall	19.250	2.763	13.834	24.666		
	H	67.580	13.500	41.120	94.040	0.000	58.840		H	50.030	14.498	21.614	78.446	0.000	41.550
	L	8.740	0.884	7.007	10.473				L	8.480	0.769	6.972	9.988		
<i>CD4-SLCO1B1-SPPL2A</i>	Overall	24.770	9.057	7.018	42.522			<i>C5-F7-ITGB2</i>	Overall	18.330	1.894	14.617	22.043		
	H	71.910	–	–	–	0.000	62.420		H	67.580	14.028	40.084	95.076	0.000	59.010
	L	9.490	1.171	7.194	11.786				L	8.570	1.316	5.991	11.149		
<i>C5-CD4-F7</i>	Overall	19.650	6.519	6.873	32.427			<i>CD4-F7-KNG1</i>	Overall	35.580	9.185	17.577	53.583		
	H	67.580	15.374	37.447	97.713	0.000	59.660		H	–	–	–	–	0.000	–
	L	7.920	1.414	5.149	10.691				L	8.740	1.206	6.376	11.104		
<i>CD4-GNMT-LMOD1</i>	Overall	21.160	5.704	9.981	32.339			<i>CD4-CETP-SELP</i>	Overall	21.550	8.293	5.295	37.805		
	H	50.030	16.348	17.987	82.073	0.000	41.550		H	66.620	14.883	37.450	95.790	0.000	56.370
	L	8.480	1.430	5.677	11.283				L	10.250	1.315	7.672	12.828		
	Overall	18.330	2.734	12.971	23.689				Overall	18.330	1.469	15.452	21.208		

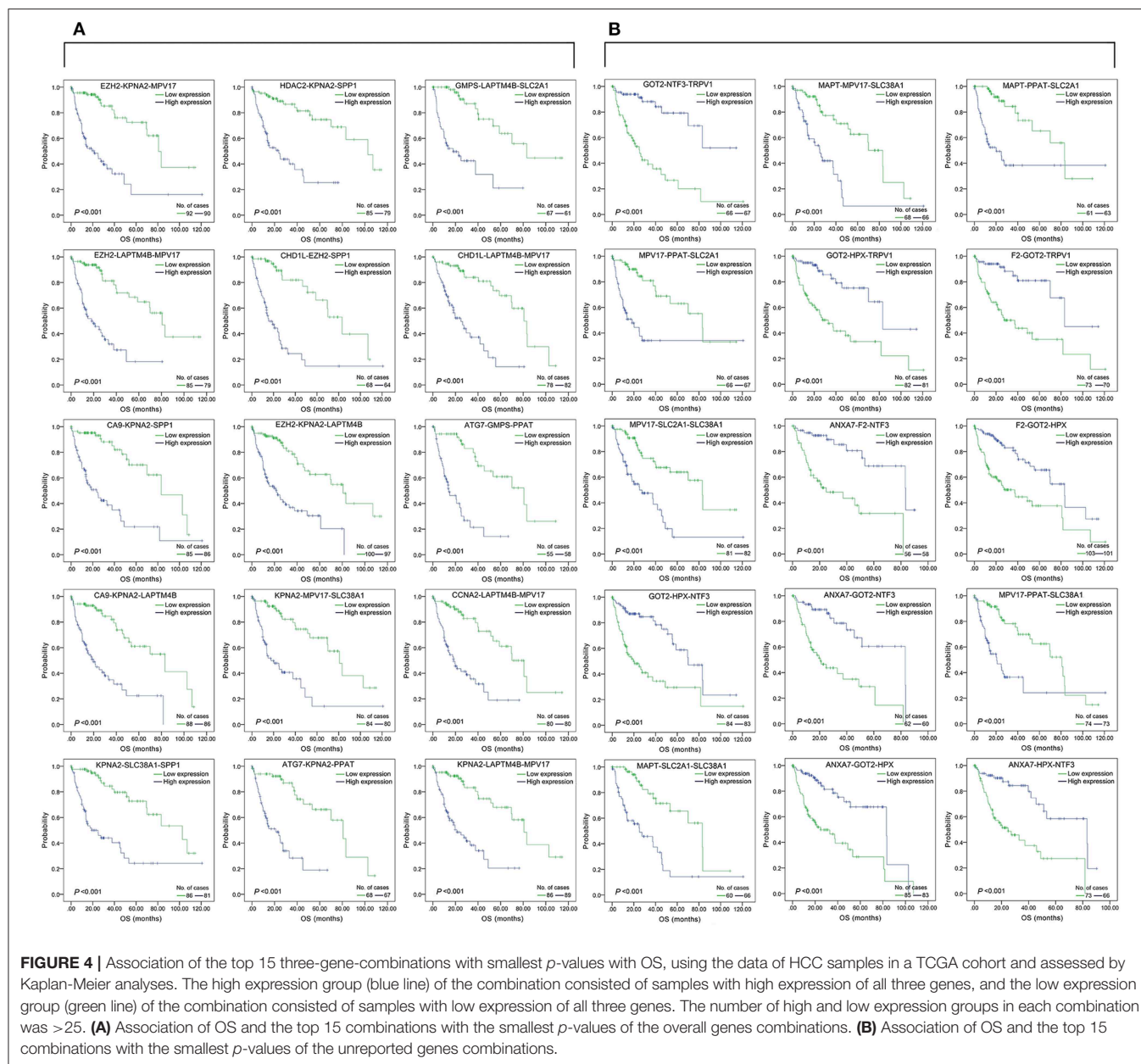
TABLE 3 | The associations of single genes contained in the multi-gene combinations with disease-free survival (DFS) and overall survival (OS) of HCC patients in a TCGA cohort, analyzed by Kaplan-Meier method.

DFS (Median) of single genes of the combinations with HCC prognosis								OS (Median) of single genes of the combinations with HCC prognosis							
		Estimate	Std. Error	95% confidence interval		P	Median survival time difference (H-L)			Estimate	Std. Error	95% Confidence Interval		P	Median survival time difference (H-L)
				Lower boundary	Upper boundary							Lower boundary	Upper boundary		
APOB	H	29.300	6.376	16.802	41.798	0.008	14.450	ANXA7	H	83.180	15.496	52.807	113.553	0.006	36.430
	L	14.850	2.049	10.834	18.866				L	46.750	7.280	32.481	61.019		
	Overall	20.930	2.318	16.387	25.473				Overall	55.650	7.925	40.116	71.184		
C5	H	29.960	6.762	16.706	43.214	0.001	16.330	ATG7	H	45.070	8.031	29.330	60.810	0.009	-35.610
	L	13.630	2.870	8.006	19.254				L	80.680	10.533	60.036	101.324		
	Overall	20.930	2.318	16.387	25.473				Overall	55.650	7.925	40.116	71.184		
CD4	H	36.700	7.693	21.622	51.778	0.000	23.070	CA9	H	37.290	8.317	20.989	53.591	0.000	-32.720
	L	13.630	2.089	9.536	17.724				L	70.010	10.210	49.999	90.021		
	Overall	20.930	2.318	16.387	25.473				Overall	55.650	7.925	40.116	71.184		
CETP	H	35.580	5.896	24.023	47.137	0.002	21.450	CCNA2	H	45.070	10.298	24.885	65.255	0.001	-24.940
	L	14.130	1.799	10.605	17.655				L	70.010	11.730	47.019	93.001		
	Overall	20.930	2.318	16.387	25.473				Overall	55.650	7.925	40.116	71.184		
COL18A1	H	27.200	4.885	17.625	36.775	0.005	11.600	CHD1L	H	39.750	6.940	26.148	53.352	0.006	-40.930
	L	15.600	3.114	9.497	21.703				L	80.680	6.587	67.770	93.590		
	Overall	20.930	2.318	16.387	25.473				Overall	55.650	7.925	40.116	71.184		
CPT2	H	29.300	4.767	19.956	38.644	0.005	14.350	EZH2	H	37.290	10.181	17.335	57.245	0.000	-43.390
	L	14.950	1.836	11.352	18.548				L	80.680	10.816	59.480	101.880		
	Overall	20.930	2.318	16.387	25.473				Overall	55.650	7.925	40.116	71.184		
DAND5	H	13.630	2.561	8.610	18.650	0.001	-16.330	F2	H	69.510	11.842	46.300	92.720	0.005	23.620
	L	29.960	5.455	19.269	40.651				L	45.890	7.020	32.132	59.648		
	Overall	20.930	2.318	16.387	25.473				Overall	55.650	7.925	40.116	71.184		
DNASE1	H	13.140	1.997	9.226	17.054	0.001	-16.160	GMPS	H	45.070	9.667	26.123	64.017	0.003	-24.440
	L	29.300	4.256	20.958	37.642				L	69.510	10.308	49.306	89.714		
	Overall	20.930	2.318	16.387	25.473				Overall	55.650	7.925	40.116	71.184		
F7	H	33.900	8.191	17.846	49.954	0.000	18.490	GOT2	H	70.010	12.025	46.441	93.579	0.000	32.260
	L	15.410	1.485	12.500	18.320				L	37.750	9.383	19.360	56.140		
	Overall	20.930	2.318	16.387	25.473				Overall	55.650	7.925	40.116	71.184		
FLT3	H	35.580	3.640	28.446	42.714	0.000	22.440	HPX	H	69.510	10.518	48.894	90.126	0.002	23.620
	L	13.140	1.833	9.547	16.733				L	45.890	10.112	26.070	65.710		
	Overall	20.930	2.318	16.387	25.473				Overall	55.650	7.925	40.116	71.184		
GNMT	H	29.300	9.167	11.334	47.266	0.002	13.370	HDAC2	H	45.070	8.365	28.675	61.465	0.002	-35.610
	L	15.930	1.821	12.360	19.500				L	80.680	12.796	55.599	105.761		
	Overall	20.930	2.318	16.387	25.473				Overall	55.650	7.925	40.116	71.184		

(Continued)

TABLE 3 | Continued

DFS (Median) of single genes of the combinations with HCC prognosis								OS (Median) of single genes of the combinations with HCC prognosis							
		Estimate	Std. Error	95% confidence interval		P	Median survival time difference (H-L)			Estimate	Std. Error	95% Confidence Interval		P	Median survival time difference (H-L)
				Lower boundary	Upper boundary							Lower boundary	Upper boundary		
ITGB2	H	35.580	4.232	27.285	43.875	0.002	19.840	KPNA2	H	33.020	8.165	17.017	49.023	0.000	−47.660
	L	15.740	2.671	10.504	20.976				L	80.680	6.908	67.139	94.221		
	Overall	20.930	2.318	16.387	25.473				Overall	55.650	7.925	40.116	71.184		
KNG1	H	25.300	6.478	12.603	37.997	0.007	9.600	LAPTM4B	H	45.070	10.511	24.468	65.672	0.000	−35.610
	L	15.700	2.458	10.882	20.518				L	80.680	12.598	55.988	105.372		
	Overall	20.930	2.318	16.387	25.473				Overall	55.650	7.925	40.116	71.184		
LMOD1	H	29.660	5.120	19.625	39.695	0.004	13.960	MAPT	H	41.750	6.888	28.249	55.251	0.006	−28.260
	L	15.700	2.655	10.497	20.903				L	70.010	9.844	50.716	89.304		
	Overall	20.930	2.318	16.387	25.473				Overall	55.650	7.925	40.116	71.184		
PER1	H	25.490	6.529	12.694	38.286	0.003	10.080	MPV17	H	37.290	6.644	24.268	50.312	0.000	−43.390
	L	15.410	3.485	8.579	22.241				L	80.680	6.504	67.933	93.427		
	Overall	20.930	2.318	16.387	25.473				Overall	55.650	7.925	40.116	71.184		
PPAT	H	14.130	2.656	8.924	19.336	0.000	−19.770	NTF3	H	70.010	12.704	45.110	94.910	0.002	29.640
	L	33.900	5.401	23.314	44.486				L	40.370	8.143	24.409	56.331		
	Overall	20.930	2.318	16.387	25.473				Overall	55.650	7.925	40.116	71.184		
PPIA	H	15.600	1.475	12.709	18.491	0.000	−13.280	PPAT	H	58.840	14.928	29.580	88.100	0.009	−10.670
	L	28.880	7.575	14.033	43.727				L	69.510	11.354	47.256	91.764		
	Overall	20.930	2.318	16.387	25.473				Overall	55.650	7.925	40.116	71.184		
PRF1	H	29.960	4.358	21.418	38.502	0.000	17.350	SLC2A1	H	45.890	6.187	33.763	58.017	0.000	−37.290
	L	12.610	2.004	8.681	16.539				L	83.180	17.113	49.638	116.722		
	Overall	20.930	2.318	16.387	25.473				Overall	55.650	7.925	40.116	71.184		
SELP	H	29.960	6.294	17.624	42.296	0.001	14.260	SLC38A1	H	45.070	3.919	37.389	52.751	0.001	−35.610
	L	15.700	2.465	10.868	20.532				L	80.680	7.141	66.684	94.676		
	Overall	20.930	2.318	16.387	25.473				Overall	55.650	7.925	40.116	71.184		
SLCO1B1	H	35.840	10.368	15.518	56.162	0.000	20.890	SPP1	H	40.370	5.288	30.005	50.735	0.000	−29.640
	L	14.950	1.359	12.286	17.614				L	70.010	13.016	44.498	95.522		
	Overall	20.930	2.318	16.387	25.473				Overall	55.650	7.925	40.116	71.184		
SPPL2A	H	27.200	5.000	17.399	37.001	0.005	11.790	TRPV1	H	80.680	7.672	65.642	95.718	0.002	35.610
	L	15.410	2.331	10.842	19.978				L	45.070	6.030	33.250	56.890		
	Overall	20.930	2.318	16.387	25.473				Overall	55.650	7.925	40.116	71.184		
TRPV1	H	29.660	6.127	17.652	41.668	0.005	13.530								
	L	16.130	1.962	12.284	19.976										
	Overall	20.930	2.318	16.387	25.473										
TUSC1	H	15.740	2.003	11.814	19.666	0.001	−18.160								
	L	33.900	8.193	17.841	49.959										
	Overall	20.930	2.318	16.387	25.473										



F2-GOT2-TRPV1 was greater than that of any of the three single genes.

The results of IHC for 90 liver cancer cases showed that the low protein expression of *F2*, *GOT2*, and *TRPV1* was significantly associated with lower 5-year survival in HCC patients (*F2*: $p = 0.033$, *GOT2*: $p = 0.035$, *TRPV1*: $p = 0.046$; K-M survival analyses). However, due to the insufficient number of events in the co-high expression group of the combination *F2-GOT2-TRPV1*, there was marginally significant difference found in the overall survival rate of HCC patients between the co-high expression group and the co-low expression group of the protein combination *F2-GOT2-TRPV1* ($p = 0.051$) (Figure 8B).

DISCUSSION

Liver cancer is characterized by inconspicuous early symptoms, a high degree of malignancy, recurrence and spread, and unsatisfactory prognosis. With limited treatment options, it is one of the common malignancies that plague the world. Therefore, identification of effective prognostic biomarkers for liver cancer is the key to improving the efficacy of targeted therapy for HCC and reducing the adverse prognostic effects of liver cancer.

In our study, by combining and searching 15 corresponding concepts of the key words “liver cancer,” “prognosis,” and “outcome,” and according to p -values < 0.05 , 1,173 genes that

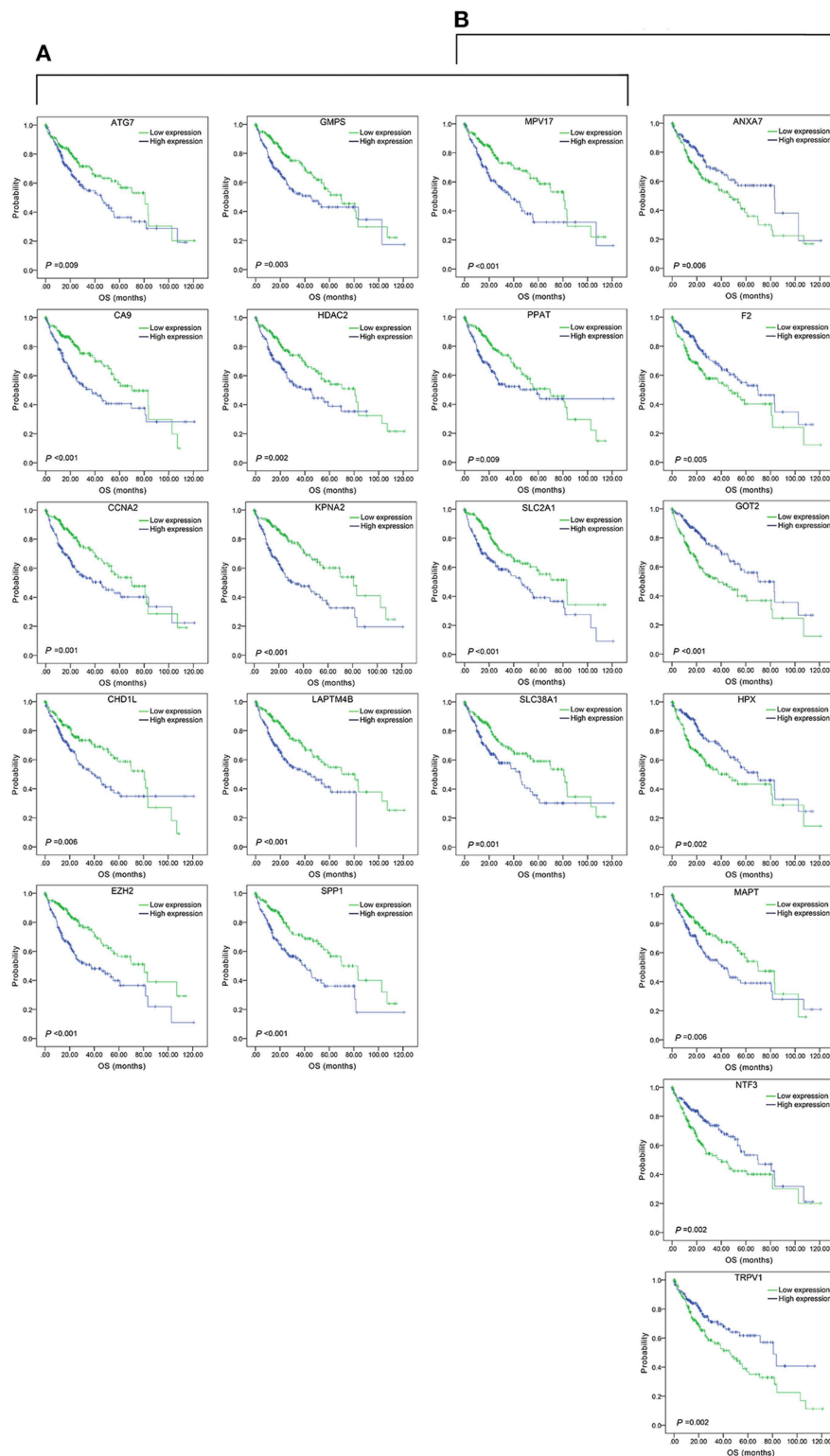


FIGURE 5 | Association of OS and the individual genes contained in the top 15 combinations with the lowest P -values, using the data of HCC samples in a TCGA cohort and assessed by Kaplan-Meier analyses. **(A)** Association of OS and the 14 single genes contained in the first 15 total-gene combinations. **(B)** Association of OS and the 11 single genes contained in the first 15 unreported-gene combinations.

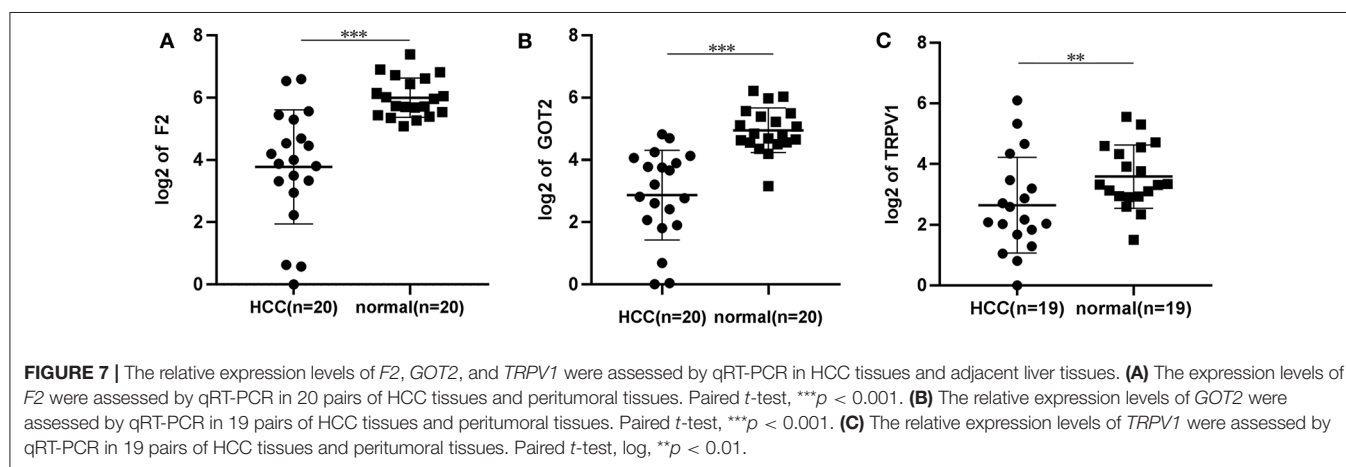
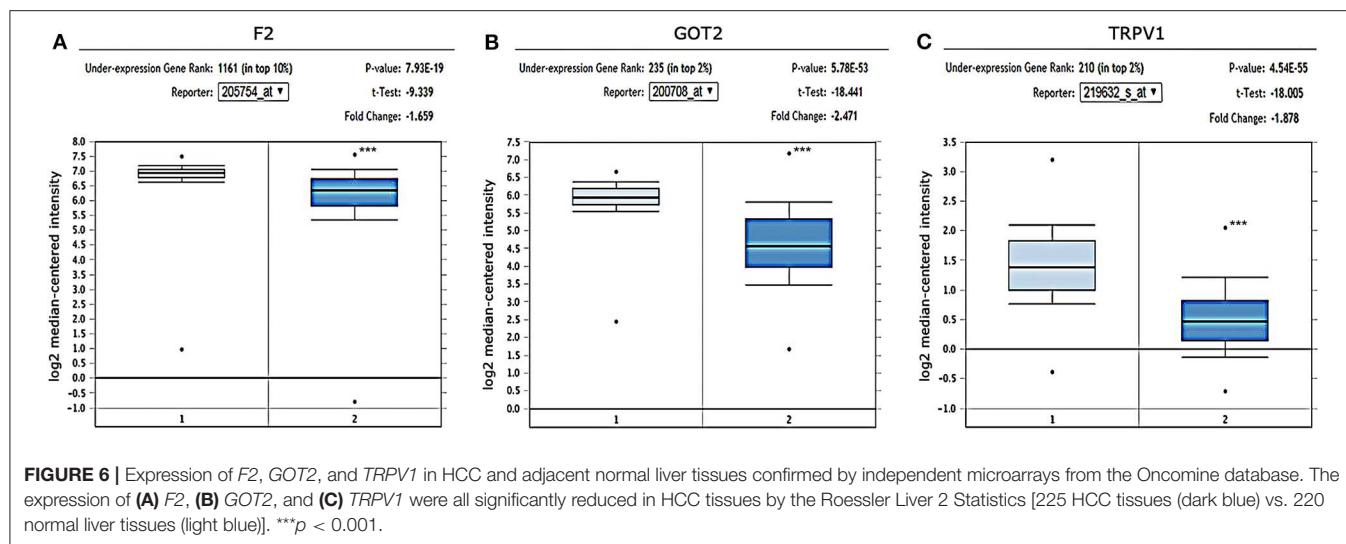
TABLE 4 | The associations of three-gene combinations with overall survival (OS) of HCC patients in a TCGA cohort, analyzed by Kaplan-Meier method.

OS (Median) of combinations of 28 genes with HCC prognosis								OS (Median) of combinations of 12 genes have unknown association with HCC prognosis							
		Estimate	Std. Error	95% confidence interval		P	Median survival time difference (H-L)			Estimate	Std. Error	95% confidence interval		P	Median survival time difference (H-L)
				Lower boundary	Upper boundary							Lower boundary	Upper boundary		
EZH2-KPNA2-MPV17	H	21.320	6.143	9.280	33.360	0.000	−59.360	GOT2-NTF3-TRPV1	H	—	—	—	—	0.000	—
	L	80.680	7.061	66.841	94.519				L	25.230	3.764	17.852	32.608		
	Overall	55.350	13.443	29.001	81.699				Overall	60.840	15.622	30.220	91.460		
EZH2-LAPTM4B-MPV17	H	18.230	5.735	6.988	29.472	0.000	−62.450	MPV17-PPAT-SLC2A1	H	18.330	4.916	8.695	27.965	0.000	−64.850
	L	80.680	7.990	65.020	96.340				L	83.180	15.794	52.224	114.136		
	Overall	48.950	10.014	29.323	68.577				Overall	53.350	15.422	23.123	83.577		
CA9-KPNA2-SPP1	H	23.780	5.368	13.259	34.301	0.000	−59.400	MPV17-SLC2A1-SLC38A1	H	25.130	9.272	6.957	43.303	0.000	−58.050
	L	83.180	16.292	51.248	115.112				L	83.180	7.322	68.829	97.531		
	Overall	51.250	11.668	28.381	74.119				Overall	46.750	6.571	33.870	59.630		
CA9-KPNA2-LAPTM4B	H	19.740	3.699	12.490	26.990	0.000	−63.440	GOT2-HPX-NTF3	H	70.010	10.631	49.174	90.846	0.000	50.430
	L	83.180	20.669	42.669	123.691				L	19.580	6.243	7.343	31.817		
	Overall	46.750	6.141	34.715	58.785				Overall	55.350	6.783	42.055	68.645		
KPNA2-SLC38A1-SPP1	H	19.090	6.876	5.614	32.566	0.000	−83.570	MAPT-SLC2A1-SLC38A1	H	25.130	9.134	7.227	43.033	0.000	−58.050
	L	102.660	21.958	59.622	145.698				L	83.180	12.085	59.493	106.867		
	Overall	69.510	10.951	48.047	90.973				Overall	45.890	7.002	32.167	59.613		
HDAC2-KPNA2-SPP1	H	23.780	5.613	12.778	34.782	0.000	−78.880	MAPT-MPV17-SLC38A1	H	25.130	4.047	17.197	33.063	0.000	−58.050
	L	102.660	14.189	74.850	130.470				L	83.180	9.720	64.128	102.232		
	Overall	83.180	21.672	40.704	125.656				Overall	45.070	6.284	32.754	57.386		
CHD1L-EZH2-SPP1	H	15.410	4.012	7.547	23.273	0.000	−67.770	GOT2-HPX-TRPV1	H	83.180	11.770	60.111	106.249	0.000	50.160
	L	83.180	9.866	63.842	102.518				L	33.020	6.971	19.356	46.684		
	Overall	46.750	12.305	22.633	70.867				Overall	70.010	14.673	41.251	98.769		
EZH2-KPNA2-LAPTM4B	H	21.680	5.445	11.008	32.352	0.000	−59.000	ANXA7-F2-NTF3	H	83.510	15.702	52.734	114.286	0.000	58.640

(Continued)

TABLE 4 | Continued

OS (Median) of combinations of 28 genes with HCC prognosis								OS (Median) of combinations of 12 genes have unknown association with HCC prognosis							
		Estimate	Std. Error	95% confidence interval		P	Median survival time difference (H-L)			Estimate	Std. Error	95% confidence interval		P	Median survival time difference (H-L)
				Lower boundary	Upper boundary							Lower boundary	Upper boundary		
KPNA2-MPV17-SLC38A1	L	80.680	7.011	66.939	94.421	0.000	−63.100	ANXA7-GOT2-NTF3	L	24.870	10.561	4.170	45.570	0.000	62.580
	Overall	46.750	10.389	26.388	67.112				Overall	53.350	14.230	25.459	81.241		
	H	17.580	5.820	6.172	28.988				H	83.180	23.271	37.569	128.791		
ATG7-KPNA2-PPAT	L	80.680	7.992	65.015	96.345	0.000	−59.560	ANXA7-GOT2-HPX	L	20.600	5.417	9.983	31.217	0.000	58.310
	Overall	53.350	11.888	30.049	76.651				Overall	48.950	7.670	33.916	63.984		
	H	21.120	6.087	9.190	33.050				H	83.180	13.677	56.373	109.987		
GMPS-LAPTM4B-SLC2A1	L	80.680	10.953	59.212	102.148	0.000	−65.210	MAPT-PPAT-SLC2A1	L	24.870	8.505	8.200	41.540	0.000	−63.070
	Overall	45.530	11.839	22.325	68.735				Overall	53.290	13.890	26.066	80.514		
	H	17.970	6.680	4.876	31.064				H	20.110	6.433	7.501	32.719		
CHD1L-LAPTM4B-MPV17	L	83.180	16.018	51.784	114.576	0.000	−55.810	F2-GOT2-TRPV1	L	83.180	14.580	54.602	111.758	0.000	55.680
	Overall	53.350	13.670	26.557	80.143				Overall	70.010	18.751	33.258	106.762		
	H	24.870	4.882	15.302	34.438				H	83.180	11.976	59.707	106.653		
ATG7-GMPS-PPAT	L	80.680	7.912	65.172	96.188	0.000	−66.720	F2-GOT2-HPX	L	27.500	6.805	14.162	40.838	0.000	45.890
	Overall	55.650	10.709	34.660	76.640				Overall	81.670	20.419	41.649	121.691		
	H	13.960	4.451	5.236	22.684				H	83.180	6.650	70.146	96.214		
CCNA2-LAPTM4B-MPV17	L	80.680	17.665	46.057	115.303	0.000	−51.680	MPV17-PPAT-SLC38A1	L	37.290	7.225	23.129	51.451	0.000	−60.080
	Overall	37.680	8.510	21.001	54.359				Overall	69.510	12.170	45.657	93.363		
	H	18.330	3.559	11.354	25.306				H	20.600	5.930	8.977	32.223		
KPNA2-LAPTM4B-MPV17	L	70.010	6.190	57.878	82.142	0.000	−59.360	ANXA7-HPX-NTF3	L	80.680	9.365	62.324	99.036	0.000	58.310
	Overall	48.950	7.272	34.697	63.203				Overall	51.250	13.888	24.030	78.470		
	H	21.320	5.082	11.359	31.281				H	83.180	26.573	31.096	135.264		
	L	80.680	7.900	65.196	96.164				L	24.870	7.244	10.672	39.068		
	Overall	51.250	14.898	22.050	80.450				Overall	48.950	5.919	37.350	60.550		



may be related to the prognosis of liver cancer were mined from the Coremine platform after merging and removing duplicates. However, due to the insufficient sample size and data related to the prognosis of liver cancer in the Coremine platform as well as the large heterogeneity among the samples, we also selected gene expression data and prognosis data of 319 samples for DFS and 370 samples for OS from the TCGA platform. We then separately conducted DFS-related and OS-related K-M survival analysis for each gene, followed by multivariate analyses, respectively. The large-scale genes mining and a large number of homogenous samples gave us a reliable analytical foundation. By far, this is the first large-scale survival analyses for hundreds of genes for subsequent screening.

In addition, the genes selected by K-M survival analyses with a low p -value ($p < 0.01$) were further screened by multivariate analyses using the Cox proportional hazards regression model. We found that 39 genes and 28 genes were reliably and significantly associated with DFS and OS, respectively, in liver cancer. Many of the above genes have been confirmed to be associated with the prognosis of HCC by previous reports. For

example, of the 39 DFS-related genes, *ALDOB* inhibits metastasis in HCC and can be a valuable novel prognosis predicting marker (30); *APOB* was found to be a prognostic biomarker for patients with radical resection of HCC (31, 32); *CCNF* is downregulated in HCC and is a promising prognostic marker (33). In addition, *CPT2* (34), *G6PD* (35), *GNMT* (36), *NEK2* (37), etc. have also been reported to be prognostic markers of HCC by affecting the occurrence or invasion of HCC. The above findings are consistent with what we identified. Other genes, such as *C5*, *CD4*, *CETP*, *COL18A1*, *DAND5*, *DNASE1*, *EBPL*, *F7*, *FLT3*, *ITGB2*, *KNG1*, *LMOD1*, *PPAT*, *PPIA*, *PRF1*, *SELP*, *SPPL2A*, and *TRPV1* that have not been systematically reported in relation to the prognosis of liver cancer, are our newly discovered prognostic markers for DFS in liver cancer. Similarly, of the 28 OS-related genes, *CA9* regulates the epithelial-mesenchymal transition and is a novel prognostic marker in HCC (38), *E2F1* expression has an impact on tumor aggressiveness and affects the prognosis of HCC (14, 15), *CYP3A4* (39), *HDAC2* (40), and *KPNA2* (41) have also been identified as prognostic markers of HCC and are reflected in our findings. The other genes, such as *ANXA7*, *F2*, *GOT2*, *HPX*,

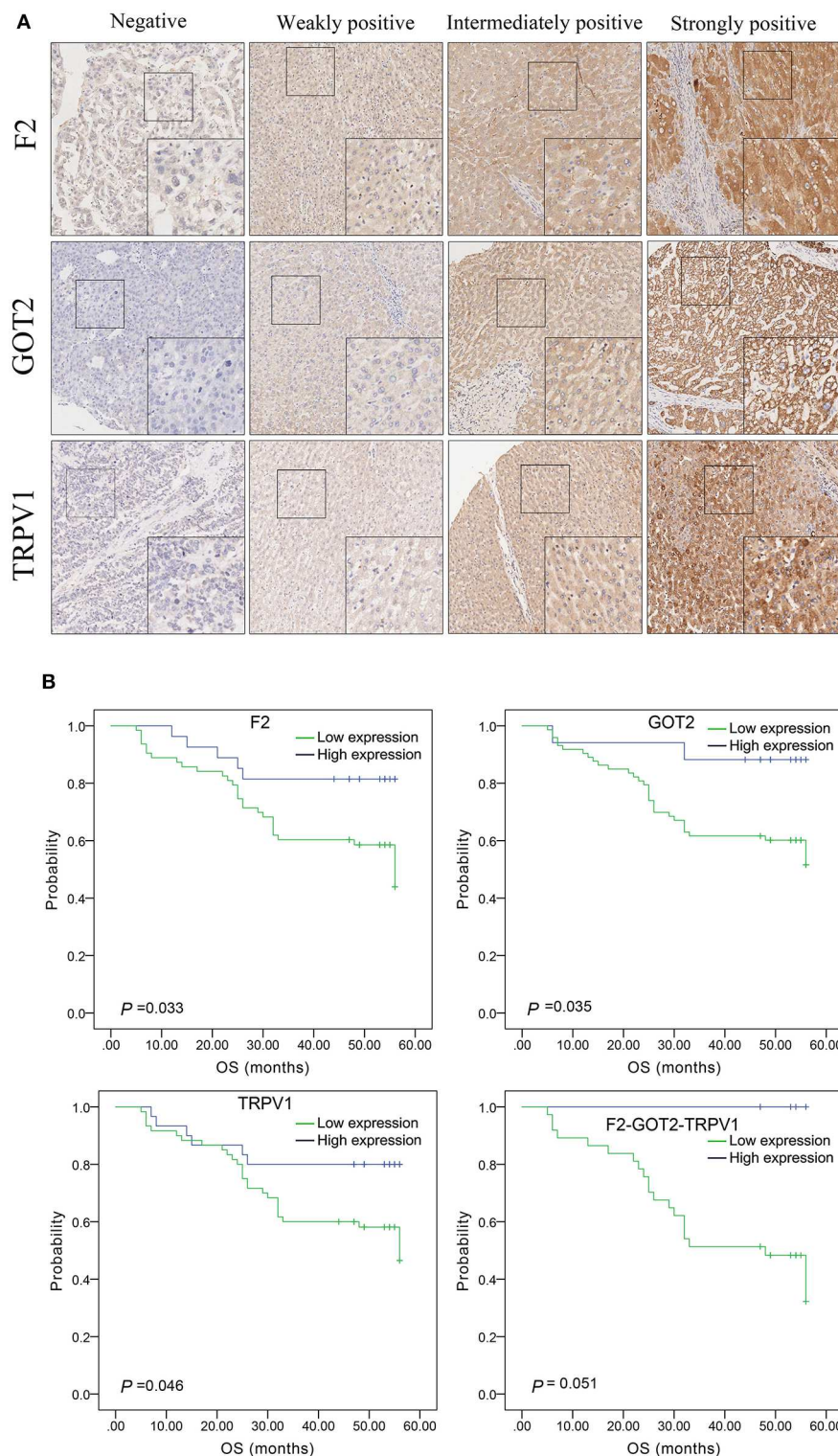


FIGURE 8 | The expression of *F2*, *GOT2*, and *TRPV1* in 90 pairs of HCC and adjacent normal liver tissues of biological tissue microarray by IHC, and the association with HCC patients prognosis. **(A)** Negative, weakly positive, intermediately positive, and strongly positive IHC staining of *F2*, *GOT2*, and *TRPV1*. *F2*, *GOT2*, and *TRPV1* were all low expressed in liver cancer. **(B)** The lower protein expression levels of *F2*, *GOT2*, and *TRPV1* were all associated with 5-year OS of 90 HCC patients, examining by Kaplan-Meier analyses and log-rank test. However, there was marginally significant association between the *F2*-*GOT2*-*TRPV1* combination protein expression levels with the OS of HCC patients. (*F2*: $p = 0.033$, *GOT2*: $p = 0.035$, *TRPV1*: $p = 0.046$, *F2*-*GOT2*-*TRPV1*: $p = 0.051$).

MAGEB3, *MAPT*, *MPV17*, *NTF3*, *PPAT*, *SLC2A1*, *SLC38A1*, and *TRPV1* are all novel prognostic markers associated with liver cancer OS found by our reliable and large-scale screening studies. Three genes (*APOB*, *PPAT*, and *TRPV1*) were associated with both DFS and OS of HCC, suggesting that *APOB*, *PPAT*, and *TRPV1* may be significant and effective in predicting both the progress and the adverse outcomes of HCC.

Moreover, there may be connections among the above selected genes and they can work together to influence the development and prognosis of liver cancer to some extent. Although there are some genes that had been reported as prognostic molecular markers of liver cancer, most reports focused on the impact of a single gene on the prognosis of liver cancer, few studies performed such a large-scale survival analysis. Studies of multiple gene combinations are more effective than the analysis of single genes in predicting the prognosis of liver cancer.

In our study, we performed three-gene combinations of the 39 DFS-related genes and 28 OS-related genes screened from the above survival analyses. In order to further study the predictive effect of the combinations constituted by the selected genes on the prognosis of liver cancer, and to compare the predictive power of single genes and corresponding gene combinations, we carried out thousands of K-M survival analyses on these combinations. To ensure the comparability and credibility, we removed the combinations of which the co-high or co-low expression group cases were fewer than 26, and screened 2,758 DFS-related combinations and 930 OS-related combinations with p -values < 0.01. Moreover, we also performed three-gene-combination models and K-M survival analyses on the 18 DFS-related genes and 12 OS-related genes we found but have not been systematically reported to be related to the prognosis of HCC. 317 unreported-gene combinations and 31 unreported-gene combinations significantly associated with DFS and OS, respectively, were screened out.

For the above four types of three-gene-combinations (the overall genes combinations associated with DFS, the unreported genes combinations associated with DFS, the overall genes combinations associated with OS, and the unreported genes combinations associated with OS), the top 15 combinations with the lowest p -values of the survival analyses and the genes they contained were, respectively, selected for comparison (Tables 2, 3, 4).

For example, for the overall gene combinations associated with OS, *KPNA2*-*SLC38A1*-*SPP1*, the median survival time difference between the co-high and the co-low expression group was 83.57 months. In contrast, that of the single genes *KPNA2*, *SLC38A1*, and *SPP1*, was 47.66, 35.61, and 29.64 months, respectively. After combining *KPNA2*, *SLC38A1*, and *SPP1*, the median survival time difference between the high and low expression groups was larger than that of any of the three single genes by at least 36 months. This shows that these three genes *KPNA2*, *SLC38A1*, and *SPP1*, after combination, may be better predictive values for liver cancer prognosis and may be more clinically useful for future treatment target selection.

We also selected genes that have not been previously reported for liver cancer prognosis and compared their prognostic efficacy with the corresponding three-gene combinations (the chart only

shows the top 15 groups with the lowest p -values of the three-gene combinations prognostic models). The expression of one of the combinations *F2*-*GOT2*-*TRPV1* had a greater effect on the median survival time of OS than any of the three individual genes (The median survival time difference: *F2*-*GOT2*-*TRPV1*: 55.68 months; *F2*: 23.62 months; *GOT2*: 32.26 months; *TRPV1*: 35.61 months).

Coagulation factor II (*F2*) plays a major role in proteolysis to form thrombin in the first step of the coagulation cascade and eventually generates hemostasis. An enrichment analysis of genetic changes during the development of HCC identified several hub genes, including *F2*, which interacts in several groups of conditional specific PPI networks (42). Additionally, it was reported that *F2* is associated with invasion in neuroendocrine prostate cancer (43). Glutamic-oxaloacetic transaminase 2 (*GOT2*) plays an important role in amino acid metabolism and the tricarboxylic acid cycle, and it affects the malate-aspartic acid shuttle activity and glycolysis in the liver under the stimulation of liver inflammation. (44, 45) *TRPV1* is a regulator of cell homeostasis, previous studies have revealed that the expression of *TRPV1* is significantly decreased in renal cell carcinoma, colorectal cancer, and melanoma. In addition, *TRPV1* can affect P53 and *TRPV1*-dependent pathways to inhibit the growth of colorectal cancer and melanoma (46–48), and can cause apoptosis in human osteosarcoma MG63 cells (49).

At present, there are few studies on the above three genes *F2*, *GOT2*, *TRPV1* and particular their combinations in the prognosis of HCC. In our study, the results of the 20 pairs of HCC and paracancerous tissues for qRT-PCR, as well as 90 pairs HCC biochips for IHC confirmed that all of the *F2*, *GOT2*, and *TRPV1* genes are significantly and consistently down-expressed in HCC tissues, and this is reconfirmed by three independent microarrays. Moreover, the low expression of *F2*, *GOT2*, and *TRPV1* were all significantly associated with poor prognosis of HCC. However, due to the number of death events in the *F2*-*GOT2*-*TRPV1* high expression group of in the HCC biochips being 0, the survival analysis of the *F2*-*GOT2*-*TRPV1* high and the expression group was marginally significant ($p = 0.051$), but this is still consistent with our above-mentioned big data-based multi-gene combination survival analysis results.

As there may be certain relationships between the genes we screened that are significantly associated with the prognosis of liver cancer, they can work together in the form of multi-gene combinations in the development of liver cancer. However, the predictive potency of different gene combinations varies. Some combinations are better predictors than individual genes, and therefore these combinations may be more valuable than individual genes in determining the target site for liver cancer prognosis. Due to limitations in human and material resources, it still remains unclear how these genes and gene combinations specifically affect the HCC survival. Further investigation and experimentations are needed to elucidate the biological mechanisms of the selected genes, particularly for the significant multi-gene combinations, in the development and progression of HCC.

Our findings cover a large gene level, and we have also explored the predictive efficacy of a number of gene

combinations for the prognosis of liver cancer. We believe that these highly significant prognostic-related genes and gene combinations derived from the above multiple screenings are promising, reliable molecular markers for the prognosis of liver cancer, and our screening methods can be extended to other tumor types.

In conclusion, based on a large sample size of public data platform, novel and effective data mining and multiple screening methods, large-scale survival analyses, as well as supplemental reliable experimental verification, we identified a series of novel genes and multi-gene combinations that are significantly associated with DFS or OS in liver cancer. Moreover, a huge difference between high and low expression group of these three-gene combination was detected. Some of the three-gene combinations can predict much longer or shorter survival time for liver cancer patients than the single genes. QRT-PCR, immunohistochemistry, and three independent microarray results confirmed our findings that three of the selected novel genes *F2*, *GOT2*, and *TRPV1*, as well as the corresponding combination *F2-GOT2-TRPV1*, showed significantly lower expression in HCC and are associated with OS in HCC. Some gene combinations may be more predictors of prognosis than single genes and can be used as potential effective therapeutic targets for liver cancer.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by The Ethics Committee of Guangxi Medical

University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

ML and XLi performed most analysis. ML led the writing of the manuscript. SL provided the clinical samples and participated in revising the manuscript. FX and JT participated in drafting and reviewing the manuscript. EG conducted a search for genes and preliminary screening work by keyword. XQ obtained and matched the TCGA samples data. ML, LW, and QL performed the single-gene and multi-gene-combination survival analyses. ZL and LL conducted an inquiry about the relevant information of the selected genes. XLu performed validation of the selected genes in three microarrays. KL and DZ performed the mRNA isolation and qRT-qPCR, and collected and analyzed experimental data. YY and XLi were subjected to immunohistochemistry and experimental data processing. FY and XZ participated in designing and reviewing the study. All the authors reviewed the manuscript and all the authors read and approved the final manuscript.

FUNDING

This study was supported by the National Natural Science Foundation of China (Grant No. 81760611) and the Key Laboratory of High-Incidence-Tumor Prevention and Treatment (Guangxi Medical University), Ministry of Education (No. GKE 2019-01).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fonc.2020.00847/full#supplementary-material>

REFERENCES

1. IARC. *Fact sheets by Population-Globocan-IARC*. (2019). Retrieved from: <http://gco.iarc.fr/today/fact-sheets-cancers> (accessed September 14, 2019).
2. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. (2018) 68:394–424. doi: 10.3322/caac.21492
3. Qiu WQ, Shi JF, Guo LW, Mao AY, Huang HY, Hu GY, et al. Medical expenditure for liver cancer in urban China: A 10-year multicenter retrospective survey (2002–2011). *J Cancer Res Ther*. (2018) 14:163–70. doi: 10.4103/jcrt.JCRT_709_16
4. Deng W, Long L, Li JL, Zheng D, Yu JH, Zhang CY, et al. Mortality of major cancers in Guangxi, China: sex, age and geographical differences from 1971 and 2005. *Asian Pac J Cancer Prev*. (2014) 15:1567–74. doi: 10.7314/APJCP.2014.15.4.1567
5. Zeng H, Chen W, Zheng R, Zhang S, Ji JS, Zou X, et al. Changing cancer survival in China during 2003–15: a pooled analysis of 17 population-based cancer registries. *Lancet Glob Health*. (2018) 6:e555–67. doi: 10.1016/S2214-109X(18)30127-X
6. Wang CY, Li S. Clinical characteristics and prognosis of 2887 patients with hepatocellular carcinoma: A single center 14 years experience from China. *Medicine*. (2019) 98:e14070. doi: 10.1097/MD.00000000000014070
7. Zhu J, Yin T, Xu Y, Lu XJ. Therapeutics for advanced hepatocellular carcinoma: Recent advances, current dilemma, and future directions. *J Cell Physiol*. (2019) 234:12122–32. doi: 10.1002/jcp.28048
8. Finn RS, Zhu AX, Farah W, Almasri J, Zaiem F, Prokop LJ, et al. Therapies for advanced stage hepatocellular carcinoma with macrovascular invasion or metastatic disease: a systematic review and meta-analysis. *Hepatology*. (2018) 67:422–35. doi: 10.1002/hep.29486
9. Cai W, Wang Z, Wei C, Wu M, Zheng W, Zhang H, et al. Prognostic evaluation of NANOG and OCT4 expression for posttransplantation hepatocellular carcinoma recurrence. *J Cell Biochem*. (2019) 120:8419–29. doi: 10.1002/jcb.28128
10. Forner A, Reig M, Bruix J. Hepatocellular carcinoma. *Lancet*. (2018) 391:1301–14. doi: 10.1016/S0140-6736(18)30010-2
11. Fan ST, Mau Lo C, Poon RT, Yeung C, Leung Liu C, Yuen WK, et al. Continuous improvement of survival outcomes of resection of hepatocellular carcinoma: a 20-year experience. *Ann Surg*. (2011) 253:745–58. doi: 10.1097/SLA.0b013e3182111195

12. Andreozzi M, Quintavalle C, Benz D, Quagliata L, Matter M, Calabrese D, et al. HMGA1 expression in human hepatocellular carcinoma correlates with poor prognosis and promotes tumor growth and migration in *in vitro* models. *Neoplasia*. (2016) 18:724–31. doi: 10.1016/j.neo.2016.10.002
13. Chen DH, Wu QW, Li XD, Wang SJ, Zhang ZM. SYPL1 overexpression predicts poor prognosis of hepatocellular carcinoma and associates with epithelial-mesenchymal transition. *Oncol Rep*. (2017) 38:1533–42. doi: 10.3892/or.2017.5843
14. Huang YL, Ning G, Chen LB, Lian YF, Gu YR, Wang JL, et al. Promising diagnostic and prognostic value of E2Fs in human hepatocellular carcinoma. *Cancer Manag Res*. (2019) 11:1725–40. doi: 10.2147/CMAR.S182001
15. Han R, Chen X, Li Y, Zhang S, Li R, Lu L. MicroRNA-34a suppresses aggressiveness of hepatocellular carcinoma by modulating E2F1, E2F3, and Caspase-3. *Cancer Manag Res*. (2019) 11:2963–76. doi: 10.2147/CMAR.S202664
16. Liang R, Lin Y, Ye JZ, Yan XX, Liu ZH, Li YQ, et al. High expression of RBM8A predicts poor patient prognosis and promotes tumor progression in hepatocellular carcinoma. *Oncol Rep*. (2017) 37:2167–76. doi: 10.3892/or.2017.5457
17. Chang L, Xi L, Liu Y, Liu R, Wu Z, Jian Z. SIRT5 promotes cell proliferation and invasion in hepatocellular carcinoma by targeting E2F1. *Mol Med Rep*. (2018) 17:342–9. doi: 10.3892/mmr.2017.7875
18. Zhou L, Zhu Y. The EpCAM overexpression is associated with clinicopathological significance and prognosis in hepatocellular carcinoma patients: A systematic review and meta-analysis. *Int J Surg*. (2018) 56:274–80. doi: 10.1016/j.ijsu.2018.06.025
19. Long H, Guo X, Qiao S, Huang Q. Tumor LXR expression is a prognostic marker for patients with hepatocellular carcinoma. *Pathol Oncol Res*. (2018) 24:339–44. doi: 10.1007/s12253-017-0249-8
20. Zhou H, Wang SC, Ma JM, Yu LQ, Jing JS. Sperm-Associated Antigen 5 expression is increased in hepatocellular carcinoma and indicates poor prognosis. *Med Sci Monit*. (2018) 24:6021–8. doi: 10.12659/MSM.911434
21. Chen D, Chen Y, Yan Y, Pan J, Xing W, Li Q, et al. Down-regulation of the tumour suppressor kappa-opioid receptor predicts poor prognosis in hepatocellular carcinoma patients. *BMC Cancer*. (2017) 17:553. doi: 10.1186/s12885-017-3541-9
22. Labib OH, Harb OA, Khalil OH, Baiomy TA, Gertallah LM, Ahmed RZ. The Diagnostic Value of Arginase-1, FTCD, and MOC-31 expression in early detection of hepatocellular carcinoma (HCC) and in differentiation between HCC and metastatic adenocarcinoma to the liver. *J Gastrointest Cancer*. (2020) 51:88–101. doi: 10.1007/s12029-019-00211-2
23. Zhang X, Wan JX, Ke ZP, Wang F, Chai HX, Liu JQ. TMEM88, CCL14 and CLEC3B as prognostic biomarkers for prognosis and palindromia of human hepatocellular carcinoma. *Tumour Biol*. (2017) 39:1010428317708900. doi: 10.1177/1010428317708900
24. Li B, Feng W, Luo O, Xu T, Cao Y, Wu H, et al. Development and validation of a three-gene prognostic signature for patients with hepatocellular carcinoma. *Sci Rep*. (2017) 7:5517. doi: 10.1038/s41598-017-04811-5
25. Long J, Zhang L, Wan X, Lin J, Bai Y, Xu W, et al. A four-gene-based prognostic model predicts overall survival in patients with hepatocellular carcinoma. *J Cell Mol Med*. (2018) 22:5928–38. doi: 10.1111/jcmm.13863
26. Zeng X, Yin F, Liu X, Xu J, Xu Y, Huang J, et al. Upregulation of E2F transcription factor 3 is associated with poor prognosis in hepatocellular carcinoma. *Oncol Rep*. (2014) 31:1139–46. doi: 10.3892/or.2014.2968
27. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*. (2012) 2:401–4. doi: 10.1158/2159-8290.CD-12-0095
28. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal*. (2013) 6:pl1. doi: 10.1126/scisignal.2004088
29. Hedditch EL, Gao B, Russell AJ, Lu Y, Emmanuel C, Beesley J, et al. ABCA transporter gene expression and poor outcome in epithelial ovarian cancer. *J Natl Cancer Inst*. (2014) 106:dju149. doi: 10.1093/jnci/dju149
30. Tao QF, Yuan SX, Yang F, Yang S, Yang Y, Yuan JH, et al. Aldolase B inhibits metastasis through Ten-Eleven Translocation 1 and serves as a prognostic biomarker in hepatocellular carcinoma. *Mol Cancer*. (2015) 14:170. doi: 10.1186/s12943-015-0437-7
31. Yan X, Yao M, Wen X, Zhu Y, Zhao E, Qian X, et al. Elevated apolipoprotein B predicts poor postsurgery prognosis in patients with hepatocellular carcinoma. *Oncotargets Ther*. (2019) 12:1957–64. doi: 10.2147/OTT.S192631
32. Lee G, Jeong YS, Kim DW, Kwak MJ, Koh J, Joo EW, et al. Clinical significance of APOB inactivation in hepatocellular carcinoma. *Exp Mol Med*. (2018) 50:147. doi: 10.1038/s12276-018-0174-2
33. Fu J, Qiu H, Cai M, Pan Y, Cao Y, Liu L, et al. Low cyclin F expression in hepatocellular carcinoma associates with poor differentiation and unfavorable prognosis. *Cancer Sci*. (2013) 104:508–15. doi: 10.1111/cas.12100
34. Fujiwara N, Nakagawa H, Enooku K, Kudo Y, Hayata Y, Nakatsuka T, et al. CPT2 downregulation adapts HCC to lipid-rich environment and promotes carcinogenesis via acylcarnitine accumulation in obesity. *Gut*. (2018) 67:1493–504. doi: 10.1136/gutjnl-2017-315193
35. Lu M, Lu L, Dong Q, Yu G, Chen J, Qin L, et al. Elevated G6PD expression contributes to migration and invasion of hepatocellular carcinoma cells by inducing epithelial-mesenchymal transition. *Acta Biochim Biophys Sin*. (2018) 50:370–80. doi: 10.1093/abbs/gmy009
36. Huang YC, Chen M, Shyr YM, Su CH, Chen CK, Li AF, et al. Glycine N-methyltransferase is a favorable prognostic marker for human cholangiocarcinoma. *J Gastroenterol Hepatol*. (2008) 23:1384–9. doi: 10.1111/j.1440-1746.2008.05488.x
37. Chang YY, Yen CJ, Chan SH, Chou YW, Lee YP, Bao CY, et al. NEK2 Promotes hepatoma metastasis and serves as biomarker for high recurrence risk after hepatic resection. *Ann Hepatol*. (2018) 17:843–56. doi: 10.5604/01.3001.0012.3146
38. Hyuga S, Wada H, Eguchi H, Otsuru T, Iwagami Y, Yamada D, et al. Expression of carbonic anhydrase IX is associated with poor prognosis through regulation of the epithelial-mesenchymal transition in hepatocellular carcinoma. *Int J Oncol*. (2017) 51:1179–90. doi: 10.3892/ijo.2017.4098
39. Ashida R, Okamura Y, Ohshima K, Kakuda Y, Uesaka K, Sugiura T, et al. CYP3A4 Gene is a novel biomarker for predicting a poor prognosis in hepatocellular carcinoma. *Cancer Genomics Proteomics*. (2017) 14:445–53. doi: 10.21873/cgp.20054
40. Quint K, Agaimy A, Di Fazio P, Montalbano R, Steindorf C, Jung R, et al. Clinical significance of histone deacetylases 1, 2, 3, and 7: HDAC2 is an independent predictor of survival in HCC. *Virchows Arch*. (2011) 459:129–39. doi: 10.1007/s00428-011-1103-0
41. Jiang P, Tang Y, He L, Tang H, Liang M, Mai C, et al. Aberrant expression of nuclear KPN2 is correlated with early recurrence and poor prognosis in patients with small hepatocellular carcinoma after hepatectomy. *Med Oncol*. (2014) 31:131. doi: 10.1007/s12032-014-0131-4
42. Xue H, Luo L, Yao YT, Wei LL, Deng SP, Huang XL. Integrated analysis of the RNA-Seq data of liver hepatocellular carcinoma. *Neoplasma*. (2018) 65:97–103. doi: 10.4149/neo_2018_170212N98
43. Choe H, Sboner A, Beltran H, Nanus D, Tagawa ST. PO-43 - Differential coagulation factor expression in neuroendocrine prostate cancer (PC), metastatic castrate-resistant PC, and localized prostatic adenocarcinoma. *Thromb Res*. (2016) 140(Suppl. 1):S192. doi: 10.1016/S0049-3848(16)30176-1
44. Wang T, Yao W, Li J, He Q, Shao Y, Huang F. Acetyl-CoA from inflammation-induced fatty acids oxidation promotes hepatic malate-aspartate shuttle activity and glycolysis. *Am J Physiol Endocrinol Metab*. (2018) 315:E496–510. doi: 10.1152/ajpendo.00061.2018
45. Yang H, Zhou L, Shi Q, Zhao Y, Lin H, Zhang M, et al. SIRT3-dependent GOT2 acetylation status affects the malate-aspartate NADH shuttle activity and pancreatic tumor growth. *Embo J*. (2015) 34:1110–25. doi: 10.15252/emboj.201591041
46. Wu YY, Liu XY, Zhuo DX, Huang HB, Zhang FB, Liao SF. Decreased expression of TRPV1 in renal cell carcinoma: association with tumor Fuhrman grades and histopathological subtypes. *Cancer Manag Res*. (2018) 10:1647–55. doi: 10.2147/CMAR.S166390

47. Hou N, He X, Yang Y, Fu J, Zhang W, Guo Z, et al. TRPV1 Induced apoptosis of colorectal cancer cells by activating calcineurin-NFAT2-p53 signaling pathway. *Biomed Res Int.* (2019) 2019:6712536. doi: 10.1155/2019/6712536
48. Yang Y, Guo W, Ma J, Xu P, Zhang W, Guo S, et al. Downregulated TRPV1 expression contributes to melanoma growth via the calcineurin-ATF3-p53 pathway. *J Invest Dermatol.* (2018) 138:2205–15. doi: 10.1016/j.jid.2018.03.1510
49. Bao Z, Dai X, Wang P, Tao Y, Chai D. Capsaicin induces cytotoxicity in human osteosarcoma MG63 cells through TRPV1-dependent and -independent pathways. *Cell Cycle.* (2019) 18:1379–92. doi: 10.1080/15384101.2019.1618119

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Liu, Liu, Liu, Xiao, Guo, Qin, Wu, Liang, Liang, Li, Zhang, Yang, Luo, Lei, Tan, Yin and Zeng. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Impact of Data Preprocessing on Integrative Matrix Factorization of Single Cell Data

Lauren L. Hsu^{1,2} and Aedin C. Culhane^{1,2*}

¹ Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, United States, ² Division of Biostatistics and Computational Biology, Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA, United States

OPEN ACCESS

Edited by:

Francesca Finotello,
Innsbruck Medical University, Austria

Reviewed by:

Valentine Svensson,
FL60 Inc, United States
Jean Fan,
Harvard University, United States
Federico Marini,
Johannes Gutenberg University
Mainz, Germany

*Correspondence:

Aedin C. Culhane
aedin@ds.dfci.harvard.edu

Specialty section:

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

Received: 20 February 2020

Accepted: 18 May 2020

Published: 23 June 2020

Citation:

Hsu LL and Culhane AC (2020)
Impact of Data Preprocessing on
Integrative Matrix Factorization of
Single Cell Data. *Front. Oncol.* 10:973.
doi: 10.3389/fonc.2020.00973

Integrative, single-cell analyses may provide unprecedented insights into cellular and spatial diversity of the tumor microenvironment. The sparsity, noise, and high dimensionality of these data present unique challenges. Whilst approaches for integrating single-cell data are emerging and are far from being standardized, most data integration, cell clustering, cell trajectory, and analysis pipelines employ a dimension reduction step, frequently principal component analysis (PCA), a matrix factorization method that is relatively fast, and can easily scale to large datasets when used with sparse-matrix representations. In this review, we provide a guide to PCA and related methods. We describe the relationship between PCA and singular value decomposition, the difference between PCA of a correlation and covariance matrix, the impact of scaling, log-transforming, and standardization, and how to recognize a horseshoe or arch effect in a PCA. We describe canonical correlation analysis (CCA), a popular matrix factorization approach for the integration of single-cell data from different platforms or studies. We discuss alternatives to CCA and why additional preprocessing or weighting datasets within the joint decomposition should be considered.

Keywords: data integration, matrix factorization, single cell, scRNA-seq, normalization, standardization, data preprocessing

INTRODUCTION

Single-cell (sc) molecular profiling provides unprecedented resolution and incredible potential to discover the heterogeneity of cell types and states and intercellular communication that drives complex cellular dynamics, homeostasis, response to environment, and disease. We will focus this review on the challenges and considerations when applying matrix factorization approaches to integration of sc RNA sequencing data (scRNA-seq). Matrix factorization methods, including principal component analysis (PCA), are central to scRNA-seq data analysis pipelines, but are often treated as “black boxes” within computational pipelines, with little consideration of what steps are included. We will “open the box” to illustrate the exact scaling and transformations that are performed on data in a PCA, and how different preprocessing steps impact data and cross-platform batch integration. These tips and considerations will also apply other single cell omics data, as well as to multi-modal integration of different omics data.

Challenging Properties of Single Cell Data

Single-cell data present a set of unique challenges for data analysis and integration (1–3). In contrast to traditional bulk RNA-seq which provides the average expression of RNA molecules across tens of thousands or millions of cells, scRNA-seq measures RNA in each cell.

The goal of scRNA-seq is frequently to define differential gene expression within specific cell types that characterize a phenotype, so cell type identification is a critical early step. In a tissue or biological sample, the population of cells is heterogeneous, containing many cell types including unidentified, new cell types, and cell states. Annotation of cell types in biological samples is challenging, as methods are still emerging and are limited by a lack of gold standard benchmarking data. To classify cell types and states, unsupervised clustering analysis is often used to partition cells into clusters, however, the biologically expected cell-to-cell variation within cell states is poorly understood, and cell clusters may be associated with systematic, batch, technical, or methodological artifacts (1). Toward the goal of creating a comprehensive cell type and state reference, the Human Cell Atlas will catalog the diversity of cell types in the human body (4) and anticipates discovering distinct tissue-specific, disease-specific, age-specific, gender-specific cell phenotypes, and will identify many new cell types and states that are yet to be defined.

Most, or at least half, of the transcriptome, is detected in a typical bulk RNAseq study. In contrast, scRNA-seq studies frequently measure <5,000 genes in a single cell (1). Most genes are not measured and these zero counts may represent zero gene expression or false negative dropout, that is, when a gene was expressed but was not detected due to technological limitations (3, 5) such as limited sequencing depth or stochastic variation. Gene expression may also be missed due to biological variance; single point-in-time measurements cannot capture dynamic processes, such as RNA transcriptional bursts. Emerging evidence suggests transcription occurs in bursts or pulses that depend on core promoter and enhancers (6) and a three-state model may be required to capture its biological complexity (7). These issues of scRNA-seq analysis underscore the importance of appropriate quality control, preprocessing, and normalization (1, 8).

Preprocessing of sc Sequencing Data

Several library preparation and read mapping approaches including genome or transcriptome mapping and pseudo-alignment can be used to generate a “raw” or unique molecular identifier (UMI) count matrix from sequencing reads (9), but in a comparison of over 3,000 preprocessing and analysis pipelines, Vieth et al. found normalization of the count matrix had greatest impact on downstream analysis (9). Standard “normalization” pipelines include scaling using sample-specific size factors, log transformation to reduce skewness, and feature filtering before PCA. The selection of a particular normalization routine will itself embed assumptions about the underlying distribution of the data. Inappropriate preprocessing may introduce artifacts that impact the ability to perform further preprocessing (e.g., alignment and integration of batches of sc data both within and between studies) and downstream biological analyses [e.g., cell type identification, classification, and differential gene expression (1, 8, 9)].

Depending upon the analysis method selected, objective defined, and the dataset itself, different approaches to preprocessing may be appropriate; various data scaling, centering, standardization, and transformation (**Figure 1**) approaches can be applied. Frequently these terms are used interchangeably even though they represent different data manipulations (11, 12). Often the goal of preprocessing steps is to generate data that meet the linearity, homoscedasticity (that the points have the same scatter, i.e., there is no relationship between mean and variance), and normality assumptions that are required for most parametric statistical methods, including linear regression. A recent review of metabolomics data includes an extensive review of scaling and transformation approaches on sparse data (13).

- **Scaling** adjusts the range of the data, by dividing by a value. There are two broad subclasses of scaling factors: size measures (e.g., mean or library size) and data dispersion measures (e.g., standard deviation). Unit or unit variance scaling uses the standard deviation as the scaling factor, such that points have a standard deviation of one and therefore the data are analyzed on the basis of correlations instead of covariances. If data are scaled by dividing by the standard deviation, then the correlation is equal to the covariance of those two variables, since the Pearson correlation coefficient of two variables is equal to dividing the covariance of these variables by the product of their standard deviations. Scaling by size measures is important when integrating multiple datasets in cases where the range of values and means of the data differ substantially.
- **Centering** is subtracting the mean of a set of points from each data point so that the new mean is 0. The scale does not change, one unit is still one unit. In **Figure 1**, we see centering produces data with a mean at zero, but the standard deviation is unchanged
- **Standardization** includes *centering* and *scaling*. A *Z-score standardization* is subtracting the mean and dividing by the standard deviation of all points. A one-unit difference after this adjustment now indicates a one-standard deviation difference. Note whilst it changes the range of the data it may not affect the distribution, and may require an additional transformation
- **Transformations**, including log transformations ($\log_2$ or $\log_{10}$) or log with pseudocount (e.g., $\log +1$), are commonly applied to sc data that increase proportionally (% or fold change) rather than linearly (8). A log transform or power transform may make skewed data look more symmetric or Gaussian (normally distributed in a bell-curve shape) and correct for heteroscedasticity (unequal scatter of points, where variance differs with mean). Recent studies reported that $\log_2+1$ transformation may distort data, introducing false variability in dimension reduction and impacting downstream analysis (8, 14, 15). Given that heteroscedasticity in omics data is both multiplicative and additive, generalized log variance-stabilizing transformations such as arcsinh (asinh) of scRNA-seq data (16, 17) and CyToF proteomic data (18, 19) are recommended. Rank-based inverse normal transformation has also been used to rescale scRNAseq gene expression (20).

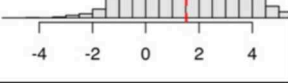
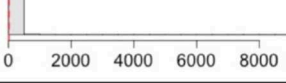
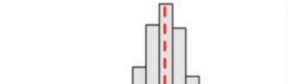
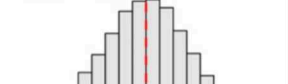
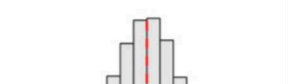
Example raw datasets		Graphical Examples	
Method	Formula	Toy data	scMix, 10X counts
		Mean = 1.5 SD = 1.5 	Mean = 14.7 SD = 73.0 
Scale	$x_{i,j}^* = \frac{x_{i,j}}{\text{scaling factor}}$ where the <i>scaling factor</i> can be a size or data dispersion measure	Mean = 0.7 SD = 0.7 	Mean = 0.2 SD = 1.0 
Center	$x_{i,j}^* = x_{i,j} - \text{column mean}$	Mean = 0 SD = 1.5 	Mean = 0 SD = 73 
Standardize	$x_{i,j}^* = \frac{x_{i,j} - \text{column mean}}{\text{scaling factor}}$ where the <i>scaling factor</i> can be a size or data dispersion measure. For example z-score subtracts means, divides by standard deviation	Mean = 0 SD = 1 	Mean = 0 SD = 1 
Transform	$x_{i,j}^* = f(x_{i,j})$ where $f(x)$ is the transformation function, for example logarithms are commonly used	Mean = 0.5 SD = 1.4  <i>Log₂ transformation</i>	Mean = 2.0 SD = 1.9  <i>Log₂ transformation, pseudocount of 1</i>

FIGURE 1 | Common data preprocessing steps include scaling, centering, standardization, and transformation. Graphical examples of these preprocessing routines are applied to two datasets (1) “toy data” with a mean and standard deviation (SD) of 1.5 generated for purposes of illustration, and (2) the 10X raw counts matrix in the scMix benchmarking dataset used in **Figure 2** (10).

- **Normalization** transforms the data points so that their distribution resembles a normal, also called Gaussian, distribution. In a normal distribution (i.e., the classic bell curve) points are distributed symmetrically around the mean, most observations are close to the mean, and the median and mean are the same. Depending upon the distribution of the original dataset, this may be achieved by a log transformation, or may require more extensive preprocessing. Two recent articles have proposed analysis of Pearson residuals rather than log normalized counts (8, 14). In bioinformatics and computational fields, this term may also refer to size and/or range scaling transformation which may not produce a normal distribution (21).

Feature selection, for instance restricting analysis to over-dispersed genes which are expected to capture a disproportionate amount of the variance in the data, is included in many analysis pipelines to reduce the computation time (16, 22). Furthermore, selecting genes with high biological variability, to exclude many genes with low biological signal and high numbers of zeros, may increase the signal to noise ratio in dimension reduction.

Dimension Reduction

Data dimension reduction is indispensable in single cell data analyses because it facilitates exploratory data analysis and visualization, and is an essential step in many downstream analysis including cell clustering (23, 24), cell-type identification, cell trajectory, lineage reconstruction, and trajectory inference (25–27). It is also a critical first step in many algorithms that align and integrate sc datasets (11, 22, 28).

Dimension reduction transforms the data to a new coordinate system (i.e., a low-dimensional shared latent space) such that the greatest variance can be identified and distinguished from background noise, or less informative variance. The output is a set of embeddings for each data point which encode their location in the low-dimensional shared latent space. It is frequently achieved using matrix factorization, a class of unsupervised techniques that provide a set of principled approaches to parsimoniously reveal the low-dimensional structure while preserving as much information as possible from the original data.

Principal component analysis (PCA) is arguably the oldest, fastest, and the most commonly used matrix factorization approach (29). PCA is a deterministic algorithm that seeks linear combinations of the variables that explain the variance in the data and ranks these such that the first component explains most of the variance or “strongest” pattern in the data. PCA uses a Gaussian likelihood and is best applied to data that are approximately normally distributed. Whilst it is not recommended to be applied to highly skewed data (**Figure 1**), nonetheless, in a recent systematic analysis of 18 linear and non-linear dimension reduction approaches, PCA and other classical linear methods performed surprisingly well in both clustering and lineage inference analysis when assessed on 30 scRNA-seq datasets (30). Linear (straight-line) analysis methods including PCA, independent component analysis (ICA), factor analysis (FA) ranked best in clustering. PCA, FA, non-negative matrix factorization [NMF, (31, 32)], and uniform manifold

approximation and projection [UMAP, (33)] ranked top in lineage inference analysis (30). We compare ICA and NMF matrix factorization in a recent review (31).

Dimension reduction methods optimized for count data that apply a better-fitting likelihood model (e.g., Poisson or negative binomial) are promising for addressing the skewed distribution of sc count data (8, 14). However, glmPCA (8), Poisson factorization (34–36), and probabilistic count matrix factorization [pCMF, (37)], as well as methods designed to model zero-inflated sparse data, including ZIFA and ZINB-WaVE (38, 39) did not outperform PCA across the full range of analyses and evaluations performed in the study Sun et al. (30). While there are particular settings where these methods may be most appropriate, they are not necessarily appropriate as “general-purpose” approaches. The high computational cost and long run time make many of these models difficult to integrate into multi-step bioinformatics pipelines.

Non-linear dimension reduction methods can identify variance in subsets of features by fitting local linear maps on subsets of points. Non-linear methods applied to sc data include diffusion maps (40), locally linear embedding, isoMap, kernel adaptations of linear methods, uniform manifold approximation and projection (UMAP) (41), and t-distributed stochastic neighbor embedding [tSNE, (42)]. However, similar to the methods that apply non-Gaussian likelihoods, non-linear dimension reduction methods are often computationally expensive and since they are not deterministic may produce different embeddings when re-applied to the same dataset. To improve computational tractability, PCA is frequently used as a preprocessing step prior to non-linear dimensionality reduction approaches including t-distributed stochastic neighbor embedding [tSNE, (43)] and UMAP (33). Although not required to run UMAP, in practice, it can be applied to accelerate computation time by significantly reducing dimensionality and noise while preserving underlying latent structure.

In this review, we focus on PCA because of its popularity, performance, and widespread use. PCA is a central step in many sc analysis algorithms and pipelines. When used with sparse-matrix representations, it can easily scale to large datasets. Excellent general tips for dimension reduction have been described (44), so we will focus on considerations and limitations when applying dimension reduction to sc data, including a step-by-step explanation of how PCA works, especially when applied to integrative sc analysis (**Figure 2A**).

The Impact of Data Preprocessing on Dimension Reduction

There are two types of PCA, which differ in data centering and scaling prior to matrix decomposition. PCA of a covariance matrix or a correlation matrix is achieved by applying matrix factorization to a centered but unscaled matrix, or a centered and scaled matrix, respectively (**Figure 2A**, Step 2). The latter is the most popular form of PCA. Linear regression using non-linear iterative partial least-squares (NIPALS), eigen analysis, or singular value decomposition (SVD) are a few of the many ways to factorize or decompose a matrix. SVD is a basic matrix operation, and fast approximations of SVD, including IRLBA, are commonly applied to sc data [extensively reviewed by (45)].

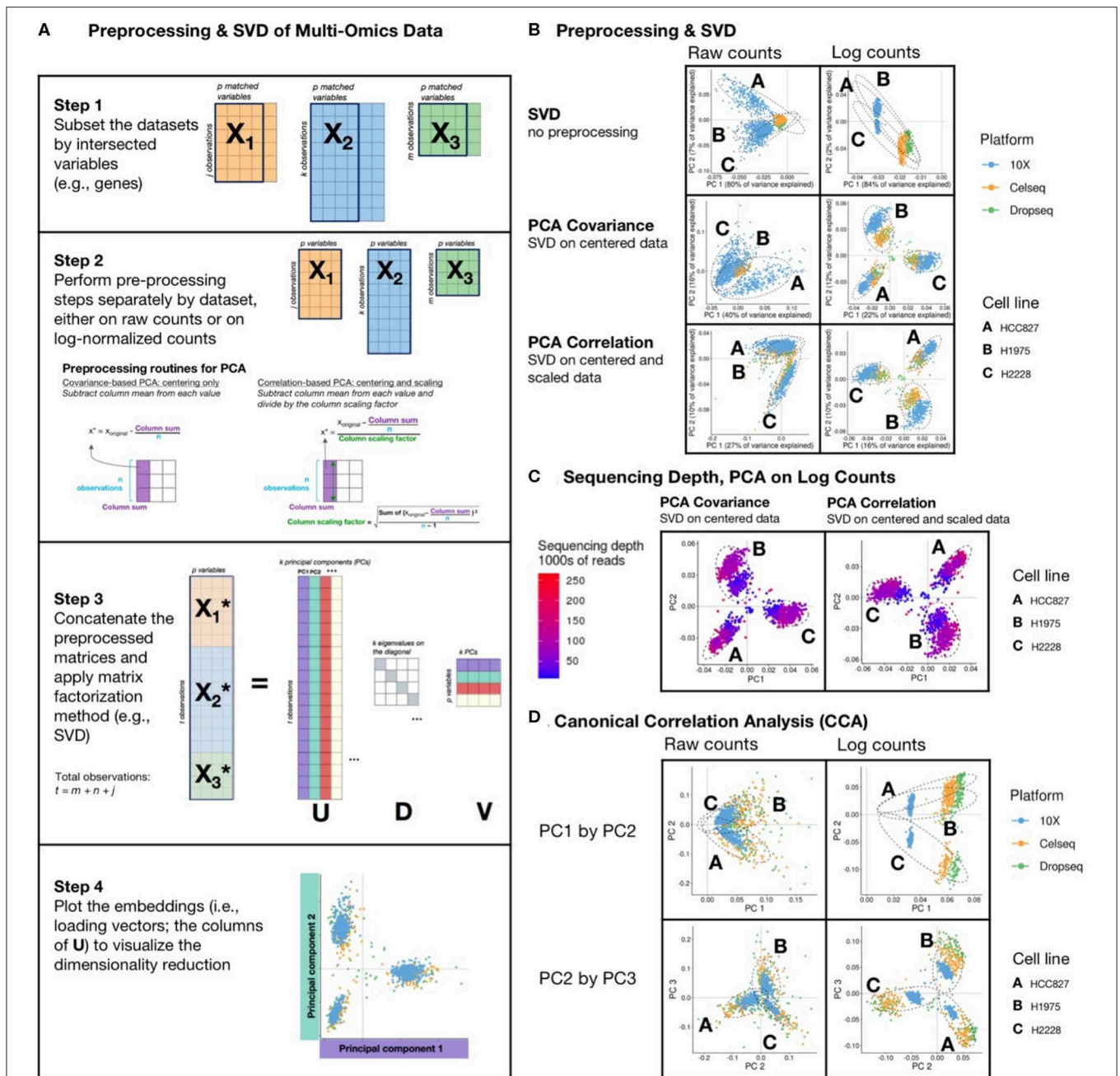


FIGURE 2 | Matrix Factorization of sc data: **(A)** schematic diagram of a PCA or CCA workflow, includes: (1) filtering of datasets to intersecting genes; (2) scaling, transformation, and normalization of individual and joint count matrices; (3) concatenating matrices and applying a matrix factorization, usually singular value decomposition (SVD); and (4) visualizing results. SVD is a matrix operation that finds for a given input matrix the left singular vectors (**U**), the right singular vectors (**V**), and the singular values (**D**), such that the product of **U** and **V** with their respective transpose matrices is the identity matrix. Each singular vector is orthogonal to the others, and they are ordered such that the first component explains the greatest variance, and each subsequent component explains less than the preceding. **(B)** The first two principal components of SVD performed on counts and log-transformed counts of the scMix benchmarking data (10), comprising 3 cell lines (HCC827, H1975, and H2228), that were unprocessed, centered, and centered and scaled, to reflect SVD, covariance-based and correlation-based PCA, respectively. Results from covariance-based and correlation-based PCA applied to log-transformed data are similarly effective, showing moderate data integration and separation by cell type but an arch effect is visible on PC1 and PC2 in SVD of the raw counts. **(C)** Covariance-based and correlation-based PCA of log-transformed data, colored by sequencing depth, show that unadjusted differences in sequencing depth limit integration, forming a gradient across each cluster. **(D)** The first three principal components from Canonical Correlation Analysis (CCA) of scMix data. In both raw counts and log-transformed data, PC1 provides poor separation by cell type and batch integration. The plot of PC2 by PC3 from CCA on log-transformed data show reasonable clustering by cell line, though exhibit poor batch integration; in contrast, PC2 by PC3 plot from CCA on raw data shows better batch integration and poorer separation by cell type.

SVD factors an input matrix into three matrices U , D , and V , as illustrated schematically in **Figure 2A** (46) (R code to perform PCA via both eigen analysis and SVD are provided in Supplementary Methods). The maximum number of principal components or rank of the analysis is the number of rows or columns of the matrix (whichever is lower, $n-1$, or $p-1$), though typically 30 or fewer components are examined in most scRNA-seq pipelines (22). Selection of the correct number of components is non-trivial and most commonly achieved by heuristic approaches. To understand the distribution of variance explained by each component, scree-plots can also be helpful visual tool (47, 48) and permutations based approaches are recommended (49, 50).

Figure 2B displays SVD of raw count or $\log_2$ transformed count matrices that were (1) unprocessed data (top row); (2) centered by subtracting column means (middle row); and (3) scaled and centered to reproduce SVD. (2) and (3) show PCA of a covariance matrix (princomp in R), and PCA of a correlation matrix (prcomp in R), respectively (**Figure 2B**). These are applied to a small, well-described benchmarking dataset (10), comprising scRNA-seq measurements of a three cell line mixture on three technological platforms (10X, Dropseq, and CELseq2). Both forms of PCA had greater success in finding structure in the data as compared to SVD alone. However, clusters of cell lines could only be distinguished in data that were log transformed. Moderate cross platform integration was observed in data that were centered, or centered and scaled (equivalent of PCA of a covariance or correlation matrix, respectively). Nonetheless, as illustrated in **Figure 2C**, we observe that systematic differences in sequencing depth between the three platforms still creates a gradient across each cluster, preventing full integration. Whilst this analysis was performed on all variables (genes), we and others have found that excluding genes with low variability and high numbers of zeros prior to dimensionality reduction may increase the signal to noise ratio (12, 48, 51).

The Horseshoe or Arch Effect

PCA is optimized for continuous, normally distributed data and is suboptimal when applied to sparse data with many zero counts. The arch or horseshoe is a common pitfall and has been described in detail in the literature (44, 52, 53). This distortion results from the presence of a gradient or sequential latent ordering in the data [Tutorial by (54)]. In the top row of **Figure 2B** all of the cell lines on the first component (PC1) are on the same side of the origin, forming a classical horseshoe pattern, characterized by a distinctive “arched” shape, with points mostly on one side of the origin and folding back on itself in one of the dimensions. This indicates that additional data preprocessing is required; cell lines cannot be distinguished, and the data are not integrated across batches. In the top right plot of **Figure 2B** which shows SVD on unprocessed log counts, the first 2 PCs appear correlated, but are by definition orthogonal—their dot product is 0. Orthogonal vectors are uncorrelated only when at least one of them has mean 0. In contrast, when data are centered (e.g., middle and bottom row of **Figure 2B**), these artifacts are gone. It is vital that such arch effects are identified, especially when PCA forms part of a computational workflow that extracts the first n principal components without inspection. As seen in **Figure 2**,

preprocessing and data normalization can remove arch artifacts and we refer the reader to excellent recent reviews on the subject (44, 52–54).

Examining PC plots can illuminate issues beyond the arch effect, in this case for instance, showing that the 10X data are located further from the origin on PC1 and PC2 as a result of difference in sequencing depth between platforms (**Figures 2B,C**). This can be corrected for by scaling the size factors by dataset to account for these systematic differences prior to log-normalization (55).

Integrating Two or More Datasets With K-table Matrix Factorization

Matrix factorization approaches have been highly effective and widely applied to removing batch effects in bulk omics data (56, 57). Whilst dimensionality reduction methods like PCA can discover batch effects (1, 11, 28), and could also be applied to remove within or even between batch effects in sc data, it is more common to explicitly define the blocks, groups, or datasets to be integrated and apply matrix factorization that is designed to extract correlated structure between groups. Emerging sc data integration and cross-study batch correction methods frequently use PCA or joint matrix decompositions as a first step.

Matrix factorization approaches that integrate multiple groups or matrices with matched rows or columns, often called K-table, multi-block component analysis or tensor decompositions (46), have been applied to both bulk and scRNA-seq data integration (46). The simplest K-table approach is possibly Procrustean analysis (58, 59). Procrustes was a figure from Greek mythology who was famous for cutting limbs or stretching unknowing passers-by such that they fit into his bed, and similarly, Procrustean analysis involves rotation or reduction of a component from one PCA to best fit a second PCA. Several other matrix factorization approaches for K-table exist (46).

Arguably the most popular K-table approach applied to omics data is canonical correlation analysis [CCA, (60, 61)], which maximizes the correlation between components, or canonical variables of each dataset, and has been widely applied to integration of bulk omics data [reviewed by (46, 62)]. Classical CCA requires more observations than features, and therefore sparse implementations that include feature selection are used in the analysis of bulk omics data (63, 64). CCA and adaptations of CCA have been applied to integrate scRNA-seq including the cross-study integration of stimulated and resting human peripheral blood mononuclear cells (PBMCs); cross-platform integration of mouse hematopoietic progenitors scRNA-seq data; and heterogeneous case-control cell populations after drug exposure (16, 22). Seurat 3 uses CCA with anchors to align datasets that are extracted using mutual nearest neighbors on the CCA subspace (65). Harmony uses PCA as a first step (66). PCA or CCA is the first step in scAlign, a neural-network based method for pairwise or data to references, alignment of single cell data (67) which was reported to outperform other single cell alignment methods (CCA in Seurat, scVI, MNN scanorama, scmap, MINT, and scMerge). Non-linear matrix factorization approaches for integration of datasets include joint NMF [LIGER, (68)] but in a recent comparative study this was reported to be computationally slow and may overlay samples

of little biological resemblance compared to the other methods (69). A benchmark comparison of 14 methods for integration of scRNA-seq datasets, on datasets from different technologies with identical cell types, non-identical cell types, multiple batches, big data, and simulated data revealed that harmony, LIGER, and Seurat 3 CCA are most performant (65).

Other matrix decomposition approaches, including multiple co-inertia analysis (48, 70), multiple factor analysis (71, 72), and consensus PCA (73–75), maximize a covariance or squared covariance criterion and are not limited by a requirement for more observations than features. These have been applied to bulk omics data and clustering, for example Meng et al., applied Westerhuis's modified implementation of consensus PCA to integrate methylation, proteomic and genomics data, reporting it was performant and faster than iCluster/iCluster+ (75). Dimension reduction methods for both single and K-table analysis, including a summary of the mathematical formulae and overview of available software packages for each mode of analysis, have been recently reviewed (46). Of note, there is also a recently described generalized framework to easily modulate between covariance and correlation-optimization in integrative matrix factorization (62, 76).

Horseshoes in CCA

Similar to PCA, a problematic arch effect is seen on PC1 and PC2 (**Figure 2D**) when CCA is applied to align and integrate raw counts or log counts of scRNA-seq measurements of three cell lines that were obtained on three technological platforms: 10X, Dropseq, and CELseq2 (10). The raw data had more platform overlap, and the log-transformed had less overlap in cell types in PC2 and PC3 (**Figure 2D**). These data demonstrate that, if CCA is used as a first step in a pipeline, it should include a check for the presence of such artifacts. For example, upon examining **Figure 2D**, one could exclude PC1, since CCA integrates the data across platforms in PC2 and PC3.

Scaling of Datasets in CCA

Simultaneous integration of multiple matrices is more complex than integrative analysis of a single dataset because each dataset may have different numbers of observations (cells), internal structure, and variance. In this CCA (**Figure 2D**) vignette the 10X dataset exhibited less correlated structure with the Dropseq and CELseq2 datasets, which had lower sequencing depth (**Figure 2C**). Therefore, in K-table matrix decomposition two levels of preprocessing are recommended. First, each individual dataset is normalized, centered, and scaled. Secondly, datasets are scaled by cross-dataset size factors (55), weighted to inflate or deflate the contribution of individual datasets, such as scaling by the square root of their total inertia, the percent variance on the first principal component, sample size, or another measure of data quality or expected contribution [reviewed by (46)].

Key Takeaways

When applying matrix factorization methods including PCA, it is recommended to consider the impact of scaling, log-transforming, standardization, and normalization. Common data challenges, and tips to address them, include:

1. *Preprocessing of data.* Consider each step in the pipeline and how it transforms the data. If necessary, consider preprocessing the data yourself. *Visualize data* after intermediate steps to ensure data are processed as expected, and to diagnose any issues that may arise.
2. *Heteroscedasticity.* Whilst widely used, $\log_2$ transformation of expression values combined with pseudocounts may not be appropriate, consider using a variance-stabilizing transformation.
3. *Arch effect in PCA.* Examine PCs if weights are not centered around the origin with negative and positive scores, to check if there is an arch artifact. This can be mitigated by scaling and/or normalization.
4. *Systematic differences in sequencing depth.* When working with data from multiple batches, we found that the *multiBatchNorm* function from the *batchelor* R/Bioconductor package corrected for the differences in sequencing depth.
5. *Uncertainty around ground truth.* Test methods using a well-characterized benchmarking dataset, if possible. The *CellBench* R/Bioconductor package provides access to several datasets, including the *scmix* dataset used in **Figure 2** (77).

SUMMARY

Single cell omics data are expanding our understanding of tumor heterogeneity, the tumor microenvironment, and tumor immunology. Algorithms for cell clustering, cell type identification, and cell trajectory analysis rely on dimension reduction to achieve computationally tractable solutions. The sparsity, noise, and high dimensionality of these data present unique challenges and underscore the importance of dimension reduction in sc analysis. PCA is widely used and popular for its speed, scalability, and performance, though it may not be the most optimal method for sc data. Matrix factorization approaches optimized for count matrices or distances matrices have been described [reviewed by (38)], and it is likely that more performant data preprocessing, scaling, and transformation approaches will continue to be developed. These methods will improve the performance of dimension reduction approaches in sc data integration and analysis.

RESOURCES

We include below a short list of single cell analysis resources, vignettes, and reference materials

<https://osca.bioconductor.org/>
<https://github.com/seandavi/awesome-single-cell>
<https://satijalab.org/seurat/>
<https://hemberg-lab.github.io/scRNA.seq.course/>
<https://github.com/SingleCellTranscriptomics>

SUPPLEMENTAL MATERIAL

R Code to reproduce these figures which describes different implementation of SVD and PCA is publicly available at <https://>

github.com/aedin/Frontiers_Supplement/. It includes a code to generate PCA, computed by SVD, eigenanalysis and PCA using R packages princomp, prcomp, ade4, FactoMineR. In each case, the relationship between these methods is described.

AUTHOR CONTRIBUTIONS

LH and AC wrote the paper. LH wrote the code and performed analysis. AC wrote the online supplemental PCA vignette code.

REFERENCES

- Hicks SC, Townes FW, Teng M, Irizarry RA. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics*. (2018) 19:562–78. doi: 10.1093/biostatistics/kxx053
- Adarabioyo MI, Ipinoyi RA. Comparing zero-inflated poisson, zero-inflated negative binomial and zero-inflated geometric in count data with excess zero. *Asian J Prob Stat*. 4, 1–10. doi: 10.9734/ajpas/2019/v4i230113
- Büttner M, Miao Z, Wolf FA, Teichmann SA, Theis FJ. A test metric for assessing single-cell RNA-seq batch correction. *Nat Methods*. (2019) 16:43–9. doi: 10.1038/s41592-018-0254-1
- Rozenblatt-Rosen O, Stubbington MJT, Regev A, Teichmann SA. The human cell atlas: from vision to reality. *Nature*. (2017) 550:451–3. doi: 10.1038/550451a
- Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, et al. Comparative analysis of single-cell RNA sequencing methods. *Mol Cell*. (2017) 65:631–43. doi: 10.1016/j.molcel.2017.01.023
- Larsson AJM, Johnsson P, Hagemann-Jensen M, Hartmanis L, Faridani OR, Reinius B, et al. Genomic encoding of transcriptional burst kinetics. *Nature*. (2019) 565:251–4. doi: 10.1038/s41586-018-0836-1
- Jia C. Kinetic foundation of the zero-inflated negative binomial model for single-cell RNA sequencing data. *arXiv [q-bio.MN]*. (2019). Available online at: <http://arxiv.org/abs/1911.00356> (accessed December 16, 2019).
- Townes FW, Hicks SC, Aryee MJ, Irizarry RA. Feature selection and dimension reduction for single cell RNA-seq based on a multinomial model. *Genome Biol*. (2019) 20:295. doi: 10.1186/s13059-019-1861-6
- Vieth B, Parekh S, Ziegenhain C, Enard W, Hellmann I. A systematic evaluation of single cell RNA-seq analysis pipelines. *Nat Commun*. (2019) 10:4667. doi: 10.1038/s41467-019-12266-7
- Tian L, Dong X, Freytag S, Lê Cao K-A, Su S, JalalAbadi A, et al. Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nat Methods*. (2019) 16:479–87. doi: 10.1038/s41592-019-0425-8
- Amezquita RA, Lun ATL, Becht E, Carey VJ, Carpp LN, Geistlinger L, et al. Orchestrating single-cell analysis with bioconductor. *Nat Methods*. (2020) 17:137–45. doi: 10.1038/s41592-019-0654-x
- Kiselev V, Andrews T, Westoby J, McCarthy D, Büttner M, Lee J, et al. *Analysis of Single Cell RNA-Seq Data*. (2019) Available online at: <http://hemberg-lab.github.io/scRNA.seq.course/> (accessed December 13, 2019)
- van den Berg RA, Hoefsloot HCJ, Westerhuis JA, Smilde AK, van der Werf MJ. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*. (2006) 7:142. doi: 10.1186/1471-2164-7-142
- Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol*. (2019) 20:1–15. doi: 10.1186/s13059-019-1874-1
- Lun A. Overcoming systematic errors caused by log-transformation of normalized single-cell RNA sequencing data. *bioRxiv*. (2018) 404962. doi: 10.1101/404962
- Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM III, et al. Comprehensive integration of single-cell data. *Cell*. (2019) 177:1888–902. doi: 10.1016/j.cell.2019.05.031
- Lin SM, Du P, Huber W, Kibbe WA. Model-based variance-stabilizing transformation for Illumina microarray data. *Nucleic Acids Res*. (2008) 36:e11. doi: 10.1093/nar/gkm1075
- Bendall SC, Simonds EF, Qiu P, Amir EAD, Krutzik PO, Finck R, et al. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science*. (2011) 332:687–96. doi: 10.1126/science.1198704
- Nowicka M, Krieg C, Crowell HL, Weber LM, Hartmann FJ, Guglietta S, et al. CyTOF workflow: differential discovery in high-throughput high-dimensional cytometry datasets. *F1000Res*. (2017) 6:748. doi: 10.12688/f1000research.11622.1
- Mohammadi S, Davila-Velderrain J, Kellis M. Reconstruction of cell-type-specific interactomes at single-cell resolution. *Cell Syst*. (2019) 9:559–68. doi: 10.1016/j.cels.2019.10.007
- Evans C, Hardin J, Stoebel DM. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief Bioinform*. (2018) 19:776–92. doi: 10.1093/bib/bbx008
- Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol*. (2018) 36:411–20. doi: 10.1038/nbt.4096
- Senabouth A, Lukowski SW, Hernandez JA, Andersen SB, Mei X, Nguyen QH, et al. ascend: R package for analysis of single-cell RNA-seq data. *Gigascience*. (2019) 8:giz087. doi: 10.1093/gigascience/giz087
- Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, et al. SC3: consensus clustering of single-cell RNA-seq data. *Nat Methods*. (2017) 14:483–6. doi: 10.1038/nmeth.4236
- Saelens W, Cannoodt R, Todorov H, Saeys Y. A comparison of single-cell trajectory inference methods. *Nat Biotechnol*. (2019) 37:547–54. doi: 10.1038/s41587-019-0071-9
- Ji Z, Ji H. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res*. (2016) 44:e117. doi: 10.1093/nar/gkw430
- Way GP, Greene CS. Bayesian deep learning for single-cell analysis. *Nat Methods*. (2018) 15:1009–10. doi: 10.1038/s41592-018-0230-9
- Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol*. (2019) 15:e8746. doi: 10.15252/msb.20188746
- Pearson K. On lines and planes of closest fit to systems of points in space. *Philos Mag*. (1901) 2:559–72. doi: 10.1080/14786440109462720
- Sun S, Zhu J, Ma Y, Zhou X. Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. *Genome Biol*. (2019) 20:269. doi: 10.1186/s13059-019-1898-6
- Stein-O'Brien GL, Arora R, Culhane AC, Favorov AV, Garmire LX, Greene CS, et al. Enter the matrix: factorization uncovers knowledge from omics. *Trends Genet*. (2018) 34:790–805. doi: 10.1016/j.tig.2018.07.003
- Shao C, Höfer T. Robust classification of single-cell transcriptome data by nonnegative matrix factorization. *Bioinformatics*. (2017) 33:235–42. doi: 10.1093/bioinformatics/btw607
- McInnes L, Healy J, Saul N, Großberger L. UMAP: uniform manifold approximation and projection. *J Open Source Softw*. (2018) 3:861. doi: 10.21105/joss.00861
- Cao Y, Zhang A, Li H. Multisample estimation of bacterial composition matrices in metagenomics data. *Biometrika*. (2020) 107:75–92. doi: 10.1093/biomet/asz062
- Salmon J, Harmany Z, Deledalle CA, Willett R. Poisson noise reduction with non-local PCA. *J Math Imaging*

FUNDING

We are grateful for funding from Stand Up to Cancer, National Institutes for Health (5U01CA214846 and 5P50CA101942), and the Assistant Secretary of Defense Health Program, through the Breast Cancer Research Program (W81XWH-15-1-0013 to AC). Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the Department of Defense. This project has been made possible in part by grant number CZF2019-002443 from the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation.

- Vision*. (2014) 48:279–94. doi: 10.1007/s10851-013-0435-6
36. Levitin HM, Yuan J, Cheng YL, Ruiz FJ, Bush EC, Bruce JN, et al. *De novo* gene signature identification from single-cell RNA-seq with hierarchical poisson factorization. *Mol Syst Biol*. (2019) 15:e8557. doi: 10.15252/msb.20188557
 37. Durif G, Modolo L, Mold JE, Lambert-Lacroix S, Picard F. Probabilistic count matrix factorization for single cell expression data analysis. *Bioinformatics*. (2019) 35:4011–19. doi: 10.1093/bioinformatics/btz177
 38. Risso D, Perraudeau F, Gribkova S, Dudoit S, Vert JP. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat Commun*. (2018) 9:284. doi: 10.1038/s41467-017-02554-5
 39. Pierson E, Yau C. ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol*. (2015) 16:241. doi: 10.1186/s13059-015-0805-z
 40. Haghverdi L, Büttner F, Theis FJ. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics*. (2015) 31:2989–98. doi: 10.1093/bioinformatics/btv325
 41. Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol*. (2018) 37:38. doi: 10.1038/nbt.4314
 42. Linderman GC, Rachh M, Hoskins JG, Steinerberger S, Kluger Y. Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nat Methods*. (2019) 16:243–5. doi: 10.1038/s41592-018-0308-4
 43. Maaten L van der, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res*. (2008) 9:2579–605. Available online at: <http://www.jmlr.org/papers/v9/vandermaaten08a.html>
 44. Nguyen LH, Holmes S. Ten quick tips for effective dimensionality reduction. *PLoS Comput Biol*. (2019) 15:e1006907. doi: 10.1371/journal.pcbi.1006907
 45. Tsuyuzaki K, Sato H, Sato K, Nikaido I. Benchmarking principal component analysis for large-scale single-cell RNA-sequencing. *Genome Biol*. (2020) 21:9. doi: 10.1186/s13059-019-1900-3
 46. Meng C, Zeleznik OA, Thallinger GG, Kuster B, Gholami AM, Culhane AC. Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief Bioinform*. (2016) 17:628–41. doi: 10.1093/bib/bbv108
 47. Holmes S, Huber W. *Modern Statistics for Modern Biology*. Cambridge University Press (2018).
 48. Meng C, Kuster B, Culhane AC, Gholami AM. A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics*. (2014) 15:162. doi: 10.1186/1471-2105-15-162
 49. Franklin SB, Gibson DJ, Robertson PA, Pohlmann JT, Fralish JS. Parallel analysis: a method for determining significant principal components. *J Veg Sci*. (1995) 6:99–106. doi: 10.2307/3236261
 50. Meng C, Basunia A, Peters B, Gholami AM, Kuster B, Culhane AC. MOGSA: integrative single sample gene-set analysis of multiple omics data. *Mol Cell Proteomics*. (2019) 18(8 Suppl. 1):S153–68. doi: 10.1074/mcp.TIR118.001251
 51. Andrews TS, Hemberg M. Identifying cell populations with scRNA-seq. *Mol Aspects Med*. (2018) 59:114–22. doi: 10.1016/j.mam.2017.07.002
 52. Legendre P, Legendre L. *Numerical Ecology*. Amsterdam: Elsevier Science (1998).
 53. Diaconis P, Goel S, Holmes S. Horseshoes in multidimensional scaling and local kernel methods. *Ann Appl Stat*. (2008) 2:777–807. doi: 10.1214/08-AOAS165
 54. Holmes S, Huber W, editors. Multivariate methods for heterogeneous data. In: *Modern Statistics for Modern Biology*. Cambridge University Press. Available online at: <http://web.stanford.edu/class/bios221/book/Chap-MultivaHetero.html#exr:ex-KernelMethods> (accessed December 16, 2019).
 55. Lun ATL, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol*. (2016) 17:75. doi: 10.1186/s13059-016-0947-7
 56. Leek JT. Sva-seq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res*. (2014) 42:e161. doi: 10.1101/006585
 57. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*. (2012) 28:882–3. doi: 10.1093/bioinformatics/bts034
 58. Dray S, Chessel D, Thioulouse J. Procrustean co-inertia analysis for the linking of multivariate datasets. *Écoscience*. (2003) 10:110–19. doi: 10.1080/11956860.2003.11682757
 59. Gower JC. Generalized procrustes analysis. *Psychometrika*. (1975) 40:35–51. doi: 10.1007/BF02291478
 60. Hotelling H. Relations between two sets of variates. *Biometrika*. (1936) 28:321–77. doi: 10.2307/2333955
 61. Carroll JD. Generalization of canonical correlation analysis to three or more sets of variables. In: *Proceedings of the American Psychological Association*. San Francisco, CA (1968). p. 227–228. doi: 10.1037/e473742008-115
 62. Tenenhaus A, Tenenhaus M. Regularized generalized canonical correlation analysis for multiblock or multigroup data analysis. *Eur J Oper Res*. (2014) 238:391–403. doi: 10.1016/j.ejor.2014.01.008
 63. Lê Cao K-A, Martin PGP, Robert-Granié C, Besse P. Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics*. (2009) 10:34. doi: 10.1186/1471-2105-10-34
 64. Witten DM, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*. (2009) 10:515–34. doi: 10.1093/biostatistics/kxp008
 65. Tran HTN, Ang KS, Chevrier M, Zhang X, Lee NYS, Goh M, et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol*. (2020) 21:12. doi: 10.1186/s13059-019-1850-9
 66. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with harmony. *Nat Methods*. (2019) 16:1289–96. doi: 10.1038/s41592-019-0619-0
 67. Johansen N, Quon G. scAlign: a tool for alignment, integration, and rare cell identification from scRNA-seq data. *Genome Biol*. (2019) 20:166. doi: 10.1186/s13059-019-1766-4
 68. Welch J, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko E. Integrative inference of brain cell similarities and differences from single-cell genomics. *bioRxiv*. (2018) doi: 10.1101/459891
 69. Lopez R, Nazaret A, Langevin M, Samaran J, Regier J, Jordan MI, et al. A joint model of unpaired data from scRNA-seq and spatial transcriptomics for imputing missing gene expression measurements. *arXiv*. (2019) Available online at: <http://arxiv.org/abs/1905.02269> (accessed December 16, 2019).
 70. Dolédec S, Chessel D. Co-inertia analysis: an alternative method for studying species-environment relationships. *Freshw Biol*. (1994) 31:277–94. doi: 10.1111/j.1365-2427.1994.tb01741.x
 71. Escoufier B, Pagès J. Méthode pour l'analyse de plusieurs groupes de variables: application à la caractérisation des vins rouges du Val de Loire. *Revue de Statistique Appliquée* (1983) 31:43–59.
 72. Abdi H, Williams LJ, Valentin D. Multiple factor analysis: principal component analysis for multitable and multiblock data sets. *WIREs Comp Stat*. (2013) 5:149–79. doi: 10.1002/wics.1246
 73. Westerhuis JA, Kourti T, MacGregor JF. Analysis of multiblock and hierarchical PCA and PLS models. *J Chemometr*. (1998) 12:301–21. doi: 10.1002/(SICI)1099-128X(199809/10)12:5<301::AID-CEM515>3.0.CO;2-S
 74. Wold S, Hellberg S, Lundstedt T, Sjöström M, Wold H. PLS model building: Theory and application. PLS modeling with latent variables in two or more dimensions. In: *PLS Symposium*. Frankfurt.
 75. Meng C, Helm D, Frejno M, Kuster B. moCluster: identifying joint patterns across multiple omics data sets. *J Proteome Res*. (2016) 15:755–65. doi: 10.1021/acs.jproteome.5b00824
 76. Garali I, Adanyeguh IM, Ichou F, Perlberg V, Seyer A, Colsch B, et al. A strategy for multimodal data integration: application to biomarkers identification in spinocerebellar ataxia. *Brief Bioinform*. (2018) 19:1356–69. doi: 10.1093/bib/bbx060
 77. Su S, Tian L, Dong X, Hickey PF, Freytag S, Ritchie ME. CellBench: R/Bioconductor software for comparing single-cell RNA-seq analysis methods. *Bioinformatics*. (2020) 36:2288–90. doi: 10.1093/bioinformatics/btz889

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Hsu and Culhane. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Toward Systems Biomarkers of Response to Immune Checkpoint Blockers

Óscar Lapuente-Santana¹ and Federica Eduati^{1,2*}

¹ Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven, Netherlands, ² Institute for Complex Molecular Systems, Eindhoven University of Technology, Eindhoven, Netherlands

OPEN ACCESS

Edited by:

Enrica Calura,
University of Padova, Italy

Reviewed by:

Howard Donninger,
University of Louisville, United States
Mirjana Efremova,
Wellcome Sanger Institute (WT),
United Kingdom

*Correspondence:

Federica Eduati
f.eduati@tue.nl

Specialty section:

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

Received: 30 January 2020

Accepted: 22 May 2020

Published: 24 June 2020

Citation:

Lapuente-Santana Ó and Eduati F
(2020) Toward Systems Biomarkers of
Response to Immune Checkpoint
Blockers. *Front. Oncol.* 10:1027.
doi: 10.3389/fonc.2020.01027

Immunotherapy with checkpoint blockers (ICBs), aimed at unleashing the immune response toward tumor cells, has shown a great improvement in overall patient survival compared to standard therapy, but only in a subset of patients. While a number of recent studies have significantly improved our understanding of mechanisms playing an important role in the tumor microenvironment (TME), we still have an incomplete view of how the TME works as a whole. This hampers our ability to effectively predict the large heterogeneity of patients' response to ICBs. Systems approaches could overcome this limitation by adopting a holistic perspective to analyze the complexity of tumors. In this Mini Review, we focus on how an integrative view of the increasingly available multi-omics experimental data and computational approaches enables the definition of new systems-based predictive biomarkers. In particular, we will focus on three facets of the TME toward the definition of new systems biomarkers. First, we will review how different types of immune cells influence the efficacy of ICBs, not only in terms of their quantification, but also considering their localization and functional state. Second, we will focus on how different cells in the TME interact, analyzing how inter- and intra-cellular networks play an important role in shaping the immune response and are responsible for resistance to immunotherapy. Finally, we will describe the potential of looking at these networks as dynamic systems and how mathematical models can be used to study the rewiring of the complex interactions taking place in the TME.

Keywords: tumor microenvironment, precision immuno-oncology, multi-omics profiling, systems biology, predictive biomarkers, cancer signaling networks, immune checkpoint blockers

A CHANGE IN THE LANDSCAPE OF BIOMARKERS DISCOVERY

Tumor cells are able to activate several mechanisms to evade the immune response by disguising themselves as “self” cells. Binding to inhibitory checkpoint molecules (i.e., immune checkpoints) they can block antitumor activities of the immune system. Immunotherapy with immune checkpoint blockers (ICBs) uses antibodies to target immune checkpoints, such as PD1, PD-L1, and CTLA-4, unleashing the immune response. In clinical trials, ICB therapy has been shown to achieve durable therapeutic response and to increase patient survival in different cancer types, although still a small number of ICBs are FDA-approved (1, 2). Even if clinically approved, ICB therapy is

effective for a small subset of patients. Given the potential immunological toxicity (3, 4) and the elevated costs (>US\$100,000 per patient per year) (5) associated with ICBs, it is of paramount importance to be able to predict which patients will likely respond to the therapy, in order to administer the optimal treatment based on biomarkers.

The investigation of mechanisms supporting immune resistance has provided a great opportunity for biomarker discovery of patient response to ICBs (Figure 1). Two biomarkers have been clinically approved for PD-1/PD-L1 blockade therapy: the first is immunohistochemistry (IHC) staining of PD-L1 in non-small-cell lung cancer (NSCLC), melanoma, renal cell carcinoma (RCC), urothelial cancer, and triple-negative breast cancer (TNBC) (6); and the second is high microsatellite instability/defective mismatch repair (MSI-H/dMMR) regardless of tumor type (7, 8). Other emerging predictive biomarkers such as tumor mutational burden (TMB) (9, 10), signatures of a T cell inflamed tumor microenvironment (TME) either alone (10) or in combination (11), and neoantigen load (12–14) are still undergoing clinical trials. In addition, T cell receptor (TCR) diversity has been used as a biomarker to monitor the clonal expansion of T cells in breast cancer, glioma, cervical cancer, and leukemia/lymphoma (15–18). Further efforts both to exploit the utility of these biomarkers and to search for additional ones are still ongoing. For a complete review of these biomarkers and in which tumors they work, we refer to Havel et al. (19).

Despite being promising, these biomarkers also present some limitations. For instance, IHC enables measuring PD-L1 expressed on tumor cells, however the expression of this biomarker fluctuates over time and varies between different tumor sites. This variability undermines the ability to evaluate PD-1/PD-L1 therapies effectiveness based on IHC, as reviewed in Topalian et al. (20) and Camidge et al. (21). Another example is TMB, which is known to correlate imperfectly with clinical response (12, 13, 22). Neoantigen burden should partially overcome this issue, however most computational tools fail to estimate true neoantigens (19, 20, 23), and additional features should be considered to better determine neoantigen immunogenicity as reviewed in Finotello et al. (24).

Above-mentioned examples shed light upon the conceptual problem of looking only at individual components of the TME. While the characterization of different parts playing a role in the interaction between tumor and immune system has been essential to elucidate the most important actionable mechanisms, further research is required to define biomarkers harnessing a more coordinated joint action of these mechanisms. Predictive biomarkers for immunotherapy with ICBs have been extensively reviewed previously (19, 20, 23, 25). In this Mini Review we

focus on how a holistic profiling of the TME can provide new opportunities for identifying systems-based biomarkers built on existing synergies between the different individual components of the TME. Such a shift toward multifaceted strategies has been favored by increasingly available multi-omics data from bulk populations, individual cells, and imaging technologies (26), that can be integrated using computational approaches. In the following sections we will describe how biomarkers can be derived by considering three increasing levels of complexity. The first is the cellular component, focusing on the immune contexture of tumors, such as immune cells quantification, functionality, and localization. The second is the network of communication between and within cells of the TME. Finally, we will elaborate on how mathematical models can be used to take the dynamic nature of these networks into account.

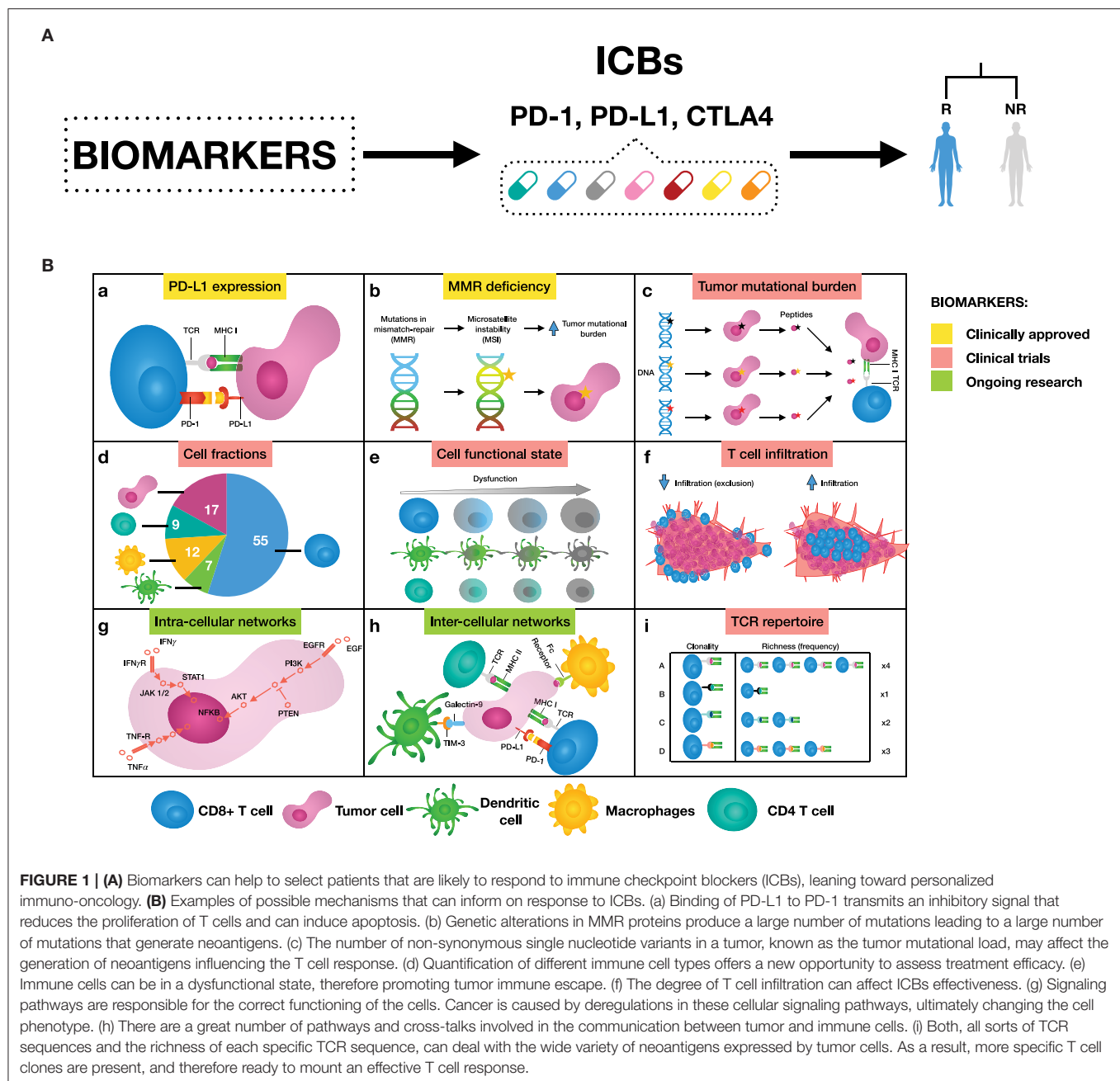
THE ROLE OF THE IMMUNE CONTEXTURE ON ICB EFFICACY

It is well-known that different types of immune cells can play a different role in the response to ICBs (27). For example, while the presence of CD8⁺ T cells within the TME is a good biomarker of ICBs efficacy, a high abundance of regulatory T (Treg) cells is generally associated with poor prognosis. Different tools have been developed to quantify tumor-infiltrating immune cells from bulk (RNA-seq) and single-cell (scRNA-seq) RNA sequencing measurements, as extensively reviewed in Finotello and Eduati (26) and Finotello and Trajanoski (28).

Apart from quantification of immune cells, their spatial localization also plays a pivotal role in the response to immunotherapy (29). For instance, CD8⁺ T cells not only need to be present, but also to be infiltrated (hot tumor) for the ICB therapy to work. In fact, pure quantification of CD8⁺ T cells is not always associated with favorable prognosis (30). Imaging techniques can be used to explore the spatial patterns of immune infiltration. A notable example of a biomarker assessing through IHC, both the abundance and the location (tumor center and invasive margin) of two lymphocyte populations (CD3⁺ and CD8⁺ T cells) is the immunoscore (31), that was shown to accurately predict patient survival in colorectal cancer patients (32). More recently, spatial information of CD8⁺ T cells from IHC was integrated with transcriptomics data to study the effect of lymphocyte infiltration in patients with TNBC, providing predictive biomarkers of ICBs response (33). Automatic approaches for image analysis could reveal useful in the future for high-throughput identification of spatial biomarkers. A first attempt in this direction was the development of tumor infiltrating lymphocytes maps by using deep learning on images from the cancer genome atlas (TCGA) (34).

Another important factor that affects patients' response to ICBs is the functional state of the different immune cells (35). Dysfunctional states of T cells can be characterized from bulk and single-cell RNA-seq (36–38) and epigenetic profiling (39–41). ICBs aim at rescuing dysfunctional T cells, therefore the investigation of their functional state can inform on ICBs therapy success and limitations (36–39, 41). Depending on the type of

Abbreviations: CTLA-4, cytotoxic T lymphocyte antigen 4; DC, dendritic cell; ICB, immune checkpoint blocker; IFN γ , interferon gamma; IHC, immunohistochemistry; MMR, DNA mismatch repair; MSI-H, high microsatellite instability; NOS2, nitric oxide synthase 2; NSCLC, non-small-cell lung cancer; PD-L1, programmed cell death-ligand 1; PD-1, programmed cell death protein 1; RCC, renal cell carcinoma; RNA-seq, RNA sequencing; scRNA-seq, single-cell RNA sequencing; TCGA, the cancer genome atlas; TCR, T cell receptor; TMB, tumor mutational burden; TME, tumor microenvironment; TNBC, triple-negative breast cancer; TNF, tumor necrosis factor; Treg, regulatory T cell.



stimulatory signal, macrophages (42, 43), and B cells (44, 45) can develop into functional subsets that have either positive or negative effects on tumors. Another example are dendritic cells (DCs), that normally control cancer antigen presentation, priming and activation of T cell responses, however the TME can compromise their ability to stimulate the immune response (46, 47). Certain computational tools for cell-type quantification can also unmask the phenotypic state of cell subpopulations in the TME by inferring the transcriptomics profiles of individual cells (48, 49). A promising research direction for biomarkers discovery is also given by new technologies that allow generation

of omics data from tissue slides preserving cell spatial identity (50, 51). These approaches would result in combined localization and characterization of the cells in the TME.

Analysis on the immune infiltrate quantification, functionality, and localization can help both to explain the diversity of the tumor immune milieu and develop informative biomarkers for ICBs (27, 52, 53). Pointing in this direction, different efforts have recently explored the use of bulk transcriptomics data to derive more complex immune-related scores to assess the likelihood of a patient to respond to ICBs (38, 54–63).

INTRA- AND INTER-CELLULAR NETWORKS ORCHESTRATE THE IMMUNE RESPONSE

The functional state of cells in the TME is defined by a complex system of communication between molecules within the cells (intra-cellular networks) and among different cells (inter-cellular networks). Looking at intra- and inter-cellular networks can provide a more holistic perspective of the TME and inform a new class of biomarkers for immunotherapy and its potential combination with other targeted therapies (64).

Intra-cellular signaling pathways play a part in shaping the interaction with the immune system [(65, 66); **Figure 2**]. Abnormalities in tumor-intrinsic signaling, involving oncogenes and tumor suppressor genes, have been associated with mechanisms of inherent immune resistance (67). Examples are PTEN loss (68) or EGFR gain of function (69), both causing PI3K-Akt pathway activation and leading to over-expression of PD-L1 and consequent immunoresistance. Due to the complexity of signaling pathways, with numerous cross-talks and feedback loops, the adoption of individual oncogenic drivers as biomarkers is not expected to be effective in most cases (20). In fact, PD-L1 signal is directly regulated by numerous oncogenic pathways such as Ras, mTOR, EGFR, MEK, ERK, and MAPK (70). Besides pathways regulating immune checkpoints, other signaling cross-talks control the immune response from different perspectives, like inactivation of TP53 or activation of β -catenin pathway, both reducing chemokine production by tumor cells and thereby reducing recruitment of immune cells into the TME (71, 72).

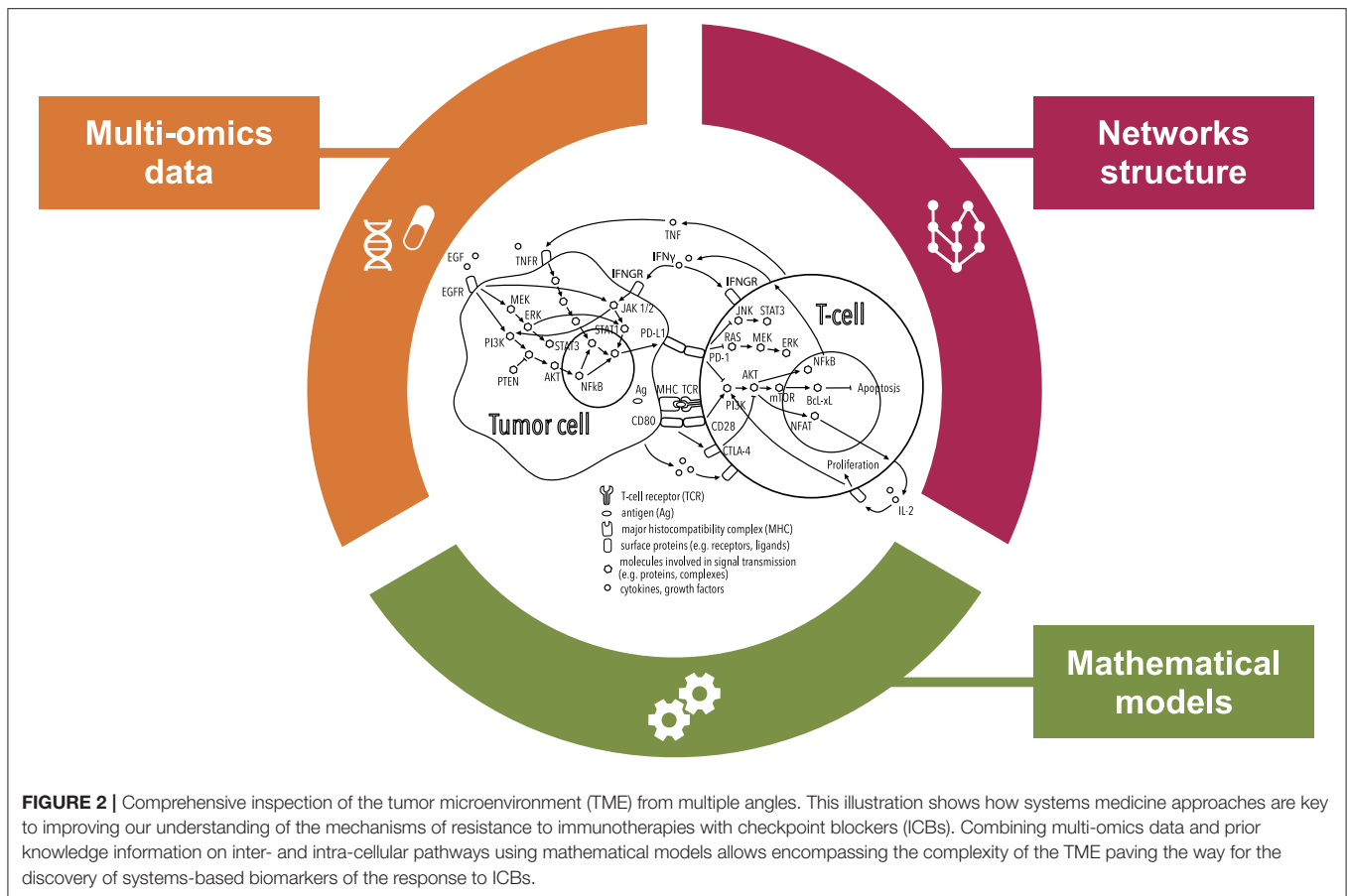
In addition, cancer cells receive signals from other cells in the TME through ligand-receptor interactions. These inter-cellular communications lead to changes in the phenotype of the regulated cells thus playing an important role in both progression and prognosis of cancer (73, 74). An example is the response elicited on cancer cells by two cytokines (TNF and IFN γ) produced by activated T cells. These cytokines induce PD-L1 expression through JAK-STAT and NF- κ B signaling, inducing acquired resistance to the immune response (75, 76). Another study identified a relationship between high expression of NOS2 and prolonged IFN signaling in tumors resistant to PD-1 blockade (77).

While collections of intra- (78) and inter-cellular (79) interactions can be derived from literature and databases, additional data are required to characterize the networks for each patient or group of patients. Transcriptomics and proteomics data can provide the basis to study intra- and inter-cellular signaling networks. Imaging data can also be integrated to improve our understanding on spatial localization of interacting cells. Computational methods have been developed to infer integrated inter- and intra-cellular networks from bulk (80, 81) and single-cell (81, 82) RNA-seq data. These tools could be exploited to derive biomarkers for immunotherapy by studying the functional effect of cell-cell communication. In a recent study, a curated database of ligand-receptor interactions (79) was integrated with gene expression data to

deconvolute the transcriptional profile of cancer and stromal cells and infer cross-talks in the TME (83). Interestingly, the authors show that for different cancer types, PD-L1 expression is higher on cancer or stromal cells which nicely correlates with the general responsiveness to immunotherapy. Further research is required to assess if this holds also for individual patients, making it potentially a more effective biomarker than bulk PD-L1 expression. In another recent publication (84), researchers performed an extensive literature curation to derive a comprehensive signaling network of innate immune response in cancer, including cell type-specific signaling in macrophages, DCs, myeloid-derived suppressor cells, and natural killer cells. Such network was then integrated with scRNAseq data from macrophages and natural killer cells in melanoma to study the heterogeneity of innate immune cell types and could potentially be used to predict patient survival and response to immunotherapies. Finally, Worzfeld et al. combined parallel bulk transcriptomics and proteomics data on tumor cell spheroids, tumor-associated T cells and macrophages to derive inter-cellular signaling networks in the ovarian cancer microenvironment (85). Such networks included several immune checkpoint regulators and appeared to have potential clinical relevance. Overall, these studies have demonstrated the enormous benefit that holistic approaches combining complex multicellular networks can bring into the immuno-oncology field, and we expect that in the forthcoming future more research efforts will be spent in this direction. The recent developments of 3D cell culture models resembling the TME, are expected to be a powerful tool for further *in vitro* and *ex vivo* investigation of intra-cellular communication, and to study their effect on the response to ICBs (86).

THE POTENTIAL OF LOOKING AT THE DYNAMICITY AND PLASTICITY OF THE TME

It is well-known that the cellular functional state changes dynamically in response to environmental changes and perturbations such as drug treatment (87, 88), calling for identification of the dynamic properties of the networks. The ideal data for dynamic functional characterization of the system's response are obtained upon perturbation (89). Functional screening of the effect of cancer drugs has been so far focused on cancer cell lines. While cell lines are a debatable model system, they proved to be a valuable tool to explore novel biomarkers of drug response (90, 91). High-throughput drug screening studies are now also being increasingly performed on organoids (92) or other 3D experimental models (86), which are more physiological human cancer models of the TME. These efforts open new ways for pre-clinical investigation of the effect of immunotherapy. Finally, more recent technologies allow screening also of patient biopsies without need for culturing steps (93–95) paving the way for functional characterization



of *ex vivo* tumor samples potentially improving personalized cancer treatment.

To capture the functional context of the immune response, statistical, and mathematical approaches are developing into more compendious methods that integrate multi-omics data and prior knowledge on network structure (Figure 2). While mathematical models do not fall into the standard definition of biomarkers, they can provide predictions of response to immunotherapy. Additionally they can be used to define dynamic biomarkers based on properties of the modeled system, as opposed to static biomarkers that only consider the initial conditions of the system (88).

Dynamic mathematical models can be used to study intra-cellular networks of the different cell types populating the TME (96). To characterize these networks at the patient-specific level, models of signaling pathways in cancer cells have been trained from perturbation experiments (97, 98), gene expression data (99), or integrating multi-omics data (100). The resulting parameters corresponding to these personalized models can be relevant biomarkers of clinical outcome (99–101). Mathematical models have also been used to study intra-cellular signaling in T cells. This includes the investigation of how PD-1 leads to deactivation of the T cell receptor signaling (102) or mechanistic understanding of T cell exhaustion (103). PD-1 is one of the main targets of ICB, and exhausted T cells have a higher number of

targetable checkpoint proteins like PD-1 and CTLA-4, therefore the investigation of these aspects could be relevant to identify possible biomarkers.

More studies are now focusing on mathematical models incorporating inter-cellular interactions to better capture the complexity of the TME. Agent-based models can be used to simulate the interactions between cells in the tumor microenvironment seen as a 2D or a 3D grid (104). Each cell is seen as an agent that can perform different tasks with a certain probability (e.g., cells can non-proliferate, divide, or die). Since the immune response can be seen as a probabilistic outcome of a complex system (88), agent-based models are an adequate mathematical approximation to capture this stochasticity. These models can be refined using a multitude of data types and used to simulate the effect of immunotherapy (105, 106), providing a variety of possible outcomes given the same initial conditions that can be interpreted as probability of success. It has been shown that tumor-bearing inbred mice, which have only minimal differences, can respond differently to immunotherapy (88), therefore having models that can incorporate stochasticity provides an interesting approximation of the *in vivo* situation. Another approach to model cell-cell communication is by using response-time modeling (107), where cells are modeled as a black-box that can receive inputs (e.g., cytokines) from other cells, process them, and change state (e.g., immune cells

can switch between inactive and active) accordingly with a certain probability. Recently, Grandclaudon et al. combined perturbation data with a multivariate quantitative model to study context dependent interactions between DCs and helper T cells (108). A different approach based on quantitative systems pharmacology was recently used to simulate the effect of ICB therapy in metastatic breast cancer patients using a four compartments (central, peripheral, tumor-draining lymph node, and tumor) model (109).

Additionally, combining mathematical models with longitudinal data, i.e., data collected at different time points, can be used to investigate the evolutionary dynamics of treatment response. This aspect is particularly relevant, especially to be able to distinguish at an early stage real tumor progression (patient should be assigned to a different treatment) from what is called pseudoprogression, i.e., temporary progression followed by a response to the treatment (patient should be kept on ICB). The latter behavior has been described using a model of immune activation incorporating the dynamics of antigen presentation (110). Based on a system of three ordinary differential equations to describe the interaction between tumor cells, Treg cells, and cytotoxic T cells, this model could explain why, in response to ICBs, the tumor can worsen before starting regressing. Other multi-cellular models have been used to derive *in silico* patients to test different possible dynamics of treatment response (111, 112), that could be compared with longitudinal measurements of tumor load from PET/CT imaging (112). Longitudinal data are often limited to non-invasive imaging and, in a few cases, to transcriptomics, IHC, TCR, and genome sequencing data (113, 114) for a limited number of time points due to invasiveness of biopsies. Computational modeling of longitudinal data is still at its infancy, but we envision that in the future more mechanistic dynamic models will be able to exploit this type of data for definition of dynamic biomarkers.

CONCLUSIONS AND FUTURE PERSPECTIVES

Current limitations in identifying predictive biomarkers for ICB therapy are partially due to overlooking the complexity

of the TME. Following the advancements in technologies to measure multi-omics data, measurements of bulk populations, individual cells, and spatial information have paved the way to a more comprehensive view of the TME. Recent efforts are focused on searching for signatures of response to ICBs that consider quantification, localization, and functionality of different immune cells in the TME, showing improved predictive power with respect to simpler biomarkers (115). However, they still miss an integrative strategy that takes a view of the whole TME, rather than examining each factor in isolation. In this respect, mechanistic models incorporating existing biological basis, e.g., on intra- and inter-cellular pathways, can accompany both therapy and biomarker development in immuno-oncology (116).

There is compelling evidence that the interplay of the immune system, tumors, organs, and external environment, harmonizes antitumor immune responses (117). Therefore, we envision that novel systems medicine approaches entailing mathematical models can gradually build up a profile of the TME, both in the lab and, more importantly, in the clinic. To this end, building patient specific models have become of increasing importance, especially when based on data that can be measured in clinical settings. Moreover, systems approaches can especially be useful to provide rationale for alternative personalized treatments such as combinatorial therapy.

AUTHOR CONTRIBUTIONS

ÓL-S and FE wrote and edited the article. All authors contributed to the article and approved the submitted version.

FUNDING

Costs related to this publication are covered by the Computational Biology group of the Department of Biomedical Engineering of Eindhoven University of Technology.

ACKNOWLEDGMENTS

The authors would like to thank the members of the CBio group for their helpful proofreading suggestions.

REFERENCES

- Hargadon KM, Johnson CE, Williams CJ. Immune checkpoint blockade therapy for cancer: an overview of FDA-approved immune checkpoint inhibitors. *Int Immunopharmacol.* (2018) 62:29–39. doi: 10.1016/j.intimp.2018.06.001
- Sun C, Mezzadra R, Schumacher TN. Regulation and function of the PD-L1 checkpoint. *Immunity.* (2018) 48:434–52. doi: 10.1016/j.immuni.2018.03.014
- Boutros C, Tarhini A, Routier E, Lambotte O, Ladurie FL, Carbone F, et al. Safety profiles of anti-CTLA-4 and anti-PD-1 antibodies alone and in combination. *Nat Rev Clin Oncol.* (2016) 13:473–86. doi: 10.1038/nrclinonc.2016.58
- Postow MA, Sidlow R, Hellmann MD. Immune-related adverse events associated with immune checkpoint blockade. *New Engl J Med.* (2018) 378:158–68. doi: 10.1056/NEJMra1703481
- Schmidt C. The benefits of immunotherapy combinations. *Nature.* (2017) 552:S67–9. doi: 10.1038/d41586-017-08702-7
- Arora S, Velichinskii R, Lesh RW, Ali U, Kubiak M, Bansal P, et al. Existing and emerging biomarkers for immune checkpoint immunotherapy in solid tumors. *Adv Ther.* (2019) 36:2638–78. doi: 10.1007/s12325-019-01051-z
- Le DT, Uram JN, Wang H, Bartlett B, Kemberling H, Eyring A, et al. PD-1 blockade in tumors with mismatch repair deficiency. *N Engl J Med.* (2015) 372:2509–20. doi: 10.1056/NEJMoa1500596
- Le DT, Durham JN, Smith KN, Wang H, Bartlett BR, Aulakh LK, et al. Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science.* (2017) 357:409–13. doi: 10.1126/science.aan6733
- Legrand FA, Gandara DR, Mariathasan S, Powles T, He X, Zhang W, et al. Association of high tissue TMB and atezolizumab efficacy across multiple tumor types. *J Clin Oncol.* (2018) 36(Suppl. 15):12000. doi: 10.1200/JCO.2018.36.15_suppl.12000

10. Ott PA, Bang Y-J, Piha-Paul SA, Abdul Razak AR, Bennouna J, Soria J-C, et al. T-cell-inflamed gene-expression profile, programmed death ligand 1 expression, and tumor mutational burden predict efficacy in patients treated with pembrolizumab across 20 cancers: KEYNOTE-028. *J Clin Oncol.* (2019) 37:318–27. doi: 10.1200/JCO.2018.78.2276
11. Cristescu R, Mogg R, Ayers M, Albright A, Murphy E, Yearley J, et al. Pan-tumor genomic biomarkers for PD-1 checkpoint blockade-based immunotherapy. *Science.* (2018) 362:eaar3593. doi: 10.1126/science.aar3593
12. Rizvi NA, Hellmann MD, Snyder A, Kvistborg P, Makarov V, Havel JJ, et al. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science.* (2015) 348:124–8. doi: 10.1126/science.aaa1348
13. Van Allen EM, Miao D, Schilling B, Shukla SA, Blank C, Zimmer L, et al. Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science.* (2015) 350:207–11. doi: 10.1126/science.aad0095
14. Lu T, Wang S, Xu L, Zhou Q, Singla N, Gao J, et al. Tumor neoantigenicity assessment with CSin score incorporates clonality and immunogenicity to predict immunotherapy outcomes. *Sci Immunol.* (2020) 5:aaz3199. doi: 10.1126/sciimmunol.aaz3199
15. Page DB, Yuan J, Redmond D, Wen YH, Durack JC, Emerson R, et al. Deep sequencing of T-cell receptor DNA as a biomarker of clonally expanded TILs in breast cancer after immunotherapy. *Cancer Immunol Res.* (2016) 4:835–44. doi: 10.1158/2326-6066.CIR-16-0013
16. Sims JS, Grinshpun B, Feng Y, Ung TH, Neira JA, Samanamud JL, et al. Diversity and divergence of the glioma-infiltrating T-cell receptor repertoire. *Proc Natl Acad Sci USA.* (2016) 113:E3529–37. doi: 10.1073/pnas.1601012113
17. Cui JH, Lin KR, Yuan SH, Jin YB, Chen XP, Su XK, et al. TCR Repertoire as a novel indicator for immune monitoring and prognosis assessment of patients with cervical cancer. *Front Immunol.* (2018) 9:2729. doi: 10.3389/fimmu.2018.02729
18. Farmanbar A, Kneller R, Firouzi S. RNA sequencing identifies clonal structure of T-cell repertoires in patients with adult T-cell leukemia/lymphoma. *NPJ Genom Med.* (2019) 4:10. doi: 10.1038/s41525-019-0084-9
19. Havel JJ, Chowell D, Chan TA. The evolving landscape of biomarkers for checkpoint inhibitor immunotherapy. *Nat Rev Cancer.* (2019) 19:133–50. doi: 10.1038/s41568-019-0116-x
20. Topalian SL, Taube JM, Anders RA, Pardoll DM. Mechanism-driven biomarkers to guide immune checkpoint blockade in cancer therapy. *Nat Rev Cancer.* (2016) 16:275–87. doi: 10.1038/nrc.2016.36
21. Camidge DR, Doebele RC, Kerr KM. Comparing and contrasting predictive biomarkers for immunotherapy and targeted therapy of NSCLC. *Nat Rev Clin Oncol.* (2019) 16:341–55. doi: 10.1038/s41571-019-0173-9
22. Snyder A, Makarov V, Merghoub T, Yuan J, Zaretsky JM, Desrichard A, et al. Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N Engl J Med.* (2014) 371:2189–99. doi: 10.1056/NEJMoa1406498
23. Keenan TE, Burke KP, Van Allen EM. Genomic correlates of response to immune checkpoint blockade. *Nat Med.* (2019) 25:389–402. doi: 10.1038/s41591-019-0382-x
24. Finotello F, Rieder D, Hackl H, Trajanoski Z. Next-generation computational tools for interrogating cancer immunity. *Nat Rev Genet.* (2019) 20:724–46. doi: 10.1038/s41576-019-0166-7
25. Duffy MJ, Crown J. Biomarkers for predicting response to immunotherapy with immune checkpoint inhibitors in cancer patients. *Clin Chem.* (2019) 65:1228–38. doi: 10.1373/clinchem.2019.303644
26. Finotello F, Eduati F. Multi-omics profiling of the tumor microenvironment: paving the way to precision immuno-oncology. *Front Oncol.* (2018) 8:430. doi: 10.3389/fonc.2018.00430
27. Fridman WH, Zitvogel L, Sautès-Fridman C, Kroemer G. The immune contexture in cancer prognosis and treatment. *Nat Rev Clin Oncol.* (2017) 14:717–34. doi: 10.1038/nrclinonc.2017.101
28. Finotello F, Trajanoski Z. Quantifying tumor-infiltrating immune cells from transcriptomics data. *Cancer Immunol Immunother.* (2018) 67:1031–40. doi: 10.1007/s00262-018-2150-z
29. Galon J, Bruni D. Approaches to treat immune hot, altered and cold tumours with combination immunotherapies. *Nat Rev Drug Discov.* (2019) 18:197–218. doi: 10.1038/s41573-018-0007-y
30. Giraldo NA, Becht E, Pagès F, Skliris G, Verkarre V, Vano Y, et al. Orchestration and prognostic significance of immune checkpoints in the microenvironment of primary and metastatic renal cell cancer. *Clin Cancer Res.* (2015) 21:3031–40. doi: 10.1158/1078-0432.CCR-14-2926
31. Galon J, Costes A, Sanchez-Cabo F, Kirilovsky A, Mlecnik B, Lagorce-Pagès C, et al. Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science.* (2006) 313:1960–64. doi: 10.1126/science.1129139
32. Pagès F, Mlecnik B, Marliot F, Bindea G, Ou F-S, Bifulco C, et al. International validation of the consensus immunoscore for the classification of colon cancer: a prognostic and accuracy study. *Lancet.* (2018) 391:2128–39. doi: 10.1016/S0140-6736(18)30789-X
33. Gruosso T, Gigoux M, Manem VSK, Bertos N, Zuo D, Perlitch I, et al. Spatially distinct tumor immune microenvironments stratify triple-negative breast cancers. *J Clin Invest.* (2019) 129:1785–1800. doi: 10.1172/JCI96313
34. Saltz J, Gupta R, Hou L, Kurc T, Singh P, Nguyen V, et al. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell Rep.* (2018) 23:181–93.e7. doi: 10.1016/j.celrep.2018.03.086
35. Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang T-H, et al. The immune landscape of cancer. *Immunity.* (2019) 51:411–12. doi: 10.1016/j.immuni.2019.08.004
36. Sade-Feldman M, Yizhak K, Bjorgaard SL, Ray JP, de Boer CG, Jenkins RW, et al. Defining T cell states associated with response to checkpoint immunotherapy in melanoma. *Cell.* (2018) 175:998–1013.e20. doi: 10.1016/j.cell.2018.10.038
37. Zheng C, Zheng L, Yoo J-K, Guo H, Zhang Y, Guo X, et al. Landscape of infiltrating T cells in liver cancer revealed by single-cell sequencing. *Cell.* (2017) 169:1342–56.e16. doi: 10.1016/j.cell.2017.05.035
38. Jiang P, Gu S, Pan D, Fu J, Sahu A, Hu X, et al. Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. *Nat Med.* (2018) 24:1550–58. doi: 10.1038/s41591-018-0136-1
39. Pauken KE, Sammons MA, Odorizzi PM, Manne S, Godec J, Khan O, et al. Epigenetic stability of exhausted T cells limits durability of reinvigoration by PD-1 blockade. *Science.* (2016) 354:1160–65. doi: 10.1126/science.12807
40. Sen DR, Kaminski J, Barnitz RA, Kurachi M, Gerdemann U, Yates KB, et al. The epigenetic landscape of T cell exhaustion. *Science.* (2016) 354:1165–9. doi: 10.1126/science.aae0491
41. Philip M, Fairchild L, Sun L, Horste EL, Camara S, Shakiba M, et al. Chromatin states define tumour-specific T cell dysfunction and reprogramming. *Nature.* (2017) 545:452–6. doi: 10.1038/nature22367
42. Quaranta V, Rainer C, Nielsen SR, Raymant ML, Ahmed MS, Engle DD, et al. Macrophage-derived granulins drives resistance to immune checkpoint inhibition in metastatic pancreatic cancer. *Cancer Res.* (2018) 78:4253–69. doi: 10.1158/0008-5472.CAN-17-3876
43. Klug F, Prakash H, Huber PE, Seibel T, Bender N, Halama N, et al. Low-dose irradiation programs macrophage differentiation to an iNOS^{hi}/M1 phenotype that orchestrates effective T cell immunotherapy. *Cancer Cell.* (2013) 24:589–602. doi: 10.1016/j.ccr.2013.09.014
44. Lindner S, Dahlke K, Sontheimer K, Hagn M, Kaltenmeier C, Barth TFE, et al. Interleukin 21-induced granzyme B-expressing B cells infiltrate tumors and regulate T cells. *Cancer Res.* (2013) 73:2468–79. doi: 10.1158/0008-5472.CAN-12-3450
45. Selitsky SR, Mose LE, Smith CC, Chai S, Hoadley KA, Dittmer DP, et al. Prognostic value of B cells in cutaneous melanoma. *Genome Med.* (2019) 11:36. doi: 10.1186/s13073-019-0647-5
46. Tesone AJ, Rutkowski MR, Brencicova E, Svoronos N, Perales-Puchalt A, Stephen TL, et al. Satb1 overexpression drives tumor-promoting activities in cancer-associated dendritic cells. *Cell Rep.* (2016) 14:1774–86. doi: 10.1016/j.celrep.2016.01.056
47. Fuertes MB, Kacha AK, Kline J, Woo S-R, Kranz DM, Murphy KM, et al. Host type I IFN signals are required for antitumor CD8⁺ T cell responses through CD8 α dendritic cells. *J Exp Med.* (2011) 208:2005–16. doi: 10.1084/jem.20101159
48. Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat Biotechnol.* (2019) 37:773–82. doi: 10.1038/s41587-019-0114-2

49. Zaitsev K, Bambouskova M, Swain A, Artyomov MN. Complete deconvolution of cellular mixtures based on linearity of transcriptional signatures. *Nat Commun.* (2019) 10:2209. doi: 10.1038/s41467-019-09990-5
50. Giesen C, Wang HAO, Schapiro D, Zivanovic N, Jacobs A, Hattendorf B, et al. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat Methods.* (2014) 11:417–22. doi: 10.1038/nmeth.2869
51. Angelo M, Bendall SC, Finck R, Hale MB, Hitzman C, Borowsky AD, et al. Multiplexed ion beam imaging of human breast tumors. *Nat Med.* (2014) 20:436–42. doi: 10.1038/nm.3488
52. Galluzzi L, Chan TA, Kroemer G, Wolchok JD, López-Soto A. The hallmarks of successful anticancer immunotherapy. *Sci Transl Med.* (2018) 10:eaat7807. doi: 10.1126/scitranslmed.aat7807
53. Binnewies M, Roberts EW, Kersten K, Chan V, Fearon DF, Merad M, et al. Understanding the tumor immune microenvironment (TIME) for effective therapy. *Nat Med.* (2018) 24:541–50. doi: 10.1038/s41591-018-0014-x
54. Auslander N, Zhang G, Lee JS, Frederick DT, Miao B, Moll T, et al. Publisher Correction: Robust prediction of response to immune checkpoint blockade therapy in metastatic melanoma. *Nat Med.* (2018) 24:1942. doi: 10.1038/s41591-018-0247-8
55. Davoli T, Uno H, Wooten EC, Elledge SJ. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *Science.* (2017) 355:eaa8399. doi: 10.1126/science.aaf8399
56. Fehrenbacher L, Spira A, Ballinger M, Kowanzet M, Vansteenkiste J, Mazieres J, et al. Atezolizumab versus docetaxel for patients with previously treated non-small-cell lung cancer (POPLAR): a multicentre, open-label, phase 2 randomised controlled trial. *Lancet.* (2016) 387:1837–46. doi: 10.1016/S0140-6736(16)00587-0
57. Huang AC, Orlowski RJ, Xu X, Mick R, George SM, Yan PK, et al. A single dose of neoadjuvant PD-1 blockade predicts clinical outcomes in resectable melanoma. *Nat Med.* (2019) 25:454–61. doi: 10.1038/s41591-019-0357-y
58. Messina JL, Fenstermacher DA, Eschrich S, Qu X, Berglund AE, Lloyd MC, et al. 12-Chemokine gene signature identifies lymph node-like structures in melanoma: potential for patient selection for immunotherapy? *Sci Rep.* (2012) 2:765. doi: 10.1038/srep00765
59. Roh W, Chen P-L, Reuben A, Spencer CN, Prieto PA, Miller JB, et al. Integrated molecular analysis of tumor biopsies on sequential CTLA-4 and PD-1 blockade reveals markers of response and resistance. *Sci Transl Med.* (2017) 9:eaa83560. doi: 10.1126/scitranslmed.aah3560
60. Rooney MS, Shukla SA, Wu CJ, Getz G, Hacohen N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell.* (2015) 160:48–61. doi: 10.1016/j.cell.2014.12.033
61. Ayers M, Luncford J, Nebozhyn M, Murphy E, Loboda A, Kaufman DR, et al. IFN- γ -related mRNA profile predicts clinical response to PD-1 blockade. *J Clin Invest.* (2017) 127:2930–40. doi: 10.1172/JCI91190
62. Ock C-Y, Hwang J-E, Keam B, Kim S-B, Shim J-J, Jang H-J, et al. Genomic landscape associated with potential response to anti-CTLA-4 treatment in cancers. *Nat Commun.* (2017) 8:1050. doi: 10.1038/s41467-017-01018-0
63. Charoentong P, Finotello F, Angelova M, Mayer C, Efremova M, Rieder D, et al. Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. *Cell Rep.* (2017) 18:248–62. doi: 10.1016/j.celrep.2016.12.019
64. Szeto GL, Finley SD. Integrative approaches to cancer immunotherapy. *Trends Cancer Res.* (2019) 5:400–10. doi: 10.1016/j.trecan.2019.05.010
65. Wellenstein MD, de Visser KE. Cancer-cell-intrinsic mechanisms shaping the tumor immune landscape. *Immunity.* (2018) 48:399–416. doi: 10.1016/j.immuni.2018.03.004
66. Spranger S, Gajewski TF. Impact of oncogenic pathways on evasion of antitumor immune responses. *Nat Rev Cancer.* (2018) 18:139–47. doi: 10.1038/nrc.2017.117
67. Sharma P, Hu-Lieskovan S, Wargo JA, Ribas A. Primary, adaptive, and acquired resistance to cancer immunotherapy. *Cell.* (2017) 168:707–23. doi: 10.1016/j.cell.2017.01.017
68. Peng W, Chen JQ, Liu C, Malu S, Creasy C, Tetzlaff MT, et al. Loss of PTEN promotes resistance to T cell-mediated immunotherapy. *Cancer Discov.* (2016) 6:202–16. doi: 10.1158/1538-7445.AM2016-4363
69. Akbay EA, Koyama S, Carretero J, Altobelli A, Tchaicha JH, Christensen CL, et al. Activation of the PD-1 pathway contributes to immune escape in EGFR-driven lung tumors. *Cancer Discov.* (2013) 3:1355–63. doi: 10.1158/2159-8290.CD-13-0310
70. Escors D, Gato-Cañás M, Zuazo M, Arasanz H, García-Granda MJ, Vera R, et al. The intracellular signalosome of PD-L1 in cancer cells. *Signal Transduct Target Ther.* (2018) 3:26. doi: 10.1038/s41392-018-0022-9
71. Quigley D, Silwal-Pandit L, Dannenfelser R, Langerød A, Vollen HKM, Vaske C, et al. Lymphocyte invasion in IC10/basal-like breast tumors is associated with wild-type TP53. *Mol Cancer Res.* (2015) 13:493–501. doi: 10.1158/1541-7786.MCR-14-0387
72. Spranger S, Bao R, Gajewski TF. Melanoma-intrinsic β -catenin signalling prevents anti-tumour immunity. *Nature.* (2015) 523:231–5. doi: 10.1038/nature14404
73. Maman S, Witz IP. A history of exploring cancer in context. *Nat Rev Cancer.* (2018) 18:359–76. doi: 10.1038/s41568-018-0006-7
74. Altan-Bonnet G, Mukherjee R. Cytokine-mediated communication: a quantitative appraisal of immune complexity. *Nat Rev Immunol.* (2019) 19:205–17. doi: 10.1038/s41577-019-0131-x
75. Zaretsky JM, Garcia-Diaz A, Shin DS, Escuin-Ordinas H, Hugo W, Hu-Lieskovan S, et al. Mutations associated with acquired resistance to PD-1 blockade in melanoma. *N Engl J Med.* (2016) 375:819–29. doi: 10.1056/NEJMoa1604958
76. Bertrand F, Montfort A, Marcheteau E, Imbert C, Gilhodes J, Filleron T, et al. TNF α blockade overcomes resistance to anti-PD-1 in experimental melanoma. *Nat Commun.* (2017) 8:2256. doi: 10.1038/s41467-017-02358-7
77. Jacquelinot N, Yamazaki T, Roberti MP, Duong CPM, Andrews MC, Verlingue L, et al. Sustained type I interferon signaling as a mechanism of resistance to PD-1 blockade. *Cell Res.* (2019) 29:846–61. doi: 10.1038/s41422-019-0224-x
78. Türei D, Korcsmáros T, Saez-Rodriguez J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat Methods.* (2016) 13:966–67. doi: 10.1038/nmeth.4077
79. Ramilowski JA, Goldberg T, Harshbarger J, Kloppmann E, Lizio M, Satagopam VP, et al. A draft network of ligand-receptor-mediated multicellular signalling in human. *Nat Commun.* (2015) 6:7866. doi: 10.1038/ncomms8866
80. Choi H, Sheng J, Gao D, Li F, Durrans A, Ryu S, et al. Transcriptome analysis of individual stromal cell populations identifies stroma-tumor crosstalk in mouse lung cancer model. *Cell Rep.* (2015) 10:1187–201. doi: 10.1016/j.celrep.2015.01.040
81. Browaeys R, Saelens W, Saeys Y. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat Methods.* (2020) 17:159–62. doi: 10.1038/s41592-019-0667-5
82. Wang S, Karikomi M, MacLean AL, Nie Q. Cell lineage and communication network inference via optimization for single-cell transcriptomics. *Nucleic Acids Res.* (2019) 47:e66. doi: 10.1093/nar/gky882
83. Ghoshdastider U, Naeini MM, Rohatgi N, Revkov E. Data-driven inference of crosstalk in the tumor microenvironment. *BioRxiv.* (2019). doi: 10.1101/835512
84. Kondratova M, Czerwinski U, Sompairac N, Amigorena SD, Soumelis V, Barillot E, et al. A multiscale signalling network map of innate immune response in cancer reveals cell heterogeneity signatures. *Nat Commun.* (2019) 10:4808. doi: 10.1038/s41467-019-12270-x
85. Worzfeld T, Finkernagel F, Reinartz S, Konzer A, Adhikary T, Nist A, et al. Proteotranscriptomics reveal signaling networks in the ovarian cancer microenvironment. *Mol Cell Proteomics.* (2018) 17:270–89. doi: 10.1074/mcp.RA117.000400
86. Modugno FD, Di Modugno F, Colosi C, Trono P, Antonacci G, Ruocco G, et al. 3D models in the new era of immune oncology: focus on T cells, CAF and ECM. *J Exp Clin Cancer Res.* (2019) 38:117. doi: 10.1186/s13046-019-1086-2
87. Barabási AL, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet.* (2004) 5:101–13. doi: 10.1038/nrg1272
88. Lesterhuis WJ, Bosco A, Millward MJ, Small M, Nowak AK, Lake RA. Dynamic versus static biomarkers in cancer immune checkpoint blockade: unravelling complexity. *Nat Rev Drug Discov.* (2017) 16:264–72. doi: 10.1038/nrd.2016.233
89. Letai A. Functional precision cancer medicine-moving beyond pure genomics. *Nat Med.* (2017) 23:1028–35. doi: 10.1038/nm.4389
90. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The cancer cell line encyclopedia enables

- predictive modelling of anticancer drug sensitivity. *Nature*. (2012) 483:603–7. doi: 10.1038/nature11003
91. Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, et al. A landscape of pharmacogenomic interactions in cancer. *Cell*. (2016) 166:740–54. doi: 10.1016/j.cell.2016.06.017
 92. Bar-Ephraim YE, Kretschmar K, Clevers H. Organoids in immunological research. *Nat Rev Immunol*. (2020) 20:279–93. doi: 10.1038/s41577-019-0248-y
 93. Wong AHH, Li H, Jia Y, Mak PI, Martins RP da S, Liu Y, et al. Drug screening of cancer cell lines and human primary tumors using droplet microfluidics. *Sci Rep*. (2017) 7:9109. doi: 10.1038/s41598-017-08831-z
 94. Eduati F, Utharala R, Madhavan D, Neumann UP, Longerich T, Cramer T, et al. A microfluidics platform for combinatorial drug screening on cancer biopsies. *Nat Commun*. (2018) 9:2434. doi: 10.1038/s41467-018-04919-w
 95. Montero J, Sarosiek KA, DeAngelo JD, Maertens O, Ryan J, Ercan D, et al. Drug-induced death signaling strategy rapidly predicts cancer response to chemotherapy. *Cell*. (2015) 160:977–89. doi: 10.1016/j.cell.2015.01.042
 96. Rohrs JA, Wang P, Finley SD. Understanding the dynamics of T-cell activation in health and disease through the lens of computational modeling. *JCO Clin Cancer Inform*. (2019) 3:1–8. doi: 10.1200/CCI.18.00057
 97. Tognetti M, Gabor A, Yang M, Cappelletti V, Windhager J, Charmpi K, et al. Deciphering the signaling network landscape of breast cancer improves drug sensitivity prediction. *bioRxiv*. (2020). doi: 10.1101/2020.01.21.907691
 98. Eduati F, Doldàn-Martelli V, Klinger B, Cokelaer T, Sieber A, Kogera F, et al. Drug resistance mechanisms in colorectal cancer dissected with cell type-specific dynamic logic models. *Cancer Res*. (2017) 77:3364–75. doi: 10.1158/0008-5472.CAN-17-0078
 99. Fey D, Halasz M, Dreidax D, Kennedy SP, Hastings JF, Rauch N, et al. Signaling pathway models as biomarkers: patient-specific simulations of JNK activity predict the survival of neuroblastoma patients. *Sci Signal*. (2015) 8:ra130. doi: 10.1126/scisignal.aab0990
 100. Béal J, Montagud A, Traynard P, Barillot E, Calzone L. Personalization of logical models with multi-omics data allows clinical stratification of patients. *Front Physiol*. (2018) 9:1965. doi: 10.3389/fphys.2018.01965
 101. Eduati F, Jaaks P, Wappler J, Cramer T, Merten CA, Garnett MJ, et al. Patient-specific logic models of signaling pathways from screenings on cancer biopsies to prioritize personalized combination therapies. *Mol Syst Biol*. (2020) 16:e8664. doi: 10.15252/msb.20188664
 102. Arulraj T, Barik D. Mathematical modeling identifies Lck as a potential mediator for PD-1 induced inhibition of early TCR signaling. *PLoS ONE*. (2018) 13:e0206232. doi: 10.1371/journal.pone.0206232
 103. Bolouri H, Young M, Beilke J, Johnson R, Fox B, Huang L, et al. Integrative network modeling reveals mechanisms underlying T cell exhaustion. *Sci Rep*. (2020) 10:1915. doi: 10.1038/s41598-020-58600-8
 104. Norton K-A, Gong C, Jamalian S, Popel AS. Multiscale agent-based and hybrid modeling of the tumor immune microenvironment. *Processes*. (2019) 7:37. doi: 10.3390/pr7010037
 105. Kather JN, Poleszczuk J, Suarez-Carmona M, Krisam J, Charoentong P, Valous NA, et al. Modeling of immunotherapy and stroma-targeting therapies in human colorectal cancer. *Cancer Res*. (2017) 77:6442–52. doi: 10.1158/0008-5472.CAN-17-2006
 106. Kather JN, Charoentong P, Suarez-Carmona M, Herpel E, Klupp F, Ulrich A, et al. High-throughput screening of combinatorial immunotherapies with patient-specific models of metastatic colorectal cancer. *Cancer Res*. (2018) 78:5155–63. doi: 10.1158/0008-5472.CAN-18-1126
 107. Thurlay K, Wu LF, Altschuler SJ. Modeling cell-to-cell communication networks using response-time distributions. *Cell Syst*. (2018) 6:355–67.e5. doi: 10.1016/j.cels.2018.01.016
 108. Grandclaudon M, Perrot-Dockès M, Trichot C, Karpf L, Abouzid O, Chauvin C, et al. A quantitative multivariate model of human dendritic cell-T helper cell communication. *Cell*. (2019) 179:432–47.e21. doi: 10.1016/j.cell.2019.09.012
 109. Wang H, Milberg O, Bartelink IH, Vicini P, Wang B, Narwal R, et al. *In silico* simulation of a clinical trial with anti-CTLA-4 and anti-PD-L1 immunotherapies in metastatic breast cancer using a systems pharmacology model. *R Soc Open Sci*. (2019) 6:190366. doi: 10.1098/rsos.190366
 110. Sontag ED. A dynamic model of immune responses to antigen presentation predicts different regions of tumor or pathogen elimination. *Cell Syst*. (2017) 4:231–41.e11. doi: 10.1016/j.cels.2016.12.003
 111. Sorribes IC, Basu A, Brady R, Enriquez-Navas PM, Feng X, Kather JN, et al. Harnessing patient-specific response dynamics to optimize evolutionary therapies for metastatic clear cell renal cell carcinoma – Learning to adapt. *bioRxiv*. (2019). doi: 10.1101/563130
 112. Perlstein D, Shlagman O, Kogan Y, Halevi-Tobias K, Yakobson A, Lazarev I, et al. Personal response to immune checkpoint inhibitors of patients with advanced melanoma explained by a computational model of cellular immunity, tumor growth, and drug. *PLoS ONE*. (2019) 14:e0226869. doi: 10.1371/journal.pone.0226869
 113. Riaz N, Havel JJ, Makarov V, Desrichard A, Urba WJ, Sims JS, et al. Tumor and microenvironment evolution during immunotherapy with nivolumab. *Cell*. (2017) 171:934–49.e16. doi: 10.1016/j.cell.2017.09.028
 114. Chen PL, Roh W, Reuben A, Cooper ZA, Spencer CN, Prieto PA, et al. Analysis of immune signatures in longitudinal tumor samples yields insight into biomarkers of response and mechanisms of resistance to immune checkpoint blockade. *Cancer Discov*. (2016) 6:827–37. doi: 10.1158/2159-8290.CD-15-1545
 115. Hwang S, Kwon AY, Jeong JY, Kim S, Kang H, Park J, et al. Immune gene signatures for predicting durable clinical benefit of anti-PD-1 immunotherapy in patients with non-small cell lung cancer. *Sci Rep*. (2020) 10:643. doi: 10.1038/s41598-019-57218-9
 116. Peskov K, Azarov I, Chu L, Voronova V, Kosinsky Y, Helmlinger G. Quantitative mechanistic modeling in support of pharmacological therapeutics development in immuno-oncology. *Front Immunol*. (2019) 10:924. doi: 10.3389/fimmu.2019.00924
 117. Eisenstein M. Making cancer immunotherapy a surer bet. *Nature*. (2017) 552:S72–S73. doi: 10.1038/d41586-017-08704-5

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Lapuente-Santana and Eduati. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Integrative Network Fusion: A Multi-Omics Approach in Molecular Profiling

Marco Chierici^{1*†}, Nicole Bussola^{1,2†}, Alessia Marcolini^{1†}, Margherita Francescatto^{1,3}, Alessandro Zandonà⁴, Lucia Trastulla⁵, Claudio Agostinelli², Giuseppe Jurman^{1*‡} and Cesare Furlanello^{1,6‡}

OPEN ACCESS

Edited by:

Chiara Romualdi,
University of Padova, Italy

Reviewed by:

Prashanth N. Suravajhala,
Birla Institute of Scientific Research,
India
Jun Zhong,
National Cancer Institute (NCI),
United States

*Correspondence:

Marco Chierici
chierici@fbk.eu
Giuseppe Jurman
jurman@fbk.eu

[†]These authors share joint first
authorship

[‡]These authors share joint last
authorship

Specialty section:

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

Received: 31 March 2020

Accepted: 28 May 2020

Published: 30 June 2020

Citation:

Chierici M, Bussola N, Marcolini A,
Francescatto M, Zandonà A,
Trastulla L, Agostinelli C, Jurman G
and Furlanello C (2020) Integrative
Network Fusion: A Multi-Omics
Approach in Molecular Profiling.
Front. Oncol. 10:1065.
doi: 10.3389/fonc.2020.01065

¹ Fondazione Bruno Kessler, Trento, Italy, ² University of Trento, Trento, Italy, ³ Department of Medical, Surgical and Health Sciences, University of Trieste, Trieste, Italy, ⁴ NIDEK Technologies Srl, Albignasego (PD), Italy, ⁵ Max Planck Institute of Psychiatry, Munich, Germany, ⁶ HK3 Lab, Milan, Italy

Recent technological advances and international efforts, such as The Cancer Genome Atlas (TCGA), have made available several pan-cancer datasets encompassing multiple omics layers with detailed clinical information in large collection of samples. The need has thus arisen for the development of computational methods aimed at improving cancer subtyping and biomarker identification from multi-modal data. Here we apply the Integrative Network Fusion (INF) pipeline, which combines multiple omics layers exploiting Similarity Network Fusion (SNF) within a machine learning predictive framework. INF includes a feature ranking scheme (rSNF) on SNF-integrated features, used by a classifier over juxtaposed multi-omics features (juXT). In particular, we show instances of INF implementing Random Forest (RF) and linear Support Vector Machine (LSVM) as the classifier, and two baseline RF and LSVM models are also trained on juXT. A compact RF model, called rSNFi, trained on the intersection of top-ranked biomarkers from the two approaches juXT and rSNF is finally derived. All the classifiers are run in a 10x5-fold cross-validation schema to warrant reproducibility, following the guidelines for an unbiased Data Analysis Plan by the US FDA-led initiatives MAQC/SEQC. INF is demonstrated on four classification tasks on three multi-modal TCGA oncogenomics datasets. Gene expression, protein expression and copy number variants are used to predict estrogen receptor status (BRCA-ER, $N = 381$) and breast invasive carcinoma subtypes (BRCA-subtypes, $N = 305$), while gene expression, miRNA expression and methylation data is used as predictor layers for acute myeloid leukemia and renal clear cell carcinoma survival (AML-OS, $N = 157$; KIRC-OS, $N = 181$). In test, INF achieved similar Matthews Correlation Coefficient (MCC) values and 97% to 83% smaller feature sizes (FS), compared with juXT for BRCA-ER (MCC: 0.83 vs. 0.80; FS: 56 vs. 1801) and BRCA-subtypes (0.84 vs. 0.80; 302 vs. 1801), improving KIRC-OS performance (0.38 vs. 0.31; 111 vs. 2319). INF predictions are generally more accurate in test than

one-dimensional omics models, with smaller signatures too, where transcriptomics consistently play the leading role. Overall, the INF framework effectively integrates multiple data levels in oncogenomics classification tasks, improving over the performance of single layers alone and naive juxtaposition, and provides compact signature sizes¹.

Keywords: multi-omics, classification, network, oncogenomics, predictive modeling

1. INTRODUCTION

The challenge of integrating multi-omics data is as old as bioinformatics itself (1, 2), but, despite the wide literature, it remains an open issue nowadays, even worth being funded by major institutions².

This study introduces Integrative Network Fusion (INF), a reproducible network-based framework for high-throughput omics data integration that leverages machine learning models to extract multi-omics predictive biomarkers. Originally conceptualized and tested on multi-omics metagenomics data in an early preliminary version (3, 4), INF combines the signatures retrieved from both the early-integration approach of variable juxtaposition (juXT) and an intermediate-integration approach [SNF, (5)], to find the optimal set of predictive features. In particular, first a set of top-ranked features is extracted by juXT by a classifier, here Random Forest (RF) and linear Support Vector Machine (LSVM). Then, a feature ranking scheme (rSNF) is computed on SNF-integrated features and finally a RF model (rSNFi) is trained on the intersection of two sets of top-ranked features from juXT and rSNF, obtaining an approach that effectively integrates multiple omics layers and provides compact predictive signatures. Selection bias and data-leakage effects are controlled by performing the experiments within a rigorous Data Analysis Plan (DAP) to warrant reproducibility, following the guidelines of the US FDA-led initiatives MAQC/SEQC (6–8). In particular, to alleviate the computational burden of the full DAP pipeline, an approximated DAP is designed to lighten computing without significantly affecting the results. Further, experiments are run on samples with randomly shuffled labels as a sanity check vs. overfitting effects and, finally, INF robustness is verified by testing on different train/test splits.

We test INF on three datasets retrieved from the TCGA repository, to predict either the estrogen receptor status (ER) or the cancer subtype on the breast invasive carcinoma (BRCA) dataset, and to predict the overall survival (OS) on the kidney renal clear cell carcinoma (KIRC) and acute myeloid leukemia (AML) datasets. Overall, INF improves over the performance of single layers and naive juxtaposition on all four oncogenomics tasks, extracting a biologically meaningful compact set of predictive biomarkers. Notably, the transcriptomics layer is

TABLE 1 | Data summary.

Dataset-task	#Samples	Layers (#features)
BRCA-ER	381	<i>gene</i> (17814), <i>cnv</i> (18050), <i>prot</i> (142)
BRCA-subtypes	305	
AML-OS	157	<i>gene</i> (10265), <i>meth</i> (2500), <i>mirna</i> (352)
KIRC-OS	181	<i>gene</i> (10265), <i>meth</i> (2500), <i>mirna</i> (484)
Synthetic-ST	380	layer1 (100), layer2 (50), layer3 (250)

BRCA, breast invasive carcinoma; *AML*, acute myeloid leukemia; *KIRC*, kidney renal clear cell carcinoma; *gene*, gene expression; *cnv*, copy number variants; *prot*, protein expression; *meth*, methylation; *mirna*, microRNA expression; *ER*, estrogen receptor; *subtypes*, breast cancer subtypes; *OS*, overall survival; *ST*, synthetic target.

prevalent inside the inferred INF signatures, consistently with published findings (9).

The INF framework is currently designed to integrate an arbitrary number of one-dimensional omics layers. We plan to further extend the framework by enabling the integration of histopathological features extracted from whole slide images (10) or deep features from radiological images (11) extracted by deep neural network architectures, carefully addressing all potential caveats (12).

2. MATERIALS AND METHODS

2.1. Data

Three multi-modal cancer datasets generated by The Cancer Genome Atlas (TCGA) Research Network (<https://www.cancer.gov/tcga>) and four classification tasks are considered in this study. Protein expression (*prot*), gene expression (*gene*), and copy number variants (*cnv*) are used to predict breast invasive carcinoma (BRCA) estrogen receptor status (0: negative; 1: positive) and subtypes (luminal A, luminal B, basal-like, HER2-enriched). Methylation (*meth*), gene expression (*gene*), and microRNA expression (*mirna*) are used to predict acute myeloid leukemia (AML) and kidney renal clear cell carcinoma (KIRC) overall survival (0: alive; 1: deceased). The number of samples and features for each omic layer and classification task are detailed in **Table 1**; class balance, split by dataset, is reported in **Table 2**.

For AML (13) and KIRC (14), gene expression is profiled using the Illumina HiSeq2000 and quantified as log₂-transformed RSEM normalized counts; miRNA mature strand expression is profiled using the Illumina Genome Analyzer and quantified as reads per million miRNA mapped; and methylation is assessed by Illumina Human Methylation 450K and expressed as beta values. For BRCA (15), gene expression is profiled with Agilent 244K custom gene expression microarrays; protein expression is

¹INF source code is publicly available on the GitLab repository <https://gitlab.fbk.eu/MPBA/INF>, while data is archived at <http://dx.doi.org/10.6084/m9.figshare.12052995.v1>

²European Call Multi-omics for genotype-phenotype associations (RIA) <https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/biotech-07-2020>

TABLE 2 | Class balance.

Dataset-task	Labels (#samples)
BRCA-ER	Negative (95), Positive (286)
BRCA-subtypes	LuminalA (170), LuminalB (102), Basal-like (81), HER2-enriched (48)
AML-OS	Dead (101), Alive (56)
KIRC-OS	Dead (133), Alive (48)

BRCA, breast invasive carcinoma; AML, acute myeloid leukemia; KIRC, kidney renal clear cell carcinoma; ER, estrogen receptor; subtypes, breast cancer subtypes; OS, overall survival.

TABLE 3 | Synthetic data summary for each simulated layer.

Layer	# Features	# Informative features	Multiplicative factor	Class separation	Random state
Layer 1	100	10	Default	1.0	1
Layer 2	50	5	Default	1.2	2
Layer 3	250	25	10	0.8	3

Multiplicative factor, class separation, and random state refer to the parameters `scale`, `class_sep`, and `random_state` of the `make_classification` function in `scikit-learn`.

assessed by reverse phase protein arrays; copy number profiles are measured using Affymetrix Genome-Wide Human SNP Array 6.0 platform, copy number variants are segmented by the TCGA Firehose pipeline using GISTIC2 method, and then mapped to genes.

The original data is publicly accessible on the National Cancer Institute GDC Data Portal (<https://portal.gdc.cancer.gov/>) and the Broad GDAC Firehose (<https://gdac.broadinstitute.org/>), where further details on data generation can be found. The data was retrieved in December, 2019 and January, 2020 using the R TCGA R library (16).

Furthermore, the INF pipeline has been tested on a synthetic dataset with 380 observations in two classes (70% class 1 and 30% class 2, defining the synthetic target ST), 3 pseudo-omics layers, and 400 features (layer 1: 100; layer 2: 50; layer 3: 250). The dataset is generated in-house using `scikit-learn`'s `make_classification` function with the arguments `shuffle=False` and `flip_y=0`. The number of informative features and the difficulty of the task were set on a per-layer basis, as summarized in Table 3.

2.2. In silico Workflow

The INF pipeline integrates two or more omics layers, e.g., gene expression, protein expression, or methylation, in a machine learning framework for improved patient classification and biomarker identification in cancer. The core consists of three main components, structured as in Figure 1, managing the integration of the omics layers and their predictive modeling. A baseline integration method (juXT) is first considered by training a Random Forest (RF) (17) or a linear Support Vector Machine (LSVM) (18) classifier on juxtaposed multi-omics data, ranking features by ANOVA *F*-value. Secondly, the multi-omics features

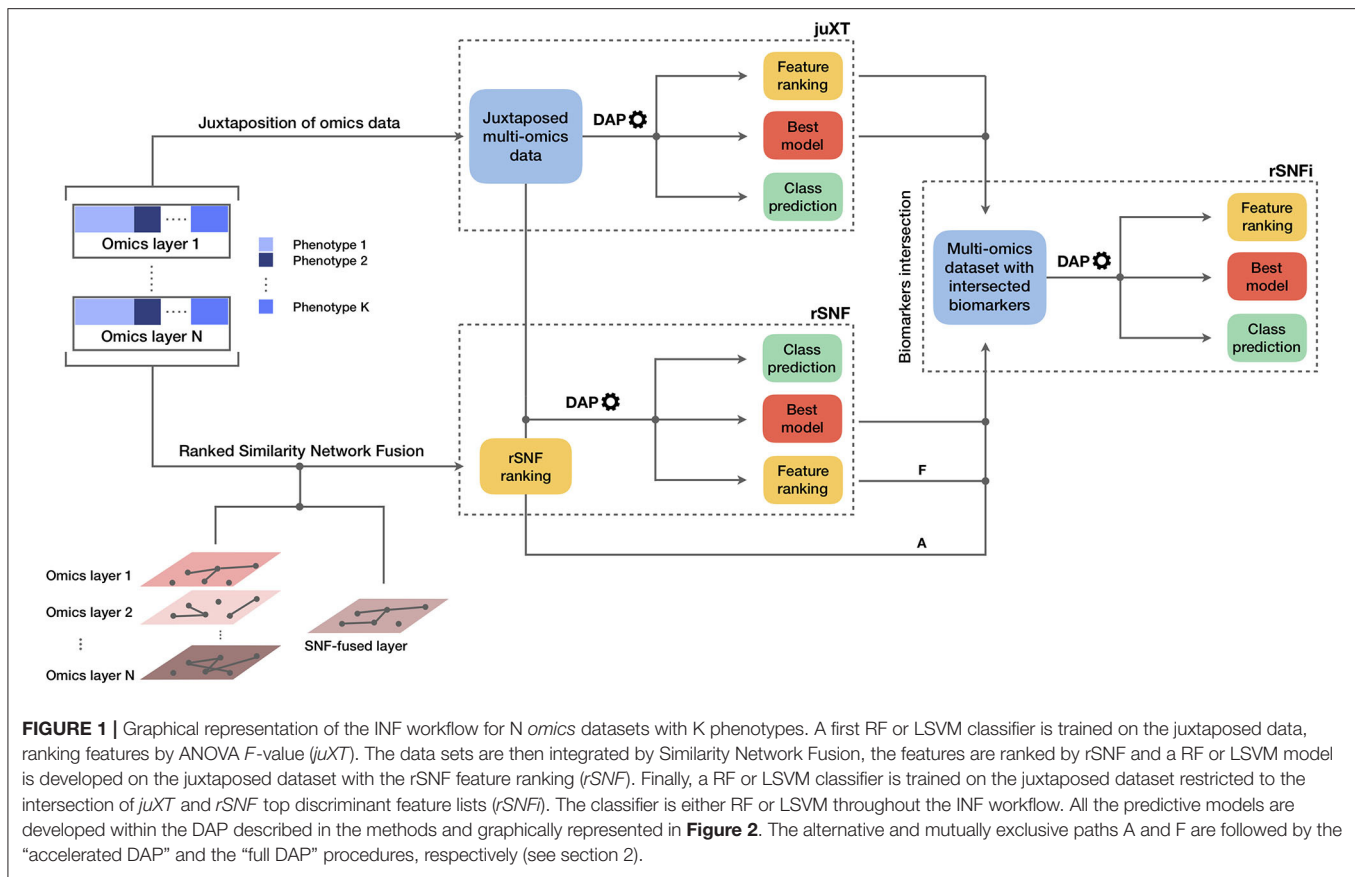
are integrated by Similarity Network Fusion (SNF) (5), a method that computes a sample similarity network for each data type and fuses them into one network. INF introduces a novel feature ranking scheme (rSNF) that sorts multi-omics features according to their contribution to the SNF-fused network structure. A RF or LSVM classifier is trained on the juxtaposed multi-omics data, ranking features by rSNF. A compact RF model (rSNFi) is finally trained on the juxtaposed dataset restricted on the intersection of top-ranked biomarkers from juXT and rSNF.

2.3. Omics Integration

In a comparative review of scientific literature, SNF (5) emerged as one of the most reliable alternatives to simple juxtaposition-based integration. SNF is a non-Bayesian network-based method that can be divided into two main steps: the first step builds a sample-similarity network for each omics dataset, where nodes represent samples and edges encode a scaled exponential Euclidean distance kernel computed on each pair of samples; the second step implements a non-linear combination of these networks into a single similarity network through an iterative procedure. The multi-omics datasets are first converted into graphs, and for each graph two matrices are computed: a patient pairwise similarity matrix ("status matrix"), and a matrix with similarity of each patient to the *K* most similar patients, through *K*-nearest neighbors ("local affinity matrix"). At each iteration, the status matrix is updated through the local affinity matrix, generating two parallel interchanging processes. The status matrices are finally fused together into a single network. Spectral clustering is performed on the fused network, in order to identify sub-communities of samples, potentially reflecting phenotypes. The clustering performance is evaluated with respect to a ground truth, i.e., the real phenotype each sample belongs to, by the Normalized Mutual Information (NMI) score. SNF integrates multiple omics datasets into a single comprehensive network in the space of samples rather than measurements (e.g., gene expression values).

This work proposes multi-omics integration as an approach to identify robust biomarkers of samples phenotypes or cancer subtypes (e.g., survival status vs. breast cancer subtyping); consequently, it is necessary to extract measurements information from the SNF-fused network of samples. To this aim, we extended SNF by implementing *rSNF* (ranked SNF), a feature-ranking scheme based on SNF-fused network clustering. In detail, a patient network W_i is built for each feature f_i , based on f_i alone, and spectral clustering is performed on it. Then, NMI score is computed comparing the samples clusters found inside W_i with those in the fused network; the higher the score, the more similar the clustering between the fused network and W_i . Thus, each feature f_i is associated to a consistency score, ranking all multi-omics features with respect to their relative contribution to the whole network structure.

The entire procedure of similarity networks inference and fusion relies on two hyperparameters: α , the scaling variance in the scaled exponential similarity kernel used for similarity networks construction, and *K*, the number of nearest neighbors in sparse kernel and scaled exponential similarity kernel construction. While the original method (5) assigned fixed values



to α and K , in this study the optimal hyperparameters are chosen among the grids $\alpha_{grid} = \{0.3, 0.35, 0.4, 0.45, \dots, 0.8\}$ and $K_{grid} = \{i \in \mathbb{N}, 10 \leq i \leq 30\}$ in a 10×5 -fold cross-validation schema.

2.4. Predictive Profiling

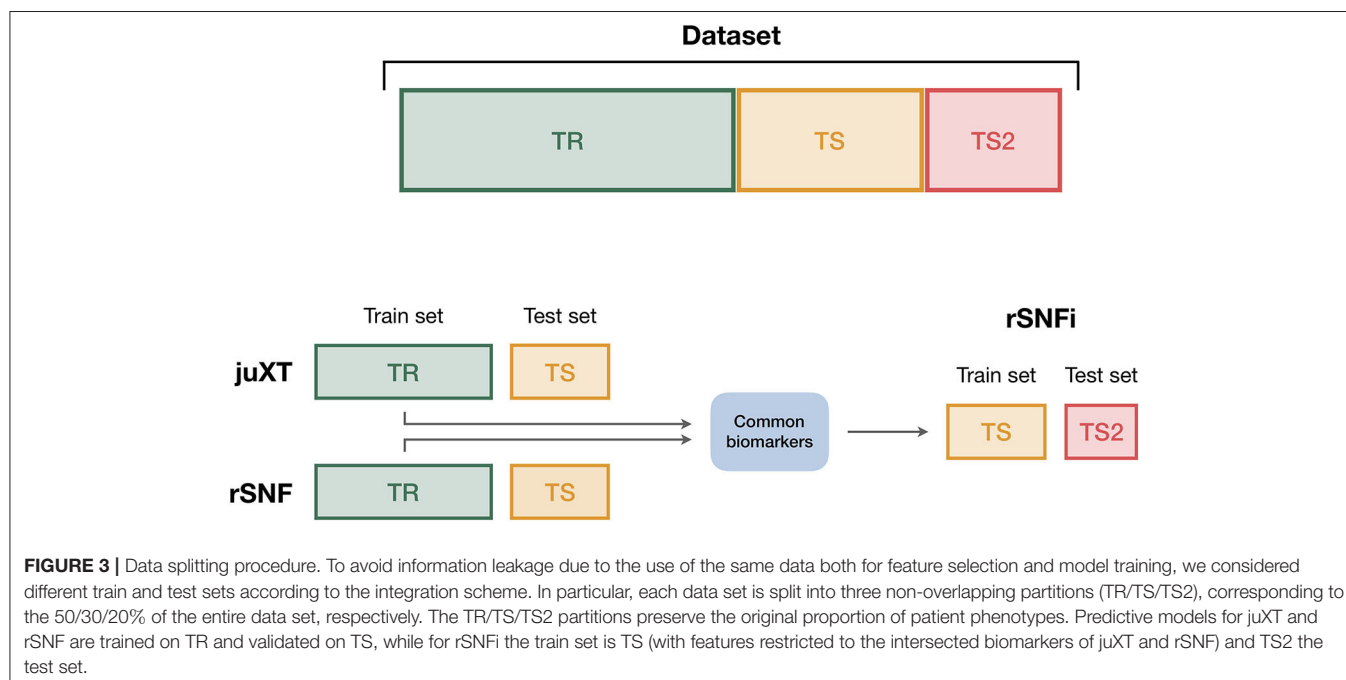
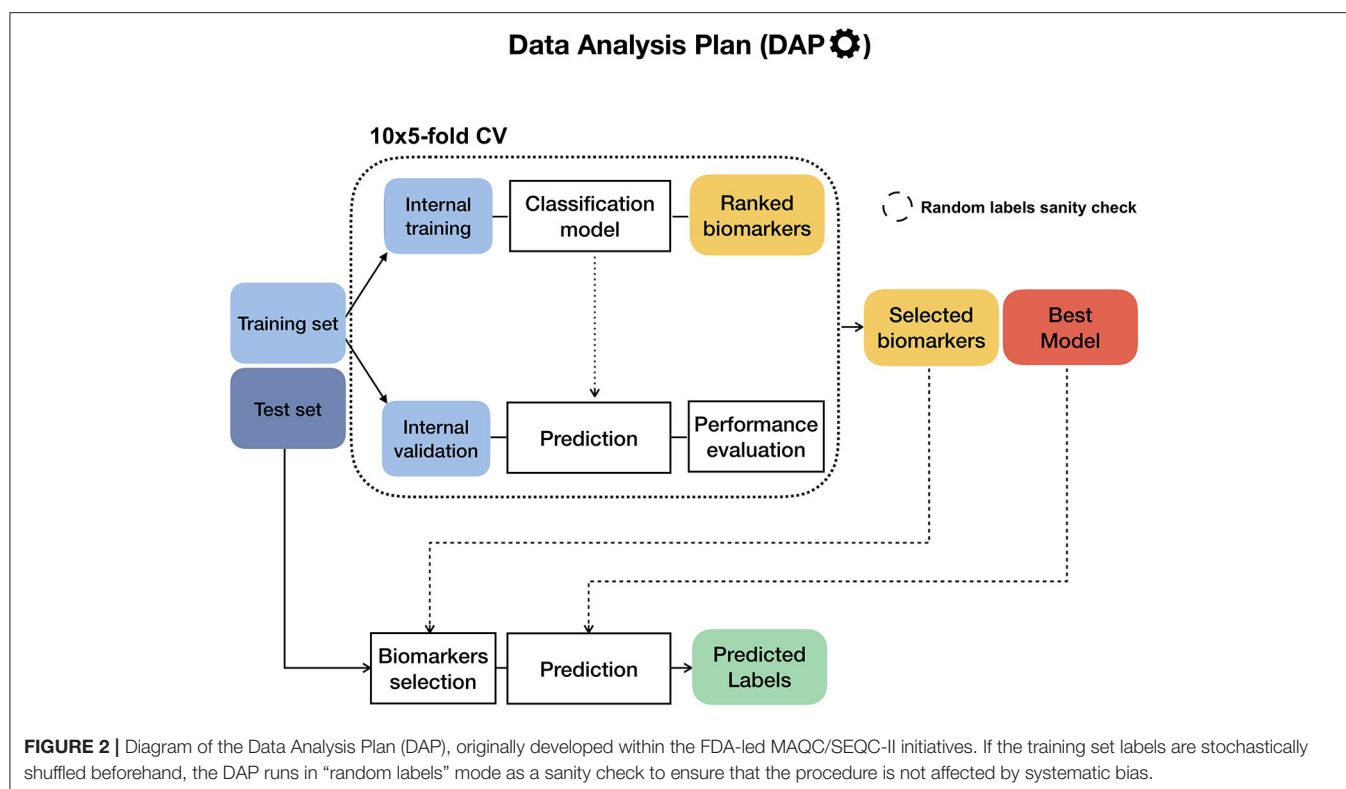
To ensure the reproducibility of results and limit overfitting, the development of classification models is performed inside a Data Analysis Plan (DAP) (**Figure 2**), following the guidelines derived by the U.S. Food and Drug Administration MAQC/SEQC studies (6, 19). Data is split in a training set (TR) and two non-overlapping test sets (TS, TS2), preserving the original proportion of patient phenotypes (classes). The TR/TS/TS2 partitions are 50/30/20 of the entire data set, respectively. The data splitting procedure is repeated 10 times so to obtain 10 different TR/TS/TS2 splits. Predictive models are trained and developed on TR and TS for $juXT$ and $rSNFi$; in the case of $rSNFi$, the models are trained and developed on TS and TS2 to avoid information leakage due to using the same data both for feature selection and model training (see **Figure 3**). For each split, Random Forest (RF) or linear kernel Support Vector Machine (LSVM) classifiers are trained on the training partition within a stratified 10×5 -fold cross-validation (10×5 -CV). The model performance is assessed in terms of average precision, recall and Matthews Correlation Coefficient (MCC) (20, 21). The MCC is generally regarded as a balanced measure of accuracy and precision that can be used both in binary and multiclass

problems (22, 23) and even when classes are imbalanced (24). MCC lies in $[-1, 1]$, with 1 meaning perfect prediction, -1 inverse prediction and 0 random guess. For binary classification tasks, MCC is calculated on true and predicted labels considering true positive (TP), true negative (TN), false positive (FP) and false negative (FN) values, as in the following:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

At each CV round, features are ranked either by ANOVA F -value (for $juXT$, $rSNFi$) or by the $rSNF$ ranking (see section 2.3) and different classification models are trained for increasing numbers of ranked features, namely 5, 10, 25, 50, 75, and 100% of the total features. A unified list of top-ranked features is then obtained by Borda aggregation of all the ranked CV lists (25, 26). The best model is later retrained on the whole training set restricted to the features yielding the maximum MCC in CV, and validated on the test partition. A global list of top-ranked features is derived for $juXT$, $rSNF$, and $rSNFi$ by Borda aggregation of the Borda lists of each TR/TS split (Borda of BORDAS, “BoB”). The signatures for $juXT$, $rSNF$, and $rSNFi$ are defined by the top N features of the corresponding BoB lists, with N being the median size of top features across all experiments.

In the “full” version of the DAP ($fDAP$), described above, the $rSNF$ ranking is performed at each CV round on



the training portion of the data. Since this procedure is quite demanding in terms of computational time, even if parallelized (≈ 9 feature/min), we devised an “accelerated” version of the DAP (*aDAP*), where the rSNF ranking is precomputed on the whole TR data and used as is at each

CV round. We assessed the *fDAP* vs. *aDAP* performance on the synthetic dataset as well as BRCA-ER and BRCA-subtypes by comparing the overall metrics and measuring the dissimilarity of the rSNF BoB of the two DAPs by the Canberra distance (25).

TABLE 4 | Summarized best predictive performances for each classification task using RF model and three omics layers.

Task	Method	MCC_cv (CI)	MCC_ts (CI)	PREC_cv (CI)	PREC_ts (CI)	REC_cv (CI)	REC_ts (CI)	Nf
BRCA-ER	juXT	0.785 (0.776, 0.795)	0.797 (0.778, 0.819)	0.935 (0.932, 0.938)	0.946 (0.935, 0.957)	0.962 (0.959, 0.965)	0.955 (0.949, 0.962)	1801
	rSNF	0.792 (0.782, 0.801)	0.804 (0.779, 0.830)	0.938 (0.935, 0.941)	0.947 (0.934, 0.961)	0.961 (0.958, 0.965)	0.958 (0.949, 0.966)	1801
	rSNFi	0.820 (0.808, 0.831)	0.830 (0.803, 0.857)	0.955 (0.951, 0.959)	0.951 (0.939, 0.962)	0.956 (0.952, 0.960)	0.967 (0.956, 0.977)	55.5
BRCA-subtypes	juXT	0.778 (0.771, 0.785)	0.795 (0.771, 0.817)					1801
	rSNF	0.769 (0.762, 0.777)	0.811 (0.787, 0.835)	-	-	-	-	1801
	rSNFi	0.788 (0.778, 0.798)	0.838 (0.794, 0.879)					301.5
KIRC-OS	juXT	0.266 (0.243, 0.289)	0.305 (0.229, 0.382)	0.540 (0.509, 0.570)	0.579 (0.494, 0.664)	0.299 (0.280, 0.317)	0.343 (0.300, 0.393)	2319
	rSNF	0.253 (0.230, 0.276)	0.274 (0.189, 0.348)	0.539 (0.505, 0.571)	0.628 (0.507, 0.739)	0.253 (0.235, 0.270)	0.257 (0.200, 0.314)	3313
	rSNFi	0.268 (0.239, 0.298)	0.378 (0.288, 0.464)	0.485 (0.449, 0.521)	0.594 (0.512, 0.668)	0.321 (0.296, 0.347)	0.490 (0.380, 0.600)	111
AML-OS	juXT	0.141 (0.120, 0.163)	0.223 (0.146, 0.307)	0.675 (0.669, 0.681)	0.704 (0.682, 0.725)	0.860 (0.849, 0.870)	0.880 (0.850, 0.907)	6559
	rSNF	0.180 (0.157, 0.202)	0.263 (0.175, 0.366)	0.685 (0.679, 0.691)	0.717 (0.692, 0.743)	0.876 (0.867, 0.886)	0.873 (0.847, 0.903)	656
	rSNFi	0.274 (0.245, 0.301)	0.176 (0.068, 0.278)	0.726 (0.718, 0.735)	0.673 (0.639, 0.706)	0.870 (0.858, 0.882)	0.835 (0.785, 0.880)	91.5

CI: 95% bootstrap confidence interval; {MCC,PREC,REC}_cv: best average MCC, precision, recall in cross-validation on training set splits; {MCC,PREC,REC}_ts: average MCC, precision, recall on test set splits; Nf: median number of features leading to MCC_cv. Bold indicates best performance (highest MCC and smallest signature size). Precision and recall were computed for binary classification tasks only.

RF models are trained using 500 trees, measuring the quality of a split as mean decrease in the Gini impurity index (17); the regularization parameter C of LSVM models is tuned over the grid $C_{grid} = \{10^i, i \in \mathbb{N}, -2 \leq i \leq 3\}$ within a $10 \times$ stratified Monte Carlo cross-validation (50% training/validation proportion). Results for RF models are summarized in **Table 4**, while LSVM models performance is detailed in the **Supplementary Tables BRCA-ER_LSVM, KIRC-OS_LSVM**.

To ensure that the predictive profiling procedure is not affected by selection bias, the whole INF workflow, including the rSNF procedure, is also repeated after randomly scrambling the training set labels (“random labels” mode): in this setup, the performance of a classifier unaffected by systematic bias should be close to that of a random predictor, with MCC close to zero.

2.5. Implementation

The complete INF pipeline is implemented through the workflow management tool Snakemake (27, 28), which allows automatic handling of all dependencies required to generate the INF output. The pipeline operates on N omics input files, one for each layer that should be integrated, and a single file describing the patient labels. The omics files are tab-separated text matrices with patients on the rows and features on the columns, with row and column identifiers. The label file is a single column file with patient phenotypes, with no header. This input structure, with one file per omic layer and a label file, simplifies the downstream analysis and reduces to a minimum the preprocessing burden for the end user.

The predictive profiling module, including the DAP, is written in Python 3.6 on top of NumPy (29) and scikit-learn methods (30). The ranked SNF (rSNF) procedure is implemented in R (31) leveraging the original R scripts provided by SNF authors (5), extended by a dedicated script for SNF tuning and a main script

for SNF analysis and the post-SNF feature selection procedure, which is parallelized over the features for efficiency using the `foreach` R library.

2.6. Computational Details

The INF computations were run on the FBK Linux high-performance computing facility KORE, on a 8-core i7 3.4 GHz Linux workstation, and on a 72-vCPU 2.7 GHz Platinum Intel Xeon 8168 Microsoft Azure cloud machine (F72s v2 series).

2.7. Data and Code Availability

To further foster reproducibility and support users and future developers, the full code of this benchmark is publicly shared on the GitLab repository <https://gitlab.fbk.eu/MPBA/INF>. Additional information is included in the **Supplementary Material** available on the publisher's website, while the full set of experimental data can be accessed at <http://dx.doi.org/10.6084/m9.figshare.12052995.v1>.

3. RESULTS

The INF workflow was run on all tasks considering 3-layer integration and all 2-layer combinations; the DAP was also run separately on all single-layer datasets in order to obtain a baseline. All results presented here refer to experiments performed with RF classifier. Experiments using LSVM were performed on BRCA-ER and KIRC-OS obtaining similar classification performances, top features and layer contributions (**Supplementary Tables BRCA-ER_LSVM, KIRC-OS_LSVM**). The classifier performance for 3-layer integration is summarized in **Table 4**, in terms of average cross-validation MCC on the 10 training set splits (MCC_cv) with 95% Studentized bootstrap confidence intervals (CI) as (MCC_min, MCC_max), average MCC on the 10 test set splits (MCC_ts) with CI, and median number of features (Nf) yielding MCC_cv. Similarly, precision (PREC) and recall

(REC) are reported in **Table 4** as average cross-validation and test set values with CI. As expected, whenever there is a non-negligible unbalance toward the positive class, the number of false positives tends to increase, with more false positives yielding a comparatively low precision with higher recall, and vice versa. In both cases, the MCC efficiently works in balancing the two effects. The classifier performance on single-layer and 2-layer data is summarized in **Figure 4**.

A comparison between the “accelerated” flavor of the DAP (*aDAP*) and the full DAP (*fDAP*) was run on synthetic data, BRCA-ER and BRCA-subtypes data, with *aDAP* yielding similar performance metrics and top-ranked biomarker lists as *fDAP* (**Supplementary Tables Synthetic_RF, BRCA_RF_fDAP, canberra_distances**), while being $\approx 30\times$ faster (for BRCA-ER, approx. 2 vs. 64 h, or 300 features/min vs. 9 features/min). All the results presented here were thus obtained using *aDAP*. Moreover, the INF workflow running in “random labels” mode achieved an average cross-validation MCC ≈ 0 , as expected by a procedure unaffected by systematic bias.

Overall, integrating multiple omics layers with INF yields better or comparable classification performance than using only features from a single layer or naïve omics juxtaposition, at the same time with much more compact signature sizes. On 3-layer BRCA-subtypes and 2- or 3-layer KIRC-OS, INF outperforms the single layers, as well as juXT and rSNF (**Figure 4, Table 4**). On 2-layer BRCA-subtypes, INF performance on *gene-cnv* and *gene-prot* is comparable to the best-performing single-layer data (*gene*) and superior to *cnv* and *prot* single layers, while INF on *cnv-prot* only improves over the *cnv* single layer. On the BRCA-ER task, the performance with INF integration of 2 or 3 layers is still better than using single layers, nevertheless to a smaller extent, except for *cnv-prot* integration which performs better than *cnv* alone but slightly worse than *gene* and *prot* single layers. The good performances achieved at the *gene* and *prot* single layers do not come unexpected, since the biological nature of the target ER-status is defined at transcriptomics level. On the more difficult AML-OS task, INF has better performance over both rSNF and juXT on *gene-mirna* and *meth-mirna* integration, still improving over single-layer performance both in terms of MCC and reduced signature sizes.

3.1. One or Multi-Omics Layers vs. juXT/rSNF/rSNFi

For BRCA-ER, three-layer INF (rSNFi) integration performs better than either rSNF or juXT (MCC test 0.830 vs. 0.804, 0.797 for rSNF and juXT, respectively). All two-layer INF integrations perform similarly to, or better than, the corresponding rSNF and juXT integrations, in particular for *cnv-prot* integration (MCC test 0.746 vs. 0.682, 0.692 resp. for rSNF and juXT).

On BRCA-subtypes, the 3-layer INF integration performs better than either rSNF or juXT (MCC test 0.838 vs. 0.811, 0.795 resp. for rSNF and juXT), nevertheless without improving over the *gene* single-layer performance (MCC test 0.821). However, the INF median signature size is only 301.5, compared to 1801 for rSNF and juXT, and 891 for the *gene* layer alone. All

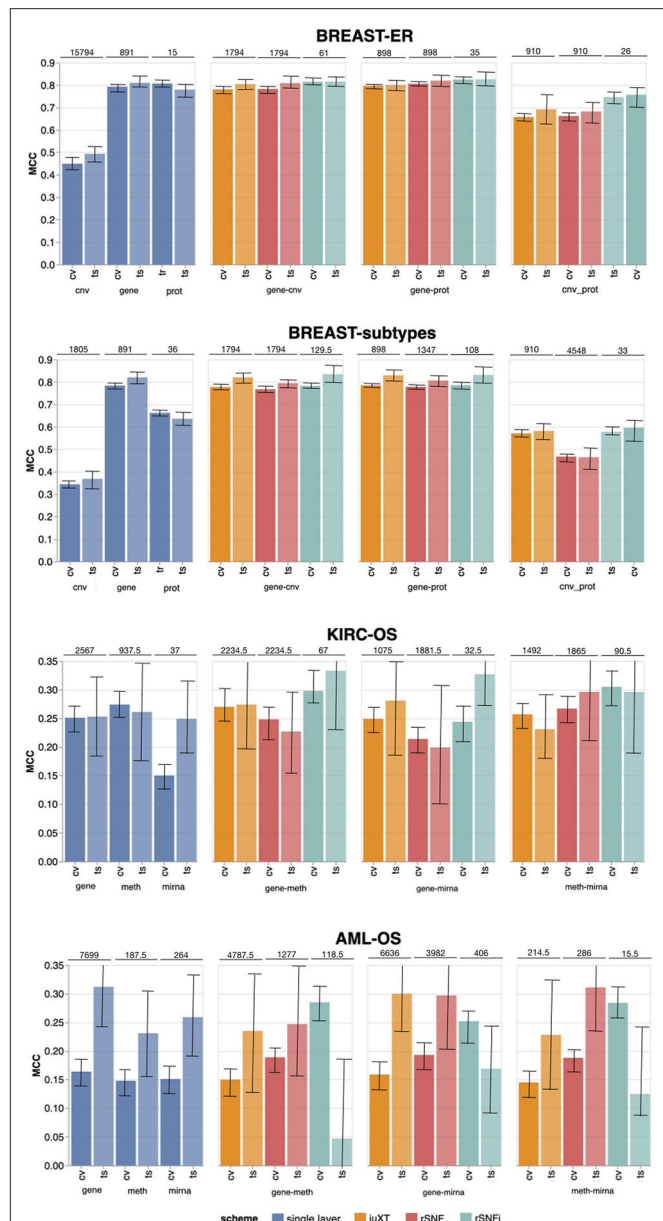


FIGURE 4 | Overview of Random Forest classification performance (MCC, Matthews Correlation Coefficient) on the four tasks in cross validation (“CV”) and test (“ts”), on single-layer (blue shades) and on all two-layer combinations for juXT (orange), rSNF (red) and rSNFi (green). Bars indicate 95% confidence intervals. On top of each CV-ts pair is the median number of features leading to best CV performance.

two-layer INF integrations yield better performance than their corresponding juXT or rSNF integrations.

Omics integration is particularly effective for KIRC-OS, as all 2- and 3-layer INF integrations outperform juXT, rSNF, and each of the single-layer classifiers. In fact, 3-layer rSNFi achieves MCC test 0.378 vs. 0.274, 0.305 (resp. for juXT, rSNF), 0.296, 0.327, 0.333 (resp. rSNFi *meth-mirna*, *gene-mirna*, *gene-meth*), and 0.253, 0.261, 0.249 (resp. *gene*, *meth*, *mirna*).

For AML-OS, INF feature sets are always more compact than either juXT or rSNF, with three-layer integration giving better MCC than any of the INF two-layer integrations (MCC test 0.176 vs. 0.125, 0.169, 0.047, respectively three-layer vs. *meth-mirna*, *gene-mirna*, *gene-meth*). Moreover, cross-validation MCCs corresponding to INF integration are better than any single layer MCC as well as rSNF and juXT.

3.2. Characterization of the Signatures Identified by INF

For all tasks, INF signatures are markedly more compact with respect to both juXT and rSNF. With 91.5 vs. 6559 (1.4%) median features (rSNFi vs. juXT), the largest reduction in size occurs for AML-OS 3-layer integration, while the least reduction is observed for BRCA-subtypes task, with 301.5 vs. 1801 (16.7%) median features (rSNFi vs. juXT).

In terms of contributions from the omics datasets being integrated, the *gene* layer generally provides the largest number of features to the signatures identified by the INF workflow. In particular for the BRCA dataset, in both ER and subtypes tasks, the *gene* layer contributes over 95% of the top features for juXT and rSNFi, with rSNF signatures being slightly more balanced (*prot* contribution remains marginal, while *cnv* provides 28.3 and 17.7% of the top features in ER and subtypes tasks respectively). This is expected as the class label is defined mainly at transcriptomics level. In AML-OS experiments, the layer contributing the most is still *gene*, accounting for ca. 78, 73, and 81% of the top feature sets for RF juXT, rSNF and rSNFi experiments, respectively. In KIRC-OS experiments, *gene* is the layer contributing the most to the top juXT and rSNF feature sets, while *meth* is the major contributor for rSNFi. The percentage of features from each omic layer contributing to the top signatures for juXT, rSNF and rSNFi 3-layer integrations are reported in **Supplementary Tables layer_contribution**. The RF rSNFi signatures for all tasks are available in **Supplementary Tables BRCA-ER_RF_rSNFi**, **BRCA-subtypes_RF_rSNFi**, **AML-OS_RF_rSNFi** and **KIRC-OS_RF_rSNFi**.

Even though a systematic biological interpretation of the identified signatures is beyond the scope of this work, to ascertain the reliability of our results we compared them with published data. The top features in the BRCA-ER rSNFi signature include multiple genes known to be associated with breast carcinoma progression and outcome such as AGR3, B3GNT, and MLPH (32–34). In addition we find the estrogen receptor gene (ESR1 from the *gene* and ER-alpha from the *prot* layer) and the transcription factor GATA3 (from both *gene* and *prot* layers) (35). Both the BRCA-ER and BRCA-subtypes signatures include genes previously identified as novel biomarkers for intrinsic breast carcinoma subtype prediction (36). Interestingly there is only partial overlap between the top features identified in BRCA ER vs. subtypes tasks. Considering AML-OS task, it is noteworthy to mention that the top feature identified has been recently reported as a potential biomarker predicting overall survival in a subset of AML patients (37).

Within the *mirna* features of the AML-OS signature, MIR-203 expression was recently found to be associated with AML patient survival (38); MIR-100 is highly expressed in AML and was found to regulate cell differentiation and survival (39); high expression of miR-504-3p was reported to be associated with favorable AML prognosis (40). Given that the rSNFi signature identified in the KIRC-OS task contains a large percentage of methylation data (86.5%), its direct interpretation is more difficult. It is however interesting to observe that all the 15 *gene* features in the signature are identified as prognostic markers for renal carcinoma according to the Human Protein Atlas (41).

3.3. Unsupervised Analysis

The features selected by juXT, rSNF and rSNFi are projected on a bi-dimensional space using the UMAP unsupervised multidimensional projection method (42, 43). Here we show an example on the BRCA-subtypes 3-layer dataset, with a UMAP projection of the features selected by juXT (**Figure 5**) compared to the UMAP projection of the INF signature (**Figure 6**) for one of the 10 data splits (the UMAP plots for the remaining 9 splits are in **Figures S1, S2**). Colors represent cancer subtypes and shapes represent training/test partitions. Using the 1801 juXT features, cancer subtypes are roughly clustered, with HER2-enriched and Luminal B being more dispersed (**Figure 5**). The clusters appear to be more sharply defined in the projection of the 302-feature INF signature: in particular, Basal-like patients form a distinct cluster, while Luminal A, Luminal B and HER2-enriched patient clusters are close to each other, slightly overlapping yet hinting to a trajectory pattern (**Figure 6**). The HER2/luminal cluster contains two patients classified as basal-like subtype, consistently with the findings of (44).

4. DISCUSSION

4.1. Background and Related Work

Ritchie et al. (45) defined omics data integration as the combination of multiple omics datasets that can be used for the development of models to predict complex traits or phenotypes. The problem of data integration in computational biology is far from having a consolidated and shared solution. Many long-standing obstacles are still far from being overcome, and the increasing availability of data [e.g., TCGA, (46)] and computational tools [see for instance (47–51) and <https://github.com/mikelove/awesome-multi-omics>], also interactive [e.g., (52)], is raising new issues that need to be addressed. In fact, not only are existing datasets still lacking standardization protocols to deal with their complexity and heterogeneity, but also the reliability, reproducibility and interpretability of new computational methods are emerging as urgent and relevant questions (53). Moreover, modern technologies allow the rapid extraction of high-dimensional, high-throughput features from different sources (e.g., gene expression, DNA sequencing, metabolomics, or high-resolution images), which in turn require collaboration between biologists, computer scientists, physicians and other experts. The lack of common methodologies and terminologies can transform this synergy into a further level of complexity in the process of data integration (54). As

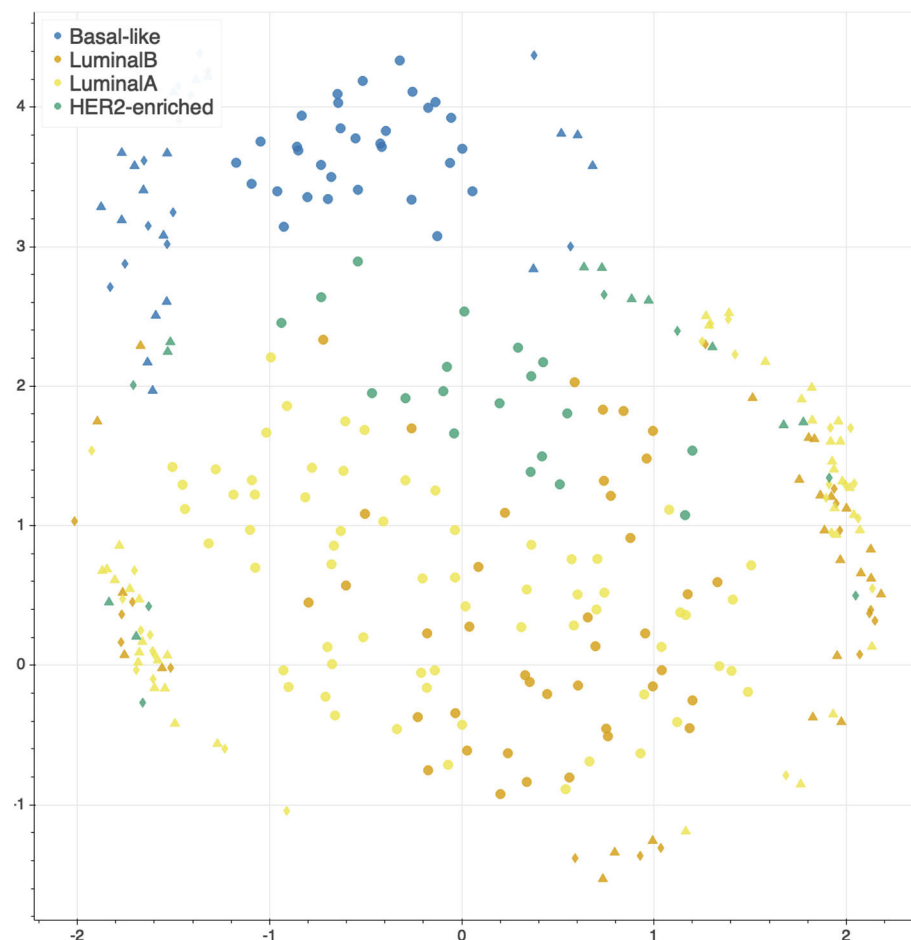


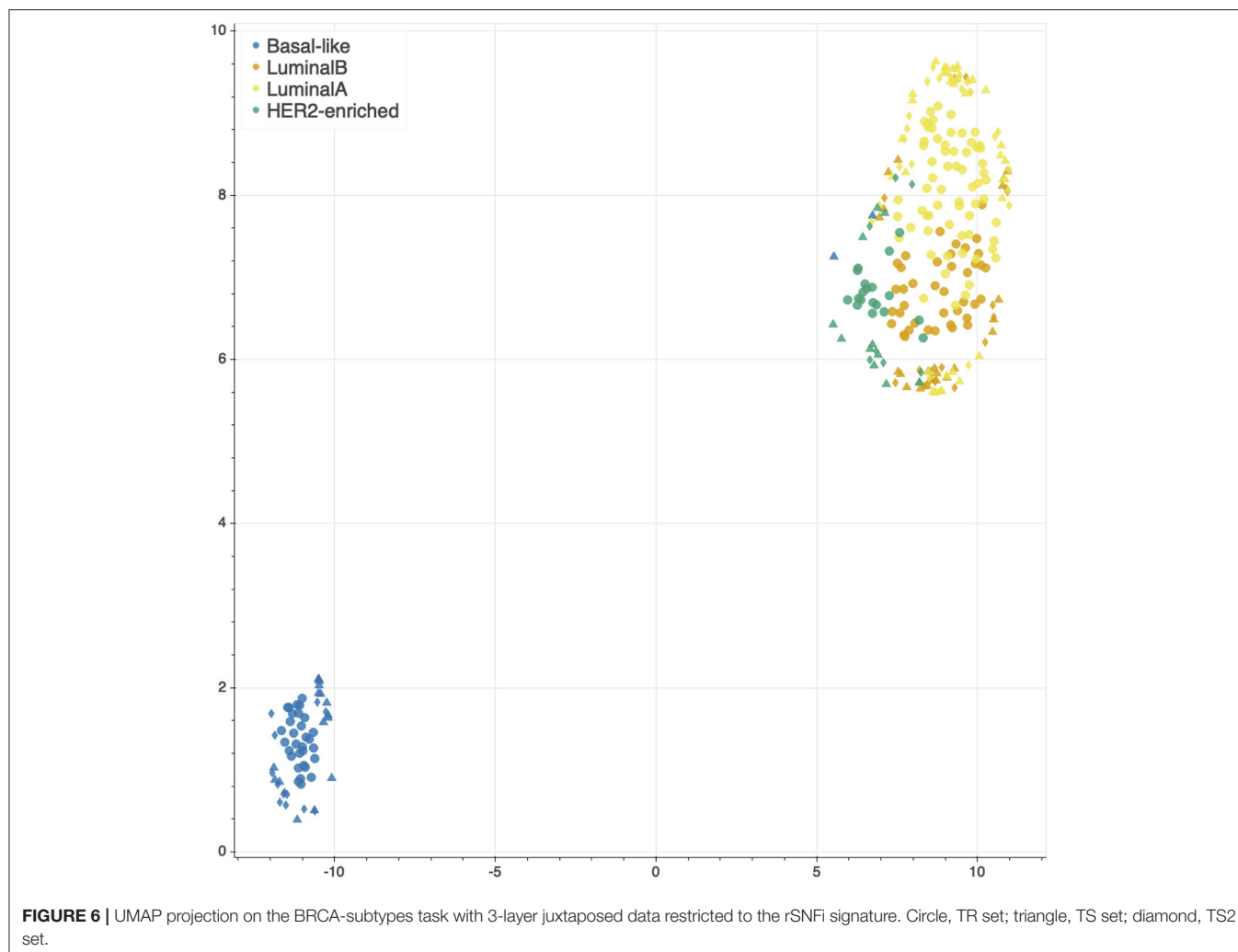
FIGURE 5 | UMAP projection on the BRCA-subtypes task with 3-layer juxtaposed data. Circle, TR set; triangle, TS set; diamond, TS2 set.

observed in (55, 56), specific technological limits, noise levels and variability ranges affect the different omics, and thus confounding the underlying biological signals, yielding that really integrative analysis is still very rare, while different methods often discover different kinds of patterns, as evidenced by the lack of consistency in the published results, although efforts in this direction have started appearing (57, 58).

Indeed, the underlying hypothesis of multi-omics integration is that different omics data can provide complementary information (56) [although sometimes redundant (9)], and thus a broader insight with respect to single-layer analysis, for a better understanding of disease mechanisms (59). This assumption has been confirmed by multiple studies on diverse diseases, such as cardiovascular disease (60), diabetes (61), liver disease (62), or mitochondrial diseases (63), and also longitudinally (64), suggesting that the more complex the disease the more advantageous the integration. As the co-occurrence of multiple causes and correlated events is a well-known characteristic of tumorigenesis and cancer development, the integration of data generated from multiple sources can thus be particularly useful for the identification of cancer hallmarks (65–68).

Many computational strategies have been introduced that combine multiple types of data to identify novel biomarkers and thus to predict a phenotype of interest or drive the development of intervention protocols. Given the heterogeneity of data and tasks, these techniques deal with the data integration at different levels of the learning process: (i) by concatenating the features before fitting a model (early-integration), (ii) by incorporating the integration step into the model training (intermediate-integration), or (iii) by combining the outputs of distinct models for the final prediction (late-integration) (69, 70).

In the early-integration approach, also known as juxtaposition-based, the multi-omics datasets are first concatenated into one matrix. To deal with the high-dimensionality of the joint dataset, these methods generally adopt matrix factorization (55, 56, 58, 71), statistical (47, 49, 58, 60, 62, 72–76), and machine learning tools (58, 76, 77). Alternatively, data models relying on polyplot approaches can be used especially in (bio)informatics applications (78, 79). Although the dimensionality reduction procedure is necessary and may improve the predictive performance, it can also cause the loss of key information (69).



Moreover, biomarkers identified purely on a computational statistics rationale from meta-omics features often lack biological plausibility (80).

In order to maximize the contribution of the single-omics layer, the late-integration methods first model each dataset individually, and then merge or average the results; they are also known as model-driven (70, 81). Although these techniques avoid the pre-selection of the features, they do not leverage the hidden correlations between the data, posing again the risk of signal loss (80, 82).

The intermediate-integration strategies aim at developing a joint model that accounts for the correlation between the omics layers, to boost their combined predictive power (83). Among these methods, the network-based models refer to the reconstruction of a graph representing the complex biological interactions (76, 84), known or predicted, between the variables to discover novel informative relationships (85). They have successfully been applied in cancer research for the identification of pan-cancer drug targets (86), the detection of subtype-specific pathways (83, 87) and of genetic aberrations (88), or the stratification of cancer patients (89–91). In particular, Koh et al. (44) predicted breast cancer subtypes by applying a modified

shrunk centroid method in the development of their network-based tool, iOmicsPASS. Further, breast cancer datasets in TCGA represent a benchmark for integrative models (92–94), as well as AML (95).

More recently, the success of deep learning algorithms in various bioinformatics fields (96) prompted the adoption of deep neural networks for omics-integration in precision oncology. Autoencoders and convolutional neural networks have been effectively trained for the prediction of prognostic outcomes (9, 97), response to chemotherapeutic drugs (50), and gene targeting (98), by adopting either an early-integration (9, 98) or a late-integration (50, 97). Although deep learning models hold the potential to include image-derived features in the integration workflow, they suffer from interpretability and generalization issues (99).

Although it is clear that no single method is consistently preferable, and that most of the proposed approaches are task and/or data dependent (80), the complexity of tumor analysis suggests that network-based approaches are needed (87, 100).

In this context, it is clear that omics-integration is one of the most promising and demanding challenges of the modern bioinformatics, and that there is an urgent need to prove the

reproducibility, interpretability, and generalization capability of the proposed methods (85, 101).

4.2. Integrative Network Fusion

We present the INF framework for the characterization of cancer patient phenotypes by integrated multi-omics signatures, combining an improved version of a state-of-the-art integration technique (5) with predictive models developed inside a Data Analysis Plan (6) for machine learning. The framework is applied to TCGA data to predict clinically relevant patient phenotypes such as the overall survival or cancer subtypes.

The simplest approach for multi-omics data integration consists in juxtaposition of normalized measurements into one joint matrix, followed by the development of a predictive model. Juxtaposition-based integration is considered as a baseline technique, since it is the most naïve approach to combine two datasets; moreover, it enables to identify multi-omics signatures by borrowing discriminatory strength from information derived by all datasets. Juxtaposition further dilutes the already possible low signal-to-noise ratio in each data type, affecting the understanding of the biological interactions at the different omics levels.

Conversely, the INF method for omics data integration is an improvement of the popular Similarity Network Fusion (SNF) approach (5), which has inspired several studies in the scientific literature, specifically in cancer genomics (77, 87, 102–106). SNF maximizes the shared or correlated information between multiple datasets by combining data through inference of a joint network-based model, accounting for how informative each data type is to the observed similarity between samples.

Two innovative solutions have been implemented in this study: (i) we devised a SNF-based procedure to rank variables according to their importance in clustering samples with similar phenotypes; and (ii) predictive models were developed exploiting the SNF-ranked variables, inside a rigorous Data Analysis Plan which ensures reproducibility (6, 19).

The performance of INF was assessed both in terms of statistical properties as well as biological interest. Concerning the statistical aspect, INF was compared with predictive models developed on the juxtaposed datasets (juXT technique), as well as on the single-layer datasets. With INF, smaller signature sizes were systematically derived to achieve comparable or even better performance both in cross-validation and in test. This is an added value for INF, as biological validation of biomarkers can definitely benefit from signatures of small size in terms of both costs and required time. This main achievement is mainly due to the novel rSNF ranking, which increases the signal-to-noise ratio from the combined layers by prioritizing the most discriminant biomarkers in terms of network mutual information. rSNF exploits two main SNF advantages: integration of heterogeneous data and clustering of sample networks. The main peculiarity of the SNF integrative procedure is its robustness to noise (5), because weak similarities among samples (low-weight edges) disappear, except for low-weight edges supported by all networks, which are conserved depending on how tightly connected their neighborhoods are across networks. Moreover, the rSNFi step further increases the signal-to-noise ratio by training a predictive

classifier on multi-omics juxtaposed data restricted to the top-ranked biomarkers shared by juXT and rSNF models. The resulting signatures are compact in size (up to 99% reduction w.r.t. juXT) while allowing predictive models to achieve equal or better performance compared to naïve juxtaposition or the single layers alone. While a comprehensive evaluation of the biological meaning of the signatures identified through the INF framework is beyond the scope of this work, we assessed their general validity with a thorough literature search. Our investigation shows that the signatures identified through the INF framework include biological markers that are relevant in the tasks under analysis and are consistent with previously published data. Further, as in (9), the largest contribution in the biomarkers' lists is provided by gene expression, while epigenomics, proteomics and miRNA transcriptomics play a minor role.

It should be noted that, especially in computational biology, multicollinearity between pairs of predictors and/or layers is intrinsic in the problem. Nevertheless, most machine learning models are indeed designed to identify the relevant predictors even in the presence of strong linear or non-linear correlations, provided that an appropriate DAP, feature ranking method, and diagnostic tools (e.g., random labels) are adopted against selection bias. To this aim, the application of a DAP derived from the MAQC-II initiative for model selection is a core attribute of the INF framework.

A fair comparison of INF results with other integration methods is currently unfeasible due to the number and variety of computational pipelines with dissimilar datasets, preprocessing methods, data analysis plans, and performance metrics.

This work is based on the original R implementation of the SNF algorithm (5). However, we are aware that Open Source implementations exist in other programming languages, in particular *snfpy* for Python (107). In a future release of the INF workflow, we plan to migrate the SNF-related parts to *snfpy* or a similar Python-based implementation, in order to drop the dependency on R and to potentially improve the overall performance.

In its current version, the INF framework supports the integration of two or more one-dimensional omics layers. As part of our future effort we will add support for the integration of medical imaging layers, for example leveraging the extraction of histopathological features from whole slide images by deep learning (10) or using radiomics or deep features from radiological images (11). In both cases, further issues will emerge from the interactions between the omics and the non-omics data, needing particular care in the integration (12).

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary files, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

CA, LT, and GJ: conceptualization. MC, NB, AM, AZ, LT, CA, and GJ: methodology. MF: interpretation. GJ:

coordination. MC, NB, AM, MF, GJ, and CF: writing. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

The authors wish to thank Dr. Valerio Maggio for helpful discussions on aspects of the machine learning workflow and for paper proofreading. The results published here are in whole or

part based upon data generated by the TCGA Research Network: <https://www.cancer.gov/tcga>.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fonc.2020.01065/full#supplementary-material>

REFERENCES

- Benton D. Bioinformatics-principles and potential of a new multidisciplinary tool. *Trends Biotechnol.* (1996) 14:261–72. doi: 10.1016/0167-7799(96)10037-8
- Chung SY, Wong L. Kleisli: a new tool for data integration in biology. *Trends Biotechnol.* (1999) 17:351–5. doi: 10.1016/S0167-7799(99)01342-6
- Zandoná A. Predictive networks for multi meta-omics data integration (Doctoral Programme in Biomolecular Sciences). University of Trento. Trento (2017). Available online at: <http://eprints-phd.biblio.unitn.it/2547/>
- Trastulla L. Techniques of Integration for High-Throughput Omics Data. Department of Mathematics, University of Trento, Trento (2016).
- Wang B, Mezlini A, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods.* (2014) 11:333–7. doi: 10.1038/nmeth.2810
- The MAQC Consortium. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat Biotechnol.* (2010) 28:827–38. doi: 10.1038/nbt.1665
- The SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequence Quality Control consortium. *Nat. Biotechnol.* (2014) 32:903–14. doi: 10.1038/nbt.2957
- Shi L, Kusko R, Wolfinger RD, Haibe-Kains B, Fischer M, Sansone SA, et al. The international MAQC Society launches to enhance reproducibility of high-throughput technologies. *Nat Biotechnol.* (2017) 35:1127–8. doi: 10.1038/nbt.4029
- Chai H, Zhou X, Cui Z, Rao J, Hu Z, Yang Y. Integrating multi-omics data with deep learning for predicting cancer prognosis. *bioRxiv. [Preprint]*. (2019) 807214. doi: 10.1101/807214
- Bizzego A, Bussola N, Chierici M, Cristoforetti M, Francescato M, Maggio V, et al. Evaluating reproducibility of AI algorithms in digital pathology with DAPPER. *PLoS Comput. Biol.* (2019) 15:e1006269. doi: 10.1371/journal.pcbi.1006269
- Bizzego A, Bussola N, Salvalai D, Chierici M, Maggio V, Jurman G, et al. Integrating deep and radiomics features in cancer bioimaging. In: *IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. Siena (2019) p. 1–8. doi: 10.1101/568170
- López de Maturana E, Alonso L, Alarcón P, Martín-Antoniano IA, Pineda S, Piorio L, et al. Challenges in the integration of omics and non-omics data. *Genes.* (2019) 10:238. doi: 10.3390/genes10030238
- The Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med.* (2013) 368:2059–74. doi: 10.1056/NEJMoa1301689
- The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature.* (2013) 499:43. doi: 10.1038/nature12222
- The Cancer Genome Atlas Research Network. Comprehensive molecular portraits of human breast tumours. *Nature.* (2012) 490:61. doi: 10.1038/nature11412
- Kosinski M, Biecek P. RTCGA: The Cancer Genome Atlas Data Integration (2019). R package version 1.16.0. Available online at: <https://rtcg.github.io/RTCGA>
- Breiman L. Random forests. *Mach Learn.* (2001) 45:5–32. doi: 10.1023/A:1010933404324
- Cortes C, Vapnik VN. Support-vector networks. *Mach Learn.* (1995) 20:273–97. doi: 10.1007/BF00994018
- Zhang W, Yu Y, Hertwig F, Thierry-Mieg J, Zhang W, Thierry-Mieg D, et al. Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. *Genome Biol.* (2015) 16:133. doi: 10.1186/s13059-015-0694-1
- Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta.* (1975) 405:442–451. doi: 10.1016/0005-2795(75)90109-9
- Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics.* (2000) 16:412–24. doi: 10.1093/bioinformatics/16.5.412
- Gorodkin J. Comparing two K-category assignments by a K-category correlation coefficient. *Comput Biol Chem.* (2004) 28:367–74. doi: 10.1016/j.compbiolchem.2004.09.006
- Jurman G, Riccadonna S, Furlanello C. A comparison of MCC and CEN error measures in multi-class prediction. *PLoS ONE.* (2012) 7:e41882. doi: 10.1371/journal.pone.0041882
- Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics.* (2020) 21:6. doi: 10.1186/s12864-019-6413-7
- Jurman G, Merler S, Barla A, Paoli S, Galea A, Furlanello C. Algebraic stability indicators for ranked lists in molecular profiling. *Bioinformatics.* (2008) 24:25864. doi: 10.1093/bioinformatics/btm550
- Jurman G, Riccadonna S, Visintainer R, Furlanello C. Algebraic comparison of partial lists in bioinformatics. *PLoS ONE.* (2012) 7:e36540. doi: 10.1371/journal.pone.0036540
- Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics.* (2012) 28:2520–2. doi: 10.1093/bioinformatics/bts480
- Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine (Erratum). *Bioinformatics.* (2018) 34:3600. doi: 10.1093/bioinformatics/bty350
- Oliphant TE. A Guide to NumPy. Vol. 1. Trelgol Publishing (2006).
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res.* Trelgol Publishing (2011) 12:2825–30.
- R Core Team. R: A Language and Environment for Statistical Computing. Vienna (2019). Available online at: <https://www.R-project.org/>
- Garczyk S, von Stillfried S, Antonopoulos W, Hartmann A, Schrauder MG, Fasching PA, et al. AGR3 in breast cancer: prognostic impact and suitable serum-based biomarker for early cancer detection. *PLoS ONE.* (2015) 10:e0122106. doi: 10.1371/journal.pone.0122106
- Potapenko IO, Lüders T, Russnes HG, Helland, Sørle T, Kristensen VN, et al. Glycan-related gene expression signatures in breast cancer subtypes; relation to survival. *Mol Oncol.* (2015) 9:861–76. doi: 10.1016/j.molonc.2014.12.013
- Thakkar A, Raj H, Ravishanker, Muthuvelan B, Balakrishnan A, Padigar M. High expression of three-gene signature improves prediction of relapse-free survival in estrogen receptor-positive and node-positive breast tumors. *Biomarker Insights.* (2015) 10:BMI.S30559. doi: 10.4137/BMI.S30559
- Guo Y, Yu P, Liu Z, Maimaiti Y, Chen C, Zhang Y, et al. Prognostic and clinicopathological value of GATA binding protein 3 in breast cancer:

- a systematic review and meta-analysis. *PLoS ONE*. (2017) 12:e0174843. doi: 10.1371/journal.pone.0174843
36. Milioli HH, Vimieiro R, Riveros C, Tishchenko I, Berretta R, Moscato P. The discovery of novel biomarkers improves breast cancer intrinsic subtype prediction and reconciles the labels in the METABRIC data set. *PLoS ONE*. (2015) 10:e0129711. doi: 10.1371/journal.pone.0129711
 37. Bai H, Zhou M, Zeng M, Han L. PLA2G4A is a potential biomarker predicting shorter overall survival in patients with Non-M3/NPM1 wildtype acute myeloid leukemia. *DNA Cell Biol*. (2020) 39:700–708. doi: 10.1089/dna.2019.5187
 38. Guo Y. Clinical significance of serum MicroRNA-203 in patients with acute myeloid leukemia. *Bioengineered*. (2019) 10:345–352. doi: 10.1080/21655979.2019.1652490
 39. Zheng YS, Zhang H, Zhang XJ, Feng DD, Luo XQ, Zeng CW, et al. MiR-100 regulates cell differentiation and survival by targeting RBP3, a phosphatase-like tumor suppressor in acute myeloid leukemia. *Oncogene*. (2012) 31:80–92. doi: 10.1038/ncr.2011.208
 40. Li SM, Zhao YQ, Hao YL, Liang YY. Upregulation of miR-504-3p is associated with favorable prognosis of acute myeloid leukemia and may serve as a tumor suppressor by targeting MTHFD2. *Eur Rev Med Pharmacol Sci*. (2019) 23:1203–13. doi: 10.26355/eurrev_201902_17013
 41. Uhlen M, Zhang C, Lee S, Sjöstedt E, Fagerberg L, Bidkhorji G, et al. A pathology atlas of the human cancer transcriptome. *Science*. (2017) 357:eaa2507. doi: 10.1126/science.aan2507
 42. McInnes L, Healy J, Saul N, Großberger L. UMAP: uniform manifold approximation and projection. *J Open Source Softw*. (2018) 3:861. doi: 10.21105/joss.00861
 43. McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. *arXiv [preprint.] arXiv:1802.03426* (2018).
 44. Koh HWL, Fermin D, Vogel C, Pui Choi K, Ewing RM, Choi H. iOmicsPASS: network-based integration of multiomics data for predictive subnetwork discovery. *NPJ Syst Biol Appl*. (2019) 5:22. doi: 10.1038/s41540-019-0099-y
 45. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet*. (2015) 16:85–97. doi: 10.1038/nrg3868
 46. Conesa A, Beck S. Making multi-omics data accessible to researchers. *Sci Data*. (2019) 6:251. doi: 10.1038/s41597-019-0258-4
 47. Rohart F, Gautier B, Singh A, Lê Cao KA. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput Biol*. (2017) 13:e1005752. doi: 10.1371/journal.pcbi.1005752
 48. Ulfenborg B. Vertical and horizontal integration of multiomics data with miodin. *BMC Bioinformatics*. (2019) 20:649. doi: 10.1186/s12859-019-3224-4
 49. Meng C, Basunia A, Peters B, Gholami AM, Kuster B, Culhane AC. MOGSA: integrative single sample gene-set analysis of multiple omics data. *Mol Cell Proteomics*. (2019) 18(8 Suppl 1):S153–68. doi: 10.1074/mcp.TIR118.010251
 50. Sharifi-Noghabi H, Zolotareva O, Collins CC, Ester M. MOLI: multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics*. (2019) 35:i501–9. doi: 10.1093/bioinformatics/btz318
 51. Zanfardino M, Franzese M, Pane K, Cavaliere C, Monti S, Esposito G, et al. Bringing radiomics into a multi-omics framework for a comprehensive genotype-phenotype characterization of oncological diseases. *J Transl Med*. (2019) 17:337. doi: 10.1186/s12967-019-2073-2
 52. Netanel D, Stern N, Laufer I, Shamir R. PROMO: an interactive tool for analyzing clinically-labeled multi-omic cancer datasets. *BMC Bioinformatics*. (2019) 20:732. doi: 10.1186/s12859-019-3142-5
 53. Lionelli S. Philosophy of biology: the challenges of big data biology. *eLife*. (2019) 8:e47381. doi: 10.7554/eLife.47381
 54. Schneider MV, Jimenez RC. Teaching the fundamentals of biological data integration using classroom games. *PLoS Comput Biol*. (2012) 8:e1002789. doi: 10.1371/journal.pcbi.1002789
 55. Pucher BM, Zeleznik OA, Thallinger GG. Comparison and evaluation of integrative methods for the analysis of multilevel omics data: a study based on simulated and experimental cancer data. *Brief Bioinform*. (2018) 28:1–11. doi: 10.1093/bib/bby027
 56. Cantini L, Zakeri P, Hernandez C, Naldi A, Thieffry D, Remy E, et al. Benchmarking joint multi-omics dimensionality reduction approaches for cancer study. *bioRxiv. [Preprint]*. (2020) 905760. doi: 10.1101/2020.01.14.905760
 57. McCabe SD, Lin DY, Love MI. Consistency and overfitting of multi-omics methods on experimental data. *Brief Bioinform*. (2019) bbz070. doi: 10.1093/bib/bbz070
 58. Wu C, Zhou F, Ren J, Li X, Jiang Y, Ma S. A selective review of multi-level omics data integration using variable selection. *High Throughput*. (2019) 8:4. doi: 10.3390/ht8010004
 59. Karczewski K, Snyder M. Integrative omics for health and disease. *Nat Rev Genet*. (2018) 19:299–310. doi: 10.1038/nrg.2018.4
 60. Leon-Mimila P, Wang J, Huertas-Vazquez A. Relevance of multi-omics studies in cardiovascular diseases. *Front Cardiovasc Med*. (2019) 6:91. doi: 10.3389/fcvm.2019.00091
 61. Prélôt L, Draisma H, Anasanti M, Balkhiyarova Z, Wielscher M, Yengo L, et al. Machine learning in multi-omics data to assess longitudinal predictors of glycaemic trait levels. *bioRxiv. [Preprint]*. (2018) 358390. doi: 10.1101/358390
 62. Del Chierico F, Nobili V, Vernocchi P, Russo A, De Stefanis C, Gnani D, et al. Gut microbiota profiling of pediatric nonalcoholic fatty liver disease and obese patients unveiled by an integrated meta-omics-based approach. *Hepatology*. (2017) 65:451–64. doi: 10.1002/hep.28572
 63. Khan S, Ince-Dunn G, Suomalainen A, Elo LL. Integrative omics approaches provide biological and clinical insights: examples from mitochondrial diseases. *J Clin Invest*. (2020) 130:20–8. doi: 10.1172/JCI129202
 64. Tarazona S, Balzano-Nogueira L, Conesa A. Chapter eighteen - multiomics data integration in time series experiments. In: Jaumot J, Bedia C, Tauler R, editors. *Comprehensive Analytical Chemistry*, Vol. 82. Elsevier (2018). p. 505–32. doi: 10.1016/bs.coac.2018.06.005
 65. Chakraborty S, Hosen MI, Ahmed M, Shekhar HU. Onco-Multi-OMICS approach: a new frontier in cancer research. *BioMed Res Int*. (2018) 2018:9836256. doi: 10.1155/2018/9836256
 66. Gallo Cantafio ME, Grillone K, Caracciolo D, Scionti F, Arbitrio M, Barbieri V, et al. From single level analysis to multi-omics integrative approaches: a powerful strategy towards the precision oncology. *High Throughput*. (2018) 7:33. doi: 10.3390/ht7040033
 67. Sathyanarayanan A, Gupta R, Thompson EW, Nyholt DR, Bauer DC, Nagaraj SH. A comparative study of multi-omics integration tools for cancer driver gene identification and tumour subtyping. *Brief Bioinform*. (2019) bbz121. doi: 10.1093/bib/bbz121
 68. Liu SH, Shen PC, Chen CY, Hsu AN, Cho YC, Lai YL, et al. DriverDBv3: a multi-omics database for cancer driver gene research. *Nucleic Acids Res*. (2019) 48:D863–70. doi: 10.1093/nar/gkz964
 69. Li Y, Wu FX, Ngom A. A review on machine learning principles for multi-view biological data integration. *Brief Bioinform*. (2018) 19:325–40. doi: 10.1093/bib/bbw113
 70. Vlahou A, Magni F, Mischak H, Zoidakis J. *Integration of Omics Approaches and Systems Biology for Clinical Applications*. Hoboken, NJ: John Wiley & Sons (2018). doi: 10.1002/9781119183952
 71. Stein-O'Brien GL, Arora R, Culhane AC, Favorov AV, Garmire LX, Greene CS, et al. Enter the matrix: factorization uncovers knowledge from omics. *Trends Genet*. (2018) 34:790–805. doi: 10.1016/j.tig.2018.07.003
 72. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, et al. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol*. (2018) 14:e8124. doi: 10.15252/msb.20178124
 73. Dao MC, Sokolovska N, Brazeilles R, Affeldt S, Pelloux V, Prifti E, et al. A data integration multi-omics approach to study calorie restriction-induced changes in insulin sensitivity. *Front Physiol*. (2019) 9:1958. doi: 10.3389/fphys.2018.01958
 74. Zeng ISL, Lumley T. Review of statistical learning methods in integrated omics studies (an integrated information science). *Bioinform Biol Insights*. (2018) 12:1–16. doi: 10.1177/1177932218759292
 75. Qiu C, Yu F, Su K, Zhao Q, Zhang L, Xu C, et al. Multi-omics data integration for identifying osteoporosis biomarkers and their biological interaction and causal mechanisms. *IScience*. (2020) 23:100847. doi: 10.1016/j.isci.2020.100847

76. Misra BB, Langefeld C, Olivier M, Cox LA. Integrated omics: tools, advances and future approaches. *J Mol Endocrinol.* (2019) 62:R21–45. doi: 10.1530/JME-18-0055
77. Speicher NK, Pfeifer N. Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. *Bioinformatics.* (2015) 31:i268–75. doi: 10.1093/bioinformatics/btv244
78. Chromiak M., Stencel K. A data model for heterogeneous data integration architecture. In: Kozielski S, Mrozek D, Kasprowski P, Mażyński-Mrozek B, Kozłowska D, editors. *Beyond Databases, Architectures, and Structures. BDAS 2014. Communications in Computer and Information Science*, Vol. 424. Springer (2014). p. 547–56. doi: 10.1007/978-3-319-06932-6_53
79. Reisman S, Hatzopoulos T, Läufer K, Thiruvathukal GK, Putonti C. A polyglot approach to bioinformatics data integration: a phylogenetic analysis of HIV-1. *Evol Bioinform Online.* (2016) 12:23–7. doi: 10.4137/EBO.S32757
80. Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Res.* (2018) 46:10546–62. doi: 10.1093/nar/gky889
81. Marin de Mas I. Chapter sixteen - multiomic data integration and analysis via model-driven approaches. In: Jaumot J, Bedia C, Tauler R, editors. *Comprehensive Analytical Chemistry*, Vol. 82. Elsevier (2018). p. 447–76. doi: 10.1016/bs.coac.2018.07.005
82. Gadepally V, Mattson T, Stonebraker M, Wang F, Luo G, Laing Y, et al. Heterogeneous data management, polystores, and analytics for healthcare: VLDB 2019. In: Workshops, Poly and DMAH. Los Angeles, CA: Springer Nature (2019). doi: 10.1007/978-3-030-33752-0
83. Vantaku V, Dong J, Ambati CR, Perera D, Donepudi SR, Amara CS, et al. Multi-omics integration analysis robustly predicts high-grade patient survival and identifies CPT1B effect on fatty acid metabolism in bladder cancer. *Clin Cancer Res.* (2019) 25:3689–701. doi: 10.1158/1078-0432.CCR-18-1515
84. Zhou G, Li S, Xia J. Network-based approaches for multi-omics integration. In: Li S, editor. *Computational Methods and Data Analysis for Metabolomics. Methods in Molecular Biology*, Vol. 2104. New York, NY: Humana (2020). p. 469–87. doi: 10.1007/978-1-0716-0239-3_23
85. Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, et al. Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics.* (2016) 17:S15. doi: 10.1186/s12859-015-0857-9
86. do Valle ÍF, Menichetti G, Simonetti G, Bruno S, Zironi I, Fernandes Durso D, et al. Network integration of multi-tumour omics data suggests novel targeting strategies. *Nat Commun.* (2018) 9:4514. doi: 10.1038/s41467-018-06992-7
87. Verbeke LPC, Van den Eynden J, Fierro AC, Demeester P, Fostier J, Marchal K. Pathway relevance ranking for tumor samples through network-based data integration. *PLoS ONE.* (2015) 10:e0133503. doi: 10.1371/journal.pone.0133503
88. Dimitrakopoulos C, Kumar Hindupur S, Häfliger L, Behr J, Montazeri H, Hall MN, et al. Network-based integration of multi-omics data for prioritizing cancer genes. *Bioinformatics.* (2018) 34:2441–8. doi: 10.1093/bioinformatics/bty148
89. Zhao L, Yan H. MCNF: a novel method for cancer subtyping by integrating multi-omics and clinical data. *IEEE/ACM Trans Comput Biol Bioinform.* (2019). doi: 10.1109/TCBB.2019.2910515. [Epub ahead of print].
90. Rappoport N, Shamir R. NEMO: cancer subtyping by integration of partial multi-omic data. *Bioinformatics.* (2019) 35:3348–56. doi: 10.1093/bioinformatics/btz058
91. Yang B, Zhang Y, Pang S, Zhang X, Zhao X, Han M. Integrating multi-omic data with deep subspace fusion clustering for cancer subtype prediction. *IEEE/ACM Trans Comput Biol Bioinform.* (2019). doi: 10.1109/TCBB.2019.2951413. [Epub ahead of print].
92. Xu A, Chen J, Peng H, Han GQ, Cai H. Simultaneous interrogation of cancer omics to identify subtypes with significant clinical differences. *Front Genet.* (2019) 10:236. doi: 10.3389/fgene.2019.00236
93. Kechavarzi BD, Wu H, Doman TN. Bottom-up, integrated -omics analysis identifies broadly dosage-sensitive genes in breast cancer samples from TCGA. *PLoS ONE.* (2019) 14:e0210910. doi: 10.1371/journal.pone.0210910
94. Kalecky K, Modisette R, Pena S, Cho YR, Taube J. Integrative analysis of breast cancer profiles in TCGA by TNBC subgrouping reveals novel microRNA-specific clusters, including miR-17-92a, distinguishing basal-like 1 and basal-like 2 TNBC subtypes. *BMC Cancer.* (2020) 20:141. doi: 10.1186/s12885-020-6600-6
95. Mehtonen J, Pölönen P, Häyrynen S, Dufva O, Lin J, Liuksiala T, et al. Data-driven characterization of molecular phenotypes across heterogeneous sample collections. *Nucleic Acids Res.* (2019) 47:e76. doi: 10.1093/nar/gkz281
96. Li Y, Huang C, Ding L, Li Z, Pan Y, Gao X. Deep learning in bioinformatics: introduction, application, and perspective in the big data era. *Methods.* (2019) 166:4–21. doi: 10.1016/j.ymeth.2019.04.008
97. Poirion O, Chaudhary K, Huang S, Garmire LX. Multi-omics-based pan-cancer prognosis prediction using an ensemble of deep-learning and machine-learning models. *medRxiv. [Preprint].* (2019) 19010082. doi: 10.1101/19010082
98. Peng C, Zheng Y, Huang DS. Capsule Network based Modeling of Multi-omics Data for Discovery of Breast Cancer-related Genes. *IEEE/ACM Trans Comput Biol Bioinform.* (2019). doi: 10.1109/TCBB.2019.2909905. [Epub ahead of print].
99. Hériché JK, Alexander S, Ellenberg J. Integrating imaging and omics: computational methods and challenges. *Annu Rev Biomed Data Sci.* (2019) 2:175–97. doi: 10.1146/annurev-biodatasci-080917-013328
100. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep.* (2019) 9:5233. doi: 10.1038/s41598-019-41695-z
101. Yu XT, Zeng T. Integrative analysis of omics big data. In: Huang T, editor. *Computational Systems Biology. Methods in Molecular Biology*, Vol. 1754. New York, NY: Humana Press (2018). p. 109–35. doi: 10.1007/978-1-4939-7717-8_7
102. Chiu AM, Mitra M, Boymoushakian L, Collier HA. Integrative analysis of the inter-tumoral heterogeneity of triple-negative breast cancer. *Sci Rep.* (2018) 8:11807. doi: 10.1038/s41598-018-29992-5
103. Cavalli FMG, Remke M, Rampasek L, Peacock J, Shih DJH, Luu B, et al. Intertumoral heterogeneity within medulloblastoma subgroups. *Cancer Cell.* (2017) 31:737–54.e6. doi: 10.1016/j.ccell.2017.05.005
104. Jiang YZ, Ma D, Suo C, Shi J, Xue M, Hu X, et al. Genomic and transcriptomic landscape of triple-negative breast cancers: subtypes and treatment strategies. *Cancer Cell.* (2019) 35:428–40.e5. doi: 10.1016/j.ccell.2019.02.001
105. Pitroda SP, Weichselbaum RR. Integrated molecular and clinical staging defines the spectrum of metastatic cancer. *Nat Rev Clin Oncol.* (2019) 16:581–8. doi: 10.1038/s41571-019-0220-6
106. Ma T, Zhang A. Affinity network fusion and semi-supervised learning for cancer patient clustering. *Methods.* (2018) 145:16–24. doi: 10.1016/j.ymeth.2018.05.020
107. Markello R. *snfpy: Similarity Network Fusion in Python.* (2019). Available online at: <https://snfpy.readthedocs.io/en/latest/>

Conflict of Interest: AZ was employed by the company NIDEK Technologies Srl. CF was employed by the company HK3 Lab.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Chierici, Bussola, Marcolini, Francescato, Zandonà, Trastulla, Agostinelli, Jurman and Furlanello. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Integrated Multi-Omics Analyses in Oncology: A Review of Machine Learning Methods and Tools

Giovanna Nicora^{1†}, Francesca Vitali^{2,3,4†}, Arianna Dagliati^{1,5,6†}, Nophar Geifman^{5,6} and Riccardo Bellazzi^{1*}

¹ Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Pavia, Italy, ² Center for Innovation in Brain Science, University of Arizona, Tucson, AZ, United States, ³ Department of Neurology, College of Medicine, University of Arizona, Tucson, AZ, United States, ⁴ Center for Biomedical Informatics and Biostatistics, University of Arizona, Tucson, AZ, United States, ⁵ Centre for Health Informatics, The University of Manchester, Manchester, United Kingdom, ⁶ The Manchester Molecular Pathology Innovation Centre, The University of Manchester, Manchester, United Kingdom

OPEN ACCESS

Edited by:

Francesca Finotello,
Innsbruck Medical University, Austria

Reviewed by:

Federica Eduati,
Eindhoven University of
Technology, Netherlands
Giuseppe Jurman,
Fondazione Bruno Kessler (FBK), Italy

*Correspondence:

Riccardo Bellazzi
riccardo.bellazzi@unipv.it

[†]These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

Received: 30 January 2020

Accepted: 26 May 2020

Published: 30 June 2020

Citation:

Nicora G, Vitali F, Dagliati A,
Geifman N and Bellazzi R (2020)
Integrated Multi-Omics Analyses in
Oncology: A Review of Machine
Learning Methods and Tools.
Front. Oncol. 10:1030.
doi: 10.3389/fonc.2020.01030

In recent years, high-throughput sequencing technologies provide unprecedented opportunity to depict cancer samples at multiple molecular levels. The integration and analysis of these multi-omics datasets is a crucial and critical step to gain actionable knowledge in a precision medicine framework. This paper explores recent data-driven methodologies that have been developed and applied to respond major challenges of stratified medicine in oncology, including patients' phenotyping, biomarker discovery, and drug repurposing. We systematically retrieved peer-reviewed journals published from 2014 to 2019, select and thoroughly describe the tools presenting the most promising innovations regarding the integration of heterogeneous data, the machine learning methodologies that successfully tackled the complexity of multi-omics data, and the frameworks to deliver actionable results for clinical practice. The review is organized according to the applied methods: Deep learning, Network-based methods, Clustering, Features Extraction, and Transformation, Factorization. We provide an overview of the tools available in each methodological group and underline the relationship among the different categories. Our analysis revealed how multi-omics datasets could be exploited to drive precision oncology, but also current limitations in the development of multi-omics data integration.

Keywords: multi-omics, machine learning, tools, systematic review, oncology, cancer

INTRODUCTION

The integration and analysis of high-throughput molecular assays is a major focus for precision medicine in enabling the understanding of patient and disease specific variations. Integrated approaches allow for comprehensive views of genetic, biochemical, metabolic, proteomic, and epigenetic processes underlying a disease that, otherwise, could not be fully investigated by single-omics approaches. Computational multi-omics approaches are based on machine learning techniques and typically aim at classifying patients into cancer subtypes (1–5), designed for biomarker discovery and drug repurposing (6, 7).

While complexities underling cancer still hampers our understanding of how this disease arises and progresses (8), multi-omics approaches have been suggested as promising tools to dissect patient's dysfunctions in multiple biological systems that may be altered by cancer mechanisms (9).

Several efforts have been made to generate comprehensive multi-omics profiles of cancer patients. The Cancer Genome Atlas (TCGA, <https://portal.gdc.cancer.gov/>) provides detailed clinical, genomics, transcriptomics, and proteomics data on about 20,000 subjects and plans to generate additional data in the next years for a variety of cancer types. Analysis of datasets generated by multi-omics sequencing requires the development of computational approaches spanning from data integration (10), statistical methods, and artificial intelligence systems to gain actionable knowledge from data.

Here we present a descriptive overview on recent multi-omics approaches in oncology, which summarizes current state-of-art in multi-omics data analysis, relevant topics in terms of machine learning approaches, and aims of each survey, such as disease subtyping, or patient similarity. We provide an overview on each methodology group, while then focusing on publicly available tools.

METHODS

Search Strategy

We retrieved publications by querying the Scopus database as: *(cancer OR tumor OR tumor OR oncolog*)AND(multi-omic* OR multiomic* OR mixomic*)AND("machine learning" OR "data fusion" OR "network analysis")*.

Eligibility Criteria

Since other review covered previous years (10, 11) we included peer-reviewed journal articles published from 2014 to 2020 (last query 04-22-2020). If a study appears in multiple publications, only the latest version was included. We selected relevant studies by screening titles and abstracts, then analyzing full-texts. We excluded papers accordingly to the following criteria:

- Review articles;
- Studies focused on non-human subjects;
- Studies intended to validate and/or apply previously developed tools;
- Studies published in conference proceedings.
- Studies that integrate different measurement of the same type of omics (such as, only proteomics measurement).

Categories and Analyses

For each article, we extracted the publication year and the number of citations. We categorize the selected publications according to:

- Data inputs (i.e., types of omics);
- Research Aims:
 1. Stratified Medicine for subgroup discovery: studies aimed at finding groups of patients that exhibit different therapeutic/prognostic outcomes;

2. Biomarker discovery: studies that detect -omics characteristics indicating a disease state;
3. Pathways analysis: studies aimed at discovering relation among -omics terms, such as genes or proteins in normal and cancer condition;
4. Drug repurposing/discovery: studies aimed at identifying new drugs to or existing effective drugs originally developed for other conditions;

- Methods and algorithms: Deep network, Networks-based methods (Bayesian and Heuristic Networks), Clustering, Features Extraction, Feature Transformation, Factorization.

We highlight successful approaches for each criterion and identify promising ones that are either nascent or unexplored as potential opportunities.

RESULTS

We retrieved 270 papers. The Scopus query did not retrieve 24 relevant works that were added manually based on our previous knowledge. After a screening of papers' abstracts, 58 papers meeting our criteria were selected. Retrieved papers were organized into a matrix table (**Table 1**) and analyzed with respect to the aforementioned categories. As highlighted in **Figure 1A**, categories are not mutually exclusive, thus we show links between groups, which relate papers applying multiple methods. **Figure 1B** depicts all considered publications by year of publication and the Field-Weighted Citation Impact, a metric that allows comparison of papers accounting for year of publication and number citations. Studies are shown with different colors and shapes according to method used and the aim/output type.

In the following sections, we describe the methodological categories that emerged from our literature review. For each methodological category, particular emphasis is placed on studies providing tools that can be exploited by other users, either with their own data or to reproduce their results.

Network-Based Methods

Network-based approaches were exploited to detect, reconstruct and study interactions among sub network modules (13, 19, 22, 25, 40); to assess functional correlation among multi-omics entities (12, 14, 20, 55, 61, 62); to integrate and fuse networks to create comprehensive view of a disease (16, 24, 32, 37, 41, 63, 65). A few work leverage Bayesian methods (4, 34) or Markov models (17, 67).

Some approaches integrate network analysis within frameworks that apply multiple algorithms (35, 51, 58). In (51) a multi-platform analysis exploited for profiling pancreatic adenocarcinoma, includes clustering and Similarity Network Fusion to integrate genomic, transcriptomic, and proteomic data from the different platforms. In (58) authors develop a framework for drug repurposing and multi-target therapies by constructing a protein network for the disease under study and fusing several data sources. In (27), a functional interaction network predicts variations in expressions caused by genomic alterations, and it is exploited to prioritize cancer genes. Few

TABLE 1 | Selected papers and categories.

References	References in Figure 1	Year	#Citation 22/04/2020	Scopus field-weighted citation impact	Method	Omics	Aim	Tool release
Agarwal et al. (12)	1	2015	2	0.34	Network	Genomics, transcriptomics	Biomarker discovery	
Amar and Shamir (13)	2	2014	16	0.70	Network	Proteomics, genomics	Pathways analysis	ModMap tool
Ao et al. (14)	3	2016	17	1.11	Network	Genomics, epigenomics	Subgroup identification	
Argelaquet et al. (15)	4	2019	57	14.40	Feature transformation	Transcriptomics, genomics	Subgroup identification	R package <i>MOFAtools</i>
Wang et al. (16)	5	2014	410	12.89	Network	Transcriptomics, epigenomics	Subgroup identification	R and MATLAB code http://compbio.cs.toronto.edu/SNF/
Beal et al. (17)	6	2018	2	1.25	Network	Transcriptomics, genomics	Subgroup identification	https://github.com/sysbio-curie/PROFILE
Benfeitas et al. (18)	7	2019	9	5.17	Clustering	Transcriptomics, proteomics, metabolomics	Subgroup identification	
Bonnet et al. (19)	8	2015	29	2.50	Network	Genomics, transcriptomics	Biomarker discovery	Lemon-Tree—command-line tool in Java http://lemon-tree.googlecode.com
Cancemi et al. (20)	9	2018	4	0.82	Network	Transcriptomics, proteomics	Pathways analysis	
Cavalli et al. (21)	10	2017	213	21.09	Clustering	Epigenomics, genomics, transcriptomics	Subgroup identification	
Champion et al. (22)	11	2018	6	1	Network	Genomics, epigenomics	Biomarker discovery	AMARETTO R package https://bitbucket.org/gevaertlab/pancanceramaretto
Chaudhary et al. (23)	12	2018	82	14.79	Deep network	Transcriptomics, epigenomics	Subgroup identification	
Cho et al. (24)	13	2016	48	6.65	Network	Genomics, proteomics	Pathways analysis	Mashup tool MATLAB code http://cb.csail.mit.edu/cb/mashup/
Costa et al. (25)	14	2018	4	0.58	Network	Genomics, epigenomics	Pathways analysis	
Costello et al. (26)	15	2014	271	14.12	Feature transformation	Genomics, transcriptomics, epigenomics, proteomics	Subgroup identification (drug response)	
Dimitrakopoulos et al. (27)	16	2018	29	6.67	Network	Genomics, transcriptomics, proteomics	Pathway analysis	https://github.com/cbg-ethz/netics
Drabovich et al. (28)	17	2019	1	0.53	Feature extraction	Transcriptomics, proteomics, secretomics, tissue specific	Subgroup identification	
Francescatto et al. (29)	18	2018	6	1.59	Deep network	Genomics, transcriptomics	Subgroup identification	
Gabasova et al. (30)	19	2017	6	0.86	Clustering	Transcriptomics, proteomics, epigenomics	Subgroup identification	Clusternomics R package https://github.com/evelinag/clusternomics
Gao et al. (31)	20	2019	0	0	Factorization	Transcriptomics, genomics	Biomarker discovery	
Griffin et al. (32)	21	2018	1	0.29	Network	Transcriptomics, epigenomics	Biomarker discovery	

(Continued)

TABLE 1 | Continued

References	References in Figure 1	Year	#Citation 22/04/2020	Scopus field-weighted citation impact	Method	Omics	Aim	Tool release
Hoadley et al. (33)	22	2014	668	32.88	Clustering	Proteomics, transcriptomics, genomics	Subgroup identification	
Hua et al. (34)	23	2016	2	0.17	Network	Genomics, epigenomics	Biomarker discovery	
Huang et al. (35)	24	2019	6	4.44	Network	Genomics, transcriptomics, epigenomics	Drug repurposing/discovery	DrugComboExplorer tool https://github.com/Roosevelt-PKU/drugcombinationprediction
Huang et al. (36)	25	2019	8	4.37	Deep network	Transcriptomics	Subgroup identification	SALMON source code https://github.com/huangzhii/SALMON/
Kim et al. (37)	26	2017	3	0.16	Network	Transcriptomics, proteomics	Drug repurposing/discovery	
Kim et al. (38)	27	2018	2	0.40	Feature extraction	Genomics, transcriptomics, epigenomics	Subgroup identification	
Kim et al. (39)	28	2019	0	0	Feature extraction	Genomics, transcriptomics	Pathways analysis	
Koh et al. (40)	29	2019	2	1.48	Network	Transcriptomics, proteomics	Subgroup identification	iOmicsPASS https://github.com/cssblab/iOmicsPASS
Lee et al. (41)	30	2018	21	3.46	Network	Genomics, transcriptomics	Drug repurposing/discovery	
Liang et al. (42)	31	2015	86	5.96	Deep network	Transcriptomics, epigenomics	Subgroup identification	
List et al. (3)	32	2014	20	2.51	Feature extraction	Transcriptomics, epigenomics	Subgroup identification	
Luo et al. (43)	33	2019	0	0	Clustering	Transcriptomics, genomics	Subgroup identification	
Ma and Zhang (44)	34	2018	4	0.71	Clustering	Transcriptomics, epigenomics	Similarity	AFN is part of the Bioconductor R package https://bioconductor.org/packages/release/bioc/html/ANF.html
Mariette and Villa-Vialaneix (45)	35	2018	8	1.90	Feature transformation	Transcriptomics, genomics	Subgroup identification	R package <i>mixKernel</i>
Meng et al. (46)	36	2014	79	5.29	Feature transformation	Transcriptomics, proteomics	Subgroup identification	R package <i>omicade4</i>
Mo et al. (47)	37	2017	18	7.03	Feature transformation	Transcriptomics, genomics	Subgroup identification	R package <i>iClusterPlus</i>
Nguyen et al. (48)	38	2017	20	2.03	Clustering	Transcriptomics, epigenomics, genomics	Subgroup identification	
O'Connell and Lock (49)	39	2016	13	1.21	Feature transformation	Transcriptomics, genomics	Subgroup identification	R Package <i>r.jive</i>
Pai et al. (50)	40	2019	6	5.23	Feature extraction	Transcriptomics, metabolomics, genomics	Similarity	

(Continued)

TABLE 1 | Continued

References	References in Figure 1	Year	#Citation 22/04/2020	Scopus field-weighted citation impact	Method	Omics	Aim	Tool release
Raphael et al. (51)	41	2017	269	26.77	Network	Transcriptomics, genomics, proteomics	Subgroup identification	
Rappoport et al. (52)	42	2019	2	1.48	Clustering	Transcriptomics, epigenomics	Subgroup identification	
Ray et al. (4)	43	2014	30	2.34	Bayesian network	Genomics, epigenomics	Biomarker discovery	MATLAB code https://sites.google.com/site/jointgenomics/
Rohart et al. (53)	44	2017	285	38.04	Feature transformation	Transcriptomics, genomics, proteomics, epigenomics	Subgroup identification	R package <i>Mixomics</i>
Sharifi-Noghabi et al. (54)	45	2019	2	6.91	Deep network	Genomics, transcriptomics	Subgroup identification (drug response)	https://github.com/hosseinshn/MOLI
Sehgal et al. (55)	46	2015	6	0.36	Network	Transcriptomics	Pathways analysis	
Song et al. (56)	47	2019	2	1.06	Feature transformation	Transcriptomics, genomics, proteomics	Biomarker discovery	R package <i>iProFun</i>
Speicher and Pfeifer (57)	48	2015	34	5.83	Clustering	Genomics, transcriptomics	Subgroup identification	
Vitali et al. (58)	49	2016	16	1.51	Network	Proteomics, transcriptomics	Drug repurposing/discovery	
Woo et al. (59)	50	2017	30	2.97	Clustering	Genomics, epigenomics	Subgroup identification	
Wu et al. (60)	51	2015	19	0.83	Clustering	Genomics, transcriptomics	Subgroup identification	
Yang et al. (61)	52	2019	2	1.23	Network	Epigenomics, transcriptomics	Biomarker discovery	
Yuan et al. (62)	53	2018	3	2.04	Network	Genomics, transcriptomics, epigenomics	Biomarker discovery	
Wang et al. (63)	54	2018	6	1	Network	Genomics, transcriptomics	Biomarker discovery	
Zhang et al. (64)	55	2018	9	1.58	Deep network	Transcriptomics, genomics	Subgroup identification	
Zhou et al. (65)	56	2015	2	0.18	Network	Genomics, epigenomics, proteomics	Biomarker discovery	
Zhu et al. (66)	57	2017	20	1.52	Feature transformation	Transcriptomics, genomics	Subgroup identification	
Žitnik and Zupan (67)	58	2015	14	2.50	Network	Transcriptomics, genomics	Biomarker discovery	

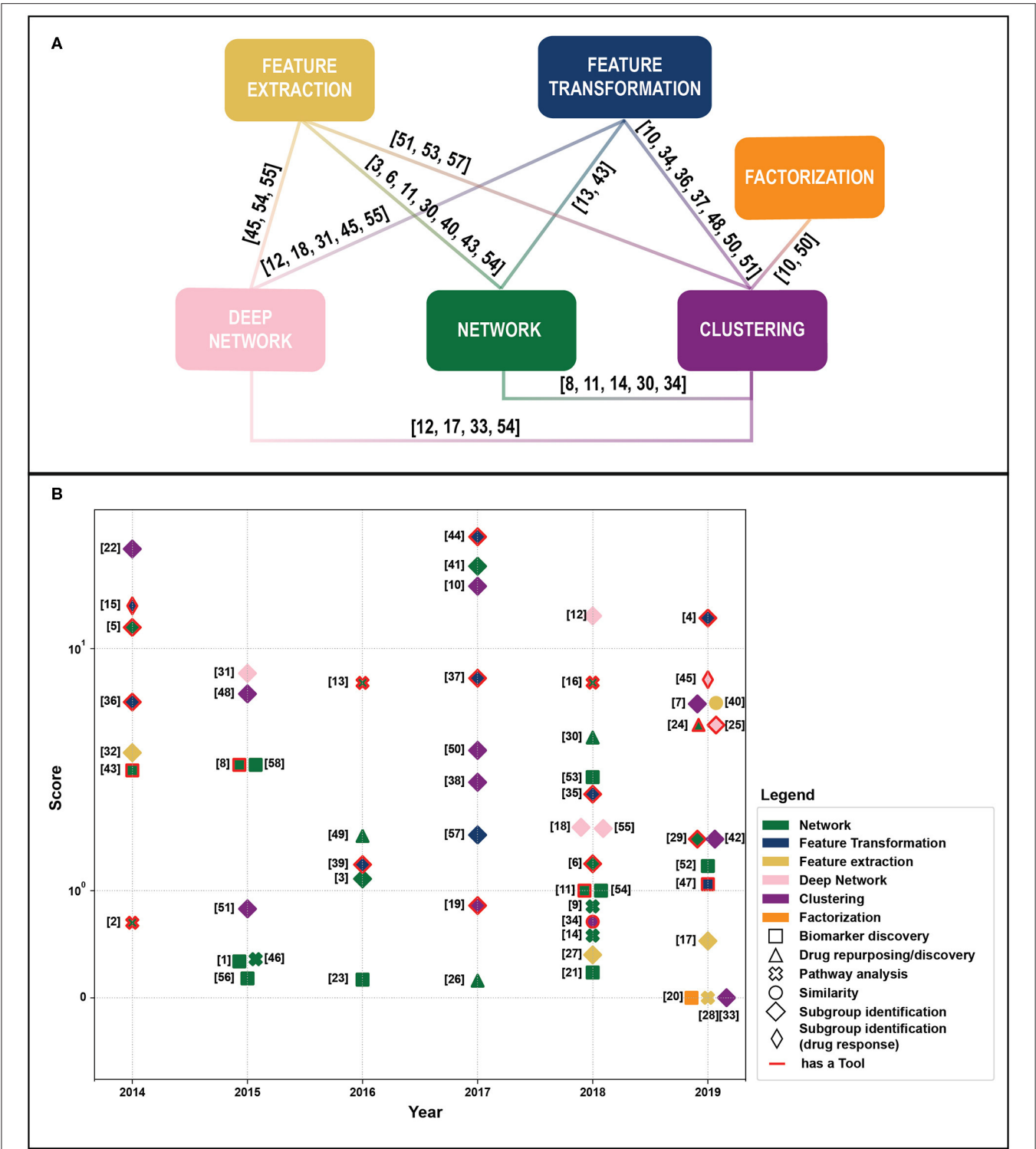


FIGURE 1 | (A) Linkage between different methodological categories. References to papers (see **Table 1**). That could be categorized in different groups are reported near the link. **(B)** Publications by year of publication and Field-Weighted Citation Impact. Different colors indicate exploited methods, shapes aims, and outputs. Papers with red borders have source code or provide a tool. Papers in the “Subgroup identification” group and/or with free tool result to be the most cited across years. The reference numbers are reported in **Table 1**.

others interesting approaches (16, 19) have been discussed in (10).

iOmicsPass

iOmicsPASS (40) implements a network-based method for integrating multi-omics profiles over genome-scale biological networks. The tool provides analysis components to transform qualitative multi-omics data into scores for biological interaction, then it uses the resulting scores as input to select predictive sub-networks; finally, it selects predictive edges for phenotypic groups using a modified nearest shrunken centroid algorithm. Authors validate iOmicsPASS on Breast Invasive Ductal Carcinoma data, integrating mRNA expression, and protein abundance, with and without the normalization of the mRNA data by the DNA Copy Number Variation (CNV). When compared with the original nearest shrunken centroid classification algorithm, iOmicsPASS outperform the baseline method, indicating the importance of selecting predictive signature forms densely connected sub networks, thus limiting the search space of predictive features to known interactions.

AMARETTO

Amaretto (22) is an algorithm developed multiple omics profiles integration across different type of cancers. Authors illustrate how the algorithm identifies cancer driver genes based on multi-omics data fusion and detects subnetworks of modules across all cancers. The algorithm identifies potential cancer driver genes by investigating significant correlations between methylation, CNV and gene expression (GE) data. When the driver genes are identified it constructs a module network connecting them with the co-expressed target genes they control. This constricts a pan-cancer network that is able to identify novel pancancer driver genes.

DrugComboExplorer

DrugComboExplorer (35) identifies candidate drug combinations targeting cancer driver signaling networks by processing DNA sequencing, CNV, DNA methylation, and RNA-seq data from individual cancer patients using an integrated pipeline of algorithms. The pipeline is based on two components: the first one extracts dysregulated networks from transcriptome and methylation profiles of specific patients using bootstrapping-based simulated annealing and weighted co-expression network analysis. The second component generates a driver network signatures for each drug, evaluates synergistic effects of drug combinations on different driver signaling networks and ranks drug combinations according the synergistic effects. In (35) authors apply DrugComboExplorer on diffuse large B-cell-lymphoma and prostate cancer, demonstrating the ability of the tool to discover synergistic drug combinations and its higher prediction accuracy compared with existing computational approaches.

Deep Network

Deep Networks (DNs) are widely used to analyse omics-data (68). In a multi-omics scenario, clustering on DNs features showed different survival groups in neuroblastoma and liver cancer (23,

29, 64). In (42) authors integrated GE, methylation and miRNA in a restricted Boltzmann machine, where hidden layers represent different survival groups in breast cancer patients. Subnetworks are used in (54) to project different omics views in latent spaces that are further concatenated and fed into a final network to predict drug response.

SALMON

SALMON (Survival Analysis Learning with Multi-Omics Neural Networks) is a Deep Learning framework that integrates omics-data (mRNA and miRNA), clinical features and cancer biomarkers (36). Instead of feeding a neural network with mRNA and miRNA data, SALMON takes as input the eigengene matrices derived from co-expression analysis. Thus, it overcomes the high-dimensionality problem, reducing input features of about 99%. Authors assume that mRNA and miRNA data affect survival outcome independently, therefore the two corresponding eigengene matrices are connected to two different hidden layers whose output is linked to the final network with a Cox proportional hazards regression network. Results on breast cancer carcinoma patients showed improvements in survival prediction ability compared to single-omics.

Clustering

Multi-omics clustering approaches are exploited to detect regularities and patterns that reveal different cancer molecular subtypes (21, 33, 43, 48, 57, 60) and prognostic groups in hepatocellular carcinoma (59). In (18) consensus clustering is performed on transcriptomics, metabolomics, and proteomics data to stratify patients with hepatocellular carcinoma based on their redox response. Clustering applications are often preceded by feature selection and/or feature transformation of multi-omics data, such as factorization, low rank approximation, and neural network. An exhaustive review on multi-omics integrative clustering approaches can be found in (69).

Nemo

NEMO (NEighborhood based Multi-Omics clustering) is a similarity-based tool that computes inter-patient similarity matrices for each omics through a radial basis function kernel. Spectral clustering is performed on the resulting average similarity matrix (52). NEMO addresses the problem of partial datasets, where not all the omics are measured for all the patients, and the final average matrix is computed on the observed omics values, without performing imputation. NEMO clustering shows higher performance compared to the same approach with imputed data, while on TCGA cancer datasets it detects significant differences in survival for six out of 10 cancer types.

Clusternomics

The main assumption of multi-omics clustering approaches relies on the existence of a consistent clustering structure across heterogeneous datasets. Alternatively, in (30) authors introduced the context-dependent clustering Clusternomics. Each omics is seen as a context describing a particular aspect of the underlying biological process. The global clustering structure is inferred from the combination of Bayesian clustering assignments. Then,

by separating cluster assignment on two levels, Clusternomics allows the number of clusters to vary on local or global structure. Its performances are evaluated on a simulated dataset, where it showed higher Adjuster Rank Index compared to other clustering techniques, but also on breast, lung and kidney cancer from TCGA repository, where it identified clinically meaningful clusters.

Affinity Network Fusion

Affinity Network Fusion (AFN) (44) is both a clustering and classification technique that applies graph clustering to a patient affinity matrix incorporating information from multiple views. For each omic, after feature selection and/or transformation, AFN computes patient pair-wise distances. kNN Graph Kernel applied to the distance metric creates a patient affinity matrix for each view. The final affinity matrix is the weighted sum of the computed affinity matrices. AFN approach showed improved clustering performance in detecting cancer subtypes on several TCGA datasets when compared to its application in single omics.

Feature Extraction

In multi-omics integration, variable selection to reduce the dimensionality of the omics dataset has a dominant role [(70), **Figure 1A**]. Recursive feature elimination was exploited to select subsets of expressed genes and methylation data to classify breast cancer disease subtypes with a Random Forest (3). Genes prioritization allowed prognosis prediction in different cancer types from epigenomics, transcriptomics, and genomics data (38), and biomarker discovery in prostate cancer (28). In (39) authors weight gene-gene interaction from transcriptomics and genomics data with a random walked-based method to select the most important interaction for survival prediction in breast cancer and neuroblastoma patients.

netDX

netDx is an algorithm that performs feature selection on Patient Similarity Networks (PSN) to classify patients in different prognostic groups (50). A PSN is built for each omics such that nodes represent patients and edges stand for the similarity of two nodes in the given view. Then netDx identifies which networks (i.e., which omics) strongly relate high- and low-risk patients through the GeneMANIA algorithm (71), which solves a regression problem to maximize the edges that connect query patients. Finally, each network is weighted according to its ability to relate patients of the same group and networks whose score exceeds a defined threshold are selected and combined in a single network by averaging their similarity scores. Authors benchmarked netDx against several machine-learning methods to predict survival outcomes on PanCancer TCGA multi-omics datasets, showing comparable results. On a breast cancer dataset, netDx selected features correspond to pathways known to be dysregulated in this type of cancer.

Feature Transformation

Feature transformation (FT) refers to algorithms that replace existing features with new features still function of the original ones. As shown in **Figure 1B**, the majority of FT techniques aims

at identifying cancer subtypes, biomarkers, omics-signatures, and key features from multi-omics data. Zhu et al. (66) proposed a kernel machine-learning method for a pan-cancer prognostic assessment by integrating multi-omics data. This work is particularly interesting since it's the only FT method we reviewed that allows multi-omics profile integration individually and in combination with clinical factors. A Kernel-based approach, combined with non-linear regression and Bayesian inference, resulted to be the best performing algorithm in a drug sensitivity prediction challenge (26).

In the following, we will report selected FT approaches, although few other tools for subgroup discovery, such as iClusterBayes (47), Multi-Omics Factor Analysis (15), JIVE (49), and MCIA (46), are available.

MixOmics

One of the most recent and biggest efforts in this field resulted in an R package called mixOmics (53). MixOmics allows for multivariate analysis of omics data including data exploration, dimension reduction, and visualization. mixOmics can be applied in numerous of studies with different aims such as integration and biomarker identification from multi-omics studies. The package includes two different types of multi-omics integration. One aimed at integrating different type of omics data of the same biological samples, while the second focus on integrating independent data measured on the same predictors to increase sample size and statistical power (53). Both frameworks aim at extracting biologically relevant features, [i.e., molecular signatures, by applying FT techniques (53)]. In (53) authors presented the results on 150 samples of mRNA, miRNA and proteomics breast cancer data and showed its ability to correctly discriminate three types of breast cancers.

mixKernel

mixKernel (45) is a R package compatible with mixOmics, which allows integration of multiple datasets by representing each dataset through a kernel that provides pairwise information between samples. The single kernels are then combined into one meta-kernel in an unsupervised framework. These new meta-kernels can be used for exploratory analyses, such as clustering or more sophisticated analysis to get insights into the data integrated. The authors showed better performances of mixKernel applied to mRNA, miRNAs and methylation breast cancer data if compared with one kernel approach.

iProFun

iProFun (56) is a method aimed at elucidating proteogenomic functional consequences of CNV and methylation alterations. The authors integrated mRNA expression levels, global protein abundances, and phosphoprotein abundances of a certain cancer. The output consists in a list of genes whose CNVs and/or DNA methylations significantly influencing some or all of the data integrated. iProFun obtains summary statistics of data integrated based on a gene-level multiple linear regression. These statistics are then used to extract genes having a cascading effect of all cis-molecular traits of interests and genes whose functional regulations are unique at global protein levels. iProFun

applied to ovarian cancer TCGA dataset showed its ability in extracting interesting genes that could be considered targets for future therapies.

Factorization

Traditional data mining methods are often inadequate to treat heterogeneous, sparse and noisy data such as multi-omics. Heavy pre-processing operations could modify, therefore loose, the inner structure of data coming from different sources. To discover latent characteristics hidden in huge amount of information, factorization techniques have been applied to highlight complex interactions among omics-data, hard to detect using standard approaches.

Gao et al. (31) developed an integrated Graph Regularized Non-negative Matrix Factorization model focused network construction by integrating gene expression data, CNV data, and methylation data. The authors used the factorization technique to decompose and fuse the multi-omics data. Then, by combining the results with network and mining analyses they showed how their method was able to find potential new cancer-related genes on two different TCGA datasets. Another method, based on factor analysis, aims at identifying latent factors in the multi-omics-data integrated in the model that can be used for subsequent analysis such as subgroup identification (15). Give its aim in extracting hidden features, we described this method in detail in the feature transformation section.

DISCUSSION

Along with technological advances in high-throughput sequencing, which characterize multiple “omes” from biological samples, holistic systems for data integration and knowledge discovery with machine-learning algorithms are still under development. Precision oncology would greatly benefit from actionable knowledge gained from multi-omics assays. In this paper we provided an overview of recent works on this topic and highlight current achievements and limitations.

We reviewed relevant tools to perform analysis based on different combination of omics, and observed their growing numbers in recent years, indicating strong commitments to develop such tools. Several issues emerged, too. The majority of the proposed techniques were applied to TCGA dataset, and data integration was mainly focused on transcriptomics and genomics. Efforts should be devoted to make new data sources available to the research community (72), such as the UKBioBank (73) and DriverDBv3 (74), and to integrate other “omes,” such as metabolome, or patient-generated, and environmental data. Research in this field would greatly benefit from the development of databases specifically developed for containing

and facilitating the analysis of multi-omics and clinical data, such as LinkedOmics (75). Another important improvement to increase usability and reproducibility would be to aim at developing methods that can be applied and generalized for all omics data type.

The complexity of multi-omics data analysis requires collaborative efforts among the clinical and machine-learning communities and the joint application of methodologies derived from heterogeneous backgrounds. We noted that some promising methods, such as matrix-factorization have not been extensively exploited, while clustering and network-based approaches are the most extensively used, probably due to their flexibility and the possibility to be integrated in comprehensive frameworks that include feature extraction and transformation to deal with the curse of dimensionality. Deep learning methods, that are flexible and achieved outstanding results in other fields, are increasingly used, even though many works share the same “pipeline” (i.e., the exploitation of autoencoder hidden layers for clustering). Interestingly, the number open source tools have increased in the very last years (Figure 1B).

We are aware of some limitations of our review. An important aspect that has not been covered by this review is the quantitative comparison among tools (76), which could highlight possible overfitting (77) and issues that may prevent the actual translation of multi-omics approaches from bench to bedside. Although, by indicating works that provide a usable tool (Table 1), our review could be a starting point for a comprehensive quantitative comparison.

AUTHOR CONTRIBUTIONS

RB conceived the study. GN, FV, and AD run the analyses and wrote the article. NG and RB revised the article. All authors contributed to the article and approved the submitted version.

FUNDING

This study was funded by Fondazione Regionale Ricerca Biomedica, Milan, Italy [FRRB project n. 2015-0042, Genomic profiling of rare hematologic malignancies, development of personalized medicine strategies, and their implementation into Rete Ematologica Lombarda (REL) clinical network] and by the NIHR Manchester BRC, MRC Molecular Pathology Node MPMPathic (grant ref MR/N00583X/1).

ACKNOWLEDGMENTS

We would like to acknowledge Simone Marini for his valuable help in the initial phases of the study.

REFERENCES

- Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*. (2009) 25:2906–12. doi: 10.1093/bioinformatics/btp543
- Lock EF, Hoadley KA, Marron JS, Nobel AB. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann Appl Stat*. (2013) 7:523–42. doi: 10.1214/12-AOAS597
- List M, Hauschild A-C, Tan Q, Kruse TA, Mollenhauer J, Baumbach J, et al. Classification of breast cancer subtypes by combining gene expression and DNA methylation data.

- J Integr Bioinform.* (2014) 11:236. doi: 10.2390/biecoll-jib-2014-236
4. Ray P, Zheng L, Lucas J, Carin L. Bayesian joint analysis of heterogeneous genomics data. *Bioinformatics.* (2014) 30:1370–6. doi: 10.1093/bioinformatics/btu064
 5. Gligorijević V, Malod-Dognin N, Pržulj N. Patient-specific data fusion for cancer stratification and personalised treatment. *Pacific Symp Biocomput.* (2016) 21:321–332. doi: 10.1142/9789814749411_0030
 6. Gottlieb A, Stein GY, Ruppén E, Sharan R. PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol Syst Biol.* (2011) 7:26. doi: 10.1038/msb.2011.26
 7. Napolitano F, Zhao Y, Moreira VM, Tagliaferri R, Kere J, D'Amato M, et al. Drug repositioning: a machine-learning approach through data integration. *J Cheminform.* (2013) 5:30. doi: 10.1186/1758-2946-5-30
 8. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* (2018) 68:394–424. doi: 10.3322/caac.21492
 9. Knox SS. From “omics” to complex disease: a systems biology approach to gene-environment interactions in cancer. *Cancer Cell Int.* (2010) 10:11. doi: 10.1186/1475-2867-10-11
 10. Huang S, Chaudhary K, Garmire LX. More is better: recent progress in multi-omics data integration methods. *Front Genet.* (2017) 8:84. doi: 10.3389/fgene.2017.00084
 11. Li Y, Wu FX, Ngom A. A review on machine learning principles for multi-view biological data integration. *Brief Bioinform.* (2018) 19:325–40. doi: 10.1093/bib/bbw113
 12. Agarwal M, Adhil M, Talukder AK. Multi-omics multi-scale big data analytics for cancer genomics. *Lect Notes Comput Sci.* (2015) 9498:228–43. doi: 10.1007/978-3-319-27057-9_16
 13. Amar D, Shamir R. Constructing module maps for integrated analysis of heterogeneous biological networks. *Nucleic Acids Res.* (2014) 42:4208–19. doi: 10.1093/nar/gku102
 14. Ao L, Song X, Li X, Tong M, Guo Y, Li J, et al. An individualized prognostic signature and multi-omics distinction for early stage hepatocellular carcinoma patients with surgical resection. *Oncotarget.* (2016) 7:24097–110. doi: 10.18632/oncotarget.8212
 15. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, et al. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol.* (2018) 14:e8124. doi: 10.15252/msb.20178124
 16. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods.* (2014) 11:333–7. doi: 10.1038/nmeth.2810
 17. Beal J, Montagud A, Traynard P, Barillot E, Calzone L. Personalization of logical models with multi-omics data allows clinical stratification of patients. *Front Physiol.* (2019) 9:1965. doi: 10.3389/fphys.2018.01965
 18. Benfeitas R, Bidkhorji G, Mukhopadhyay B, Kleivstig M, Arif M, Zhang C, et al. Characterization of heterogeneous redox responses in hepatocellular carcinoma patients using network analysis. *EBioMedicine.* (2019) 40:471–87. doi: 10.1016/j.ebiom.2018.12.057
 19. Bonnet E, Calzone L, Michoel T. Integrative multi-omics module network inference with lemon-tree. *PLoS Comput Biol.* (2015) 11:3983. doi: 10.1371/journal.pcbi.1003983
 20. Cancemi P, Buttacavoli M, Cara GD, Albanese NN, Bivona S, Pucci-Minafra I, et al. A multiomics analysis of S100 protein family in breast cancer. *Oncotarget.* (2018) 9:29064–81. doi: 10.18632/oncotarget.25561
 21. Cavalli FMG, Remke M, Rampasek L, Peacock J, Shih DJH, Luu B, et al. Intertumoral heterogeneity within medulloblastoma subgroups. *Cancer Cell.* (2017) 31:737–54.e6. doi: 10.1016/j.ccell.2017.05.005
 22. Champion M, Brennan K, Croonenborghs T, Gentles AJ, Pochet N, Gevaert O. Module analysis captures pancancer genetically and epigenetically deregulated cancer driver genes for smoking and antiviral response. *EBioMedicine.* (2018) 27:156–66. doi: 10.1016/j.ebiom.2017.11.028
 23. Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep learning-based multi-omics integration robustly predicts survival in liver cancer. *Clin Cancer Res.* (2018) 24:1248–59. doi: 10.1158/1078-0432.CCR-17-0853
 24. Cho H, Berger B, Peng J. Compact integration of multi-network topology for functional analysis of genes. *Cell Syst.* (2016) 3:540–8.e5. doi: 10.1016/j.cels.2016.10.017
 25. Costa RL, Boroni M, Soares MA. Distinct co-expression networks using multi-omic data reveal novel interventional targets in HPV-positive and negative head-and-neck squamous cell cancer. *Sci Rep.* (2018) 8:5. doi: 10.1038/s41598-018-33498-5
 26. Costello JC, Heiser LM, Georgii E, Gönen M, Menden MP, Wang NJ, et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nat Biotechnol.* (2014) 32:1202–12. doi: 10.1038/nbt.2877
 27. Dimitrakopoulos C, Hindupur SK, Hafliger L, Behr J, Montazeri H, Hall MN, et al. Network-based integration of multi-omics data for prioritizing cancer genes. *Bioinformatics.* (2018) 34:2441–8. doi: 10.1093/bioinformatics/bty148
 28. Drabovich AP, Saraon P, Drabovich M, Karakosta TD, Dimitromanolakis A, Hyndman ME, et al. Multi-omics biomarker pipeline reveals elevated levels of protein-glutamine gamma-glutamyltransferase 4 in seminal plasma of prostate cancer patients. *Mol Cell Proteomics.* (2019) 18:1807–23. doi: 10.1074/mcp.RA119.001612
 29. Francescato M, Chierici M, Rezvan Dezfouli S, Zandonà A, Jurman G, Furlanello C, et al. Multi-omics integration for neuroblastoma clinical endpoint prediction. *Biol Direct.* (2018) 13:8. doi: 10.1186/s13062-018-0207-8
 30. Gabasova E, Reid J, Wernisch L. Clusternomics: integrative context-dependent clustering for heterogeneous datasets. *PLoS Comput Biol.* (2017) 13:e1005781. doi: 10.1371/journal.pcbi.1005781
 31. Gao Y-L, Hou M-X, Liu J-X, Kong X-Z. An integrated graph regularized non-negative matrix factorization model for gene co-expression network analysis. *IEEE Access.* (2019) 7:126594–602. doi: 10.1109/ACCESS.2019.2939405
 32. Griffin PJ, Zhang Y, Johnson WE, Kolaczyk ED. Detection of multiple perturbations in multi-omics biological networks. *Biometrics.* (2018) 74:1351–61. doi: 10.1111/biom.12893
 33. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell.* (2014) 158:929–44. doi: 10.1016/j.cell.2014.06.049
 34. Hua L, Zheng WY, Xia H, Zhou P. Detecting the potential cancer association or metastasis by multi-omics data analysis. *Genet Mol Res.* (2016) 15:e038987. doi: 10.4238/gmr.15038987
 35. Huang L, Brunell D, Stephan C, Mancuso J, Yu X, He B, et al. Driver network as a biomarker: systematic integration and network modeling of multi-omics data to derive driver signaling pathways for drug combination prediction. *Bioinformatics.* (2019) 35:3709–17. doi: 10.1093/bioinformatics/btz109
 36. Huang Z, Zhan X, Xiang S, Johnson TS, Helm B, Yu CY, et al. Salmon: survival analysis learning with multi-omics neural networks on breast cancer. *Front Genet.* (2019) 10:166. doi: 10.3389/fgene.2019.00166
 37. Kim JY, Lee H, Woo J, Yue W, Kim K, Choi S, et al. Reconstruction of pathway modification induced by nicotinamide using multi-omic network analyses in triple negative breast cancer. *Sci Rep.* (2017) 7:7. doi: 10.1038/s41598-017-03322-7
 38. Kim M, Oh I, Ahn J. An improved method for prediction of cancer prognosis by network learning. *Genes.* (2018) 9:1–11. doi: 10.3390/genes9100478
 39. Kim SY, Jeong HH, Kim J, Moon JH, Sohn KA. Robust pathway-based multi-omics data integration using directed random walks for survival prediction in multiple cancer studies. *Biol Direct.* (2019) 14:8. doi: 10.1186/s13062-019-0239-8
 40. Koh HWL, Fermin D, Vogel C, Choi KP, Ewing RM, Choi H, et al. iOmicsPASS: network-based integration of multiomics data for predictive subnetwork discovery. *npj Syst Biol Appl.* (2019) 5:22. doi: 10.1038/s41540-019-0099-y
 41. Lee S-I, Celik S, Logsdon BA, Lundberg SM, Martins TJ, Oehler VG, et al. A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. *Nat Commun.* (2018) 9:5. doi: 10.1038/s41467-017-02465-5
 42. Liang M, Li Z, Chen T, Zeng J. Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach. *IEEE/ACM Trans Comput Biol Bioinforma.* (2015) 12:928–37. doi: 10.1109/TCBB.2014.2377729

43. Luo Z, Wang W, Li F, Songyang Z, Feng X, Xin C, et al. Pan-cancer analysis identifies telomerase-associated signatures and cancer subtypes. *Mol Cancer*. (2019) 18:106. doi: 10.1186/s12943-019-1035-x
44. Ma T, Zhang A. Affinity network fusion and semi-supervised learning for cancer patient clustering. *Methods*. (2018) 145:16–24. doi: 10.1016/j.jymeth.2018.05.020
45. Mariette J, Villa-Vialaneix N. Unsupervised multiple kernel learning for heterogeneous data integration. *Bioinformatics*. (2018) 34:1009–15. doi: 10.1093/bioinformatics/btx682
46. Meng C, Kuster B, Culhane AC, Gholami AM. A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics*. (2014) 15:162. doi: 10.1186/1471-2105-15-162
47. Mo Q, Shen R, Guo C, Vannucci M, Chan KS, Hilsenbeck SG. A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. *Biostatistics*. (2018) 19:71–86. doi: 10.1093/biostatistics/kxx017
48. Nguyen T, Tagett R, Diaz D, Draghici S. A novel approach for data integration and disease subtyping. *Genome Res*. (2017) 27:2025–39. doi: 10.1101/gr.215129.116
49. O'Connell MJ, Lock EF. R. JIVE for exploration of multi-source molecular data. *Bioinformatics*. (2016) 32:2877–9. doi: 10.1093/bioinformatics/btw324
50. Pai S, Hui S, Isserlin R, Shah MA, Kaka H, Bader GD, et al. netDx: interpretable patient classification using integrated patient similarity networks. *Mol Syst Biol*. (2019) 15:e8497. doi: 10.15252/msb.20188497
51. Raphael BJ, Hruban RH, Aguirre AJ, Moffitt RA, Yeh JJ, Stewart C, et al. Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer Cell*. (2017) 32:185–203.e13. doi: 10.1016/j.ccell.2017.07.007
52. Rappoport N, Shamir R, Schwartz R. NEMO: cancer subtyping by integration of partial multi-omic data. *Bioinformatics*. (2019) 35:3348–56. doi: 10.1093/bioinformatics/btz058
53. Rohart F, Gautier B, Singh A, Lê Cao KA. mixOmics: an R package for 'omics feature selection and multiple data integration. *PLoS Comput Biol*. (2017) 13:e1005752. doi: 10.1371/journal.pcbi.1005752
54. Sharifi-Noghabi H, Zolotareva O, Collins CC, Ester M. MOLI: multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics*. (2019) 35:i501–9. doi: 10.1093/bioinformatics/btz318
55. Sehgal V, Seviour EG, Moss TJ, Mills GB, Azencott R, Ram PT. Robust selection algorithm (RSA) for multi-omic biomarker discovery; integration with functional network analysis to identify miRNA regulated pathways in multiple cancers. *PLoS ONE*. (2015) 10:72. doi: 10.1371/journal.pone.0140072
56. Song X, Ji J, Gleason KJ, Yang F, Martignetti JA, Chen LS, et al. Insights into impact of DNA copy number alteration and methylation on the proteogenomic landscape of human ovarian cancer via a multi-omics integrative analysis. *Mol Cell Proteomics*. (2019) 18(8 Suppl.1):S52–65. doi: 10.1074/mcp.RA118.001220
57. Speicher NK, Pfeifer N. Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. *Bioinformatics*. (2015) 31:i268–75. doi: 10.1093/bioinformatics/btv244
58. Vitali F, Cohen LD, Demartini A, Amato A, Eterno V, Zambelli A, et al. A network-based data integration approach to support drug repurposing and multi-Target therapies in triple negative breast cancer. *PLoS ONE*. (2016) 11:e0162407. doi: 10.1371/journal.pone.0162407
59. Woo HG, Choi J-H, Yoon S, Jee BA, Cho EJ, Lee J-H, et al. Integrative analysis of genomic and epigenomic regulation of the transcriptome in liver cancer. *Nat Commun*. (2017) 8:839. doi: 10.1038/s41467-017-00991-w
60. Wu D, Wang D, Zhang MQ, Gu J. Fast dimension reduction and integrative clustering of multi-omics data using lowrank approximation: application to cancer molecular classification. *BMC Genomics*. (2015) 16:1022. doi: 10.1186/s12864-015-2223-8
61. Yang Z, Liu B, Lin T, Zhang Y, Zhang L, Wang M, et al. Multiomics analysis on DNA methylation and the expression of both messenger RNA and microRNA in lung adenocarcinoma. *J Cell Physiol*. (2019) 234:7579–86. doi: 10.1002/jcp.27520
62. Yuan L, Guo LH, Yuan CA, Zhang Y, Han K, Nandi AK, et al. Integration of multi-omics data for gene regulatory network inference and application to breast cancer. *IEEE/ACM Trans Comput Biol Bioinforma*. (2019) 16:782–91. doi: 10.1109/TCBB.2018.2866836
63. Wang Z, Wei Y, Zhang R, Su L, Gogarten SM, Liu G, et al. Multi-omics analysis reveals a HIF network and hub gene EPAS1 associated with lung adenocarcinoma. *EBioMedicine*. (2018) 32:93–101. doi: 10.1016/j.ebiom.2018.05.024
64. Zhang L, Lv C, Jin Y, Cheng G, Fu Y, Yuan D, et al. Deep learning-based multi-omics data integration reveals two prognostic subtypes in high-risk neuroblastoma. *Front Genet*. (2018) 9:477. doi: 10.3389/fgene.2018.00477
65. Zhou Y, Liu Y, Li K, Zhang R, Qiu F, Zhao N, et al. ICan: an integrated co-alteration network to identify ovarian cancer-related genes. *PLoS ONE*. (2015) 10:e0116095. doi: 10.1371/journal.pone.0116095
66. Zhu B, Song N, Shen R, Arora A, Machiela MJ, Song L, et al. Integrating clinical and multiple omics data for prognostic assessment across human cancers. *Sci Rep*. (2017) 7:8. doi: 10.1038/s41598-017-17031-8
67. Žitnik M, Zupan B. Gene network inference by fusing data from diverse distributions. *Bioinformatics*. (2015) 31:i230–9. doi: 10.1093/bioinformatics/btv258
68. Tang B, Pan Z, Yin K, Khateeb A. Recent advances of deep learning in bioinformatics and computational biology. *Front Genet*. (2019) 10:214. doi: 10.3389/fgene.2019.00214
69. Wang D, Gu J. Integrative clustering methods of multi-omics data for molecule-based cancer classifications. *Quant Biol*. (2016) 4:58–67. doi: 10.1007/s40484-016-0063-4
70. Wu C, Zhou F, Ren J, Li X, Jiang Y, Ma S, et al. A selective review of multi-level omics data integration using variable selection. *High-Throughput*. (2019) 8:4. doi: 10.3390/ht8010004
71. Mostafavi S, Ray D, Warde-Farley D, Grouios C, Morris Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol*. (2008) 9:S4. doi: 10.1186/gb-2008-9-s1-s4
72. Conesa A, Beck S. Making multi-omics data accessible to researchers. *Sci Data*. (2019) 6:251. doi: 10.1038/s41597-019-0258-4
73. Ollier W, Sprosen T, Peakman T. UK Biobank: from concept to reality. *Pharmacogenomics*. (2005) 6:639–46. doi: 10.2217/14622416.6.6.639
74. Liu SH, Shen PC, Chen CY, Hsu AN, Cho YC, Lai YL, et al. DriverDBv3: a multi-omics database for cancer driver gene research. *Nucleic Acids Res*. (2020) 48:D863–70. doi: 10.1093/nar/gkz964
75. Vasaiyar SV, Straub P, Wang J, Zhang B. LinkedOmics: analyzing multi-omics data within and across 32 cancer types. *Nucleic Acids Res*. (2018) 46:D956–63. doi: 10.1093/nar/gkx1090
76. Sathyanarayanan A, Gupta R, Thompson EW, Nyholt DR, Bauer DC, Nagaraj SH, et al. A comparative study of multi-omics integration tools for cancer driver gene identification and tumour subtyping. *Brief Bioinform*. (2019). doi: 10.1093/bib/bbz121. [Epub ahead of print].
77. McCabe SD, Lin DY, Love MI. Consistency and overfitting of multi-omics methods on experimental data. *Brief Bioinform*. (2019). doi: 10.1093/bib/bbz070. [Epub ahead of print].

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Nicora, Vitali, Dagliati, Geifman and Bellazzi. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Integrated Transcriptome Analysis of Human Visceral Adipocytes Unravels Dysregulated microRNA-Long Non-coding RNA-mRNA Networks in Obesity and Colorectal Cancer

Sabrina Tait¹, Antonella Baldassarre², Andrea Masotti², Enrica Calura³, Paolo Martini³, Rosaria Vari¹, Beatrice Scazzocchio¹, Sandra Gessani¹ and Manuela Del Cornò^{1*}

¹ Center for Gender-Specific Medicine, Istituto Superiore di Sanità, Rome, Italy, ² Bambino Gesù Children's Hospital-IRCCS, Research Laboratories, Rome, Italy, ³ Department of Biology, University of Padua, Padua, Italy

OPEN ACCESS

Edited by:

Margaret Jane Currie,
University of Otago, Christchurch,
New Zealand

Reviewed by:

Weifeng Ding,
Nantong University, China
Olga Brovkina,
Federal Medical-Biological
Agency, Russia

*Correspondence:

Manuela Del Cornò
manuela.delcornò@iss.it

Specialty section:

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

Received: 06 March 2020

Accepted: 01 June 2020

Published: 02 July 2020

Citation:

Tait S, Baldassarre A, Masotti A, Calura E, Martini P, Vari R, Scazzocchio B, Gessani S and Del Cornò M (2020) Integrated Transcriptome Analysis of Human Visceral Adipocytes Unravels Dysregulated microRNA-Long Non-coding RNA-mRNA Networks in Obesity and Colorectal Cancer. *Front. Oncol.* 10:1089. doi: 10.3389/fonc.2020.01089

Obesity, and the obesity-associated inflammation, represents a major risk factor for the development of chronic diseases, including colorectal cancer (CRC). Dysfunctional visceral adipose tissue (AT) is now recognized as key player in obesity-associated morbidities, although the biological processes underpinning the increased CRC risk in obese subjects are still a matter of debate. Recent findings have pointed to specific alterations in the expression pattern of non-coding RNAs (ncRNAs), such as microRNAs (miRNAs), and long non-coding RNAs (lncRNAs), as mechanisms underlying dysfunctional adipocyte phenotype in obesity. Nevertheless, the regulatory networks and interrelated processes relevant for adipocyte functions, that may contribute to a tumor-promoting microenvironment, are poorly known yet. To this end, based on RNA sequencing data, we identified lncRNAs and miRNAs, which are aberrantly expressed in visceral adipocytes from obese and CRC subjects, as compared to healthy lean control, and validated a panel of modulated ncRNAs by real-time qPCR. Furthermore, by combining the differentially expressed lncRNA and miRNA profiles with the transcriptome analysis dataset of adipocytes from lean and obese subjects affected or not by CRC, lncRNA-miRNA-mRNA adipocyte networks were defined for obese and CRC subjects. This analysis highlighted several ncRNAs modulation that are common to both obesity and CRC or unique of each disorder. Functional enrichment analysis of network-related mRNA targets, revealed dysregulated pathways associated with metabolic processes, lipid and energy metabolism, inflammation, and cancer. Moreover, adipocytes from obese subjects affected by CRC exhibited a higher complexity, in terms of number of genes, lncRNAs, miRNAs, and biological processes found to be dysregulated, providing evidence that the transcriptional and post-transcriptional program of adipocytes from CRC patients is deeply affected by obesity. Overall, this study adds further evidence for a central role of visceral adipocyte dysfunctions in the obesity-cancer relationship.

Keywords: obesity, colorectal cancer, adipocyte, RNASeq, microRNAs, long non-coding RNAs, networks

INTRODUCTION

The increase of obesity is a major health problem afflicting nowadays adults and children worldwide (1). Obesity is a complex condition, characterized by excessive expansion and functional alteration of white adipose tissue (AT), that increases the risk of life threatening diseases such as cardiovascular disease, diabetes and cancer, including colorectal cancer (CRC). Indeed, white AT, particularly visceral fat, is a complex endocrine and immunocompetent organ, homing adipocytes and resident immune cells, exhibiting secretory as well as immunological, metabolic, and endocrine regulatory activities and playing a central role in obesity-associated morbidities (2). Its functional units, the adipocytes, produce and secrete a large array of mediators including cytokines/chemokines, extracellular matrix proteins, hormones, growth and angiogenic factors that influence, either locally or systemically, a variety of physiological and pathological processes, such as immune functions, cell proliferation, migration, angiogenesis (3, 4). In addition of being an established risk factor (5), excess adiposity is also associated with CRC worse outcomes (6, 7), although the mechanisms underlying the detrimental link between obesity and CRC are complex and not yet precisely defined. In this respect, it has been postulated that this association may be due to the large spectrum of cytokines and metabolites that are produced by AT showing pro-inflammatory and cancer prone features. Moreover, obesity-related metabolic alterations (i.e., triggering of insulin resistance, impairment in lipid metabolism, endocrinologic changes and oxidative stress) may contribute to CRC initiation and progression (8). More recently, emerging evidence point to the role of non-coding RNAs (ncRNAs) in many obesity-related disorders including cardiovascular and metabolic diseases, inflammation, and cancer (9), and more specifically in CRC (10).

NcRNAs are transcripts that are not translated into proteins. They are present in all organisms, where they regulate gene expression and, therefore, biological processes, at the transcriptional and post-transcriptional level (11). Multiple types of regulatory ncRNAs are emerging as key elements of cellular homeostasis and diseases. Among these long ncRNAs (lncRNAs) (>200 nts) and small ncRNAs (<200 nts), such as microRNAs (miRNAs), small interfering-, Piwi interacting-, small nucleolar-, small nuclear-, extracellular-RNAs, are arbitrarily classified according to their nucleotide length (12). Among them, microRNAs (miRNAs) are evolutionarily conserved small ncRNAs (18–25 nt in length) playing a crucial role in cell transcriptional regulation (13, 14). Their expression correlates with different obesity relevant parameters, such as body mass index (BMI), adipocyte size and metabolic parameters, highlighting important regulatory role in obesity (15–18). The importance of miRNAs in mediating the initiation, growth, and development of CRC was also reported (19). In contrast with small ncRNAs, lncRNAs undergo post-transcriptional modifications, such as polyadenylation and splicing, although they lack protein-coding capacity (20). They are emerging as miRNA sponges and inhibitors, thus releasing downstream genes from the miRNA control (21). Furthermore, lncRNAs can also

interact with DNA, RNA and proteins, overall regulating gene expression and epigenetic status (12). Accumulating evidence has revealed that the expression of lncRNAs is involved in the occurrence and development of many major diseases, including human cancers (22, 23), and that lncRNA-miRNA-mRNA networks are specifically associated with CRC (24). High-throughput methods and bioinformatics approaches have significantly contributed to the identification of new transcripts, including ncRNAs. However, only few studies have described miRNAs and lncRNAs in human AT under obesity (9, 25–27). Moreover, no studies have reported the expression of miRNAs and lncRNAs in AT from CRC patients. In this regard, we recently reported that obesity and CRC, conditions characterized by the common denominator of inflammation, are associated with changes in the transcriptional program of adipocytes mostly involving pathways and biological processes linked to fibrosis, inflammation and metabolism of pyruvate, lipids, and glucose (28). In this study, we analyzed the ncRNA expression profiles, specifically miRNAs and lncRNAs, of lean and obese subjects affected or not by CRC, by RNASeq/Small RNASeq analysis. This approach allowed to highlight changes in adipocyte miRNA and lncRNA profiles that are specifically associated with obesity or CRC, or shared by both conditions. Finally, by integrating bioinformatics prediction, functional enrichment analysis, and data on differential mRNA expression previously described (28), we identified lncRNA-miRNA-mRNA regulatory networks and defined multiple pathways characterizing visceral adipocytes, that are altered in obesity and/or CRC. Overall, this might contribute to set the basis for a more tumor-prone microenvironment, thus adding further evidence for the central role of AT functional alterations in linking obesity to cancer.

METHODS

Ethics Statement

Investigation has been conducted in accordance with the ethical standards and with the Declaration of Helsinki, and according to national and international guidelines. It was approved by the institutional review board of Istituto Superiore di Sanità. All enrolled subjects were provided with complete information about the study and asked to sign an informed consent.

Patient and Sample Collection

As previously described (28), “human visceral adipose tissue (VAT) was collected from age-matched lean and obese subjects undergoing abdominal surgery or laparoscopy for benign (i.e., gallbladder disease without icterus, umbilical hernia, and uterine fibromatosis) or CRC conditions (histologically proved primary colon adenocarcinoma, stage TNM 0–III). The exclusion's criteria were: clinical evidence of active infection, recent (within 14 days) use of antibiotics/anti-inflammatory drugs, pregnancy, hormonal therapies, severe mental illness, autoimmune diseases, family history of cancer, others neoplastic diseases. In the normal weight group, the BMI range was 20–25 Kg/m². In the obese group the BMI was ≥ 30 Kg/m², and waist circumference >

95 cm for men and > 80 cm for women. The total number of subjects was six/category.”

Adipocyte Isolation, RNA Preparation and Sequencing

Adipocytes were isolated from human VAT as previously described (29). Total RNA was isolated with Total RNA Purification Plus Kit (Norgen Biotek, Canada). RNA quality and quantity was assessed by Agilent 2,100 Bioanalyzer and samples stored at -80°C until use. Total RNA (2 μg) was used to prepare the library for Illumina sequencing (Illumina TruSeq Small RNA Sample Preparation). Single-end reads (> 10 M reads per sample) were produced by Illumina HiSeq 2000.

RNASeq Data Preprocessing and Differential Expression Analysis

Libraries were then processed with Illumina cBot for cluster generation on the flowcell, following the manufacturer's instructions and sequenced on single-end mode at the multiplexing level requested on HiSeq2000 (Illumina, San Diego, CA). The CASAVA 1.8.2 version of the Illumina pipeline was used to process raw data for both format conversion and de-multiplexing. Adapters were removed and low-quality bases were trimmed by the script TrimGalore. Per sample, per read and per base quality of raw sequence data have been assessed with FastQC version 0.11.3 (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) and all the included samples passed the initial quality checks. All the sequencing data had all the range of the per base quality values into very good quality calls, lower than the 0.02% of the total sequences showed a per sequence low quality score and no adapter content. Thus, no quality trimming where performed during preprocessing. The percentage of mapped reads resulted high with the mean value of 97.5% (min 94.08% and max 98.41%).

The transcriptome reconstruction was performed as previously described (28). Re-annotation of previously unknown transcripts was performed using the bioMart package (30) into R 3.6 (31), querying available Ensemble transcript IDs and retrieving Gene Names, Entrez gene IDs, gene and transcript biotypes thus allowing the identification of a higher number of lncRNAs. Multiple testing controlling procedure was applied following Benjamini & Hochberg method hereafter referred as False Discovery Rate (FDR). We then extracted the list of differentially expressed lncRNAs (DEL) with a False Discovery Rate (FDR) ≤ 0.05 . For small RNASeq analysis, raw reads where pre-processed using cutadapt 1.9.1 (<http://code.google.com/p/cutadapt/>) and reads shorter than 17 bases were excluded. MiRNA expression quantification was carried out using MirDeep2 (version 2.0.0.8, Bowtie version 1.1.2) (32) using hg38.p2 genome version and 79 Ensembl version. MiRNA mature/hairpin sequences were downloaded from Mirbase 21 version (33), then, raw counts were filtered to keep only miRNA with more or equal to 10 reads in at least one sample. MiRNA expression was normalized with upper quantile normalization (EDASeq version 2.10.0) (34) while differential expression (all comparisons) was computed using edgeR (3.18.1) (35) from raw

counts. All analyses were carried out in R and Bioconductor 3.5 version (<https://bioconductor.org>). Due to the limited differential expression of miRNAs (DEM), the threshold of FDR was set ≤ 0.06 . For small RNA sequencing, six biological replicates per category were prepared and the raw sequence data are available from the NCBI Sequence Read Archive (SRA) (<http://www.ncbi.nlm.nih.gov/sra>) under accession number SRA: PRJNA632999. For long RNA sequencing, we employed the RNASeq datasets previously published and available under accession number SRA: PRJNA508473 (28).

mRNA-miRNA-lncRNA Regulation Network Construction

Target genes of the identified differentially expressed miRNAs (DEM) were searched in the TarBase v.8 (36) and miRTarBase 7.0 (37) databases which feature up-to-date experimentally validated miRNA-targets interactions. Interactions between DEL and DEM were retrieved in both the DIANA-lncBase v2.0 database (38), using the prediction module and a score ≥ 0.6 as cut-off, and the ENCORI database (39) featuring experimentally verified RNA-RNA interactions. The ENCORI database was also used to search for DEL-mRNA verified interactions. The overall targets of DEM and DEL were filtered against the lists of differentially expressed transcripts (DET) and integrated to define specific mRNA-miRNA-lncRNA interactions networks for each condition. The Cytoscape software (40) was used to visualize the obtained networks.

Functional Analysis

The cumulative list of DEM and DEL targets within the DET of each condition was explored for significantly enriched pathways with the Cytoscape plug-in ClueGO and CluePEDIA (41) querying the KEGG, WikiPathways and Reactome databases. Default settings were used for the pathways selection, connectivity and grouping. A two-sided enrichment analysis was performed, adjusting the *p*-values with the Benjamini-Hochberg correction and considering significant only pathways with *p* < 0.05.

Real-Time qPCR Validation of Differentially Expressed lncRNAs and miRNAs

Twelve candidate ncRNAs, found differentially expressed by RNASeq, were selected for validation by real time qPCR (RT-qPCR). The validation of lncRNA expression was performed by qPCR using SYBRGreen assays (Supplemental Table 1). The synthesis of cDNA was performed by using 300–500 ng of total RNA in 20 μL reaction volume using the Superscript III kit (ThermoFisher Scientific) following the manufacturer's instructions. The reverse transcription conditions were as follows: 5 min at 25°C , 60 min at 50°C , and 15 min at 70°C . cDNA was mixed with $2 \times$ SensiFast SYBR low rox (Bioline), lncRNA expression values were normalized to the expression of GUSB as the endogenous control. For the validation of miRNA expression levels, we started the reverse transcription of 6 miRNAs by using 2 μL (5 ng/ μL) of total RNA with the miRCURY LNA RT Kit (Qiagen). The reverse transcription conditions were as follows: 60 min at 42°C and 5 min at 95°C . cDNA was

mixed with 2 × miRCURY SYBR Green Master Mix (Qiagen) following the manufacturer's instructions. The expression values of miRNAs were normalized to the expression of let-7a-5p as the endogenous control. For each sample, the relative expression level was determined according to the $2^{-\Delta\Delta CT}$ method after running the samples on a QuantStudio 12 K Flex Real-Time PCR System (ThermoFisher Scientific) following the manufacturer's instructions. For each sample, the relative gene expression level was determined according to the $2^{-\Delta\Delta CT}$ method. Statistical comparisons of means from six biological replicates, matched with RNASeq analysis, was performed between the various subject groups (five for the NwCRC group) by one-way analysis of variance (ANOVA) with LSD *post hoc* tests by using SPSS software (Ver.20). Differences were considered statistically significant when *p*-values were ≤ 0.05 . Analysis of correlation between qPCR and RNASeq data was performed by Spearman's rank test setting significance at $p < 0.05$.

RESULTS

Long and Small RNA Sequencing Analysis Identify Differentially Expressed lncRNAs and miRNAs That Are Associated With Obesity and/or CRC

We have previously analyzed the transcriptome profiles of human adipocytes isolated from visceral AT (VAT) biopsies obtained from healthy control lean (normal weight, Nw) and obese (Ob) subjects, or CRC patients (normal weight or obese, NwCRC, and ObCRC, respectively), by RNA sequencing (28). Along with the protein coding transcripts, the long RNASeq analysis detected also a total of 90 differentially expressed lncRNAs (DEL, FDR ≤ 0.05), 35 of which were novel transcripts (Table 1). In NwCRC subjects, 45 DEL were found dysregulated (11 downregulated, 33 upregulated and one DEL with two transcripts inversely modulated, NUTM2A-AS1) compared to Nw healthy controls. In Ob group, we found 27 DEL (3 downregulated, 23 upregulated and one DEL with two inversely modulated transcripts, RASSF8-AS1). Finally, when comparing ObCRC group with the control lean group, a total of 52 DEL, including 13 downregulated, 38 upregulated and one with three transcripts (MIR4435-2HG, one up- and two downregulated), were found. Among the overall 90 DEL, 10 were shared by all the three subject categories (AC109460.3, AL031429.1, AL139260.1, APTR, FAM198B-AS1, LINC00968, LINC01106, LINC01348, MIR4435-2HG, SNHG16), 6 were shared by NwCRC and Ob patients (AC008105.3, AC021092.1, HIF1A-AS1, LINC00926, RASSF8-AS1, ZNF883), 12 were shared by NwCRC and ObCRC patients (AC009022.1, AC010457.1, AC016582.2, AC068888.1, AL356056.1, AP000317.2, FAM27E3, MINCR, MIR100HG, SLC14A2-AS1, STAG3L5P-PVRIG2P-PILRB, TPTEP1), and only one was shared by Ob and ObCRC patients (AC022007.1). On the other hand, a number of lncRNAs were selectively modulated in each subject category, with the ObCRC group exhibiting the highest number of specific DEL (Table 1). In parallel, small RNASeq analysis revealed a total of 58 differentially expressed miRNAs (DEM, FDR ≤ 0.06) in adipocytes of

TABLE 1 | Differentially expressed lncRNAs in normal weight affected by CRC (NwCRC), obese (Ob), and obese affected by CRC (ObCRC) individuals vs. healthy lean control.

Gene Name	Entrez Gene ID	Log ₂ FC (FDR ≤ 5%)		
		NwCRC	Ob	ObCRC
Novel lncRNA				
AC004241.1		4.4723		
AC004477.3		1.9133		
AC006504.5		2.3449		
AC007098.1				6.6690
AC008105.3		4.2391	4.4356	
AC009022.1		8.7676		8.7671
AC010457.1		2.2489		2.8844
AC016582.2		5.1047		
AC021092.1		1.1310	1.0531	
AC022007.1			9.0501	9.3198
AC023421.1				5.4800
AC061992.1		−4.2812		
AC068473.5		1.4616		
AC068888.1		7.4252		8.3630
AC084757.3				−2.9022
AC092279.1		−5.9393		
AC099518.3				−7.8032
AC109460.3		6.9064	6.8707	6.2148
AC114956.3				2.6267
AC139256.1			2.8280	
AC144548.1				−1.0320
AC141930.1		3.2551		
AL031429.1		3.3503	3.0813	4.1653
AL078612.1				2.7046
AL138828.1				2.3053
AL138963.4				2.2073
AL139260.1		7.4131	6.2839	7.3094
AC016582.1				5.2659
AL161772.1				1.9376
AL355607.2				1.9507
AL356056.1		5.4758		6.9848
AL591848.3		1.5321		
AP000317.2				−11.2572
AP000790.1				6.1063
FP236383.3				8.4412
Known lncRNA				
AGAP11	119385			1.8547
APTR	100505854	3.3044	4.0329	2.8546
ARHGEF7-AS2	100874238			4.6621
BCYRN1	618			3.1417
CFLAR-AS1	65072	−6.3218		
DLGAP1-AS1	649446		1.4003	
FAM198B-AS1	285505	3.8088	4.4575	4.6582
FAM27E3	100131997			3.9401
FOXP4-AS1	101060264			6.3680
H19	283120	−22.6445		

(Continued)

TABLE 1 | Continued

Gene Name	Entrez Gene ID	Log ₂ FC (FDR ≤ 5%)		
		NwCRC	Ob	ObCRC
HIF1A-AS1	100750246	3.7185	2.9894	
HOXB-AS3	404266		-7.1867	
LINC00486	285045			6.9600
LINC00926	283663	4.5816	4.6596	
LINC00968	100507632	8.9218	8.3768	8.4248
LINC01106	151009	19.5614	19.7102	22.0251
LINC01106	151009	20.9418	20.6120	19.8254
LINC01140	339524	-3.7389		
LINC01140	339524	-3.5970		
LINC01184	644873			-8.4369
LINC01239	441389			-2.1368
LINC01291	102724515			-21.0782
LINC01348	731656	8.9118	8.2512	7.3795
LINC01619	256021			-8.8258
LNCOG	105369848		1.8031	
LUCAT1	100505994		9.5244	
MALINC1	100505636		2.0735	
MAP4K3-DT	728730			5.1283
MINCR	100507316	3.7292		4.1920
MIR100HG	399959	-8.2382		-21.7228
MIR100HG	399959			-9.8115
MIR3142HG	107075116		3.1109	
MIR4435-2HG	541471	9.2841	9.7786	9.1337
MIR4435-2HG	541471			-9.4517
MIR4435-2HG	541471			-8.6134
MSC-AS1	100132891	-23.2039		
NUTM2A-AS1	728190	-6.5505		
NUTM2A-AS1	728190	9.2961		
OLMALINC	90271			-2.9871
PGM5P3-AS1	101929127	-7.6345		
RASSF8-AS1	100506451	19.9364	22.0934	
RASSF8-AS1	100506451		-9.9462	
SCAT8	112935969	-3.9716		
SLC14A2-AS1	101927980	6.4340		9.1498
SNHG16	100507246	4.6908	4.6101	5.3027
SNHG29	125144			-7.6114
SNORD3C	780853		6.8661	
STAG3L5P-PVRIG2P-PILRB	101752399			20.8685
TMEM161B-AS1	100505894			4.6379
TPRG1-AS1	100874043		-1.6872	
TPTEP1	387590			7.4005
TPT1-AS1	100190939	-6.8438		
UBA6-AS1	550112			-8.0421
USP9Y	8287	8.9477		
XIST	7503			-25.1297
ZFAS1	441951		-5.2589	
ZNF295-AS1	150142			1.8551
ZNF883	169834	6.9421	7.4847	

TABLE 2 | Differentially expressed miRNAs in normal weight affected by CRC (NwCRC), obese (Ob), and obese affected by CRC (ObCRC) individuals vs. healthy lean control.

miRNA	Log ₂ FC (FDR ≤ 6%)		
	NwCRC	Ob	ObCRC
let-7c-5p		0.8158	
let-7e-3p		-0.5273	-0.5072
let-7f-5p	0.6367		
let-7i-3p			-0.7816
miR-100-5p			-0.8363
miR-107			0.4816
miR-10b-3p			0.9854
miR-10b-5p			1.0690
miR-1246	1.4342		
miR-1247-5p	-1.1469	-1.2365	-1.0182
miR-125a-5p	-0.6525	-0.6889	-0.7270
miR-125b-1-3p	-1.1204		-1.1041
miR-1287-5p		1.1111	1.2275
miR-1296-5p		-0.9913	
miR-1299			1.2495
miR-1323			-1.4956
miR-144-5p		1.3677	
miR-152-3p		0.6263	0.6534
miR-181c-3p		1.1703	
miR-181c-5p		1.2190	1.2047
miR-181d-5p		1.2244	1.2720
miR-185-5p		0.784	0.8778
miR-193b-3p	-0.7937	-0.7410	-0.7744
miR-22-5p	0.6981		0.7611
miR-24-3p		0.6745	0.7660
miR-28-5p			0.4959
miR-29b-2-5p	0.6554		0.7330
miR-29b-3p	0.9662		
miR-30c-5p			-0.6740
miR-3182	1.2731		
miR-328-3p	-0.8400		
miR-33b-3p	-1.2430		
miR-345-5p	-0.6699	-0.8146	
miR-34a-5p		0.9133	1.1756
miR-361-3p		-0.4835	
miR-3622a-5p	-1.249		
miR-374a-3p	0.7650		
miR-374b-5p	0.9486		
miR-378f			-1.0918
miR-421		0.8311	0.8296
miR-4455	-1.3104		-1.2491
miR-451a		1.2621	
miR-452-5p	0.6354		0.7434
miR-483-5p		1.0648	
miR-508-3p			1.1567
miR-512-3p			-1.3976
miR-515-5p			-1.1957
miR-516a-5p			-1.4569

(Continued)

TABLE 2 | Continued

miRNA	Log ₂ FC (FDR ≤ 6%)		
	NwCRC	Ob	ObCRC
miR-516b-5p			−1.3241
miR-517a-3p			−1.5517
miR-517b-3p			−1.5517
miR-548az-5p			1.3631
miR-598-3p	0.9845		
miR-664a-3p	0.9304		
miR-7706	−0.8991		−0.6495
miR-92a-3p			−0.5834
miR-98-5p	0.7211		0.6373
miR-99a-3p			0.5715

NwCRC, Ob, and ObCRC subjects compared to Nw individuals (Table 2). Specifically, 22 DEM were found in NwCRC (12 upregulated and 10 downregulated), 20 DEM were detected in Ob subjects (13 upregulated and 7 downregulated), while the comparison of ObCRC with Nw control revealed a higher number of dysregulated miRNAs (39 DEM, 20 upregulated and 19 downregulated), suggesting that the conditions of obesity and CRC interact concurrently, thus influencing the miRNA expression profile in adipocyte from ObCRC subjects. Among the overall modulated 58 DEM, only 3 were common to all group of subjects (miR-1247-5p, miR-125a-5p, miR-193b-3p), 7 were shared by NwCRC and ObCRC subjects (miR-125b-1-3p, miR-22-5p, miR-29b-2-5p, miR-4455, miR-452-5p, miR-7706, miR-98-5p), 9 were shared by ObCRC and Ob subjects (let-7e-3p, miR-1287-5p, miR-152-3p, miR-181c-5p, miR-181d-5p, miR-185-5p, miR-24-3p, miR-34a-5p, miR-421), while only one was common to NwCRC and Ob subjects (miR-345-5p). As regards the subject group-specific DEM, again the ObCRC category exhibited the highest number of selectively dysregulated miRNAs (Table 2). A Venn diagram was then generated to discover the common or unique lncRNAs and miRNAs among the three experimental groups (Ob, NwCRC, and ObCRC subjects) (Figure 1). By intersecting DEL and DEM data from the three comparisons (NwCRC, Ob and ObCRC individuals compared to Nw subjects), 13 ncRNAs were found to be shared between cancer and obese conditions. The identification of these differentially expressed ncRNAs, likely involved directly in creating a tumor-promoting microenvironment, may provide clues on the epigenetic mechanisms by which obesity favor CRC onset, as well as on how CRC development in obesity differs from that in lean individuals.

Identification of Target Genes Regulated by Differentially Expressed miRNAs

To investigate the potential involvement of the aforementioned DEM in the pathogenic events related to obesity and/or CRC, we next analyzed dysregulated miRNAs and validated consistency of differential expression of their targets. For each identified DEM, we extracted the list of experimentally validated mRNA

targets from TarBase and miRTarBase repositories. Based on our previously obtained gene expression dataset (28), we considered only those targets included in the list of differentially expressed transcripts (DET). We then assembled an interaction network between DEM and their target genes for each group (Figure 2). The complete list of DEM-DET interactions for each condition is reported in Supplemental Table 2.

In detail, interaction analysis showed 713 nodes (21 DEM and 692 target DET) and 1,669 edges in the NwCRC network (Supplemental Table 2), with two DEM having a number of directed edges ≥ 200 (hsa-let-7f-5p and hsa-miR-98-5p) and five DEM having $< 200 \geq 100$ directed edges (hsa-miR-193b-3p, hsa-miR-29b-3p, hsa-miR-125a-5p, hsa-miR-22-5p, and hsa-miR-374b-5p). Among the modulated genes, BTG2, and SON genes were the target of 10 DEM and other 33 DET interacted with more than five DEM. In the interaction network of Ob subjects, 808 nodes (20 DEM and 788 DET) and 1,759 edges were found (Supplemental Table 2), with hsa-miR-34a-5p having 420 directed edges and six DEM having over 100 directed edges (hsa-let-7c-5p, hsa-miR-24-3p, hsa-miR-193b-3p, hsa-miR-185-5p, hsa-miR-181c-5p, hsa-miR-125a-5p). SON was the target genes of 10 DEM and 33 DET interacted with more than five DEM. In ObCRC subjects 1,056 nodes (37 DEM and 1,019 DET) and 3,449 edges were found (Supplemental Table 2). hsa-miR-34a-5p and hsa-miR-107 had, respectively, 464 and 357 targets, four DEM had over 200 direct edges (hsa-miR-92a-3p, hsa-miR-24-3p, hsa-miR-98-5p, hsa-miR-30c-5p), seven DEM had $< 200 \geq 100$ directed edges (hsa-miR-10b-5p, hsa-miR-193b-3p, hsa-miR-22-5p, hsa-miR-125a-5p, hsa-miR-185-5p, hsa-miR-181c-5p, hsa-miR-181d-5p). The top interacting DET was again SON and other 19 DET had more than 10 directed edges.

Identification of Target Genes and microRNAs Regulated by Differentially Expressed lncRNAs

In addition to the miRNA regulatory networks, the dysregulation of lncRNA expression was recently associated with obesity and CRC (27, 42). Therefore, we constructed lncRNA-mRNA regulatory networks through an integrated analysis of the new identified DEL and the previously described DET (28), for each category of subjects.

As shown in Figure 3, only for a subgroup of DEL at least one experimentally validated interaction was found in the ENCORI database. In particular, in NwCRC subjects, the up-regulated DEL SNHG16, AC109460.3, NUTM2A-AS1, and STAG3L5P-PVRIG2P-PILRB, as well as the down-regulated AP000317.2, were relevant hubs each interacting with more than three DET (Figure 3A). In Ob subjects, main nodes were represented by the up-regulated SNHG16, AC109460.3, and MIR3142HG (Figure 3B). In ObCRC subjects, the down-regulated XIST interacted with 264 DET while the up-regulated SNHG16 and AC109460.3 interacted with more than 10 DET. Other 3 DEL (LINC01184, STAG3L5P-PVRIG2P-PILRB, AP000317.2) had more or equal than 5 directed edges (Figure 3C).

Since lncRNAs can bind to miRNAs to “communicate” with other RNA targets as well as to be reciprocally regulated by

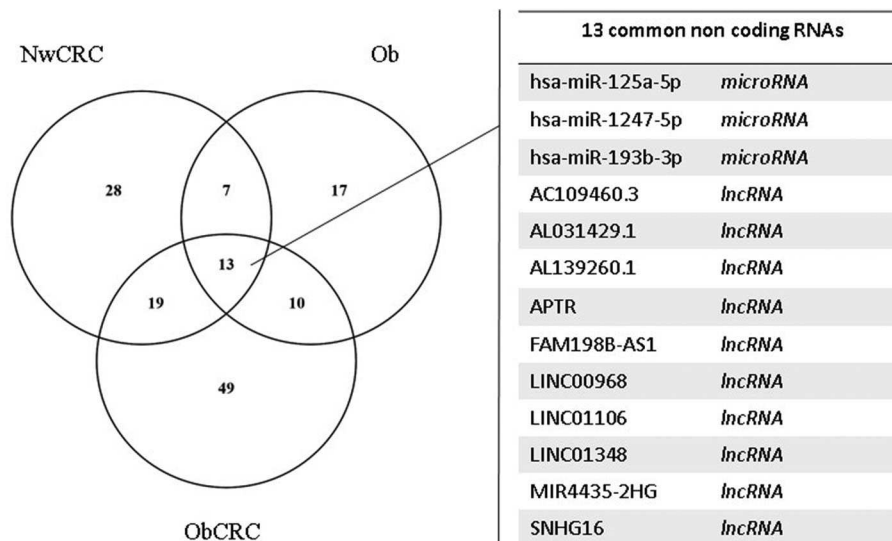


FIGURE 1 | Analysis of lncRNAs and miRNAs shared by obese and CRC-affected individuals or unique for each condition. Venn diagram showing unique or shared ncRNAs resulting by the comparison of DEL and DEM from all pathological conditions vs. healthy lean subjects. Each comparison is represented by a circle. The numbers in the region of the overlapping circles indicate the ncRNAs that are expressed in two or more conditions. The complete list of the 13 ncRNAs shared by obesity and CRC is shown on the right.

miRNAs (21), we then explored the ENCORI database for experimentally validated DEL-DEM interactions. As shown in **Figure 4**, DEL-DEM interaction networks in CRC patients, both lean and obese, displayed more interconnections than in obese individuals not affected by CRC. In particular, 95 relationship pairs between 28 DEL and 19 DEM were found in NwCRC patients, with the DEL USP9Y and AC006504.5 interacting with 12 and 11 DEM respectively; besides, hsa-miR-664a-3p and hsa-miR-22-5p were the top interaction DEM with 8 direct connections to DEL (**Figure 4A**). Likewise, in ObCRC subjects, 146 relationship pairs between 34 DEL and 28 DEM were found, with XIST and STAG3L5P-PVRIG2P-PILRB interacting with 23 and 11 DEM, respectively, and the top DEM hsa-miR-515-5p and hsa-miR-516b-5p interacted with 15 and 8 DEL, respectively (**Figure 4C**). Conversely, in Ob subjects only 37 relationship pairs between 16 DEL and 15 DEM were found, with AC021092.1 interacting with 5 DEM and hsa-miR-181d-5p and hsa-miR-181c-5p interacting with 5 DEL (**Figure 4B**).

mRNA-miRNA-lncRNA Regulatory Networks

In order to identify novel key regulators in the transcriptional and post-transcriptional adipocyte reprogramming under obesity and CRC conditions, integrated lncRNA-miRNA-mRNA networks were constructed for each conditions taking into account and combining the interactions described between miRNA/mRNA, lncRNA/miRNA, and lncRNA/mRNA.

In this regard, it is reported that a stronger connectivity of RNA nodes in the network can reflect the importance of the biological functions of these RNAs in the network. Therefore, hub nodes with degree exceeding 5 represent key players in biological networks (43). Based on this criterion, different

number and distribution of hubs, according to the RNA type, were identified in the three integrated networks. Specifically, we described 9 lncRNAs, 20 miRNAs, and 79 mRNAs hubs in the NwCRC network, 3 lncRNAs, 18 miRNAs, and 70 mRNAs hubs in the Ob network, and 10 lncRNAs, 36 miRNAs, and 308 mRNAs hubs in the ObCRC network, according to the higher complexity already described for the ObCRC condition in term of DEM-DET, DEL-DET, DEL-DEM interactions. Due to the complexity of the networks, only nodes with degree equal or higher than 6 are shown in **Figure 5**, whereas results description refers to the whole network. Focusing on ncRNAs, the most highly connected hubs in the NwCRC network were let-7f-5p, miR-98-5p, miR-193b-3p, miR-29b-3p, while SNHG16, and NUTM2A-AS1 had higher degrees compared with the other lncRNAs (**Figure 5A**). In the Ob network, miR-34a-5p, let-7c-5p, miR-24-3p, miR-193b-3p, and SNHG16, along with the novel lncRNA AC109460.3, were the most highly connected hubs ncRNAs (**Figure 5B**). Predominant nodes in the ObCRC network were miR-34a-5p, miR-107, miR-92a-3p, miR-24-3p, while the lncRNA XIST represents the main key interactor in the network (**Figure 5C**).

Searching for common key regulators, 23 genes were found to be the pivotal nodes in all networks, which include two miRNAs, miR-193b-3p, and miR-125a-5p, a known (SNHG16) and a novel (AC109460.3) lncRNA, and 19 mRNAs, indicating that these common elements and their interactors could be involved in relevant processes in obesity and CRC. Among the shared mRNA nodes, we found key players involved in the adipocyte transcriptional program (e.g., STAT3, RORA, CNOT1), in adipogenesis and lipogenesis processes (e.g., SEC31A, BMP2), and in food intake and hypothalamic signaling (e.g., SON, PRRC2A, CUX1).

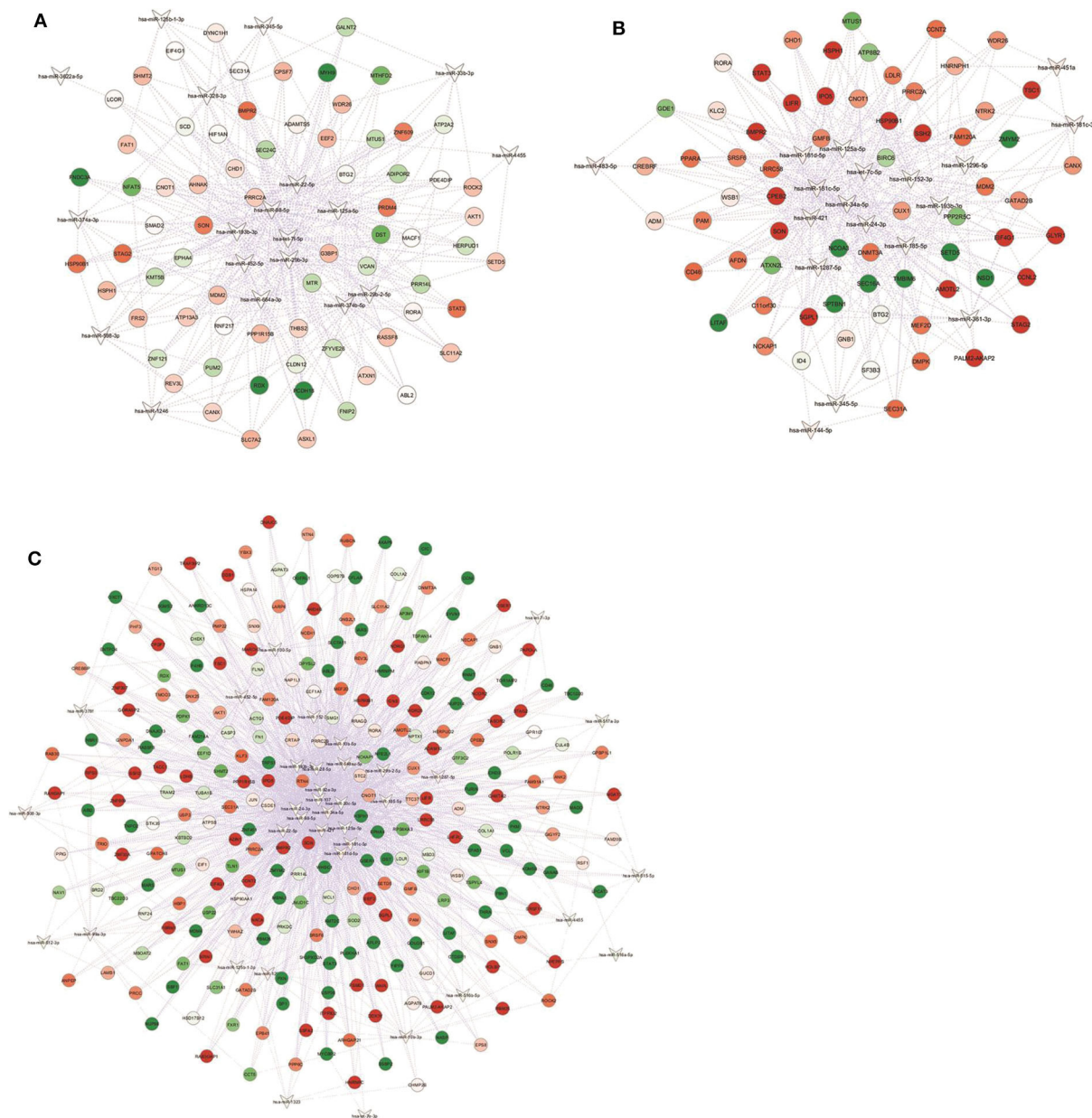


FIGURE 2 | miRNA-gene targets regulatory networks. Interaction networks between deregulated miRNAs (DEM; V-shape) and validated mRNA targets among modulated genes (DET; circles) in **(A)** NwCRC, **(B)** Ob, and **(C)** ObCRC patients in comparison to healthy lean subjects. Every node represents one gene, and each edge represents the interaction between genes. Only nodes with a number of directed edges ≥ 5 are shown (see **Supplemental Table 2** for the extended network). Shades of green and red indicate, respectively, down- or up-regulated DEM/DET.

Functional Enrichment Analysis of Networks-Related mRNA Targets

The biological function of a miRNA-lncRNA-mRNA network may be explained by the functions of the included target mRNAs. Thus, target genes of DEM and/or DEL found in the interaction networks of each subject group, were subjected to functional enrichment analysis combining different databases (KEGG, WikiPathways, and Reactome). The detailed list of terms, along with the genes involved in each term, are

reported in **Supplemental Table 3** and results are summarized in **Figure 6**.

In NwCRC patients numerous pathway terms associated with metabolic processes (e.g., *One-carbon metabolism*, *Purine metabolism*, *Cysteine/Methionine metabolism*), lipid metabolism (e.g., *Fatty acyl-coA biosynthesis*, *AMPK*, and *SREBP signaling*) and pathways involved in cancer (e.g., *Signaling by FGFR1 in disease*, *Pathway in clear cell renal cell carcinoma*) were obtained (**Figure 6A**). While the cancer pathways mainly

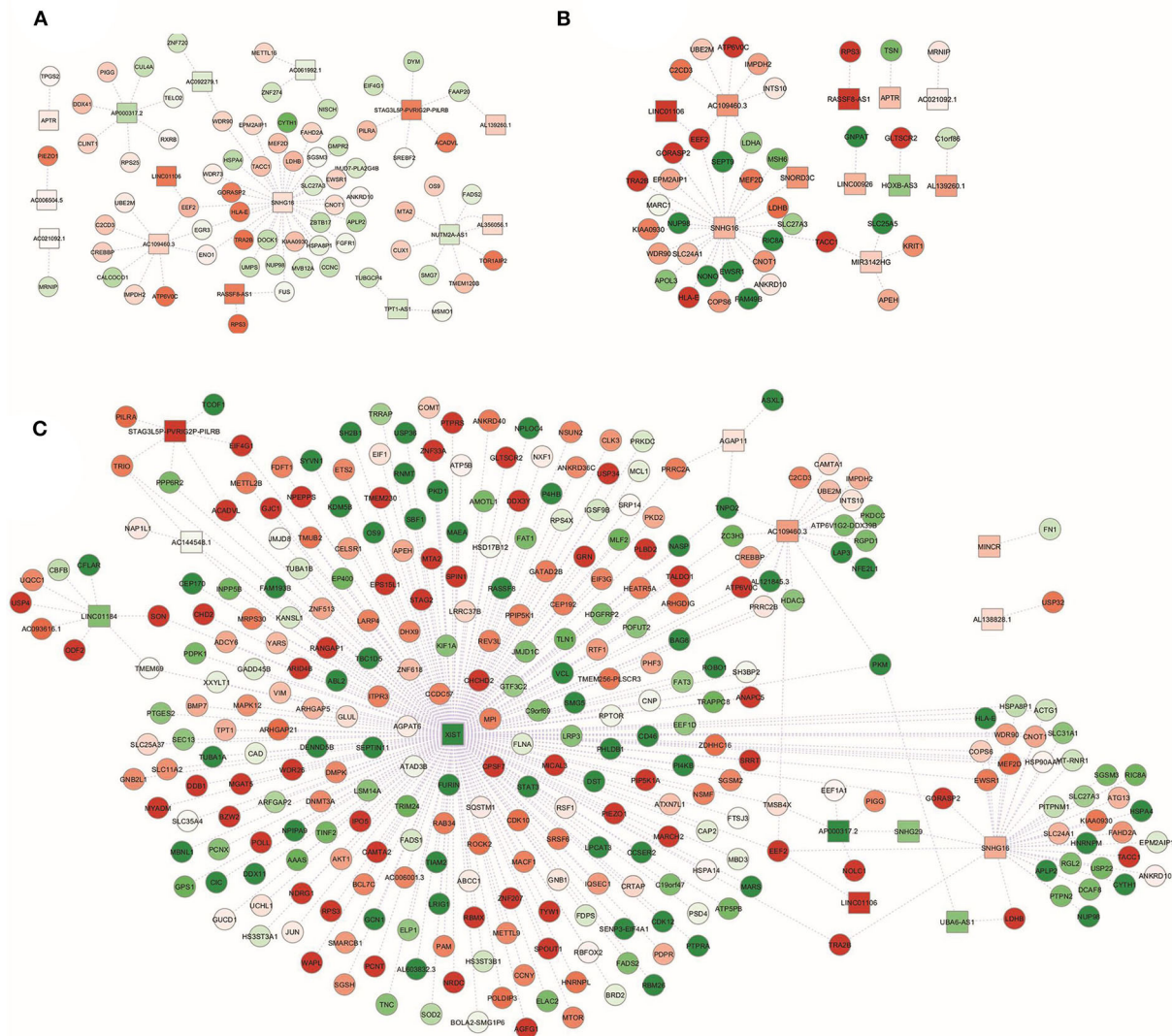


FIGURE 3 | lncRNA-gene targets regulatory networks. Interaction networks between deregulated lncRNAs (DEL; squares) and validated mRNA targets among modulated genes (DET; circles) in (A) NwCRC, (B) Ob, and (C) ObCRC patients in comparison to healthy lean subjects. Every node represents one gene, and each edge represents the interaction between genes. Shades of green and red indicate, respectively, down- or up-regulated DEL/DET.

featured up-regulated genes, the lipid metabolism pathways mainly included down-regulated genes. As expected, also the obesity-associated network was enriched in terms related to lipid metabolism (e.g., *Cholesterol biosynthesis*, *Glycerophospholipid Biosynthetic Pathway*). Further, in Ob individuals, we found enriched cancer pathways shared with CRC lean subjects (e.g., *Signaling by FGFR1*, *Pathway in clear cell renal cell carcinoma*, *Integrated Breast Cancer Pathway*), or unique of obese condition, such as a TP53-related pathway, all induced (Figure 6B). Finally, the ObCRC network (Figure 6C) was primarily enriched by fundamental biological functions that are implicated in inflammatory signaling pathway (e.g., *Platelet degranulation*, *TGF-beta signaling*, *IL-4*, and *IL13 signaling*), tumor suppression and insulin sensitivity (e.g., *Regulation*

of *PTEN* gene transcription, *Interleukin-37 signaling*, *Insulin resistance*), along with categories related to metabolism (e.g., *Pyruvate metabolism* and *Citric Acid cycle*, *AMPK signaling*) and cancer (e.g., *FGFR1 mutant receptor activation; signaling by VEGF*). Interestingly, in contrast to what observed for Ob and NwCRC networks, the majority of enriched categories featured under-expressed genes in ObCRC patients, with the exception, among others, of pathways related to energy metabolism (e.g., *mTOR signaling* and *AMPK signaling*), to the growth factor EGF (e.g., *EGF/EGFR signaling pathway*) and to neuronal development (e.g., *Netrin-1 signaling*).

Finally, pathways related to type I interferon signaling (e.g., *Interferon type I signaling pathway*, *ISG15 antiviral mechanism*, *antiviral mechanisms by IFN-stimulated genes*) are shared by

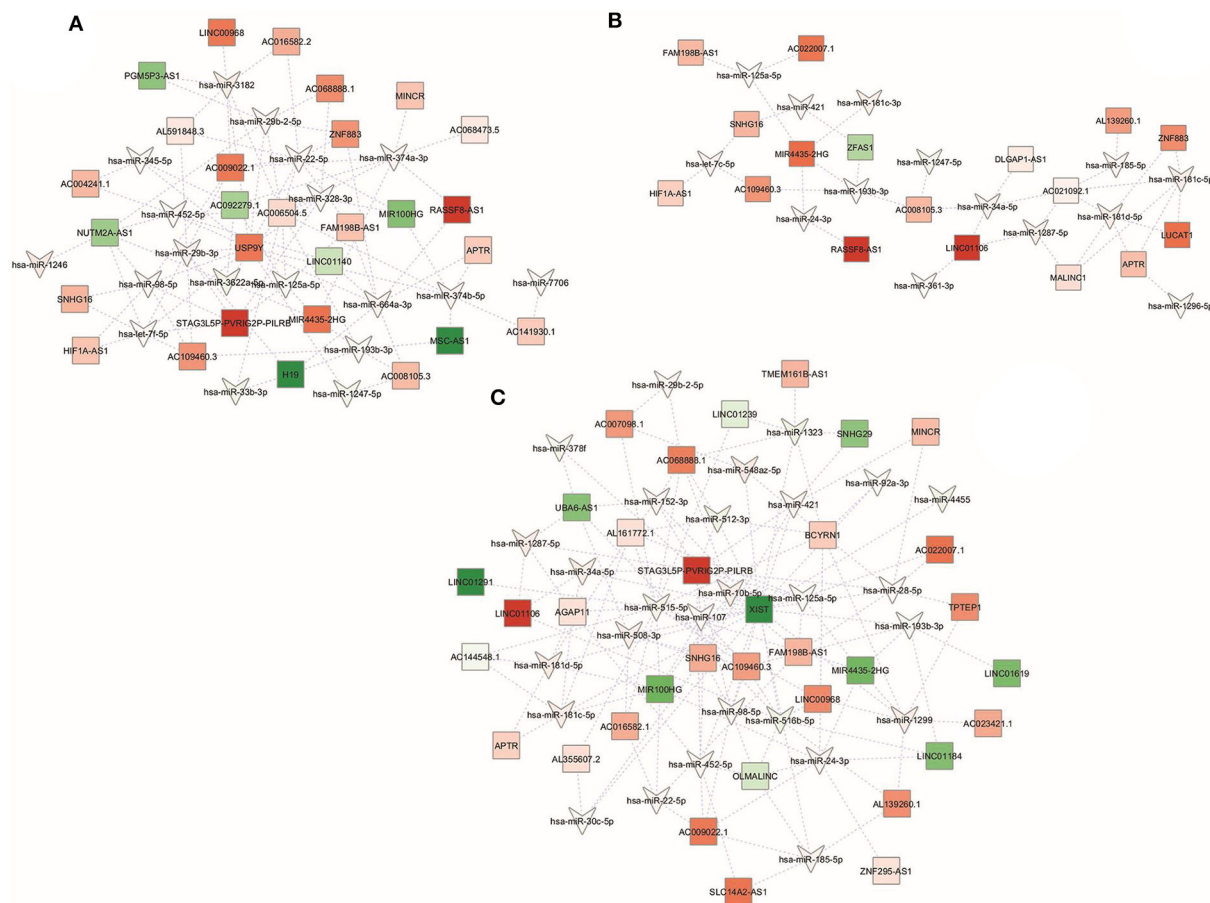


FIGURE 4 | lncRNA-miRNA interaction networks. Interaction networks between deregulated lncRNAs (DEL; squares) and miRNA (DEM; V-shapes) in (A) NwCRC, (B) Ob, and (C) ObCRC patients in comparison to healthy lean subjects. Every node represents one gene, and each edge represents the interaction between genes. Shades of green and red indicate, respectively, down- or up-regulated DEL/DEM.

obese and CRC networks. Furthermore, all networks described showed dysregulated genes belonging to processes involved in RNA regulation (e.g., *metabolism of RNA*), endocytosis and vesicle-mediated transport (e.g., *Membrane trafficking*, *Vesicle budding*, *Endocytosis*, *Extracellular matrix organization*) and sumoylation (e.g., *SUMO E3 ligases SUMOylate target proteins*). Interestingly, in ObCRC patients we observed a predominant pathway repression state, again indicating that the interplay between obesity- and CRC results in a specific modulation of adipocyte transcriptional and post-transcriptional program.

Validation by Using Real Time qPCR

The expression levels of pivotal transcripts were validated by RT-qPCR. Candidate transcripts were selected among those DEL and DEM found to be shared between cancer and obese conditions (e.g., LINC01106, LINC00968, SNHG16, miR-125a-5p, miR-193b-3p, miR-1247-5p), along with those of ncRNAs specific for CRC or obese subjects (e.g., XIST, H19, MINCR, miR-29b, miR-125b-1-3p, miR-181d-5p), on the basis of their relevance in the described regulatory networks. As shown in **Figure 7**,

the lncRNAs belonging to all categories of subjects (obese and CRC affected) were found to be significantly modulated compared to healthy lean subjects. Specifically, LINC01106 was significantly up-modulated in Ob and ObCRC, while H19 was significantly down-modulated in NwCRC patients. We failed to observe a significant up-regulation of LINC00698, MINCR and SNHG16 in NwCRC patients, although we confirmed their up-regulation in the other subject groups (Ob and ObCRC for LINC00698 and ObCRC for MINCR and SNHG16), according to RNASeq analysis (**Figure 7A**). Overall, RNASeq and qPCR data displayed a significant positive correlation ($Rho = 0.829$; $p < 0.0001$). Similarly, in the case of miRNAs (**Figure 7B**), qPCR analysis confirmed the down-modulation of miR-125b-1-3p in all conditions and miR-193b only in Ob and ObCRC, whereas the under-expression of miR-1247 and miR-125a-5p was validated in Ob and ObCRC groups or ObCRC group, respectively. We also reported an up-regulation of miR-181d-5p in both Ob and CRC affected subjects, although RNASeq data showed its over-expression in Ob subjects only. In contrast to what observed from RNASeq analysis, no differential expression

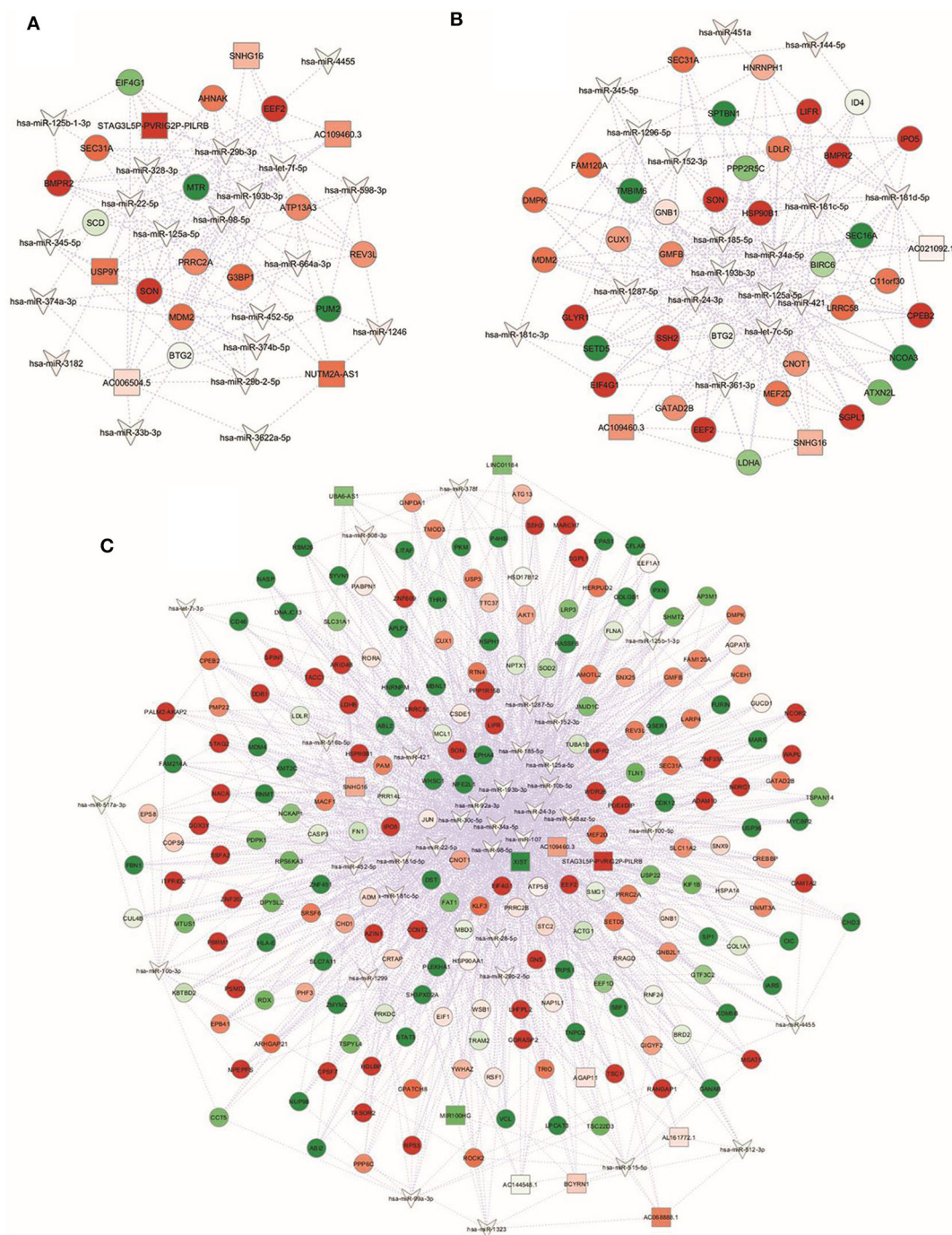
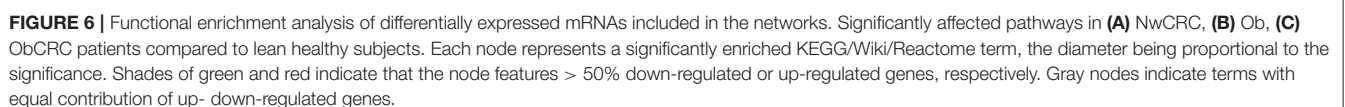


FIGURE 5 | Global view of the miRNA-lncRNA-mRNA interaction networks. miRNAs are indicated with a V-shape, lncRNAs are indicated with squares, and mRNAs are indicated with circles. Only nodes with a number of directed edges ≥ 6 are shown. Shades of green and red indicate, respectively, down- or up-regulated DEM/DET/DEL. (A) NwCRC, (B) Ob, (C) ObCRC individuals in comparison to healthy lean subjects.

of miR-29b-3p was found in all groups of subjects. Overall, although we did not achieve a complete correspondence between miRNA expression data from the two different techniques, qPCR and RNASeq results were significantly correlated ($Rho = 0.6079$; $p = 0.0074$).

DISCUSSION

The prevalence of obesity and obesity-associated diseases, including CRC, is in constant increase, accounting for a large portion of public health challenges. These multifactorial and



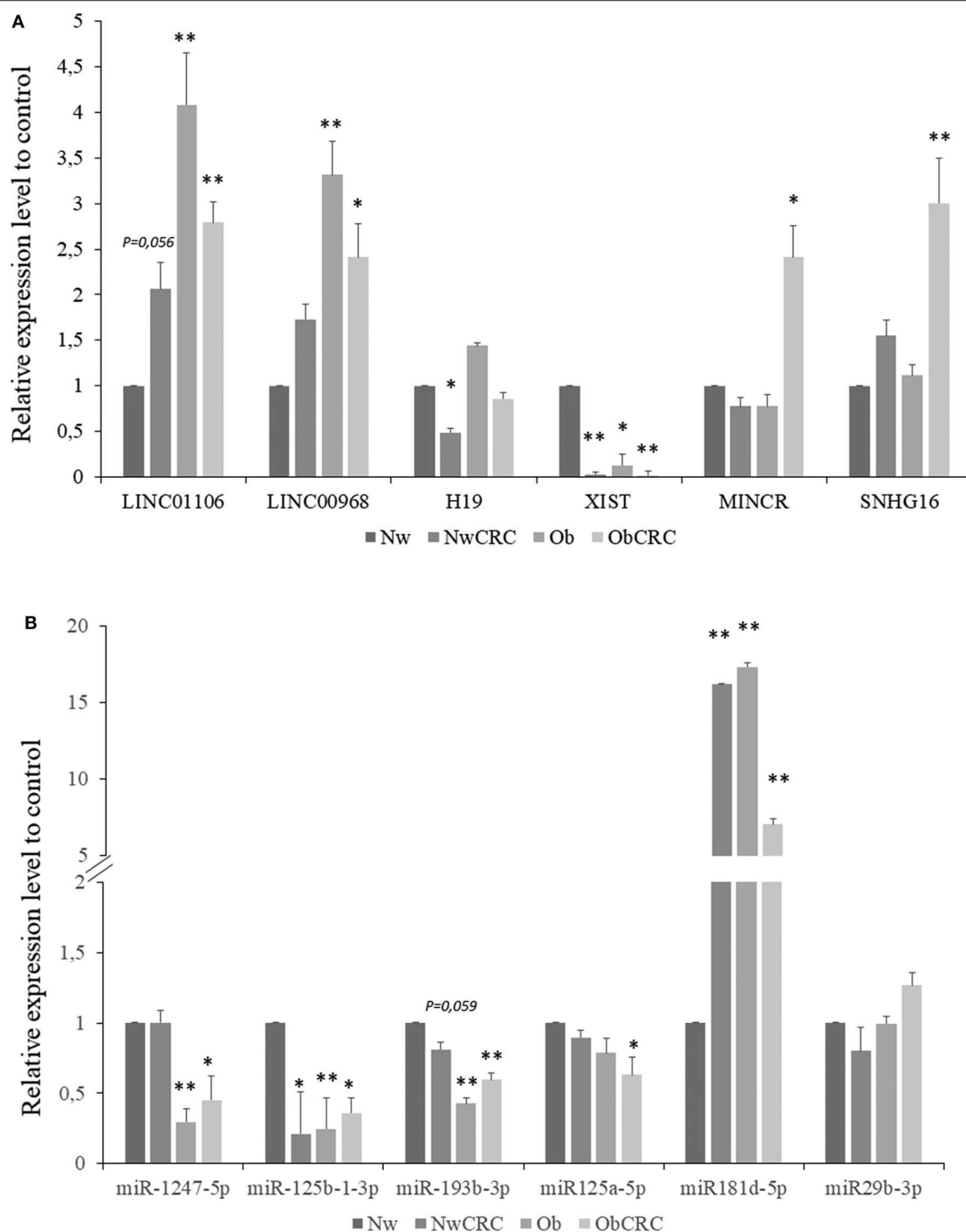


FIGURE 7 | Validation by real-time qPCR of selected lncRNAs and miRNAs. Expression level of selected lncRNAs (A) or miRNAs (B) for NwCRC, Ob, and ObCRC subjects were normalized to healthy lean control. Statistical significance is indicated with * for $p \leq 0.05$, and ** for $p \leq 0.005$ vs. Nw control.

complex disorders are strongly interconnected, although the mechanisms underlying the higher susceptibility to cancer development and the poorer cancer prognosis in obese individuals are still a matter of debate. Different components of the AT microenvironment, such as chronic inflammation,

vascularity and fibrosis, altered levels of sex hormones, insulin resistance, are nowadays recognized as important determinants of CRC risk. Moreover, adipocytes release lipids acting as an energy reservoir for cancer cells, while the rapid expansion of AT in obesity produces hypoxia and promotes angiogenesis, favoring

the tumor spread (7, 44, 45). Recent findings in epigenetics emphasized an important functional role of miRNAs, as well as of lncRNAs, in pathophysiological processes. The dysregulation of these transcripts, in fact, has been found in pathological conditions such as cancer and dysmetabolic disorders including obesity. In AT, miRNAs regulate all aspects of the adipocyte biology, including inflammation and adipokines production, metabolic responses, lipolysis and lipogenesis, adipogenesis and browning (9, 46). Likewise, the total number of lncRNAs identified in AT and found to modulate adipose function, is rapidly increasing (26, 47–49). Several studies reported the involvement of lncRNAs in adipogenesis and lipid metabolism (27, 50) as well as in AT function and development in mouse models (51, 52). Nevertheless, their implication in human adipocytes remains largely unknown. Likewise, no definitive conclusions regarding the molecular factors and the mechanistic processes underlying the relationships among obesity, AT dysfunction and CRC have been reached so far. To the best of our knowledge, this is the first comparative study that performed an integrated multi-omic analysis on human visceral adipocytes to assess how obesity, alone or combined to CRC, affects miRNA, and lncRNA expression and networks, as a potential mechanism linking obesity and CRC.

The expression of miRNAs of obese subjects with respect to lean individuals has been previously investigated in both VAT and subcutaneous AT (SAT), the two main fat depots that exhibit significant differences in anatomical, cellular and molecular features (6, 53). Heterogeneity of subjects (fat depots, BMI), type of samples (isolated adipocytes compared to adipose tissue), together with the use of different high-throughput techniques (arrays, RNA sequencing) has rendered difficult to identify a specific “miRNA signature” altered in obesity (9). In this regard, differences in miRNAs expression were observed when comparing visceral and subcutaneous fat (17, 54, 55), or isolated adipocytes and whole AT (56). In our study, we performed a whole analysis of miRNAs in human adipocytes isolated from the visceral fat. Among the miRNAs dysregulated in obese subjects compared to the normal weight controls, we found those involved in adipogenesis (e.g., let-7 family, miR-193b, –483-5p), in lipid metabolism (e.g., miR-181d), or in glucose and insulin metabolism (e.g., miR-34a-5p, –24-3p, –144-5p, –361-3p), previously described in different AT depots of obese subjects (9, 57–59), further supporting a role of these miRNAs in the functional alterations of adipocytes occurring in obesity. Additionally, in obese adipocytes we also reported the dysregulation of those miRNAs previously found to be involved in the regulation of immune response, adipokine secretion and inflammation (e.g., miR-125a-5p; –181 family, –193b) (15, 60, 61) or implicated in many aspects of carcinogenesis in several cancer types, including CRC (e.g., miR-34a, let7e-3p, –144-5p, –193b, –361-3p, –451a) (54, 62). Specifically, we found that miR-125a-5p and miR-193b-3p were downregulated in both obesity and CRC, in keeping with their previously reported down-regulation in VAT of obese subjects (63, 64), although contrasting results on miR-193b expression have been showed in human SAT (56, 64). Notably, we have previously described an up-regulation of the target genes of miR-193b (i.e., CCL2

and miR-125a-5p (i.e., STAT3), as an important mechanism underlying obesity-associated inflammation (29, 65), according to the literature (54, 56). Furthermore, we also report the characterization of 35 novel and 55 known lncRNAs in visceral adipocytes. An important property of lncRNAs is their cell- and tissue- specific expression (66). Therefore, the current annotation of lncRNAs is far from being complete. Alterations in the expression of some lncRNAs have been reported in both SAT and VAT, as important regulators of AT functions (26, 27, 67). In our study, we report the first analysis of lncRNAs in purified visceral adipocytes and this could explain the discrepancies observed with previous studies mainly conducted in whole AT (26, 27, 67). In general, we identified known and novel lncRNAs not previously described in other reports. Specifically, in obese subjects we found several lncRNAs (e.g., ZFAS1, LUCAT1, HIF1A-AS1, HOXB-AS3) already identified in the setting of different type of cancers, but not previously reported in human AT. Moreover, the lncRNA MIR3142HG, recently described as important mediator of the inflammatory response in Idiopathic Pulmonary Lung Fibroblasts positively regulating CXCL8 and CCL2 release (68), is specifically up-modulated in obesity. Notably, we previously reported an upregulation of both CCL2 and CXCL8 in adipocytes from Ob individuals (29), suggesting a role of this lncRNA in the AT inflammation. Other two lncRNAs, SNHG16, and LINC01106, were found to be upregulated in obesity, and this modulation was shared between obese and cancer conditions. In this regard, an abnormal expression of SNHG16 has been observed in multiple cancers and usually correlates with worse pathological features (69), while the novel lncRNA LINC01106 has been recently reported to be related to the overall survival of CRC patients by acting as inflammatory mediator in inflammatory bowel disease (IBD)-related CRC. This lncRNA showed also an intimate interaction with miR-193a in epithelial tissue from IBD and CRC patients (70).

Despite the well-known link between AT related inflammation and CRC development, no previous studies considered the expression of ncRNAs in the AT of CRC patients. When overlapping the data from NwCRC, Ob, and ObCRC individuals, the down-regulation of miRNAs, such as miR-193b-3p, miR-125a-5p, and miR-1247-5p, was found to be shared between cancer and obese conditions. Interestingly, both miR-193b-3p, and miR-1247-5p act as tumor suppressors in CRC or other types of cancer (71, 72), suggesting that their repression in AT from Ob and CRC individuals could have a potential pro-tumorigenic role. Beside common features, some ncRNAs are unique of tumor conditions. For instance, lncRNA H19, among others, was repressed only in NwCRC patients, with respect to healthy control. Interestingly, H19 has been described to play a role in obesity-induced cancer and to promote epithelial-mesenchymal transition of CRC, with a reported poor prognosis for cancer patients exhibiting H19 induction (73, 74). However, we observed an opposite expression in AT compared to cancer cells, suggesting a different role of this lncRNA in visceral adipocytes, that could potentially involve H19 target genes STAT3 and SPARC (75, 76). Indeed, we and others previously reported a key role of STAT3 and SPARC in AT dysfunctions both in obese (28, 65, 77) and CRC conditions (28, 65). Similarly,

the lncRNA XIST is highly down-modulated in the AT from CRC group, although its up-regulation in CRC tissues and cell lines was reported (75, 78, 79). Remarkably, XIST can act as oncogene or tumor suppressor depending on the human malignancies (80) and was recently identified as a candidate in mediating glucose metabolism in glioma and contributing to cancer progression (81).

In this study, we not only identified some specific lncRNAs and miRNAs across the adipocyte genome, but we also described miRNA-lncRNA-mRNA interaction networks and the functional analysis of the pathways in which the target genes are involved. The target genes we identified in the networks were mainly enriched in several pathways, associated with metabolic processes, lipid and energy metabolism, inflammation, and cancer. Specifically, the SREBP pathway was remarkably inhibited in the NwCRC network, with implications not only on lipid metabolism but also on inflammation-mediated metabolic diseases, as well as on immune responses (82). Of note, the lncRNA SNHG16, that we have identified as a main hub of this network, has been reported to modulate the lipogenesis via regulation of SREBP2 expression (83), and to affect others genes involved in lipid metabolism (84). Another intriguing connection identified in Ob network is the upregulated TP53 transcriptional regulation pathway. The activation of this pathway has been previously observed in obesity and correlated to the release of inflammatory cytokines fueling cancer initiation and progression (85), thus potentially setting the basis for a more tumor-prone AT microenvironment in obese subjects. Furthermore, p53 in human AT was shown to be involved in insulin resistance, adipogenesis, lipid metabolism and nutrient sensing (86).

We also previously reported the influence of obesity on the adipocyte transcriptional program in CRC, with ObCRC subjects showing a higher number of dysregulated genes and processes (28). Likewise, in this study we observe a higher complexity of ObCRC network in terms of lncRNA and miRNA profiles. Interestingly, we describe in ObCRC patients the deregulation of fundamental biological functions that are mainly implicated in inflammatory signaling pathways, such as IL-37 and IL-13 signaling. In this regard, an increase expression of the cytokine IL-13, contributing to AT inflammation, has been reported to play an important role in obesity-related colon carcinogenesis (87), while IL-37 signaling has been described to play an inhibitory role in innate immune responses. In fact, it acts by reducing systemic and local inflammation, whereas its expression in SAT was negatively correlated with BMI (88). Other enriched categories in ObCRC network are: (i) TGF- β signaling that has been reported to regulate multiple aspects of AT biology (i.e., vascularization, inflammation and fibrosis) (89), (ii) Netrin-1 signaling, recently described to play a role in tissue regulation outside the nervous system, specifically in tumor development (i.e., angiogenesis and inflammation) and (iii) PTEN regulation, for which a dual role as tumor suppressor and metabolic regulator has been reported (90). Finally, the networks described in all subject groups were enriched in: (i) type I IFN signaling, recently identified as essential in the regulation of metabolism and in maintaining AT function (91), (ii) SUMOylation, a post-translational modification mechanism that plays an emerging role in cellular metabolism and metabolic disease (92) and

(iii) pathways involved in RNA metabolism, as expected. The identification of these pathways in both obese and cancer groups strongly points to the local metabolic alterations in AT as key element in colorectal carcinogenesis.

Additionally, pathways related to membrane trafficking, vesicle budding and endocytosis processes were also found to be dysregulated in both obesity and CRC networks. In this regard, it is worth to note that in addition to act locally, adipocytes influence and communicate with distant organs and tissues, by releasing bioactive molecules, such as triglycerides, adipokines, cytokines, and free fatty acids (93). This ability allows even tumors with no direct contact with AT to be affected by obesity, as indicated by epidemiological studies linking obesity with several types of cancers (94). Among adipocytes products that could sustain cancer cell growth, circulating miRNAs, both naked or associated to exosomes, may regulate the function of the immune system and distant organs and could potentially be used as biomarkers of diagnosis and prognosis of obesity and cancer (15). Likewise, exosomal lncRNAs have been shown to promote angiogenesis, cell proliferation and drug resistance and can be found in several body fluids, being highly stable, thus considered potential tumor biomarkers (95).

In conclusion, the importance of understanding the role of lncRNAs and miRNAs in AT of obese and CRC affected subjects extends beyond the description of gene regulation mechanisms. The results obtained in this study, through a multi-omics approach and computational analysis, contribute to the identification of candidate genes, ncRNAs and their regulatory networks relevant to many AT biological processes, although the direct causality remains to be established, requiring further experimental and functional studies. Nonetheless, the identification of AT miRNAs and lncRNAs as key components of interrelated processes and pathways may not only better define their role in human AT, but also identify promising mechanism-based targets, to disrupt the relationship between obesity, metabolic dysregulation, and cancer, potentially improving intervention and treatment plans.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the NCBI Sequence Read Archive (SRA) (<http://www.ncbi.nlm.nih.gov/sra>) (PRJNA632999, PRJNA508473).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the institutional review board of Istituto Superiore di Sanità. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

RV and BS isolated adipocytes from human visceral adipose tissue biopsies. AB prepared samples for RNA Sequencing and performed real-time qPCR for gene validation. ST, AM, EC, and PM performed bioinformatics and statistical analyses of RNASeq

data. ST and AM provided intellectual input throughout the study. MD and SG provided substantial contributions to the conception of the work as well as interpretation of data and manuscript writing. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by a grant of the Italian Association for Cancer Research (AIRC) (IG 2013 N14185) to SG.

REFERENCES

- WHO. Obesity and Overweight. Fact sheet (2015). Available online at: <http://www.who.int/mediacentre/factsheets/fs311/en/>
- Unamuno X, Gomez-Ambrosi J, Rodriguez A, Becerril S, Fruhbeck G, Catalan V. Adipokine dysregulation and adipose tissue inflammation in human obesity. *Eur J Clin Invest.* (2018) 48:e12997. doi: 10.1111/eci.12997
- O'Sullivan J, Lysaght J, Donohoe CL, Reynolds JV. Obesity and gastrointestinal cancer: the interrelationship of adipose and tumour microenvironments. *Nat Rev Gastroenterol Hepatol.* (2018) 15:699–714. doi: 10.1038/s41575-018-0069-7
- van Kruijsdijk RC, van der Wall E, Visseren FL. Obesity and cancer: the role of dysfunctional adipose tissue. *Cancer Epidemiol Biomarkers Prev.* (2009) 18:2569–78. doi: 10.1158/1055-9965.EPI-09-0372
- Keum N, Giovannucci E. Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. *Nat Rev Gastroenterol Hepatol.* (2019) 16:713–32. doi: 10.1038/s41575-019-0189-8
- Bardou M, Barkun AN, Martel M. Republished: obesity and colorectal cancer. *Postgrad Med J.* (2013) 89:519–33. doi: 10.1136/postgradmedj-2013-304701rep
- Park J, Morley TS, Kim M, Clegg DJ, Scherer PE. Obesity and cancer—mechanisms underlying tumour progression and recurrence. *Nat Rev Endocrinol.* (2014) 10:455–65. doi: 10.1038/nrendo.2014.94
- Martinez-Useros J, Garcia-Foncillas J. Obesity and colorectal cancer: molecular features of adipose tissue. *J Transl Med.* (2016) 14:21. doi: 10.1186/s12967-016-0772-5
- Lorente-Cebrian S, Gonzalez-Muniesa P, Milagro FI, Martinez JA. MicroRNAs and other non-coding RNAs in adipose tissue and obesity: emerging roles as biomarkers and therapeutic targets. *Clin Sci.* (2019) 133:23–40. doi: 10.1042/CS20180890
- Ayers D, Boughanem H, Macias-Gonzalez M. Epigenetic influences in the obesity/colorectal cancer axis: a novel theragnostic Avenue. *J Oncol.* (2019) 2019:7406078. doi: 10.1155/2019/7406078
- Rinn JL, Chang HY. Genome regulation by long noncoding RNAs. *Annu Rev Biochem.* (2012) 81:145–66. doi: 10.1146/annurev-biochem-051410-092902
- Marchese FP, Raimondi I, Huarte M. The multidimensional mechanisms of long noncoding RNA function. *Genome Biol.* (2017) 18:206. doi: 10.1186/s13059-017-1348-2
- Ambros V. The functions of animal microRNAs. *Nature.* (2004) 431:350–5. doi: 10.1038/nature02871
- Bartel DP. MicroRNAs: target recognition and regulatory functions. *Cell.* (2009) 136:215–33. doi: 10.1016/j.cell.2009.01.002
- Arner P, Kulyte A. MicroRNA regulatory networks in human adipose tissue and obesity. *Nat Rev Endocrinol.* (2015) 11:276–88. doi: 10.1038/nrendo.2015.25
- Zampetaki A, Mayr M. MicroRNAs in vascular and metabolic disease. *Circ Res.* (2012) 110:508–22. doi: 10.1161/CIRCRESAHA.111.247445
- Kloting N, Berthold S, Kovacs P, Schon MR, Fasshauer M, Ruschke K, et al. MicroRNA expression in human omental and subcutaneous adipose tissue. *PLoS ONE.* (2009) 4:e4699. doi: 10.1371/journal.pone.0004699
- Rottiers V, Naar AM. MicroRNAs in metabolism and metabolic disorders. *Nat Rev Mol Cell Biol.* (2012) 13:239–50. doi: 10.1038/nrm3313

ACKNOWLEDGMENTS

We are indebted to Drs. R. Persiani, G. Silecchia, and A. Iacovelli for kindly providing clinical samples.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fonc.2020.01089/full#supplementary-material>

- Tam C, Wong JH, Tsui SKW, Zuo T, Chan TF, Ng TB. LncRNAs with miRNAs in regulation of gastric, liver, and colorectal cancers: updates in recent years. *Appl Microbiol Biotechnol.* (2019) 103:4649–77. doi: 10.1007/s00253-019-09837-5
- Mercer TR, Dinger ME, Mattick JS. Long non-coding RNAs: insights into functions. *Nat Rev Genet.* (2009) 10:155–9. doi: 10.1038/nrg2521
- Salmerna L, Poliseno L, Tay Y, Kats L, Pandolfi PP. A ceRNA hypothesis: the rosetta stone of a hidden RNA language? *Cell.* (2011) 146:353–8. doi: 10.1016/j.cell.2011.07.014
- Huarte M. The emerging role of lncRNAs in cancer. *Nat Med.* (2015) 21:1253–61. doi: 10.1038/nm.3981
- Li X, Wu Z, Fu X, Han W. Long noncoding RNAs: insights from biological features and functions to diseases. *Med Res Rev.* (2013) 33:517–53. doi: 10.1002/med.21254
- Yuan W, Li X, Liu L, Wei C, Sun D, Peng S, et al. Comprehensive analysis of lncRNA-associated ceRNA network in colorectal cancer. *Biochem Biophys Res Commun.* (2019) 508:374–9. doi: 10.1016/j.bbrc.2018.11.151
- Landrier JF, Derghal A, Mounien L. MicroRNAs in obesity and related metabolic disorders. *Cells.* (2019) 8:859. doi: 10.3390/cells8080859
- Gao H, Kerr A, Jiao H, Hon CC, Ryden M, Dahlman I, et al. Long non-coding RNAs associated with metabolic traits in human white adipose tissue. *EBioMedicine.* (2018) 30:248–60. doi: 10.1016/j.ebiom.2018.03.010
- Sun L, Lin JD. Function and mechanism of long noncoding RNAs in adipocyte biology. *Diabetes.* (2019) 68:887–96. doi: 10.2337/dbi18-0009
- Del Corno M, Baldassarre A, Calura E, Conti L, Martini P, Romualdi C, et al. Transcriptome profiles of human visceral adipocytes in obesity and colorectal cancer unravel the effects of body mass index and polyunsaturated fatty acids on genes and biological processes related to tumorigenesis. *Front Immunol.* (2019) 10:265. doi: 10.3389/fimmu.2019.00265
- Del Corno M, D'Archivio M, Conti L, Scazzocchio B, Vari R, Donninelli G, et al. Visceral fat adipocytes from obese and colorectal cancer subjects exhibit distinct secretory and omega6 polyunsaturated fatty acid profiles and deliver immunosuppressive signals to innate immunity cells. *Oncotarget.* (2016) 7:63093–105. doi: 10.18632/oncotarget.10998
- Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc.* (2009) 4:1184–91. doi: 10.1038/nprot.2009.97
- R Core Team (2019). *The Comprehensive R Archive Network*. Vienna: R Core Team. Available online at: <https://cran.r-project.org/>
- Friedlander MR, Mackowiak SD, Li N, Chen W, Rajewsky N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.* (2012) 40:37–52. doi: 10.1093/nar/gkr688
- Griffiths-Jones S. miRBase: the microRNA sequence database. *Methods Mol Biol.* (2006) 342:129–38. doi: 10.1385/1-59745-123-1:129
- Risso D, Schwartz K, Sherlock G, Dudoit S. GC-content normalization for RNA-Seq data. *BMC Bioinformatics.* (2011) 12:480. doi: 10.1186/1471-2105-12-480
- Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* (2010) 26:139–40. doi: 10.1093/bioinformatics/btp616
- Karagkouni D, Paraskevopoulou MD, Chatzopoulos S, Vlachos IS, Tastsoglou S, Kanellos I, et al. DIANA-TarBase v8: a decade-long collection of

- experimentally supported miRNA-gene interactions. *Nucleic Acids Res.* (2018) 46:D239–45. doi: 10.1093/nar/gkx1141
37. Chou CH, Shrestha S, Yang CD, Chang NW, Lin YL, Liao KW, et al. miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res.* (2018) 46:D296–302. doi: 10.1093/nar/gkx1067
 38. Paraskevopoulou MD, Vlachos IS, Hatzigeorgiou AG. DIANA-TarBase and DIANA suite tools: studying experimentally supported microRNA targets. *Curr Protoc Bioinformatics.* (2016) 55:12.14.1–18. doi: 10.1002/cpbi.12
 39. Li JH, Liu S, Zhou H, Qu LH, Yang JH. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.* (2014) 42:D92–7. doi: 10.1093/nar/gkt1248
 40. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* (2003) 13:2498–504. doi: 10.1101/gr.1239303
 41. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics.* (2009) 25:1091–3. doi: 10.1093/bioinformatics/btp101
 42. He Q, Long J, Yin Y, Li Y, Lei X, Li Z, et al. Emerging roles of lncRNAs in the formation and progression of colorectal cancer. *Front Oncol.* (2020) 9:1542. doi: 10.3389/fonc.2019.01542
 43. Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, et al. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature.* (2004) 430:88–93. doi: 10.1038/nature02555
 44. Oh TH, Byeon JS, Myung SJ, Yang SK, Choi KS, Chung JW, et al. Visceral obesity as a risk factor for colorectal neoplasm. *J Gastroenterol Hepatol.* (2008) 23:411–7. doi: 10.1111/j.1440-1746.2007.05125.x
 45. Ritchie SA, Connell JM. The link between abdominal obesity, metabolic syndrome and cardiovascular disease. *Nutr Metab Cardiovasc Dis.* (2007) 17:319–26. doi: 10.1016/j.numecd.2006.07.005
 46. Brandao BB, Guerra BA, Mori MA. Shortcuts to a functional adipose tissue: the role of small non-coding RNAs. *Redox Biol.* (2017) 12:82–102. doi: 10.1016/j.redox.2017.01.020
 47. Zhang XT, Pan SX, Wang AH, Kong QY, Jiang KT, Yu ZB. Long non-coding RNA (lncRNA) X-Inactive Specific Transcript (XIST) plays a critical role in predicting clinical prognosis and progression of colorectal cancer. *Med Sci Monit.* (2019) 25:6429–35. doi: 10.12659/MSM.915329
 48. Ding C, Lim YC, Chia SY, Walet ACE, Xu S, Lo KA, et al. *De novo* reconstruction of human adipose transcriptome reveals conserved lncRNAs as regulators of brown adipogenesis. *Nat Commun.* (2018) 9:1329. doi: 10.1038/s41467-018-03754-3
 49. Kornfeld JW, Bruning JC. Regulation of metabolism by long, non-coding RNAs. *Front Genet.* (2014) 5:57. doi: 10.3389/fgene.2014.00057
 50. Chen C, Cui Q, Zhang X, Luo X, Liu Y, Zuo J, et al. Long non-coding RNAs regulation in adipogenesis and lipid metabolism: emerging insights in obesity. *Cell Signal.* (2018) 51:47–58. doi: 10.1016/j.cellsig.2018.07.012
 51. Zhao XY, Li S, DelProposto JL, Liu T, Mi L, Porsche C, et al. The long noncoding RNA Blnc1 orchestrates homeostatic adipose tissue remodeling to preserve metabolic health. *Mol Metab.* (2018) 14:60–70. doi: 10.1016/j.molmet.2018.06.005
 52. Alvarez-Dominguez JR, Bai Z, Xu D, Yuan B, Lo KA, Yoon MJ, et al. *De novo* reconstruction of adipose tissue transcriptomes reveals long non-coding RNA regulators of brown adipocyte development. *Cell Metab.* (2015) 21:764–76. doi: 10.1016/j.cmet.2015.04.003
 53. Klimcakova E, Roussel B, Marquez-Quinones A, Kovacova Z, Kovacikova M, Combes M, et al. Worsening of obesity and metabolic status yields similar molecular adaptations in human subcutaneous and visceral adipose tissue: decreased metabolism and increased immune response. *J Clin Endocrinol Metab.* (2011) 96:E73–82. doi: 10.1210/jc.2010-1575
 54. Kurylowicz A, Wicik Z, Owczarz M, Jonas MI, Kotlarek M, Swierniak M, et al. NGS reveals molecular pathways affected by obesity and weight loss-related changes in miRNA levels in adipose tissue. *Int J Mol Sci.* (2017) 19:66. doi: 10.3390/ijms19010066
 55. Oger F, Gheeraert C, Mogilenko D, Benomar Y, Molendi-Coste O, Bouchaert E, et al. Cell-specific dysregulation of microRNA expression in obese white adipose tissue. *J Clin Endocrinol Metab.* (2014) 99:2821–33. doi: 10.1210/jc.2013-4259
 56. Arner E, Mejhert N, Kulyte A, Balwierz PJ, Pachkov M, Cormont M, et al. Adipose tissue microRNAs as regulators of CCL2 production in human obesity. *Diabetes.* (2012) 61:1986–93. doi: 10.2337/db11-1508
 57. Abu-Farha M, Cherian P, Al-Khairi I, Nizam R, Alkandari A, Arefanian H, et al. Reduced miR-181d level in obesity and its role in lipid metabolism via regulation of ANGPTL3. *Sci Rep.* (2019) 9:11866. doi: 10.1038/s41598-019-48371-2
 58. Jones A, Danielson KM, Benton MC, Ziegler O, Shah R, Stubbs RS, et al. miRNA signatures of insulin resistance in obesity. *Obesity.* (2017) 25:1734–44. doi: 10.1002/oby.21950
 59. Chen K, He H, Xie Y, Zhao L, Zhao S, Wan X, et al. miR-125a-3p and miR-483-5p promote adipogenesis via suppressing the RhoA/ROCK1/ERK1/2 pathway in multiple symmetric lipomatosis. *Sci Rep.* (2015) 5:11909. doi: 10.1038/srep11909
 60. Belarbi Y, Mejhert N, Lorente-Cebrian S, Dahlman I, Arner P, Ryden M, et al. MicroRNA-193b controls adiponectin production in human white adipose tissue. *J Clin Endocrinol Metab.* (2015) 100:E1084–8. doi: 10.1210/jc.2015-1530
 61. Lorente-Cebrian S, Mejhert N, Kulyte A, Laurencikiene J, Astrom G, Heden P, et al. MicroRNAs regulate human adipocyte lipolysis: effects of miR-145 are linked to TNF-alpha. *PLoS ONE.* (2014) 9:e86800. doi: 10.1371/journal.pone.0086800
 62. Li J, Zhou C, Li J, Su Z, Sang H, Jia E, et al. Global correlation analysis for microRNA and gene expression profiles in human obesity. *Pathol Res Pract.* (2015) 211:361–8. doi: 10.1016/j.prp.2014.11.014
 63. Diawara MR, Hue C, Wilder SP, Venticlef N, Aron-Wisniewsky J, Scott J, et al. Adaptive expression of microRNA-125a in adipose tissue in response to obesity in mice and men. *PLoS ONE.* (2014) 9:e91375. doi: 10.1371/journal.pone.0091375
 64. Meerson A, Traurig M, Ossowski V, Fleming JM, Mullins M, Baier LJ. Human adipose microRNA-221 is upregulated in obesity and affects fat metabolism downstream of leptin and TNF-alpha. *Diabetologia.* (2013) 56:1971–9. doi: 10.1007/s00125-013-2950-9
 65. D'Archivio M, Scacciocchio B, Giammarioli S, Fiani ML, Vari R, Santangelo C, et al. omega3-PUFAs exert anti-inflammatory activity in visceral adipocytes from colorectal cancer patients. *PLoS ONE.* (2013) 8:e77432. doi: 10.1371/journal.pone.0077432
 66. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* (2011) 25:1915–27. doi: 10.1101/gad.17446611
 67. Giroud M, Scheidele M. Long non-coding RNAs in metabolic organs and energy homeostasis. *Int J Mol Sci.* (2017) 18:2578. doi: 10.3390/ijms18122578
 68. Hadjicharalambous MR, Roux BT, Feghali-Bostwick CA, Murray LA, Clarke DL, Lindsay MA. Long non-coding RNAs are central regulators of the IL-1beta-induced inflammatory response in normal and idiopathic pulmonary lung fibroblasts. *Front Immunol.* (2018) 9:2906. doi: 10.1038/s41598-019-42292-w
 69. Gong CY, Zhang HH. Diverse roles of SNHG16 in human cancers. *Clin Chim Acta.* (2019) 503:175–80. doi: 10.1016/j.cca.2019.12.023
 70. Sun F, Liang W, Tang K, Hong M, Qian J. Profiling the lncRNA-miRNA-mRNA ceRNA network to reveal potential crosstalk between inflammatory bowel disease and colorectal cancer. *PeerJ.* (2019) 7:e7451. doi: 10.7717/peerj.7451
 71. Liang J, Zhou W, Sakre N, DeVecchio J, Ferrandon S, Ting AH, et al. Epigenetically regulated miR-1247 functions as a novel tumour suppressor via MYCBP2 in methylator colon cancers. *Br J Cancer.* (2018) 119:1267–77. doi: 10.1038/s41416-018-0249-9
 72. Mazzu YZ, Hu Y, Soni RK, Mojica KM, Qin LX, Agius P, et al. miR-193b-Regulated signaling networks serve as tumor suppressors in liposarcoma and promote adipogenesis in adipose-derived stem cells. *Cancer Res.* (2017) 77:5728–40. doi: 10.1158/0008-5472.CAN-16-2253
 73. Yau MY, Xu L, Huang CL, Wong CM. Long non-coding RNAs in obesity-induced cancer. *Noncoding RNA.* (2018) 4:19. doi: 10.3390/ncrna4030019

74. Liang WC, Fu WM, Wong CW, Wang Y, Wang WM, Hu GX, et al. The lncRNA H19 promotes epithelial to mesenchymal transition by functioning as miRNA sponges in colorectal cancer. *Oncotarget*. (2015) 6:22513–25. doi: 10.18632/oncotarget.4154
75. Liu L, Liu L, Lu S. lncRNA H19 promotes viability and epithelial-mesenchymal transition of lung adenocarcinoma cells by targeting miR-29b-3p and modifying STAT3. *Int J Oncol*. (2019) 54:929–41. doi: 10.3892/ijo.2019.4695
76. Zhou S, Lei D, Bu F, Han H, Zhao S, Wang Y. MicroRNA-29b-3p targets SPARC gene to protect cardiocytes against autophagy and apoptosis in hypoxic-induced H9c2 cells. *J Cardiovasc Transl Res*. (2019) 12:358–65. doi: 10.1007/s12265-018-9858-1
77. Kos K, Wilding JP. SPARC: a key player in the pathologies associated with obesity and diabetes. *Nat Rev Endocrinol*. (2010) 6:225–35. doi: 10.1038/nrendo.2010.18
78. Song H, He P, Shao T, Li Y, Li J, Zhang Y. Long non-coding RNA XIST functions as an oncogene in human colorectal cancer by targeting miR-132-3p. *J BUON*. (2017) 22:696–703.
79. Chen DL, Chen LZ, Lu YX, Zhang DS, Zeng ZL, Pan ZZ, et al. Long noncoding RNA XIST expedites metastasis and modulates epithelial-mesenchymal transition in colorectal cancer. *Cell Death Dis*. (2017) 8:e3011. doi: 10.1038/cddis.2017.421
80. Yang Z, Jiang X, Jiang X, Zhao H. X-inactive-specific transcript: a long noncoding RNA with complex roles in human cancers. *Gene*. (2018) 679:28–35. doi: 10.1016/j.gene.2018.08.071
81. Cheng Z, Luo C, Guo Z. LncRNA-XIST/microRNA-126 sponge mediates cell proliferation and glucose metabolism through the IRS1/PI3K/Akt pathway in glioma. *J Cell Biochem*. (2020) 121:2170–83. doi: 10.1002/jcb.29440
82. Park HY, Kang HS, Im SS. Recent insight into the correlation of SREBP-mediated lipid metabolism and innate immune response. *J Mol Endocrinol*. (2018) 61:R123–31. doi: 10.1530/JME-17-0289
83. Yu Y, Dong JT, He B, Zou YF, Li XS, Xi CH, et al. LncRNA SNHG16 induces the SREBP2 to promote lipogenesis and enhance the progression of pancreatic cancer. *Future Oncol*. (2019) 15:3831–44. doi: 10.2217/fon-2019-0321
84. Christensen LL, True K, Hamilton MP, Nielsen MM, Damas ND, Damgaard CK, et al. SNHG16 is regulated by the Wnt pathway in colorectal cancer and affects genes involved in lipid metabolism. *Mol Oncol*. (2016) 10:1266–82. doi: 10.1016/j.molonc.2016.06.003
85. Zwezdaryk K, Sullivan D, Saifudeen Z. The p53/Adipose-Tissue/Cancer Nexus. *Front Endocrinol*. (2018) 9:457. doi: 10.3389/fendo.2018.00457
86. Krstic J, Reinisch I, Schupp M, Schulz TJ, Prokesch A. p53 functions in adipose tissue metabolism and homeostasis. *Int J Mol Sci*. (2018) 19:2622. doi: 10.3390/ijms19092622
87. Matsui S, Okabayashi K, Tsuruta M, Shigeta K, Seishima R, Ishida T, et al. Interleukin-13 and its signaling pathway is associated with obesity-related colorectal tumorigenesis. *Cancer Sci*. (2019) 110:2156–65. doi: 10.1111/cas.14066
88. Jia H, Liu J, Han B. Reviews of interleukin-37: functions, receptors, and roles in diseases. *Biomed Res Int*. (2018) 2018:3058640. doi: 10.1155/2018/3058640
89. Lee MJ. Transforming growth factor beta superfamily regulation of adipose tissue biology in obesity. *Biochim Biophys Acta Mol Basis Dis*. (2018) 1864:1160–71. doi: 10.1016/j.bbadis.2018.01.025
90. Chen CY, Chen J, He L, Stiles BL. PTEN: tumor suppressor and metabolic regulator. *Front Endocrinol*. (2018) 9:338. doi: 10.3389/fendo.2018.00338
91. Wieser V, Adolph TE, Grander C, Grabherr F, Enrich B, Moser P, et al. Adipose type I interferon signalling protects against metabolic dysfunction. *Gut*. (2018) 67:157–5. doi: 10.1136/gutjnl-2016-313155
92. Kamynina E, Stover PJ. The roles of SUMO in metabolic regulation. *Adv Exp Med Biol*. (2017) 963:143–68. doi: 10.1007/978-3-319-50044-7_9
93. Stern JH, Rutkowski JM, Scherer PE. Adiponectin, leptin, and fatty acids in the maintenance of metabolic homeostasis through adipose tissue crosstalk. *Cell Metab*. (2016) 23:770–84. doi: 10.1016/j.cmet.2016.04.011
94. Quail DF, Dannenberg AJ. The obese adipose tissue microenvironment in cancer development and progression. *Nat Rev Endocrinol*. (2019) 15:139–54. doi: 10.1038/s41574-018-0126-x
95. de Los Santos MC, Dragomir MP, Calin GA. The role of exosomal long non-coding RNAs in cancer drug resistance. *Cancer Drug Resist*. (2019) 2:1178–92. doi: 10.20517/cdr.2019.74

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Tait, Baldassarre, Masotti, Calura, Martini, Vari, Scazzocchio, Gessani and Del Cornò. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Unraveling the Complexity of the Cancer Microenvironment With Multidimensional Genomic and Cytometric Technologies

Natasja L. de Vries^{1,2}, Ahmed Mahfouz^{3,4,5}, Frits Koning² and Noel F. C. C. de Miranda^{1*}

¹ Pathology, Leiden University Medical Center, Leiden, Netherlands, ² Immunohematology and Blood Transfusion, Leiden University Medical Center, Leiden, Netherlands, ³ Human Genetics, Leiden University Medical Center, Leiden, Netherlands, ⁴ Delft Bioinformatics Laboratory, Delft University of Technology, Delft, Netherlands, ⁵ Leiden Computational Biology Center, Leiden University Medical Center, Leiden, Netherlands

OPEN ACCESS

Edited by:

Francesca Finotello,
Innsbruck Medical University, Austria

Reviewed by:

Itai Yanai,
New York University, United States
Christina Stuelten,
National Cancer Institute (NCI),
United States
Pablo G. Camara,
University of Pennsylvania,
United States

*Correspondence:

Noel F. C. C. de Miranda
N.F.de_Miranda@lumc.nl

Specialty section:

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

Received: 01 April 2020

Accepted: 17 June 2020

Published: 23 July 2020

Citation:

de Vries NL, Mahfouz A, Koning F and
de Miranda NFCC (2020) Unraveling
the Complexity of the Cancer
Microenvironment With
Multidimensional Genomic and
Cytometric Technologies.
Front. Oncol. 10:1254.
doi: 10.3389/fonc.2020.01254

Cancers are characterized by extensive heterogeneity that occurs intratumorally, between lesions, and across patients. To study cancer as a complex biological system, multidimensional analyses of the tumor microenvironment are paramount. Single-cell technologies such as flow cytometry, mass cytometry, or single-cell RNA-sequencing have revolutionized our ability to characterize individual cells in great detail and, with that, shed light on the complexity of cancer microenvironments. However, a key limitation of these single-cell technologies is the lack of information on spatial context and multicellular interactions. Investigating spatial contexts of cells requires the incorporation of tissue-based techniques such as multiparameter immunofluorescence, imaging mass cytometry, or *in situ* detection of transcripts. In this Review, we describe the rise of multidimensional single-cell technologies and provide an overview of their strengths and weaknesses. In addition, we discuss the integration of transcriptomic, genomic, epigenomic, proteomic, and spatially-resolved data in the context of human cancers. Lastly, we will deliberate on how the integration of multi-omics data will help to shed light on the complex role of cell types present within the human tumor microenvironment, and how such system-wide approaches may pave the way toward more effective therapies for the treatment of cancer.

Keywords: cancer microenvironment, single-cell, data integration, multi-omics, mass cytometry, spatial analysis, immunophenotyping

INTRODUCTION – HETEROGENEITY OF CANCER AND NEED FOR MULTIDIMENSIONAL APPROACHES

A genetic basis for cancer development was first proposed by the German zoologist Theodor Boveri who speculated that malignant tumors might be the result of abnormal chromosome alterations in cells (1). By then, a cancer cell-centric vision dominated, where tumorigenesis was thought to be exclusively driven by multistep alterations in cellular genomes. During the last decades, however, it has become increasingly apparent that the study of cancers must also encompass other constituents of the cancer microenvironment including immune cells, fibroblasts, and other stromal components, to capture all aspects of cancer biology (2). The immune system, for example,

plays a dichotomous role in cancer development and progression, as different cells can antagonize or promote tumorigenesis (3). The mapping and understanding of the interplay between cancer cells and other constituents of the cancer microenvironment is thus fundamental for the clinical management of this disease.

The study of cancers as complex systems is further complicated by cancer heterogeneity that can occur at different levels; intratumorally, between lesions, and across patients. Intratumoral heterogeneity involves the near-stochastic generation of both genetic (e.g., mutations, chromosomal aberrations) and epigenetic (e.g., DNA methylation, chromatin remodeling, post-translational modification of histones) modifications. Within tumors, distinct niches can favor the outgrowth of different cancer cell clones that acquired characteristics compatible with regional microenvironments (e.g., nutrient and oxygen availability, exposure to immune cells). Other intrinsic sources of heterogeneity such as self-renewal of cancer cells and cell differentiation processes contribute further to intratumoral heterogeneity (4, 5). In addition, the immune system is a major part of the tumor microenvironment and contains many different types of adaptive (e.g., CD4⁺ and CD8⁺ T lymphocytes) and innate (e.g., macrophages and innate lymphoid cells) immune cells that also contribute to cancer heterogeneity (6). Their location within a tumor has been shown to significantly impact their anti- or pro-tumorigenic effects (7). In addition, the density of immune cell infiltration in tumors is a well-known determinant for the prognosis of cancer patients (8). Inter-lesional heterogeneity can be observed between multiple primary tumors at time of diagnosis, between a primary tumor and metastasis, and between different metastatic niches in individual patients. They can be a result of the outgrowth of subclones that can be (epi)genetically distinguished by mutations or structural variations (9). Moreover, the structure and composition of the cancer microenvironment can vary between the primary tumor and metastases. Upon extravasation, cancer cells from primary tumors are exposed to different types of immune cells, stromal cells, platelets, and metabolic stress, and have to adapt to the new tissue microenvironment. As such, the metastatic tissue (“soil”) plays a critical role in regulating the growth of metastases (“seed”) (10). Finally, interpatient heterogeneity is, on top of the aforementioned variables, also fueled by distinct germline genetic backgrounds and environmental and stochastic factors that can affect cancer progression but also immunity.

Major challenges in the field of cancer research are the identification of predictive biomarkers to select patients that are likely to respond to specific treatments, the detection of mechanisms of resistance to therapy, and the development of novel treatments to improve cancer survival. Here, we review the rise of cutting-edge multidimensional technologies such as spectral flow cytometry, multiparameter immunofluorescence, (imaging) mass cytometry, single-cell RNA-sequencing (scRNA-seq), and RNA spatial profiling that may play a crucial role to address the former problems. We will discuss how multi-omics of dissociated cells as well as of spatial data can be obtained (**Figure 1A**) and the importance of integrating them to reveal the full cellular landscape of the cancer microenvironment

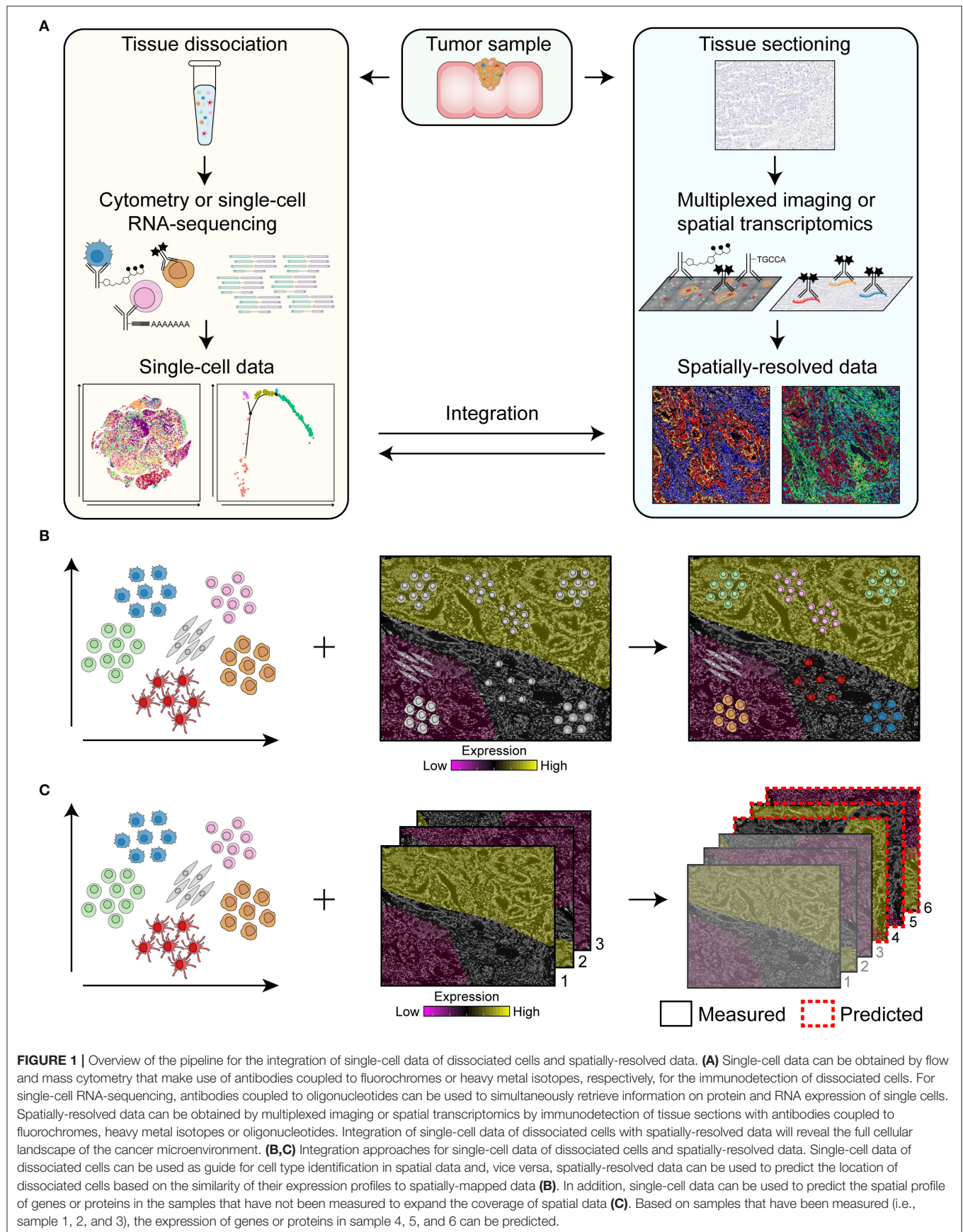
(**Figures 1B,C**). For example, single-cell data of dissociated cells can be used as guide for cell type identification in spatial data (11) and, vice versa, spatial data can be used to predict the location of dissociated cells based on the similarity of their expression profiles to spatially-mapped data (12–14) (**Figure 1B**). In addition, mapping can be used to predict the spatial profile of genes or proteins which have not been experimentally measured to expand the coverage of spatial data (**Figure 1C**) (15–17).

MULTIDIMENSIONAL SINGLE-CELL TECHNOLOGIES AND THEIR STRENGTHS AND WEAKNESSES

Single-Cell DNA- and RNA-Sequencing

Next-generation sequencing (NGS) approaches have revolutionized our ability to generate high-throughput genomic data where individual RNA and DNA molecules are represented by sequencing reads thereby retaining information on genotypes, phenotypes, cellular states, and sub-clonal alterations. Traditional molecular profiling has, until recently, largely relied on the analysis of bulk cell populations. Deep sequencing of DNA and RNA from tissues enabled reconstruction of “average” genomes and “average” transcriptomes that could then be deconstructed by employing bioinformatic algorithms to perform clonal evolution analysis or determine the composition of cancer microenvironments (18–21). For an unbiased and systematic characterization of cells, high-throughput single-cell DNA- and RNA-sequencing have emerged as powerful tools. With single-cell DNA-sequencing, the genomic heterogeneity of tissues can be explored in detail. It can be used to detect nucleotide variations and chromosomal copy number alterations as well as more complex genomic rearrangements and cellular fractions. Single-cell genome sequencing involves whole-genome amplification of single cells, of which the three main methods are MDA (22), MALBAC (23), and DOP-PCR (24). In 2011, the first study of DNA-sequencing of human breast cancer single cells was published (25), which was followed by many single-cell studies charting genetic heterogeneity within individual tumors as well as between primary tumors and their metastases, thereby allowing for a detailed understanding of the evolution processes occurring in a tumor. Single-cell DNA-sequencing has myriad applications in cancer research including examining intratumoral heterogeneity (26–28), investigating clonal evolution during tumorigenic processes (25, 29–32), tracing metastatic dissemination (33), genomic profiling of circulating tumor cells (34–36), measuring mutation rates (37), and gain insight into resistance to therapy (38). By defining, in detail, the genetic composition of tumors, the rationalization of targeted cancer therapies is made possible. However, drawbacks of single-cell DNA-sequencing methods are non-uniform coverage and allelic dropout events as well as artifacts introduced during genomic amplification, all of which contribute to a high rate of false negative and false positive findings (39).

The first single-cell RNA-sequencing (scRNA-seq) experiment was published in 2009 by Tang and colleagues who profiled the



transcriptome of a single cell from early embryonic development (40). Rapid technological advances resulted in an exponential increase in the number of cells that can be studied by scRNA-seq analyses (41). Just 8 years later, 10x Genomics published a scRNA-seq dataset of more than one million individual cells from embryonic mice brains (42). There are many different scRNA-seq library preparation platforms, which can be categorized into plate-based, droplet-based, and microwell-based (41). The selection of the method depends on the research question, the number of input cells, the sequencing depth, the need for full-length coverage of transcriptomes, and costs, among others [reviewed by (43, 44)]. ScRNA-seq has demonstrated to be a powerful technique to decipher cancer biology. In 2012, Ramskold et al. applied scRNA-seq to study circulating tumor cells in melanoma, and could identify potential biomarkers for melanoma as well as SNPs and mutations in this relatively rare circulating tumor cell population (45). Thereafter, scRNA-seq has been used to study the microenvironment of several cancer types including prostate cancer (46), breast cancer (47), glioma (48–50), renal cancer (51), lung cancer (52), melanoma (53–56), colorectal cancer (57–59), pancreatic ductal adenocarcinoma (60), liver cancer (61), head and neck cancer (62), leukemia (63), and glioma (64). A pioneering study that applied scRNA-seq to primary glioblastomas uncovered inherent variability in oncogenic signaling, proliferation, immune responses, and regulators of stemness across cells sorted from five tumors (48). However, this study was restricted to cancer cells and did not further investigate other cell types of the cancer microenvironment. Subsequently, another scRNA-seq study examined distinct genotypic and phenotypic states of malignant, immune, stromal, and endothelial cells of melanomas from 19 patients (53). They identified cell states linked to resistance to targeted therapy, interactions between stromal factors and immune cell abundance, and potential biomarkers for distinguishing dysfunctional and cytotoxic T cells. A recent study in colorectal cancer broadened such scRNA-seq analysis by including a comparison of primary tumors to matched normal mucosa samples (58). By projecting their scRNA-seq data to a large reference panel, the authors identified distinct subtypes of cancer-associated fibroblasts and new expression signatures that were predictive of prognosis in colorectal cancer. Further, scRNA-seq has been applied to investigate changes in the tumor microenvironment of cancer patients treated with immune checkpoint blockade to find signatures associated with positive responses to this therapy (65, 66).

Currently scRNA-seq can be combined with sequencing of T cell receptor and immunoglobulin repertoires thereby allowing to connect information of B- and T cell specificity and phenotype. High-throughput single-cell B cell receptor sequencing of more than 250,000 B cells from different species has recently been pioneered to obtain paired antibody heavy- and light chain information at the single-cell level, and revealed a rapid discovery of antigen-reactive antibody candidates (67). By a novel approach called RAGE-seq (Repertoire and Gene Expression by Sequencing), gene expression profiles can be paired with targeted full-length mRNA transcripts providing BCR and TCR sequences (68). This method has been applied to study cells

from the primary tumor and tumor-associated lymph node of a breast cancer patient and demonstrated the ability to track clonally related lymphocytes across tissues and link TCR and BCR clonotypes with gene expression features (68). A limitation of scRNA-seq is that RNA levels are not fully representative of protein amounts. The advent of CITE-seq, REAP-seq, and Abseq overcame this limitation by enabling simultaneous detection of gene expression and protein levels in single cells by combining oligonucleotide-labeled antibodies against cell surface proteins with transcriptome profiling of thousands of single cells in parallel (69–71). scRNA-seq, when employed in a discovery setting, can inform on the best markers to be used for the study of specific populations by complementary technologies such as flow or mass cytometry. However, aspects of sample preparation and handling have been shown to induce significant alterations in the transcriptome (72). Furthermore, throughput is limited by cost, protocol complexity, available sequencing depth, and dropout events. Together, this can affect the downstream analysis pipeline such as clustering of cell populations and the inference of cellular relationships.

Computational analysis of scRNA-seq data is challenging and involves multiple steps, e.g., quality control, normalization, clustering, and identification of differentially expressed genes and/or trajectory inferences. Multiple unsupervised clustering analyses are available to identify putative cell types, of which graph-based clustering is most widely used (73). For each of these steps, numerous computational tools are available, but in addition software packages have implemented the entire clustering workflow such as Seurat (16), scanpy (74), and SINCERA (75).

Single-Cell Epigenetic Characterization

Although most high-throughput profiling studies to date have focused on DNA, RNA, and protein expression, recent progress in studying the epigenetic regulation of gene expression, at single-cell level, has been made. Over the last decades, methods have been developed including ATAC-seq to measure chromatin accessibility (76), bisulfite sequencing to measure DNA methylation (77), ChIC-sequencing to measure histone modifications (78), and chromosome conformation capture (3C) to analyze the spatial organization of chromatin in a cell (79). Several studies revealed epigenetic programs that regulate T cell differentiation and dysfunction in tumors. Analysis of chromatin accessibility by ATAC-seq revealed that CD8⁺ T cell dysfunction is accompanied with a broad remodeling of the enhancer landscape and transcription factor binding as compared to functional CD8⁺ T cells in tumors (80–83). Also, an increased chromatin accessibility at the enhancer site of the *PDCD1* gene (encoding for PD-1) has been found in the context of dysfunctional CD8⁺ T cells (82). In addition, studies have applied epigenetics to determine mechanisms of resistance to cancer immunotherapies by characterizing chromatin regulators of intratumoral T cell dysfunction before and after PD-1, PD-L1, or CTLA-4 blockade therapy (84, 85). Lastly, DNA hypermethylation may result in the inactivation of genes, such as mismatch repair gene *MLH1* associated with microsatellite instability in colorectal cancer (86). Until recently,

studies on epigenetic modifications depended on correlations between bulk cell populations. Since 2013, with the development of single-cell technologies, epigenomic techniques have been modified for application to single cells to study cell-to-cell variability in for instance chromatin organization in hundreds or thousands of single cells simultaneously. Several single-cell epigenomic techniques have been reported on recently, including measurements of DNA methylation patterns (scRRBS, scBS-seq, scWHBS) (87–89), chromatin accessibility (scATAC-seq) (90), chromosomal conformations (scHi-C) (91), and histone modifications (scChIC-seq) (92). A recent study applied scATAC-seq to characterize chromatin profiles of more than 200,000 single cells in peripheral blood and basal cell carcinoma. By analyzing tumor biopsies before and after PD-1 blockade therapy, Satpathy et al. could identify chromatin regulators of therapy-responsive T cell subsets at the level of individual genes and regulatory DNA elements in single cells (93). Interestingly, variability in histone modification patterns in single cells have also been studied by mass cytometry, which was denominated EpiTOF (94). In this way, Cheung et al. identified a variety of different cell-type and lineage-specific profiles of chromatin marks that could predict the identity of immune cells in humans. Lastly, scATAC-seq has been combined with scRNA-seq and CITE-seq analyses to find distinct and shared molecular mechanisms of leukemia (95). These single-cell strategies will allow to further understand how the epigenome drives differentiation at the single-cell level and unravel drivers of epigenetic states that could be used as target for the treatment of cancer. Additionally, these methods may be used to measure genome structure in single cells to define the 3D structure of the genome. However, for many of these single-cell epigenetic techniques, disadvantages are the low coverage of regulatory regions such as enhancers (scRRBS), low coverage of sequencing reads (scChIP-seq, scATAC-seq), and low sequencing resolution (scHi-C) (96, 97).

Single-Cell Protein Measurements

Flow cytometry has been, in the past decades, the method of choice for high-throughput analysis of protein expression in single cells. The number of markers that can be simultaneously assayed was limited to ~14 markers due to the broad emission spectra of the fluorescent dyes. Recent developments with spectral flow cytometer machines enable the detection of up to 34 markers in a single experiment by measuring the full spectra from each cell, which are unmixed by reference spectra of the fluorescent dyes and the autofluorescence spectrum (98). Fluorescence emission is registered by detectors consisting of avalanche photodiodes instead of photomultiplier tubes used in conventional flow cytometry. A variety of cellular features can be detected by flow cytometry including DNA and RNA content, cell cycle stage, detailed immunophenotypes, apoptotic states, activation of signaling pathways, and others [reviewed by (99)]. This technique has thus been paramount in characterizing cell types, revealing the existence of previously unrecognized cell subsets, and for the isolation of functionally distinct cell subsets for the characterization of tumors. However, the design of multiparameter flow cytometry antibody panels is a challenging

and laborious task, and most flow cytometry studies have so far focused on the in-depth analysis of specific cellular lineages, instead of a broad and system-wide approach.

In 2009, the advent of a new cytometry methodology, mass cytometry (CyTOF, cytometry by time-of-flight), overcame the limitation of spectral overlap by using metal-isotope-conjugated antibodies to detect antigens (100). The metal isotopes attached to each cell are distinguished by mass and quantified in a quadrupole time-of-flight mass spectrometer. A mass cytometer is theoretically capable of detecting over 100 parameters per cell, but current chemical methods limit its use to ~40–50 parameters, simultaneously. Mass cytometry has expanded the breadth of single-cell data in each experiment, making it highly suitable for systems-level analyses such as immunophenotyping of cancer microenvironments. By allowing the examination of large datasets at single-cell resolution, mass cytometry can be applied for the discovery of novel cell subsets as well as for the detection and identification of rare cells. Further advantages of mass cytometry are the irrelevance of autofluorescence, the low biological background as heavy metals are not naturally present in biological systems, and limited signal spillover between heavy metals, thereby reducing the complexity of panel design. Conversely, as compared to flow cytometry, mass cytometry suffers from a higher cell loss during acquisition, is more expensive, and is low-throughput, with a flow rate of up to 500 cells *per sec* as compared to thousands of cells *per sec* in flow cytometry. In addition, cells cannot be sorted for further analysis and forward- and side-scattered light is not detected.

Several studies have applied mass cytometry to further characterize immune cell profiles in peripheral blood or tissues from patients with breast cancer (101), renal cancer (102), melanoma (55, 56, 103–105), lung cancer (52, 106, 107), glioma (49, 50), colorectal cancer (57, 106, 108, 109), liver cancer (61, 110), ovarian cancer (111), and myeloma (112–115), among others. In addition to characterizing immune cell profiles of different tissue types, mass cytometry has also been used to characterize immunophenotypes in tumors and monitor changes during immunotherapy (56, 103–105, 114). In this way, factors that influence response to immunotherapy can be investigated and mechanisms at play during treatment can be characterized. This information can be used to understand and facilitate the identification and classification of responder *vs.* non-responders to cancer immunotherapy. Most of the studies so far have focused on the CTLA-4 and PD-1/PD-L1 axis of cancer immunotherapy, but novel immunotherapeutic targets such as co-inhibitory molecules LAG-3 or TIM-3 or co-stimulatory molecules such as OX40 and GITR are currently being explored in mice models and clinical trials (116). Moreover, mass cytometry has been employed to study antigen-specific T cells with a multiplex MHC class I tetramer staining approach, which has led to the identification of phenotypes associated with tumor antigen-specific T cells (106). Most studies applied mass cytometry for measuring cell surface or intracellular markers, but it can also be used to evaluate cell signaling processes relying on the analysis of protein phosphorylation (117). Altogether, these studies showed that immune responses in cancer are extremely diverse, within tumors from individual patients as well as between patients

with equivalent tumor types. Hence, finding clinically-relevant characteristics based on overall differences can be challenging because of inter-patient variability; differences between cancer patients can be so large that they compromise the discovery of biomarkers.

Because the number of potential phenotypes (resulting from the combination of different markers) increases exponentially with the rise in number of antibodies being measured simultaneously, computational tools for the analysis and visualization of multidimensional data have become key in this field. Traditional workflows for analyzing flow cytometry datasets by manual gating are not efficient to capture the phenotypic differences in mass cytometry and complex flow cytometry data and suffer from individual user bias. In addition, flow and mass cytometry datasets can easily contain millions of cells, illustrating the need for scalable clustering algorithms for efficient analysis. Current single-cell computational tools employed for complex flow cytometry and mass cytometry datasets include unsupervised clustering-based algorithms such as SPADE (118), Phenograph (119), and FlowSOM (120). However, these clustering-based tools do not provide single-cell resolution of the data. On the other hand, non-linear dimensionality reduction-based algorithms such as t-SNE (121) are widely used tools but limited by the number of cells that they can analyze simultaneously, resulting in down-sampled datasets and non-classified cells. Recently, a hierarchical approach of the t-SNE dimensionality-reduction-based technique, HSNE, was described to be scalable to tens of millions of cells (122, 123). In addition, a novel algorithm has recently been implemented in the single-cell analysis field as a dimensionality reduction tool, called uniform manifold approximation and projection (UMAP) (124).

Spatially-Resolved Data

Most of the multidimensional single-cell techniques such as flow cytometry, mass cytometry, and scRNA-seq require cellular dissociation to obtain cell suspensions prior to measuring the individual cells. Different dissociation methods are used, both mechanical and enzymatic, and may result in the loss of certain cell types and affect the expression of specific cell surface markers. Moreover, tissue specimens are often contaminated with blood or other tissues that are processed along with the tissue of interest. As such, not all subsets identified in single-cell data may be representative of the sample of interest. Another key limitation is the lack of information on spatial localization and cellular interactions within a tissue. Analysis of tissue sections by traditional IHC- and immunofluorescence-based methods are useful in providing spatial information (125), but are severely limited in the number of markers that can be measured simultaneously. Recent technological advances have greatly expanded the number of markers that can be captured on tissue slides. For instance, by applying the principles of secondary ion mass spectrometry to image antibodies conjugated to heavy metal isotopes in tissue sections with imaging mass cytometry (IMC) (126) and multiplexed ion beam imaging by time-of-flight (MIBI-TOF) (127). In both imaging approaches, conventional IHC workflows are used but with metal-isotope-conjugated antibodies that are detected through a time-of-flight

mass spectrometer. In IMC, a pulsed laser is used to ablate a tissue section by rasterizing over a selected region of interest. The liberated antibody-bound ions are subsequently introduced into the inductively coupled plasma time-of-flight mass spectrometer. IMC can currently image up to 40 proteins with a subcellular resolution of 1 μm . The principle of MIBI-TOF is similar, but it makes use of a time-of-flight mass spectrometer equipped with a duoplasmatron primary oxygen ion beam rather than a laser. It currently enables simultaneous imaging of 36 proteins at resolutions down to 260 nm (128). Both techniques are, however, low-throughput due to the relatively long imaging time of 2 h per field of 1 mm^2 in IMC and 1 h 12 min per field of 1 mm^2 in MIBI-TOF (129). IMC has been applied to study tumor heterogeneity in several types of cancers, such as pancreatic cancer (130), biliary tract cancer (131), breast cancer (126, 132, 133), and colorectal cancer (108, 134). MIBI-TOF has been used to study the tumor-immune microenvironment of breast cancer (127, 128, 135, 136) and the metabolic state of T cells in colorectal cancer (109). These spatially-resolved, single-cell analyses have great potential to characterize the spatial inter- and intratumoral phenotypic heterogeneity, which can guide cancer diagnosis, prognosis and the selection of treatment. A recent study was able to extend IMC data by integration with genomic characterization of breast tumors and could, in this way, investigate the effect of genomic alterations on multidimensional tumor phenotypes of breast cancer (137).

Other multiplexed imaging techniques such as the Digital Spatial Profiling (DSP) system from NanoString and co-detection by indexing (CODEX) make use of DNA oligonucleotides. In DSP, antibodies or probes are tagged with unique ultraviolet-photocleavable DNA oligos that are released after ultraviolet exposure in specific ROIs and quantified (138). It enables simultaneous detection of up to 40 proteins or over 90 RNA targets from a tissue section and theoretically allows unlimited multiplexing using the NGS readout, but is time-consuming, does not allow for a reconstruction of the image, and has a lower resolution (10 μm) (129). In CODEX, antibodies conjugated to unique oligonucleotide sequences are detected in a cyclic manner by sequential primer extension with fluorescent dye-labeled nucleotides. CODEX currently allows the detection of over 50 markers with an automated fluidic setup platform including a three-color fluorescence microscope (139). Of note, throughput is limited by sequential detection of antibody binding. A disadvantage of CODEX, but also of IMC, is the lack of signal amplification which hampers the detection of lowly abundant antigens. A novel imaging technique, called Immuno-SABER, overcame this limitation by implementing a signal amplification step using primer exchange reactions. Immuno-SABER makes use of multiple DNA-barcoded primary antibodies that are hybridized to orthogonal single-stranded DNA concatemers, generated via primer exchange reactions (140). These primer exchange reactions allow multiplexed signal amplification with rapid exchange cycles of fluorophore-bearing imager strands. The Nanostring DSP platform has been used to study the tumor microenvironment and the outcome of various clinical trials of combination therapy for melanoma (141–144), interactions between macrophages and lymphocytes in metastatic uveal

melanoma (145), immune cell subsets in lung cancer (129, 143), and tumor microenvironments of different metastases in prostate cancer (146). CODEX has been applied to study the immune tumor microenvironment of colorectal cancer patients with 56 protein markers simultaneously (147).

These multiplexed imaging techniques can be applied to snap-frozen as well as FFPE samples that are usually stored in clinical repositories. However, they raise new analysis challenges such as the visualization of 40 markers simultaneously, the image segmentation for cell determination, and algorithms for image-based expression profiles. To understand the tissue architecture, it is necessary to have prior knowledge on which cell types can be present and what their physical relationship to one another could be. Several computational approaches have been developed to enable data analysis of spatially-resolved multiplexed tissue measurements including HistoCAT (148) and ImaCytE (149). These approaches apply cell segmentation masks [using a combination of Ilastik (150) and CellProfiler (151)] to extract single-cell data from each image, which allow for deep characterization using multidimensional reduction tools such as t-SNE combined with the assessment of spatial localization and cellular interactions. In addition to cell-based analysis such imaging technologies also allow the employment of pixel-based analysis that do not depend on cell segmentation.

Integration of single cell transcriptome profiles with their spatial position in tissue context can be achieved by labeling of DNA, RNA, or probes using *in situ* hybridization (ISH). Traditional ISH techniques have been improved to allow the detection of single molecules, named single-molecule fluorescence ISH (smFISH) that can be used to quantitate RNA transcripts at single-cell resolution within a tissue context (152, 153). However, only a small number of genes can be measured simultaneously and a main limitation is the lack of cellular resolution to hundreds of micrometers. To improve the throughput, several highly multiplex methods of *in situ* RNA visualization have been developed such as osmFISH (154), sequential FISH [seqFISH (155) and seqFISH+ (156)] and error-robust FISH [MERFISH (157)]. These allow the subcellular detection of 100–10,000 transcripts simultaneously in single cells *in situ* by using sequential rounds of hybridization with temporal barcodes for each transcript, but require a high number of probes and are time-consuming. Furthermore, ISH can suffer from probe-specific noise due to sequence specificity and background binding. Another approach which may be more applicable for tumors is *in situ* RNA sequencing on tissue sections. STARmap (158) and FISSEQ (159) can profile a few hundreds to thousands of transcripts by using enzymatic amplification methods, but at lower resolution and sensitivity compared to seqFISH and MERFISH. Spatial Transcriptomics (160) and Slide-seq (161) can profile whole transcriptomes by using spatially barcoded oligo-dT microarrays. The spatial transcriptomics method has been used to study and visualize the distribution of mRNAs within tissue sections of breast cancer (160, 162), metastatic melanoma (56, 163), prostate cancer (164), and pancreatic cancer (165). These studies highlight the potential of gene expression profiling of cancer tissue sections to reveal the complex transcriptional landscape in its

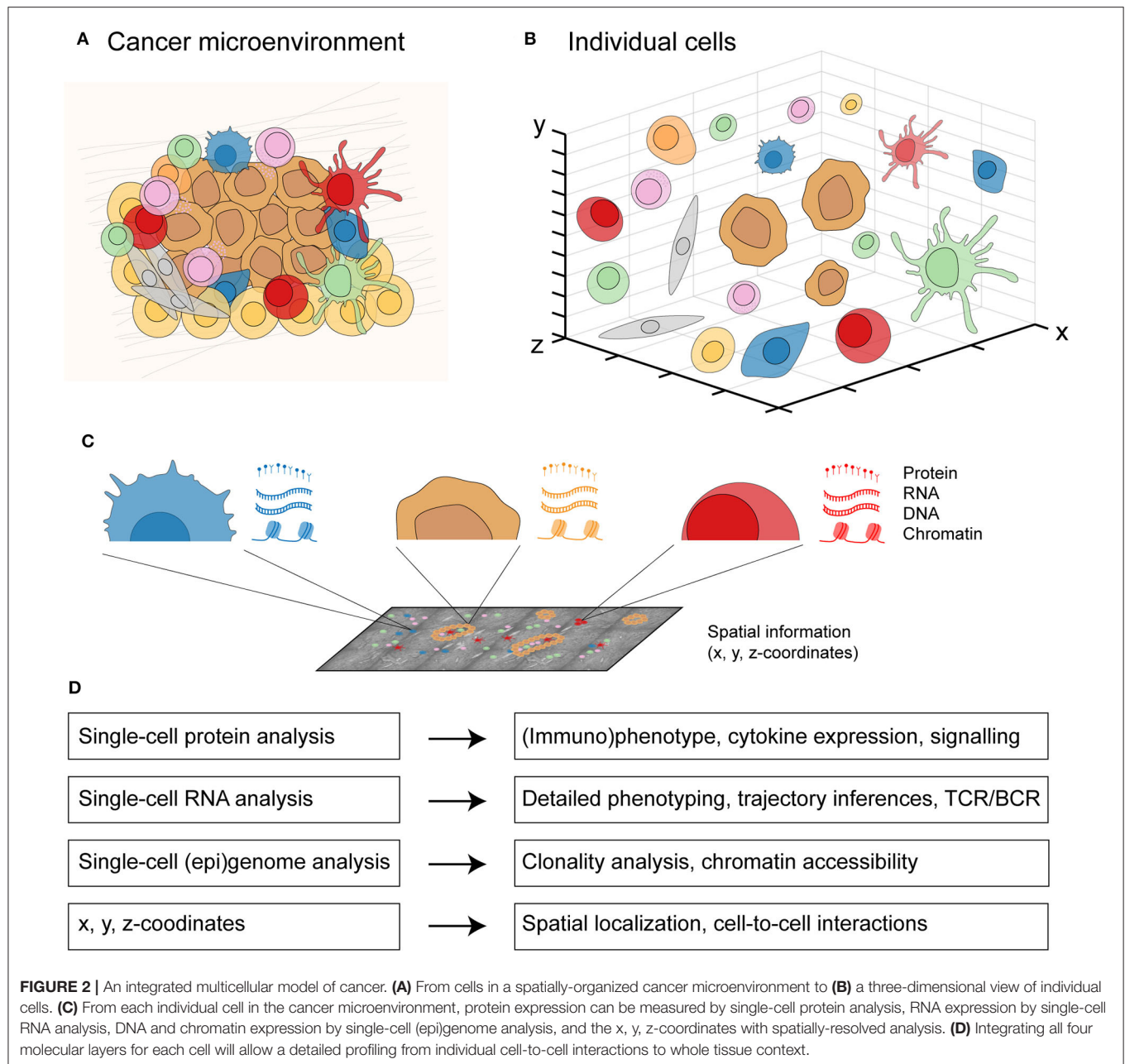
spatial context to gain insight into tumor progression and therapy outcome.

INTEGRATION OF TRANSCRIPTOMIC, (EPI)GENOMIC, PROTEOMIC, AND SPATIALLY-RESOLVED SINGLE-CELL DATA

Traditionally, each type of single-cell data has been considered independently to investigate a biological system. However, cancer is a spatially-organized system composed of many distinct cell types (Figure 2A). These different cell types including immune cells, stromal cells, and malignant cells can be visualized and investigated in an interactive manner (Figure 2B). By applying multi-omics to individual cells in the cancer microenvironment, the molecular landscape of every cell (44) can be defined with its proteome (proteins), transcriptome (RNA sequence), genome (DNA sequence), epigenome (DNA methylation, chromatin accessibility), and spatial localization (x, y, z-coordinates) (Figure 2C). Integrating these different molecular layers for each cell will allow a detailed profiling of cancer as a complex biological system (Figure 2D). Data integration approaches have classically been categorized in three groups: early (concatenation-based), intermediate (transformation-based), and late (model-based) stage integration (166). Early or intermediate stage integration approaches are more powerful than late stage integration since they can capture interactions between different molecular data-types. However, such approaches are also more challenging methodologically given the different data distributions across data types.

A number of studies have used complementary forms of multidimensional analysis on the same sample type in the context of cancer. We have performed a search strategy in PubMed, Web of Science, and Embase databases to find studies that have used mass cytometry in concert with scRNA-seq in the context of human cancer (Supplementary Table 1). An overview of the eight relevant studies that applied mass cytometry together with scRNA-seq to study human cancer and their integration stage is shown in Table 1. In addition, we performed a search strategy in PubMed, Web of Science, and Embase databases on studies that applied single-cell mass cytometry in concert with spatially-resolved data obtained by IMC or MIBI-TOF in human cancer (Supplementary Table 1). An overview of the two relevant studies and their integration stage is shown in Table 2. Notably, all different multidimensional datasets in these studies were analyzed separately and follow a late (model-based) stage integration. Only Goveia and colleagues applied an integration of clustered mass cytometry and scRNA-seq data (107). They merged scaled average gene expression data for each scRNA-seq cluster with scaled average protein expression data for each CyTOF cluster, an approach based on a recently described method from Giordani et al. (167). As they integrated the data only after clustering each modality separately, it is still considered late stage integration.

Integrating multiple single-cell datasets is a challenging task because of the inherently high levels of noise and the large



amount of missing data. Furthermore, the ever-expanding scale of single-cell experiments to millions of cells poses additional challenges. Several methods have been proposed to integrate multimodal single-cell data. State-of-the-art methods focus on embedding both spatial and standard datasets into a latent space using dimensionality reduction, such as Seurat (16), LIGER (17), and Harmony (168), or by employing factor analysis, such as MOFA (169), MOFA+ (170), scMerge (171), and scCoGAPS (172). In addition, a recent study introduced gimVI as a model for integrating spatial transcriptomics data with scRNA-seq data to impute missing gene expression measurements (15). Of note, most of the methods so far follow an intermediate

or late integration approach (166). As such, these methods overcome challenges due to the different data distributions across data types, but they are less powerful in capturing interactions between different molecular data types.

Several methodologies have been developed to simultaneously acquire multiple measurements from the same cell (**Box 1**). Although obtaining simultaneous measurements from the same cell is becoming more feasible, it is still more common to perform subsequent measurements from the same sample (different sets of cells). Integrating spatial-based assays with mRNA or protein expression measurements can be beneficial for several reasons. For instance, spatial measurements are often limited in

TABLE 1 | Overview of studies applying mass cytometry together with single-cell RNA-sequencing to study human cancer heterogeneity.

References	Methods for single-cell profiling	Cancer type	Integration stage
Lavin et al. 2017 (52)	Mass cytometry and scRNA-seq	Lung cancer	Late
De Vries et al. 2019 (57)	Mass cytometry and scRNA-seq	Colorectal cancer	Late
Zhang et al. 2019 (61)	Mass cytometry and scRNA-seq	Liver cancer	Late
Sankowski et al. 2019 (49)	Mass cytometry and scRNA-seq	Glioma	Late
Halaby et al. 2019 (55)	Mass cytometry and scRNA-seq	Melanoma	Late
Goswami et al. 2020 (50)	Mass cytometry and scRNA-seq	Glioblastoma	Late
Goveia et al. 2020 (107)	Mass cytometry and scRNA-seq	Lung cancer	Late
Helmink et al. 2020 (56)	Mass cytometry and scRNA-seq	Melanoma	Late

scRNA-seq, single-cell RNA-sequencing.

TABLE 2 | Overview of studies applying mass cytometry together with imaging mass cytometry or MIBI-TOF to study human cancer heterogeneity.

References	Methods for single-cell profiling	Cancer type	Integration stage
Zhang et al. 2019 (108)	Mass cytometry and IMC	Colon cancer	Late
Hartmann et al. 2020 (109)	Mass cytometry and MIBI-TOF	Colorectal cancer	Late

IMC, imaging mass cytometry; MIBI-TOF, multiplexed ion beam imaging by time-of-flight.

terms of the number of features they can assess simultaneously, although the latest generations of MERFISH and seqFISH(+) can measure around 10,000 transcripts per cell. By integrating these imaging techniques with scRNA-seq, the amount of biologically-relevant information can be enhanced. Moncada et al. presented an integration of scRNA-seq with the spatial transcriptomics method generated from the same sample to study pancreatic cancer (165). A clear challenge when integrating spatial protein (e.g., IMC, MIBI-TOF, CODEX) with scRNA-seq data is the need to model relationships between mRNA and protein expression levels, thus adding an extra layer of complexity. The advent of CITE-seq, combining antibody-based detection of protein markers with transcriptome profiling, could be used to bridge this gap since it allows simultaneous measurement of both mRNA and surface protein marker expression. We foresee an important role for CITE-seq data in the integration of IMC, MIBI-TOF, and CODEX spatial data with scRNA-seq data. Recently, the integration of CITE-seq with CODEX as well as with IMC has been pioneered by Govek et al. (173).

POTENTIAL AVENUES OF HOW THE INTEGRATED DATA WILL HELP TO SHED LIGHT ON THE COMPLEX ROLE OF THE MICROENVIRONMENT IN CANCER

Cancer heterogeneity has long been recognized as a factor complicating the study and treatment of cancer but, until recently, it was difficult to account for in cancer research. The advent of multidimensional single-cell technologies has shed light on the tremendous cellular diversity that exists in cancer tissues and heterogeneity across patients. Moving forward, it will be important to work on the integration of available (spectral) flow cytometry, mass cytometry, scRNA-seq,

and spatially-resolved datasets to investigate commonalities and differences in cellular landscapes between cancer tissues. Multiple flow and mass cytometry datasets can be matched if they include a shared marker set between panels, thereby extending the number of markers per cell and allowing meta-analysis of different mass cytometry datasets with a common core of markers (174). In addition, cell-type references from different single-cell datasets can improve the functional characterization of cells (175). Such a system-wide approach will improve insights into how different components of the cancer microenvironment interact in a tissue context. This requires an extensive collaboration between multi-disciplinary research fields such as oncology, immunology, pathology, and bioinformatics.

Nevertheless, the development and widespread use of innovative methodologies also implies the development of analytical tools for the interpretation of complex datasets and their standardization across laboratories. Furthermore, systems-level analyses challenge a researcher’s capacity to reconnect findings to their biological relevance. Studies should focus on the removal of unwanted variation and experimental noise in high-throughput single-cell technologies as well as the development of cell-type references, such as the Human Cell Atlas (176) and the Allen Brain Atlas (177) principles. We need to further develop algorithms to integrate data from different imaging and non-imaging single-cell technologies. Alternatively, technological developments should allow the acquisition of molecular profiles from single cells without the need of dissociating them from their tissue context. Lastly, it would be of great value to correlate multi-omics techniques with cell-to-cell signaling networks such as CellPhoneDB (178) and NicheNet (179). We expect that this integrated and comprehensive data can be used to create a multicellular model of cancer, from single cells to its tissue context, to understand and exploit cancer heterogeneity for improved precision medicine for cancer patients.

BOX 1 | Methods for the integration of transcriptomic, (epi)genomic, and proteomic single-cell data.

The analysis of protein expression has been extended to include transcript measurements at the single-cell level. CITE-seq (69), REAP-seq (70), and PLAYR (180) can be used to detect mRNA and protein levels simultaneously in single-cells. In CITE-seq and REAP-seq, oligonucleotide-labeled antibodies are used to integrate cellular protein and transcriptome measurements. In PLAYR, mass spectrometry is used to simultaneously analyze the transcriptome and protein expression levels. The analysis of mRNA expression and methylation status in single cells can be achieved by scM&Tseq (181). In addition, mRNA expression and chromatin accessibility of single cells can be analyzed by sci-CAR (182), SNARE-seq (183), and Paired-seq (184). Chromatin organization and DNA methylation from a single nucleus can jointly be profiled by snm3C-seq (185). DR-seq (186) and G&T-seq (187) can assay genomic DNA and mRNA expression simultaneously in single cells, allowing correlations between genomic aberrations and transcriptional levels. Moreover, recent studies have reported on the development of single-cell triple-omics sequencing techniques, such as scTrio-seq (188) and scNMT-seq (189). In scTrio-seq, the transcriptome, genome, and DNA methylome of individual cells are jointly captured. Lastly, scNMT-seq jointly profiles transcription, DNA methylation, and chromatin accessibility, allowing for a thorough investigation of different epigenomic layers with transcriptional status.

How will such system-wide approaches contribute toward more effective therapies for the treatment of cancer? With the advent of targeted therapy and immunotherapy, remarkable advances have been made that changed the management of oncologic treatment for a significant number of patients. However, still only a minority of cancer patients benefit from these therapies, and resistance to treatment remains a major complication in the clinical management of advanced cancer patients. Integrated multi-omics data can help to improve our understanding of the variability in treatment response and resistance mechanisms. By linking detailed molecular and immunological profiles of cells in the cancer microenvironment with sensitivity to specific therapies, potential targets for cancer treatments and associated biomarkers can be identified. This

would also support a rational selection of patients that are most likely to benefit from specific treatments. Furthermore, integrated multi-omics data has the potential to guide the development of alternative therapies, for instance through the identification of resistance mechanisms. We expect that such system-wide approaches, with technologies that include spatial information, will become standard methodologies in cancer research in the coming years.

AUTHOR CONTRIBUTIONS

NV and AM performed the bibliographic research for the manuscript and designed the figures. NV, AM, FK, and NM jointly wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

The authors acknowledge funding from the European Commission under a MSCA-ITN award (675743: ISPIC), the KWF Bas Mulder Award UL (2015-7664), the ZonMw Veni grant (916171144), and the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (grant agreement no. 852832).

ACKNOWLEDGMENTS

We thank J.W. Schoones from Walaeus library of Leiden University Medical Center for his help with developing the literature search strategies and M.E. Ijsselstein from the Department of Pathology of Leiden University Medical Center for providing imaging mass cytometry images of colorectal cancer.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fonc.2020.01254/full#supplementary-material>

REFERENCES

- Boveri T. *Zur Frage Der Entstehung Maligner Tumoren*. Jena: Gustav Fischer (1914).
- Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell*. (2011) 144:646–74. doi: 10.1016/j.cell.2011.02.013
- Schreiber RD, Old LJ, Smyth MJ. Cancer immunoediting: integrating immunity's roles in cancer suppression and promotion. *Science*. (2011) 331:1565–70. doi: 10.1126/science.1203486
- Vladoiu MC, El-Hamamy I, Donovan LK, Farooq H, Holgado BL, Sundaravadanam Y, et al. Childhood cerebellar tumours mirror conserved fetal transcriptional programs. *Nature*. (2019) 572:67–73. doi: 10.1038/s41586-019-1158-7
- Tirosh I, Venteicher AS, Hebert C, Escalante LE, Patel AP, Yizhak K, et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature*. (2016) 539:309–13. doi: 10.1038/nature20123
- Fridman WH, Pages F, Sautes-Fridman C, Galon J. The immune contexture in human tumours: impact on clinical outcome. *Nat Rev Cancer*. (2012) 12:298–306. doi: 10.1038/nrc3245
- Binnewies M, Roberts EW, Kersten K, Chan V, Fearon DF, Merad M, et al. Understanding the tumor immune microenvironment (TIME) for effective therapy. *Nat Med*. (2018) 24:541–50. doi: 10.1038/s41591-018-0014-x
- Fridman WH, Zitvogel L, Sautes-Fridman C, Kroemer G. The immune contexture in cancer prognosis and treatment. *Nat Rev Clin Oncol*. (2017) 14:717–34. doi: 10.1038/nrclinonc.2017.101
- Angelova M, Mlecnik B, Vasaturo A, Bindea G, Fredriksen T, Lafontaine L, et al. Evolution of metastases in space and time under immune selection. *Cell*. (2018) 175:751–65.e16. doi: 10.1016/j.cell.2018.09.018
- Paget S. The distribution of secondary growths in cancer of the breast. *Lancet*. (1889) 133:571–3. doi: 10.1016/S0140-6736(00)49915-0
- Qian X, Harris KD, Hauling T, Nicoloutsopoulos D, Muñoz-Manchado AB, Skene N, et al. Probabilistic cell typing enables fine mapping of closely related cell types *in situ*. *Nat Methods*. (2019) 17:101–6. doi: 10.1038/s41592-019-0631-4
- Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol*. (2015) 33:495–502. doi: 10.1038/nbt.3192

13. Achim K, Pettit J-B, Saraiva LR, Gavriouchkina D, Larsson T, Arendt D, et al. High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nat Biotechnol.* (2015) 33:503–9. doi: 10.1038/nbt.3209
14. Nitzan M, Karaïskos N, Friedman N, Rajewsky N. Gene expression cartography. *Nature.* (2019) 576:132–7. doi: 10.1038/s41586-019-1773-3
15. Lopez R, Nazaret A, Langevin M, Samaran J, Regier J, Jordan MI, et al. A joint model of unpaired data from scRNA-seq and spatial transcriptomics for imputing missing gene expression measurements. *arXiv.* (2019). arXiv:1905.02269.
16. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, et al. Comprehensive integration of single-cell data. *Cell.* (2019) 177:1888–902.e21. doi: 10.1016/j.cell.2019.05.031
17. Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko EZ. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell.* (2019) 177:1873–87.e17. doi: 10.1016/j.cell.2019.05.006
18. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods.* (2015) 12:453–7. doi: 10.1038/nmeth.3337
19. Finotello F, Mayer C, Plattner C, Laschober G, Rieder D, Hackl H, et al. Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. *Genome Med.* (2019) 11:34. doi: 10.1186/s13073-019-0655-5
20. Gerlinger M, Rowan AJ, Horswell S, Math M, Larkin J, Endesfelder D, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med.* (2012) 366:883–92. doi: 10.1056/NEJMoa1113205
21. Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang TH, et al. The immune landscape of cancer. *Immunity.* (2018) 48:812–30.e14. doi: 10.1016/j.immuni.2018.03.023
22. Dean FB, Nelson JR, Giesler TL, Lasken RS. Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res.* (2001) 11:1095–9. doi: 10.1101/gr.180501
23. Zong C, Lu S, Chapman AR, Xie XS. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science.* (2012) 338:1622–6. doi: 10.1126/science.1229164
24. Telenius H, Carter NP, Bebb CE, Nordenskjöld M, Ponder BA, Tunnacliffe A. Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. *Genomics.* (1992) 13:718–25. doi: 10.1016/0888-7543(92)90147-K
25. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, et al. Tumour evolution inferred by single-cell sequencing. *Nature.* (2011) 472:90–4. doi: 10.1038/nature09807
26. Xu X, Hou Y, Yin X, Bao L, Tang A, Song L, et al. Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell.* (2012) 148:886–95. doi: 10.1016/j.cell.2012.02.025
27. Francis JM, Zhang C-Z, Maire CL, Jung J, Manzo VE, Adalsteinsson VA, et al. EGFR variant heterogeneity in glioblastoma resolved through single-nucleus sequencing. *Cancer Discov.* (2014) 4:956–71. doi: 10.1158/2159-8290.CD-13-0879
28. Gawad C, Koh W, Quake SR. Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. *Proc Natl Acad Sci USA.* (2014) 111:17947–52. doi: 10.1073/pnas.1420822111
29. Casasent AK, Schalck A, Gao R, Sei E, Long A, Pangburn W, et al. Multiclonal invasion in breast tumors identified by topographic single cell sequencing. *Cell.* (2018) 172:205–17.e12. doi: 10.1016/j.cell.2017.12.007
30. Yu C, Yu J, Yao X, Wu WKK, Lu Y, Tang S, et al. Discovery of biclonal origin and a novel oncogene SLC12A5 in colon cancer by single-cell sequencing. *Cell Res.* (2014) 24:701–12. doi: 10.1038/cr.2014.43
31. Hughes AEO, Magrini V, Demeter R, Miller CA, Fulton R, Fulton LL, et al. Clonal architecture of secondary acute myeloid leukemia defined by single-cell sequencing. *PLoS Genet.* (2014) 10:e1004462. doi: 10.1371/journal.pgen.1004462
32. Jan M, Snyder TM, Corces-Zimmerman MR, Vyas P, Weissman IL, Quake SR, et al. Clonal evolution of preleukemic hematopoietic stem cells precedes human acute myeloid leukemia. *Sci Transl Med.* (2012) 4:149ra18. doi: 10.1126/scitranslmed.3004315
33. Leung ML, Davis A, Gao R, Casasent A, Wang Y, Sei E, et al. Single-cell DNA sequencing reveals a late-dissemination model in metastatic colorectal cancer. *Genome Res.* (2017) 27:1287–99. doi: 10.1101/gr.209973.116
34. Heitzer E, Auer M, Gasch C, Pichler M, Ulz P, Hoffmann EM, et al. Complex tumor genomes inferred from single circulating tumor cells by array-CGH and next-generation sequencing. *Cancer Res.* (2013) 73:2965–75. doi: 10.1158/0008-5472.CAN-12-4140
35. Lohr JG, Adalsteinsson VA, Cibulskis K, Choudhury AD, Rosenberg M, Cruz-Gordillo P, et al. Whole-exome sequencing of circulating tumor cells provides a window into metastatic prostate cancer. *Nat Biotechnol.* (2014) 32:479–84. doi: 10.1038/nbt.2892
36. Ni X, Zhuo M, Su Z, Duan J, Gao Y, Wang Z, et al. Reproducible copy number variation patterns among single circulating tumor cells of lung cancer patients. *Proc Natl Acad Sci USA.* (2013) 110:21083–8. doi: 10.1158/1538-7445.AM2014-3577
37. Wang Y, Waters J, Leung ML, Unruh A, Roh W, Shi X, et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature.* (2014) 512:155–60. doi: 10.1038/nature13600
38. Suzuki A, Matsushima K, Makinoshima H, Sugano S, Kohno T, Tsuchihara K, et al. Single-cell analysis of lung adenocarcinoma cell lines reveals diverse expression patterns of individual cells invoked by a molecular target drug treatment. *Genome Biol.* (2015) 16:66. doi: 10.1186/s13059-015-0636-y
39. Wang Y, Navin NE. Advances and applications of single-cell sequencing technologies. *Mol Cell.* (2015) 58:598–609. doi: 10.1016/j.molcel.2015.05.005
40. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods.* (2009) 6:377–82. doi: 10.1038/nmeth.1315
41. Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade. *Nat Protoc.* (2018) 13:599–604. doi: 10.1038/nprot.2017.149
42. Genomics X. 10X Genomics Single Cell Gene Expression Datasets. (2017). Available online at: <https://support.10xgenomics.com/single-cell-gene-expression/datasets> (accessed February 27, 2020).
43. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med.* (2018) 50:96. doi: 10.1038/s12276-018-0071-8
44. Wagner A, Regev A, Yosef N. Revealing the vectors of cellular identity with single-cell genomics. *Nat Biotechnol.* (2016) 34:1145–60. doi: 10.1038/nbt.3711
45. Ramsköld D, Luo S, Wang Y-C, Li R, Deng Q, Faridani OR, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol.* (2012) 30:777–82. doi: 10.1038/nbt.2282
46. Miyamoto DT, Zheng Y, Wittner BS, Lee RJ, Zhu H, Broderick KT, et al. RNA-Seq of single prostate CTCs implicates noncanonical Wnt signaling in antiandrogen resistance. *Science.* (2015) 349:1351–6.
47. Brady SW, McQuerry JA, Qiao Y, Piccolo SR, Shrestha G, Jenkins DF, et al. Combating subclonal evolution of resistant cancer phenotypes. *Nat Commun.* (2017) 8:1231. doi: 10.1038/s41467-017-01174-3
48. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science.* (2014) 344:1396–401. doi: 10.1126/science.1254257
49. Sankowski R, Böttcher C, Masuda T, Geirsdottir L, Sagar, Sindram E, et al. Mapping microglia states in the human brain through the integration of high-dimensional techniques. *Nat Neurosci.* (2019) 22:2098–110. doi: 10.1038/s41593-019-0532-y
50. Goswami S, Walle T, Cornish AE, Basu S, Anandhan S, Fernandez I, et al. Immune profiling of human tumors identifies CD73 as a combinatorial target in glioblastoma. *Nat Med.* (2020) 26:39–46. doi: 10.1038/s41591-019-0694-x
51. Kim K-T, Lee HW, Lee H-O, Song HJ, Jeong DE, Shin S, et al. Application of single-cell RNA sequencing in optimizing a combinatorial therapeutic strategy in metastatic renal cell carcinoma. *Genome Biol.* (2016) 17:80. doi: 10.1186/s13059-016-0945-9
52. Lavin Y, Kobayashi S, Leader A, Amir ED, Elefant N, Bigenwald C, et al. Innate immune landscape in early lung adenocarcinoma by paired single-cell analyses. *Cell.* (2017) 169:750–65.e17. doi: 10.1016/j.cell.2017.04.014
53. Tirosh I, Izar B, Prakadan SM, Wadsworth MH, 2nd, Treacy D, Trombetta JJ, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science.* (2016) 352:189–96. doi: 10.1126/science.aad0501

54. Li H, van der Leun AM, Yofe I, Lubling Y, Gelbard-Solodkin D, van Akkooi ACJ, et al. Dysfunctional CD8 T cells form a proliferative, dynamically regulated compartment within human melanoma. *Cell*. (2019) 176:775–89.e18. doi: 10.1016/j.cell.2018.11.043
55. Halaby MJ, Hezaveh K, Lamorte S, Ciudad MT, Kloetgen A, MacLeod BL, et al. GCN2 drives macrophage and MDSC function and immunosuppression in the tumor microenvironment. *Sci Immunol*. (2019) 4:eax8189. doi: 10.1126/sciimmunol.aax8189
56. Helmink BA, Reddy SM, Gao J, Zhang S, Basar R, Thakur R, et al. B cells and tertiary lymphoid structures promote immunotherapy response. *Nature*. (2020) 577:549–55. doi: 10.1038/s41586-019-1922-8
57. de Vries NL, van Unen V, Ijsselstein ME, Abdelaal T, van der Breggen R, Farina Sarasqueta A, et al. High-dimensional cytometric analysis of colorectal cancer reveals novel mediators of antitumour immunity. *Gut*. (2019) 69:691–703. doi: 10.1136/gutjnl-2019-318672
58. Li H, Courtois ET, Sengupta D, Tan Y, Chen KH, Goh JJJ, et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat Genet*. (2017) 49:708–18. doi: 10.1038/ng.3818
59. Zhang Y, Zheng L, Zhang L, Hu X, Ren X, Zhang Z. Deep single-cell RNA sequencing data of individual T cells from treatment-naïve colorectal cancer patients. *Sci Data*. (2019) 6:131. doi: 10.1038/s41597-019-0131-5
60. Elyada E, Bolisetty M, Laise P, Flynn WF, Courtois ET, Burkhart RA, et al. Cross-species single-cell analysis of pancreatic ductal adenocarcinoma reveals antigen-presenting cancer-associated fibroblasts. *Cancer Discov*. (2019) 9:1102–23. doi: 10.1158/2159-8290.CD-19-0094
61. Zhang Q, Lou Y, Yang J, Wang J, Feng J, Zhao Y, et al. Integrated multiomic analysis reveals comprehensive tumour heterogeneity and novel immunophenotypic classification in hepatocellular carcinomas. *Gut*. (2019) 68:2019–31. doi: 10.1136/gutjnl-2019-318912
62. Puram SV, Tirosh I, Parkhi AS, Patel AP, Yizhak K, Gillespie S, et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell*. (2017) 171:1611–24.e24. doi: 10.1016/j.cell.2017.10.044
63. Giustacchini A, Thongjuea S, Barkas N, Woll PS, Povinelli BJ, Booth CAG, et al. Single-cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia. *Nat Med*. (2017) 23:692–702. doi: 10.1038/nm.4336
64. Filbin MG, Tirosh I, Hovestadt V, Shaw ML, Escalante LE, Mathewson ND, et al. Developmental and oncogenic programs in H3K27M gliomas dissected by single-cell RNA-seq. *Science*. (2018) 360:331–5. doi: 10.1126/science.aao4750
65. House IG, Savas P, Lai J, Chen AXY, Oliver AJ, Teo ZL, et al. Macrophage-derived CXCL9 and CXCL10 are required for antitumor immune responses following immune checkpoint blockade. *Clin Cancer Res*. (2020) 26:487–504. doi: 10.1158/1078-0432.CCR-19-1868
66. Jerby-Arnon L, Shah P, Cuoco MS, Rodman C, Su MJ, Melms JC, et al. A cancer cell program promotes T cell exclusion and resistance to checkpoint blockade. *Cell*. (2018) 175:984–97.e24. doi: 10.1016/j.cell.2018.09.006
67. Goldstein LD, Chen YJ, Wu J, Chaudhuri S, Hsiao YC, Schneider K, et al. Massively parallel single-cell B-cell receptor sequencing enables rapid discovery of diverse antigen-reactive antibodies. *Commun Biol*. (2019) 2:304. doi: 10.1038/s42003-019-0551-y
68. Singh M, Al-Eryani G, Carswell S, Ferguson JM, Blackburn J, Barton K, et al. High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes. *Nat Commun*. (2019) 10:3120. doi: 10.1038/s41467-019-11049-4
69. Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods*. (2017) 14:865–8. doi: 10.1038/nmeth.4380
70. Peterson VM, Zhang KX, Kumar N, Wong J, Li L, Wilson DC, et al. Multiplexed quantification of proteins and transcripts in single cells. *Nat Biotechnol*. (2017) 35:936–9. doi: 10.1038/nbt.3973
71. Shahi P, Kim SC, Haliburton JR, Gartner ZJ, Abate AR. Abseq: ultrahigh-throughput single cell protein profiling with droplet microfluidic barcoding. *Sci Rep*. (2017) 7:44447. doi: 10.1038/srep44447
72. van den Brink SC, Sage F, Vértessy Á, Spanjaard B, Peterson-Maduro J, Baron CS, et al. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nat Methods*. (2017) 14:935–6. doi: 10.1038/nmeth.4437
73. Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat Rev Genet*. (2019) 20:273–82. doi: 10.1038/s41576-018-0088-9
74. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol*. (2018) 19:15. doi: 10.1186/s13059-017-1382-0
75. Guo M, Wang H, Potter SS, Whitsett JA, Xu Y. SINCERA: a pipeline for single-cell RNA-seq profiling analysis. *PLoS Comput Biol*. (2015) 11:e1004575. doi: 10.1371/journal.pcbi.1004575
76. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods*. (2013) 10:1213–8. doi: 10.1038/nmeth.2688
77. Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, et al. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci USA*. (1992) 89:1827–31. doi: 10.1073/pnas.89.5.1827
78. Schmid M, Durussel T, Laemmli UK. ChIC and ChEC; genomic mapping of chromatin proteins. *Mol Cell*. (2004) 16:147–57. doi: 10.1016/S1097-2765(04)00540-4
79. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. *Science*. (2002) 295:1306–11. doi: 10.1126/science.1067799
80. Mognol GP, Spreafico R, Wong V, Scott-Browne JP, Togher S, Hoffmann A, et al. Exhaustion-associated regulatory regions in CD8⁺ tumor-infiltrating T cells. *Proc Natl Acad Sci USA*. (2017) 114:E2776–e85. doi: 10.1073/pnas.1620498114
81. Philip M, Fairchild L, Sun L, Horste EL, Camara S, Shakiba M, et al. Chromatin states define tumour-specific T cell dysfunction and reprogramming. *Nature*. (2017) 545:452–6. doi: 10.1038/nature22367
82. Sen DR, Kaminski J, Barnitz RA, Kurachi M, Gerdemann U, Yates KB, et al. The epigenetic landscape of T cell exhaustion. *Science*. (2016) 354:1165–9. doi: 10.1126/science.aae0491
83. Pauken KE, Sammons MA, Odorizzi PM, Manne S, Godec J, Khan O, et al. Epigenetic stability of exhausted T cells limits durability of reinvigoration by PD-1 blockade. *Science*. (2016) 354:1160–5. doi: 10.1126/science.aaf2807
84. Pan D, Kobayashi A, Jiang P, Ferrari de Andrade L, Tay RE, Luoma AM, et al. A major chromatin regulator determines resistance of tumor cells to T cell-mediated killing. *Science*. (2018) 359:770–5. doi: 10.1126/science.aao1710
85. Miao D, Margolis CA, Gao W, Voss MH, Li W, Martini DJ, et al. Genomic correlates of response to immune checkpoint therapies in clear cell renal cell carcinoma. *Science*. (2018) 359:801–6. doi: 10.1126/science.aan5951
86. Deng G, Peng E, Gum J, Terdiman J, Sleisenger M, Kim YS. Methylation of hMLH1 promoter correlates with the gene silencing with a region-specific manner in colorectal cancer. *Br J Cancer*. (2002) 86:574–9. doi: 10.1038/sj.bjc.6600148
87. Guo H, Zhu P, Wu X, Li X, Wen L, Tang F. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Res*. (2013) 23:2126–35. doi: 10.1101/gr.161679.113
88. Smallwood SA, Lee HJ, Angermueller C, Krueger F, Saadeh H, Peat J, et al. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat Methods*. (2014) 11:817–20. doi: 10.1038/nmeth.3035
89. Farlik M, Sheffield NC, Nuzzo A, Datlinger P, Schönegger A, Klughammer J, et al. Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. *Cell Rep*. (2015) 10:1386–97. doi: 10.1016/j.celrep.2015.02.001
90. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*. (2015) 523:486–90. doi: 10.1038/nature14590
91. Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*. (2013) 502:59–64. doi: 10.1038/nature12593
92. Ku WL, Nakamura K, Gao W, Cui K, Hu G, Tang Q, et al. Single-cell chromatin immunocleavage sequencing (scChIC-seq)

- to profile histone modification. *Nat Methods*. (2019) 16:323–5. doi: 10.1038/s41592-019-0361-7
93. Satpathy AT, Granja JM, Yost KE, Qi Y, Meschi F, McDermott GP, et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat Biotechnol*. (2019) 37:925–36. doi: 10.1038/s41587-019-0206-z
 94. Cheung P, Vallania F, Warsinske HC, Donato M, Schaffert S, Chang SE, et al. Single-cell chromatin modification profiling reveals increased epigenetic variations with aging. *Cell*. (2018) 173:1385–97.e14. doi: 10.1016/j.cell.2018.03.079
 95. Granja JM, Klemm S, McGinnis LM, Kathiria AS, Mezger A, Corces MR, et al. Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nat Biotechnol*. (2019) 37:1458–65. doi: 10.1038/s41587-019-0332-7
 96. Lo PK, Zhou Q. Emerging techniques in single-cell epigenomics and their applications to cancer research. *J Clin Genom*. (2018) 1. doi: 10.1017/jcg.1000103
 97. Clark SJ, Lee HJ, Smallwood SA, Kelsey G, Reik W. Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity. *Genome Biol*. (2016) 17:72. doi: 10.1186/s13059-016-0944-x
 98. Futamura K, Sekino M, Hata A, Ikebuchi R, Nakanishi Y, Egawa G, et al. Novel full-spectral flow cytometry with multiple spectrally-adjacent fluorescent proteins and fluorochromes and visualization of *in vivo* cellular movement. *Cytometry A*. (2015) 87:830–42. doi: 10.1002/cyto.a.22725
 99. Irish JM, Doxie DB. High-dimensional single-cell cancer biology. *Curr Top Microbiol Immunol*. (2014) 377:1–21. doi: 10.1007/82_2014_367
 100. Bandura DR, Baranov VI, Ornatsky OI, Antonov A, Kinach R, Lou X, et al. Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Anal Chem*. (2009) 81:6813–22. doi: 10.1021/ac901049w
 101. Wagner J, Rapsomaniki MA, Chevrier S, Anzeneder T, Langwieder C, Dykgers A, et al. A single-cell atlas of the tumor and immune ecosystem of human breast cancer. *Cell*. (2019) 177:1330–45.e18. doi: 10.1016/j.cell.2019.03.005
 102. Chevrier S, Levine JH, Zanotelli VRT, Silina K, Schulz D, Bacac M, et al. An immune atlas of clear cell renal cell carcinoma. *Cell*. (2017) 169:736–49.e18. doi: 10.1016/j.cell.2017.04.016
 103. Wei SC, Levine JH, Cogdill AP, Zhao Y, Anang N-AAS, Andrews MC, et al. Distinct cellular mechanisms underlie anti-CTLA-4 and anti-PD-1 checkpoint blockade. *Cell*. (2017) 170:1120–33.e17. doi: 10.1016/j.cell.2017.07.024
 104. Krieg C, Nowicka M, Guglietta S, Schindler S, Hartmann FJ, Weber LM, et al. High-dimensional single-cell analysis predicts response to anti-PD-1 immunotherapy. *Nat Med*. (2018) 24:144–53. doi: 10.1038/nm.4466
 105. Spitzer MH, Carmi Y, Reticker-Flynn NE, Kwek SS, Madhiredy D, Martins MM, et al. Systemic immunity is required for effective cancer immunotherapy. *Cell*. (2017) 168:487–502.e15. doi: 10.1016/j.cell.2016.12.022
 106. Simoni Y, Becht E, Fehlings M, Loh CY, Koo SL, Teng KWW, et al. Bystander CD8⁺ T cells are abundant and phenotypically distinct in human tumour infiltrates. *Nature*. (2018) 557:575–9. doi: 10.1038/s41586-018-0130-2
 107. Goveia J, Rohlenova K, Taverna F, Treps L, Conradi L-C, Pircher A, et al. An integrated gene expression landscape profiling approach to identify lung tumor endothelial cell heterogeneity and angiogenic candidates. *Cancer Cell*. (2020) 37:21–36.e13. doi: 10.1016/j.ccell.2019.12.001
 108. Zhang T, Lv J, Tan Z, Wang B, Warden AR, Li Y, et al. Immunocyte profiling using single-cell mass cytometry reveals EpCAM⁺ CD4⁺ T cells abnormal in colon cancer. *Front Immunol*. (2019) 10:1571. doi: 10.3389/fimmu.2019.01571
 109. Hartmann FJ, Mrdjen D, McCaffrey E, Glass DR, Greenwald NF, Bharadwaj A, et al. Multiplexed single-cell metabolic profiles organize the spectrum of cytotoxic human T cells. *bioRxiv*. (2020). doi: 10.1101/2020.01.17.909796
 110. Chew V, Lai L, Pan L, Lim CJ, Li J, Ong R, et al. Delineation of an immunosuppressive gradient in hepatocellular carcinoma using high-dimensional proteomic and transcriptomic analyses. *Proc Natl Acad Sci USA*. (2017) 114:E5900–9. doi: 10.1073/pnas.1706559114
 111. Gonzalez VD, Samusik N, Chen TJ, Savig ES, Aghaeepour N, Quigley DA, et al. Commonly occurring cell subsets in high-grade serous ovarian tumors identified by single-cell mass cytometry. *Cell Rep*. (2018) 22:1875–88. doi: 10.1016/j.celrep.2018.01.053
 112. Baughn LB, Sachs Z, Noble-Orcutt KE, Mitra A, Van Ness BG, Linden MA. Phenotypic and functional characterization of a bortezomib-resistant multiple myeloma cell line by flow and mass cytometry. *Leuk Lymphoma*. (2017) 58:1931–40. doi: 10.1080/10428194.2016.1266621
 113. Hansmann L, Blum L, Ju C-H, Liedtke M, Robinson WH, Davis MM. Mass cytometry analysis shows that a novel memory phenotype B cell is expanded in multiple myeloma. *Cancer Immunol Res*. (2015) 3:650–60. doi: 10.1158/2326-6066.CIR-14-0236-T
 114. Adams HC 3rd, Stevenaert F, Krejci J, Van der Borgh K, Smets T, Bald J, et al. High-parameter mass cytometry evaluation of relapsed/refractory multiple myeloma patients treated with daratumumab demonstrates immune modulation as a novel mechanism of action. *Cytometry A*. (2019) 95:279–89. doi: 10.1002/cyto.a.23693
 115. Marsh-Wakefield F, Kruzins A, McGuire HM, Yang S, Bryant C, Fazekas de St. Groth B, et al. Mass cytometry discovers two discrete subsets of CD39⁺Treg which discriminate MGUS from multiple myeloma. *Front Immunol*. (2019) 10:1596. doi: 10.3389/fimmu.2019.01596
 116. Marin-Acevedo JA, Dholaria B, Soyano AE, Knutson KL, Chumsri S, Lou Y. Next generation of immune checkpoint therapy in cancer: new developments and challenges. *J Hematol Oncol*. (2018) 11:39. doi: 10.1186/s13045-018-0582-8
 117. Bodenmiller B, Zunder ER, Finck R, Chen TJ, Savig ES, Bruggner RV, et al. Multiplexed mass cytometry profiling of cellular states perturbed by small-molecule regulators. *Nat Biotechnol*. (2012) 30:858–67. doi: 10.1038/nbt.2317
 118. Qiu P, Simonds EF, Bendall SC, Gibbs KD Jr, Bruggner RV, et al. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat Biotechnol*. (2011) 29:886–91. doi: 10.1038/nbt.1991
 119. Levine JH, Simonds EF, Bendall SC, Davis KL, Amir el AD, Tadmor MD, et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell*. (2015) 162:184–97. doi: 10.1016/j.cell.2015.05.047
 120. Van Gassen S, Callebaut B, Van Helden MJ, Lambrecht BN, Demeester P, Dhaene T, et al. FlowSOM: using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry A*. (2015) 87:636–45. doi: 10.1002/cyto.a.22625
 121. van der Maaten LJP, Hinton GE. Visualizing high-dimensional data using t-SNE. *J Mach Learn Res*. (2008) 9:2579–605.
 122. Pezzotti N, Höllt T, Lelieveldt B, Eisemann E, Vilanova A. Hierarchical stochastic neighbor embedding. *Comput Graph Forum*. (2016) 35:21–30. doi: 10.1111/cgf.12878
 123. van Unen V, Hollt T, Pezzotti N, Li N, Reinders MJT, Eisemann E, et al. Visual analysis of mass cytometry data by hierarchical stochastic neighbour embedding reveals rare cell types. *Nat Commun*. (2017) 8:1740. doi: 10.1038/s41467-017-01689-9
 124. Becht E, McInnes L, Healy J, Dutertre C-A, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol*. (2018). doi: 10.1038/nbt.4314. [Epub ahead of print].
 125. Ijsselsleijn ME, Brouwer TP, Abdulrahman Z, Reidy E, Ramalheiro A, Heeren AM, et al. Cancer immunophenotyping by seven-colour multispectral imaging without tyramide signal amplification. *J Pathol Clin Res*. (2019) 5:3–11. doi: 10.1002/cjp.2.113
 126. Giesen C, Wang HA, Schapiro D, Zivanovic N, Jacobs A, Hattendorf B, et al. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat Methods*. (2014) 11:417–22. doi: 10.1038/nmeth.2869
 127. Angelo M, Bendall SC, Finck R, Hale MB, Hitzman C, Borowsky AD, et al. Multiplexed ion beam imaging of human breast tumors. *Nat Med*. (2014) 20:436–42. doi: 10.1038/nm.3488
 128. Keren L, Bosse M, Thompson S, Risom T, Vijayaragavan K, McCaffrey E, et al. MIBI-TOF: a multiplexed imaging platform relates cellular phenotypes and tissue structure. *Sci Adv*. (2019) 5:eaax5851. doi: 10.1126/sciadv.aax5851
 129. Decalf J, Albert ML, Ziai J. New tools for pathology: a user's review of a highly multiplexed method for *in situ* analysis of protein and RNA expression in tissue. *J Pathol*. (2019) 247:650–61. doi: 10.1002/path.5223

130. Chang Q, Ornatsky OI, Siddiqui I, Loboda A, Baranov VI, Hedley DW. Imaging mass cytometry. *Cytometry A*. (2017) 91:160–9. doi: 10.1002/cyto.a.23053
131. Umemoto K, Togashi Y, Arai Y, Nakamura H, Takahashi S, Tanegashima T, et al. The potential application of PD-1 blockade therapy for early-stage biliary tract cancer. *Int Immunol*. (2019) 32:273–81. doi: 10.1093/intimm/dxz080
132. Jackson HW, Fischer JR, Zanotelli VRT, Ali HR, Mechera R, Soysal SD, et al. The single-cell pathology landscape of breast cancer. *Nature*. (2020) 578:615–20. doi: 10.1038/s41586-019-1876-x
133. Schulz D, Zanotelli VRT, Fischer JR, Schapiro D, Engler S, Lun XK, et al. Simultaneous multiplexed imaging of mRNA and proteins with subcellular resolution in breast cancer tissue samples by mass cytometry. *Cell Syst*. (2018) 6:25–36.e5. doi: 10.1016/j.cels.2017.12.001
134. Ijsselstein ME, van der Breggen R, Farina Sarasqueta A, Koning F, de Miranda NFCC. A 40-marker panel for high dimensional characterization of cancer immune microenvironments by imaging mass cytometry. *Front Immunol*. (2019) 10:2534. doi: 10.3389/fimmu.2019.02534
135. Keren L, Bosse M, Marquez D, Angoshtari R, Jain S, Varma S, et al. A structured tumor-immune microenvironment in triple negative breast cancer revealed by multiplexed ion beam imaging. *Cell*. (2018) 174:1373–87.e19. doi: 10.1016/j.cell.2018.08.039
136. Rost S, Giltman J, Bordeaux JM, Hitzman C, Koeppen H, Liu SD. Multiplexed ion beam imaging analysis for quantitation of protein expression in cancer tissue sections. *Lab Invest J Tech Methods Pathol*. (2017) 97:992–1003. doi: 10.1038/labinvest.2017.50
137. Ali HR, Jackson HW, Zanotelli VRT, Danenberg E, Fischer JR, Bardwell H, et al. Imaging mass cytometry and multiplatform genomics define the phenogenomic landscape of breast cancer. *Nat Cancer*. (2020) 1:163–75. doi: 10.1038/s43018-020-0026-6
138. Merritt CR, Ong GT, Church S, Barker K, Geiss G, Hoang M, et al. High multiplex, digital spatial profiling of proteins and RNA in fixed tissue using genomic detection methods. *Methods Mol Biol*. (2019) 2055:563–83. doi: 10.1101/559021
139. Goltsev Y, Samusik N, Kennedy-Darling J, Bhate S, Hale M, Vazquez G, et al. Deep profiling of mouse splenic architecture with CODEX multiplexed imaging. *Cell*. (2018) 174:968–81.e15. doi: 10.1016/j.cell.2018.07.010
140. Saka SK, Wang Y, Kishi JY, Zhu A, Zeng Y, Xie W, et al. Immuno-SABER enables highly multiplexed and amplified protein imaging in tissues. *Nat Biotechnol*. (2019) 37:1080–90. doi: 10.1038/s41587-019-0207-y
141. Cabrita R, Lauss M, Sanna A, Donia M, Skaarup Larsen M, Mitra S, et al. Tertiary lymphoid structures improve immunotherapy and survival in melanoma. *Nature*. (2020) 577:561–5. doi: 10.1038/s41586-019-1914-8
142. Amaria RN, Reddy SM, Tawbi HA, Davies MA, Ross MI, Glitza IC, et al. Neoadjuvant immune checkpoint blockade in high-risk resectable melanoma. *Nat Med*. (2018) 24:1649–54. doi: 10.1038/s41591-018-0197-1
143. Toki MI, Merritt CR, Wong PF, Smithy JW, Kluger HM, Syrgis KN, et al. High-plex predictive marker discovery for melanoma immunotherapy-treated patients using digital spatial profiling. *Clin Cancer Res*. (2019) 25:5503–12. doi: 10.1158/1078-0432.CCR-19-0104
144. Blank CU, Rozeman EA, Fanchi LF, Sikorska K, van de Wiel B, Kvistborg P, et al. Neoadjuvant versus adjuvant ipilimumab plus nivolumab in macroscopic stage III melanoma. *Nat Med*. (2018) 24:1655–61. doi: 10.1038/s41591-018-0198-0
145. Figueiredo CR, Kalirai H, Sacco JJ, Azevedo RA, Duckworth A, Slupsky JR, et al. Loss of BAP1 expression is associated with an immunosuppressive microenvironment in uveal melanoma, with implications for immunotherapy development. *J Pathol*. (2020) 250:420–39. doi: 10.1002/path.5384
146. Ihle CL, Provera MD, Straign DM, Smith EE, Edgerton SM, Van Bokhoven A, et al. Distinct tumor microenvironments of lytic and blastic bone metastases in prostate cancer patients. *J Immunother Cancer*. (2019) 7:293. doi: 10.1186/s40425-019-0753-3
147. Schürch CM, Bhate S, Barlow GL, Phillips DJ, Noti L, Zlobec I, et al. Coordinated Cellular Neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive front cell. *bioRxiv*. (2019). doi: 10.1101/743989
148. Schapiro D, Jackson HW, Raghuraman S, Fischer JR, Zanotelli VRT, Schulz D, et al. histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data. *Nat Methods*. (2017) 14:873–6. doi: 10.1038/nmeth.4391
149. Somarakis A, Van Unen V, Koning F, Lelieveldt BPF, Holtt T. ImaCytE: visual exploration of cellular microenvironments for imaging mass cytometry data. *IEEE Trans Vis Comput Graph*. (2019). doi: 10.1109/TVCG.2019.2931299. [Epub ahead of print].
150. Sommer C, Straehle C, Köthe U, Hamprecht FA. Ilastik: interactive learning and segmentation toolkit. In: *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. (2011). doi: 10.1109/ISBI.2011.5872394
151. Carpenter AE, Jones TR, Lamprecht MR, Clarke C, Kang IH, Friman O, et al. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol*. (2006) 7:R100. doi: 10.1186/gb-2006-7-10-r100
152. Femino AM, Fay FS, Fogarty K, Singer RH. Visualization of single RNA transcripts *in situ*. *Science*. (1998) 280:585–90. doi: 10.1126/science.280.5363.585
153. Raj A, van den Bogaard P, Rifkin SA, van Oudenaarden A, Tyagi S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat Methods*. (2008) 5:877–9. doi: 10.1038/nmeth.1253
154. Codeluppi S, Borm LE, Zeisel A, La Manno G, van Lunteren JA, Svensson CI, et al. Spatial organization of the somatosensory cortex revealed by osmFISH. *Nat Methods*. (2018) 15:932–5. doi: 10.1038/s41592-018-0175-z
155. Lubeck E, Coskun AF, Zhiyentayev T, Ahmad M, Cai L. Single-cell *in situ* RNA profiling by sequential hybridization. *Nat Methods*. (2014) 11:360–1. doi: 10.1038/nmeth.2892
156. Eng C-HL, Lawson M, Zhu Q, Dries R, Koulina N, Takei Y, et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature*. (2019) 568:235–9. doi: 10.1038/s41586-019-1049-y
157. Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science (New York, NY)*. (2015) 348:aaa6090. doi: 10.1126/science.aaa6090
158. Wang X, Allen WE, Wright MA, Sylvestrak EL, Samusik N, Vesuna S, et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*. (2018) 361:eaat5691. doi: 10.1126/science.aat5691
159. Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Ferrante TC, Terry R, et al. Fluorescent *in situ* sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nat Protoc*. (2015) 10:442–58. doi: 10.1038/nprot.2014.191
160. Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*. (2016) 353:78–82. doi: 10.1126/science.aaf2403
161. Rodriques SG, Stickels RR, Goeva A, Martin CA, Murray E, Vanderburg CR, et al. Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. *Science*. (2019) 363:1463–7. doi: 10.1126/science.aaw1219
162. Vickovic S, Eraslan G, Salmen F, Klughammer J, Stenbeck L, Schapiro D, et al. High-definition spatial transcriptomics for *in situ* tissue profiling. *Nat Methods*. (2019) 16:987–90. doi: 10.1038/s41592-019-0548-y
163. Thrane K, Eriksson H, Maaskola J, Hansson J, Lundeberg J. Spatially resolved transcriptomics enables dissection of genetic heterogeneity in stage III cutaneous malignant melanoma. *Cancer Res*. (2018) 78:5970–9. doi: 10.1158/0008-5472.CAN-18-0747
164. Berglund E, Maaskola J, Schultz N, Friedrich S, Marklund M, Bergenstråhle J, et al. Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nat Commun*. (2018) 9:2419. doi: 10.1038/s41467-018-04724-5
165. Moncada R, Barkley D, Wagner F, Chiodin M, Devlin JC, Baron M, et al. Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nat Biotechnol*. (2020) 38:333–42. doi: 10.1038/s41587-019-0392-8
166. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet*. (2015) 16:85–97. doi: 10.1038/nrg3868
167. Giordani L, He GJ, Negroni E, Sakai H, Law JYC, Siu MM, et al. High-dimensional single-cell cartography reveals novel skeletal

- muscle-resident cell populations. *Mol Cell*. (2019) 74:609–21.e6. doi: 10.1016/j.molcel.2019.02.026
168. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods*. (2019) 16:1289–96. doi: 10.1038/s41592-019-0619-0
 169. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, et al. Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol*. (2018) 14:e8124. doi: 10.15252/msb.20178124
 170. Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, et al. MOFA+: a probabilistic framework for comprehensive integration of structured single-cell data. *Genome Biol*. (2019) 21:111. doi: 10.1101/837104
 171. Lin Y, Ghazanfar S, Wang KYX, Gagnon-Bartsch JA, Lo KK, Su X, et al. scMerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell RNA-seq datasets. *Proc Natl Acad Sci USA*. (2019) 116:9775–84. doi: 10.1073/pnas.1820006116
 172. Stein-O'Brien GL, Clark BS, Sherman T, Zibetti C, Hu Q, Sealfon R, et al. Decomposing cell identity for transfer learning across cellular measurements, platforms, tissues, and species. *Cell Syst*. (2019) 8:395–411.e8. doi: 10.1016/j.cels.2019.04.004
 173. Govek KW, Troisi EC, Miao Z, Woodhouse S, Camara PG. Single-cell transcriptomic analysis of mIHC images via antigen mapping. *bioRxiv*. (2020). doi: 10.1101/672501
 174. Abdelaal T, Holtt T, van Unen V, Lelieveldt BPF, Koning F, Reinders MJT, et al. CyTOFmerge: Integrating mass cytometry data across multiple panels. *Bioinformatics*. (2019) 35:4063–71. doi: 10.1093/bioinformatics/btz180
 175. Gomes T, Teichmann SA, Talavera-López C. Immunology driven by large-scale single-cell sequencing. *Trends Immunol*. (2019) 40:1011–21. doi: 10.1016/j.it.2019.09.004
 176. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, et al. The human cell atlas. *Elife*. (2017) 6:e27041. doi: 10.7554/eLife.27041
 177. Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, et al. Genome-wide atlas of gene expression in the adult mouse brain. *Nature*. (2007) 445:168–76. doi: 10.1038/nature05453
 178. Efremova M, Vento-Tormo M, Teichmann SA, Vento-Tormo R. CellPhoneDB v2.0: Inferring cell-cell communication from combined expression of multi-subunit receptor-ligand complexes. *Nat Protoc*. (2019) 15:1484–506. doi: 10.1038/s41596-020-0292-x
 179. Bonnardel J, T'Jonck W, Gaublot D, Browaeys R, Scott CL, Martens L, et al. Stellate cells, hepatocytes, and endothelial cells imprint the kupffer cell identity on monocytes colonizing the liver macrophage niche. *Immunity*. (2019) 51:638–54.e9. doi: 10.1016/j.immuni.2019.08.017
 180. Frei AP, Bava F-A, Zunder ER, Hsieh EWY, Chen S-Y, Nolan GP, et al. Highly multiplexed simultaneous detection of RNAs and proteins in single cells. *Nat Methods*. (2016) 13:269–75. doi: 10.1038/nmeth.3742
 181. Angermueller C, Clark SJ, Lee HJ, Macaulay IC, Teng MJ, Hu TX, et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat Methods*. (2016) 13:229–32. doi: 10.1038/nmeth.3728
 182. Cao J, Cusanovich DA, Ramani V, Aghamirzaie D, Pliner HA, Hill AJ, et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*. (2018) 361:1380–5. doi: 10.1126/science.aau0730
 183. Chen S, Lake BB, Zhang K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat Biotechnol*. (2019) 37:1452–7. doi: 10.1038/s41587-019-0290-0
 184. Zhu C, Yu M, Huang H, Juric I, Abnoui A, Hu R, et al. An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nat Struct Mol Biol*. (2019) 26:1063–70. doi: 10.1038/s41594-019-0323-x
 185. Lee D-S, Luo C, Zhou J, Chandran S, Rivkin A, Bartlett A, et al. Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. *Nat Methods*. (2019) 16:999–1006. doi: 10.1038/s41592-019-0547-z
 186. Dey SS, Kester L, Spanjaard B, Bienko M, van Oudenaarden A. Integrated genome and transcriptome sequencing of the same cell. *Nat Biotechnol*. (2015) 33:285–9. doi: 10.1038/nbt.3129
 187. Macaulay IC, Haerty W, Kumar P, Li YI, Hu TX, Teng MJ, et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat Methods*. (2015) 12:519–22. doi: 10.1038/nmeth.3370
 188. Hou Y, Guo H, Cao C, Li X, Hu B, Zhu P, et al. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Res*. (2016) 26:304–19. doi: 10.1038/cr.2016.23
 189. Clark SJ, Argelaguet R, Kapourani C-A, Stubbs TM, Lee HJ, Alda-Catalinas C, et al. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat Commun*. (2018) 9:781. doi: 10.1038/s41467-018-03149-4

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor declared a past co-authorship with one of the authors NM.

Copyright © 2020 de Vries, Mahfouz, Koning and de Miranda. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Multi-Omics Characterization of the 4T1 Murine Mammary Gland Tumor Model

Barbara Schrörs¹, Sebastian Boegel², Christian Albrecht¹, Thomas Bukur¹, Valesca Bukur¹, Christoph Holtsträter¹, Christoph Ritzel², Katja Manninen¹, Arbel D. Tadmor¹, Mathias Vormehr^{2,3}, Ugur Sahin^{1,4} and Martin Löwer^{1*}

¹ TRON gGmbH - Translationale Onkologie an der Universitätsmedizin der Johannes Gutenberg-Universität Mainz Gemeinnützige GmbH, Mainz, Germany; ² University Medical Center of the Johannes Gutenberg, University Mainz, Mainz, Germany; ³ BioNTech SE, Mainz, Germany; ⁴ HI-TRON - Helmholtz-Institut für Translationale Onkologie Mainz, Mainz, Germany

OPEN ACCESS

Edited by:

Francesca Finotello,
Innsbruck Medical University, Austria

Reviewed by:

Christina Stuelten,
National Cancer Institute (NCI),
United States

Timothy O'Donnell,
Icahn School of Medicine at
Mount Sinai, United States

*Correspondence:

Martin Löwer
martin.loewer@tron-mainz.de

Specialty section:

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

Received: 29 January 2020

Accepted: 12 June 2020

Published: 23 July 2020

Citation:

Schrörs B, Boegel S, Albrecht C, Bukur T, Bukur V, Holtsträter C, Ritzel C, Manninen K, Tadmor AD, Vormehr M, Sahin U and Löwer M (2020) Multi-Omics Characterization of the 4T1 Murine Mammary Gland Tumor Model. *Front. Oncol.* 10:1195. doi: 10.3389/fonc.2020.01195

Background: Tumor models are critical for our understanding of cancer and the development of cancer therapeutics. The 4T1 murine mammary cancer cell line is one of the most widely used breast cancer models. Here, we present an integrated map of the genome, transcriptome, and immunome of 4T1.

Results: We found Trp53 (Tp53) and Pik3g to be mutated. Other frequently mutated genes in breast cancer, including Brca1 and Brca2, are not mutated. For cancer related genes, Nav3, Cenpf, Muc5Ac, Mpp7, Gas1, MageD2, Dusp1, Ros, Polr2a, Rragd, Ros1, and Hoxa9 are mutated. Markers for cell proliferation like Top2a, Birc5, and Mki67 are highly expressed, so are markers for metastasis like Msln, Ect2, and Plk1, which are known to be overexpressed in triple-negative breast cancer (TNBC). TNBC markers are, compared to a mammary gland control sample, lower (Esr1), comparably low (Erbb2), or not expressed at all (Pgr). We also found testis cancer antigen Pbk as well as colon/gastrointestinal cancer antigens Gpa33 and Epcam to be highly expressed. Major histocompatibility complex (MHC) class I is expressed, while MHC class II is not. We identified 505 single nucleotide variations (SNVs) and 20 insertions and deletions (indels). Neoantigens derived from 22 SNVs and one deletion elicited CD8⁺ or CD4⁺ T cell responses in IFN γ -ELISpot assays. Twelve high-confidence fusion genes were observed. We did not observe significant downregulation of mismatch repair (MMR) genes or SNVs/indels impairing their function, providing evidence for 6-thioguanine resistance. Effects of the integration of the murine mammary tumor virus were observed at the genome and transcriptome level.

Conclusions: 4T1 cells share substantial molecular features with human TNBC. As 4T1 is a common model for metastatic tumors, our data supports the rational design of mode-of-action studies for pre-clinical evaluation of targeted immunotherapies.

Keywords: immunotherapy, cancer models, computational immunology, triple negative breast cancer, 4T1 murine mammary gland tumor cell line

INTRODUCTION

The translational value of pre-clinical cancer studies is dependent on the availability of model systems that mimic the situation in the patient. The murine mammary carcinoma cell line 4T1 is widely used as syngeneic tumor model for human breast cancer [e.g., (1–3)], a tumor entity with the world-wide highest incidence¹. This cell line was originally derived from a subpopulation of a spontaneously arising mammary tumor of a mouse mammary tumor virus (MMTV) positive BALB/c mouse foster nursed on a C3H mother (BALB/BfC3H) (4, 5). 4T1 can easily be transplanted into the mammary gland and was already described as poorly immunogenic, highly tumorigenic, invasive, and spontaneously metastasizing to distant organs (6). Thus, the location of the primary tumor and its metastatic spreading closely resemble the clinical course in patients. Moreover, 4T1 cells are used to specifically investigate triple-negative breast cancer (TNBC) [e.g., (7–9)] lacking protein expression of estrogen receptor (ER), progesterone receptor (PgR), and epidermal growth factor receptor 2 (ErbB2) (10). This triple-negative phenotype is estimated for more than 17% of breast cancers that are annually diagnosed (11).

In spite of being such a widely used system, until now mainly phenotypic characteristics of 4T1 have been compared to human (triple-negative) breast cancer in the literature, while no comprehensive genomic, transcriptomic, and immunomic overview has been provided that would complement the evaluation of 4T1 as adequate breast cancer or even TNBC model. In our study, we examined the 4T1 cell line from a multi-omic point of view to complete the picture.

MATERIALS AND METHODS

Samples

BALB/cJ mice (Charles River) were kept in accordance with legal and ethical policies on animal research. The animal study was reviewed and approved by the federal authorities of Rhineland-Palatinate, Germany and all mice were kept in accordance with federal and state policies on animal research at the University of Mainz and BioNTech SE. Germline BALB/cJ DNA was extracted from mouse tail. 4T1 WT cells were purchased from ATCC. Third and 4th passages of cells were used for tumor experiments.

Data

ENCODE RNA Sequencing data of adult BALB/c mammary gland tissue for differential expression analysis against 4T1 expression profiles was downloaded from the UCSC Genome Browser (12) repository:

- URL: <http://hgdownload.cse.ucsc.edu/goldenPath/mm9/encodeDCC/wgEncodeCshlLongRnaSeq>
- Files:
wgEncodeCshlLongRnaSeqMamgAdult8wksFastqRd1Rep1.fastq.tgz
wgEncodeCshlLongRnaSeqMamgAdult8wksFastqRd1Rep2.fastq.tgz

¹<http://gco.iarc.fr/today/data/factsheets/populations/900-world-fact-sheets.pdf>

wgEncodeCshlLongRnaSeqMamgAdult8wksFastqRd2Rep1.fastq.tgz
wgEncodeCshlLongRnaSeqMamgAdult8wksFastqRd2Rep2.fastq.tgz

Female BALB/c RNA-Seq data sets for the comparison of the MHC expression were described before (13) and are available in the European Nucleotide Archive (see Data Availability Statement).

High-Throughput Sequencing and Read Alignment

Exome capture from 4T1 and BALB/cJ mice were sequenced in duplicate using the Agilent Sure-Select solution-based mouse protein coding exome capture assay. 4T1 oligo(dT)-isolated RNA for gene expression profiling was prepared in duplicate. Libraries were sequenced on an Illumina HiSeq2500 (2 × 50 nt). DNA-derived sequence reads were aligned to the mm9 genome using bwa [(14); default options, version 0.5.9_r16]. Ambiguous reads mapping to multiple locations of the genome were removed. RNA-derived sequence reads were aligned to the mm9 genome using STAR [(15); default options, version 2.1.4a]. The sequencing reads are available in the European Nucleotide Archive (see Data Availability Statement).

Mutation Detection

Somatic SNV and short insertion/deletion (indel) calling was performed using Strelka [(16); default options for whole exome sequencing, version 2.0.14] on each cell line or normal library replicate pair individually. The individual analysis runs resulted in 1,115 and 1,108 SNV candidates, with an overlap of 886 SNVs (66%) and in 60 and 58 indel candidates, with an overlap of 50 (74%).

Transcriptome Profiling

Transcript abundance estimation was done with kallisto [(17); default options, version 0.42.4] on each cell line or normal sample library replicate individually using the mean transcripts per million (TPM) per transcript final value. Differential expression analysis was performed with edgeR [(18); default options, version 3.26.8] using the reported transcript counts of kallisto, summarized by adding up the counts of the respective transcripts associated with each gene. The TPM values of the technical replicates have a Pearson's correlation coefficient of more than 0.99. Enriched pathways (KEGG 2019 Mouse²) and gene ontologies (GO Biological Process 2018³) in differentially up- or downregulated genes were determined using Enrichr (19). The associated Enrichr libraries were used as background lists for comparison with enrichment analysis in TNBC subtypes (20).

²https://amp.pharm.mssm.edu/Enrichr/geneSetLibrary?mode=text&libraryName=KEGG_2019_Mouse

³https://amp.pharm.mssm.edu/Enrichr/geneSetLibrary?mode=text&libraryName=GO_Biological_Process_2018

Data from human TNBC studies (20–22) was obtained from the respective journal websites^{4,5,6}. Data for mapping human and mouse gene symbols was obtained from the Jackson Laboratory⁷. TNBC and breast tissue short read data in fastq format was obtained from the short read archive (TNBC: accession number PRJNA607061, sample accession numbers are documented in Table S7).

TCGA BRCA expression values for ERBB2, ESR1, and PGR was obtained from the UCSC Xena browser (<http://xena.ucsc.edu>), using the “HTSeq FPKM-UQ” dataset. The clinical annotation including immunohistochemistry results was downloaded from the GDC Legacy site⁸. These tables were merged using the patient barcodes keeping only patients with non-missing and non-inconclusive results for the immunohistochemistry status of “Her2”, “Pr”, and “Esr”. This resulted in 808 data points. Principal component analysis was done in R with the “prcomp” function.

Fusion Gene Detection

Fusion genes were detected with an in-house pipeline: We employed the “wisdom of crowds” approach (23), and applied four fusion detection tools, SOAPFuse, MapSplice2, InFusion and STARFusion (23–26) to two technical replicates of the 4T1 cell line. We used Ensembl GRCm38.95 as reference. SOAPFuse and STARFusion were run with default parameters, MapSplice2 was run with “–qual-scale phred33 –bam –seglen 20 –min-map-len 40” as additional parameters, and InFusion was run with “–skip-finished –min-unique-alignment-rate 0 –min-unique-split-reads 0 –allow-non-coding” as additional parameters. For run time improvement, we did a first manual pass of a STAR alignment to the mm10 reference genome and retained only non-matching and chimeric reads for further processing by the four fusion detection tools. In order to combine the eight resulting datasets (four tools applied to two replicates) we first created the union of results of all four tools for each replicate, followed by the intersection of both independent runs (one per replicate cell line RNA library). This was considered as high confidence result set.

DNA Copy Number Calling

Absolute copy numbers were detected from exome capture data using Control-FREEC [(27), version 11.5]. Control-FREEC was run multiple times with different ploidy input parameters (ploidy = x for values of $x = 2, 3, 4$, or 5) on the merged alignment files (merged with the “merge” command from samtools). In addition, the following non-default parameters were set: forceGCcontentNormalization = 1, intercept = 0, minCNAlength = 3, sex = XX, step = 0, uniqueMatch = TRUE, contaminationAdjustment = FALSE.

⁴<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4911051/bin/pone.0157368.s007.xlsx>

⁵<http://downloads.hindawi.com/journals/bmri/2018/2760918.f1.docx>

⁶https://static-content.springer.com/esm/art%3A10.1186%2Fs13058-016-0690-8/MediaObjects/13058_2016_690_MOESM1_ESM.docx

⁷http://www.informatics.jax.org/downloads/reports/HOM_MouseHumanSequence.rpt

⁸file URL: <https://portal.gdc.cancer.gov/legacy-archive/files/735bc5ff-86d1-421a-8693-6e6f92055563>

The CNV calls were processed with custom Python and R scripts: The output segment copy numbers were assigned to gene symbols by intersection with gene coordinates. Using the gene symbols, the previously detected SNVs were mapped to the copy numbers. Computed variant allele frequencies (VAF) from read alignments were then compared to the expected allele frequency distribution based on discrete copy numbers. For e.g., for a copy number of 3 (as predicted by Control-FREEC), one would expect SNV VAFs in associated genes clustered around values of 0.33 (one allele mutated), 0.66 (two alleles mutated), and 1 (three alleles mutated). The best match was manually determined for a Control-FREEC ploidy value of 5.

Transcript Assembly

RNA-Seq transcript assembly was done using trinity [(28); default options, version r20140413p1]. Assembled transcript contigs were mapped to human transcript sequences and the MMTV genome (GenBank accession number NC_001503.1) with blat (29).

MHC Typing

MHC type of the 4T1 cells was determined from RNA-Seq reads as described in Castle et al. (13).

MHC Expression

MHC expression was quantified using Sailfish [(30); default options, version 0.6.2] on an mm9 transcriptome index which represents C57BL/6 mice, combined with the expected BALB/cJ MHC sequences.

Mutation Signatures

Mutation signatures (31) were computed with the R package YAPSA (default settings, version 1.4.0).

Expression Profiling of Viral Genes

Virus genomes were downloaded in FASTA format from the NCBI Virus Genomes resource (32). Sequence reads were aligned using STAR [(15); version 2.5] to a combined reference genome containing murine genome sequences (mm9) and 7,807 virus genomes. We used a maximum mismatch ratio of 0.2, reporting ambiguous alignments only when the alignment scores matched the best alignment of the read.

For each of the virus accession numbers, the GenBank features “mRNA” and “CDS” were extracted from NCBI sources to create a virus gene database for expression analysis. Taxonomic information was extracted for filtering closely related viruses with lower read counts.

Viral gene expression was calculated using the built virus gene database and an in-house software as previously described (33). Any read overlapping a union model of all of a gene’s isoforms was counted. All read counts were normalized to reads per kilobase of gene model per million mapped reads (RPKM) for all murine and viral genes.

Neoantigen Selection for Immunogenicity Testing

The selection for the initial immunogenicity assessment was described earlier (34). For the subsequent testing of 11 additional

4T1-WT SNVs, the following more strict criteria were applied: (i) present in both replicates, (ii) hitting a transcript outside the untranslated region (UTR), (iii) resulting in a non-synonymous amino acid exchange (no stop gain or loss), (iv) mean expression in replicates > 0 , (v) VAF in 4T1 DNA > 0 , (vi) VAF in 4T1 RNA > 0.1 , and (vii) VAF in RNA of an independent control mammary gland sample was 0. Indels were selected accordingly, but with a less stringent filter on the variant allele frequency in the tumor RNA (VAF_in_RNA > 0). Indels were subjected to confirmation via Sanger sequencing [performed as in (34)] which left two of the three pre-filtered indels for further experiments.

Immunogenicity Testing

The immunogenicity assessment of SNV-derived neoantigens was performed as described earlier (34). For the testing of indel-derived mutated peptides, mice ($n = 3$) were vaccinated with repetitive intravenous injections of 40 μ g RNA lipoplexes (35) on days 0, 7, and 14. Five days after the last immunization, splenocytes of mice were tested for recognition of 15-mer peptides spanning the complete mutated sequence (11 amino acid overlap). T-cell responses were measured via IFN- γ enzyme-linked immunospot assay (ELISpot) as previously described (34). In brief, 5×10^5 splenocytes were stimulated overnight by addition of 2 μ g/mL peptide at 37°C in anti-IFN- γ (10 μ g/mL, clone AN18, Mabtech) coated Multiscreen 96-well plates (Millipore) and cytokine secretion was detected with an anti-IFN- γ antibody (1 μ g/mL, clone R4-6A2, Mabtech). For subtyping of T-cell responses, CD8 $^+$ T cells were isolated from splenocytes via magnetic-bead based cell separation [Miltenyi Biotech, CD8a (Ly-2) MicroBeads] according to the manufacturer's recommendations. CD8 $^+$ T cell-depleted splenocytes served as a source for CD4 $^+$ T cells. 1.5×10^5 isolated CD8 $^+$ T cells and 5×10^5 cells derived from the CD4 $^+$ T cell containing flow-through were restimulated in an IFN- γ ELISpot as described above. 1×10^5 syngeneic bone marrow derived-dendritic cells (34) served as antigen-presenting cells for CD8 $^+$ T cells.

RESULTS

The 4T1 Tumor Genome

Using whole exome and RNA-Seq data, we assessed genomic variation patterns by comparing 4T1 to BALB/c DNA, examining copy number aberrations, indels, SNVs, and gene fusions. Moreover, we determined absolute DNA copy numbers.

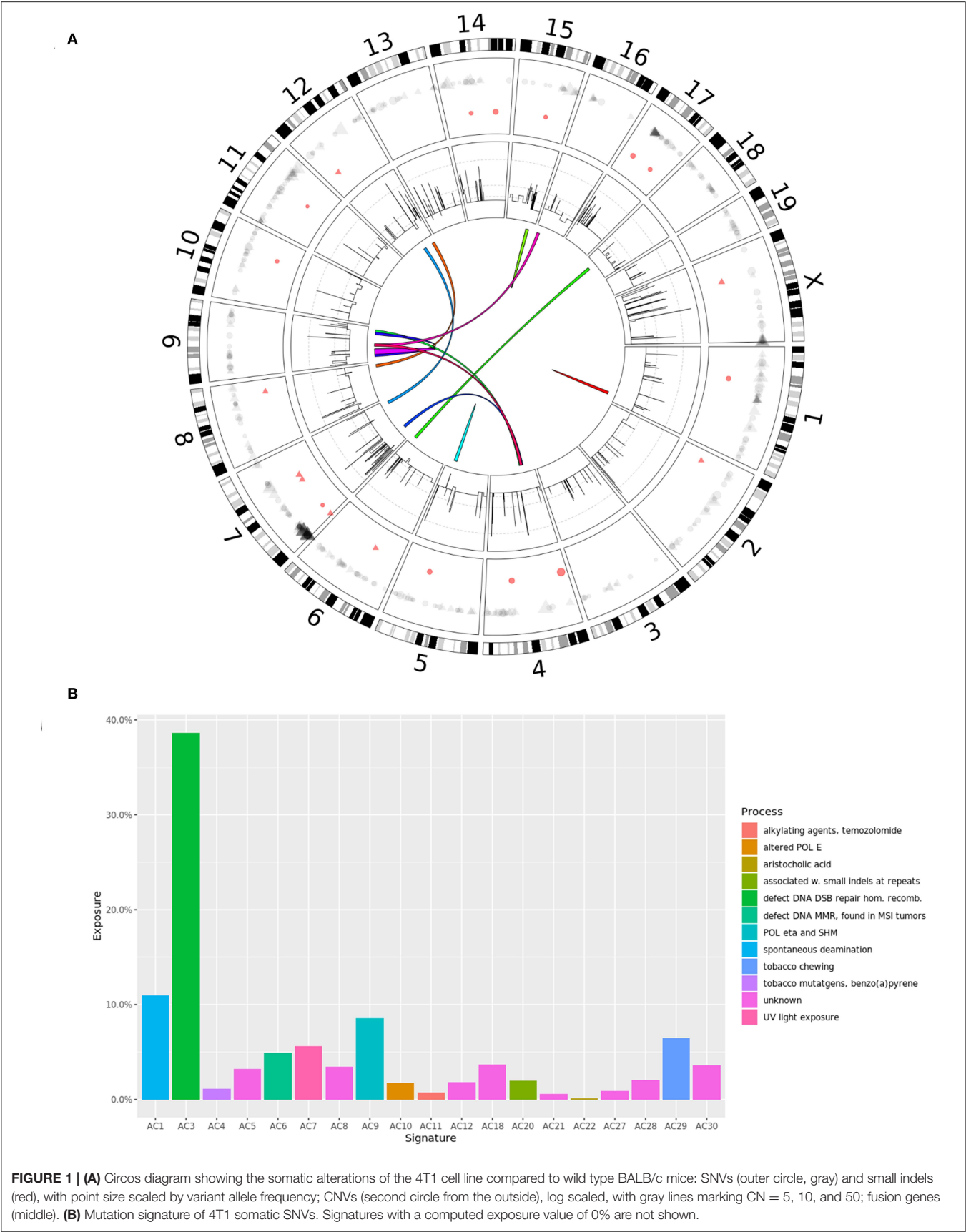
No reads mapped to Y chromosome (DNA or RNA), which is expected as 4T1 originated from a female mouse. The analysis of the copy number profile revealed a median gene copy number of four, suggesting a tetraploid genome, although a sizable fraction of the genome seemed to be present in five copies (Figure 1A, second circle from the outside; Table S1). The findings were confirmed by a good agreement between the observed SNV allele frequencies and the allele frequency profile expected by the predicted gene copy number (e.g., for a copy number of four we expected SNV VAFs to be clustered around the values of 0.25, 0.5, 0.75, and 1). We observed known breast cancer oncogenes Akt1 and Sf3b1 (36) with focal amplifications (copy number six

and seven, respectively), while pan-cancer oncogene Myc had a copy number of 11 (Table S1). Several known human tumor suppressor genes had a predicted copy number of less than four, with a possible functional impact (Table S1).

We identified 505 SNVs (Table S2, Figure 1A, outer circle, gray) and 20 short indels (Table S3, Figure 1A, outer circle, red) in transcripts, as well as 12 fusion events (Table S4, Figure 1A, middle). The majority of SNVs caused non-synonymous protein changes outside UTRs (264; 52%) including 248 missense and 16 non-sense variations (15 premature stops and one stop loss). Relative to the mouse genome (32 million protein-coding nucleotides), the 4T1 variation rate was 1.1 mutations per MB, which is within the range observed for human breast cancer (31). This number is an order of magnitude lower compared to the murine colon cancer model CT26, which suggests that CT26 is more likely to encode immunogenic epitopes than 4T1. The observed difference in the mutational load was in agreement with previous studies (37, 38), even though we detected a higher number of somatic mutations in both tumor models. We confirmed 45 of 47 (96%) and 193 of 246 (78%) previously reported SNVs in our data. Of the 264 non-synonymous SNVs, we found 91 (34%) mutations to be expressed (VAF > 0), which is comparable with a study in human TBNC that found ~36% of mutations to be expressed (39). We have recently shown a high correlation between the DNA and RNA mutation allele frequencies in three murine tumor models (including 4T1) (13). Here, using updated methods for transcript quantification and mutation calling, we were able to reproduce these results ($R^2 = 0.98$, Figure S1), thus further corroborating that genes are equally transcribed from all alleles, mutated and wild-type (WT), in proportion to their DNA allele frequency.

Examining the mutational landscape in the 4T1 exome (Figure 1B), we found a higher prevalence of C>T, C>G, and C>A SNVs (Figure S2), which is in concordance with the somatic mutational signatures in human breast cancers (40). Interestingly, we found an overrepresentation of C>T transitions at XCG triplets (Figure S2; C is the mutated base, preceded by any nucleotide and followed by G), which is a known mutational mechanism due to deamination of methylated cytosines to thymine and has been observed in human breast cancers (41). C>T transitions showed the largest contribution to the mutational signatures in 4T1 and has been attributed to the activity of the APOBEC family of cytidine deaminases (42). Of note, Apobec3 has been found to provide partial protection in mice against infection with the oncogenic retrovirus MMTV (43), suggesting activation of this gene during MMTV infection and genome integration with subsequent cytosine deamination, resulting in the observed mutation pattern. The mutational signatures revealed a strong signal for signature AC3 (Figure 1B), which is associated with breast cancer and colloquially called “BRCAness,” followed by signature AC1, which is associated with spontaneous deamination. In contrast, signature AC2 was not found at all (and therefore not shown in Figure 1B), which would further strengthen the potential connection to APOBEC cytidine deaminases, as described above.

Of the most frequently mutated genes recently identified in breast cancer in general (41) and TNBC in particular (39) (Tp53,



Pik3ca, Myc, Ccnd1, Pten, Erbb2, Znf703/Fgfr1 locus, Gata3, RB1, and Map3k1, Egfr), we only identified mutations in Trp53 (frameshift insertion of “A”) and Pik3cg (synonymous SNV) which is the catalytic subunit of class I PI3 kinases (similar to Pik3ca). In addition, we did not find mutations in breast cancer susceptibility genes Brca1 and Brca2. Further mutations in cancer-related genes included Nav3 (V1129L), Cenpf (D1327E), Muc5ac (A429P), Mpp7 (Q158R), Gas1 (G326R), Maged2 (A473S), Dusp1 (C24R), Ros1 (W1875C), Polr2a (M1102I), Rragd (L385P), and Hoxa9 (insertion of “G” in UTR). Variations in immune-relevant genes included Tlr8 (R613H), Tlr9 (N332K), and Lilrb3 (S91R).

Using RNA-Seq data of 4T1 replicates, we identified 12 fusion events (Table S4), including a fusion of Siva1 and Gas8, one regulating cell cycle progression/proliferation and apoptosis, the other being a putative tumor suppressor gene. None of them have been reported before in breast cancer (44, 45).

MMTV Integration

MMTV is a milk-transmitted retrovirus that is oncogenic through integration into the host genome, thereby activating the expression of nearby genes (46). Multiple common insertion sites (CIS) have been identified and associated with candidate oncogenes and pathways involved in mammary tumorigenesis, including the Wnt and Fgf clusters (47, 48). A subset of CIS was significantly correlated with overexpression and deregulation of candidate oncogenes (49). We collected a set of 54 candidate genes for MMTV integration and compared their expression in 4T1 cells to that in normal mammary gland (Figure S3). About 68.5% of these genes showed significant down- or upregulation, while only about 42% of all genes of 4T1 cells were differentially expressed, suggesting MMTV integration as a possible cause. However, many of the 54 candidate genes are involved in oncogenic pathways, so it is not clear if the observed differential expression are caused by the integration, potentially dysregulating a pathway or effect of the dysregulated pathway in the first place.

Moreover, we had direct evidence from RNA-Seq based transcriptome assembly of an integration site 5' to the Fgfr2 gene (Figure S4). A CIS near Fgfr2 was associated with an increased copy number and overexpression of Fibroblast growth factor receptor 2 (Fgfr2) (47). While we just observed a copy number of four, three of eleven isoforms were significantly overexpressed in 4T1. Fgfr2 is a transmembrane tyrosine kinase receptor and its activation triggers a complex signal transduction network (via e.g., Ras-Raf-Mapk or Pik3-Akt pathway), which leads to transcription of genes involved in angiogenesis, cell migration, proliferation, differentiation and survival. There is evidence of deregulated activation of FGFR signaling in the pathogenesis of human cancers (46). FGFR2 amplifications have been found in 10% of gastric cancers (50) and were also found in a subset of human TNBC patients (39, 51); FGFR2 amplifications are estimated to occur in ~4% of TNBC samples, resulting in constitutive activation of FGFR2 (52). Increased expression of this gene is associated with poor overall survival and disease-free survival (53). This amplification is targetable with high sensitivity to FGFR inhibitors *in vitro*

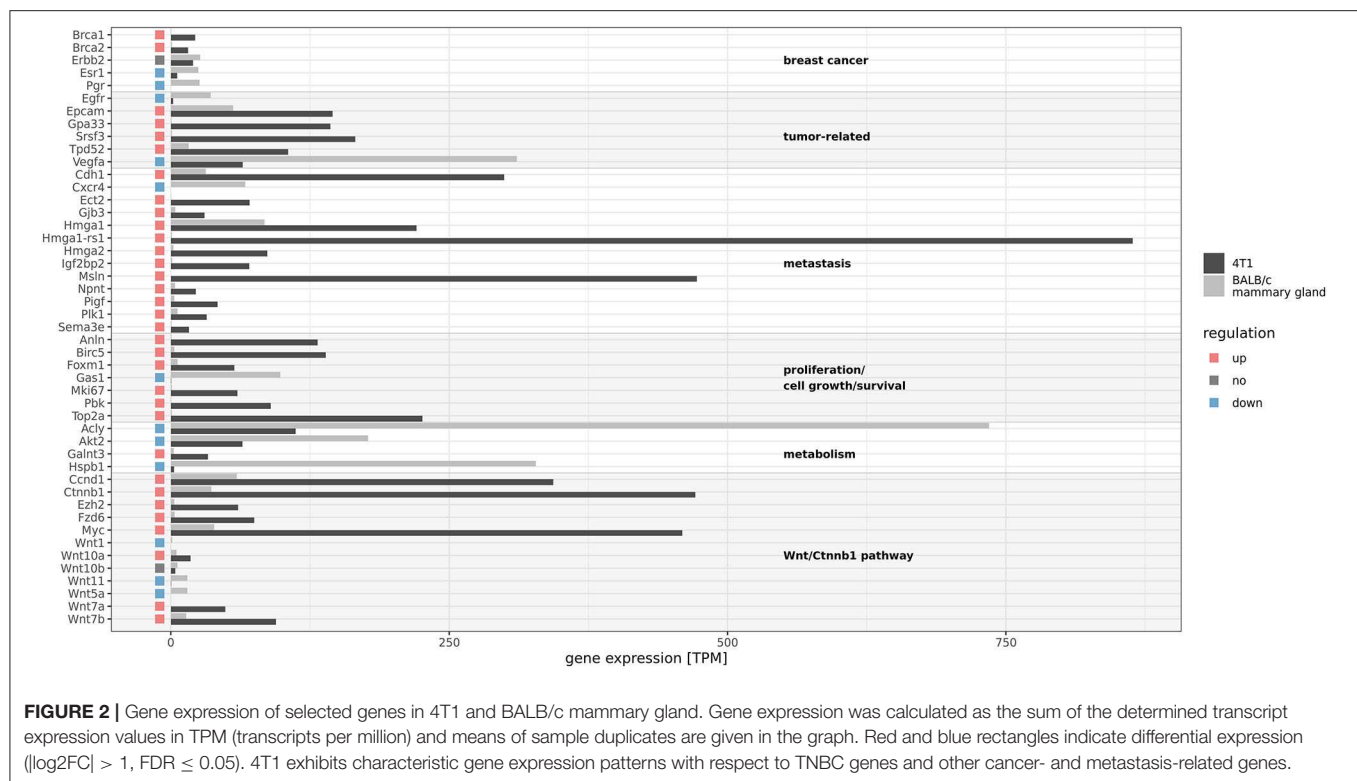
(52), an FGFR2-targeting antibody showed potent antitumor activity against human cancers in pre-clinical studies (54) and several FGFR tyrosine-kinase inhibitors are in clinical trials (54–56). However, the contribution of MMTV infection and initiation to human mammary carcinogenesis in general and FGFR2 amplification in particular is still highly debated (57). Of note, Notch4 and Krüppel-like factor 15 (Klf15) have been shown to be associated with MMTV CIS and although both genes are expressed in normal murine mammary gland, we do not find any isoform expressed in 4T1 possibly due to MMTV integration. Interestingly, while KLF15 has been recently proposed to be a tumor suppressor in breast cancer (58) and silencing this transcription factor results in a fitness advantage for the tumor, Notch-4 is a potent breast oncogene, overexpressed in TNBC (59) and Notch signaling is involved in mammary gland tumorigenesis (60).

The 4T1 Transcriptome

Differential expression analysis of 4T1 cell RNA expression vs. healthy mammary gland tissue RNA revealed 12810 differentially expressed genes ($FDR \leq 5\%$, absolute $\log_2$ fold-change >1) out of 29,955 total genes in mm9 (Tables S5, S6). This set of differentially expressed genes is very similar to differentially expressed genes in human breast cancer: we compared the gene sets of two studies comparing TNBC epithelium to adjacent microdissected stroma (21) and TNBC to non-TNBC cancers (22). These studies allowed a gene set enrichment test, yielding p -values of 2.2×10^{-16} and 0.001002 (Fisher's exact test), respectively. Next, we compared pathways and gene ontologies (GO) that were significantly enriched ($FDR \leq 0.05$, Table S7) in 4T1 differentially expressed genes to a study including different TNBC subtypes (20). Here, we only found significant overlap with top pathways and GO terms reported for subtype “Basal-like and immune suppressed (BLIS)” ($p_{\text{pathway}} = 0.04506$ and $p_{GO} = 0.0142$, Fisher's exact test). Furthermore, we analyzed RNA-Seq data of 57 TNBC breast cancer samples from the short read archive (PRJNA607061) and 66 breast tissue samples from the GTEx project (Table S8). All analysis steps were performed in analogy to the analysis of the 4T1 data. Here, we computed a p -value of 2.2×10^{-16} with Fisher's exact test when comparing the sets of differentially expressed genes. Moreover, the mean gene expression in TNBC is well-correlated to the gene expression in 4T1, as demonstrated by a Pearson's correlation coefficient of 0.727 (Figure S5).

Figure 2 shows the expression of a selection of relevant genes discussed below. The murine homologs of the typical genes associated with TNBC are Esr1, Pgr, and Erbb2. While Esr1 was about 2-fold downregulated and Pgr showed zero expression, Erbb2 had a comparable expression in 4T1 vs. the non-cancer mammary gland samples (about 20 TPM). However, compared to the ERBB2 expression in the TCGA human breast cancer (BRCA) cohort, this value was on the lower end of the expression level spectrum [not shown⁹ and (61)]. In order to investigate this detected mRNA expression, we compared the ERBB2, ESR1, and PGR mRNA expression in available TCGA breast cancer samples

⁹<http://gepia.cancer-pku.cn/detail.php?gene=ERBB2>



and grouped the expression values by the annotated result of the immunohistochemistry (IHC) assay. A principal component analysis (**Figure S6**) showed, that mRNA expression can separate IHC positives from negatives (albeit not perfectly). The data also showed that a negative IHC result is not necessarily associated with zero mRNA expression (**Figure S7**). With copy numbers of five, the three genes also did not divert from the general genomic copy number level. Moreover, genes *Brca1* and *Brca2* were highly overexpressed.

4T1 is a widely used model for metastatic breast cancer (62) and consistently, we found known metastasis-associated genes such as the differentiation antigen *Msln* (mesothelin), *Cdh1*, *Sema3e*, *Gjb3*, and *Ect2* to be overexpressed. The latter one is known to be a key factor in progression of breast cancer (63) as well as in metastasis, and high expression is associated with poor prognosis for TNBC patients (64, 65). Overexpression of mesothelin was shown to promote invasion and metastasis in breast cancer cells (66). Interestingly, we found also High-mobility group protein HMG-I/HMG-Y (*Hmga1*) and *Hmga*-related sequence 1 (*Hmga1-rs1*) to be upregulated in 4T1 cells. *Hmga1* is involved in promoting metastatic processes in breast cancer (67) and it has also been found to stimulate retroviral integration (68). *Hmga2* is a driver of tumor metastasis (69) and *Igf2bp2* is a downstream target gene (70). Both genes were highly expressed in 4T1 cells. In addition, we found a 6-fold overexpression of Nephronectin (*Npnt*) in 4T1 compared to the normal murine breast samples examined, in which we detected only weak signals of this gene (22.4 vs. 3.6 TPM). *Npnt* plays a role in kidney development, is associated with embryonic precursors of the urogenital system (71) as well as

with integrin expression (72). High expression levels of *Npnt* have been observed in human thyroid (median: 277 TPM), human blood vessels (e.g., aorta, 200 TPM), human lung (161 TPM) and to a much lesser extent in human mammary tissue (14 TPM)¹⁰. Furthermore, *Npnt* has been suggested to have a role in promoting metastasation, as decreased expression in 4T1 tumors significantly inhibited spontaneous metastasis to the lung (73), further indicating the highly metastatic phenotype of 4T1. In contrast, we found an extremely low expression of *Gas1*, which plays a role in growth suppression. Also, growth factor *Vegfa* and growth factor receptor *Egfr* were downregulated.

Other deregulated genes are also described as being cancer-related, including *Srsf3*, which has a proto-oncogenic function and is frequently upregulated in various types of cancer (74). *FOXM1* is a proto-oncogene involved in regulating the expression of genes that are specific for the G2/M DNA damage checkpoint during cell cycle prior to mitosis. *Foxm1* has been found overexpressed in a variety of solid tumors, including breast cancer (75) and indeed, we also observed a 9-fold increase in 4T1 cells. *PLK1* is also involved in the G2/M transition, found to be significantly overexpressed in TNBC and targeting this gene has been described as a potential therapeutic option for TNBC patients (76). Tumor protein D52 (*Tpd52*) was 6-fold upregulated, which is in consistence with reports showing high overexpression in many solid tumors and in particular breast cancer (77). Of note, we found the colon cancer antigen *Gpa33* (78) to be highly expressed in 4T1 (143 TPM), not in normal

¹⁰<https://gtexportal.org/home/gene/NPNT> (accessed January 9, 2020)

murine breast (<1 TPM) and not in any other human non-cancer tissue except colon (median: 111 TPM) and small intestine (median: 75 TPM) (data from¹¹).

Among factors associated with a poor prognosis, proliferation markers *Top2a*, *Mki67*, and *Birc5* (79–81) were highly expressed in 4T1, while almost absent in normal murine breast tissues. *Pbk* is also considered a marker for cellular proliferation (82) and is associated with poorer prognosis in lung cancer (83). *Anln* is highly expressed in breast cancer tissues (84) and a marker of poor prognosis in breast cancer (85) and indeed, we also found high expression of this gene in 4T1 (131.8 TPM). In addition, *Pigf*, which has been shown to enhance breast cancer motility (86) was overexpressed in 4T1 (42 TPM vs. 32.7 TPM). Genes related to metabolic regulation, such as *Acly* and *Akt2*, were downregulated. Polypeptide N-acetylgalactosaminyltransferase 3 (*Galnt3*) was upregulated in 4T1 and overexpression of this gene is associated with shorter progression-free survival in advanced ovarian cancer (87).

Moreover, *Wnt7a* and *Wnt7b* were upregulated in 4T1 cells, while other components of the Wnt/β-catenin pathway were downregulated (*Wnt1*, *Wnt11*, and *Wnt5a*). The role of *Wnt10b* in TNBC has been described before (88), indicating a direct effect on *Hmga2* expression (see above). Furthermore, the gene *Ezh2*, known for its deregulatory activity of the Wnt pathway, was upregulated as well. Consequently, we found the Wnt target genes including the proto-oncogene *Myc* and the genes *Ctnnb1*, *Ccnd1*, and *Fzd6* (Frizzled) to be upregulated (89).

As reported before (90), we found expression of the Murine Leukemia Virus (MuLV) gene coding for gp70, as well as of genes of the Murine osteosarcoma virus (NC_001506.1) and (confirming the genomic findings on MMTV integration) of all MMTV genes (Table S9).

6-Thioguanine Resistance

Due to the resistance to 6-thioguanine (6-TG) treatment, metastatic 4T1 cells can be precisely quantified even in distant organs (6). The cytotoxicity of 6-TG is based on the conversion of 6-TG into 2'-deoxy-6-thioguanosine triphosphate which can be incorporated into DNA (91). Deficiency in MMR, which is found in various cancer types (92), is associated with resistance to 6-TG (91). In 4T1, we did observe significant downregulation of *Pold4* only, but none of the other MMR genes (*Exo1*, *Lig1*, *Mlh1*, *Mlh3*, *Msh2*, *Msh3*, *Msh6*, *Pcna*, *Pcna-ps2*, *Pms2*, *Pold1*, *Pold2*, *Pold3*, *Rfc1*, *Rfc2*, *Rfc3*, *Rfc4*, *Rfc5*, *Rpa1*, *Rpa2*, *Rpa3*, and *Ssbp1*; MSigDB: C2 curated gene sets, KEGG_MISMATCH_REPAIR, mouse orthologs obtained from¹²) at mRNA level (Table S6). Moreover, no non-synonymous SNVs or indels were detected in these genes, which might have impaired their function. In addition, mutational signatures AC6 and AC20 (associated with defective MMR) are present, but with relatively weak signals of about 5% and less (Figure 1B). Signatures AC15 and AC26 (also associated with defective MMR) are not detected. Diouf et al. (93) observed in human leukemia cells that MMR deficiency and thus an increased resistance to thiopurines can also result from a

deregulated MSH2 degradation. While we again did not detect any mutations in the genes involved in regulating the stability of MSH2 (*Mtor*, *Herc1*, *Prkcz*, and *Pik3c2b*), we found *Pik3c2b* to be downregulated (Table S6). As the knockdown of *PIK3C2B* in human leukemia CCRF-CEM cells decreased sensitivity to 6-TG in comparison to control (93), lacking or reduced expression of *Pik3c2b* mRNA in 4T1 might explain the resistance to 6-TG treatment.

MHC Expression

The key players of the mammalian adaptive immune system are the major histocompatibility complex (MHC) molecules with the primary task to bind and present self, abnormal self, and foreign peptides derived from intracellular (MHC class I) or from extracellular proteins (MHC class II) on the surface of nucleated cells for recognition by T lymphocytes. Novel cancer immunotherapy concepts target tumor-specific antigens (either tumor-associated antigens or neo-epitopes) presented by MHC molecules of tumor cells. In general, non-cancer murine tissues show variable expression of MHC class I and class II, with lymphatic organs (i.e., lymph node, spleen) showing highest abundance of MHC transcripts and brain having the lowest MHC expression (Figure 3), which is in agreement with expression patterns of the human MHC system (94).

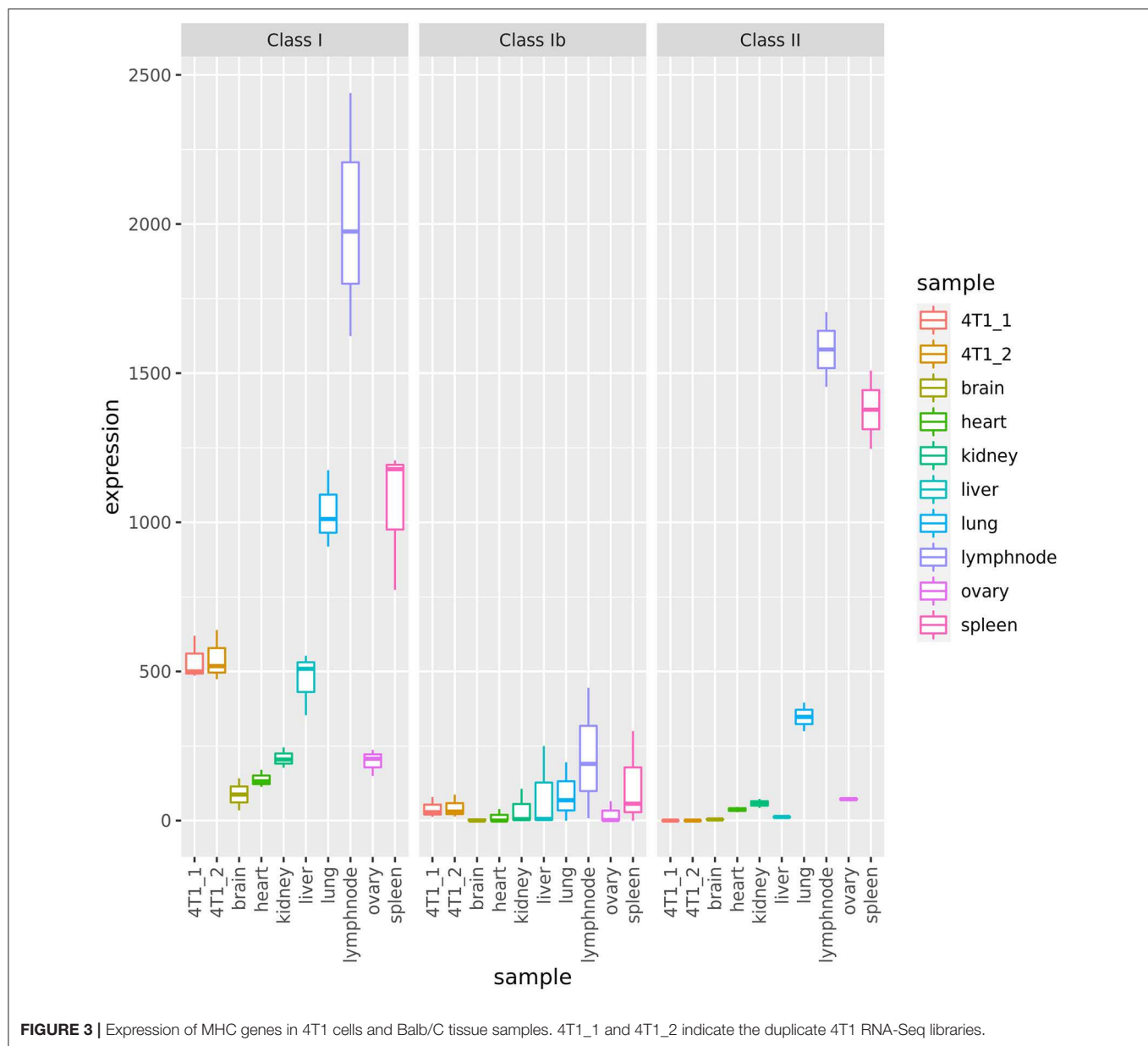
We confirmed that 4T1 cells have the same class I MHC haplotype as the parental BALB/c mice: H-2D^d, H-2K^d, and H-2L^d. MHC class II could not be typed from RNA-Seq reads due to lack of expression. In 4T1, we found MHC class I and Ib loci to be expressed at comparable levels to normal (non-lymphatic) tissues (Figure 3, Table S10). In addition, β₂-microglobulin (*B2m*), essential component of the MHC class I complex, and members of the MHC class I antigen presenting pathway were expressed (Figure S8). This suggests that MHC class I antigen presentation is functional and thus 4T1 cells are capable of presenting peptides and neo-epitopes to T effector cells. In contrast, 4T1 cells expressed neither MHC class II nor the MHC class II master regulator and transcriptional coactivator *Ciita* [Figure 3, Figure S8; (95)]. Both findings suggest that 4T1 cells do not have functional MHC class II antigen presentation.

4T1 Neoantigens

To investigate the mutations with regard to their potential to elicit immune responses *in vivo*, experiments in mice were conducted. In a previous study (34), we already examined 38 SNVs detected in the 4T1-luc2-tdtomato mammary carcinoma (4T1-Luc) cell line. Thirty-six of these were also present in the WT 4T1 cell line, 16 of which were immunogenic. Based on the subsequent re-analysis of WT 4T1, we selected additional eleven SNVs and two indels for immunogenicity assessment (Figure 4A). This selection was done by filtering the available set of potential neoantigens in order to enrich for likely immunogenic peptide sequences (see Methods). To this end, a vaccine for each of the newly selected 13 mutations was engineered using antigen-encoding pharmacologically optimized lipoplexed RNA as vaccine format. As before, SNVs were flanked by 13 amino acids of WT sequence, in-frame indel

¹¹<https://gtexportal.org/home/gene/GPA33> (accessed January 9, 2020)

¹²<http://bioinf.wehi.edu.au/software/MSigDB/>

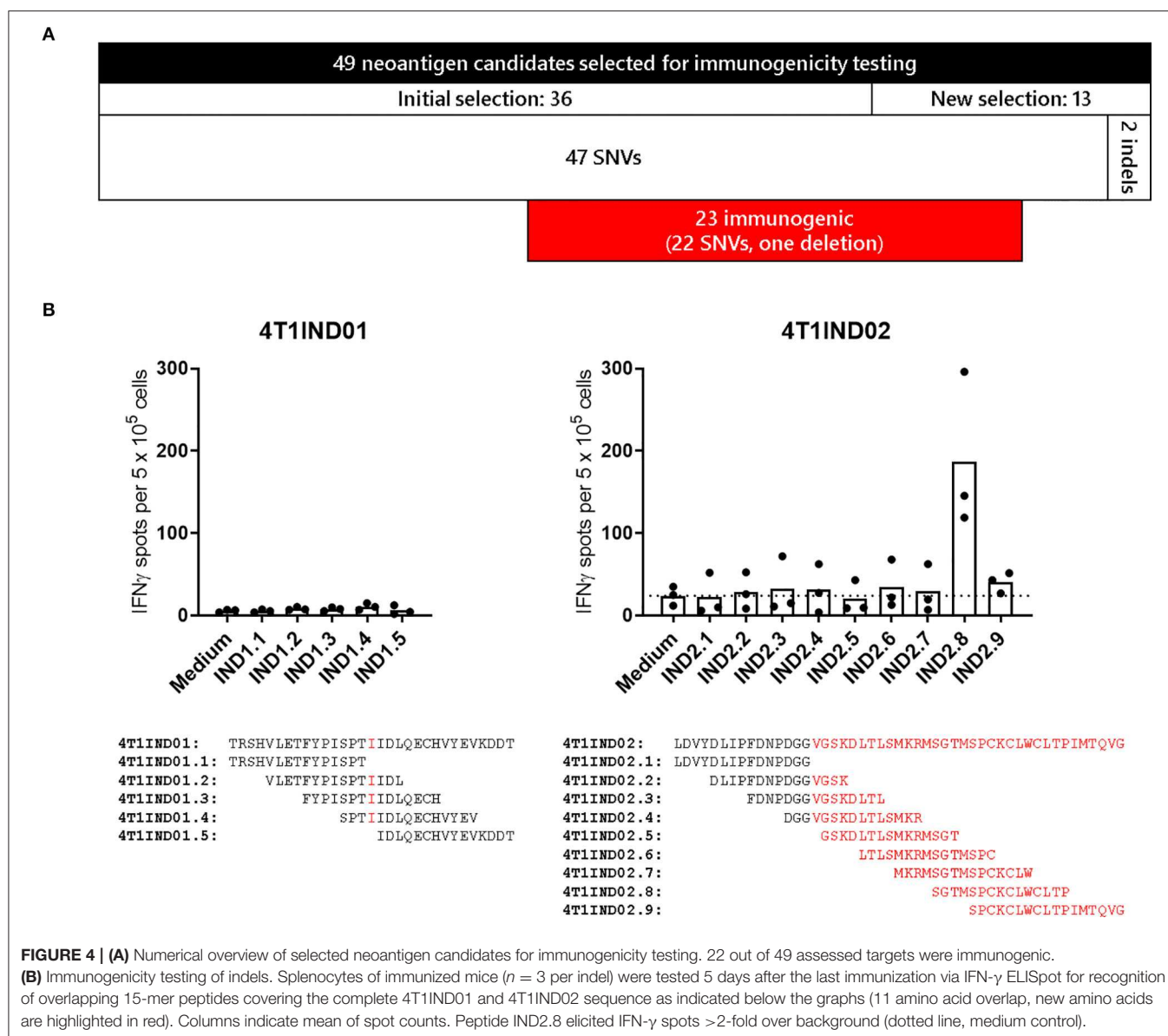


mutations were flanked by 15 amino acids of WT sequence and frameshift mutations were investigated covering 15 WT amino acids upstream of the mutations as well as the whole sequence of new amino acids until reaching a stop codon. Mice ($n = 3-5$) were immunized intravenously three times within a 2-week timeframe. IFN- γ ELISpot of splenocytes stimulated with overlapping 15-mer peptides covering the respective vaccinated sequence was performed 5 days after the last immunization. With this, we found immune responses against additional six SNVs and one deletion (see **Figure 4B** for the results on the indels, **Table S11** summarizes all immune responses). In total, we can thus report 22 SNVs and one deletion identified in 4T1 triggering immune responses in immunized mice. Of note, only four and 14 of these were derived from SNVs already reported

before (37, 38). For a subset of 15 SNVs the WT counterpart was tested, which revealed that 10 responses were clearly specific for the mutated sequence. As already observed (34), most of the reactivities were elicited by CD4⁺ T cells (15 out of 21 analyzed mutations). Two SNVs were targeted by CD4⁺ and CD8⁺ T cells.

CONCLUSION

The murine mammary cancer cell line 4T1 is one of the most often used model systems for breast cancer and in particular TNBC. Here, we could confirm that 4T1 indeed resembles metastatic TNBC at the transcriptional level with respect to key markers *Esr1*, *ErbB2*, and *Pgr*. In addition, compared to



human TNBC data, we found good concordance on the level of differentially expressed genes and pathways and a reasonable correlation of raw expression values. The expression profile was in agreement with the metastatic phenotype of 4T1, as we found Msln, Ect2, and Plk1, and other genes associated to metastasis to be highly overexpressed in comparison to normal mammary gland. As described above, also a number of genes involved in proliferation and survival were deregulated. Moreover, it is known that the Wnt/ β -catenin (Ctnnb1) pathway plays an important role in human breast cancers (96) with high activation rates and association with a poor prognosis (97). Some components of this pathway including Wnt target genes were upregulated in 4T1 cells. Overall, the observed profile reflected the complex interplay of various factors of tumorigenesis- and metastasis-driving signaling and allows for further mode-of-action investigation in the 4T1 tumor model.

On the mutation level, the raw numbers of mutations compare well against the CT26 colon cancer model. CT26 has 3,023 SNVs and 362 short indels, and in 4T1 we found an order of magnitude less variants (505 SNVs and 20 short indels). This is a similar relationship as observed for human colorectal and breast cancer (31) and supports previous findings (37, 38) as mentioned above. Differences in the absolute numbers in comparison to these reports might be due to genetic diversification of *in vitro* cell lines investigated at different laboratories at differing passage numbers (98) or different sequencing and mutation calling strategies.

Here and in a previous study (34), we determined *in vivo* immune responses against 22 SNVs (out of 49 tested, 45%) as well as one deletion (out of two indels tested) upon vaccination of BALB/c mice and 10 mutations (out of 15 immunogenic SNVs) showed mutation specificity. Although we did not examine all possible candidate neoantigens, the low mutational burden

and the similarity to the basal-like and immune suppressed TNBC subtype suggest that 4T1 is a tumor model exhibiting relatively low immunogenicity. This is in agreement with others (37), while different studies argue the opposite, showing upregulation of many immune activation genes (38, 99) and thus immune cell infiltration in transplanted 4T1 tumors. Our 4T1 RNA-Seq data, however, was generated from the pure cell line. Accordingly, we could not see upregulation of immune-related genes. Nonetheless, 4T1 cells can secrete a plethora of inflammatory mediators and thereby modulate not only lymphocyte-mediated immune responses against the tumor, but also the innate microbial host defense (100–102). In future studies, the identified fusion transcripts might also be viable and interesting candidates for immunogenicity testing.

Besides the expression of MMTV at the RNA level and the deregulation of known genes with nearby insertion sites, we found direct evidence of MMTV integration near the gene *Fgfr2*. Combined with the relatively low mutational burden, we hypothesize that the MMTV infection and integration is the major genomic change causing eventually the TNBC-like phenotype. Interestingly, despite no observed somatic mutations in *Brca1* or *Brca2*, a “BRCAness” mutation signature could be found (Figure 1B, signature AC3).

A very recent publication (38) underlined the importance of profiling tumor models to appropriately translate pre-clinical findings. The here presented genome, transcriptome, and immunome data serves as a baseline for further studies, examining e.g., tumor-host interactions in terms of immunogenicity and TNBC in general. Although the data sources are highly heterogeneous (resulting from different studies and sequencing experiments), a distinct overlap between our qualitative and quantitative findings and studies on human TNBC can be found and confirms our approach. Together, our study supports the rational design of pre-clinical studies with an important and established tumor model.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the European Nucleotide Archive (<https://www.ebi.ac.uk/ena/browser/home>) with the study accession number PRJEB36287.

ETHICS STATEMENT

This animal study was reviewed and approved by the federal authorities of Rhineland-Palatinate, Germany and all mice were kept in accordance with federal and state policies on animal research at the University of Mainz and BioNTech SE.

AUTHOR CONTRIBUTIONS

ML, BS, SB, and US contributed to the conception and design of the study. CA, VB, and KM were responsible for NGS sequencing. ML, BS, SB, TB, CH, and CR performed bioinformatics analyses. ML, BS, MV, and AT selected neoantigen candidates. MV planned and performed immunogenicity testing. ML, BS, and SB

interpreted the data and wrote the first draft of the manuscript. CH, TB, and MV wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (Grant Agreement No. 789256).

ACKNOWLEDGMENTS

We thank John C. Castle, Sebastian Kreiter, and Mustafa Diken for discussions and advice, Luisa Bresadola and Jonas Ibn-Salem for proofreading the manuscript, and Karen Chu for project management support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fonc.2020.01195/full#supplementary-material>

Figure S1 | Comparison of DNA and RNA variant allele frequency (VAF) in 4T1 cells. The Pearson correlation coefficient is 0.977.

Figure S2 | Abundance of nucleotide substitutions in 4T1 cells with respect to nucleotide triplets.

Figure S3 | Differential expression of MMTV integration effector genes. Colored dots indicate differential expression in 4T1 vs. BALB/c mammary gland. Red gene labels indicate genes that are described as upregulated in the literature.

Figure S4 | Schematic view of proposed MMTV integration in *Fgfr2* gene. Upper panel shows a UCSC Genome Browser view of an alignment of assembled sequence c75264_g4_i1 to the mm9 genome. The middle part shows the assembled sequence (blue) and the part mapping to *Fgfr2* (red). Numbers indicate parts of the sequence mapping to *Fgfr2* (red) and MMTV (green). The lower panel shows a schematic of Betaretrovirus genome, for which MMTV is a reference strain (taken from <https://viralzone.expasy.org/66>).

Figure S5 | Mean gene expression of TNBC plotted against mean gene expression in orthologous genes of 4T1. Counts per million (cmp) were computed by edgeR.

Figure S6 | Scatterplot of a principal component analysis of TCGA BRCA gene expression of genes *ERBB2*, *ESR1*, and *PGR*. Shown are the first two principal components (PC1 and PC2). Ellipses indicate normal-probability contours.

Figure S7 | Boxplot of TCGA BRCA gene expression of genes *ERBB2*, *ESR1*, and *PGR*, separated by TNBC status. Expression on y-axis is given as $\log_2$ (FPKM+1) units.

Figure S8 | Gene expression of members of the MHC class I and II antigen presenting pathway in 4T1 and BALB/c mammary gland.

Table S1 | Raw Control-FREEC output (sheet 1) and predicted absolute gene copy numbers of 4T1 genes (sheet 2).

Table S2 | Somatic SNVs in 4T1, including annotation on amino acid substitutions, affected genes/transcripts, expression of these, and coverage/VAF in the DNA/RNA NGS libraries.

Table S3 | Somatic INDELs in 4T1, including annotation on frameshift, affected genes/transcripts and coverage/VAF in the DNA/RNA NGS libraries. A VAF of –1 means “not covered,” while a VAF of 0 indicates coverage but absence of the variant allele.

Table S4 | Fusion genes in 4T1, including predicted positions of breakpoints, number of junction reads, and spanning read pairs and the program that detected a fusion.

Table S5 | Gene expression in 4T1 and BALB/c mammary gland in TPM.

Table S6 | Differential gene expression in 4T1 vs. BALB/c mammary gland, showing log fold change, FDR, and baseline expression values.

Table S7 | Gene set and pathway enrichment in differentially expressed genes of 4T1 cells for upregulated and downregulated genes in GO gene sets and KEGG pathways, respectively (sheets are labeled "up GO", "up KEGG", "down GO", and "down KEGG", respectively).

Table S8 | Differential gene expression in human TNBC vs. breast tissue, showing log fold change, FDR and baseline expression values.

Table S9 | Expression in RPKM of MMTV genes for two replicates of 4T1 RNA-Seq libraries.

Table S10 | Expression values in TPM of MHC genes in 4T1 and BALB/c tissues. The used reference sequences from the UCSC known genes or Genbank are also listed.

Table S11 | Results of immunogenicity testing, including details on mutation, amino acid substitution, the result of the ELISpot assay, the subtype of the T-cell response, and the specificity when compared to a WT control.

REFERENCES

- Gravekamp C, Leal B, Denny A, Bahar R, Lampkin S, Castro F, et al. *In vivo* responses to vaccination with Mage-b, GM-CSF and thioglycollate in a highly metastatic mouse breast tumor model, 4T1. *Cancer Immunol Immunother.* (2008) 57:1067–77. doi: 10.1007/s00262-007-0438-5
- Gupta SK, Tiwari AK, Gandham RK, Sahoo AP. Combined administration of the apoptin gene and poly (I:C) induces potent anti-tumor immune response and inhibits growth of mouse mammary tumors. *Int Immunopharmacol.* (2016) 35:163–73. doi: 10.1016/j.intimp.2016.03.034
- Tiash S, Chua MJ, Chowdhury EH. Knockdown of ROS1 gene sensitizes breast tumor growth to doxorubicin in a syngeneic mouse model. *Int J Oncol.* (2016) 48:2359–66. doi: 10.3892/ijo.2016.3452
- Dexter DL, Kowalski HM, Blazar BA, Fligiel Z, Vogel R, Heppner GH. Heterogeneity of tumor cells from a single mouse mammary tumor. *Cancer Res.* (1978) 38:3174–81.
- Aslakson CJ, Miller FR. Selective events in the metastatic process defined by analysis of the sequential dissemination of subpopulations of a mouse mammary tumor. *Cancer Res.* (1992) 52:1399–405.
- Pulaski BA, Ostrand-Rosenberg S. Mouse 4T1 breast tumor model. *Curr Protoc Immunol.* (2001) Chapter 20:Unit 20.2. doi: 10.1002/0471142735.im2002s39
- Chiu H-W, Yeh Y-L, Wang Y-C, Huang W-J, Chen Y-A, Chiou Y-S, et al. Suberoylanilide hydroxamic acid, an inhibitor of histone deacetylase, enhances radiosensitivity and suppresses lung metastasis in breast cancer *in vitro* and *in vivo*. *PLoS ONE.* (2013) 8:e76340. doi: 10.1371/journal.pone.0076340
- Parvani JG, Davuluri G, Wendt MK, Espinosa C, Tian M, Danielpour D, et al. Depror enhances triple-negative breast cancer metastasis and chemoresistance through coupling to survivin expression. *Neoplasia.* (2015) 17:317–28. doi: 10.1016/j.neo.2015.02.003
- Simone BA, Dan T, Palagani A, Jin L, Han SY, Wright C, et al. Caloric restriction coupled with radiation decreases metastatic burden in triple negative breast cancer. *Cell Cycle.* (2016) 15:2265–74. doi: 10.1080/15384101.2016.1160982
- Kaur P, Nagaraja GM, Zheng H, Gizachew D, Galukande M, Krishnan S, et al. A mouse model for triple-negative breast cancer tumor-initiating cells (TNBC-TICs) exhibits similar aggressive phenotype to the human disease. *BMC Cancer.* (2012) 12:120. doi: 10.1186/1471-2407-12-120
- Anders CK, Carey LA. Biology, metastatic patterns, and treatment of patients with triple-negative breast cancer. *Clinical Breast Cancer.* (2009) 9 (Suppl. 2):S73–81. doi: 10.3816/CBC.2009.s.008
- Rosenbloom KR, Sloan CA, Malladi VS, Dreszer TR, Learned K, Kirkup VM, et al. ENCODE data in the UCSC genome browser: year 5 update. *Nucleic Acids Res.* (2013) 41:D56–63. doi: 10.1093/nar/gks1172
- Castle JC, Loewer M, Boegel S, Graaf J, de Bender C, Tadmor AD, et al. Immunomic, genomic and transcriptomic characterization of CT26 colorectal carcinoma. *BMC Genomics.* (2014) 15:190. doi: 10.1186/1471-2164-15-190
- Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics.* (2009) 25:1754–60. doi: 10.1093/bioinformatics/btp324
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* (2013) 29:15–21. doi: 10.1093/bioinformatics/bts635
- Saunders CT, Wong WSW, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics.* (2012) 28:1811–7. doi: 10.1093/bioinformatics/bts271
- Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol.* (2016) 34:525–7. doi: 10.1038/nbt.3519
- Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* (2010) 26:139–40. doi: 10.1093/bioinformatics/btp616
- Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* (2016) 44:W90–7. doi: 10.1093/nar/gkw377
- Liu YR, Jiang YZ, Xu XE, Yu KD, Jin X, Hu X, et al. Comprehensive transcriptome analysis identifies novel molecular subtypes and subtype-specific RNAs of triple-negative breast cancer. *Breast Cancer Res.* (2016) 18:33. doi: 10.1186/s13058-016-0690-8
- Lehmann BD, Jovanović B, Chen X, Estrada MV, Johnson KN, Shyr Y, et al. Refinement of triple-negative breast cancer molecular subtypes: implications for neoadjuvant chemotherapy selection. *PLoS ONE.* (2016) 11:e0157368. doi: 10.1371/journal.pone.0157368
- Dong P, Yu B, Pan L, Tian X, Liu F. Identification of key genes and pathways in triple-negative breast cancer by integrated bioinformatics analysis. *Biomed Res Int.* (2018) 2018:2760918. doi: 10.1155/2018/2760918
- Haas BJ, Dobin A, Li B, Stransky N, Pochet N, Regev A. Accuracy assessment of fusion transcript detection via read-mapping and *de novo* fusion transcript assembly-based methods. *Genome Biol.* (2019) 20:213. doi: 10.1186/s13059-019-1842-9
- Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.* (2010) 38:e178. doi: 10.1093/nar/gkq622
- Jia W, Qiu K, He M, Song P, Zhou F, et al. SOAPfuse: an algorithm for identifying fusion transcripts from paired-end RNA-Seq data. *Genome Biol.* (2013) 14:R12. doi: 10.1186/gb-2013-14-2-r12
- Okonechnikov K, Imai-Matsushima A, Paul L, Seitz A, Meyer TF, Garcia-Alcalde F. InFusion: advancing discovery of fusion genes and chimeric transcripts from deep RNA-sequencing data. *PLoS ONE.* (2016) 11:e0167417. doi: 10.1371/journal.pone.0167417
- Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, Schleiermacher G, et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics.* (2012) 28:423–5. doi: 10.1093/bioinformatics/btr670
- Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol.* (2011) 29:644–52. doi: 10.1038/nbt.1883
- Kent WJ. BLAT—the BLAST-like alignment tool. *Genome Res.* (2002) 12:656–64. doi: 10.1101/gr.229202
- Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat Biotechnol.* (2014) 32:462–4. doi: 10.1038/nbt.2862

31. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. *Nature*. (2013) 500:415–21. doi: 10.1038/nature12477
32. Brister JR, Ako-Adjei D, Bao Y, Blinkova O. NCBI viral genomes resource. *Nucleic Acids Res*. (2015) 43:D571–7. doi: 10.1093/nar/gku1207
33. Scholtalbers J, Boegel S, Bukur T, Byl M, Goerges S, Sorn P, et al. TCLP: An online cancer cell line catalogue integrating HLA type, predicted neo-epitopes, virus and gene expression. *Genome Med*. (2015) 7:118. doi: 10.1186/s13073-015-0240-5
34. Kreiter S, Vormehr M, van de Roemer N, Diken M, Löwer M, Diekmann J, et al. Mutant MHC class II epitopes drive therapeutic immune responses to cancer. *Nature*. (2015) 520:692–6. doi: 10.1038/nature14426
35. Kranz LM, Diken M, Haas H, Kreiter S, Loquai C, Reuter KC, et al. Systemic RNA delivery to dendritic cells exploits antiviral defence for cancer immunotherapy. *Nature*. (2016) 534:396–401. doi: 10.1038/nature18300
36. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive characterization of cancer driver genes and mutations. *Cell*. (2018) 173:371–85.e18. doi: 10.1016/j.cell.2018.02.060
37. Kim K, Skora AD, Li Z, Liu Q, Tam AJ, Blosser RL, et al. Eradication of metastatic mouse cancers resistant to immune checkpoint blockade by suppression of myeloid-derived cells. *Proc Natl Acad Sci USA*. (2014) 111:11774–9. doi: 10.1073/pnas.1410626111
38. Zhong W, Myers JS, Wang F, Wang K, Lucas J, Rosfjord E, et al. Comparison of the molecular and cellular phenotypes of common mouse syngeneic models with human tumors. *BMC Genomics*. (2020) 21:2. doi: 10.1186/s12864-019-6344-3
39. Shah SP, Roth A, Goya R, Oloumi A, Ha G, Zhao Y, et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature*. (2012) 486:395–9. doi: 10.1038/nature10933
40. Stephens PJ, Tarpey PS, Davies H, van Loo P, Greenman C, Wedge DC, et al. The landscape of cancer genes and mutational processes in breast cancer. *Nature*. (2012) 486:400–4. doi: 10.1038/nature11017
41. Nik-Zainal S, Alexandrov LB, Wedge DC, van Loo P, Greenman CD, Raine K, et al. Mutational processes molding the genomes of 21 breast cancers. *Cell*. (2012) 149:979–93. doi: 10.1016/j.cell.2012.04.024
42. Burns MB, Lackey L, Carpenter MA, Rathore A, Land AM, Leonard B, et al. APOBEC3B is an enzymatic source of mutation in breast cancer. *Nature*. (2013) 494:366–70. doi: 10.1038/nature11881
43. Okeoma CM, Lovsin N, Peterlin BM, Ross SR. APOBEC3 inhibits mouse mammary tumour virus replication *in vivo*. *Nature*. (2007) 445:927–30. doi: 10.1038/nature05540
44. Edgren H, Murumagi A, Kangaspeska S, Nicorici D, Hongisto V, Kleivi K, et al. Identification of fusion genes in breast cancer by paired-end RNA-sequencing. *Genome Biol*. (2011) 12:R6. doi: 10.1186/gb-2011-12-1-r6
45. Banerji S, Cibulskis K, Rangel-Escareno C, Brown KK, Carter SL, Frederick AM, et al. Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature*. (2012) 486:405–9. doi: 10.1038/nature11154
46. Jain VK, Turner NC. Challenges and opportunities in the targeting of fibroblast growth factor receptors in breast cancer. *Breast Cancer Res*. (2012) 14:208. doi: 10.1186/bcr3139
47. Theodorou V, Kimm MA, Boer M, Wessels L, Theelen W, Jonkers J, et al. MMTV insertional mutagenesis identifies genes, gene families and pathways involved in mammary cancer. *Nat Genet*. (2007) 39:759–69. doi: 10.1038/ng2034
48. Klijn C, Koudijs MJ, Kool J, Hoeve J, ten Boer M, Moes J, et al. Analysis of tumor heterogeneity and cancer gene networks using deep sequencing of MMTV-induced mouse mammary tumors. *PLoS ONE*. (2013) 8:e62113. doi: 10.1371/journal.pone.0062113
49. Callahan R, Mudunur U, Bargo S, Raafat A, McCurdy D, Boulanger C, et al. Genes affected by mouse mammary tumor virus (MMTV) proviral insertions in mouse mammary tumors are deregulated or mutated in primary human mammary tumors. *Oncotarget*. (2012) 3:1320–34. doi: 10.18632/oncotarget.682
50. Kunii K, Davis L, Gorenstein J, Hatch H, Yashiro M, Di Bacco A, et al. FGFR2-amplified gastric cancer cell lines require FGFR2 and Erbb3 signaling for growth and survival. *Cancer Res*. (2008) 68:2340–8. doi: 10.1158/0008-5472.CAN-07-5229
51. Yates LR, Gerstung M, Knappskog S, Desmedt C, Gundem G, van Loo P, et al. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat Med*. (2015) 21:751–9. doi: 10.1038/nm.3886
52. Turner N, Lambros MB, Horlings HM, Pearson A, Sharpe R, Natrajan R, et al. Integrative molecular profiling of triple negative breast cancers identifies amplicon drivers and potential therapeutic targets. *Oncogene*. (2010) 29:2013–23. doi: 10.1038/nc.2009.489
53. Sun S, Jiang Y, Zhang G, Song H, Zhang X, Zhang Y, et al. Increased expression of fibroblastic growth factor receptor 2 is correlated with poor prognosis in patients with breast cancer. *J Surg Oncol*. (2012) 105:773–9. doi: 10.1002/jso.22120
54. Bai A, Meetze K, Vo NY, Kollipara S, Mazza EK, Winston WM, et al. GP369, an FGFR2-IIIb-specific antibody, exhibits potent antitumor activity against human cancers driven by activated FGFR2 signaling. *Cancer Res*. (2010) 70:7630–9. doi: 10.1158/0008-5472.CAN-10-1489
55. Turner N, Grose R. Fibroblast growth factor signalling: from development to cancer. *Nat Rev Cancer*. (2010) 10:116–29. doi: 10.1038/nrc2780
56. André F, Cortés J. Rationale for targeting fibroblast growth factor receptor signaling in breast cancer. *Breast Cancer Res Treat*. (2015) 150:1–8. doi: 10.1007/s10549-015-3301-y
57. Mason AL, Gilady SY, Mackey JR. Mouse mammary tumor virus in human breast cancer red herring or smoking gun? *Am J Pathol*. (2011) 179:1588–90. doi: 10.1016/j.ajpath.2011.08.003
58. Yoda T, McNamara KM, Miki Y, Onodera Y, Takagi K, Nakamura Y, et al. KLF15 in breast cancer: a novel tumor suppressor? *Cell Oncol*. (2015) 38:227–35. doi: 10.1007/s13402-015-0226-8
59. Clementz AG, Rogowski A, Pandya K, Miele L, Osipo C. NOTCH-1 and NOTCH-4 are novel gene targets of PEA3 in breast cancer: novel therapeutic implications. *Breast Cancer Res*. (2011) 13:R63. doi: 10.1186/bcr2900
60. Callahan R, Raafat A. Notch signaling in mammary gland tumorigenesis. *J Mammary Gland Biol Neoplasia*. (2001) 6:23–36. doi: 10.1023/A:1009512414430
61. Tang Z, Li C, Kang B, Gao G, Li C, Zhang Z. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res*. (2017) 45:W98–102. doi: 10.1093/nar/gkx247
62. Yang S, Zhang JJ, Huang X-Y. Mouse models for tumor metastasis. *Methods Mol Biol*. (2012) 928:221–8. doi: 10.1007/978-1-62703-008-3_17
63. Fields AP, Justilien V. The guanine nucleotide exchange factor (GEF) Ect2 is an oncogene in human cancer. *Adv Enzyme Regul*. (2010) 50:190–200. doi: 10.1016/j.advenzreg.2009.10.010
64. Mansour M, Haupt S, Chan A-L, Godde N, Rizzitelli A, Loi S, et al. The E3-ligase E6AP represses breast cancer metastasis via regulation of ECT2-Rho signaling. *Cancer Res*. (2016) 76:4236–48. doi: 10.1158/0008-5472.CAN-15-1553
65. Wang HK, Liang JF, Zheng HX, Xiao H. Expression and prognostic significance of ECT2 in invasive breast cancer. *J Clin Pathol*. (2018) 71:442–5. doi: 10.1136/jclinpath-2017-204569
66. Wang Y, Wang L, Li D, Wang HB, Chen QF. Mesothelin promotes invasion and metastasis in breast cancer cells. *J Int Med Res*. (2012) 40:2109–16. doi: 10.1177/030006051204000608
67. Pegoraro S, Ros G, Piazza S, Sommaggio R, Ciani Y, Rosato A, et al. HMGA1 promotes metastatic processes in basal-like breast cancer regulating EMT and stemness. *Oncotarget*. (2013) 4:1293–308. doi: 10.18632/oncotarget.1136
68. Beitzel B, Bushman F. Construction and analysis of cells lacking the HMGA gene family. *Nucleic Acids Res*. (2003) 31:5025–32. doi: 10.1093/nar/gkg684
69. Morishita A, Zaidi MR, Mito A, Sankarasharma D, Szabolcs M, Okada Y, et al. HMGA2 is a driver of tumor metastasis. *Cancer Res*. (2013) 73:4289–99. doi: 10.1158/0008-5472.CAN-12-3848
70. Brants JR, Ayoubi TAY, Chada K, Marchal K, van de Ven WJM, Petit MMR. Differential regulation of the insulin-like growth factor II mRNA-binding protein genes by architectural transcription factor HMGA2. *FEBS Lett*. (2004) 569:277–83. doi: 10.1016/j.febslet.2004.05.075
71. Miner JH. Mystery solved: discovery of a novel integrin ligand in the developing kidney. *J Cell Biol*. (2001) 154:257–9. doi: 10.1083/jcb.200106124

72. Fujiwara H, Ferreira M, Donati G, Marciano DK, Linton JM, Sato Y, et al. The basement membrane of hair follicle stem cells is a muscle cell niche. *Cell*. (2011) 144:577–89. doi: 10.1016/j.cell.2011.01.014
73. Eckhardt BL, Parker BS, van Laar RK, Restall CM, Natoli AL, Tavaría MD, et al. Genomic analysis of a spontaneous model of breast cancer metastasis to bone reveals a role for the extracellular matrix. *Mol Cancer Res*. (2005) 3:1–13.
74. Ajiro M, Jia R, Yang Y, Zhu J, Zheng Z-M. A genome landscape of SRSF3-regulated splicing events and gene expression in human osteosarcoma U2OS cells. *Nucleic Acids Res*. (2016) 44:1854–70. doi: 10.1093/nar/gkv1500
75. Wonsey DR, Follettie MT. Loss of the forkhead transcription factor FoxM1 causes centrosome amplification and mitotic catastrophe. *Cancer Res*. (2005) 65:5181–9. doi: 10.1158/0008-5472.CAN-04-4059
76. Maire V, Némati F, Richardson M, Vincent-Salomon A, Tesson B, Rigault G, et al. Polo-like kinase 1: a potential therapeutic option in combination with conventional chemotherapy for the management of patients with triple-negative breast cancer. *Cancer Res*. (2013) 73:813–23. doi: 10.1158/0008-5472.CAN-12-2633
77. Tennstedt P, Böhlch C, Strobel G, Minner S, Burkhardt L, Grob T, et al. Patterns of TPD52 overexpression in multiple human solid tumor types analyzed by quantitative PCR. *Int J Oncol*. (2014) 44:609–15. doi: 10.3892/ijo.2013.2200
78. Rageul J, Mottier S, Jarry A, Shah Y, Théoleyre S, Masson D, et al. KLF4-dependent, PPARgamma-induced expression of GPA33 in colon cancer cell lines. *Int J Cancer*. (2009) 125:2802–9. doi: 10.1002/ijc.24683
79. Lv Y-G, Yu F, Yao Q, Chen J-H, Wang L. The role of survivin in diagnosis, prognosis and treatment of breast cancer. *J Thorac Dis*. (2010) 2:100–10.
80. Shigematsu H, Ozaki S, Yasui D, Yamamoto H, Zaitzu J, Taniyama D, et al. Overexpression of topoisomerase II alpha protein is a factor for poor prognosis in patients with luminal B breast cancer. *Oncotarget*. (2018) 9:26701–10. doi: 10.18632/oncotarget.25468
81. Luo Z-W, Zhu M-G, Zhang Z-Q, Ye F-J, Huang W-H, Luo X-Z. Increased expression of Ki-67 is a poor prognostic marker for colorectal cancer patients: a meta analysis. *BMC Cancer*. (2019) 19:123. doi: 10.1186/s12885-019-5324-y
82. Ayllón V, O'Connor R. PBK/TOPK promotes tumour cell proliferation through p38 MAPK activity and regulation of the DNA damage response. *Oncogene*. (2007) 26:3451–61. doi: 10.1038/sj.onc.1210142
83. Shih MC, Chen JY, Wu YC, Jan YH, Yang BM, Lu PJ, et al. TOPK/PBK promotes cell migration via modulation of the PI3K/PTEN/AKT pathway and is associated with poor prognosis in lung cancer. *Oncogene*. (2012) 31:2389–400. doi: 10.1038/onc.2011.419
84. Zhou W, Wang Z, Shen N, Pi W, Jiang W, Huang J, et al. Knockdown of ANLN by lentivirus inhibits cell growth and migration in human breast cancer. *Mol Cell Biochem*. (2015) 398:11–9. doi: 10.1007/s11010-014-2200-6
85. O'Leary PC, Penny SA, Dolan RT, Kelly CM, Madden SF, Rexhepaj E, et al. Systematic antibody generation and validation via tissue microarray technology leading to identification of a novel protein prognostic panel in breast cancer. *BMC Cancer*. (2013) 13:175. doi: 10.1186/1471-2407-13-175
86. Taylor AP, Leon E, Goldenberg DM. Placental growth factor (PlGF) enhances breast cancer cell motility by mobilising ERK1/2 phosphorylation and cytoskeletal rearrangement. *Br J Cancer*. (2010) 103:82–9. doi: 10.1038/sj.bjc.6605746
87. Wang Z-Q, Bachvarova M, Morin C, Plante M, Gregoire J, Renaud M-C, et al. Role of the polypeptide N-acetylgalactosaminyltransferase 3 in ovarian cancer progression: possible implications in abnormal mucin O-glycosylation. *Oncotarget*. (2014) 5:544–60. doi: 10.18632/oncotarget.1652
88. Wend P, Runke S, Wend K, Anchondo B, Yesayan M, Jardon M, et al. WNT10B/ β -catenin signalling induces HMGA2 and proliferation in metastatic triple-negative breast cancer. *EMBO Mol Med*. (2013) 5:264–79. doi: 10.1002/emmm.201201320
89. Acebron SP, Karaulanov E, Berger BS, Huang Y-L, Niehrs C. Mitotic wnt signaling promotes protein stabilization and regulates cell size. *Mol Cell*. (2014) 54:663–74. doi: 10.1016/j.molcel.2014.04.014
90. Scrimieri F, Askew D, Corn DJ, Eid S, Bobanga ID, Bjelac JA, et al. Murine leukemia virus envelope gp70 is a shared biomarker for the high-sensitivity quantification of murine tumor burden. *Oncoimmunology*. (2013) 2:e26889. doi: 10.4161/onci.26889
91. Fink D, Aebi S, Howell SB. The role of DNA mismatch repair in drug resistance. *Clin Cancer Res*. (1998) 4:1–6.
92. Le DT, Uram JN, Wang H, Bartlett BR, Kemberling H, Eyring AD, et al. PD-1 blockade in tumors with mismatch-repair deficiency. *N Engl J Med*. (2015) 372:2509–20. doi: 10.1056/NEJMoa1500596
93. Diouf B, Cheng Q, Krynetskaia NF, Yang W, Cheok M, Pei D, et al. Somatic deletions of genes regulating MSH2 protein stability cause DNA mismatch repair deficiency and drug resistance in human leukemia cells. *Nat Med*. (2011) 17:1298–303. doi: 10.1038/nm.2430
94. Boegel S, Löwer M, Bukur T, Sorn P, Castle JC, Sahin U. HLA and proteasome expression body map. *BMC Med Genomics*. (2018) 11:36. doi: 10.1186/s12920-018-0354-x
95. LeibundGut-Landmann S, Waldburger J-M, Krawczyk M, Otten LA, Suter T, Fontana A, et al. Mini-review: specificity and expression of CIITA, the master regulator of MHC class II genes. *Eur J Immunol*. (2004) 34:1513–25. doi: 10.1002/eji.200424964
96. Zhan T, Rindtorff N, Boutros M. Wnt signaling in cancer. *Oncogene*. (2017) 36:1461–73. doi: 10.1038/onc.2016.304
97. Lin SY, Xia W, Wang JC, Kwong KY, Spohn B, Wen Y, et al. Beta-catenin, a novel prognostic marker for breast cancer: Its roles in cyclin D1 expression and cancer progression. *Proc Natl Acad Sci USA*. (2000) 97:4262–6. doi: 10.1073/pnas.060025397
98. Ben-David U, Siranosian B, Ha G, Tang H, Oren Y, Hinohara K, et al. Genetic and transcriptional evolution alters cancer cell line drug response. *Nature*. (2018) 560:325–30. doi: 10.1038/s41586-018-0409-3
99. Lechner MG, Karimi SS, Barry-Holton K, Angell TE, Murphy KA, Church CH, et al. Immunogenicity of murine solid tumor models as a defining feature of *in vivo* behavior and response to immunotherapy. *J Immunother*. (2013) 36:477–89. doi: 10.1097/01.cji.0000436722.46675.4a
100. Cho HJ, Jung JI, Lim DY, Kwon GT, Her S, Park JH, et al. Bone marrow-derived, alternatively activated macrophages enhance solid tumor growth and lung metastasis of mammary carcinoma cells in a Balb/C mouse orthotopic model. *Breast Cancer Res*. (2012) 14:R81. doi: 10.1186/bcr3195
101. Rego SL, Helms RS, Dréau D. Breast tumor cell TACE-shed MCSF promotes pro-angiogenic macrophages through NF- κ B signaling. *Angiogenesis*. (2014) 17:573–85. doi: 10.1007/s10456-013-9405-2
102. Madera L, Greenshields A, Coombs MRP, Hoskin DW. 4T1 Murine Mammary Carcinoma Cells Enhance Macrophage-Mediated Innate Inflammatory Responses. *PLoS ONE*. (2015) 10:e0133385. doi: 10.1371/journal.pone.0133385

Conflict of Interest: MV and US are employed by the company BioNTech SE.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Schrörs, Boegel, Albrecht, Bukur, Bukur, Holtsträter, Ritzel, Manninen, Tadmor, Vormehr, Sahin and Löwer. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Advantages of publishing in Frontiers



OPEN ACCESS

Articles are free to read
for greatest visibility
and readership



FAST PUBLICATION

Around 90 days
from submission
to decision



HIGH QUALITY PEER-REVIEW

Rigorous, collaborative,
and constructive
peer-review



TRANSPARENT PEER-REVIEW

Editors and reviewers
acknowledged by name
on published articles

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org

Contact us: info@frontiersin.org | +41 21 510 17 00



REPRODUCIBILITY OF RESEARCH

Support open data
and methods to enhance
research reproducibility



DIGITAL PUBLISHING

Articles designed
for optimal readership
across devices



FOLLOW US

[@frontiersin](https://twitter.com/frontiersin)



IMPACT METRICS

Advanced article metrics
track visibility across
digital media



EXTENSIVE PROMOTION

Marketing
and promotion
of impactful research



LOOP RESEARCH NETWORK

Our network
increases your
article's readership