BIOINFORMATICS ANALYSIS OF SINGLE CELL SEQUENCING DATA AND APPLICATIONS IN PRECISION MEDICINE

EDITED BY: Jialiang Yang, Liao Bo, Tuo Zhang and Yifei Xu

PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88963-528-3
DOI 10.3389/978-2-88963-528-3

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

BIOINFORMATICS ANALYSIS OF SINGLE CELL SEQUENCING DATA AND APPLICATIONS IN PRECISION MEDICINE

Topic Editors:
Jialiang Yang, Geneis (Beijing) Co. Ltd, China
Liao Bo, Hainan Normal University, China
Tuo Zhang, Cornell University, United States
Yifei Xu, University of Oxford, United Kingdom

Citation: Yang, J., Bo, L., Zhang, T., Xu, Y., eds. (2020). Bioinformatics Analysis of Single Cell Sequencing Data and Applications in Precision Medicine. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88963-528-3

Table of Contents

frontiers in Genetics

EDITORIAL
published: 23 January 2020
doi: 10.3389/fgene.2019.01358

Editorial: Bioinformatics Analysis of Single Cell Sequencing Data and Applications in Precision Medicine

Jialiang Yang[1,2*], *Bo Liao*[1], *Tuo Zhang*[3] *and Yifei Xu*[4]

[1] School of Mathematics and Statistics, Hainan Normal University, Haikou, China, [2] Department of Sciences, Geneis Beijing Co., Ltd., Beijing, China, [3] Department of Microbiology and Immunology, Weill Cornell Medicine, New York, NY, United States, [4] Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom

Keywords: single cell RNA sequencing, bioinformatics, precision medicine, cell cluster, trajectory analyses

Editorial on the Research Topic

Bioinformatics Analysis of Single Cell Sequencing Data and Applications in Precision Medicine

Next-generation sequencing (NGS) technology has been successfully applied in disease diagnostics, oncological immunotherapy, and drug repurposing, especially for precision medicine where optimized medication is tailored to individual patients. Recently, the development of single cell techniques makes it possible to examine gene expression and mutation at individual cell resolution, which provides an unprecedented opportunity to study cell development and differentiation, and reveal cell-to-cell heterogeneity during disease development, treatment, and drug response for individual patients. With the exponential increase of single cell sequencing data, it is critical to develop appropriate bioinformatics and machine learning tools to mine the rules behind them. However, due to the technical barriers in single cell sequencing and the noisy nature of raw sequencing data, this task is challenging especially in the context of disease diagnosis and drug development.

To promote the translation and efficient usage of single cell sequencing data to precision medicine, it is necessary to develop new analysis tools for analyzing and integrating multi-level single cell data including DNA, RNA, protein, and so on, comparing existing methods and results derived from different studies, and enhancing disease diagnostics and drug development. For example, the quality control, normalization, differential gene calling, and clustering methods are quite different between single cell sequencing and traditional bulk cell sequencing. Thus, it is critical to develop a best practice specifically for dealing with single cell sequencing data. For disease treatment, it is also important to identify disease driver genes common to all cell types as well as those specific to a particular cell type or subgroup as revealed by single cell techniques, based on existing or novel network and machine learning-based methods. Finally, more translational work should be done to bridge the bioinformatics analyses and clinical applications for single cell researchers.

To provide a platform bridging single cell analysis and translational studies, we organized this special issue, in which 11 manuscripts have been accepted for publication. Firstly, Zen and Dai presented a comprehensive review on scRNA-seq associated biological experiments as well as computational methods for evaluating disease heterogeneity. They described the early impact of such technologies as well as a variety of common methods applicable to upstream and downstream processes. Upstream processes include several computational methods related to the detection and removal of technical noise given commonly assumed statistical distributions. In addition, the

Edited and reviewed by:
Richard D. Emes, University of Nottingham, United Kingdom

***Correspondence:**
Jialiang Yang
yangjl@geneis.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *10 November 2019*
Accepted: *11 December 2019*
Published: *23 January 2020*

Citation:
Yang J, Liao B, Zhang T and Xu Y (2020) Editorial: Bioinformatics Analysis of Single Cell Sequencing Data and Applications in Precision Medicine. Front. Genet. 10:1358. doi: 10.3389/fgene.2019.01358

authors overviewed the recent adoption of methods to combat the statistical effects of batched experiments and zero-inflated data. Downstream processes include methods to integrate transcriptomic information with several other types of data such as epigenetic or spatial. They also introduced several clustering and pseudotemporal ordering methods. Finally, the authors conducted a small study comparing a handful of clustering and pseudotemporal analysis methods on four marginally related datasets due to their significance to disease systems.

Z. Wang et al. introduced a newly built database, SCDevDB, which provides the analysis results of single-cell gene expression profiles in different human developmental processes. This database mainly contains the gene expression profiles across 35 development stages as well as the differential gene analysis for 24 developmental pathways.

The manuscript by Finnegan et al. utilized single-cell RNA-seq data from 22,338 human foreskin keratinocytes to study transcription factor networks during the keratinocyte transition from the basal to the differentiated state. Their analysis uncovered novel players and novel roles of transcription factors in the intricate orchestration of keratinocyte differentiation and shed lights in elucidating disease and cancer processes.

Wu et al. designed a framework for evaluating 14 commonly used gene expression normalization methods, achieving consistency in the evaluation results using both bulk RNA-seq and scRNA-seq data. This framework was implemented as R package for researchers to choose the best normalization method.

X. Wang et al. identified 91 ethylene-responsive factors (ERFs) in *F. vesca*, based on which they provided evolutionary analysis, expansion analysis and expression analysis, especially for the influences of tandem duplication mechanism on expansion of ERF gene family.

Yin et al. focused on identification of novel breast cancer predisposition genes, which is of great significance in understanding the pathogenesis of breast cancer. The authors reanalyzed published whole exon sequencing data to screen susceptible genes, followed with experimental and functional validation. The most striking finding in the article is the discovery of *NCK1* as a novel breast cancer gene and the authors successfully correlated its expression and function with carcinogenesis.

Bope et al. provides a comprehensive review of genomic resources that have been established with respect to African individuals and their genomic data. This review presents an interesting perspective, and road map concerning the developments and studies that are needed in order to complement and promote efforts related to implementing Clinical Genomics in Africa.

Mercatelli et al. conducted a pan-cancer analysis to investigate the predictive power of gene expression on somatic mutations and copy number variations. They showed that genomic alterations could be modeled by gene expression across several human cancers using machine learning algorithms, and single-cell sequencing data can increase the performance of the model.

Chen et al. investigated the gene expression profiles of patient-derived tumor xenograft (PDX) models originated from eight tissues using machine learning algorithms, and showed that the specificity of primary tumor site was preserved in PDX models.

Cheng et al. provided a comprehensive review of recently technologies and literature of human microbiome. Firstly, the technologies producing the microbiome big data were reviewed. Secondly, the connections of the microbiota with different host organs were discussed. After that, the association of microbiota with the clinical medicine was discussed, with a special focus on a few major microbiota-associated diseases. Lastly, the future research trends were proposed.

Liu et al. conducted an integrated bioinformatics analysis on the public epithelial ovarian cancer (EOC) data collected from GEO; they identified potential biomarkers for evaluating EOC prognosis and bioactivate compounds for EOC treatment. Their study provides an example of bioinformatics analysis in promoting cancer research.

AUTHOR CONTRIBUTIONS

JY, BL, TZ and YX organized this special issue and wrote the editorial. All authors have approved the final version of the editorial.

ACKNOWLEDGMENTS

We thank the authors for contributing their valuable work to this special issue and the reviewers for their constructive comments. We are also grateful to the editorial board for approving this topic and hope this issue will advance the research on single cell RNA-seq analysis and its applications.

Conflict of Interest: JY is currently employed by Geneis Beijing Co. Ltd.

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.

frontiers in Genetics

METHODS
published: 30 April 2019
doi: 10.3389/fgene.2019.00400

NormExpression: An R Package to Normalize Gene Expression Data Using Evaluated Methods

Zhenfeng Wu[1,2†], Weixiang Liu[3†], Xiufeng Jin[2], Haishuo Ji[2], Hua Wang[2], Gustavo Glusman[4], Max Robinson[4], Lin Liu[2], Jishou Ruan[1] and Shan Gao[2*]*

[1] School of Mathematical Sciences, Nankai University, Tianjin, China, [2] College of Life Sciences, Nankai University, Tianjin, China, [3] School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen, China, [4] Institute for Systems Biology, Washington, DC, United States

OPEN ACCESS

Edited by:
Tuo Zhang,
Cornell University, United States

Reviewed by:
Yudong Cai,
Shanghai University, China
Naibin Duan,
Shandong Academy of Agricultural Sciences, China

***Correspondence:**
Jishou Ruan
jsruan@nankai.edu.cn
Shan Gao
gao_shan@mail.nankai.edu.cn

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *24 December 2018*
Accepted: *12 April 2019*
Published: *30 April 2019*

Citation:
Wu Z, Liu W, Jin X, Ji H, Wang H, Glusman G, Robinson M, Liu L, Ruan J and Gao S (2019) NormExpression: An R Package to Normalize Gene Expression Data Using Evaluated Methods. Front. Genet. 10:400. doi: 10.3389/fgene.2019.00400

Data normalization is a crucial step in the gene expression analysis as it ensures the validity of its downstream analyses. Although many metrics have been designed to evaluate the existing normalization methods, different metrics or different datasets by the same metric yield inconsistent results, particularly for the single-cell RNA sequencing (scRNA-seq) data. The worst situations could be that one method evaluated as the best by one metric is evaluated as the poorest by another metric, or one method evaluated as the best using one dataset is evaluated as the poorest using another dataset. Here raises an open question: principles need to be established to guide the evaluation of normalization methods. In this study, we propose a principle that one normalization method evaluated as the best by one metric should also be evaluated as the best by another metric (the consistency of metrics) and one method evaluated as the best using scRNA-seq data should also be evaluated as the best using bulk RNA-seq data or microarray data (the consistency of datasets). Then, we designed a new metric named Area Under normalized CV threshold Curve (AUCVC) and applied it with another metric mSCC to evaluate 14 commonly used normalization methods using both scRNA-seq data and bulk RNA-seq data, satisfying the consistency of metrics and the consistency of datasets. Our findings paved the way to guide future studies in the normalization of gene expression data with its evaluation. The raw gene expression data, normalization methods, and evaluation metrics used in this study have been included in an R package named NormExpression. NormExpression provides a framework and a fast and simple way for researchers to select the best method for the normalization of their gene expression data based on the evaluation of different methods (particularly some data-driven methods or their own methods) in the principle of the consistency of metrics and the consistency of datasets.

Keywords: gene expression, normalization, evaluation, R package, scRNA-seq

INTRODUCTION

Global gene expression analysis provides quantitative information about the population of RNA species in cells and tissues (Lovén et al., 2012). High-throughput technologies to measure global gene expression levels started with Serial Analysis of Gene Expression (SAGE) and are widely used with microarray and RNA-seq (Gao et al., 2014). Recently, single-cell RNA sequencing (scRNA-seq) has been used to simultaneously measure the expression levels of genes from a single cell, providing a higher resolution of cellular differences than what can be achieved by bulk RNA-seq, which can only produce an expression value for each gene by averaging its expression levels across a large population of cells (Gao, 2018). Raw gene expression data from these high-throughput technologies must be normalized to remove technical variation so that meaningful biological comparisons can be made. Data normalization is a crucial step in the gene expression analysis as it ensures the validity of its downstream analyses (Lovén et al., 2012). The differential expression analysis or the co-expression analysis using the same dataset could produce significant different genes using different data normalization methods. Although the significance of data normalization in the gene expression analysis has been demonstrated (Bullard et al., 2010), how to select a successful normalization method is still an open question, particularly for scRNA-seq data.

Basically, two classes of methods are available to normalize gene expression data using global normalization factors. They are the control-based normalization and the average-bulk normalization. The former class of methods assumes the total expression level summed over a pre-specified group of genes is approximately the same across all the samples. The latter class of methods assumes most genes are not significantly Differentially Expressed (DE) across all the samples. The control-based normalization often uses RNA from a group of internal control genes (e.g., housekeeping genes) or external spike-in RNA [e.g., ERCC RNA (Jiang et al., 2011)], while the average-bulk normalization is more commonly used for their universality. Five average-bulk normalization methods designed to normalize bulk RNA-seq data are library size, median of the ratios of observed counts that is also referred to as DESeq (Anders and Huber, 2010), Relative Log Expression (RLE), upper quartile (UQ), and Trimmed Mean of M values (TMM) (Robinson et al., 2010). Recently, three new methods were introduced as Total Ubiquitous (TU), Network Centrality Scaling (NCS), and Evolution Strategy (ES) with the best performance among 15 tested methods (Glusman et al., 2013). To improve scRNA-seq data normalization, Lun et al. (2016) introduced a new method using the pooled size factors (Pooled) and claimed that their method outperformed the library size method, DESeq and TMM. Bacher et al. (2017) addressed that using existing normalization methods on scRNA-seq data introduced artifacts that bias downstream analyses. Then, another new method SCnorm was introduced and claimed to outperform MR, Transcripts Per Million (TPM), scran, SCDE, and BASiCS using both simulated and case study data (Bacher et al., 2017).

Although many metrics have been designed to evaluate the relative success of these methods, different metrics or different datasets yield inconsistent evaluation results. Here raises another open question: principles need to be established to guide the evaluation of normalization methods. Glusman et al. (2013) proposed that a successful normalization method should simultaneously maximize the number of uniform genes and minimize the correlation between the expression profiles of gene pairs. Based on this criterion, they presented two novel and mutually independent metrics to evaluate 15 normalization methods and achieved consistent results using bulk RNA-seq data (Glusman et al., 2013). In this study, we designed a new metric named Area Under normalized CV threshold Curve (AUCVC) and applied it with another metric mSCC (see section "Materials and Methods") to evaluate 14 commonly used normalization methods using both scRNA-seq and bulk RNA-seq data from the same library construction protocol. The evaluation results by both AUCVC and mSCC achieved consistency. In addition, the evaluation results using both scRNA-seq and bulk RNA-seq data also achieved consistency. So, we propose a principle that one normalization method evaluated as the best by one metric should also be evaluated as the best by another metric (the consistency of metrics) and one method evaluated as the best using one dataset should also be evaluated as the best using another dataset (the consistency of datasets). The datasets using different protocols (RNA-seq, scRNA-seq, or microarray) need to be used to validate the consistency, which is beyond the scope of this study. As many new normalization methods are being developed, researchers need a fast and simple way to evaluate different methods, particularly some data-driven methods or their own methods, rather than obtain information from published evaluation results, which could have biases or mistakes, e.g., misunderstanding of RLE, UQ and TMM (see section "Results"). To satisfy this demand, we developed an R package NormExpression including the raw gene expression data, normalization methods and evaluation metrics used in this study. This tool provides a framework for researchers to select the best method for the normalization of their gene expression data based on the evaluation of different methods in the principle proposed in this study.

RESULTS

Basic Concepts

In total, 14 normalization methods have been evaluated in this study. They are Housekeeping Genes (HG7), External RNA Control Consortium (ERCC), Total Read Number (TN), Total Read Count (TC), Cellular RNA (CR), Nuclear RNA (NR), median of the ratios of observed counts (DESeq), Relative Log Expression (RLE), UQ, Trimmed Mean of M values (TMM), Total Ubiquitous (TU), Network Centrality Scaling (NCS), Evolution Strategy (ES), and SCnorm (see section "Materials and Methods"). Currently, most methods with a few exceptions (e.g., SCnorm) are used to normalize a raw gene expression matrix (n samples by m genes) by multiplying a global normalization factor to each of its columns, yielding a normalized

gene expression matrix (**Figure 1A**). In different methods, the definitions of normalization factor, scaling factor and size factor are inconsistent and need to be explained here. Both the normalization factor defined in the package NormExpression and the scaling factor defined in a previous study (Glusman et al., 2013) are the global normalization factors (**Figure 1A**). As the library size methods, TN, TC, CR, or NR can be used to estimate a library size, which represents the amount of total RNA in a cDNA library from a sample. HG7, ERCC, DESeq, TU, NCS, and ES produce a pseudo library size (in **Figure 1B**), which represents the relative amount of total RNA. Library size is also named as size factor in the Bioconductor package DESeq (Anders and Huber, 2010). In general, HG7, ERCC, TN, TC, CR, NR, DESeq, TU, NCS, and ES produce the global normalization factor by the reciprocal of library size or pseudo library size. RLE, UQ, and TMM in the Bioconductor package edgeR (Robinson et al., 2010) produce normalization factors to normalize the library sizes and the global normalization factors for data normalization should be calculated by one million multiplying the reciprocal of normalized library sizes (**Figure 1B**). However, the normalization factors produced by RLE, UQ, and TMM have been wrongly used as the global normalization factors in previous studies (Li et al., 2015). The NormExpression package includes such modifications as below to integrate the above normalization methods. DESeq, RLE, UQ, and TMM have been modified to ignore zero values to be fit for the scRNA-seq data processing. As NR is the best among the library size methods (TN, TC, CR, and NR), RLE, UQ, and TMM use NR to estimate library sizes. As HG7 and ERCC produce pseudo library sizes (**Figure 1B**) as TN, TC, CR, and NR, their normalization factors are amplified by one million for a uniform representation (**Figure 1B**). The resulting normalization factors of all 14 methods except SCnorm need to be further normalized by their geometric mean values (**Figure 1B**). After further normalization, RLE is identical to DESeq and presented as DESeq (RLE) or DESeq* in this study. It has been confirmed that all the modifications do not change the evaluation results.

A

$$\begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{in} \\ \vdots & & \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mj} & \cdots & x_{mn} \end{bmatrix} \times \begin{bmatrix} f_1 \\ \vdots \\ f_j \\ \vdots \\ f_n \end{bmatrix}$$

B

$$For\ all\ methods: \quad f_j = f_j \Big/ exp\left[\frac{1}{n}\sum_{j=1}^{n} log\left(f_j\right)\right]$$

$$Library\ size: f_j = 10^6 / N_j$$

$$DESeq: f_j = 1/N_j^*$$

$$N_j^* = exp\left[median_i\left(log\left(x_{ij}\right) - \frac{1}{n}\sum_{k=1}^{n} log\left(x_{ik}\right)\right)\right]$$

$$RLE: f_j = 10^6 / N_j s_j$$

$$s_j = median_i\left[x_{ij} \Big/ exp\left(\frac{1}{n}\sum_{k=1}^{n} log\left(x_{ik}\right)\right)\right] \Big/ N_j$$

$$UQ: f_j = 10^6 / N_j s_j \qquad s_j = Q_{i\,75}\left(x_{ij}/N_j\right)$$

FIGURE 1 | Basic concepts. **(A)** A raw gene expression matrix can be transformed into a normalized gene expression matrix by the multiplication of a global factor f_j to each column. Each column represents the expression levels of all genes from a sample and each row represents the expression levels of a gene across all samples. **(B)** As the library size methods, TN, TC, CR, or NR can be used to estimate a library size N_j. The library size methods (TN, TC, CR, and NR) produce the global normalization factor f_j by the reciprocal of library size N_j. HG7, ERCC, DESeq, TU, NCS, and ES produce a pseudo library size N_j*, which represents the relative amount of total RNA. RLE, UQ, and TMM produce a normalization factor s_j to normalize the library size N_j and the global normalization factor for data normalization should be $10^6/N_j s_j$. Q_{75} represents the third quartile Q_3. For all methods, log represents the natural logarithm.

Evaluation of 14 Normalization Methods

In the previous study, Glusman et al. (2013) had quantified the success of normalization methods by the number of uniform genes (see section "Materials and Methods") and used the Coefficient of Variation (CV) cutoff 0.25 to determine the number of uniform genes for each method. This metric was designed based on the theory that the relative values among different normalization methods are quite stable, although the absolute number of uniform genes depend on the cutoff value. However, it is almost impossible to determine a CV cutoff for scRNA-seq data as CV in scRNA-seq data has a much larger dynamic range than in bulk RNA-seq data. Inspired by Area Under the receiver operating characteristic Curve (AUC) (Gao et al., 2009), we designed a new metric named Area Under normalized CV threshold Curve (AUCVC) to evaluate normalization methods. Using one scRNA-seq dataset scRNA663 and one bulk RNA-seq dataset bkRNA18 (see section "Materials and Methods"), we applied AUCVC and another metric mSCC (see section "Materials and Methods") to evaluate 14 normalization methods and then we compared the evaluation results by mSCC with those by AUCVC to assess the consistency of datasets and the consistency of metrics.

The non-zero ratio cutoffs (see section "Materials and Methods") from 0.2 to 0.9 for scRNA663 and from 0.7 to 1 for bkRNA18 were used to produce AUCVCs of all methods (**Figures 2A,B**). Among 14 methods, TU, NCS, and ES are parameter-dependent approaches, which use the occurrence rate, upper and lower cutoffs as three parameters (see section "Materials and Methods"). For each non-zero ratio cutoff, TU used the maximum AUCVC to determine the optimal ones by testing all possible combinations of three parameters. In addition, the calculation only considered each combination of three parameters which produced more than 100 ubiquitous genes (see section "Materials and Methods") for scRNA663 and

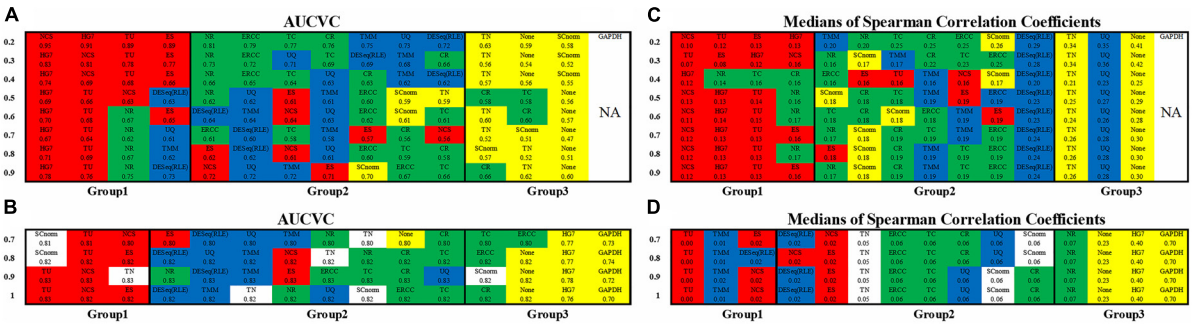


FIGURE 2 | Consistency of metrics and consistency of datasets. The non-zero ratio cutoffs from 0.2 to 0.9 for scRNA663 and from 0.7 to 1 for bkRNA18 were used to produce AUCVCs and mSCCs. All the normalization methods were classified into three groups based on their AUCVC values sorted in descending order (from the best to the poorest) using one scRNA-seq dataset scRNA663 **(A)** and one bulk RNA-seq dataset bkRNA18 **(B)**. These methods were also classified into three groups based on their mSCC values sorted in ascending order (from the best to the poorest) using one scRNA-seq dataset scRNA663 **(C)** and one bulk RNA-seq dataset bkRNA18 **(D)**. GAPDH is not applicable to scRNA-seq data due to zero counts of GAPDH present in many samples. All the numbers are accurate to two decimal places, the marginal differences are reflected by their orders. The raw gene expression data (None) was also used to produce AUCVCs and mSCCs for comparison. After further normalization, RLE is identical to DESeq and presented as DESeq (RLE) or DESeq* in this study.

more than 1,000 for bkRNA18. The occurrence rate cutoff was tested from 0.2 to 0.6 for scRNA663 at interval of 0.1 and set to 1 for bkRNA18. The lower cutoff was tested from 5 to 40% and the upper cutoff was tested from 60 to 95% at interval of 5%. For each non-zero ratio cutoff, NCS and ES used the occurrence rate, lower and upper cutoffs determined by TU, when TU achieved the maximum AUCVC.

The evaluation results using both scRNA663 and bkRNA18 showed consistency (the consistency of datasets) that all methods except HG7, TN and SCnorm were classified into three groups, based on their AUCVC values sorted in descending order (**Figures 2A,B**). The first group including TU, NCS and ES achieved the best performances. TU, NCS and ES, which had only been evaluated using bulk RNA-seq data in the previous study (Glusman et al., 2013), were evaluated by our new metric AUCVC as the best normalization methods using both scRNA-seq and bulk RNA-seq data. The second group including ERCC, TC, CR, NR, DESeq, RLE, UQ, and TMM achieved medial performances using both scRNA663 and bkRNA18. In the second group, ERCC, TC, CR, and NR outperformed DESeq, RLE, UQ, and TMM using scRNA663 but underperformed them using bkRNA18. The third group achieved the poorest performances, including TN, SCnorm and None (the raw gene expression data) for scRNA663 (**Figure 2A**), and HG7, GAPDH and None for bkRNA18 (**Figure 2B**). The evaluation results of HG7, TN, and SCnorm did not achieve the consistency using scRNA663 and bkRNA18. HG7 and GAPDH achieved the poorest performances using bkRNA18, suggesting that a predefined set of housekeeping genes may not be appropriate guides for the normalization of bulk RNA-seq data. However, it could be coincidental that HG7 was classified into the first group using scRNA663. TN underperformed the second group of methods using scRNA663 but outperformed it using bkRNA18. SCnorm was designed to improve the normalization of scRNA-seq data but it performed poorer using scRNA-seq data than bulk RNA-seq data. Particularly, SCnorm ranked the first in the best group by its AUCVC to normalize bkRNA18 when the non-zero ratio cutoffs were set to 0.7 or 0.8 (**Figure 2B**), but ranked the last in the poorest group to normalize scRNA663 when the non-zero ratio cutoffs were set to 0.2–0.4 (**Figure 2A**). SCnorm claimed that it is not designed to process datasets containing more than 80% zero counts. However, scRNA663 was build using the Smart-seq2 scRNA-seq protocol, which contained the least zero counts among current scRNA-seq protocols.

The evaluation results (**Figures 2C,D**) by mSCC were consistent with those by AUCVC (the consistency of metrics). This proved that a successful normalization method simultaneously maximizes the number of uniform genes and minimizes the correlation between the expression profiles of gene pairs. We selected the best evaluation results using scRNA-seq data (the none-zero ratio cutoff = 0.2) and bulk RNA-seq data (the none-zero ratio cutoff = 1) in **Figure 2** for visualization using NormExpression (**Figure 3**). Then, we calculated the Spearman's rank Correlation Coefficients (SCCs) between all the normalization factors except that using SCnorm. Using 1-SCCs as distances, hierarchical clustering of 13 normalization factors showed equivalent classification into the same groups (**Figures 3E,F**) as those by AUCVC and by mSCC (**Figure 2**). From **Figure 3**, it can be seen that while all normalization methods except HG7, GAPDH and None performed without much differences using bulk RNA-seq data (**Figures 3B,D**), they had significant differences in the performances using scRNA-seq data (**Figures 3A,C**). These differences suggest that although scRNA-seq provides a higher resolution of cellular differences, it is more challenging to select the best method for the normalization of scRNA-seq data. In addition, these differences provided an explanation as to why scRNA-seq data and bulk RNA-seq data from the same samples resulted in different results in many previous studies. From **Figure 3**, it also can be seen that the best mSCC (TU-normalized) using scRNA-seq data still had a certain distance from 0 (**Figure 3C**), while the best mSCC (TU-normalized) using bulk RNA-seq data was close to 0 (**Figure 3D**). Therefore, further studies need to be conducted to investigate whether we can obtain the best mSCC using scRNA-seq data close to 0 as that using bulk RNA-seq

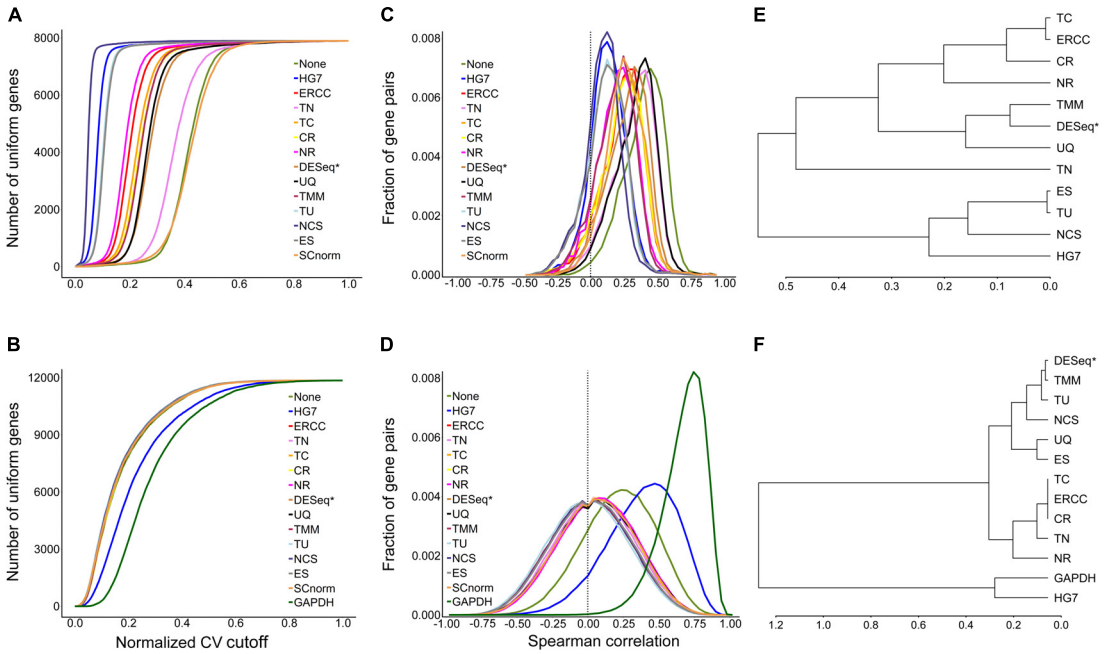


FIGURE 3 | Visualization of evaluation results. A normalization method with a higher AUCVC produced a lower median of Spearman's rank Correlation Coefficient (mSCC) between the normalized expression profiles of ubiquitous gene pairs using both scRNA-seq **(A,C)** and bulk RNA-seq data **(B,D)**. Using 1-SCCs as distances, hierarchical clustering of 13 normalization factors showed equivalent classification into the same groups **(E,F)** as those by AUCVC and by mSCC (**Figure 2**). SCnorm was not applicable to be used to calculate SCCs, as it produced a factor matrix rather than a factor vector as 13 other methods. GAPDH is not applicable to scRNA-seq data due to zero counts of GAPDH present in many samples. The raw gene expression data (None) was also used to produce AUCVCs and mSCCs for comparison. After further normalization, RLE is identical to DESeq and presented as DESeq (RLE) or DESeq* in this study.

data. If we cannot, what is the reason? And is it the nature of scRNA-seq data that result in this bias from 0?

To further test our principle, we searched other performance metrics in the published papers. The Bioconductor package scone (Cole et al., 2018) provides eight metrics to evaluate the normalization methods using scRNA-seq data. Among eight metrics, three are based on clustering properties and three other metrics are associated with control genes or QC metrics. Only two metrics based on global distributional properties can be used as general metrics. These two metrics are named as mean squared median relative log-expression (RLE_MED) and variance of inter-quartile range (IQR) of RLE (RLE_IQR). The evaluation results (**Supplementary File 1**) of three groups (particularly TU and ES) by RLE_MED were consistent with those by mSCC and by AUCVC using both scRNA-seq and bulk RNA-seq data. However, the evaluation results (**Supplementary File 1**) by RLE_IQR were not consistent with those by mSCC and by AUCVC. This suggests that mSCC, AUCVC, RLE_MED can be used together for method evaluation to test the consistency of metrics.

Implementation and Availability

The raw gene expression data, normalization methods (except NCS, ES and SCnorm) and evaluation metrics (AUCVC and mSCC) have been included in the R package NormExpression. The data process in this study is provided in detail (**Supplementary File 1**). All the methods except NCS and ES have been implemented in R programs for their running on R platforms of any version. DESeq uses an R program from the Bioconductor package DESeq (Anders and Huber, 2010), which has been modified to process scRNA-seq data. RLE, UQ and TMM use R programs from the Bioconductor package edgeR (Robinson et al., 2010), which have been modified to process scRNA-seq data. NCS and ES had been implemented in Perl programs with multiple dependencies on Perl modules (Glusman et al., 2013), which have been modified into stand-alone programs for Linux systems (**Supplementary File 2**). SCnorm uses the Bioconductor package SCnorm (Bacher et al., 2017).

NormExpression can be used in three modes: normalization without evaluation, normalization with simple evaluation or normalization with complete evaluation. In the first mode, TU is recommended for the normalization of gene expression data, as it has been already ranked as the best method for both scRNA-seq and bulk RNA-seq data. In the second mode, AUCVC is used to select the best method from 10 normalization methods, which are HG7, ERCC (if available), TN, TC, CR, NR, DESeq (RLE), UQ, and TMM. TN, NCS, ES, and SCnorm are not used in the second mode, as the evaluation results of TN and SCnorm cannot achieve consistency, and NCS and ES have similar performances to TU but are much more time consuming. In the third mode, AUCVC and mSCC are used to select the best method from TU and at least 10 normalization methods. The normalization with simple evaluation determines the best method based on AUCVC values, while the normalization with complete evaluation determines the

best method in the principle of the consistency of metrics and the consistency of datasets. As a result of a complete evaluation, the tables of AUCVC and mSCC (**Figure 2**) are required for the method selection.

MATERIALS AND METHODS

Datasets

In a previous study (SRA: SRP113436), 831 single-cell samples and 18 bulk samples had been sequenced using the Smart-seq2 scRNA-seq protocol. In this study, we built a scRNA-seq dataset including 663 single cells from colon tumor tissues and 10 single cells from distal tissues (>10 cm) as control. The data of 166 single-cell samples were removed, as each of them contained NR less than 100,000 reads. The data of two single-cell samples were removed, as each of them contained simulated ERCC RNA less than 0 reads. We also built a bulk RNA-seq dataset including nine samples from colon tumor tissues and nine samples from distal tissues. The cleaning and quality control of both scRNA-seq and bulk RNA-seq data were performed using the pipeline Fastq_clean (Zhang et al., 2014) that was optimized to clean the raw reads from Illumina platforms. Using the software STAR (Dobin et al., 2013) v2.5.2b, we aligned all the cleaned scRNA-seq and bulk RNA-seq reads to the human genome GRCh38/hg38 and quantified the expression levels of 57,992 annotated genes (57,955 nuclear and 37 mitochondrial). Mitochondrial RNAs should have been, but were not discarded to test the robustness of normalization methods. Non-polyA RNAs and small RNAs (<200 bp) were not discarded either, although the Smart-seq2 protocol theoretically had only captured polyA RNAs. In addition, the expression levels of 92 ERCC RNAs and the long non-coding RNA (lncRNA) MDL1 in the human mitochondrial genome (Gao et al., 2017) were also quantified. ERCC RNA had been spiked into 208 single-cell samples before library construction; the expression levels of 92 ERCC RNAs in other 455 single-cell samples and 18 bulk samples were simulated by linear regression. Finally, the two datasets were named scRNA663 (58085 × 663) and bkRNA18 (58085 × 18), and used as raw gene expression data in this study. As these two datasets were obtained by sequencing the libraries using the same protocol and samples from the same group of patients, they had great values to be used to evaluate normalization methods and assess the consistency of datasets. Researchers can select the best method for the normalization of their gene expression data or evaluate different methods using the data of 57,955 nuclear genes.

Normalization Methods

The library size methods (TN, TC, CR, and NR) use the gene expression level summed over total genes in a sample as the library size to calculate the normalization factor. HG7, ERCC and TU use the gene expression level summed over these pre-selected genes in a sample as the pseudo library size (see section "Results"). NR only counts reads which can be aligned to nuclear genomes, while CR counts reads which can be aligned to both nuclear and mitochondrial genomes. TC counts reads which can be aligned to 92 ERCC RNAs, nuclear and mitochondrial genomes (TC = CR + ERCC). TN uses the number of all reads which can be aligned to 92 ERCC RNAs, nuclear and mitochondrial genomes. The pre-selected genes used by HG7, ERCC and TU are seven housekeeping genes, 92 ERCC RNAs and the ubiquitous genes (described below), respectively. Seven genes (UBC, HMBS, TBP, GAPDH, HPRT1, RPL13A, and ACTB) in HG7 had been used to achieve the best evaluation result among those using all possible combinations of tested housekeeping genes in the previous study by Glusman et al. (2013). ERCC RNA is a set of commonly used spike-in RNA consisting of 92 polyadenylated transcripts with short 3′ polyA tails but without 5′ caps (Jiang et al., 2011). A single housekeeping gene GAPDH was used for comparison in the evaluation of normalization methods using bulk RNA-seq data, but it was not applicable to scRNA-seq data due to zero counts of GAPDH present in many samples. The raw gene expression data (None) was also used to produce AUCVCs and mSCCs for comparison.

Uniform Genes and Ubiquitous Genes

A gene is defined as uniform when the Coefficient of Variation (CV, Formula 1) of its expression values across all samples is not more than a cutoff (Glusman et al., 2013). To determine the number of uniform genes using scRNA-seq data containing a high frequency of zeros, NormExpression only considers genes with non-zero ratios not less than a cutoff. The non-zero ratio of one gene should be calculated as the number of its all non-zero expression values divided by the number of total samples.

Ubiquitous genes are defined as the intersection of a trimmed sets of all samples (Glusman et al., 2013). This trimmed set of genes are selected for each sample by (1) excluding genes with zero values, (2) sorting the non-zero genes by their expression levels in that sample, and (3) removing the upper and lower ends of the sample-specific expression distribution. Glusman et al. (2013) determined the optimal parameters by testing all possible combinations of lower and upper cutoffs at interval of 5% to maximize the number of resulting uniform genes using one bulk RNA-seq dataset. The size of a scRNA-seq dataset is usually very large, which could result in a very small or even empty set of ubiquitous genes, as the number of ubiquitous genes depends on the sizes of datasets. To identify the ubiquitous genes using scRNA-seq data, we defined a parameter named occurrence rate, governing the minimal fraction of trimmed sets in which a gene must appear to be considered ubiquitous. In NormExpression, TU includes three parts. The first part determines the optimal parameters by testing all possible combinations of occurrence rate, lower and upper cutoffs to maximize AUCVC (described below) instead of the number of resulting uniform genes. The second part uses the optimal occurrence rate, upper and lower cutoffs to obtain the ubiquitous genes. The third part uses the ubiquitous genes to calculate the TU normalization factor. NormExpression only use the raw gene expression data to obtain the ubiquitous genes, which are used to calculate the TU normalization factor and to evaluate all methods. In addition, the same ubiquitous genes are used by NCS and ES to obtain the NCS and ES normalization factors, respectively. These ubiquitous genes are also used by TU, NCS and ES to produce their mSCCs for method evaluation.

AUCVC and mSCC

In the previous study, Glusman et al. (2013) designed two novel and mutually independent metrics, which were the number of uniform genes and Spearman's rank Correlation Coefficients (SCCs) between expression profiles of gene pairs. The basic theory underlying these two evaluation metrics is that a successful normalization method simultaneously maximizes the number of uniform genes and minimizes the correlation between the expression profiles of gene pairs. In this study, we designed a new metric AUCVC instead of the number of uniform genes and used the median of Spearman's rank Correlation Coefficients between the normalized expression profiles of ubiquitous gene pairs (mSCC) instead of observation of SCC distributions for method evaluation. On default settings, NormExpression randomly selected 1,000,000 ubiquitous gene pairs to calculate the mSCCs for method evaluation (**Figures 3C,D**).

AUCVC (**Figures 3A,B**) is created by plotting the number of uniform genes (y-axis) at each normalized CV (Formula 2) cutoff (x-axis). As a high or a low normalized CV cutoff produces more false or less true uniform genes, it is reasonable to consider the overall performance of each method at various cutoff settings instead of that at one specific cutoff setting. In Formula 1 and 2, symbols have the same meanings as those in **Figure 1** and n* does not count zero elements in each sample.

$$CV_i = \left(\sqrt{\frac{1}{n-1} \sum_{j=1}^{n} \left(x_{ij} - \bar{x}_i \right)^2} \right) \Big/ \bar{x}_i, \bar{x}_i = \frac{1}{n} \sum_{j=1}^{n} x_{ij} \quad (1)$$

$$\begin{aligned} &Normalized\ CV_i = \left\{ CV_i - \min_i \left(CV_i \right) \right\} \Big/ \left\{ \max_i \left(CV_i \right) - \min_i \left(CV_i \right) \right\} \\ &CV_i = \left(\sqrt{\frac{1}{n^*-1} \sum_{j=1}^{n^*} \left(log_2 \left(x_{ij} \right) - \bar{x}_i \right)^2} \right) \Big/ \bar{x}_i, \ \ \bar{x}_i = \frac{1}{n^*} \sum_{j=1}^{n^*} log_2 \left(x_{ij} \right), \\ &x_{ij} > 0 \end{aligned} \quad (2)$$

AUTHOR CONTRIBUTIONS

SG conceived this project. SG and JR supervised this project. ZW, WL, and XJ performed the programming. ZW and HJ analyzed the data. HW prepared all the figures and **Supplementary Files**. SG drafted the main manuscript. GG, MR, and LL revised the manuscript. All authors read and approved the manuscript.

FUNDING

This work was supported by grants from National Key Research and Development Program of China (2016YFC0502304-03) to Defu Chen, National Natural Science Foundation of China (11701296) to Jianzhao Gao, and Internationalization of Outstanding Postdoctoral Training Program from Tianjin Government to SG.

ACKNOWLEDGMENTS

We thank Professor Sui Huang and Dr. Joseph Zhou from Institute for Systems Biology (ISB) their hosts on SG' visiting to ISB. This manuscript has been released as a preprint (Wu et al., 2018).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00400/full#supplementary-material

FILE S1 | How to use NormExpression.

FILE S2 | Perl scripts to run NCS and ES.

REFERENCES

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11:R106. doi: 10.1186/gb-2010-11-10-r106

Bacher, R., Chu, L., Leng, N., Gasch, A. P., Thomson, J. A., Stewart, R. M., et al. (2017). SCnorm: robust normalization of single-cell RNA-seq data. *Nat. Methods* 14:584. doi: 10.1038/nmeth.4263

Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 11:94. doi: 10.1186/1471-2105-11-94

Cole, M. B., Risso, D., Wagner, A., DeTomaso, D., Ngai, J., Purdom, E., et al. (2018). Performance assessment and selection of normalization procedures for single-cell RNA-seq. *bioRxiv* [Preprint]. doi: 10.1101/235382

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi: 10.1093/bioinformatics/bts635

Gao, S. (2018). Data analysis in single-cell transcriptome sequencing. *Methods Mol. Biol.* 1754:18.

Gao, S., Ou, J., and Xiao, K. R. (2014). *language and Bioconductor in Bioinformatics Applications(Chinese Edition)*. Tianjin: Tianjin Science and Technology Translation Publishing Ltd.

Gao, S., Tian, X., Chang, H., Sun, Y., Wu, Z., Cheng, Z., et al. (2017). Two novel lncRNAs discovered in human mitochondrial DNA using PacBio full-length transcriptome data. *Mitochondrion* 38, 41–47. doi: 10.1016/j.mito.2017.08.002

Gao, S., Zhang, N., Duan, G. Y., Yang, Z., Ruan, J. S., and Zhang, T. (2009). Prediction of function changes associated with single-point protein mutations using support vector machines (SVMs). *Hum. Mutat.* 30, 1161–1166. doi: 10.1002/humu.21039

Glusman, G., Caballero, J., Robinson, M., Kutlu, B., and Hood, L. (2013). Optimal scaling of digital transcriptomes. *PLoS One* 8:e77885. doi: 10.1371/journal.pone.0077885

Jiang, L., Schlesinger, F., Davis, C. A., Zhang, Y., Li, R., Salit, M., et al. (2011). Synthetic spike-in standards for RNA-seq experiments. *Genome Res.* 21:1543. doi: 10.1101/gr.121095.111

Li, P., Piao, Y., Shon, H. S., and Ryu, K. H. (2015). Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data. *BMC Bioinformatics* 16:347. doi: 10.1186/s12859-015-0778-7

Lovén, J., Orlando, D. A., Sigova, A. A., Lin, C. Y., Rahl, P. B., Burge, C. B., et al. (2012). Revisiting global gene expression analysis. *Cell* 151, 476–482. doi: 10.1016/j.cell.2012.10.012

Lun, A. T., Karsten, B., and Marioni, J. C. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* 17:75. doi: 10.1186/s13059-016-0947-7

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Wu, Z., Liu, W., Jin, X., Yu, D., Wang, H., Glusman, G., et al. (2018). NormExpression: an R package to normalize gene expression data using evaluated methods. *bioRxiv* [Preprint]. doi: 10.1101/251140

Zhang, M., Zhan, F., Sun, H., Gong, X., Fei, Z., et al. (2014). "Fastq_clean: an optimized pipeline to clean the Illumina sequencing data with quality control," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, (Piscataway, NJ: IEEE).

Conflict of Interest Statement: 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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 04 June 2019
doi: 10.3389/fgene.2019.00527

Revisiting Non-BRCA1/2 Familial Whole Exome Sequencing Datasets Implicates *NCK1* as a Cancer Gene

*Jie Yin†, Kai Wu†, Qingyang Ma, Hang Dong, Yufei Zhu, Landian Hu and Xiangyin Kong**

State Key Laboratory of Medical Genomics, Institute of Health Sciences, Shanghai Jiao Tong University School of Medicine and Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China

Through linkage and candidate gene screening, many breast cancer (BC) predisposition genes have been identified in the past 20 years. However, the majority of genetic risks that contribute to familial BC remains undetermined. In this study, we revisited whole exome sequencing datasets from non-BRCA1/2 familial BC patients, to search for novel BC predisposition genes. Based on the infinite mutation model, we supposed that rare non-silent variants that cooccurred between familial and TCGA-germline datasets, might play a predisposition contributing role. In our analysis, we not only identified novel potential pathogenic variants from known cancer predisposition genes, such as *MRE11*, *CTR9* but also identified novel candidate predisposition genes, such as *NCK1*. According to the TCGA mRNA expression dataset of BC, *NCK1* was significantly upregulated in basal-like subtypes and downregulated in luminal subtypes. *In vitro*, NCK1 mutants (D73H and R42Q) transfected MCF7 cell lines, which attributed to the luminal subtype, were much more viable and invasive than the wild type. On the other side, our results also showed that overall survival and disease-free survival of patients with *NCK1* variations might be dependent on the genomic context. In conclusion, genetic heterogeneity exists among non-BRCA1/2 BC pedigrees and *NCK1* could be a novel BC predisposition gene.

Keywords: breast cancer, non-BRCA1/2, *NCK1*, predisposition gene, invasion

OPEN ACCESS

Edited by:
Jialiang Yang,
Icahn School of Medicine at Mount Sinai, United States

Reviewed by:
Quan Zou,
University of Electronic Science and Technology of China, China
Ian Campbell,
Peter MacCallum Cancer Centre, Australia
Guang Wu,
Guangxi Academy of Sciences, China

***Correspondence:**
Xiangyin Kong
xykong@sibs.ac.cn

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *03 February 2019*
Accepted: *14 May 2019*
Published: *04 June 2019*

Citation:
Yin J, Wu K, Ma Q, Dong H, Zhu Y, Hu L and Kong X (2019) Revisiting Non-BRCA1/2 Familial Whole Exome Sequencing Datasets Implicates NCK1 as a Cancer Gene. Front. Genet. 10:527. doi: 10.3389/fgene.2019.00527

INTRODUCTION

Breast cancer (BC) is the most malignant cancer type, affecting women worldwide (30%) and is the secondary cause of death in women (14%) (Siegel et al., 2018). Although most BC patients are sporadic, about 10–15% of BC s show familial aggregation (Kiiski et al., 2014; Lynch et al., 2015). High penetrance genes, such as *BRCA1* and *BRCA2*, contribute about 20% to the etiology of familial BC (Mavaddat et al., 2010; Rizzolo et al., 2011; Melchor and Benitez, 2013). While linkage analyses failed to identify any compelling evident region of linkage in non-BRCA1/2 BC pedigrees (Antoniou and Easton, 2006). According to candidate gene screening, other high or moderate penetrance genes, such as *TP53*, *PALB2*, *STK11*, *ATM*, and *CHEK2* have been identified (Stratton and Rahman, 2008; Melchor and Benitez, 2013). With the application of Whole Exome Sequencing (WES), several novel BC predisposition genes have been identified from BC pedigrees, which further confirms that non-BRCA1/2 familial BC is highly heterogeneous.

An evaluation of potential predisposition roles of germline variants is challenging. First, to distinguish disease-causative variants from the non-pathogenic ones during WES analysis usually involves a series of filtering steps, including *in silico* prediction; however, such filtering steps

might cause over-filtering or be misleading (Bamshad et al., 2011). For instance, on one hand, *in silico* predictions might not be sensitive enough to detect all deleterious or damaging variants; on the other hand, the *in silico* predicted damaging variants might not be clinically pathogenic (Rahman, 2014). Second, to identify predisposition factors usually starts with an inspection of familial aggregation datasets, followed by a case-cohort confirmation (Kiiski et al., 2014); however, variants may be misclassified as having a uncertain significance due to their extreme rarity and heterogeneity. The efficiency of predisposition gene identification cannot be promoted significantly by simply increasing sample size. Third, incidental findings, which are not related to the observed phenotype of the patient, also complicate the analysis of the WES result (Kohane et al., 2006).

The American College of Medical Genetics and Genomics-Association for Molecular Pathology (ACMG-AMP) based guidelines have been widely used in variant classification (Hampel et al., 2015). Recently, ACMG-AMP-based variant classification rules have also been used in familial BC (Maxwell et al., 2016) and pan-cancer datasets (Huang et al., 2018). Of note, a co-segregation status of a germline variant is also important for variant classification (Jarvik and Browning, 2016). Pan-cancer studies have provided valuable sources to inspect tumor initiation and progression (Weinstein et al., 2013). An integrative analysis of germline and somatic variants could help to decipher tumor progression (Kanchi et al., 2014). We supposed that the co-occurrence between non-silent familial co-segregation variants and TCGA derived germline datasets could provide supporting evidence for a predisposition. Furthermore, pan-cancer datasets would also provide additional clues and evidence. Given that, we reanalyzed the WES datasets including 10 familial non-BRCA1/BRCA2 BC pedigrees (Gracia-Aznarez et al., 2013; Hilbers et al., 2013), manually evaluated variants as recommended (Hampel et al., 2015), and performed data mining on pan-cancer datasets.

In our analysis, some recently published BC predisposition genes, including *MRE11* (Bartkova et al., 2008), *CTR9* (Hanks et al., 2014), were recalibrated in our results, but were missed in the original publication. In addition, we identified novel cancer predisposition genes, such as *NCK1*. *NCK1* encodes the cytoplasmic adaptor protein NCK1, which contains Src homolog2 and 3 (SH2 and SH3) domains. As an adaptor, NCK1 mediates multiple signals from receptors, including EGFR, PDGFR, to downstream effectors and the overexpression of Nck in the NIH 3T3 cell line showed oncogenic features (Li et al., 1992). In mammalians, most Nck1 effectors are involved in cytoskeletal dynamics (Li et al., 2001). For instance, Nck1 is involved in actin cytoskeletal remodeling via the WASp/Arp2/3 complex, which in turn causes the polarization and directional migration of the cell (Lapetina et al., 2009). Interestingly, the mutation *NCK1* (p.D73H) identified from the BC pedigree (F2887) is located in an N-WASP activation motif (Okrut et al., 2015). Therefore, we supposed that *NCK1* (p.D73H) might impact cell invasion. MCF7 cell lines, which are non-invasive, transfected with *NCK1* mutants and were much more viable and invasive, *in vitro*. In conclusion, our results support that *NCK1* could be a candidate cancer predisposition gene.

MATERIALS AND METHODS

Whole Exome Sequencing Datasets

In this study, we reanalyzed WES data of non-*BRCA1/BRCA2* BC pedigrees (Gracia-Aznarez et al., 2013; Hilbers et al., 2013). Ten pedigrees with at least two independent patients applied to whole exome sequencing were involved in this study. The raw data of pedigrees (2887, 3311, RUL36, and RUL153) are available at National Centre for Biotechnology Information (NCBI) Sequence Read Archive (SRA) database (Project ID: PRJEB3235). The raw data of pedigrees (NIJM6, NIJM8, RUL39, RUL70, RUL79, and RUL154) were transmitted with permission. The authority of the datasets about those pedigrees belongs to the original authors.

Variant Calling, Annotation, and Evaluation

We mapped the WES reads against the human reference genome (hg19) using BWA *mem* mode, with parameters set as default (Li and Durbin, 2009) and preprocessed as recommended (McKenna et al., 2010). Mindful that highly quality off-target variants could be identified from WES (Guo et al., 2012), we generated all exon regions with flanking 100 bp via UCSC Table browser supplied to GATK for variant calling. We combined VQSR (Variant Quality Score Recalibration) and a hard filters to filter out potential false positive variants. The parameters are summarized in the **Supplementary Table S1**. The variants were then annotated with ANNOVAR (Wang et al., 2010) and classified as recommended (Hampel et al., 2015). The databases involved in annotation and the variant classification methods are summarized in the **Supplementary Table S1**.

Vector Construction, Cell Culture, and Transfection

Full-length *NCK1* was cloned from pLX304 to MSCV-5′HA (3×). We generated point mutants of *NCK1* (p.D73H and p.R42Q) via site-directed mutagenesis with primers designed by Primer X[1]. All the vectors were confirmed via Sanger sequencing. For lentivirus production, the *NCK1* mutants containing MSCV vectors were co-transfected with pCMV-VSVg and GAG/pol plasmids into 293FT cells by Lipo2000. Cell lines were cultured at 37°C under 5% CO_2 in DMEM, high glucose medium (Gibco) with 10% (v/v) fetal bovine serum (FBS; Gibco) and penicillin G (100 U/ml, Gibco) and streptomycin (100 ug/ml, Gibco).

Cell Viability Assay and Transwell Invasion Assay

Cell viability was assessed with MTT colorimetric assay (Ameresco), at time periods of 6 days. The optical absorbance was measured at 562 nm on a spectrophotometer (Biotek), and the reference wavelength at 630 nm. All the experiments were performed in triplicate and repeated three times. Cell invasion assays were performed using 24-well transwell (8 μm pore, Corning) that were coated with 1:10 diluted Matrigel Matrix (BD Biosciences). A total of 2×10^4 cells, in 200 μL of serum-free

[1]http://www.bioinformatics.org/primerx/

DMEM medium, were added into the upper transwell chamber, and 500 μL of 10% FBS DMEM medium containing 1 μg/mL EGF was added into the lower chamber. After incubation for 48 h, the cells were fixed in 4% paraformaldehyde and stained with 0.1% crystal violet. The cell images were taken at five random microscopic fields (Olympus, 10×). All experiments were repeated three times. The Student's *t*-test was used to test whether the difference was significant.

NCK1 Mutation Analysis

TCGA-germline variants were retrieved by subtracting the non-TCGA variants (ExAC-non-TCGA) from the whole dataset (ExAC) (Lek et al., 2016). Pan-cancer somatic mutations of *NCK1* were retrieved from cBioportal (Cerami et al., 2012). We performed a hotspot analysis on *NCK1* somatic mutations via the R package DominoEffect (Buljan et al., 2018). The flanking regions were determined after normalizing the gene length and impaired residues by function *calculate boundary* (Buljan et al., 2018). In order to evaluate substitution tolerance of *NCK1* mutations, position specific score matrix (PSSM) was generated by PSI-BLAST (Altschul et al., 1997). For a given missense mutation, we obtained the score difference between the mutation and wild type residue: $\Delta S = S_{mutation} - S_{wild-type}$. We generated 10,000 sets of three random mutations of *NCK1* and evaluated the mean score for each set.

NCK1 Mutation Burden Analysis

To perform mutation burden analysis of *NCK1* germline mutations in a cancer-cohort and normal controls, we retrieved the allele count and allele number of corresponding *NCK1* mutations from the general cohort, control-cohort, non-cancer cohort collected from the Genome Aggregation Database (genomAD) (Karczewski et al., 2019). The cancer-cohort specific allele count and allele number of *NCK1* mutations was obtained by deducting the non-cancer cohort from the general cohort. A Fisher test was used to test the occurrence of non-silent mutations in *NCK1* across the cohorts mentioned above.

NCK1 Expression Analysis

As described before (Chen et al., 2016), the mRNA expression level in *NCK1* (RNA-seq V2) of 99 tumor-normal matched BC samples were retrieved from the Cancer Genome Atlas database (Weinstein et al., 2013) and the RSEM normalized result were applied to the downstream analysis. Among them, 95 patients owned inferred PAM50 subtypes (Netanely et al., 2016).

RESULTS

Re-evaluation Variants Identified From Familial Breast Cancer Patients

We reanalyzed published Whole Exome Sequencing datasets from 10 non-*BRCA1/2* BC pedigrees (Gracia-Aznarez et al., 2013; Hilbers et al., 2013). Two samples per pedigree were applied to whole exome sequencing, and the kinship of the samples varied from 0.016 to 0.25 (**Table 1**). We set those rare

TABLE 1 | Candidate predisposition genes identified from non-BRCA1/2 pedigrees.

Pedigree	Kinship	Evaluation	Genotype	Gene	Type	Transcript	Exon	HGVS(.c)	HGVS(.p)
F3311	0.0625	Likely pathogenic	Het	*MRE11*	Missense	NM_005591.3	Exon 3	c.94A > G	p.R32G
NJM6	0.25	Likely pathogenic$^{+,*}$	Het	*XRCC2*	Missense	NM_005431.1	Exon 3	c.271C > T	p.R91W
NJM6	0.25	Pathogenic	Het	*CHEK2*	Frameshift deletion	NM_007194.3	Exon 11	c.1100delC	p.T367fs
NJM8	0.125	Likely pathogenic	Het	*CHRNA3*	Missense	NM_000743.4	Exon 5	c.877A > G	p.T293A
RUL036	0.125	Likely pathogenic*	Het	*ATM*	Missense	NM_000051.3	Exon 57	c.8393C > A	p.A2798D
RUL036	0.125	Pathogenic	Het	*CTR9*	Splicing	NM_014633.4	Exon 22	c.2728-1G > A	–
RUL153	0.016	Pathogenic^{+}	Het	*CHEK2*	Frameshift deletion	NM_007194.3	Exon 11	c.1100delC	p.T367fs
RUL39	0.0625	Likely pathogenic	Het	*IGF2R*	Missense	NM_000876.3	Exon 27	c.3718C > G	p.L1240V
RUL79	0.25	Pathogenic^{+}	Het	*MUTYH*	Missense	NM_001128425.1	Exon 8	c.667A > G	p.I223V
F2887	0.03	Uncertain of significance*	Het	*NCK1*	Missense	NM_006153.5	Exon 2	c.217G > C	p.D73H

*+, indicates the corresponding variant reported in the original publication; *, the corresponding variants that occurred in TCGA-germline database; Het, Heterozygous; Kinship of the samples applied to WES were estimated based on the pedigree structure; all the variants were shared between the two samples applied to WES per pedigree.*

non-silent variants, shared between patients per pedigree, as candidate co-segregated ones. To reduce incidental findings, we first focused on the genes that had been assigned with pathogenic supporting evidence (**Supplementary Table S1**), especially the known cancer predisposition genes (Rahman, 2014). Second, we filtered for variants with uncertain clinical significance, which must show in both the familial and TCGA germline dataset. The detailed variant filtering and classification parameters are summarized in **Supplementary Table S1**.

In our analysis, we found that seven out of 10 pedigrees had potential co-segregated pathogenic variants in known cancer-associated genes (**Table 1** and **Supplementary Table S1**), including *CHEK2*, *ATM*, *MRE11*, and *CTR9*, and some other cancer-associated genes, such as *IGF2R* and *CHRNA3* (**Table 1** and **Supplementary Table S1**). Interestingly, we found that *XRCC2* (p.R91W) and *ATM* (p.A2798D) co-occurred in the ExAC TCGA-germline dataset (**Table 1** and **Supplementary Table S1**). Furthermore, the *XRCC2* (p.R91W) was also reported in the original publication (Hilbers et al., 2012) and an independent pedigree (Park et al., 2012), which further confirmed our approach was effective. Finally, we identified a novel candidate gene, *NCK1*, from pedigree F2887 (**Table 1** and **Supplementary Table S1**). *NCK1* (p.D73H) occurred once in about 7000 TCGA samples, but did not show up in more than 60,000 control samples (**Supplementary Table S2**). Generally, we succeeded in identifying potential cancer predisposition variants from eight in 10 pedigrees in the evaluation.

Most of the Somatic and Germline Mutations in *NCK1* Were Intolerant

So far, few publications have reported the cancer predisposition role of *NCK1*. First, we inspected the *NCK1* variants in the genome aggregation database (genomAD), which contained the cancer patient cohort and provided detailed cohort information, such as non-cancer, control (**Supplementary Table S2**). We could therefore retrieve the allele counts and allele numbers of the corresponding variants recorded in genomAD for enrichment analysis (**Supplementary Table S2**). Additionally, we only focused on the high-quality variants, which were marked as a pass in both the exome and genome datasets. The *NCK1* mutations were significantly enriched in the cancer cohort, non-cancer cohort, and general cohort in comparison to the control cohort (Fisher-test; $P < 0.001$) (**Supplementary Table S2**).

Second, we inspected the occurrence of somatic mutations in *NCK1* among pan-cancer datasets since the somatic event is another important factor involved in cancer progression. According to pan-cancer datasets, 0.3% of patients had *NCK1* somatic mutations, including 102 non-silent mutations from 97 patients, and four fusion variants impaired *NCK1* in four patients (**Figure 1A**). *NCK1* mutations were enriched in some cancer types, including uterine endometrioid carcinoma ($P = 1e^{-16}$), stomach adenocarcinoma ($P = 6.679e^{-06}$), cutaneous melanoma ($P = 2.63e^{-05}$), but not BC ($P > 0.05$) (**Supplementary Figure S1**; Binominal test). Among these mutations, residue 42 is the most frequent in somatic, and residue 73 mutated in both the BC pedigree and the cancer cohort (**Figures 1A,B**). Given the rarity of the *NCK1* germline and somatic mutations, we supposed that mutations in *NCK1* might be intolerant.

To confirm that supposition, we generated a position specific score matrix (PSSM) via PSI-BLAST (Altschul et al., 1997) and predicted the damaging effect with SIFT (Ng and Henikoff, 2003) and PolyPhen2 (Adzhubei et al., 2010). *In silico*, PolyPhen2 predicted that those two were possibly damaging, and SIFT predicted that those two mutations were tolerant. Paradoxically, the residue D73 and R42 are conserved among 100 vertebrates according to MultiZ alignment (**Supplementary Figure S2**; Rosenbloom et al., 2015), and the residue R42 and D73 are both conserved in NCK1 and NCK2, which is the paralog of NCK1, but not conserved in the orthologs in *Caenorhabditis elegans* and *Drosophila melanogaster* (**Figure 1C**). According to PSSM, both germline and somatic mutations of *NCK1* were more intolerant than randomly modeling mutations (**Figure 1D**), and substitution score of NCK1 D73H ($\Delta S = -3$) and NCK1 R42Q ($\Delta S = -1$) both are negative. *In vitro*, we found that both the mutants could increase cell viability (**Figures 1E,F**); therefore, both the *NCK1* mutations should be deleterious.

Role of *NCK1* Variations in Tumor Progression

Based on the "20/20" rule (Vogelstein et al., 2013), which means that more than 20 percent missense were located in recurred residues (**Figures 1A,B**), we supposed that *NCK1* might have an oncogenic role. According to hotspot analysis of *NCK1* somatic mutations, we found that the residue 42 turned to be a hotspot site ($P < 0.001$) (**Supplementary Table S3**). Indeed, NCK1-D73H and NCK1-R42Q transfected MCF7 cell lines showed significantly increased cell viability in comparison with wild type (**Figures 1E,F**). In addition, NCK1 contains an N-WASP activation motif (Okrut et al., 2015), where the residue D73 locates. Given this, we supposed that *NCK1* might involve in tumor invasion.

To further prove that, we assessed the *NCK1* mRNA expression level among 99 tumor-normal matched samples from TCGA-BRCA. However, the expression of *NCK1* mRNA in tumor samples was significantly lower than the matched normal samples (**Figure 2A**), which was also observed across different tumor stages (**Figures 2B–D**). Mindful that BC is a molecular heterogenous cancer type, we retrieved PAM50 subtypes of the corresponding samples (Netanely et al., 2016). We found that *NCK1* was significantly upregulated in the basal-like subtype (**Figure 2E**). No significant difference was observed in the Her2 subtype (**Figure 2F**), but the expression of *NCK1* was still significantly downregulated in the Luminal A (**Figure 2G**) and Luminal B subtype (**Figure 2H**), especially in Luminal A. In this study, both NCK1-R42Q and NCK1-D73H transfected MCF cell lines, which are luminal subtypes, and showed a significantly increased invasion ability (**Figures 3A,B**). Recently, Morris et al. (2017) reported that the deficiency of *Nck* in MDA-MB-231, which is a basal-like subtype, could delay BC progression and metastasis, which was consistent with our results - given that *NCK1* also plays a vital role in tumor invasion. Finally, we inspected the survival status of the patients with *NCK1* variations, including CNVs, somatic mutations, and a Z-score normalized mRNA expression level, via cBioPortal (Gao et al., 2013). We found that the

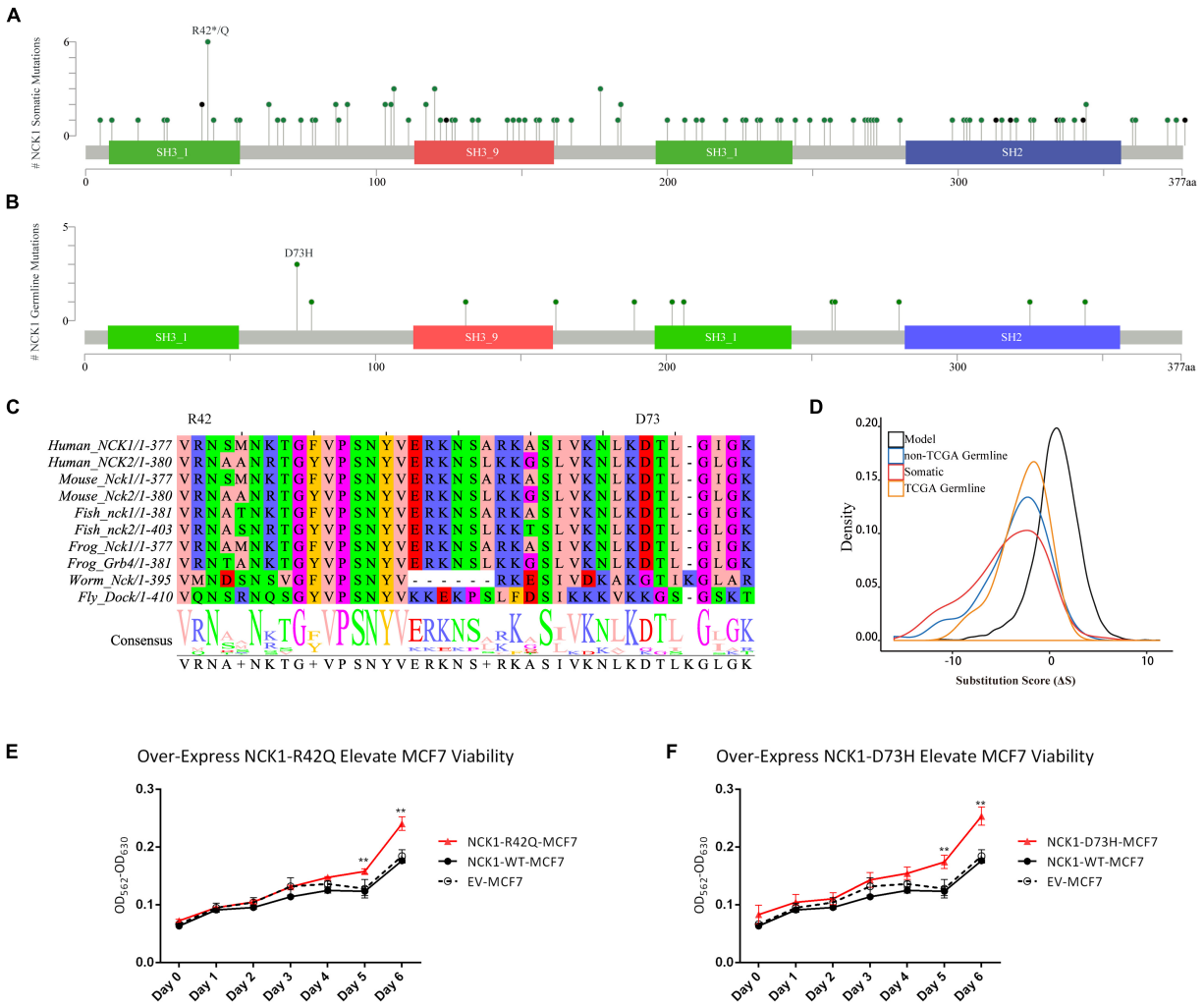


FIGURE 1 | *NCK1* mutation diagram and potential functional effect. **(A)** Mutation diagram of *NCK1* collected in cBioportal (Pan-Cancer). **(B)** Mutation diagram of germline mutations in *NCK1*, including all TCGA-germline variants and *NCK1* D73H, identified in familial breast cancer pedigree (F2887). **(C)** Multiple sequence alignment of sequence flanking NCK1 D73 residue. **(D)** Distribution of substitution score (ΔS) of NCK1 based on Position Specific Score Matrix. **(E and F)** The cell viabilities in all groups of mutant over-expression assay about R42Q **(E)** and D73H **(F)** at different time points (0, 1, 2, 3, 4, 5, and 6 days). Data were expressed as mean ± standard deviation (SD) of experiments with triplicates. Asterisks indicate significant increasing of cell viability in mutant (R42Q and D73H) transfected MCF7 cells compared with wild type transfected MCF7 cells (Student's *t*-test; $P < 0.01$). Model: random mutations generated by *in silico*, non-TCGA germline: variants collected in ExAC non-TCGA dataset; TCGA-germline: variants collected in ExAC, but not in the ExAC non-TCGA dataset; Somatic: somatic variants collected in cBioportal.

patients with both *NCK1* variations and *TP53* mutations had poorer overall survival ($P < 0.05$) and disease-free survival ($P < 0.05$) (**Figures 3C,D**). In general, the roles of *NCK1* in tumor progression could be genomic context dependent and differentiated in cancer types.

DISCUSSION

Intense efforts have been dedicated to identifying BC genes; however, more than 50% of familial BC heritability is still undetermined (Melchor and Benitez, 2013). Furthermore, non-BRCA1/2 familial BC patients are highly heterogeneous. For instance, we found *CHEK2* mutations from four pedigrees, including pedigree RUL153, NIJM6, NIJM8 and RUL70 (**Supplementary Table S4**). The *CHEK2* (p.T367fs) in pedigree NIJM8 appears to be homozygous but was only identified in one patient. Two separate *CHEK2* variants were identified from members of pedigree NIJM8 (**Supplementary Table S4**). In RUL70, we also identified a *CHEK2* mutation from only one patient. However, the confident predisposition variant in *XRCC2* (**Table 1**) identified from another *CHEK2* positive pedigree (NIJM6) further complicate the evaluation. *CHEK2* (p.T367fs) was not co-segregated across all patients in RUL153, which was explained as a phenocopy (Gracia-Aznarez et al., 2013). Although *CHEK2* (p.T367fs) is a well-known BC predisposition gene (Meijers-Heijboer et al., 2002), the co-segregation status of

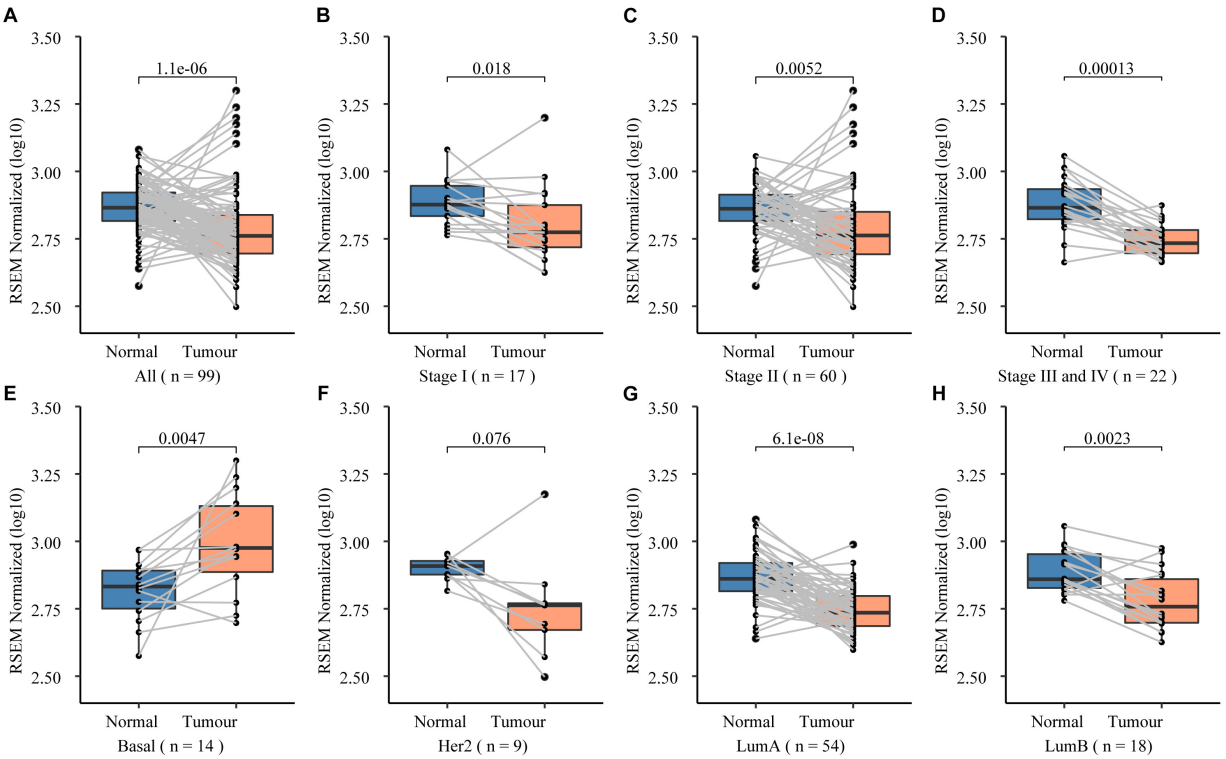


FIGURE 2 | Expression spectrum of *NCK1* in 99 tumor-normal paired samples across different stage and subtypes. **(A)** All; **(B)** Stage I; **(C)** Stage II; **(D)** Stage III and IV; **(E)** basal-like; **(F)** Her2; **(G)** LumA (LuminalA); and **(H)** LumB (LuminalB).

the variant has turned out to be negative among those pedigrees. Due to the patients in RUL70 and NIJM6, NIJM8 has been reported with a chromosome 22 gain like profile (Hilbers et al., 2013), where *CHEK2* locates, and we therefore suppose that structural variants might also contribute.

During our analysis, we also identified some likely pathogenic variants in recently established cancer predisposition genes, such as *MRE11* (Bartkova et al., 2008; Damiola et al., 2014) and *CTR9* (Hanks et al., 2014). MRE11A, encoded by *MRE11*, acting as a component of the MRN (MRE11A-RAD50-NBN) complex, which plays a vital role in DNA double-strand break repair (Yuan et al., 2012). Dysfunction of the MRN complex could promote BC invasion and metastasis (Gupta et al., 2013). In pedigree RUL036, we identified two candidate predisposition genes, including *ATM* and *CTR9*. Although the *ATM* variant occurred in the TCGA-germline dataset, multiple *in silico* tools predicted it to be benign or tolerant. *CTR9* was first reported as a Wilms tumor predisposition gene, and the mutations are almost truncated (Hanks et al., 2014). As it occurs in the Wilms tumor, we also identified a splicing site mutation in *CTR9*. Interestingly, evidence indicates that CTR9 plays an import role in regulating the estrogen signaling pathway, which promotes estrogen receptor α (ERα) positive BC progression (Zeng and Xu, 2015). In addition, we found a rare non-silent mutation in *IGF2R*. *IGF2R* is a polymorphic imprinting locus in humans (Xu et al., 1993), which indicates that individuals with *IGF2R* imprinted, might have increased cancer susceptibility (Feinberg, 1993). *CHRNA3* encodes an α type subunit of the nicotinic acetylcholine receptor. Polymorphisms in *CHRNA3* have been associated with increased smoking initiation risk and increases susceptibility to lung cancer (Hung et al., 2008). Given the heterogeneity in BC, the predisposition genes might have different disease-causative mechanisms and predisposition factors of non-BRCA1/2 pedigrees might be multifactorial, such as gene-environment interaction.

In our study, we mainly focused on gene *NCK1*, because few reports suggest the underlying predisposition role of *NCK1* mutations. As an adaptor, NCK1 mediated multiple signaling pathways, especially actin dynamic and organization involved in invadopodia formation and maturation (Stylli et al., 2009; Oser et al., 2010). The SH2 domain of NCK1 involves the recognition of cell surface receptors and transduces signals to downstream effectors (Li et al., 2001). The SH3 domain of NCK1 usually interacts with downstream effectors, most of which involves the actin cytoskeletal dynamic. For instance, NCK1 is required for EGFR-mediated cell migration and tumor metastasis (Huang et al., 2012). And the metastasis-promoting role of *NCK1* has been reported in multiple cancer types, such as colorectal cancer (Zhang et al., 2017) and BC (Morris et al., 2017). Interestingly, *NCK1* also have connections to the hotspot mutation of *PIK3CA*. Wu et al. reported that oncogenic mutations of *PIK3CA* mediate tumor cell invasion through cortactin (Wu et al., 2014), which is a partner of NCK1 in invadopodia maturation (Oser et al., 2010). Therefore, *NCK1* might be an invisible participant in tumor progression, because *NCK1* mutations rarely occur in cancer patients.

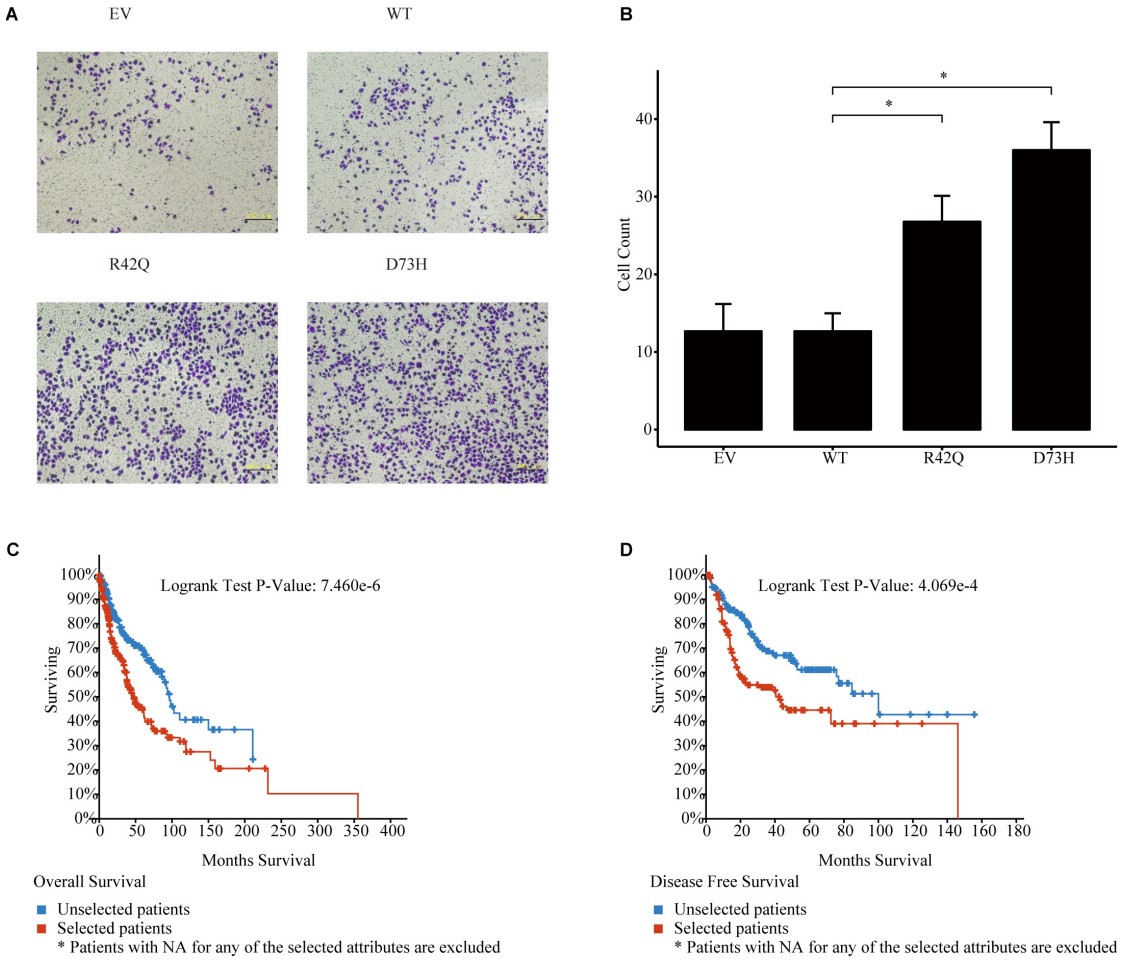


FIGURE 3 | Roles of *NCK1* in tumor progression might be context dependent. **(A)** Images of MCF7 cells migrated from transwell membrane **(B)** Cell count and quantitative analysis of the migrated MCF7 cells. Patients with both *NCK1* aberrations and TP53 mutations showed a much poorer overall survival **(C)** and disease-free survival **(D)**. Selected patients: patients with both *NCK1* aberrations and *TP53* mutation. Unselected patients: patients with only *NCK1* aberrations. Scale bar: 200 μm. Data are depicted as mean ± standard deviation (SD).

On the one hand, overexpression of *NCK1* shows oncogenic roles (Li et al., 1992), and the high expression of *NCK1*, at least in basal-like BC, contributes to tumor proliferation and metastasis (Morris et al., 2017). In our study, we identified a mutation in a motif that is involved in N-WASP activation, which is involved in invadopodia maturation (Okrut et al., 2015). Our results showed that both the NCK1 mutants (D73H and R42Q) indeed promote cell proliferation and invasion *in vitro*. We propose that *NCK1* not only contributes to cancer predisposition but is also involved in cancer progression and prognosis. In addition, our results also suggest that the tumor-promoting role of *NCK1* might be a cancer subtype dependent. On the other hand, downregulation of *NCK1* might also be pathogenic, but in different mechanisms. For instance, *Nck* degradation could prevent cancer cells from apoptosis (Li et al., 2013) and regulate actin dynamics (Buvall et al., 2013). Furthermore, *NCK1* played important roles in angiogenesis (Zhang et al., 2017; Xia et al., 2018) and even has an unexpected link to CHEK2 activation (Kremer et al., 2007).

Traditional approaches to identify underlying predisposition genes usually involves allele frequency filtering and *in silico* prediction and the sequences involved in the comparative analysis could also impact the final accuracy. Although we identified some novel candidate cancer predisposition variants, the power to confirm the predisposition role of those variants was limited. Because most of candidate cancer predisposition variants identified in our analysis turn out to be familial specific, which indicates that the power to establish a novel predisposition variant depends on an extremely large sample size (Guo et al., 2016). For instance, the variant *NCK1* (p.D73H), identified from the pedigree F2887, occurred once in about 7,000 cancer samples, but not in about 60,000 controls according to the genomAD datasets. The predisposition role of *NCK1* mutations was ignored probably because of its rare occurrence. In general, our results support *NCK1* as a candidate cancer gene; however, the underlying mechanisms requirefurther investigation. In addition, we imagine that many more cancer genes like *NCK1* might exist.

AUTHOR CONTRIBUTIONS

JY, LH, and XK aided in data collection and developed the concepts. JY performed the data analysis, variants evaluation, vector construction, table, created the figures, and drafted and revised the manuscript. KW aided in variants evaluation, vector construction, *in vitro* experiments, and manuscript revision. QM and HD aided in experiments design and manuscript revision. YZ and LH aided in data collection and manuscript revision. XK supervised all the study. All the authors read and approved the final manuscript.

FUNDING

This work was supported by the Key Programs of the Chinese Academy of Sciences (Grant No. QYZDJ-SSW-SMC01), the National Natural Science Foundation of China (Grant Nos. 31371499, 31471224, and 81570827), and the National Basic Research Program of China (Grant No. 2011CB510102).

ACKNOWLEDGMENTS

We are grateful to all patients and families for their participation in the original study. We thank Florentine S. Hilbers for sharing the Whole Exome Sequencing data of the non-BRCA1/2 breast cancer pedigrees. The pan-cancer analysis result here are in whole based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. The pan-cancer data mining analysis are mainly based on cBioportal: https://www.cbioportal.org. The large cohort analysis result about germline variants here was based on data generated by both the Exome Aggregation Consortium (ExAC): http://exac.broadinstitute.org and the Genome Aggregation Database (genomAD): https://gnomad.broadinstitute.org.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00527/full#supplementary-material

FIGURE S1 | Summary of *NCK1* variations across TCGA studies.

FIGURE S2 | Multiple sequence alignment of residue flanking R42 and D73 across 100 vertebrates.

TABLE S1 | Variant filtering and classification parameters and the whole list of candidate genes.

TABLE S2 | The allele count and allele number of the non-silent *NCK1* mutations retrieved from genomAD and the burden test result.

TABLE S3 | Hotspot analysis of *NCK1* somatic mutations.

TABLE S4 | All *CHEK2* variants identified from the pedigrees.

REFERENCES

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., et al. (2010). A method and server for predicting damaging missense mutations. *Nat. Methods* 7, 248–249. doi: 10.1038/nmeth0410-248

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Antoniou, A. C., and Easton, D. F. (2006). Models of genetic susceptibility to breast cancer. *Oncogene* 25, 5898–5905. doi: 10.1038/sj.onc.1209879

Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M. J., Nickerson, D. A., et al. (2011). Exome sequencing as a tool for mendelian disease gene discovery. *Nat. Rev. Genet.* 12, 745–755. doi: 10.1038/nrg3031

Bartkova, J., Tommiska, J., Oplustilova, L., Aaltonen, K., Tamminen, A., Heikkinen, T., et al. (2008). Aberrations of the MRE11-RAD50-NBS1 DNA damage sensor complex in human breast cancer: MRE11 as a candidate familial cancer-predisposing gene. *Mol. Oncol.* 2, 296–316. doi: 10.1016/j.molonc.2008.09.007

Buljan, M., Blattmann, P., Aebersold, R., and Boutros, M. (2018). Systematic characterization of pan-cancer mutation clusters. *Mol. Syst. Biol.* 14:e7974. doi: 10.15252/msb.20177974

Buvall, L., Rashmi, P., Lopez-Rivera, E., Andreeva, S., Weins, A., Wallentin, H., et al. (2013). Proteasomal degradation of Nck1 but not Nck2 regulates RhoA activation and actin dynamics. *Nat. Commun.* 4:2863. doi: 10.1038/ncomms3863

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* 2, 401–404. doi: 10.1158/2159-8290.cd-12-0095

Chen, X., Cao, X., Sun, X., Lei, R., Chen, P., Zhao, Y., et al. (2016). Bcl-3 regulates TGFbeta signaling by stabilizing Smad3 during breast cancer pulmonary metastasis. *Cell Death Dis.* 7:e2508. doi: 10.1038/cddis.2016.405

Damiola, F., Pertesi, M., Oliver, J., Le Calvez-Kelm, F., Voegele, C., Young, E. L., et al. (2014). Rare key functional domain missense substitutions in MRE11A, RAD50, and NBN contribute to breast cancer susceptibility: results from a Breast Cancer Family Registry case-control mutation-screening study. *Breast Cancer Res.* 16:R58. doi: 10.1186/bcr3669

Feinberg, A. P. (1993). Genomic imprinting and gene activation in cancer. *Nat. Genet.* 4, 110–113. doi: 10.1038/ng0693-110

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* 6:l1. doi: 10.1126/scisignal.2004088

Gracia-Aznarez, F. J., Fernandez, V., Pita, G., Peterlongo, P., Dominguez, O., de la Hoya, M., et al. (2013). Whole exome sequencing suggests much of non-BRCA1/BRCA2 familial breast cancer is due to moderate and low penetrance susceptibility alleles. *PLoS One* 8:e55681. doi: 10.1371/journal.pone.0055681

Guo, M. H., Dauber, A., Lippincott, M. F., Chan, Y. M., Salem, R. M., and Hirschhorn, J. N. (2016). Determinants of power in gene-based burden testing for monogenic disorders. *Am J Hum. Genet.* 99, 527–539. doi: 10.1016/j.ajhg.2016.06.031

Guo, Y., Long, J., He, J., Li, C. I., Cai, Q., Shu, X. O., et al. (2012). Exome sequencing generates high quality data in non-target regions. *BMC Genomics* 13:194. doi: 10.1186/1471-2164-13-194

Gupta, G. P., Vanness, K., Barlas, A., Manova-Todorova, K. O., Wen, Y. H., and Petrini, J. H. (2013). The Mre11 complex suppresses oncogene-driven breast tumorigenesis and metastasis. *Mol. Cell* 52, 353–365. doi: 10.1016/j.molcel.2013.09.001

Hampel, H., Bennett, R. L., Buchanan, A., Pearlman, R., and Wiesner, G. L. (2015). A practice guideline from the american college of medical genetics and genomics and the national society of genetic counselors: referral indications

for cancer predisposition assessment. *Genet. Med.* 17, 70–87. doi: 10.1038/gim.2014.147

Hanks, S., Perdeaux, E. R., Seal, S., Ruark, E., Mahamdallie, S. S., Murray, A., et al. (2014). Germline mutations in the PAF1 complex gene CTR9 predispose to Wilms tumour. *Nat. Commun.* 5:4398. doi: 10.1038/ncomms5398

Hilbers, F. S., Meijers, C. M., Laros, J. F., van Galen, M., Hoogerbrugge, N., Vasen, H. F., et al. (2013). Exome sequencing of germline DNA from non-BRCA1/2 familial breast cancer cases selected on the basis of aCGH tumor profiling. *PLoS One* 8:e55734. doi: 10.1371/journal.pone.0055734

Hilbers, F. S., Wijnen, J. T., Hoogerbrugge, N., Oosterwijk, J. C., Collee, M. J., Peterlongo, P., et al. (2012). Rare variants in XRCC2 as breast cancer susceptibility alleles. *J. Med. Genet.* 49, 618–620. doi: 10.1136/jmedgenet-2012-101191

Huang, K. L., Mashl, R. J., Wu, Y., Ritter, D. I., Wang, J., Oh, C., et al. (2018). Pathogenic germline variants in 10,389 adult cancers. *Cell* 173, 355.e–370.e. doi: 10.1016/j.cell.2018.03.039

Huang, M., Anand, S., Murphy, E. A., Desgrosellier, J. S., Stupack, D. G., Shattil, S. J., et al. (2012). EGFR-dependent pancreatic carcinoma cell metastasis through Rap1 activation. *Oncogene* 31, 2783–2793. doi: 10.1038/onc.2011.450

Hung, R. J., McKay, J. D., Gaborieau, V., Boffetta, P., Hashibe, M., Zaridze, D., et al. (2008). A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature* 452, 633–637. doi: 10.1038/nature06885

Jarvik, G. P., and Browning, B. L. (2016). Consideration of cosegregation in the pathogenicity classification of genomic variants. *Am. J. Hum. Genet.* 98, 1077–1081. doi: 10.1016/j.ajhg.2016.04.003

Kanchi, K. L., Johnson, K. J., Lu, C., McLellan, M. D., Leiserson, M. D., Wendl, M. C., et al. (2014). Integrated analysis of germline and somatic variants in ovarian cancer. *Nat. Commun.* 5:3156. doi: 10.1038/ncomms4156

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., et al. (2019). Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv*

Kiiski, J. I., Pelttari, L. M., Khan, S., Freysteinsdottir, E. S., Reynisdottir, I., Hart, S. N., et al. (2014). Exome sequencing identifies FANCM as a susceptibility gene for triple-negative breast cancer. *Proc Natl Acad Sci U.S.A.* 111, 15172–15177. doi: 10.1073/pnas.1407909111

Kohane, I. S., Masys, D. R., and Altman, R. B. (2006). The incidentalome: a threat to genomic medicine. *Jama* 296, 212–215. doi: 10.1001/jama.296.2.212

Kremer, B. E., Adang, L. A., and Macara, I. G. (2007). Septins regulate actin organization and cell-cycle arrest through nuclear accumulation of NCK mediated by SOCS7. *Cell* 130, 837–850. doi: 10.1016/j.cell.2007.06.053

Lapetina, S., Mader, C., Machida, K., Mayer, B., and Koleske, A. (2009). Arg interacts with cortactin to promote adhesion-dependent cell edge protrusion. *J. Cell Biol.* 185, 503–519. doi: 10.1083/jcb.200809085

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285–291. doi: 10.1038/nature19057

Li, B., Pi, Z., Liu, L., Zhang, B., Huang, X., Hu, P., et al. (2013). FGF-2 prevents cancer cells from ER stress-mediated apoptosis via enhancing proteasome-mediated Nck degradation. *Biochem. J.* 452, 139–145. doi: 10.1042/bj20121671

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, W., Fan, J., and Woodley, D. T. (2001). Nck/Dock: an adapter between cell surface receptors and the actin cytoskeleton. *Oncogene* 20, 6403–6417. doi: 10.1038/sj.onc.1204782

Li, W., Hu, P., Skolnik, E. Y., Ullrich, A., and Schlessinger, J. (1992). The SH2 and SH3 domain-containing Nck protein is oncogenic and a common target for phosphorylation by different surface receptors. *Mol. Cell Biol.* 12, 5824–5833. doi: 10.1128/mcb.12.12.5824

Lynch, H., Synder, C., and Wang, S. M. (2015). Considerations for comprehensive assessment of genetic predisposition in familial breast cancer. *Breast J.* 21, 67–75. doi: 10.1111/tbj.12358

Mavaddat, N., Antoniou, A. C., Easton, D. F., and Garcia-Closas, M. (2010). Genetic susceptibility to breast cancer. *Mol. Oncol.* 4, 174–191. doi: 10.1016/j.molonc.2010.04.011

Maxwell, K. N., Hart, S. N., Vijai, J., Schrader, K. A., Slavin, T. P., Thomas, T., et al. (2016). Evaluation of ACMG-guideline-based variant classification of cancer susceptibility and non-cancer-associated genes in families affected by breast cancer. *Am. J. Hum. Genet.* 98, 801–817. doi: 10.1016/j.ajhg.2016.02.024

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

Meijers-Heijboer, H., van den Ouweland, A., Klijn, J., Wasielewski, M., de Snoo, A., Oldenburg, R., et al. (2002). Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of BRCA1 or BRCA2 mutations. *Nat. Genet.* 31, 55–59. doi: 10.1038/ng879

Melchor, L., and Benitez, J. (2013). The complex genetic landscape of familial breast cancer. *Hum. Genet.* 132, 845–863. doi: 10.1007/s00439-013-1299-y

Morris, D. C., Popp, J. L., Tang, L. K., Gibbs, H. C., Schmitt, E., Chaki, S. P., et al. (2017). Nck deficiency is associated with delayed breast carcinoma progression and reduced metastasis. *Mol. Biol. Cell* 28, 3500–3516. doi: 10.1091/mbc.E17-02-0106

Netanely, D., Avraham, A., Ben-Baruch, A., Evron, E., and Shamir, R. (2016). Expression and methylation patterns partition luminal-A breast tumors into distinct prognostic subgroups. *Breast Cancer Res.* 18:74. doi: 10.1186/s13058-016-0724-722

Ng, P. C., and Henikoff, S. (2003). SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 31, 3812–3814. doi: 10.1093/nar/gkg509

Okrut, J., Prakash, S., Wu, Q., Kelly, M. J., and Taunton, J. (2015). Allosteric N-WASP activation by an inter-SH3 domain linker in Nck. *Proc. Natl. Acad. Sci. U.S.A.* 112, E6436–E6445. doi: 10.1073/pnas.1510876112

Oser, M., Mader, C. C., Gil-Henn, H., Magalhaes, M., Bravo-Cordero, J. J., Koleske, A. J., et al. (2010). Specific tyrosine phosphorylation sites on cortactin regulate Nck1-dependent actin polymerization in invadopodia. *J. Cell. Sci.* 123(Pt 21), 3662–3673. doi: 10.1242/jcs.068163

Park, D. J., Lesueur, F., Nguyen-Dumont, T., Pertesi, M., Odefrey, F., Hammet, F., et al. (2012). Rare mutations in XRCC2 increase the risk of breast cancer. *Am. J. Hum. Genet.* 90, 734–739. doi: 10.1016/j.ajhg.2012.02.027

Rahman, N. (2014). Realizing the promise of cancer predisposition genes. *Nature* 505, 302–308. doi: 10.1038/nature12981

Rizzolo, P., Silvestri, V., Falchetti, M., and Ottini, L. (2011). Inherited and acquired alterations in development of breast cancer. *Appl. Clin. Genet.* 4, 145–158. doi: 10.2147/tacg.s13226

Rosenbloom, K. R., Armstrong, J., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., et al. (2015). The UCSC genome browser database: 2015 update. *Nucleic Acids Res.* 43, D670–D681. doi: 10.1093/nar/gku1177

Siegel, R. L., Miller, K. D., and Jemal, A. (2018). Cancer statistics, 2018. *Can. J. Clin.* 68, 7–30. doi: 10.3322/caac.21442

Stratton, M. R., and Rahman, N. (2008). The emerging landscape of breast cancer susceptibility. *Nat. Genet.* 40, 17–22. doi: 10.1038/ng.2007.53

Stylli, S. S., Stacey, T. T., Verhagen, A. M., Xu, S. S., Pass, I., Courtneidge, S. A., et al. (2009). Nck adaptor proteins link Tks5 to invadopodia actin regulation and ECM degradation. *J. Cell Sci.* 122(Pt 15), 2727–2740. doi: 10.1242/jcs.046680

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A. Jr., Kinzler, K. W., et al. (2013). Cancer genome landscapes. *Science* 339, 1546–1558. doi: 10.1126/science.1235122

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38, e164. doi: 10.1093/nar/gkq603

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* 45, 1113–1120. doi: 10.1038/ng.2764

Wu, X., Renuse, S., Sahasrabuddhe, N. A., Zahari, M. S., Chaerkady, R., Kim, M. S., et al. (2014). Activation of diverse signalling pathways by oncogenic PIK3CA mutations. *Nat. Commun.* 5:4961. doi: 10.1038/ncomms5961

Xia, P., Huang, M., Zhang, Y., Xiong, X., Yan, M., Xiong, X., et al. (2018). NCK1 promotes the angiogenesis of cervical squamous carcinoma via Rac1/PAK1/MMP2 signal pathway. *Gynecol. Oncol.* 152, 387–395. doi: 10.1016/j.ygyno.2018.11.013

Xu, Y., Goodyer, C. G., Deal, C., and Polychronakos, C. (1993). Functional polymorphism in the parental imprinting of the human IGF2R gene. *Biochem. Biophys. Res. Commun.* 197, 747–754. doi: 10.1006/bbrc.1993.2542

Yuan, S. S., Hou, M. F., Hsieh, Y. C., Huang, C. Y., Lee, Y. C., Chen, Y. J., et al. (2012). Role of MRE11 in cell proliferation, tumor invasion, and DNA repair in breast cancer. *J. Natl. Cancer Inst.* 104, 1485–1502. doi: 10.1093/jnci/djs355

Zeng, H., and Xu, W. (2015). Ctr9, a key subunit of PAFc, affects global estrogen signaling and drives ERalpha-positive breast tumorigenesis. *Genes Dev.* 29, 2153–2167. doi: 10.1101/gad.268722.115

Zhang, F., Lu, Y. X., Chen, Q., Zou, H. M., Zhang, J. M., Hu, Y. H., et al. (2017). Identification of NCK1 as a novel downstream effector of STAT3 in colorectal cancer metastasis and angiogenesis. *Cell Signal.* 36, 67–78. doi: 10.1016/j.cellsig.2017.04.020

Conflict of Interest Statement: 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.

frontiers in Genetics

PERSPECTIVE
published: 25 June 2019
doi: 10.3389/fgene.2019.00601

Dissecting *in silico* Mutation Prediction of Variants in African Genomes: Challenges and Perspectives

Christian Domilongo Bope[1,2], Emile R. Chimusa[1], Victoria Nembaware[1], Gaston K. Mazandu[1], Jantina de Vries[3] and Ambroise Wonkam[1,3,4]*

[1] Department of Pathology, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa, [2] Departments of Mathematics and Computer Sciences, Faculty of Sciences, University of Kinshasa, Kinshasa, Democratic Republic of Congo, [3] Department of Medicine, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa, [4] Institute of Infectious Diseases and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa

Genomic medicine is set to drastically improve clinical care globally due to high throughput technologies which enable speedy *in silico* detection and analysis of clinically relevant mutations. However, the variability in the *in silico* prediction methods and categorization of functionally relevant genetic variants can pose specific challenges in some populations. *In silico* mutation prediction tools could lead to high rates of false positive/negative results, particularly in African genomes that harbor the highest genetic diversity and that are disproportionately underrepresented in public databases and reference panels. These issues are particularly relevant with the recent increase in initiatives, such as the Human Heredity and Health (H3Africa), that are generating huge amounts of genomic sequence data in the absence of policies to guide genomic researchers to return results of variants in so-called actionable genes to research participants. This report (i) provides an inventory of publicly available Whole Exome/Genome data from Africa which could help improve reference panels and explore the frequency of pathogenic variants in actionable genes and related challenges, (ii) reviews available *in silico* prediction mutation tools and the criteria for categorization of pathogenicity of novel variants, and (iii) proposes recommendations for analyzing pathogenic variants in African genomes for their use in research and clinical practice. In conclusion, this work proposes criteria to define mutation pathogenicity and actionability in human genetic research and clinical practice in Africa and recommends setting up an African expert panel to oversee the proposed criteria.

Keywords: African genome, incidental findings, actionable variants, whole exome sequencing, whole genome sequencing, precision medicine, pathogenicity

OPEN ACCESS

Edited by:
Tuo Zhang,
Cornell University, United States

Reviewed by:
Erik Garrison,
Wellcome Sanger Institute, United Kingdom
Marcelo Adrian Marti,
University of Buenos Aires, Argentina

****Correspondence:***
Christian Domilongo Bope
christian.bope@uct.ac.za

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *28 February 2019*
Accepted: *05 June 2019*
Published: *25 June 2019*

Citation:
Bope CD, Chimusa ER, Nembaware V, Mazandu GK, de Vries J and Wonkam A (2019) Dissecting in silico Mutation Prediction of Variants in African Genomes: Challenges and Perspectives. Front. Genet. 10:601. doi: 10.3389/fgene.2019.00601

INTRODUCTION

High throughput technologies in "omics" research are expected to improve clinical care globally through genomic medicine. However, the categorization and criteria to infer variants' pathogenicity differs around the world and can pose specific challenges in some populations (Dorschner et al., 2013; Green et al., 2013; MacArthur et al., 2014; Amendola et al., 2015;

Hunter et al., 2016; Ichikawa et al., 2017; Kwak et al., 2017; Lacaze et al., 2017; Tang et al., 2018). Particularly, in African genomes that harbor the highest genetic diversity, it is possible that most *in silico* prediction tools could lead to the highest rate of false positive/negative results (Martin et al., 2018). The H3Africa Consortium has significantly contributed to reducing the dearth of genomic research on the African continent by supporting African genomics researchers and developing policies (Dandara et al., 2014; H3Africa, 2017). However, in the current genomics landscape, it is particularly challenging to interpret some variants found in African genomes, i.e., to determine whether that variant is common or rare, benign or pathogenic. Firstly, approaches to determine the rareness of a variant are based on exploring publicly available genome reference databases in which African data are under-represented (Lek et al., 2016; Popejoy and Fullerton, 2016). In addition, most of the current well-established bioinformatics tools, variant calling pipelines, are benchmarked using non-African populations and most of the variants deposited in the public database are from non-African populations (Pabinger et al., 2012; Bao et al., 2014). Secondly, the high genetic diversity of African populations means that genomic studies are likely to detect many novel variants that are yet to be described in current public databases (Lebeko et al., 2017). Thirdly, there is a lack of evidence-based policies and guidelines to inform the characterization of actionable genes in African genomic research. A guideline on feeding back findings was recently developed by H3Africa; while this is a commendable achievement, it lacks the support of published empirical evidence[1]. This latter point is particularly important given the recent call from the American College of Medical Genetics (ACMG) to investigate pathogenic variants in so-called actionable genes that could potentially have direct clinical benefit, and to return the results to research participants (ACMG, 2013). This will open up a series of ethically relevant questions (Kiezun et al., 2012; MacArthur et al., 2014; Parker and Kwiatkowski, 2016), such as the definition of actionability and relevance to personalized medicine in a context of often scarce human and material resources, and ill-equipped healthcare systems (Masimirembwa et al., 2014).

To address these multiple challenges, and particularly that of variant interpretation in African genomes, it is appropriate to develop new pipelines using African genetics data or to benchmark existing bioinformatics pipeline tools using African populations to account for African genetic diversity. This paper aims to (i) provide an inventory of existing Whole Exome/Genome data from Africans that could help develop an African reference genome build, improve reference panels, and explore the frequency of pathogenic variants in actionable genes and related challenges; (ii) review available *in silico* prediction mutation tools and criteria for categorization of pathogenicity of novel variants; and (iii) propose recommendations for analyzing pathogenic variants in African genomes for their use in research and clinical practice.

CURRENT CHALLENGES OF WES/WGS DATA INTERPRETATION IN AFRICANS

Mastering of genome sequencing pipelines and downstream analysis are important for inferring meaningful information, such as detection of variants in medically relevant genes, from high throughput data such as Next Generation Sequencing (NGS), Whole Exome Sequencing (WES), or Whole Genome Sequencing (WGS). However, data processing, deep sequencing, and meticulous downstream analysis of WES/WGS still constitute a challenge in most of the current pipelines and tools. In addition, there are still some challenges, such as the interpretation of rare missense variants, reliability, and accuracy of pipelines for sequence alignments, variant calling, and data analysis, for the WES and WGS data of African populations (Wang et al., 2013; Rabbani et al., 2014; Bertier et al., 2016; Popejoy and Fullerton, 2016). To address some of these challenges, a plethora of bioinformatics algorithms and pipelines have been developed (Pabinger et al., 2012; Hentzsche et al., 2016; Xu, 2018). Current practice is to use existing variant calling pipelines, but this raises a number of questions, including how are universally reliable and accurate current WES/WGS bioinformatics tools and pipelines benchmarked using non-African data? What is the true proportion of African population data in the current reference genome builds that are publicly available, taking into account the variable level of admixture of African Americans who tend to be considered proxies of Africans in these databases [the Genome Reference Consortium Human Genome (GRCh3) and University of California, Santa Cruz (UCSC)] (Kuhn et al., 2009; Fujita et al., 2011; Leipzig, 2017)? Addressing these challenges will require that genomic research communities from the African continent develop an African benchmark bioinformatics pipeline to analyze genomic data that includes genetic diversity found in the African populations, and engage in a major effort in constructing an African-specific reference panel.

African populations in current reference panels are not representative of more differentiated population groups within Africa. Variant calling from NGS data is based on alignment to a single reference genome, which is problematic for diverse regions or populations, such as African populations. There is great opportunity in improving read alignment and variant calling for African genomes. A genome reference graph for alignment and variant calling may capture natural variation among populations, particularly populations of high diversity with low level of linkage disequilibrium.

Repetitive DNA sequences are abundant in a broad range of species, from bacteria to mammals, and they cover nearly half of the human genome. The other main issue is that repeats have always presented technical challenges for sequence alignment and assembly programs. NGS projects, with their short read lengths and high data volumes, have made these challenges more difficult. From a computational perspective, repeats create ambiguities in

[1]https://h3africa.org/wp-content/uploads/2018/05/H3Africa%20Feedback%20of%20Individual%20Genetic%20Results%20Policy.pdf

alignment and assembly, which, in turn, can produce biases and errors when interpreting results. Simply ignoring repeats is not an option, as this creates problems of its own and may mean that important biological phenomena are missed. Variation in repeats can alter the expression of genes, and changes in the number of repeats have been linked to certain human diseases. Unfortunately, the molecular characterization of these repeats has been hampered by technical limitations related to cloning, sequencing techniques, and alignment algorithms (Dilthey et al., 2014; Marcus et al., 2014; Church et al., 2015; Paten et al., 2017).

Fortunately, the number of genomic researchers in Africa is on the rise, which has led to an increase in African genomic data and publications (The H3Africa Consortium, 2014; Uthman et al., 2015; Mulder et al., 2016; Ndiaye Diallo et al., 2017). The increase in African genomic research has the potential to narrow the research gap between Africa and the rest of the world and can also improve implementation of genomic medicine. Therefore, we propose to use the available data to (i) develop Bioinformatics tools using African data, particularly for populations from sub-Saharan Africa who have the highest genetic diversity and low levels of admixture with European or Asian populations; (ii) benchmark existing tools using available African population data; and (iii) there is an urgent need for a centralized repository of publicly available African genomic data with annotated variants based on their pathogenicity, in order to increase our understanding of continental genomic diversity (Jongeneel et al., 2017; Mulder et al., 2017; Ahmed et al., 2018). To help initiate such endeavors we have provided here an inventory of African Whole Exome and Whole Genome data that are currently available to our knowledge (**Table 1**).

IN SILICO PREDICTION OF MUTATIONS AND CHALLENGES

The accuracy of variant calling pipelines (Li et al., 2009; DePristo et al., 2011; Wei et al., 2011; Garrison and Marth, 2012; Koboldt et al., 2012; Wilm et al., 2012; Lai et al., 2016) is a major step prior to the downstream *in silico* prediction of mutations. Nevertheless, a challenge remains in downstream NGS variant calling analysis, i.e., to distinguish pathogenic mutations and rare non-pathogenic variants from most of the annotating variant calling pipelines. The accuracy of *in silico* prediction of rare and actionable disease-causing genetic variants for the detection of pathogenic rare mutations and polymorphisms is the greatest challenge. Variant calling pipelines generate large numbers variations erroneously, which may contain rare, common genetic variants, false positives, and false negatives (Dong et al., 2015). Further downstream analysis such as variant annotations, variant filtrations, and prioritization methods are conducted to annotate variant genomic features, gene symbols, exonic functions, and amino acid modifications (Bao et al., 2014). Different *in silico* prediction algorithms are implemented to annotate disease-causing mutations based on the following information from the variants: (i) sequence homology (Reva et al., 2011), (ii) protein structure (Ng and Henikoff, 2006; Teng et al., 2009), (iii) evolutionary conservation (Cooper et al., 2010), (iv) the frequency of pathogenicity (Kobayashi et al., 2017), and (v) change in ancestry. Most of the *in silico* prediction methods interact with public databases to incorporate updated variant information in order to enhance annotation prediction efficiency. The incorporated information is mainly the minor allele frequency (MAF), experimental clinical assay information and deleterious prediction of variants (Pabinger et al., 2012). The majority of *in silico* prediction tools provide a reduced number of annotations from large background errors of detected variants. To annotate, filter, and prioritize accurately variant calling, researchers developed pipelines combined with different annotation tools and databases. Germline and somatic mutation databases, such as ANNOVAR (Wang et al., 2010; Yang and Wang, 2015), Human Gene Mutation Database[2], dbSNP[3] (Sherry et al., 2001), and GENEKEEPER[4] and others are important for evaluating variants. Liu et al. (2011) developed a robust database called dbNSFP, which combines the prediction scores of six prediction algorithms namely SIFT (Kumar et al., 2009), PolyPhen-2 (Adzhubei et al., 2010), LRT (Chun and Fay, 2009), MutationTaster (Schwarz et al., 2010), Mutation Assessor (Reva et al., 2011), FATHMM (Chun and Fay, 2009; Shihab et al., 2013), and conservative score tools namely GERP++ (Davydov et al., 2010), SiPhy (Garber et al., 2009), and PhyloP (Doerks et al., 2002) and then compiles the scores of these tools into one (Liu et al., 2011). ClinVar is a commonly used database for germline variants, namely pathogenic and benign and provides related clinical and experimental information[5] (Landrum et al., 2016).

After annotation, it is recommended to filter annotated variants from many tools using two approaches (i) free hypothesis, to cast the vote of the annotated variant filters for "Deleterious or damaging disease-causing (D)" or "disease-causing automatic (A)" among annotation prediction tools based on a defined cut-off (∼50%); and (ii) non-free hypothesis, which provides a list of known genes of the studies with another level of prediction cut-off (∼25%). The cut-off for both hypotheses is study related.

In silico prediction of mutations in the context of African populations introduces additional specific challenges that are partly related to the use of non-African populations to benchmark *in silico* prediction pipelines and the low proportion of African population data in most of the interrogated databases. Another challenge when working with African population data is the annotation of common variants specific to African populations, which can be considered as pathogenic variants when using public databases. This emphasizes the need for a guideline, which defines approaches to infer pathogenicity variants in African populations.

Predicting Pathogenic Variants and Challenges

In the literature and in most annotation databases, the classification of pathogenicity differs (Sherry et al., 2001;

[2]http://www.hgmd.cf.ac.uk/ac/index.php
[3]https://www.ncbi.nlm.nih.gov/projects/SNP/
[4]https://kewinc.com/analytics/
[5]https://www.ncbi.nlm.nih.gov/clinvar/

TABLE 1 | Published whole exome and genomes data from sub-Saharan Africa.

Country	Region	Individuals	References
	Exomes		
Botswana	Southern Africa	164	Retshabile et al., 2018
Uganda	Southern Africa	150	Retshabile et al., 2018
Ghanaian	Western Africa	1032	Kodaman et al., 2017
Tunisia	North Africa	7 19	Hamdi et al., 2018 Ben Rekaya et al., 2018
Morocco	North Africa	3	Bousfiha et al., 2017
Cameroonian	Central Africa	179	Sickle Cell Disease Project (Unpublished)
Congo	Central Africa	23	Sickle Cell Disease Project (Unpublished)
Black Xhosa (SA)	Southern Africa	25	Hearing Impairment Project (Unpublished)
African Caribbean (ACB)	Caribbean	98	1000 Genomes project
Esan in Nigeria (ESN)	Western Africa	111	1000 Genomes project
Mende in Sierra Leone (MSL)	Western Africa	98	1000 Genomes project
Yoruba in Ibadan, Nigeria (YRI)	Western Africa	100	1000 Genomes project
Luhya in Webuye, Kenya (LWK)	East Africa	106	1000 Genomes project
African Ancestry in Southwest USA (ASW)	North America	68	1000 Genomes project
Gambia in Western Division, The Gambia (GWD)	Western Africa	120	1000 Genomes project
African American	North America	761	Auer et al., 2012
	Genomes		
Baganda, Banyarwanda, Barundi (Uganda); Luhya, Kikuyu, Kalenjin (Kenya); Sotho, Zulu (South Africa); Yoruba, Igbo (Nigeria); Ga-Adangbe (Ghana); Jola, Fula, Wolof, Mandika (Gambia); Amhara, Oromo, Somali (Ethiopia)	Sub Sahara (Eastern Africa, Southern Africa, Western Africa, Southern Africa)	320	Gurdasani et al., 2015
Kombo (Gambia)	Western Africa	2560	Jallow et al., 2009
South Africa	Southern Africa	13	Kramvis et al., 2002
Ovamboland (Namibia) Angola Madagascar	Southern Africa Southern Africa Madagascar	23 2 1	Kramvis et al., 2005
Algeria, Morocco, Libya, Tunisia, Tuareg Congo, Gabon, Cameroun, Nigeria	Northern Africa Sub Sahara	25 59	Fadhlaoui-Zid et al., 2013
Uganda Zimbabwean	Sub Sahara	112 174	Venner et al., 2016
Mandinka II, Serehule, Bambara, Malike, FulaII, FulaI, MandikaI, WollofI, Serere, Manjogo, Jola Mossi, Kasem, Yoruba, Namkam, Semi-Bantu, Akans, Bantu Kauma, Chonyi, Wabondel, Kambe, Luhya, Maasai, Wasambaa, Giriama, Mzigua Ari, Anuak, Sudanese, Gumuz Oromo, Somali, Wolayta, Afar, Tigray, Amhara Nama, Karretjie, Khomani, Malawi, Herero, Khwe, Ixu, HU/'Hoansi, Amaxhosa Sebantu	West African Niger-Congo Central West African Niger-Congo East Africa Niger-Congo East Africa Nilo-Saharan East Africa Afroasiatic Khoesan	2504	Busby et al., 2016
African American	African American	35370	Ng et al., 2014
Ethiopian (Weth)	Sub Sahara	120	Tekola-Ayele et al., 2015
South Africa	Southern Africa	24 (8 Colored, 16 Black)	Choudhury et al., 2017

Wang et al., 2010; Yang and Wang, 2015; Landrum et al., 2016; McLaren et al., 2016). Nevertheless, a common strategy to define pathogenicity involves combining results from many annotation pipelines (Lebeko et al., 2017). Further downstream analyses are gene network analysis and gene enrichment. The purpose of these analyses is to investigate the level of interactions between genes and the annotated variants associated with human phenotypes and then mine affected biological processes, networks, pathways, and molecular functions (Bindea et al., 2009; Warde-Farley et al., 2010; Lebeko et al., 2017).

In the comprehensive standards and guidelines, ACMG and the Association for Molecular Pathology (AMP) define the nomenclature for variants (**Table 2**). Recommendations for laboratories and clinicians to return incidental findings (IFs) has led to interest toward defining criteria and mechanisms for evaluating pathogenicity and the frequencies of IFs in different populations. For example, Dorschner et al. (2013) analyzed actionable pathogenic variants in 500 European and 500 participants of African descent using exome data. The classifications for pathogenicity (**Table 2**) included allele

TABLE 2 | Variants pathogenicity categorization.

References	Categorization	Interpretation
Dorschner et al., 2013	Pathogenic	Allele frequency of the identified variant is below cutoff AND segregation can be found in at least two unrelated families
	Likely pathogenic variant of uncertain significance (VUS)	Allele frequency of the identified variant is below cutoff AND identified in at least three unrelated individuals
	VUS	Allele frequency of the discovered variant is below cutoff AND present in less than three unrelated affected individuals
	Likely benign VUS	Allele frequency of the identified variant is below cutoff AND/OR seen in combination with a known pathogenic mutation
Richards et al., 2015 ACGM and AMP	Pathogenic very strong	Null variant (non-sense, frameshift, initiation codon, single or multiexon deletion) in a gene is a known mechanism of disease
	Pathogenic strong	Similar amino acid modification which was previously considered as pathogenic variant independent of nucleotide change
	Pathogenic moderate	Localize in a mutational critical region and very important functional domain without benign variation
	Pathogenic supporting	Cosegregation with disease located in many affected family members in a gene well known to be the cause the disease
	Benign standalone	Allele frequency is greater than 5% in Exome Sequencing Project, 1000 Genomes Project, or Exome Aggregation Consortium
	Benign strong	Allele frequency is greater than expected for disorder
	Benign supporting	Missense variant in a gene for which primarily truncating variants are known to cause disease

frequency of the variants, segregation evidence, and the number of the patients affected with the variants and their status as a *de novo* mutation. The results showed major discrepancies in the frequencies of pathogenic variants among Europeans versus Africans, with an estimated frequency of ∼3.4% for those of European descent and ∼1.2% for those of African descent. In a similar study, Amendola et al. (2015) investigated IFs in 6503 with 4300 Europeans and 2203 individuals of African descent. In addition, functionally disruptive variant categories were added which represent the expected pathogenic variants as truncating and misplace-causing variants. To validate the results, a comparative analysis was conducted with other clinical and research genetic laboratories and *in silico* pathogenicity scores. The results also showed that those of African descent had a scientifically lower proportion (nearly 50%) of a pathogenic variant in actionable genes compared to European participants. This lower proportion found in both studies could be due to the underrepresentation of populations of African descent in the literature and publicly available databases.

Taking into account the high level of admixture of European ancestry among African Americans and the highest level of diversity among Africans, and poor representativity in public databases as well little clinical genetic research from Africa that is publicly available, it is likely that a similar study could even lead to a much lower proportion of IFs in sub-Saharan African populations. This indicates that there is an urgent need to improve criteria to categorize the pathogenicity when studying African populations, stressing for example investigating an appropriate number of ethnically matched control populations.

Variants Actionability and Challenges

The Clinical Genome Resource (ClinGen) defines actionability as clinically prescribed interventions specific to the genetic disorder under consideration that is effective for prevention or delay of clinical disease, lowered clinical burden, or improved clinical outcomes in a previously undiagnosed adult and suggested a metric to score clinical actionability (Hunter et al., 2016). Interventions include patient management (e.g., risk-reducing surgery), surveillance, or specific circumstances the patients should avoid (e.g., certain types of anesthesia). The actionability includes interventions to improve outcomes for at-risk family members. Genetic testing recommendations for at-risk family members alone, however, were not considered sufficient to meet the criteria for actionability. In addition, actionability did not include reproductive decision-making.

Alternatively, the 100,000 Genomes Project protocol defines actionable genes as variants with a significant potential to prevent disease morbidity and mortality, if identified before symptoms become apparent. The variants with potentially severe impacts are clinically actionable causes of rare disease, where a healthcare intervention or screening programs might prevent an untoward outcome. The variants are known to result in illness or disability that is clinically significant, severely or moderately life threatening and clinically actionable. It should be emphasized that the exact criteria for considering whether a variant is considered actionable or not, and serious or not, is context-dependent and in some instances only emerges during the process of seeking ethical approval for the study (Genomics England, 2017).

The accepted process consists of defining actionability of the variant and a pathogenicity classification criterion. Both processes are evaluated, inspected and validated by a group of experts (Richards et al., 2015; Hunter et al., 2016). In the African context with highly genetically diverse populations, there is a need to update the proposed scoring metric to take into account the scarcity of health care professionals with medical genetics and

genetic counseling skills, poorly equipped health facilities with a major disparity between urban and rural setting, and generally inadequate health systems.

RETURN OF INCIDENTAL FINDINGS AND CHALLENGES IN AFRICA

Next Generation Sequencing analysis could contribute to the improvement of patient care. This development has blurred the line between genomics and healthcare; the global recommendations on the identification and the return of IFs have raised some ethical concerns for genomic researchers, clinicians, and the public health authority. Prior to returning IFs, there is a need to have clear guidelines and recommendations on a list of potentially actionable genes and define how, what and when IFs should be returned (Ness, 2008; ACMG, 2013, 2015; Souzeau et al., 2016; Nowak et al., 2018). Wolf et al. (2008) published a paper, proposing a framework supporting disclosure of IFs to guide researchers particularly on informed consent, the handling process and the responsibility of institutional review boards. The process on informed consent regarding incidental findings returns is a separate ethical debate that will require appropriate consideration by various stakeholders through, for example, an African and international experts panel meeting with the aim to address (a) the definition of actionability in the context of Africa, (b) the priority list of conditions and related gene variants that are actionable in Africa, (c) the criteria for molecular validation of the variants found in genomic research for clinical use, (d) the clinical environment necessary for returning such results and by which category of health professionals, as most African settings do not have medical genetic services, and (e) the process of wording and integrating informed consent for incidental findings in genomic research in Africa. In the United States, the ACMG has provided a guideline and recommendations to evaluate the cost-effectiveness of returning pathogenic variants for 56 specific genes considered medically actionable (ACMG, 2013, 2015). In Europe, the EuroGenTest and the European Society of Human Genetics recently presented guidelines for diagnostic NGS, including a rating system for diagnostic tests (Matthijs et al., 2016). In the United Kingdom, the Association for Clinical genetic Science (ACGS) has also released a guideline for the evaluation of pathogenicity and reporting of sequence variants in clinical molecular genetics (Wallis et al., 2013). To the best of our knowledge, there are no evidence-based recommendations for African researchers and clinicians on how to report IFs (de Vries and Pepper, 2012; Sookrajh et al., 2015). This is not a surprise due to the fact that African populations and the diaspora are underrepresented in most of the genetics studies, which questions the universal applicability of the genetic findings in large genome studies, disease association and evolutionary genetic studies (Need and Goldstein, 2009; Rosenberg et al., 2010; Dorschner et al., 2013; Tiffin, 2014; Manrai et al., 2016).

Prior to proposing guidelines on the return of IFs for the African populations, researchers and clinicians should first conduct multiple genetics studies to characterize the nature of genes for both monogenic and complex diseases on multiple African populations. The results of such studies should first identify the frequency of pathogenic variants in actionable gene lists as defined, e.g., by the ACMG, annotate, and filter genes. An expert panel should validate the list of pathogenic and actionable variants, then conduct a comparative analysis with results from non-African populations (ACMG, 2013; Green et al., 2013; Kalia et al., 2017). The next step could be to define novel actionable genes and variants that are relevant to Africa, e.g., sickle cell disease or *APOL1* variants. Only after completing the aforementioned steps, African researchers and clinicians will be able to provide a comprehensive and clear guideline on which putative pathogenic genes may be returned. It should be noted that the framework on the return of IFs should covert different aspects such as ethical guidelines and genetic counseling. Due to the high diversity in the African population, the classification of pathogenic and actionable variants for the return of secondary findings is more challenging due to the following additional factors: (i) contextualizing the African definition of pathogenicity and actionable genes, (ii) the choice of control cohort for the validation among African populations (iii) the power of the sample size for the case and control cohort, and (iv) a list of actionable genes of the most prominent diseases in the African populations. These questions need to be considered and addressed prior to the development of African actionable gene standards and guideline for IFs. The guidelines and the list of African populations' actionable genes to be returned as IFs is a major milestone toward personalized medicine.

CONCLUSION AND PERSPECTIVES

The power of high-throughput genomic technologies, particularly DNA sequencing, has potential to bridge the gap between genomic research and clinical care. However, this blurry line has opened several technical and ethical questions and concerns, especially in the context of African genomic research. With the highest genetic diversity found in individuals and communities across the African continent, the use of personalized medicine will be beneficial both to the continent and worldwide. The state of WES and WGS on the continent is in the early stages in terms of available genetic data, publications on genetic conditions, appropriately designed pipelines and bioinformatics tools. The process of handling IFs should be clearly discussed and defined by the African research community, clinicians, specifically on the categorization of the pathogenicity, and actionability of genes and variants in order to take advantage of the genomic technology.

We have provided a list of available WES and WGS data that can help in initiating, the development of bioinformatics pipelines suitable for African population genomic data, quantify the frequency of pathogenic and so-called actionable genes, and to develop appropriate policies for their investigation in genomic research. This requires African researchers and experts to be encouraged to share and make data available in public databases. This once again is an urgent call to set an African expert panel to categorize and refine criteria for pathogenicity

and African actionability in human genetic research in Africa. We recommend that experts should prioritize the following steps: (1) define better criteria for classification of pathogenicity, and actionability, including relevant genes lists, that can be explored and return as IFs to research participant in Africa; (2) benchmark existing variant calling and *in silico* prediction pipelines for African genomic data or develop new pipelines using African data; (3) use hypothesis and non-hypothesis approaches *in silico* mutation prediction to avoid false positive mutation; (4) develop an African reference panel; and (5) Sanger sequencing to be done on the new variants for validation.

DATA AVAILABILITY

All datasets analyzed for this study are cited in the manuscript and the Supplementary Files.

AUTHOR CONTRIBUTIONS

All authors contributed in conceiving, preparing, and revising the manuscript, approved the manuscript, and agreed to be accountable for all aspects of the presented work.

FUNDING

This work was supported by NIH Common Fund H3Africa Initiative, award number U54HG009790 to AW and JdV, Wellcome Trust/AAS Ref: H3A/18/001 to AW, and the NIH, United States, Grant Numbers U01HG009716, U01HG007459, and NIH/NHLBI U24HL135600 to AW. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

ACMG (2013). ACMG practice guidelines: incidental findings in clinical genomics: a clarifcation. *Genet. Med.* 15, 664–666.

ACMG (2015). ACMG policy statement: updated recommendations regarding analysis and reporting of secondary findings in clinical genome-scale sequencing. *Genet. Med.* 17, 68–69. doi: 10.1038/gim.2013.82

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., et al. (2010). A method and server for predicting damaging missense mutations. *Nat. Methods* 7, 248–249. doi: 10.1038/nmeth0410-248

Ahmed, A. E., Mpangase, P. T., Panji, S., Baichoo, S., Botha, G., Fadlelmola, F. M., et al. (2018). Organizing and running bioinformatics hackathons within Africa: the H3ABioNet cloud computing experience. *AAS Open Res.* 1, 1–14. doi: 10.12688/aasopenres.12847.1

Amendola, L. M., Dorschner, M. O., Robertson, P. D., Salama, J. S., Hart, R., Shirts, B. H., et al. (2015). Actionable exomic incidental findings in 6503 participants-challenges of variant classification. *Genome Res.* 25, 305–315. doi: 10.1101/gr.183483.114

Auer, P. L., Johnsen, J. M., Johnson, A. D., Logsdon, B. A., Lange, L. A., Nalls, M. A., et al. (2012). Imputation of exome sequence variants into population- based samples and blood-cell-trait-associated loci in African Americans: NHLBI GO exome sequencing project. *Am. J. Hum. Genet.* 91, 794–808. doi: 10.1016/j.ajhg.2012.08.031

Bao, R., Huang, L., Andrade, J., Tan, W., Kibbe, W. A., Jiang, H., et al. (2014).). Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing. *Cancer Inform.* 13, 67–82. doi: 10.4137/CIn.s13779.RECEIvED

Ben Rekaya, M., Naouali, C., Messaoud, O., Jones, M., Bouyacoub, Y., Nagara, M., et al. (2018). Whole Exome Sequencing allows the identification of two novel groups of Xeroderma pigmentosum in Tunisia. XP-D and XP-E: impact on molecular diagnosis. *J. Dermatol. Sci.* 89, 172–180. doi: 10.1016/j.jdermsci.2017.10.015

Bertier, G., Hetu, M., and Joly, Y. (2016). Unsolved challenges of clinical whole-exome sequencing: a systematic literature review of end-users' views. *BMC Med. Genomics.* 9:52. doi: 10.1186/s12920-016-0213-6

Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., et al. (2009). ClueGO: a cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* 25, 1091–1093. doi: 10.1093/bioinformatics/btp101

Bousfiha, A., Bakhchane, A., Charoute, H., Detsouli, M., Rouba, H., Charif, M., et al. (2017). Novel compound heterozygous mutations in the GPR98 (USH2C) gene identified by whole exome sequencing in a Moroccan deaf family. *Mol. Biol. Rep.* 44, 429–434. doi: 10.1007/s11033-017-4129-9

Busby, G. B., Band, G., Si Le, Q., Jallow, M., Bougama, E., Mangano, V. D., et al. (2016). Admixture into and within sub-Saharan Africa. *Elife* 5:e15266. doi: 10.7554/eLife.15266

Choudhury, A., Ramsay, M., Hazelhurst, S., Aron, S., Bardien, S., Botha, G., et al. (2017). Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. *Nat. Commun.* 8:2062. doi: 10.1038/s41467-017-00663-9

Chun, S., and Fay, J. C. (2009). Identification of deleterious mutations within three human genomes. *Genome Res.* 19, 1553–1561. doi: 10.1101/gr.092619.109

Church, D. M., Schneider, V. A., Steinberg, K. M., Schatz, M. C., Quinlan, A. R., Chin, C. S., et al. (2015). Extending reference assembly models. *Genome Biol.* 16:13. doi: 10.1186/s13059-015-0587-3

Cooper, G. M., Goode, D. L., Ng, S. B., Sidow, A., Bamshad, M. J., Shendure, J., et al. (2010). Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nat. Methods* 7, 250–251. doi: 10.1038/nmeth0410-250

Dandara, C., Huzair, F., Borda-Rodriguez, A., Chirikure, S., Okpechi, I., Warnich, L., et al. (2014). H3Africa and the African life sciences ecosystem: building sustainable innovation. *OMICS* 18, 733–739. doi: 10.1089/omi.2014.0145

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* 6:e1001025. doi: 10.1371/journal.pcbi.1001025

de Vries, J., and Pepper, M. (2012). Genomic sovereignty and the African promise: mining the African genome for the benefit of Africa. *J. Med. Ethics* 38, 474–478. doi: 10.1136/medethics-2011-100448

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–498. doi: 10.1038/ng.806492VOLUME

Dilthey, A., Cox, C. J., Iqbal, Z., Cox, C., Nelson, M. R., and McVean, G. (2014). Improved genome inference in the MHC using a population reference graph. *BioRxiv*

Doerks, T., Copley, R. R., Schultz, J., Ponting, C. P., and Bork, P. (2002). Systematic identification of novel protein domain families associated with nuclear functions. *Genome Res.* 12, 47–56. doi: 10.1101/gr.203201

Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., et al. (2015). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum. Mol. Genet.* 24, 2125–2137. doi: 10.1093/hmg/ddu733

Dorschner, M. O., Amendola, L. M., Turner, E. H., Robertson, P. D., Shirts, B. H., Gallego, C. J., et al. (2013). Actionable, pathogenic incidental findings in 1,000 participants' exomes. *Am. J. Hum. Genet.* 93, 631–640. doi: 10.1016/j.ajhg.2013.08.006

Fadhlaoui-Zid, K., Haber, M., Martinez-Cruz, B., Zalloua, P., Benammar Elgaaied, A., and Comas, D. (2013). Genome-wide and paternal diversity reveal a recent origin of human populations in North Africa. *PLoS One* 8:e80293. doi: 10.1371/journal.pone.0080293

Fujita, P. A., Rhead, B., Zweig, A. S., Hinrichs, A. S., Karolchik, D., Cline, M. S., et al. (2011). The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.* 39, D876–D882. doi: 10.1093/nar/gkq963

Garber, M., Guttman, M., Clamp, M., Zody, M. C., Friedman, N., and Xie, X. (2009). Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* 25, i54–i62. doi: 10.1093/bioinformatics/btp190

Garrison, F., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv*

Genomics England (2017). *Genomics England K1000 Project Protocol*. Charterhouse Square: Genomics England.

Green, R. C., Berg, J. S., Grody, W. W., Kalia, S. S., Korf, B. R., Martin, C. L., et al. (2013). ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet. Med.* 15, 565–574. doi: 10.1038/gim.2013.73

Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., et al. (2015). The African genome variation project shapes medical genetics in Africa. *Nature* 517, 327–332. doi: 10.1038/nature13997

H3Africa (2017). *Ethics and Governance Framework for Best Practice in Genomic Research and Biobanking in Africa*. Available at: https://www.sun.ac.za/english/faculty/healthsciences/rdsd/Documents/Final%20Framework%20for%20African%20genomics%20and%20biobanking_SC_February%202017%20II%20.pdf

Hamdi, Y., Boujemaa, M., Ben Rekaya, M., Ben Hamda, C., Mighri, N., El Benna, H., et al. (2018). Family specific genetic predisposition to breast cancer: results from Tunisian whole exome sequenced breast cancer cases. *J. Transl. Med.* 16:158. doi: 10.1186/s12967-018-1504-9

Hentzsche, J. D., Robinson, W. A., and Tan, A. C. (2016). A survey of computational tools to analyze and interpret whole exome sequencing data. *Int. J. Genomics* 2016:7983236. doi: 10.1155/2016/7983236

Hunter, J. E., Irving, S. A., Biesecker, L. G., Buchanan, A., Jensen, B., Lee, K., et al. (2016). A standardized, evidence-based protocol to assess clinical actionability of genetic disorders associated with genomic variation. *Genet. Med.* 18, 1258–1268. doi: 10.1038/gim.2016.40

Ichikawa, H., Nagahashi, M., Shimada, Y., Hanyu, T., Ishikawa, T., Kameyama, H., et al. (2017). Actionable gene-based classification toward precision medicine in gastric cancer. *Genet. Med.* 9:93. doi: 10.1186/s13073-017-0484-3

Jallow, M., Teo, Y. Y., Small, K. S., Rockett, K. A., Deloukas, P., Clark, T. G., et al. (2009). Genome-wide and fine-resolution association analysis of malaria in West Africa. *Nat. Genet.* 41, 657–665. doi: 10.1038/ng.388

Jongeneel, C. V., Achinike-Oduaran, O., Adebiyi, E., Adebiyi, M., Adeyemi, S., Akanle, B., et al. (2017). Assessing computational genomics skills: our experience in the H3ABioNet African bioinformatics network. *PLoS Comput. Biol.* 13:e1005419. doi: 10.1371/journal.pcbi.1005419

Kalia, S. S., Adelman, K., Bale, S. J., Chung, W. K., Eng, C., Evans, J. P., et al. (2017). Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American college of medical genetics and genomics. *Genet. Med.* 19, 249–255. doi: 10.1038/gim.2016.190

Kiezun, A., Garimella, K., Do, R., Stitziel, N. O., Neale, M. B., McLaren, P. J., et al. (2012). Exome sequencing and the genetic basis of complex traits. *Nat. Genet.* 44, 623–630. doi: 10.1038/ng.2303

Kobayashi, Y., Yang, S., Nykamp, K., Garcia, J., Lincoln, S. E., and Topper, S. E. (2017). Pathogenic variant burden in the ExAC database: an empirical approach to evaluating population data for clinical variant interpretation. *Genome Med.* 9:13. doi: 10.1186/s13073-017-0403-7

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., et al. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 22, 568–576. doi: 10.1101/gr.129684.111

Kodaman, N., Sobota, R. S., Asselbergs, F. W., Oetjens, M. T., Moore, J. H., Brown, N. J., et al. (2017). Genetic effects on the correlation structure of CVD risk factors: exome-wide data from a ghanaian population. *Glob. Heart* 12, 133–140. doi: 10.1016/j.gheart.2017.01.013

Kramvis, A., Restorp, K., Norder, H., Botha, J. F., Magnius, L. O., and Kew, M. C. (2005). Full genome analysis of hepatitis B virus genotype E strains from South-Western Africa and Madagascar reveals low genetic variability. *J. Med. Virol.* 77, 47–52. doi: 10.1002/jmv.20412

Kramvis, A., Weitzmann, L., Owiredu, K. B. A., and Kew, C. M. (2002). Analysis of the complete genome of subgroup AHhepatitis Bvirus isolates from South Africa. *Gen. Virol.* 83, 835–839. doi: 10.1099/0022-1317-83-4-835

Kuhn, R. M., Karolchik, D., Zweig, A. S., Wang, T., Smith, K. E., Rosenbloom, K. R., et al. (2009). The UCSC genome browser database: update 2009. *Nucleic Acids Res.* 37, D755–D761. doi: 10.1093/nar/gkn875

Kumar, P., Henikoff, S., and Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* 4, 1073–1081. doi: 10.1038/nprot.2009.86

Kwak, S. H., Chae, J., Choi, S., Kim, M. J., Choi, M., Chae, J. H., et al. (2017). Findings of a 1303 Korean whole-exome sequencing study. *Exp. Mol. Med.* 49:e356. doi: 10.1038/emm.2017.142

Lacaze, P., Ryan, J., Woods, R., Winship, I., and McNeil, J. (2017). Pathogenic variants in the healthy elderly: unique ethical and practical challenges. *J. Med.Ethics* 43, 714–722. doi: 10.1136/medethics-2016-103967

Lai, Z., Markovets, A., Ahdesmaki, M., Chapman, B., Hofmann, O., McEwen, R., et al. (2016). VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.* 44:e108. doi: 10.1093/nar/gkw227

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., et al. (2016). ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* 44, D862–D868. doi: 10.1093/nar/gkv1222

Lebeko, K., Manyisa, N., Chimusa, E. R., Mulder, N., Dandara, C., and Wonkam, A. (2017). A genomic and protein-protein interaction analyses of nonsyndromic hearing impairment in cameroon using targeted genomic enrichment and massively parallel sequencing. *OMICS* 21, 90–99. doi: 10.1089/omi.2016.0171

Leipzig, J. (2017). A review of bioinformatic pipeline frameworks. *Brief. Bioinform.* 18, 530–536. doi: 10.1093/bib/bbw020

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285–291. doi: 10.1038/nature19057

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Liu, X., Jian, X., and Boerwinkle, E. (2011). dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat.* 32, 894–899. doi: 10.1002/humu.21517

MacArthur, D. G., Manolio, T. A., Dimmock, D. P., Rehm, H. L., Shendure, J., Abecasis, G. R., et al. (2014). Guidelines for investigating causality of sequence variants in human disease. *Nature* 508, 469–475. doi: 10.1038/nature13127

Manrai, K. A., Funke, B. H., Rehm, H. L., Olesen, M. S., Maron, B. A., Szolovits, P., et al. (2016). Genetic misdiagnoses and the potential for health disparities. *N Engl. J. Med.* 375, 655–665. doi: 10.1056/NEJMsa1507092

Marcus, S., Lee, H., and Schatz, M. C. (2014). SplitMEM: a graphical algorithm for pan-genome analysis with suffix skips. *Bioinformatics* 30, 3476–3483. doi: 10.1093/bioinformatics/btu756

Martin, A. R., Teferra, S., Moller, M., Hoal, E. G., and Daly, M. J. (2018). The critical needs and challenges for genetic architecture studies in Africa. *Curr. Opin. Genet. Dev.* 53, 113–120. doi: 10.1016/j.gde.2018.08.005

Masimirembwa, C., Dandara, C., and Hasler, J. (2014). "Population diversity and pharmacogenomics in Africa," in *Handbook of Pharmacogenomics and Stratified Medicine*, ed. S. Padmanabhan (Amsterdam: Elsevier Science), 971–998. doi: 10.1016/b978-0-12-386882-4.00043-8

Matthijs, G., Souche, E., Alders, M., Corveleyn, A., Eck, S., Feenstra, I., et al. (2016). Guidelines for diagnostic next-generation sequencing. *Eur. J. Hum. Genet.* 24, 2–5. doi: 10.1038/ejhg.2015.226

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., et al. (2016). The ensembl variant effect predictor. *Genome Biol.* 17:122. doi: 10.1186/s13059-016-0974-4

Mulder, N. J., Adebiyi, E., Adebiyi, M., Adeyemi, S., Ahmed, A., Ahmed, R., et al. (2017). Development of bioinformatics infrastructure for genomics research. *Glob. Heart* 12, 91–98. doi: 10.1016/j.gheart.2017.01.005

Mulder, N. J., Adebiyi, E., Alami, R., Benkahla, A., Brandful, J., Doumbia, S., et al. (2016). H3ABioNet, a sustainable pan-African bioinformatics network for human heredity and health in Africa. *Genome Res.* 26, 271–277. doi: 10.1101/gr.196295.115

Ndiaye Diallo, R., Gadji, M., Hennig, B. J., Gueye, M. V., Gaye, A., Diop, J. P. D., et al. (2017). Strengthening human genetics research in Africa: report of the 9th

meeting of the African society of human genetics in Dakar in May 2016. *Glob. Health Epidemiol. Genom.* 2:e10. doi: 10.1017/gheg.2017.3

Need, A. C., and Goldstein, D. B. (2009). Next generation disparities in human genomics: concerns and remedies. *Trends Genet.* 25, 489–494. doi: 10.1016/j.tig.2009.09.012

Ness, B. V. (2008). Genomic research and incidental findings. *J. Law Med. Ethics* 36, 292–312. doi: 10.1111/j.1748-720X.2008.00272.x

Ng, M. C., Shriner, D., Chen, B. H., Li, J., Chen, W. M., Guo, X., et al. (2014). Meta-analysis of genome-wide association studies in African Americans provides insights into the genetic architecture of type 2 diabetes. *PLoS Genet.* 10:e1004517. doi: 10.1371/journal.pgen.1004517

Ng, P. C., and Henikoff, S. (2006). Predicting the effects of amino acid substitutions on protein function. *Annu. Rev. Genomics Hum. Genet.* 7, 61–80. doi: 10.1146/annurev.genom.7.080505.115630

Nowak, K. J., Bauskis, A., Dawkins, H. J., and Baynam, G. (2018). Incidental inequity. *Eur. J. Hum. Genet.* 26, 616–617. doi: 10.1038/s41431-018-0101-y

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., et al. (2012). A survey of tools for variant analysis of next-generation genome sequencing data. *Brief. Bioinform.* 15, 256–278. doi: 10.1093/bib/bbs086

Parker, M., and Kwiatkowski, D. P. (2016). The ethics of sustainable genomic research in Africa. *Genome Biol.* 17:44. doi: 10.1186/s13059-016-0914-3

Paten, B., Novak, M. A., Eizenga, M. J., and Garrisson, E. (2017). Genome graphs and the evoluation of genome inference. *Genome Res* 27, 665–676. doi: 10.1101/gr.214155.116

Popejoy, A. B., and Fullerton, S. M. (2016). Genomics is failing on diversity. *Nature* 538, 161–164. doi: 10.1038/538161a

Rabbani, B., Tekin, M., and Mahdieh, N. (2014). The promise of whole-exome sequencing in medical genetics. *J. Hum. Genet.* 59, 5–15. doi: 10.1038/jhg.2013.114

Retshabile, G., Mlotshwa, B. C., Williams, L., Mwesigwa, S., Mboowa, G., Huang, Z., et al. (2018). Whole-exome sequencing reveals uncaptured variation and distinct ancestry in the Southern African population of Botswana. *Am. J. Hum. Genet.* 102, 731–743. doi: 10.1016/j.ajhg.2018.03.010

Reva, B., Antipin, Y., and Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* 39:e118. doi: 10.1093/nar/gkr407

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American college of medical genetics and genomics and the association for molecular pathology. *Genet. Med.* 17, 405–424. doi: 10.1038/gim.2015.30

Rosenberg, N. A., Huang, L., Jewett, E. M., Szpiech, Z. A., Jankovic, I., and Boehnke, M. (2010). Genome-wide association studies in diverse populations. *Nat. Rev. Genet.* 11, 356–366. doi: 10.1038/nrg2760

Schwarz, J. M., Rodelsperger, C., Schuelke, M., and Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods* 7, 575–576. doi: 10.1038/nmeth0810-575

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., et al. (2001). dbSNP- the NCBI database of genetic variation. *Nucleic Acids Res.* 29, 308–311. doi: 10.1093/nar/29.1.308

Shihab, H. A., Gough, J., Cooper, D. N., Day, I. N., and Gaunt, T. R. (2013). Predicting the functional consequences of cancer-associated amino acid substitutions. *Bioinformatics* 29, 1504–1510. doi: 10.1093/bioinformatics/btt182

Sookrajh, Y., Naidoo, S., Ramjee, G., and Team, M. D. P. (2015). Shared responsibility for ensuring appropriate management of incidental findings: a case study from South Africa. *J. Med. Ethics* 41, 281–283. doi: 10.1136/medethics-2013-101561

Souzeau, E., Burdon, K. P., Mackey, D. A., Hewitt, A. W., Savarirayan, R., Otlowski, M., et al. (2016). Ethical Considerations for the Return of Incidental Findings in Ophthalmic Genomic Research. *Transl. Vis. Sci. Technol.* 5, 1–11. doi: 10.1167/tvst.5.1.3

Tang, C.-M., Dattani, S., So, M.-T., Cherny, S. S., Tam, P. K. H., Sham, P. C., et al. (2018). Actionable secondary findings from whole-genome sequencing of 954 East Asians. *Hum. Genet.* 137, 31–37. doi: 10.1007/s00439-017-1852-1

Tekola-Ayele, F., Adeyemo, A., Aseffa, A., Hailu, E., Finan, C., Davey, G., et al. (2015). Clinical and pharmacogenomic implications of genetic variation in a Southern Ethiopian population. *Pharmacogenomics J.* 15, 101–108. doi: 10.1038/tpj.2014.39

Teng, S., Madej, T., Panchenko, A., and Alexov, E. (2009). Modeling effects of human single nucleotide polymorphisms on protein-protein interactions. *Biophys. J.* 96, 2178–2188. doi: 10.1016/j.bpj.2008.12.3904

The H3Africa Consortium (2014). Enabling the genomic revolution in Africa. *Science* 344, 1347–1348. doi: 10.1126/science.1251546

Tiffin, N. (2014). Unique considerations for advancing genomic medicine in African populations. *Per. Med.* 11, 187–196. doi: 10.2217/pme.13.105

Uthman, O. A., Wiysonge, C. S., Ota, M. O., Nicol, M., Hussey, G. D., Ndumbe, P. M., et al. (2015). Increasing the value of health research in the WHO African Region beyond 2015–reflecting on the past, celebrating the present and building the future: a bibliometric analysis. *BMJ Open* 5:e006340. doi: 10.1136/bmjopen-2014-006340

Venner, C. M., Nankya, I., Kyeyune, F., Demers, K., Kwok, C., Chen, P. L., et al. (2016). Infecting HIV-1 Subtype Predicts Disease Progression in Women of Sub-Saharan Africa. *EBioMedicine* 13, 305–314. doi: 10.1016/j.ebiom.2016.10.014

Wallis, Y., Payne, S., McAnulty, C., Bodmer, D., Sistermans, E., Robertson, K., et al. (2013). *Practice Guidelines for Evaluations of Pathogenecity and the Reporting of Sequencing Variants in Clinical Molecular Genetics*. Birmingham: Association for Clinical Genetic Science.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38:e164. doi: 10.1093/nar/gkq603

Wang, Z., Liu, X., Yang, B.-Z., and Gelernter, J. (2013). The role and challenges of exome sequencing in studies of human diseases. *Front. Genet.* 4:160. doi: 10.3389/fgene.2013.00160

Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., et al. (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* 38, W214–W220. doi: 10.1093/nar/gkq537

Wei, Z., Wang, W., Hu, P., Lyon, G. J., and Hakonarson, H. (2011). SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Res* .39:e132. doi: 10.1093/nar/gkr599

Wilm, A., Aw, P. P., Bertrand, D., Yeo, G. H., Ong, S. H., Wong, C. H., et al. (2012). LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res.* 40, 11189–11201. doi: 10.1093/nar/gks918

Wolf, S. M., Lawrenz, F. P., Nelson, C. A., Kahn, J. P., Cho, M. K., Clayton, E. W., et al. (2008). Managing Incidental Findings in Human Subjects Research-Analysis and Recommendations. *J. Law Med. Ethics* 36, 219–248.

Xu, C. (2018). A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Comput. Struct. Biotechnol. J.* 16, 15–24. doi: 10.1016/j.csbj.2018.01.003

Yang, H., and Wang, K. (2015). Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nat. Protoc.* 10, 1556–1566. doi: 10.1038/nprot.2015.105

Conflict of Interest Statement: 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.

frontiers in Genetics

HYPOTHESIS AND THEORY
published: 12 July 2019
doi: 10.3389/fgene.2019.00629

Single-Cell RNA Sequencing-Based Computational Analysis to Describe Disease Heterogeneity

Tao Zeng and Hao Dai*

Key Laboratory of Systems Biology, Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, Shanghai, China

The trillions of cells in the human body can be viewed as elementary but essential biological units that achieve different body states, but the low resolution of previous cell isolation and measurement approaches limits our understanding of the cell-specific molecular profiles. The recent establishment and rapid growth of single-cell sequencing technology has facilitated the identification of molecular profiles of heterogeneous cells, especially on the transcription level of single cells [single-cell RNA sequencing (scRNA-seq)]. As a novel method, the robustness of scRNA-seq under changing conditions will determine its practical potential in major research programs and clinical applications. In this review, we first briefly presented the scRNA-seq-related methods from the point of view of experiments and computation. Then, we compared several state-of-the-art scRNA-seq analysis frameworks mainly by analyzing their performance robustness on independent scRNA-seq datasets for the same complex disease. Finally, we elaborated on our hypothesis on consensus scRNA-seq analysis and summarized the potential indicative and predictive roles of individual cells in understanding disease heterogeneity by single-cell technologies.

Keywords: cellular heterogeneity, complex diseases, single-cell RNA sequencing, network, integration

OPEN ACCESS

Edited by:
Richard D. Emes, University of Nottingham, United Kingdom

Reviewed by:
Jidong Lang, Geneis (Beijing) Co. Ltd, China
Ken Lau, Vanderbilt University, United States

****Correspondence:***
Tao Zeng
zengtao@sibs.ac.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *17 January 2019*
Accepted: *17 June 2019*
Published: *12 July 2019*

Citation:
Zeng T and Dai H (2019) Single-Cell RNA Sequencing-Based Computational Analysis to Describe Disease Heterogeneity. Front. Genet. 10:629. doi: 10.3389/fgene.2019.00629

INTRODUCTION

It is known that an adult human body consists of trillion cells of different types and origins, and each of them plays its respective role in the body system. These cells can be viewed as basic but essential biological units supporting different body states, e.g., health, disease, or the response to therapy. Decades ago, the low resolution of cell isolation and measurement technologies limited our understanding of the cell-specific molecular profiles and their importance in cellular systems, causing humans to always underestimate disease heterogeneity.

In recent years, the establishment and the rapid growth of single-cell sequencing technology have led to the efficient and inexpensive identification of molecular profiles of individual cells (Bose et al., 2015; Baran-Gale et al., 2018; Svensson et al., 2018). In particular, the transcription of single cells (Wu et al., 2014; Ziegenhain et al., 2017) is a novel and fast evolving field. Single-cell RNA sequencing (scRNA-seq) attracts increasing attention to the identification and characterization of cells on an individual level rather than on a population level (Saliba et al., 2014; McDavid et al., 2016; Raj et al., 2018; Torre et al., 2018).

The research field of single cells, e.g., identifying cell types, recognizing cell markers, and tracing cell origins, is currently undergoing rapid development. New knowledge on cells can improve our understanding of biological systems by changing our perspective from the traditional population level

to the individual cellular level. It can further provide novel insights into old biological and biomedical questions (Raj et al., 2018). For example, with scRNA-seq data rather than bulk transcriptome data, we can detect genes with conserved expression levels across individual cells (Lin Y. et al., 2017). Single-cell transcriptomics could even uncover the diverse transcriptional states of immune cells and their coordination during immune responses (Vegh and Haniffa, 2018). In addition, simultaneous measurements of transcription along with genomic and epigenetic profiling at the single-cell level (Clark et al., 2016) is expected to be developed soon and will provide groundbreaking biological insights into these basic blocks building the biological body (Hemberg 2018).

In this quickly evolving field, many reviews have focused on the biotechnological applications of scRNA-seq and in silico gene expression analysis. The program goals of the Common Fund-supported Single Cell Analysis Program from the National Institutes of Health point out the impact of resolving tissue heterogeneity at the cellular level (Roy et al., 2018). Different scRNA-seq protocols have their strengths and disadvantages under respective settings (Saliba et al., 2014; Bacher and Kendziorski, 2016). The pre-processing approaches of sparse and row-rank scRNA-seq data (Zhang L. et al., 2018), normalization methods (Vallejos et al., 2017), and batch effect corrections (Dal Molin and Di Camillo, 2018; Haghverdi et al., 2018) have all been carried out for a wide range of comparisons and evaluations. Finally, the cell type clustering algorithms, cell marker identification, and cell trajectory reference also have their target-specific evaluation approaches for the deconvolution of biological system heterogeneity (Menon 2018; Papalexi and Satija, 2018). In addition, integrative impacts of whole scRNA-seq protocols and analysis methodologies have undergone in-depth assessments (Dal Molin et al., 2017; Svensson et al., 2017; Todorov and Saeys, 2018).

These current developments and achievements of scRNA-seq motivated us to investigate the individual cell types, cell signatures, cell origins in time and space, and cell communication strategies. Meanwhile, as a novel method, its robustness under different conditions (e.g., when applied to different datasets) will determine its actual practical potential in major research programs (e.g., the Precision Medicine Initiative or the Human Brain Project) (Poo et al., 2016; Sankar and Parker, 2016) or in clinical applications (e.g., diagnosis or prognosis of complex diseases) (Zeng et al., 2016). Thus, in this review paper, we discussed scRNA-seq from the point of view of experiments and computation. Then, on independent scRNA-seq datasets for the same complex disease (i.e., diabetes), we compared several state-of-the-art scRNA-seq analysis frameworks mainly by the robustness of their performances in the identification of cell types and markers. Lastly, we elaborated on our hypothesis on consensus scRNA-seq analysis and summarized the potential indicative and predictive roles of characteristic cells in understanding disease heterogeneity by single-cell technologies.

MATERIALS AND METHODS

A recent review has demonstrated the principle and potential of scRNA-seq in a wide range of studies, including development, physiology, and disease (Potter 2018). It concluded that the data noise and cell number are the main limitations in scRNA-seq studies, and many research fields would benefit from its continuous development. In contrast, this work concentrated on the scRNA-seq-based study from the two angles of experiments and computation. Especially, the robustness of scRNA-seq under changing conditions will decide its practical potential, e.g., in precision medicine. Thus, different from a previous report (Potter 2018), we further compared several state-of-the-art scRNA-seq analysis frameworks and included our hypothesis on the performance consensus.

scRNA-seq-Associated Biological Experiments

scRNA-seq is becoming a widely used genome-wide technology to detect cellular identities and dynamics, e.g., cell subpopulations, cell state marker genes and pathways, cell state transitions, and cell trajectories (Nguyen et al., 2018). This sustained improvement of the sensitivity, flexibility, and efficiency of scRNA-seq will help to resolve many biological and biomedical research questions on the individual cell level.

On the one hand, the rapid development of experimental protocols of scRNA-seq expands the measurement of mRNA levels to many associated fields of study (Fuzik et al., 2016; Hashimshony et al., 2016; Ilicic et al., 2016; Bagnoli et al., 2018; Han et al., 2018; Hayashi et al., 2018; Sasagawa et al., 2018). Especially, scRNA-seq applications have provided new insights into conventional biological questions, e.g., cellular heterogeneity. New cell types have been more widely recognized than previously expected (Burns et al., 2015; Usoskin et al., 2015; Rheaume et al., 2018), and gene expression levels corresponding to old and new cell types have uncovered many biological functions and mechanisms that were overlooked in conventional cell population studies (Nelson et al., 2016; Li H. et al., 2017); single-cell transcriptomic characteristics can reveal more time-dependent features of a biological system (Zeisel et al., 2015; Zeng et al., 2017; Lescroart et al., 2018; Liu D. et al., 2018), whereas the pseudo-time of single cells would mimic the actual dynamic biological process (Kowalczyk et al., 2015; Cacchiarelli et al., 2018). Taking all of the above novelties together, we can deepen our understanding on the complex mechanisms underlying cell-to-cell variation. These complex dynamic responses are controlled by regulatory cell-to-cell communication, which is also responsible for cellular heterogeneity (Shalek et al., 2014).

Measuring Regulatory Elements in a Single Cell

Cell-specific transcriptional signals might be regulated by the high-order structural folding of nucleosomes (Nagano et al., 2017; Lando et al., 2018), which can be investigated by combining scRNA-seq with other single-cell approaches (Stevens et al., 2017; Liu T. et al., 2018; Mezger et al., 2018). Of note, current scRNA profiling methods usually destroy cells during the analysis process, hindering the measurement of temporal gene expression changes. However, some information on biological dynamics will always be present in the data. For example, the continuum of molecular states in a population can reflect the trajectory or pseudo-time of a typical cell, so various methods increase their

power by reconstructing the trajectory by quantification of a group of cells in multiple static snapshots (Weinreb et al., 2018).

Measuring Post-transcriptional Regulations in a Single Cell

Understanding nongenetic cellular heterogeneity will help to characterize complete biological mechanisms in live cells, but little knowledge is available on the heterogeneity of regulatory modifications between individual cells. For example, microRNAs (miRNAs) are small RNAs that regulate gene expression in a post-transcriptional manner and might reduce cell-to-cell variability on the protein level by repressing mRNA translation or promoting mRNA degradation. Although the wet experimental evidence for the roles of miRNA in individual cells is limited, great efforts have been made to investigate such regulatory modifications in single cells (Fan et al., 2015). For instance, single-cell Quartz-Seq technology was developed to identify different kinds of nongenetic cellular heterogeneity in a quantitative manner (Sasagawa et al., 2013). Single-cell small RNA sequencing and analysis techniques have supplied much evidence that miRNAs could be potential molecular biomarkers for indicating the type and state of particular cells (Faridani et al., 2016). Moreover, using a combination of scRNA-seq data and mathematical modeling, it is also possible to detect key miRNAs as cell type-specific post-transcriptional regulators (Rzepiela et al., 2018).

Measuring Upstream Regulatory Factors in a Single Cell

Individual cells within different subpopulations can show significant variations when responding to external stresses, but the nature of this cellular heterogeneity is not clear, especially the remarkable alterations in the transcriptional architecture (Xue et al., 2013; Edsgard et al., 2016; Gasch et al., 2017). Fortunately, scRNA-seq provides high resolution to genetics by linking phenotypes to cell-specific gene functions, and the genetic screening of single cells can even be realized now (Birnbaum 2018; Raj et al., 2018). For example, the Perturb-seq was designed to combine scRNA-seq and CRISPR-based perturbations to detect individual perturbations causing target gene changes, gene signature appearances, genetic interaction rewiring, and cell state transitions (Dixit et al., 2016), e.g., discovering previously unknown immune circuits (Jaitin et al., 2016). Next, the allele-sensitive scRNA-seq could recognize clonal and dynamic monoallelic expression patterns (Reinius et al., 2016) or analyze allele-specific cis-control in genome-wide expressions (Deng et al., 2014; Jiang et al., 2017). Besides, focusing on the quantitative trait locus (QTL), the computational tool demuxlet was implemented to perform expression QTL (eQTL) analysis, which can identify natural genetic variation within multiplexed droplet scRNA-seq to evaluate cell type-specific gene expression changes (Kang et al., 2018). Similarly, some new cell type-specific "co-expression QTLs" have even been detected according to the genetic variants, significantly altering co-expression relationships (van der Wijst et al., 2018).

Measuring Downstream Regulation in a Single Cell

The cell-to-cell regulatory communication plays important roles in cellular diversity across diverse biological systems, which is an important factor in the evolution of observed cell types. scRNA-seq provides a powerful tool to analyze particular regulatory mechanisms and their downstream influence in a corresponding subset of cells (Chu et al., 2016; Korthauer et al., 2016; Enge et al., 2017; Severo et al., 2018). For example, the integration of transcription factor expression, chromatin profiling, and sequence motif analysis can be effective to identify the cell-specific genomic regulation underlying cell-specific gene expression (Sebe-Pedros et al., 2018). Similarly, the integration of information about single-cell transcriptomics and cell-free plasma RNA provides the potential to uncover longitudinal cellular dynamics of cells in complex biological processes or pathological development (Tsang et al., 2017). Next, a two-part method combining a generalized linear model and gene set enrichment analysis on single-cell data provided evolutionary insights in gene co-expression by experimental treatments (Finak et al., 2015). In addition, benefitting from time-course data obtained by scRNA-seq, it is possible to characterize the fate decision and transcriptional control of self-renewal, differentiation, and maturation of particular cells (Su et al., 2017), and transient cellular states corresponding to asynchronous cellular responses can be observed under conditional perturbations (Rizvi et al., 2017).

scRNA-seq-Associated Analytic Computations

As seen in the above summary, scRNA-seq technologies are swiftly developing. They are greatly beneficial to the investigation of transcriptional landscapes at the single-cell level, where they are able to profile cell-to-cell variability in cell populations and characterize unexpected heterogeneity of transcription in originally thought homogeneous cell populations. Although many computational methods for analyzing scRNA-seq data have been extensively developed, tested, and validated on simulated datasets, scRNA-seq protocols are still complex so that bias will easily occur in downstream analysis. In fact, computational models and tools available for the design and analysis of scRNA-seq experiments (**Table 1**) have their advantages and disadvantages in various settings, and many questions have yet to be solved in this exciting area (Bacher and Kendziorski, 2016). Similar to other high-throughput sequencing technologies, the general actions on scRNA-seq data include several key steps before the follow-up analysis for single cells (Jia et al., 2017; Li Y. H. et al., 2017; McCarthy et al., 2017; Chen W. et al., 2018; Vu et al., 2018), i.e., pre-procession (e.g., zero imputation) (Li and Li, 2018; Van den Berge et al., 2018), quality control (e.g., variation analysis) (Brennecke et al., 2013; Ding et al., 2015; Jiang et al., 2016; Eling et al., 2018; Lu et al., 2018), normalization (Bacher et al., 2017; Cole et al., 2017; Haghverdi et al., 2018; Tian et al., 2018), and visualization/simulation (Zappia et al., 2017). Although scRNA-seq studies have provided revolutionary tools to assist researchers to address scientific questions previously hard to investigate directly, several computational challenges are beginning to arise.

Challenge of Cluster Analysis of Single Cells

The detection of cell types from heterogeneous cells is an important step in the development of scRNA-seq data analysis

TABLE 1 | List of computational tools for single-cell RNA sequencing (scRNA-seq) analysis.

Category	ID	Access	Code and citation
Pre-procession	scater	Bioconductor	R (McCarthy et al., 2017)
	scPipe	Bioconductor	R (Tian et al., 2018)
	GRM	http://wanglab.ucsd.edu/star/GRM	R (Ding et al., 2015)
Cell clustering	SAFEclustering	http://yunliweb.its.unc.edu/safe/	R (Yang et al., 2018)
	DendroSplit	Github	Python (Zhang J. et al., 2018)
	clusterExperiment	Bioconductor	R (Risso et al., 2018)
	scmap	Bioconductor	R (Kiselev et al., 2018)
	scVDMC	Github	Matlab (Zhang H. et al., 2018)
	CIDR	Github	R (Lin P. et al., 2017)
	scClustBench	http://www.maths.usyd.edu.au/u/SMS/bioinformatics/software.html	R (Kim et al., 2018)
	SNN-Cliq	http://bioinfo.uncc.edu/SNNCliq	Matlab & Python (Xu and Su, 2015)
Cell marking	MAST	Github	R (Finak et al., 2015)
	SC2P	Github	R (Wu et al., 2018)
	DEsingle	Bioconductor	R (Miao et al., 2018)
	powsimR	Github	R (Vieth et al., 2017)
	BPSC	Github	R (Vu et al., 2016)
	Sincell	Bioconductor	R (Julia et al., 2015)
Cell ordering	dynverse	Github	R (Saelens et al., 2018)
	Progra	Github	R (Gong et al., 2018)
	p-Creode	Github	Python (Herring et al., 2018)
Pipeline	SINCERA	https://research.cchmc.org/pbge/sincera.html	R (Guo et al., 2015)
	SCell	Github	Exe (Diaz et al., 2016)
	Falco	Github	Python (Yang et al., 2017)
	ASAP	Github	R & python (Gardeux et al., 2017)
	SIMLR	Github	R & Matlab (Wang et al., 2017; Wang B. et al., 2018)
	SEURAT	http://satijalab.org/seurat/	R (Butler et al., 2018)
	Monocle	Bioconductor	R (Trapnell et al., 2014; Qiu et al., 2017a; Qiu et al., 2017b)
	DPT	http://www.helmholtz-muenchen.de/icb/dpt	R & Matlab (Haghverdi et al., 2016)
B-cell receptor reconstruction	VDJPuzzle	bitbucket	R & Python (Rizzetto et al., 2018)
	bracer	Github	Python (Lindeman et al., 2018)
Network inference	SCODE	Github	R (Matsumoto et al., 2017)
	LEAP	CRAN	R (Specht and Li, 2017)

in biological research (Marinov et al., 2014; Lin C. et al., 2017; Jin et al., 2018; Kiselev et al., 2019). Different methods use distinct characteristics of data and gain varying outcomes in terms of both the number of clusters and the cluster assignment of cells (Ntranos et al., 2016; Kim et al., 2018; Risso et al., 2018). Many approaches, such as SAFE clustering (Yang et al., 2018), DendroSplit (Zhang J. et al., 2018), scmap (Kiselev et al., 2018), MetaNeighbor (Crow et al., 2018), scVDMC (Zhang H. et al., 2018), CIDR (Lin P. et al., 2017), SC3 (Kiselev et al., 2017), scLVM (Buettner et al., 2015), and RaceID (Grun et al., 2015), have been developed to promote the efficiency of clustering single cells. They promote the clustering consensus, interpretability, subjectivity, comparability, and replicability. However, the biological significance, number estimation, and computational speed of such clustering analysis still require significant improvements (Duan et al., 2018).

Challenge of Identity Analysis of Single Cells

scRNA-seq has brought transcriptome research to a higher resolution as the "up or down" expression pattern can be examined at the single-cell level (Chen L. et al., 2018; Xie et al., 2019). The projection of high-dimensional data into a low-dimensional subspace will be a powerful strategy for mining such extensive data (Zeng et al., 2016; Yip et al., 2018; Yu and Zeng, 2018). Statistic-based approaches, such as PowsimR (Vieth et al., 2017), BPSC (Vu et al., 2016), Linnorm (Yip et al., 2017), and Oscope (Leng et al., 2015), have been established to evaluate differential expression among individual cells. Especially, latent factor-based analysis will be useful to find hidden biological signals and corresponding gene components from scRNA-seq samples (Buettner et al., 2017; Yu, 2018). However, to guarantee the biological meaning of detected cell identities, it is still necessary to discriminate the real and dropout zeros in scRNA-seq data (Miao et al., 2018). It is also essential to identify the combination of binary and continuous regulation in individual cells (Wu et al., 2018) and to integrate the nonlinear projection with prior-known biological knowledge (Li X. et al., 2017).

Challenge of Trajectory Analysis of Single Cells

The single-cell experiments provide a great chance to rebuild a sequence of changes in a dynamical process of the biological system from individual "snapshots" of cells (Matsumoto et al., 2017; Gong et al., 2018). The construction of a pseudo-temporal path as cell orders would be a useful way to characterize dynamical gene expression in a heterogeneous cell population, assuming the existence hypothesis of gradual transition of the cell transcriptome (Specht and Li, 2017; Herring et al., 2018; Shindo et al., 2018; Strauss et al., 2018). For example, based on the minimum spanning tree approach, the Tools for Single Cell

Analysis is developed for in silico pseudo-time reconstruction in scRNA-seq analysis (Ji and Ji, 2016). As an iterative supervised learning algorithm, FateID can recognize the cell fate preference by quantifying the lineage-specific probabilistic biases (Herman et al., 2018). By unsupervisedly selecting feature genes and judging the location and number of branches and loops, SLICER is able to infer highly nonlinear trajectories (Welch et al., 2016). However, many opportunities still exist to develop these current methods, particularly detecting complex trajectory topologies, linking pseudo-time and real-world time, determining baseline points, estimating transition possibility, and recognizing progression trends with tipping point (Zeng et al., 2013).

Challenge of Origin Analysis of Single Cells

The origin and nature of signals leading to pattern formation and self-organization is an essential question in developmental or stem cell biology. The answer would be recovered from the gene expressions of individual cells with spatial locations in a particular tissue (Vergara et al., 2017; Chen Q. et al., 2018). On the one hand, from the technological point of view, several methods have been designed for recording the spatial information of cells. The spatial transcriptomic technology and computational deconvolution can be combined to detect distinct expression profiles corresponding to different tissue components (Berglund et al., 2018). One technique that performs RT-LAMP reactions on a histological tissue section can preserve the original spatial location of the nucleic acid molecules to become an effective tissue analysis tool (Ganguli et al., 2018). Another technique is based on a panel of zonated landmark genes, where the lobule coordinates of mouse liver cells can be inferred according to their transcriptome, whereas the zonation profiles of all liver genes can also be characterized with high spatial resolution (Halpern et al., 2017). On the other hand, from the analytic point of view, supervised methods have been shown to be efficient, inferring the potential spatial distribution of cells. On the foundation of a reference gene expression database, e.g., the gene expression atlas for positional gene expression profiles within cells, an scRNA-seq-based high-throughput method has been applied to identify the spatial origin of cells (Achim et al., 2015). Obviously, spatial labeling technologies still need further technological developments for more easy and accurate testing, and the spatial classification and prediction of cells require more elaborate and efficient mathematical and computational models.

Challenge of Integrative Analysis of Single Cells

Understanding the genetic and cellular processes and programs driving the differentiation of diverse cell types and organ formation is a major challenge in developmental biology (Kelsey et al., 2017, Velten et al., 2017, Duren et al., 2018, Liu L. et al., 2018). Frameworks and software are required to perform dimension reduction, clustering, and visualization on scRNA-seq data to improve biological interpretability (Gardeux et al., 2017; Wang et al., 2017). Numerous methods have been implemented for analyzing scRNA-seq data in a whole life-cycle manner (Guo et al., 2015; Diaz et al., 2016; Leng et al., 2016; Yang et al., 2017). SparseDC solves a unified optimization problem so that it can carry out three tasks simultaneously, e.g., identifying cell types, tracing expression changes across conditions, and identifying marker genes for these changes (Barron et al., 2018). BigSCale implements a scalable analytical framework to handle millions of cells, so it can overcome large data challenges by the directed down-sampling strategy on index cell transcriptomes (Iacono et al., 2018). In addition to these usual analytic routines for conventional targets, more diverse integration models are required for data-driven, model-driven, hypothesis-driven, and combinatory bioinformatics mining in single-cell data.

Understanding Disease Heterogeneity by scRNA-seq Analysis

For questions in the biological and biomedical fields, human cancers are especially considered complex ecosystems where the basic elements (cells) exist in different disease states characterized by phenotypes and genotypes. As is well known, conventional methods have their limits when measuring and quantifying the diverse tumor (cell) composition in patients, e.g., traditional bulk expression profiles have to average the cells within each tumor. Nowadays, scRNA-seq provides a powerful technique to detect critical cell differences and deconvolve such cellular heterogeneity in disease tissues. Therefore, one important benefit obtained from scRNA-seq is the possibility to decipher tumor architecture (Cloney 2017), so that it might overcome intratumoral heterogeneity, which hampers the success of precision medicine and is therefore a huge challenge in cancer treatment (Patel et al., 2014; Kim et al., 2016; Zong, 2017). Actually, in the context of cancer, mRNA can be used to identify malignant cells and diverse tumor-tissue compositions; such tumor compositions could indicate the cancer-associated cells and types determining tumor characteristics (Young et al., 2018). Thus, scRNA-seq-based methods could be widely applied in clinical decision support (Tirosh et al., 2016a; Filbin et al., 2018; Krieg et al., 2018; Pellegrino et al., 2018).

i) *Tumor mechanism investigation*. One general framework can be used to decipher differences between multiple classes of human tumors by decoupling cancer cell genotypes, phenotypes, and the composition of tumor microenvironment (Venteicher et al., 2017). One single-cell analysis method has provided some insights into the cellular architecture of oligodendrogliomas and their function in development regulation, which potentially is compatible with the cancer stem cell model and its consideration in disease management (Tirosh et al., 2016b).
ii) *Tumor subtype recognition*. To deconvolve the cellular composition of a solid tumor from bulk gene expression data using reference gene expression profiles from tumor-derived scRNA-seq data, many cell types or subtypes must be identified accurately (Schelker et al., 2017). For example, one scRNA-seq study of triple-negative breast cancer identified the individual subpopulations with respective gene expression phenotypes and corresponding genotype driver candidates, whose associated signature genes can predict long-term outcomes (Karaayvaz et al., 2018).
iii) *Tumor immune therapy*. Single-cell analyses have suggested distinct patterns in the tumor microenvironment, e.g., the breast cancer transcriptome has shown a wide range of

intratumoral heterogeneity that is reshaped by both immune and tumor cells in a closely communicated microenvironment at a single-cell resolution (Chung et al., 2017). An unbiased scRNA-seq analysis has detected human dendritic cells and several monocyte subtypes in the human blood to permit more accurate immune monitoring in health and disease (Villani et al., 2017). In a more special field, the single-cell transcriptional information in B-cell lineages might have broad applications involved in vaccine design, antibody development, and cancer treatment (Rizzetto et al., 2018; Upadhyay et al., 2018).

iv) *Tumor virus-environment recognition*. Indeed, the interaction between a host and a pathogen is a highly dynamical process, so the potential association between a pathogen and cancer is worthy of profound investigation. An scRNA-seq-based method, scDual-Seq, has been proposed to capture host and pathogen transcriptomes simultaneously (Avital et al., 2017). In different mouse models, the hypothetical virus-host interaction events have been found to play some key regulatory role in virus phenotypes involved in complex diseases by tracking viral RNA at single-cell resolution within the immune system (Douam et al., 2017).

Of course, the translational usage of scRNA-seq is not limited to the field of tumor biology or complex human diseases; it is expected to have great potential and to enjoy a wide range of applications in biological and biomedical fields, such as infant development, health and wellness, and disease monitoring.

Design of Hypothesis and Theory Study on scRNA-seq Analysis Robustness

As is well known, scRNA-seq analysis is used to compare the expression levels of multiple genes at single-cell resolution (Tang et al., 2009). Different from the conventional population-based biological technologies for gene expression measurement (e.g., bulk gene expression), scRNA-seq is able to distinguish the expression differences between individual cells rather than tissues. With the continuous development of such technology, the testing cost is decreasing, whereas the number of cells that can simultaneously be tested is increasing exponentially. Some recent reviews have summarized these technological developments and protocol improvements in scRNA-seq analyses (Svensson et al., 2017; Ziegenhain et al., 2017; Svensson et al., 2018). An inspiring observation is that the number of tested cells and the number of detected genes can vary significantly depending on the corresponding experimental platforms. For example, SMART-seq2 is able to detect about 10,000 genes and achieve the highest accuracy, but the number of cells analyzed by this method is only 100 to 1,000 (Picelli et al., 2013; Picelli et al., 2014). In contrast, Drop-seq is able to test more than 10,000 cells simultaneously, but the number of genes detected is usually less than 5,000 (Macosko et al., 2015). Recently, several commercial platforms, such as 10X Genomics Chromium, Fluidigm C1, and Wafergen ICELL8, were available for scRNA-seq analysis with the capability to measure hundreds to millions of cells through a simple and fast workflow.

Researchers are usually required to select the suitable experimental protocol to design the follow-up scRNA-seq analysis based on corresponding biological questions:

i) If one aims to discover new cell types with distinct expression patterns, more cells should be tested because it is impossible to find rare cell types from only a few hundred cells by chance.
ii) If one aims to analyze the changes in gene expression between different cell types or developmental stages or to analyze the gene interactions to find some key regulatory genes, more genes have to be measured with high accuracy.
iii) If one aims to analyze particular cell types by isolating a subset of cells for sequencing, fluorescence-activated cell sorting or a similar technology needs to be used to select the cells with cell type-specific cell surface markers.

To evaluate and investigate the robustness of different scRNA-seq analysis methods, we have carried out two comparisons on multiple scRNA-seq datasets.

The aim of the first comparison is to discuss the experimental factors for scRNA-seq analysis. As is well known, the accuracy of RNA-seq data analysis is dependent on the experimental methods, especially the sequencing depth and dropout rate. To test these experimental factors before further evaluation, we compared four datasets on two different experimental platforms: GSE81608 (Xin et al., 2016) and GSE83139 (Wang et al., 2016) on an Illumina HiSeq 2500 and GSE86469 (Lawlor et al., 2017) and GSE81547 (Enge et al., 2017) on an Illumina NextSeq 500. All of these datasets come from the single-cell studies of human pancreatic islet cells so that their computational results will be comparable, and the number of clusters for each method was fixed to be the same as the number of biological classes corresponding to each dataset, as shown in **Table 2**.

The aim of the second comparison is to discuss the analytic approaches for scRNA-seq analysis. The performance of dissimilar methods on different real datasets of the same complex disease is important to evaluate, because performance robustness will be strictly required for biomedical studies and applications. Thus, we have employed several widely used methods in a few public scRNA-seq datasets from complex disease studies, which are listed in **Table 3**. According to the above summary, we actually evaluated the performances on cell cluster, cell identity, and cell trajectory. These methods' parameter settings are listed in the supplementary files (**Supp 1**).

i) For cell clustering analysis, traditional methods, such as hierarchical clustering, k-means, and scRNA-seq-induced SIMLR (Wang et al., 2017; Wang B. et al., 2018), SNN-Cliq (Xu and Su, 2015), and SEURAT (Butler et al., 2018) have been evaluated and compared.
ii) For cell pseudo-time analysis, the Monocle (Trapnell et al., 2014; Qiu et al., 2017a; Qiu et al., 2017b) and diffusion pseudo-time (DPT) (Haghverdi et al., 2016) have been tested and compared.

Of note, to quantitatively measure and compare the analysis accuracy of cell clusters from different methods, the conventional adjusted rand index (ARI) is applied. Given a dataset of n cells,

TABLE 2 | Clustering performances of four datasets with different experiment methods represented as adjusted rand index (ARI).

	GSE81547	GSE83139	GSE81608	GSE86469
Experiment platforms	NextSeq 500	HiSeq 2500	HiSeq 2500	NextSeq 500
Number of cells	2,282	635	1,600	617
Number of detected genes per cell on average	3,281	5,638	5,706	8,339
Number of potential cell types*	6	8	4	7
Hierarchical clustering	0.34	0.25	0.46	0.63
k-means	0.34	0.27	0.44	0.48
tSNE+k-means	0.37	0.34	0.54	0.72
SIMLR	0.34	0.32	0.51	0.61
SNN-Cliq	0.10	0.31	0.05	0.61
SEURAT	0.31	0.31	0.45	0.89

**GSE81547 includes alpha cells, beta cells, delta cells, acinar cells, mesenchyme cells, and ductal cells. GSE83139 includes alpha cells, beta cells, delta cells, PP cells, acinar cells, mesenchyme cells, ductal cells, and dropped cells. GSE81608 includes alpha cells, beta cells, delta cells, and PP cells. GSE86469 includes alpha cells, beta cells, delta cells, PP cells, acinar cells, stellate cells, and ductal cells.*

TABLE 3 | Summary of evaluation datasets on human complex diseases.

Data ID	Purpose	Platform	#scRNA-Seq	#Class
GSE69405	scRNA-seq identifies subclonal heterogeneity in anticancer drug responses of lung adenocarcinoma cells	HiSeq 2500	176	3
GSE73121	scRNA-seq in optimizing a combinatorial therapeutic strategy in metastatic renal cell carcinoma	HiSeq 2500	118	3
GSE81608	scRNA-seq on human islet cells revealing type 2 diabetes genes	HiSeq 2500	1600	4
GSE83139	scRNA-seq of the human endocrine pancreas	HiSeq 2500	635	8

the experimentally determined cell types are X_1, X_2, …, X_r and the calculated clusters are Y_1, Y_2, …, Y_s. The number of cells that belong to cell type X_i is denoted as a_i, the number of cells that belong to cluster Y_j is denoted as b_j, and the number of cells that belong to both X_i and Y_j is denoted as n_{ij}, which means $n_{ij} = |X_i \cap Y_j|$. Then, the ARI is calculated as follows:

$$ARI = \frac{\Sigma_{ij}\binom{n_{ij}}{2} - \left[\Sigma_i \binom{a_i}{2} \cdot \Sigma_j \binom{b_j}{2}\right] \Big/ \binom{n}{2}}{\frac{1}{2}\left[\Sigma_i \binom{a_i}{2} + \Sigma_j \binom{b_j}{2}\right] - \left[\Sigma_i \binom{a_i}{2} \cdot \Sigma_j \binom{b_j}{2}\right] \Big/ \binom{n}{2}}.$$

RESULTS AND DISCUSSION

Experimental Factors for scRNA-seq Analysis

The experimental processes of the four datasets presented in **Table 2** are briefly summarized below.

1) For GSE81608 (Xin et al., 2016), islets were handpicked and enzymatically digested; during RNA *in situ* hybridization, the cells were permeabilized and hybridized with combinations of mRNA probes and a multiplex fluorescent kit was used to amplify the mRNA signal. Sequencing was performed on an Illumina HiSeq2500 in rapid mode by multiplexed single-read run with 50 cycles.
2) For GSE83139 (Wang et al., 2016), human islets require careful sample acquisition and preparation; the SMART-seq method was used for first-strand cDNA synthesis and polymerase chain reaction (PCR) amplification. All of the libraries were sequenced on the Illumina HiSeq 2500 with 100 bp single-end reads.
3) For GSE86469 (Lawlor et al., 2017), islets are systematically acquired, processed, and dissociated; then, single-cell processing is carried out on the C1 single-cell Autoprep system. All of the sequencing was performed on an Illumina NextSeq500 using the 75-cycle high-output chip.
4) For GSE81547 (Enge et al., 2017), the experimental models and human pancreas or islet samples were conducted in accordance with guidelines; during flow cytometry, isolated human islets were dissociated into single cells by enzymatic digestion using Accumax (Invitrogen). Next, single-cell RNA-seq libraries were generated as described in the literature, and barcoded libraries were pooled and subjected to 75 bp paired-end sequencing on the Illumina NextSeq instrument.

Of course, the whole experimental process should be consistent; however, the scRNA-seq wet experiments in different studies were conducted with different parameters and under different circumstances, which are worthy of future evaluation. Although sequencing platforms are only one part of the scRNA-seq experiment, we tried to include them for the comparison study in this work. In **Table 2**, we see that there is no obvious performance difference between two experiment platforms; however, the accuracy (i.e., ARI) seems to increase when the number of detected genes becomes large for almost all of the tested methods, which is consistent with a previous conclusion (Potter, 2018) and implies that the influence of sequencing depth is very important

in the experimental protocol for follow-up data analysis. Of note, the parameter setting for each compared method in this work is outlined in the supplementary files (**Supp 1**).

Analytic Approaches for scRNA-seq Analysis

First, it can be seen that the datasets after dimension reduction by t-distributed stochastic neighbor embedding (tSNE) (Maaten and Hintton, 2008) exhibit better performances in conventional k-means clustering than the initial dataset, which is due to the noise reduction of scRNA-seq data. Dimension reduction can be used in the visualization of such phenomena, which reduces one dataset from high-dimensional data space to two- or three-dimensional data space. **Figure 1A** illustrates the performances of principal component analysis (PCA) and tSNE on multiple datasets. It is clear that tSNE, a nonlinear method, can usually achieve better visualization effects than PCA, a linear method. This is because tSNE can group the cell points from one class cluster together and keep the cell points from different classes separated from each other. The quantitative measurement of the influence of PCA and tSNE by the Davies-Bouldin index also supported this conclusion, as shown in the supplementary files (**Supp 2**). Of note, due to the large computational complexity of nonlinear methods, the general strategy for large data analysis includes two steps. The first is to reduce the dimension to 20 to 50 by PCA, and the second is to reduce such moderate dimension to 2 to 3 by tSNE. This strategy is expected to achieve a good balance between computational performance and resource consumption.

Second, in the cell clustering analysis, the analyzed genes are selected that exhibit expression in at least three cells, so that most genes have actually been used. For hierarchical clustering, k-means, tSNE+k-means, and SIMLR, the number of clusters for each method was fixed to be the same as the number of biological classes corresponding to each dataset, as shown in **Table 3**. For SNN-Cliq and SEURAT, the parameters were adjusted to guarantee that the number of final clusters was the same as the number of biological classes in those datasets, as shown in **Table 3**. In other words, the number of clusters for every method is the same for one dataset to make different methods fairly comparable to ARI. As seen in **Figure 1B**, it is obvious the performances of tSNE+k-means, SIMLR, and SEURAT were better than those of others with higher ARI values in most scRNA-seq datasets. In addition, although tSNE+k-means, SIMLR, and SEURAT have similar performances with regard to ARI, they usually accurately detected different true classes (**Figure 1C**). This means different methods would have different analysis preferences due to different underlying mathematical or biological frameworks and explanations of scRNA information.

Third, scRNA-seq data follow a time series and the expression of cells may change continuously. For this kind of dataset, some statistical methods can be used to order the cells one by one along a trajectory, which is called pseudo-time or pseudo-trajectory. This mathematical model has been widely applied in developmental biology to reconstruct the differentiation processes and find the key time point of differentiation (Cannoodt et al., 2016). In addition, cell pseudo-time analysis can also be used in studies of cancer and diabetes to reconstruct the occurrence and transformation processes of complex diseases. Thus, the Monocle and DPT have been carried out for pseudo-time analysis on multiple scRNA-seq datasets; these two computational methods are dependent on entirely different principles. In this cell pseudo-time analysis, the most expression-variable genes are selected as feature genes for downstream analysis. As shown in **Figure 1D**, the feature genes exhibit great differences between datasets with different biological backgrounds; however, the two datasets on similar biological phenotypes still have much overlap (i.e., the feature genes from two datasets related to tumor cells with treatments or those from two datasets associated with diabetes). Of note, using human pancreas scRNA-seq datasets in another platform (i.e., GSE86469 and GSE81547; **Table 2**) as controls, the top 50 selected feature genes from the total four datasets indeed had more overlapping genes, as listed in the supplementary files (**Supp 3**). In **Figure 1E**, it is seen that both Monocle and DPT are able to reconstruct the pseudo-time with branches, and DPT seems to obtain more accurate results as the cells of the same cell type tend to group together. Meanwhile, the pseudo-time and branch point seem to be clearer in the analyses of Monocle. Of note, the performance of pseudo-time analysis will be strongly influenced by the selected feature genes. In this comparison, the most expression-variable genes were used, but usually it would be much better to select the feature genes based on the prior biological knowledge in each case study. Furthermore, the consistency of pseudo-time results from different methods is considered and evaluated. As shown in **Figure 2**, the correlations between the first principal components of the pseudo-time results from Monocle and DPT have been calculated. Then, the estimation similarities of cell orders in particular cell classes from different methods are compared. It is obvious that the cell order correlations have huge variances in a wide range among different prior-known cell classes. In addition, two other pseudo-time methods, Wanderlust (Bendall et al., 2014) and SCUBA (Marco et al., 2014), were also applied to reconstruct the pseudo-time trajectory of single cells without branch, as discussed in the supplementary files (**Supp 4**). The observations and conclusions were similar. Thus, in the pseudo-time analysis, consensus performance of dissimilar methods is weak currently.

CONCLUSION

scRNA-seq has opened a new way to study complex biological phenomena on the single-cell level, which will be especially helpful in the research of complex diseases. However, to enhance its performance in actual applications, e.g., in the clinic, several improvements are still required. For cell clustering and identification, gene networks rather than separate genes would be more important and reliable to characterize cell states (e.g., network biomarkers for disease subtypes) (Zeng et al., 2014; Zeng et al., 2016). For the cell order, the start or end point of pseudo-time is still a manual judgment, and the auto-determination of these time points will render these methods more flexible and applicable (e.g., temporal driving for disease causality) (Yu et al., 2017; Wang et al., 2018; Setty et al., 2019). The branch point of pseudo-time also requires more models on critical transitions (e.g., tipping point for

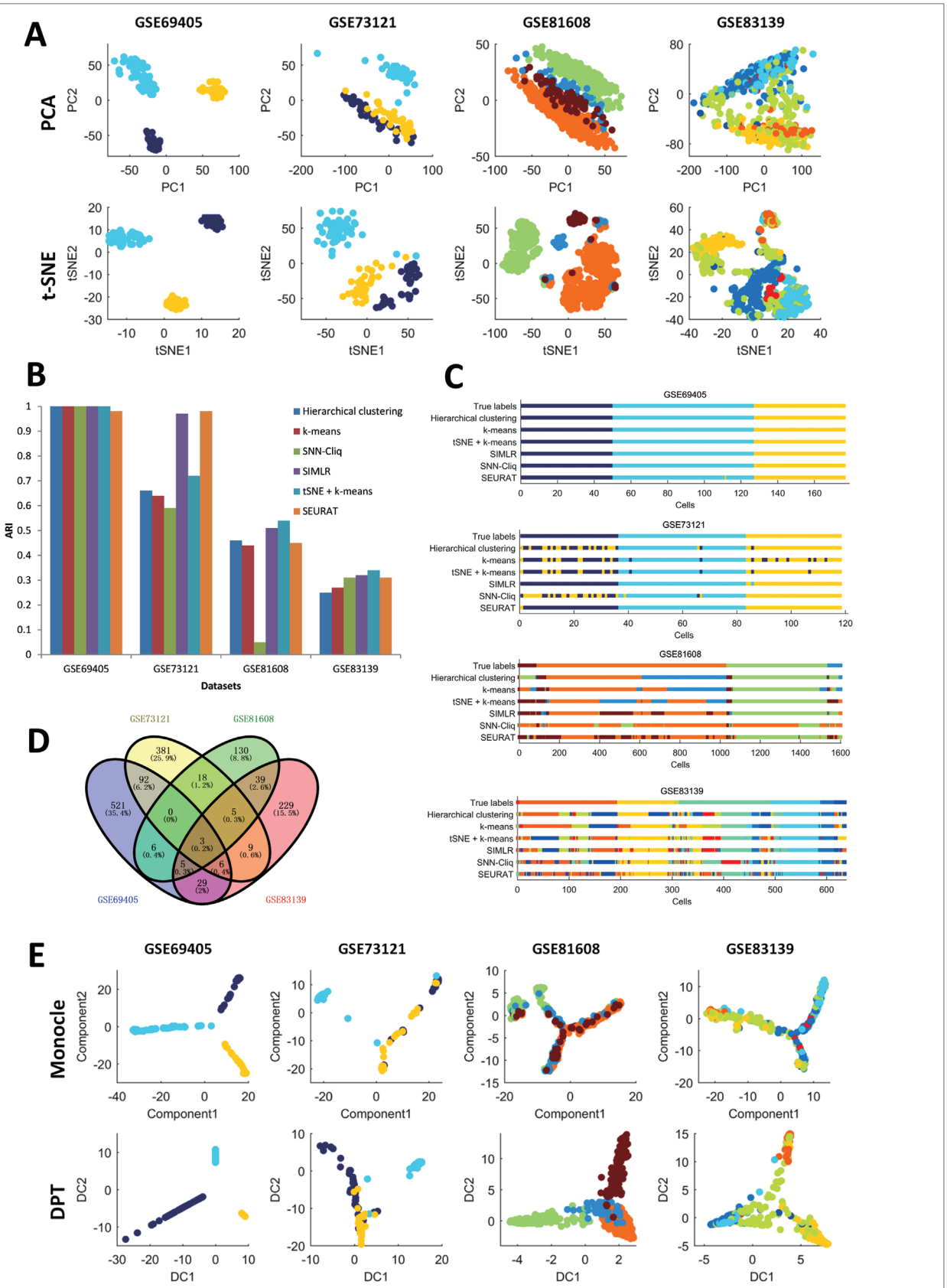


FIGURE 1 | Summary of performance comparison.

ID	GSE69405	GSE73121	GSE81608	GSE83139
Class 1	*H358 human lung cancer cells, 50 cells* **0.15**	*PDX - metastatic renal cell carcinoma, 36 cells* **0.65**	*Pancreatic PP cells, 93 cells* **0.06**	*Acinar cells, 6 cells* **0.77**
Class 2	*Tumor cell-enriched PDX cells, 77 cells* **0.79**	*PDX - primary renal cell carcinoma, 47 cells* **0.84**	*Pancreatic α cells, 946 cells* **0.62**	*α cells, 190 cells* **0.62**
Class 3	*Another lung cancer PDX case, 49 cells* **0.14**	*Parental - metastatic renal cell carcinoma, 35 cells* **0.78**	*Pancreatic β cells, 503 cells* **0.23**	*β cells, 111 cells* **0.65**
Class 4			*Pancreatic δ cells, 58 cells* **0.13**	*δ cells, 9 cells* **0.81**
Class 5				*Dropped cells, 178 cells* **0.40**
Class 6				*Duct cells, 96 cells* **0.26**
Class 7				*Mesenchyme cells, 27 cells* **0.54**
Class 8				*PP cells, 18 cells* **0.80**

FIGURE 2 | Summary of robustness comparison.

disease transition) (Zeng et al., 2013; Li et al., 2014). Particularly, the assembling method with good consensus on different datasets is expected to provide more robust integrative scRNA-seq methods for biological and biomedical studies (e.g., pattern fusion for disease heterogeneity) (Shi et al., 2017; Guo et al., 2018).

AUTHOR CONTRIBUTIONS

TZ conceived the concept and design of the work. HD and TZ performed the experiments. TZ and HD analyzed the results. TZ drafted the manuscript. TZ and HD revised the paper.

FUNDING

This study was supported by the National Key R&D Program Special Project on Precision Medicine (2016YFC0903400), the National Natural Science Foundation of China (11871456 and 61803360), the Shanghai Municipal Science and Technology Major Project (2017SHZDZX01), and the Natural Science Foundation of Shanghai (17ZR1446100).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00629/full#supplementary-material

SUPP 1 | Parameter settings of scRNA-seq analysis methods.

SUPP 2 | Feature genes of four human pancreas scRNA-seq datasets.

SUPP 3 | Quantitative measurement of PCA and tSNE.

SUPP 4 | Additional pseudo-time methods on four scRNA-seq datasets.

REFERENCES

Achim, K., Pettit, J. B., Saraiva, L. R., Gavriouchkina, D., Larsson, T., Arendt, D., et al. (2015). High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nat. Biotechnol.* 33 (5), 503–509. doi: 10.1038/nbt.3209

Avital, G., Avraham, R., Fan, A., Hashimshony, T., Hung, D. T., and Yanai, I. (2017). scDual-Seq: mapping the gene regulatory program of *Salmonella* infection by host and pathogen single-cell RNA-sequencing. *Genome Biol.* 18 (1), 200. doi: 10.1186/s13059-017-1340-x

Bacher, R., and Kendziorski, C. (2016). Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol.* 17, 63. doi: 10.1186/s13059-016-0927-y

Bacher, R., Chu, L. F., Leng, N., Gasch, A. P., Thomson, J. A., Stewart, R. M., et al. (2017). SCnorm: robust normalization of single-cell RNA-seq data. *Nat. Methods* 14 (6), 584–586. doi: 10.1038/nmeth.4263

Bagnoli, J. W., Ziegenhain, C., Janjic, A., Wange, L. E., Vieth, B., Parekh, S., et al. (2018). Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. *Nat. Commun.* 9 (1), 2937. doi: 10.1038/s41467-018-05347-6

Baran-Gale, J., Chandra, T., and Kirschner, K. (2018). Experimental design for single-cell RNA sequencing. *Brief Funct. Genom.* 17 (4), 233–239. doi: 10.1093/bfgp/elx035

Barron, M., Zhang, S., and Li, J. (2018). A sparse differential clustering algorithm for tracing cell type changes *via* single-cell RNA-sequencing data. *Nucleic Acids Res.* 46 (3), e14. doi: 10.1093/nar/gkx1113

Bendall, S. C., Davis, K. L., Amir el, A. D., Tadmor, M. D., Simonds, E. F., Chen, T. J., et al. (2014). Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* 157 (3), 714–725. doi: 10.1016/j.cell.2014.04.005

Berglund, E., Maaskola, J., Schultz, N., Friedrich, S., Marklund, M., Bergenstrahle, J., et al. (2018). Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nat. Commun.* 9 (1), 2419. doi: 10.1038/s41467-018-04724-5

Birnbaum, K. D. (2018). Power in numbers: single-cell RNA-seq strategies to dissect complex tissues. *Annu. Rev. Genet.* 23 (52), 203–221. doi: 10.1146/annurev-genet-120417-031247

Bose, S., Wan, Z., Carr, A., Rizvi, A. H., Vieira, G., Pe'er, D., et al. (2015). Scalable microfluidics for single-cell RNA printing and sequencing. *Genome Biol.* 16, 120. doi: 10.1186/s13059-015-0684-3

Brennecke, P., Anders, S., Kim, J. K., Kolodziejczyk, A. A., Zhang, X., Proserpio, V., et al. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* 10 (11), 1093–1095. doi: 10.1038/nmeth.2645

Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., et al. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* 33 (2), 155–160. doi: 10.1038/nbt.3102

Buettner, F., Pratanwanich, N., McCarthy, D. J., Marioni, J. C., and Stegle, O. (2017). f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq. *Genome Biol.* 18 (1), 212. doi: 10.1186/s13059-017-1334-8

Burns, J. C., Kelly, M. C., Hoa, M., Morell, R. J., and Kelley, M. W. (2015). Single-cell RNA-seq resolves cellular complexity in sensory organs from the neonatal inner ear. *Nat. Commun.* 6, 8557. doi: 10.1038/ncomms9557

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* 36 (5), 411–420. doi: 10.1038/nbt.4096

Cacchiarelli, D., Qiu, X., Srivatsan, S., Manfredi, A., Ziller, M., Overbey, E., et al. (2018). Aligning single-cell developmental and reprogramming trajectories identifies molecular determinants of myogenic reprogramming outcome. *Cell Syst.* 7 (3), 258–268. doi: 10.1016/j.cels.2018.07.006

Cannoodt, R., Saelens, W., and Saeys, Y. (2016). Computational methods for trajectory inference from single-cell transcriptomics. *Eur. J. Immunol.* 46 (11), 2496–2506. doi: 10.1002/eji.201646347

Chen, L., and Zheng, S. (2018). BCseq: accurate single cell RNA-seq quantification with bias correction. *Nucleic Acids Res.* 46 (14), e82. doi: 10.1093/nar/gky308

Chen, Q., Shi, J., Tao, Y., and Zernicka-Goetz, M. (2018). Tracing the origin of heterogeneity and symmetry breaking in the early mammalian embryo. *Nat. Commun.* 9 (1), 1819. doi: 10.1038/s41467-018-04155-2

Chen, W., Li, Y., Easton, J., Finkelstein, D., Wu, G., and Chen, X. (2018). UMI-count modeling and differential expression analysis for single-cell RNA sequencing. *Genome Biol.* 19 (1), 70. doi: 10.1186/s13059-018-1438-9

Chu, L. F., Leng, N., Zhang, J., Hou, Z., Mamott, D., Vereide, D. T., et al. (2016). Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biol.* 17 (1), 173. doi: 10.1186/s13059-016-1033-x

Chung, W., Eum, H. H., Lee, H. O., Lee, K. M., Lee, H. B., Kim, K. T., et al. (2017). Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nat. Commun.* 8, 15081. doi: 10.1038/ncomms15081

Clark, S. J., Lee, H. J., Smallwood, S. A., Kelsey, G., and Reik, W. (2016). Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity. *Genome Biol.* 17, 72. doi: 10.1186/s13059-016-0944-x

Cloney, R. (2017). Cancer genomics: single-cell RNA-seq to decipher tumour architecture. *Nat. Rev. Genet.* 18 (1), 2–3. doi: 10.1038/nrg.2016.151

Cole, M. B., Risso, D., Wagner, A., DeTomaso, D., Ngai, J., Purdom, E., et al. (2017). Performance assessment and selection of normalization procedures for single-cell RNA-Seq. *bioRxiv*. doi: 10.1101/235382. [Epub ahead of print].

Crow, M., Paul, A., Ballouz, S., Huang, Z. J., and Gillis, J. (2018). Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. *Nat. Commun.* 9 (1), 884. doi: 10.1038/s41467-018-03282-0

Dal Molin, A., Baruzzo, G., and Di Camillo, B. (2017). Single-cell RNA-sequencing: assessment of differential expression analysis methods. *Front. Genet.* 8, 62. doi: 10.3389/fgene.2017.00062

Dal Molin, A., and Di Camillo, B. (2018). How to design a single-cell RNA-sequencing experiment: pitfalls, challenges and perspectives. *Brief Bioinform.* doi: 10.1093/bib/bby007. [Epub ahead of print].

Deng, Q., Ramskold, D., Reinius, B., and Sandberg, R. (2014). Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* 343 (6167), 193–196. doi: 10.1126/science.1245316

Diaz, A., Liu, S. J., Sandoval, C., Pollen, A., Nowakowski, T. J., Lim, D. A., et al. (2016). SCell: integrated analysis of single-cell RNA-seq data. *Bioinformatics* 32 (14), 2219–2220. doi: 10.1093/bioinformatics/btw201

Ding, B., Zheng, L., Zhu, Y., Li, N., Jia, H., Ai, R., et al. (2015). Normalization and noise reduction for single cell RNA-seq experiments. *Bioinformatics* 31 (13), 2225–2227. doi: 10.1093/bioinformatics/btv122

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., et al. (2016). Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* 167 (7), 1853–1866 e1817. doi: 10.1016/j.cell.2016.11.038

Douam, F., Hrebikova, G., Albrecht, Y. E., Sellau, J., Sharon, Y., Ding, Q., et al. (2017). Single-cell tracking of flavivirus RNA uncovers species-specific interactions with the immune system dictating disease outcome. *Nat. Commun.* 8, 14781. doi: 10.1038/ncomms14781

Duan, T., Pinto, J. P., and Xie, X. (2018). Parallel clustering of single cell transcriptomic data with split-merge sampling on Dirichlet process mixtures. *Bioinformatics* 35 (6), 953–961. doi: 10.1093/bioinformatics/bty702

Duren, Z., Chen, X., Zamanighomi, M., Zeng, W., Satpathy, A. T., Chang, H. Y., et al. (2018). Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. *Proc. Natl. Acad. Sci. U. S. A.* 115 (30), 7723–7728. doi: 10.1073/pnas.1805681115

Edsgard, D., Reinius, B., and Sandberg, R. (2016). scphaser: haplotype inference using single-cell RNA-seq data. *Bioinformatics* 32 (19), 3038–3040. doi: 10.1093/bioinformatics/btw484

Eling, N., Richard, A. C., Richardson, S., Marioni, J. C., and Vallejos, C. A. (2018). Correcting the mean-variance dependency for differential variability testing using single-cell RNA sequencing data. *Cell Syst.* 7 (3), 284–294. doi: 10.1016/j.cels.2018.06.011

Enge, M., Arda, H. E., Mignardi, M., Beausang, J., Bottino, R., Kim, S. K., et al. (2017). Single-cell analysis of human pancreas reveals transcriptional signatures of aging and somatic mutation patterns. *Cell* 171 (2), 321–330 e314. doi: 10.1016/j.cell.2017.09.004

Fan, X., Zhang, X., Wu, X., Guo, H., Hu, Y., Tang, F., et al. (2015). Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. *Genome Biol.* 16, 148. doi: 10.1186/s13059-015-0706-1

Faridani, O. R., Abdullayev, I., Hagemann-Jensen, M., Schell, J. P., Lanner, F., and Sandberg, R. (2016). Single-cell sequencing of the small-RNA transcriptome. *Nat. Biotechnol.* 34 (12), 1264–1266. doi: 10.1038/nbt.3701

Filbin, M. G., Tirosh, I., Hovestadt, V., Shaw, M. L., Escalante, L. E., Mathewson, N. D., et al. (2018). Developmental and oncogenic programs in H3K27M gliomas dissected by single-cell RNA-seq. *Science* 360 (6386), 331–335. doi: 10.1126/science.aao4750

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., et al. (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* 16, 278. doi: 10.1186/s13059-015-0844-5

Fuzik, J., Zeisel, A., Mate, Z., Calvigioni, D., Yanagawa, Y., Szabo, G., et al. (2016). Integration of electrophysiological recordings with single-cell RNA-seq data identifies neuronal subtypes. *Nat. Biotechnol.* 34 (2), 175–183. doi: 10.1038/nbt.3443

Ganguli, A., Ornob, A., Spegazzini, N., Liu, Y., Damhorst, G., Ghonge, T., et al. (2018). Pixelated spatial gene expression analysis from tissue. *Nat. Commun.* 9 (1), 202. doi: 10.1038/s41467-017-02623-9

Gardeux, V., David, F. P. A., Shajkofci, A., Schwalie, P. C., and Deplancke, B. (2017). ASAP: a web-based platform for the analysis and interactive visualization of single-cell RNA-seq data. *Bioinformatics* 33 (19), 3123–3125. doi: 10.1093/bioinformatics/btx337

Gasch, A. P., Yu, F. B., Hose, J., Escalante, L. E., Place, M., Bacher, R., et al. (2017). Single-cell RNA sequencing reveals intrinsic and extrinsic regulatory heterogeneity in yeast responding to stress. *PLoS Biol.* 15 (12), e2004050. doi: 10.1371/journal.pbio.2004050

Gong, W., Kwak, I. Y., Koyano-Nakagawa, N., Pan, W., and Garry, D. J. (2018). TCM visualizes trajectories and cell populations from single cell data. *Nat. Commun.* 9 (1), 2749. doi: 10.1038/s41467-018-05112-9

Grun, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., et al. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* 525 (7568), 251–255. doi: 10.1038/nature14966

Guo, M., Wang, H., Potter, S. S., Whitsett, J. A., and Xu, Y. (2015). SINCERA: a pipeline for single-cell RNA-Seq profiling analysis. *PLoS Comput. Biol.* 11 (11), e1004575. doi: 10.1371/journal.pcbi.1004575

Guo, W. F., Zhang, S. W., Liu, L. L., Liu, F., Shi, Q. Q., Zhang, L., et al. (2018). Discovering personalized driver mutation profiles of single samples in cancer by network control strategy. *Bioinformatics* 34 (11), 1893–1903. doi: 10.1093/bioinformatics/bty006

Haghverdi, L., Lun, A. T. L., Morgan, M. D., and Marioni, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* 36 (5), 421–427. doi: 10.1038/nbt.4091

Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F., and Theis, F. J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* 13 (10), 845–848. doi: 10.1038/nmeth.3971

Halpern, K. B., Shenhav, R., Matcovitch-Natan, O., Toth, B., Lemze, D., Golan, M., et al. (2017). Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. *Nature* 542 (7641), 352–356. doi: 10.1038/nature21065

Han, X., Chen, H., Huang, D., Chen, H., Fei, L., Cheng, C., et al. (2018). Mapping human pluripotent stem cell differentiation pathways using high throughput single-cell RNA-sequencing. *Genome Biol.* 19 (1), 47. doi: 10.1186/s13059-018-1426-0

Hashimshony, T., Senderovich, N., Avital, G., Klochendler, A., de Leeuw, Y., Anavy, L., et al. (2016). CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* 17, 77. doi: 10.1186/s13059-016-0938-8

Hayashi, T., Ozaki, H., Sasagawa, Y., Umeda, M., Danno, H., and Nikaido, I. (2018). Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. *Nat. Commun.* 9 (1), 619. doi: 10.1038/s41467-018-02866-0

Hemberg, M. (2018). Single-cell genomics. *Brief Funct. Genom.* 17 (4), 207–208. doi: 10.1093/bfgp/ely025

Herman, J. S., Sagar, , and Grun, D. (2018). FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data. *Nat. Methods* 15 (5), 379–386. doi: 10.1038/nmeth.4662

Herring, C. A., Banerjee, A., McKinley, E. T., Simmons, A. J., Ping, J., Roland, J. T., et al. (2018). Unsupervised trajectory analysis of single-cell RNA-Seq and imaging data reveals alternative tuft cell origins in the gut. *Cell Syst.* 6 (1), 37–51 e39. doi: 10.1016/j.cels.2017.10.012

Iacono, G., Mereu, E., Guillaumet-Adkins, A., Corominas, R., Cusco, I., Rodriguez-Esteban, G., et al. (2018). bigSCale: an analytical framework for big-scale single-cell data. *Genome Res.* 28 (6), 878–890. doi: 10.1101/gr.230771.117

Ilicic, T., Kim, J. K., Kolodziejczyk, A. A., Bagger, F. O., McCarthy, D. J., Marioni, J. C., et al. (2016). Classification of low quality cells from single-cell RNA-seq data. *Genome Biol.* 17, 29. doi: 10.1186/s13059-016-0888-1

Jaitin, D. A., Weiner, A., Yofe, I., Lara-Astiaso, D., Keren-Shaul, H., David, E., et al. (2016). Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-Seq. *Cell* 167 (7), 1883–1896 e1815. doi: 10.1016/j.cell.2016.11.039

Ji, Z., and Ji, H. (2016). TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res.* 44 (13), e117. doi: 10.1093/nar/gkw430

Jia, C., Hu, Y., Kelly, D., Kim, J., Li, M., and Zhang, N. R. (2017). Accounting for technical noise in differential expression analysis of single-cell RNA sequencing data. *Nucleic Acids Res.* 45 (19), 10978–10988. doi: 10.1093/nar/gkx754

Jiang, P., Thomson, J. A., and Stewart, R. (2016). Quality control of single-cell RNA-seq by SinQC. *Bioinformatics* 32 (16), 2514–2516. doi: 10.1093/bioinformatics/btw176

Jiang, Y., Zhang, N. R., and Li, M. (2017). SCALE: modeling allele-specific gene expression by single-cell RNA sequencing. *Genome Biol.* 18 (1), 74. doi: 10.1186/s13059-017-1200-8

Jin, S., MacLean, A. L., Peng, T., and Nie, Q. (2018). scEpath: energy landscape-based inference of transition probabilities and cellular trajectories from single-cell transcriptomic data. *Bioinformatics* 34 (12), 2077–2086. doi: 10.1093/bioinformatics/bty058

Julia, M., Telenti, A., and Rausell, A. (2015). Sincell: an R/Bioconductor package for statistical assessment of cell-state hierarchies from single-cell RNA-seq. *Bioinformatics* 31 (20), 3380–3382. doi: 10.1093/bioinformatics/btv368

Kang, H. M., Subramaniam, M., Targ, S., Nguyen, M., Maliskova, L., McCarthy, E., et al. (2018). Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* 36 (1), 89–94. doi: 10.1038/nbt.4042

Karaayvaz, M., Cristea, S., Gillespie, S. M., Patel, A. P., Mylvaganam, R., Luo, C. C., et al. (2018). Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. *Nat. Commun.* 9 (1), 3588. doi: 10.1038/s41467-018-06052-0

Kelsey, G., Stegle, O., and Reik, W. (2017). Single-cell epigenomics: recording the past and predicting the future. *Science* 358 (6359), 69–75. doi: 10.1126/science.aan6826

Kim, K. T., Lee, H. W., Lee, H. O., Song, H. J., Jeong da, E., Shin, S., et al. (2016). Application of single-cell RNA sequencing in optimizing a combinatorial therapeutic strategy in metastatic renal cell carcinoma. *Genome Biol.* 17, 80. doi: 10.1186/s13059-016-0945-9

Kim, T., Chen, I. R., Lin, Y., Wang, A. Y., Yang, J. Y. H., and Yang, P. (2018). Impact of similarity metrics on single-cell RNA-seq data clustering. *Brief Bioinform.* doi: 10.1093/bib/bby076

Kiselev, V. Y., Andrews, T. S., and Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.* 20 (5), 273–282. doi: 10.1038/s41576-018-0088-9

Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., et al. (2017). SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods* 14 (5), 483–486. doi: 10.1038/nmeth.4236

Kiselev, V. Y., Yiu, A., and Hemberg, M. (2018). scmap: projection of single-cell RNA-seq data across data sets. *Nat. Methods* 15 (5), 359–362. doi: 10.1038/nmeth.4644

Korthauer, K. D., Chu, L. F., Newton, M. A., Li, Y., Thomson, J., Stewart, R., et al. (2016). A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biol.* 17 (1), 222. doi: 10.1186/s13059-016-1077-y

Kowalczyk, M. S., Tirosh, I., Heckl, D., Rao, T. N., Dixit, A., Haas, B. J., et al. (2015). Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Res.* 25 (12), 1860–1872. doi: 10.1101/gr.192237.115

Krieg, C., Nowicka, M., Guglietta, S., Schindler, S., Hartmann, F. J., Weber, L. M., et al. (2018). High-dimensional single-cell analysis predicts response to anti-PD-1 immunotherapy. *Nat. Med.* 24 (2), 144–153. doi: 10.1038/nm.4466

Lando, D., Stevens, T. J., Basu, S., and Laue, E. D. (2018). Calculation of 3D genome structures for comparison of chromosome conformation capture experiments with microscopy: An evaluation of single-cell Hi-C protocols. *Nucleus* 9 (1), 190–201. doi: 10.1080/19491034.2018.1438799

Lawlor, N., George, J., Bolisetty, M., Kursawe, R., Sun, L., Sivakamasundari, V., et al. (2017). Single-cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes. *Genome Res.* 27 (2), 208–222. doi: 10.1101/gr.212720.116

Leng, N., Choi, J., Chu, L. F., Thomson, J. A., Kendziorski, C., and Stewart, R. (2016). OEFinder: a user interface to identify and visualize ordering effects in single-cell RNA-seq data. *Bioinformatics* 32 (9), 1408–1410. doi: 10.1093/bioinformatics/btw004

Leng, N., Chu, L. F., Barry, C., Li, Y., Choi, J., Li, X., et al. (2015). Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments. *Nat. Methods* 12 (10), 947–950. doi: 10.1038/nmeth.3549

Lescroart, F., Wang, X., Lin, X., Swedlund, B., Gargouri, S., Sanchez-Danes, A., et al. (2018). Defining the earliest step of cardiovascular lineage segregation by single-cell RNA-seq. *Science* 359 (6380), 1177–1181. doi: 10.1126/science.aao4174

Li, H., Horns, F., Wu, B., Xie, Q., Li, J., Li, T., et al. (2017). Classifying *Drosophila* olfactory projection neuron subtypes by single-cell RNA sequencing. *Cell* 171 (5), 1206–1220 e1222. doi: 10.1016/j.cell.2017.10.019

Li, M., Zeng, T., Liu, R., and Chen, L. (2014). Detecting tissue-specific early warning signals for complex diseases based on dynamical network biomarkers: study of type 2 diabetes by cross-tissue analysis. *Brief Bioinform.* 15 (2), 229–243. doi: 10.1093/bib/bbt027

Li, W. V., and Li, J. J. (2018). An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. Commun.* 9 (1), 997. doi: 10.1038/s41467-018-03405-7

Li, X., Chen, W., Chen, Y., Zhang, X., Gu, J., and Zhang, M. Q. (2017). Network embedding-based representation learning for single cell RNA-seq data. *Nucleic Acids Res.* 45 (19), e166. doi: 10.1093/nar/gkx750

Li, Y. H., Li, D., Samusik, N., Wang, X., Guan, L., Nolan, G. P., et al. (2017). Scalable multi-sample single-cell data analysis by partition-assisted clustering and multiple alignments of networks. *PLoS Comput. Biol.* 13 (12), e1005875. doi: 10.1371/journal.pcbi.1005875

Lin, C., Jain, S., Kim, H., and Bar-Joseph, Z. (2017). Using neural networks for reducing the dimensions of single-cell RNA-Seq data. *Nucleic Acids Res.* 45 (17), e156. doi: 10.1093/nar/gkx681

Lin, P., Troup, M., and Ho, J. W. (2017). CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol.* 18 (1), 59. doi: 10.1186/s13059-017-1188-0

Lin, Y., Ghazanfar, S., Strbenac, D., Wang, A., Patrick, E., Speed, T., et al. (2017). Housekeeping genes, revisited at the single-cell level. *bioRxiv*. doi: 10.1101/229815. [Epub ahead of print].

Lindeman, I., Emerton, G., Mamanova, L., Snir, O., Polanski, K., Qiao, S. W., et al. (2018). BraCeR: B-cell-receptor reconstruction and clonality inference from single-cell RNA-seq. *Nat. Methods* 15 (8), 563–565. doi: 10.1038/s41592-018-0082-3

Liu, D., Wang, X., He, D., Sun, C., He, X., Yan, L., et al. (2018). Single-cell RNA-sequencing reveals the existence of naive and primed pluripotency in pre-implantation rhesus monkey embryos. *Genome Res.* 28 (10), 1481–1493. doi: 10.1101/gr.233437.117

Liu, L., Liu, C., Wu, L., Quintero, A., Yuan, Y., Wang, M., et al. (2018). Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity. *bioRxiv*. doi: 10.1101/316208. [Epub ahead of print].

Liu, T., and Wang, Z. (2018). scHiCNorm: a software package to eliminate systematic biases in single-cell Hi-C data. *Bioinformatics* 34 (6), 1046–1047. doi: 10.1093/bioinformatics/btx747

Lu, H., Li, J., Martinez Paniagua, M. A., Bandey, I. N., Amritkar, A., Singh, H., et al. (2018). TIMING 2.0: High-throughput single-cell profiling of dynamic cell-cell interactions by time-lapse imaging microscopy in nanowell grids. *Bioinformatics* 35 (4), 706–708. doi: 10.1093/bioinformatics/bty676

Maaten, L. v. d., and Hintton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605. doi: 10.1007/s10846-008-9235-4

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161 (5), 1202–1214. doi: 10.1016/j.cell.2015.05.002

Marco, E., Karp, R. L., Guo, G., Robson, P., Hart, A. H., Trippa, L., et al. (2014). Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proc. Natl. Acad. Sci. U. S. A.* 111 (52), E5643–E5650. doi: 10.1073/pnas.1408993111

Marinov, G. K., Williams, B. A., McCue, K., Schroth, G. P., Gertz, J., Myers, R. M., et al. (2014). From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome Res.* 24 (3), 496–510. doi: 10.1101/gr.161034.113

Matsumoto, H., Kiryu, H., Furusawa, C., Ko, M. S. H., Ko, S. B. H., Gouda, N., et al. (2017). SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics* 33 (15), 2314–2321. doi: 10.1093/bioinformatics/btx194

McCarthy, D. J., Campbell, K. R., Lun, A. T., and Wills, Q. F. (2017). Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* 33 (8), 1179–1186. doi: 10.1093/bioinformatics/btw777

McDavid, A., Finak, G., and Gottardo, R. (2016). The contribution of cell cycle to heterogeneity in single-cell RNA-seq data. *Nat. Biotechnol.* 34 (6), 591–593. doi: 10.1038/nbt.3498

Menon, V. (2018). Clustering single cells: a review of approaches on high-and low-depth single-cell RNA-seq data. *Brief Funct. Genom.* 17 (4), 240–245. doi: 10.1093/bfgp/elx044

Mezger, A., Klemm, S., Mann, I., Brower, K., Mir, A., Bostick, M., et al. (2018). High-throughput chromatin accessibility profiling at single-cell resolution. *Nat. Commun.* 9 (1), 3647. doi: 10.1038/s41467-018-05887-x

Miao, Z., Deng, K., Wang, X., and Zhang, X. (2018). DEsingle for detecting three types of differential expression in single-cell RNA-seq data. *Bioinformatics* 34 (18), 3223–3224. doi: 10.1093/bioinformatics/bty332

Nagano, T., Lubling, Y., Varnai, C., Dudley, C., Leung, W., Baran, Y., et al. (2017). Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature* 547 (7661), 61–67. doi: 10.1038/nature23001

Nelson, A. C., Mould, A. W., Bikoff, E. K., and Robertson, E. J. (2016). Single-cell RNA-seq reveals cell type-specific transcriptional signatures at the maternal-foetal interface during pregnancy. *Nat. Commun.* 7, 11414. doi: 10.1038/ncomms11414

Nguyen, Q. H., Lukowski, S. W., Chiu, H. S., Senabouth, A., Bruxner, T. J. C., Christ, A. N., et al. (2018). Single-cell RNA-seq of human induced pluripotent stem cells reveals cellular heterogeneity and cell state transitions between subpopulations. *Genome Res.* 28 (7), 1053–1066. doi: 10.1101/gr.223925.117

Ntranos, V., Kamath, G. M., Zhang, J. M., Pachter, L., and Tse, D. N. (2016). Fast and accurate single-cell RNA-seq analysis by clustering of transcript-compatibility counts. *Genome Biol.* 17 (1), 112. doi: 10.1186/s13059-016-0970-8

Papalexi, E., and Satija, R. (2018). Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat. Rev. Immunol.* 18 (1), 35–45. doi: 10.1038/nri.2017.76

Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* 344 (6190), 1396–1401. doi: 10.1126/science.1254257

Pellegrino, M., Sciambi, A., Treusch, S., Durruthy-Durruthy, R., Gokhale, K., Jacob, J., et al. (2018). High-throughput single-cell DNA sequencing of acute myeloid leukemia tumors with droplet microfluidics. *Genome Res.* 28 (9), 1345–1352. doi: 10.1101/gr.232272.117

Picelli, S., Björklund, Å. K., Faridani, O. R., Sagasser, S., Winberg, G., and Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* 10, 1096–1098. doi: 10.1038/nmeth.2639

Picelli, S., Faridani, O. R., Björklund, Å. K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. Nat. Prot. 9, 171–181. doi: 10.1038/nprot.2014.006

Poo, M. M., Du, J. L., Ip, N. Y., Xiong, Z. Q., Xu, B., and Tan, T. (2016). China Brain Project: basic neuroscience, brain diseases, and brain-inspired computing. *Neuron* 92 (3), 591–596. doi: 10.1016/j.neuron.2016.10.050

Potter, S. S. (2018). Single-cell RNA sequencing for the study of development, physiology and disease. *Nat. Rev. Nephrol.* 14 (8), 479–492. doi: 10.1038/s41581-018-0021-7

Qiu, X., Hill, A., Packer, J., Lin, D., Ma, Y.-A., and Trapnell, C. (2017a). Single-cell mRNA quantification and differential analysis with Census. *Nat. Methods* 14 (3), 309–315. doi: 10.1038/nmeth.4150

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H. A., et al. (2017b). Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* 14 (10), 979–982. doi: 10.1038/nmeth.4402

Raj, B., Wagner, D. E., McKenna, A., Pandey, S., Klein, A. M., Shendure, J., et al. (2018). Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* 36 (5), 442–450. doi: 10.1038/nbt.4103

Reinius, B., Mold, J. E., Ramskold, D., Deng, Q., Johnsson, P., Michaelsson, J., et al. (2016). Analysis of allelic expression patterns in clonal somatic cells by single-cell RNA-seq. *Nat. Genet.* 48 (11), 1430–1435. doi: 10.1038/ng.3678

Rheaume, B. A., Jereen, A., Bolisetty, M., Sajid, M. S., Yang, Y., Renna, K., et al. (2018). Single cell transcriptome profiling of retinal ganglion cells identifies cellular subtypes. *Nat. Commun.* 9 (1), 2759. doi: 10.1038/s41467-018-05134-3

Risso, D., Purvis, L., Fletcher, R. B., Das, D., Ngai, J., Dudoit, S., et al. (2018). clusterExperiment and RSEC: A Bioconductor package and framework for clustering of single-cell and other large gene expression datasets. *PLoS Comput. Biol.* 14 (9), e1006378. doi: 10.1371/journal.pcbi.1006378

Rizvi, A. H., Camara, P. G., Kandror, E. K., Roberts, T. J., Schieren, I., Maniatis, T., et al. (2017). Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development. *Nat. Biotechnol.* 35 (6), 551–560. doi: 10.1038/nbt.3854

Rizzetto, S., Koppstein, D. N. P., Samir, J., Singh, M., Reed, J. H., Cai, C. H., et al. (2018). B-cell receptor reconstruction from single-cell RNA-seq with VDJPuzzle. *Bioinformatics* 34 (16), 2846–2847. doi: 10.1093/bioinformatics/bty203

Roy, A. L., Conroy, R., Smith, J., Yao, Y., Beckel-Mitchener, A. C., Anderson, J. M., et al. (2018). Accelerating a paradigm shift: the Common Fund Single Cell Analysis Program. *Sci. Adv.* 4 (8), eaat8573. doi: 10.1126/sciadv.aat8573

Rzepiela, A. J., Ghosh, S., Breda, J., Vina-Vilaseca, A., Syed, A. P., Gruber, A. J., et al. (2018). Single-cell mRNA profiling reveals the hierarchical response of miRNA targets to miRNA induction. *Mol. Syst. Biol.* 14 (8), e8266. doi: 10.15252/msb.20188266

Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2018). A comparison of single-cell trajectory inference methods: towards more accurate and robust tools. *bioRxiv*. doi: 10.1101/276907. [Epub ahead of print].

Saliba, A. E., Westermann, A. J., Gorski, S. A., and Vogel, J. (2014). Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res.* 42 (14), 8845–8860. doi: 10.1093/nar/gku555

Sankar, P. L., and Parker, L. S. (2016). The Precision Medicine Initiative's All of Us Research Program: an agenda for research on its ethical, legal, and social issues. *Genet. Med.* 19 (7), 743–750. doi: 10.1038/gim.2016.183

Sasagawa, Y., Danno, H., Takada, H., Ebisawa, M., Tanaka, K., Hayashi, T., et al. (2018). Quartz-Seq2: a high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads. *Genome Biol.* 19 (1), 29. doi: 10.1186/s13059-018-1407-3

Sasagawa, Y., Nikaido, I., Hayashi, T., Danno, H., Uno, K. D., Imai, T., et al. (2013). Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome Biol.* 14 (4), R31. doi: 10.1186/gb-2013-14-4-r31

Schelker, M., Feau, S., Du, J., Ranu, N., Klipp, E., MacBeath, G., et al. (2017). Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. *Nat. Commun.* 8 (1), 2032. doi: 10.1038/s41467-017-02289-3

Sebe-Pedros, A., Saudemont, B., Chomsky, E., Plessier, F., Mailhe, M. P., Renno, J., et al. (2018). Cnidarian cell type diversity and regulation revealed by whole-organism single-cell RNA-Seq. *Cell* 173 (6), 1520–1534 e1520. doi: 10.1016/j.cell.2018.05.019

Setty, M., Kiseliovas, V., Levine, J., Gayoso, A., Mazutis, L., and Pe'er, D. (2019). Characterization of cell fate probabilities in single-cell data with Palantir. *Nat. Biotechnol.* 37 (4), 451–460. doi: 10.1038/s41587-019-0068-4

Severo, M. S., Landry, J. J. M., Lindquist, R. L., Goosmann, C., Brinkmann, V., Collier, P., et al. (2018). Unbiased classification of mosquito blood cells by single-cell genomics and high-content imaging. *Proc. Natl. Acad. Sci. U. S. A.* 115 (32), E7568–E7577. doi: 10.1073/pnas.1803062115

Shalek, A. K., Satija, R., Shuga, J., Trombetta, J. J., Gennert, D., Lu, D., et al. (2014). Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* 510 (7505), 363–369. doi: 10.1038/nature13437

Shi, Q., Zhang, C., Peng, M., Yu, X., Zeng, T., Liu, J., et al. (2017). Pattern fusion analysis by adaptive alignment of multiple heterogeneous omics data. *Bioinformatics* 33 (17), 2706–2714. doi: 10.1093/bioinformatics/btx176

Shindo, Y., Kondo, Y., and Sako, Y. (2018). Inferring a nonlinear biochemical network model from a heterogeneous single-cell time course data. *Sci. Rep.* 8 (1), 6790. doi: 10.1038/s41598-018-25064-w

Specht, A. T., and Li, J. (2017). LEAP: constructing gene co-expression networks for single-cell RNA-sequencing data using pseudotime ordering. *Bioinformatics* 33 (5), 764–766. doi: 10.1093/bioinformatics/btw729

Stevens, T. J., Lando, D., Basu, S., Atkinson, L. P., Cao, Y., Lee, S. F., et al. (2017). 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* 544 (7648), 59–64. doi: 10.1038/nature21429

Strauss, M. E., Reid, J. E., and Wernisch, L. (2018). GPseudoRank: a permutation sampler for single cell orderings. *Bioinformatics* 35 (4), 611–618. doi: 10.1093/bioinformatics/bty664

Su, X., Shi, Y., Zou, X., Lu, Z. N., Xie, G., Yang, J. Y. H., et al. (2017). Single-cell RNA-seq analysis reveals dynamic trajectories during mouse liver development. *BMC Genomics* 18 (1), 946. doi: 10.1186/s12864-017-4342-x

Svensson, V., Natarajan, K. N., Ly, L.-H., Miragaia, R. J., Labalette, C., Macaulay, I. C., et al. (2017). Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods* 14 (4), 387–387. doi: 10.1038/nmeth.4220

Svensson, V., Vento-Tormo, R., and Teichmann, S. A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Prot.* 13 (4), 599–604. doi: 10.1038/nprot.2017.149

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., et al. (2009). mRNA-seq whole-transcriptome analysis of a single cell. *Nat. Methods* 6 (5), 377–382. doi: 10.1038/nmeth.1315

Tian, L., Su, S., Dong, X., Amann-Zalcenstein, D., Biben, C., Seidi, A., et al. (2018). scPipe: a flexible R/Bioconductor preprocessing pipeline for single-cell RNA-sequencing data. *PLoS Comput. Biol.* 14 (8), e1006361. doi: 10.1371/journal.pcbi.1006361

Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth, M. H., Treacy, D., Trombetta, J. J., et al. (2016a). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* 352 (6282), 189–196. doi: 10.1126/science.aad0501

Tirosh, I., Venteicher, A. S., Hebert, C., Escalante, L. E., Patel, A. P., Yizhak, K., et al. (2016b). Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature* 539 (7628), 309–313. doi: 10.1038/nature20123

Todorov, H., and Saeys, Y. (2018). Computational approaches for high-throughput single-cell data analysis. *FEBS J.* 286 (8), 1451–1467. doi: 10.1111/febs.14613

Torre, E., Dueck, H., Shaffer, S., Gospocic, J., Gupte, R., Bonasio, R., et al. (2018). Rare cell detection by single-cell RNA sequencing as guided by single-molecule RNA FISH. *Cell Syst.* 6 (2), 171–179 e175. doi: 10.1016/j.cels.2018.01.014

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., et al. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* 32 (4), 381–386. doi: 10.1038/nbt.2859

Tsang, J. C. H., Vong, J. S. L., Ji, L., Poon, L. C. Y., Jiang, P., Lui, K. O., et al. (2017). Integrative single-cell and cell-free plasma RNA transcriptomics elucidates placental cellular dynamics. *Proc. Natl. Acad. Sci. U. S. A.* 114 (37), E7786–E7795. doi: 10.1073/pnas.1710470114

Upadhyay, A. A., Kauffman, R. C., Wolabaugh, A. N., Cho, A., Patel, N. B., Reiss, S. M., et al. (2018). BALDR: a computational pipeline for paired heavy and light chain immunoglobulin reconstruction in single-cell RNA-seq data. *Genome Med.* 10 (1), 20. doi: 10.1186/s13073-018-0528-3

Usoskin, D., Furlan, A., Islam, S., Abdo, H., Lonnerberg, P., Lou, D., et al. (2015). Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nat. Neurosci.* 18 (1), 145–153. doi: 10.1038/nn.3881

Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S., and Marioni, J. C. (2017). Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat. Methods* 14 (6), 565–571. doi: 10.1038/nmeth.4292

Van den Berge, K., Perraudeau, F., Soneson, C., Love, M. I., Risso, D., Vert, J. P., et al. (2018). Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. *Genome Biol.* 19 (1), 24. doi: 10.1186/s13059-018-1406-4

van der Wijst, M. G. P., Brugge, H., de Vries, D. H., Deelen, P., Swertz, M. A., LifeLines Cohort, S., et al. (2018). Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.* 50 (4), 493–497. doi: 10.1038/s41588-018-0089-9

Vegh, P., and Haniffa, M. (2018). The impact of single-cell RNA sequencing on understanding the functional organization of the immune system. *Brief Funct. Genom.* 17 (4), 265–272. doi: 10.1093/bfgp/ely003

Velten, L., Haas, S. F., Raffel, S., Blaszkiewicz, S., Islam, S., Hennig, B. P., et al. (2017). Human haematopoietic stem cell lineage commitment is a continuous process. *Nat. Cell Biol.* 19 (4), 271–281. doi: 10.1038/ncb3493

Venteicher, A. S., Tirosh, I., Hebert, C., Yizhak, K., Neftel, C., Filbin, M. G., et al. (2017). Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. *Science* 355, 6332. doi: 10.1126/science.aai8478

Vergara, H. M., Bertucci, P. Y., Hantz, P., Tosches, M. A., Achim, K., Vopalensky, P., et al. (2017). Whole-organism cellular gene-expression atlas reveals conserved cell types in the ventral nerve cord of Platynereis dumerilii. *Proc. Natl. Acad. Sci. U. S. A.* 114 (23), 5878–5885. doi: 10.1073/pnas.1610602114

Vieth, B., Ziegenhain, C., Parekh, S., Enard, W., and Hellmann, I. (2017). powsimR: power analysis for bulk and single cell RNA-seq experiments. *Bioinformatics* 33 (21), 3486–3488. doi: 10.1093/bioinformatics/btx435

Villani, A. C., Satija, R., Reynolds, G., Sarkizova, S., Shekhar, K., Fletcher, J., et al. (2017). Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* 356, 6335. doi: 10.1126/science.aah4573

Vu, T. N., Wills, Q. F., Kalari, K. R., Niu, N., Wang, L., Pawitan, Y., et al. (2018). Isoform-level gene expression patterns in single-cell RNA-sequencing data. *Bioinformatics* 34 (14), 2392–2400. doi: 10.1093/bioinformatics/bty100

Vu, T. N., Wills, Q. F., Kalari, K. R., Niu, N., Wang, L., Rantalainen, M., et al. (2016). Beta-Poisson model for single-cell RNA-seq data analyses. *Bioinformatics* 32 (14), 2128–2135. doi: 10.1093/bioinformatics/btw202

Wang, B., Zhu, J., Pierson, E., Ramazzotti, D., and Batzoglou, S. (2017). Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat. Methods* 14 (4), 414–416. doi: 10.1038/nmeth.4207

Wang, B., Ramazzotti, D., Sano, L. D., Zhu, J., Pierson, E., and Batzoglou, S. (2018). SIMLR: a tool for large-scale genomic analyses by multi-kernel learning. *Proteomics* 18, 2. doi: 10.1002/pmic.201700232

Wang, L., Yu, X., Zhang, C., and Zeng, T., (2018). Detecting personalized determinants during drug treatment from Omics big data. *Curr. Pharm. Des*. doi: 10.2174/1381612824666181106102111

Wang, Y. J., Schug, J., Won, K. J., Liu, C., Naji, A., Avrahami, D., et al. (2016). Single-cell transcriptomics of the human endocrine pancreas. *Diabetes* 65 (10), 3028–3038. doi: 10.2337/db16-0405

Weinreb, C., Wolock, S., Tusi, B. K., Socolovsky, M., and Klein, A. M. (2018). Fundamental limits on dynamic inference from single-cell snapshots. *Proc. Natl. Acad. Sci. U. S. A.* 115 (10), E2467–E2476. doi: 10.1073/pnas.1714723115

Welch, J. D., Hartemink, A. J., and Prins, J. F. (2016). SLICER: inferring branched, nonlinear cellular trajectories from single cell RNA-seq data. *Genome Biol.* 17 (1), 106. doi: 10.1186/s13059-016-0975-3

Wu, A. R., Neff, N. F., Kalisky, T., Dalerba, P., Treutlein, B., Rothenberg, M. E., et al. (2014). Quantitative assessment of single-cell RNA-sequencing methods. *Nat. Methods* 11 (1), 41–46. doi: 10.1038/nmeth.2694

Wu, Z., Zhang, Y., Stitzel, M. L., and Wu, H. (2018). Two-phase differential expression analysis for single cell RNA-seq. *Bioinformatics* 34 (19), 3340–3348. doi: 10.1093/bioinformatics/bty329

Xie, P., Gao, M., Wang, C., Zhang, J., Noel, P., Yang, C., et al. (2019). SuperCT: a supervised-learning framework for enhanced characterization of single-cell transcriptomic profiles. *Nucleic Acids Res.* 47 (8), e48 doi: 10.1093/nar/gkz116

Xin, Y., Kim, J., Okamoto, H., Ni, M., Wei, Y., Adler, C., et al. (2016). RNA sequencing of single human islet cells reveals type 2 diabetes genes. *Cell Metab.* 24 (4), 608–615. doi: 10.1016/j.cmet.2016.08.018

Xu, C., and Su, Z. (2015). Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* 31 (12), 1974–1980. doi: 10.1093/bioinformatics/btv088

Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C. Y., Feng, Y., et al. (2013). Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* 500 (7464), 593–597. doi: 10.1038/nature12364

Yang, A., Troup, M., Lin, P., and Ho, J. W. (2017). Falco: a quick and flexible single-cell RNA-seq processing framework on the cloud. *Bioinformatics* 33 (5), 767–769. doi: 10.1101/064006

Yang, Y., Huh, R., Culpepper, H. W., Lin, Y., Love, M. I., and Li, Y. (2018). SAFE-clustering: Single-cell Aggregated (From Ensemble) clustering for single-cell RNA-seq data. *Bioinformatics*. 35 (8), 1269–1277. doi: 10.1093/bioinformatics/bty793

Yip, S. H., Sham, P. C., and Wang, J. (2018). Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data. *Brief Bioinform.* doi: 10.1093/bib/bby011. [Epub ahead of print].

Yip, S. H., Wang, P., Kocher, J. A., Sham, P. C., and Wang, J. (2017). Linnorm: improved statistical analysis for single cell RNA-seq expression data. *Nucleic Acids Res.* 45 (22), e179. doi: 10.1093/nar/gkx1189

Young, M. D., Mitchell, T. J., Vieira Braga, F. A., Tran, M. G. B., Stewart, B. J., Ferdinand, J. R., et al. (2018). Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. *Science* 361 (6402), 594–599. doi: 10.1126/science.aat1699

Yu, T. (2018). A new dynamic correlation algorithm reveals novel functional aspects in single cell and bulk RNA-seq data. *PLoS Comput. Biol.* 14 (8), e1006391. doi: 10.1371/journal.pcbi.1006391

Yu, X., Zhang, J., Sun, S., Zhou, X., Zeng, T., and Chen, L. (2017). Individual-specific edge-network analysis for disease prediction. *Nucleic Acids Res.* 45 (20), e170. doi: 10.1093/nar/gkx787

Yu, X. T., and Zeng, T. (2018). Integrative analysis of omics big data. *Methods Mol. Biol.* 1754, 109–135. doi: 10.1007/978-1-4939-7717-8_7

Zappia, L., Phipson, B., and Oshlack, A. (2017). Splatter: simulation of single-cell RNA sequencing data. *Genome Biol.* 18 (1), 174. doi: 10.1186/s13059-017-1305-0

Zeisel, A., Munoz-Manchado, A. B., Codeluppi, S., Lonnerberg, P., La Manno, G., Jureus, A., et al. (2015). Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 347 (6226), 1138–1142. doi: 10.1126/science.aaa1934

Zeng, C., Mulas, F., Sui, Y., Guan, T., Miller, N., Tan, Y., et al. (2017). Pseudotemporal ordering of single cells reveals metabolic control of postnatal beta cell proliferation. *Cell Metab.* 25 (5), 1160–1175 e1111. doi: 10.1016/j.cmet.2017.04.014

Zeng, T., Sun, S. Y., Wang, Y., Zhu, H., and Chen, L. (2013). Network biomarkers reveal dysfunctional gene regulations during disease progression. *FEBS J.* 280 (22), 5682–5695. doi: 10.1111/febs.12536

Zeng, T., Wang, D. C., Wang, X., Xu, F., and Chen, L. (2014). Prediction of dynamical drug sensitivity and resistance by module network rewiring-analysis based on transcriptional profiling. *Drug Resist. Updat.* 17 (3), 64–76. doi: 10.1016/j.drup.2014.08.002

Zeng, T., Zhang, W., Yu, X., Liu, X., Li, M., and Chen, L. (2016). Big-data-based edge biomarkers: study on dynamical drug sensitivity and resistance in individuals. *Brief Bioinform.* 17 (4), 576–592. doi: 10.1093/bib/bbv078

Zhang, H., Lee, C. A. A., Li, Z., Garbe, J. R., Eide, C. R., Petegrosso, R., et al. (2018). A multitask clustering approach for single-cell RNA-seq analysis in recessive dystrophic epidermolysis bullosa. *PLoS Comput. Biol.* 14 (4), e1006053. doi: 10.1371/journal.pcbi.1006053

Zhang, J. M., Fan, J., Fan, H. C., Rosenfeld, D., and Tse, D. N. (2018). An interpretable framework for clustering single-cell RNA-Seq datasets. *BMC Bioinform.* 19 (1), 93. doi: 10.1186/s12859-018-2092-7

Zhang, L., and Zhang, S. (2018). Comparison of computational methods for imputing single-cell RNA-sequencing data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi: 10.1109/TCBB.2018.2848633. [Epub ahead of print].

Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., et al. (2017). Comparative analysis of single-cell RNA sequencing methods. *Mol. Cell* 65 (4), 631–643 e634. doi: 10.1016/j.molcel.2017.01.023

Zong, C. C. (2017). Single-cell RNA-seq study determines the ontogeny of macrophages in glioblastomas. *Genome Biol.* 18 (1), 235. doi: 10.1186/s13059-017-1375-z

Conflict of Interest Statement: 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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 18 July 2019
doi: 10.3389/fgene.2019.00671

Pan-Cancer and Single-Cell Modeling of Genomic Alterations Through Gene Expression

Daniele Mercatelli[1], *Forest Ray*[2] *and Federico M. Giorgi*[1]*

[1] Department of Pharmacy and Biotechnology, University of Bologna, Bologna, Italy, [2] Department of Systems Biology, Columbia University Medical Center, New York, NY, United States

Cancer is a disease often characterized by the presence of multiple genomic alterations, which trigger altered transcriptional patterns and gene expression, which in turn sustain the processes of tumorigenesis, tumor progression, and tumor maintenance. The links between genomic alterations and gene expression profiles can be utilized as the basis to build specific molecular tumorigenic relationships. In this study, we perform pan-cancer predictions of the presence of single somatic mutations and copy number variations using machine learning approaches on gene expression profiles. We show that gene expression can be used to predict genomic alterations in every tumor type, where some alterations are more predictable than others. We propose gene aggregation as a tool to improve the accuracy of alteration prediction models from gene expression profiles. Ultimately, we show how this principle can be beneficial in intrinsically noisy datasets, such as those based on single-cell sequencing.

Keywords: NGS (next generation sequencing), genomics, cancer, TCGA, single-cell sequencing

OPEN ACCESS

Edited by:
Yifei Xu,
University of Oxford,
United Kingdom

Reviewed by:
Ashok Sharma,
Augusta University,
United States
Hauke Busch,
Universität zu Lübeck,
Germany

***Correspondence:**
Federico M. Giorgi
federico.giorgi@unibo.it

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *01 March 2019*
Accepted: *27 June 2019*
Published: *18 July 2019*

Citation:
Mercatelli D, Ray F and Giorgi FM (2019) Pan-Cancer and Single-Cell Modeling of Genomic Alterations Through Gene Expression. Front. Genet. 10:671. doi: 10.3389/fgene.2019.00671

INTRODUCTION

Cancer is a molecular disease occurring when a cell or group of cells acquire uncontrolled proliferative behavior, conferred by a multitude of deregulations in specific pathways (Hanahan and Weinberg, 2011). As is implied by such a broad definition, cancer is a highly heterogeneous disease, showing remarkably different molecular, histological, genetic, and clinical properties, even when comparing tumors originating from the same tissue (Meacham and Morrison, 2013). Many cancers are characterized by the presence of single nucleotide or short indel mutations and/or copy number alterations, which appear somatically at the early stages of oncogenesis and can drive tumor progression (Bozic et al., 2010). Cancers can be broadly divided in two classes: the M class, where point mutations are prevalent, and the C class, where copy number variations (CNVs) are more numerous and are often associated with TP53 mutations. Tumor class influences anatomic location. Most ovarian cancers, for example, belong to the C class, while most colorectal cancers belong to the M class, although many exceptions do exist (Ciriello et al., 2013).

The Cancer Genome Atlas (TCGA) project (Chang et al., 2013) has recently undergone a major effort to collect vast amounts of information on thousands of distinct tumor samples. The TCGA data collection, commonly referred to as the "pan-cancer" dataset, provided the scientific community with an avalanche of data on DNA alterations, gene expression, methylation status, and protein abundances among others, with the critical mass necessary to identify rarer driver tumorigenesis effects in many types of cancers (Brennan et al., 2013; Cancer Genome Atlas

Network, 2015; Leiserson et al., 2015). By combining all 33 TCGA datasets, Bailey and colleagues (Bailey et al., 2018) recently outlined a pan-cancer map of which mutations can be drivers for the progression of cancer.

The availability of thousands of samples measuring many different variables in cancer has allowed scientists to generate statistical models of relationships between different molecular species. A pan-cancer correlation network between coding genes and long noncoding RNAs, for example, sheds light on the function of non-coding parts of the transcriptome (Liu and Zhao, 2016). More recently, mutations on transcription factors (TFs) have been linked to altered gene expressions and phosphoprotein levels in 12 TCGA tumor type datasets (Osmanbeyoglu et al., 2017). Network approaches have been applied to identify clusters of coexpressed genes, shared by multiple cancer types (Kim and Kim, 2018). Several studies have sought to characterize the relationships between genomic status and expression levels in cancer, trying to identify commonalities across different cancer types (Ghazanfar and Yang, 2016; Sharma et al., 2018). In particular, Alvarez and colleagues (Alvarez et al., 2016) have postulated that the effect of genomic alterations in cancer can be more readily assessed by aggregating gene expression profiles into transcriptional networks, rather than by profiles taken separately.

While the association between genomic events and gene expression is proven in several scenarios, it remains to be seen if it can be assessed in scenarios where fully quantitative readouts are unavailable, such as low-coverage samples. One of these scenarios is single-cell sequencing (Nawy, 2013), often carried out in experiments where thousands of mutations are generated *via* a system of pooled CRISPR-Cas9 knockouts (Datlinger et al., 2017).

To our knowledge, there is no study trying to identify relationships between all genomic alteration events (somatic mutations/indels and CNVs) and global gene expression across cancers. In this study, we use 24 TCGA tumor datasets to investigate whether gene expression can be used to predict the presence of specific genomic alterations in several cancer tissue contexts. To this end, we leverage the current availability of a vast family of machine learning algorithms (Kuhn, 2008). We investigate whether some gene alterations can be better modeled than others and whether using grouped gene expression profiles as aggregated variables can effectively identify specific genomic alterations. Finally, we test whether predicting mutations and CNVs can be carried out in an intrinsically noisy single-cell RNA-Seq (scRNA-Seq) transcriptomics datasets.

RESULTS

Collection of Pan-Cancer Dataset

We downloaded the most recent version of the TCGA datasets available on Firehose (v2016_01_28), encompassing mutational, CNV, and gene expression data. Initially, we organized the expression data as a matrix of 9,642 samples and 20,531 genes, visualized in **Figure 1A** using T-distributed stochastic neighbor embedding (TSNE; van der Maaten and Hinton, 2008) clustering and two-dimensional (2D)-density estimates for each tumor type. As observed before (Chen et al., 2018), the transcriptional properties of TCGA tumors separate tumor types by tissue of origin. In particular, two tumor types segregate into two subgroups: breast cancer, which subdivides into a major luminal cluster and a smaller (in terms of samples collected) basal cluster (Perou et al., 2000); and esophageal carcinoma, which roughly subdivides into adenocarcinomas and squamous cell carcinomas (TCGA network, 2017).

We then aggregated the single nucleotide and short indel somatic mutation data from the same samples for which we had collected gene expression. As is widely known, TP53 is the most mutated gene in human cancer (**Figure 1B**), followed by PIK3CA, SYNE1, and KRAS. As shown before (Ciriello et al., 2013), some tumor types are characterized by a high presence of somatic mutations. In particular, lung squamous carcinoma (LUSC), mesothelioma, and esophageal cancer carry at least one of these events in almost 100% of the samples in the TCGA dataset. In the figure, we filtered out commonly known nondriver mutations (Lawrence et al., 2013), such as those happening in long genes like TTN and OBSCN, but we kept them in all following analyses for the sake of completion. A representation of all mutated genes, including blacklisted ones, is available in **Figure S1**. Some tumors are characterized by the prevalence of a mutation in a specific gene, such as the G-protein coding BRAF in thyroid carcinoma (Kimura et al., 2003) or IDH1, translating into isocitrate dehydrogenase, in low-grade glioma (Yan et al., 2009).

Finally, we obtained readouts of CNV status for all TCGA samples. CNVs can have different extensions in terms of nucleotides affected and can sometimes encompass entire chromosomes (Shlien and Malkin, 2009) and the thousands of genes therein. In order to limit the number of variables to a more meaningful subset, we assigned a CNV score to every gene, according to the copy number score of the genomic region most overlapping with the University of California, Santa Cruz-annotated gene boundaries (genome version hg19). We then tested models for all genes affected by a CNV in at least 10 samples [extending what was previously done in Chen et al. (2014)]. In order to make CNV variables comparable with the mutational ones, we defined a cutoff for presence or absence by using the $\log_2$(CNV) threshold of 0.5, which roughly corresponds to at least one copy gain for amplifications, and at least one copy loss for deletions (see Materials and Methods). We then reported their abundance in the pan-cancer dataset, distinguishing between amplifications (**Figure 1C**) and deletions (**Figure 1D**). As previously shown (Ciriello et al., 2013), virtually all ovarian cancer samples are characterized by at least one CNV event. Among the most amplified genes, we find the oncogenes SOX2 (Bass et al., 2009), EGFR (Bell et al., 2005), and MDM2 (Momand et al., 1998), and also a noncoding gene, PVT1, the most amplified gene in breast cancer, with proven but as-of-yet uncharacterized

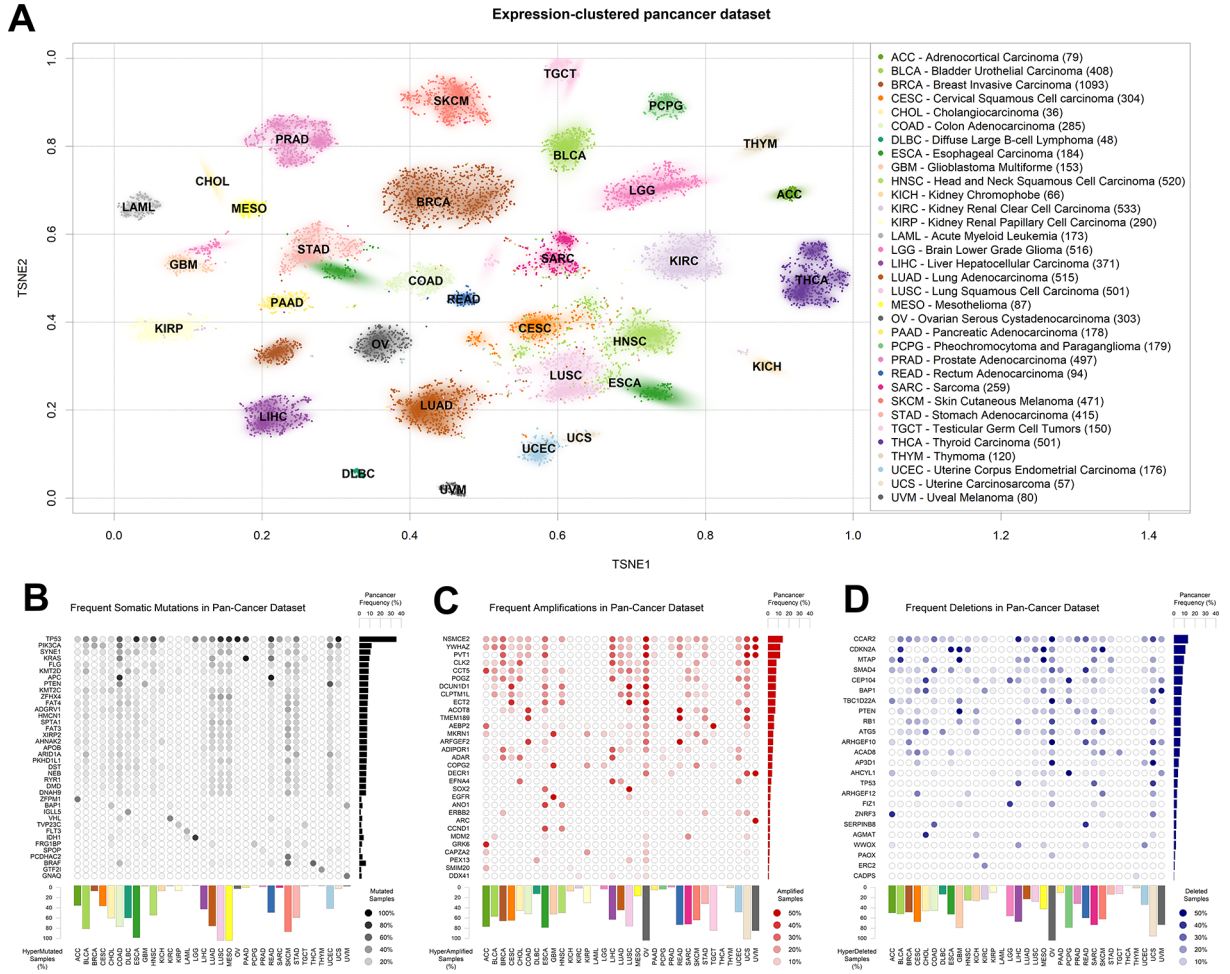


FIGURE 1 | The Cancer Genome Atlas (TCGA) dataset used. **(A)** T-distributed stochastic neighbor embedding (TSNE) clustering of TCGA samples based on gene expression profiles. The 2D median of each tumor type is indicated using the TCGA tumor code. Subset size is indicated in brackets next to tumor type names to the right. **(B)** Table of most somatically mutated genes across TCGA tumor samples, in terms of number of samples where the gene is somatically mutated with altered protein product sequence. **(C)** Table of most amplified genes across TCGA tumor samples. **(D)** Table of most deleted genes across TCGA tumor samples. The fraction of total TCGA samples carrying a gene-targeting event is indicated to the right of panels **(B–D)**, and the fraction of samples where more than 0.5% of the genes is affected by the panel event type is indicated to the bottom of panels **(B–D)**.

proto-oncogenic effects (Colombo et al., 2015; Li et al., 2017). Among the most deleted genes (**Figure 1D**), we observe well-known tumor-suppressor genes, such as CDKN2A (Usvasalo et al., 2008; Mistry et al., 2015) and PTEN (Zhao et al., 2017; Wang et al., 2018).

Modeling Cancer Alterations With Gene Expression

After collecting all the expression and genomic alteration data from TCGA, we set out to generate models that are able to predict the presence or absence of each event by virtue of gene expression data in the contexts of all collected tumor types.

We tested several modeling algorithms for classification using the aggregator platform for machine learning caret (Kuhn, 2008) in the bladder cancer mutational dataset (Robertson et al., 2017). In our rationale, we tested at least one algorithm from every major machine learning family (decision trees, support vector machine, neural networks, and linear models; see Methods for a full list). We observed that all models provide better-than-random predictions for the majority of mutational events, in terms of area under the ROC curve (AUROC) (**Figure 2**) (Fawcett, 2006). For the bulk of the subsequent analysis, we selected the top-scoring algorithm in this test, the gradient boost modeling algorithm (gbm), a well-established tree-based boosting model (Friedman, 2001), due to its robustness and speed of implementation. In all our test runs (**Figure 2** for bladder cancer and Figure S2 for liver hepatocellular carcinoma), gbm models are not significantly different (in terms of AUROC comparison, two-tailed Wilcoxon Test $p > 0.1$) from other well-performing algorithms, such as linear discriminant analysis or support vector machine.

We therefore calculated gbm models for all tumor types of at least 100 samples with co-measured expression and CNV or

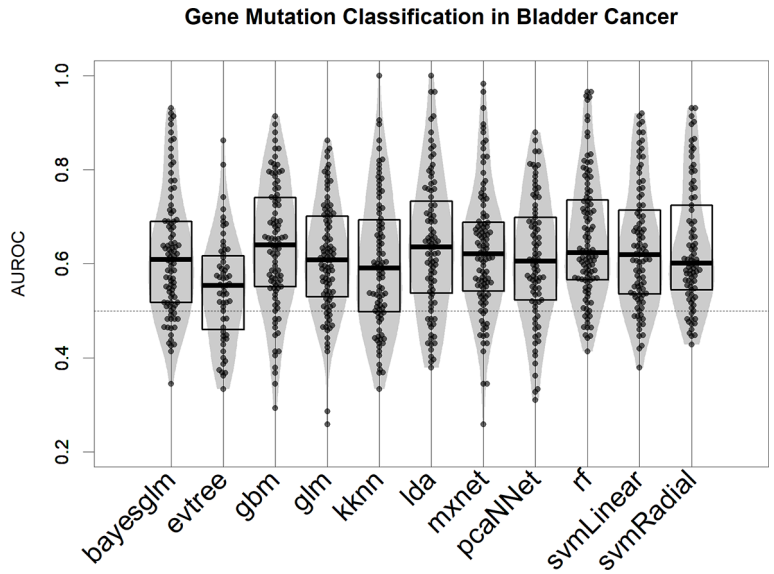


FIGURE 2 | Performance of 11 machine learning algorithms in binary classification of mutated/nonmutated samples using gene expression predictor variables in the bladder cancer dataset. Each point corresponds to a specific mutation/model. Performance is indicated as AUROC: area under the receiver operating characteristic curve.

mutations, which included 24 of the 33 TCGA tumor types. The models were predictive of genomic events observed in no less than 5% and no more than 95% of the patients in the dataset, and at least in 10 samples. Our results show that in all tumor types, a machine learning algorithm based on gene expression is consistently better than a random predictor (AUROC line at 0.5) at correctly classifying tumor samples for the presence or absence of specific genomic alteration events (**Figure 3** and **Supplementary Table S1**).

We focused on TP53 somatic alteration models not only because this tumor suppressor gene is frequently mutated or lost in cancer (**Figure 1**) but also because its loss of function is one of the most common driver events associated to tumorigenesis (Petitjean et al., 2007). In our study, TP53 mutations are well modeled in many of these tumor types (**Figure 3**), being the most well-predicted mutational event in both acute myeloid leukemia and low-grade glioma. In these tumors, loss-of-function somatic mutations of TP53 have been recurrently found as driver events for tumor initiation (Venneti and Huse, 2015; Metzeler et al., 2016). We could also model the presence of a copy loss of TP53 in sarcoma, which can be predicted with an accuracy of 70%. Ovarian and pancreatic cancer datasets presented exceptional cases, where TP53 is mutated virtually in all patients (next to 95%) (Cole et al., 2016; Cicenas et al., 2017). This presents a challenge for the modeling algorithm, as there are not enough wild-type samples to perform a robust training (TP53 model performances in these tumors are close to 0.5, i.e. randomness).

We further focused on models predicting KRAS, a very important oncogene whose protein product is fundamental in transmitting proliferation signals in the early steps of the mitogen-activated protein kinase cascade (Tsuchida et al., 1982). KRAS's role in cancer is caused by specific point mutations in its guanosine triphosphate-binding domain, which make it constantly active and therefore a deregulated signal transducer for proto-oncogenic pathways (Kranenburg, 2005). Our results confirm the key role of KRAS-targeting somatic mutations, which are well modeled by gene expression in KRAS-driven tumors: colon, lung, pancreas, stomach, and testicular cancers, as well as cervical squamous carcinoma (Prior et al., 2012) (**Figure 3**). Less commonly, the oncogenic activity of KRAS can be increased by amplification in ovarian cancer (Huang et al., 2012) and LUSC (Wagner et al., 2011). Our results show that patients can be well separated between KRAS-amplified and KRAS-normal using gene expression in these two tumor types, confirming the presence of a transcriptionally defined subset of patients with KRAS copy number gains.

In general, the observed high variability between somatic mutations and CNVs roots is due to the fact that not all genomic alterations are disease drivers, and some are simply passenger events (Bozic et al., 2010), located either close to the amplified oncogene/deleted tumor suppressor gene, or hypermutated due to deficits in the DNA damage repair mechanisms (Chae et al., 2016), such as the case of skin melanoma (Guan et al., 2015). Differences between mutation and CNV model performances in individual cancer types may be due to the specific characteristics of these. For example, LUSC initiation and progression tend to depend on copy number alterations (Ciriello et al., 2013) rather than somatic mutations, which is highlighted by the highest performance of CNV-predicting transcription-based models over mutation-predicting ones (**Figure 3**). However, the biological heterogeneity observed within cancer datasets does not allow for perfect generalizations, such as tumor types driven exclusively by CNVs or mutations (Smith and Sheltzer, 2018).

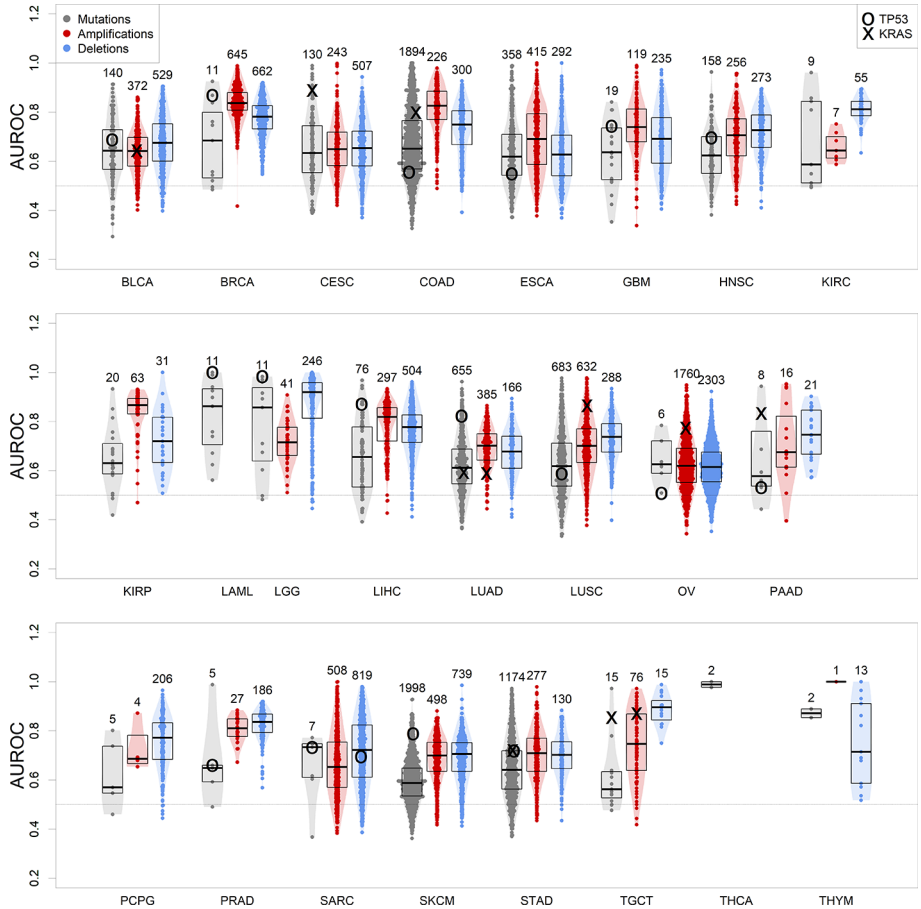


FIGURE 3 | Performance of gbm models for each genomic alteration event in TCGA, predicted as a function of each tumor gene expression. Boxplots indicate distribution median, upper and lower quartile. Alterations targeting TP53 and KRAS are indicated. Numbers on top of the violin plots indicate the number of models generated.

We noted a tendency where models for more frequent CNV events yielded a greater predictive power (**Figure S3**), a tendency not observed for somatic mutation models. We then tested if known tumor-related genes, such as those curated by the Cancer Gene Census (Futreal et al., 2004) are better modeled than the rest of the genome. There is no difference in mutation and amplification results, but for deletion events, oncogenes yield weaker models (Wilcoxon test, $p = 0.0037$, **Figure S4**), and tumor suppressor genes yield generally stronger models ($p = 0.00050$). This is in agreement with the central paradigm of cancer, where a tumor suppressor gene deletion can be one of the driving events of tumorigenesis and tumor progression (Sager, 1989). On the other hand, deletion of tumor-promoting oncogenes is generally unfavorable for tumor progression, and so, generally speaking, it should be present only as a passenger event, unlikely to determine global gene expression and tumor fate.

Modeling Specific Alterations With Noise Addition

In order to understand whether cancer-related genomic alterations can be modeled by gene expression in scenarios with lower signal-to-noise ratio, we artificially perturbed the TCGA gene expression dataset *via* the addition of Gaussian noise and then proceeded to build models to predict the presence of TP53 mutations in breast cancer, the largest dataset in TCGA by number of samples.

As expected, the addition of uniform random Gaussian noise to the gene expression matrix has a detrimental effect on the amount of information left for modeling the presence of TP53 somatic mutations (**Figure 4A**).

We then decided to test several permutations of noise addition on the same breast cancer expression data, by each time aggregating genes into networks defined *a priori* in the same context, using a Tukey biweight robust average method (Irizarry et al., 2006) on weighted gene correlation network analysis (WGCNA) clusters (Langfelder and Horvath, 2008) and the VIPER algorithm (Alvarez et al., 2016) on ARACNe-AP networks (Lachmann et al., 2016). It is important to note that WGCNA clusters are completely nonoverlapping and yield generally a lower number of aggregated variables than VIPER clusters, which are groups of genes possibly shared by other TF clusters and that collectively yield the global expression of a TF target set (dubbed as a proxy for "TF activity" in the original VIPER manuscript; Alvarez et al., 2016).

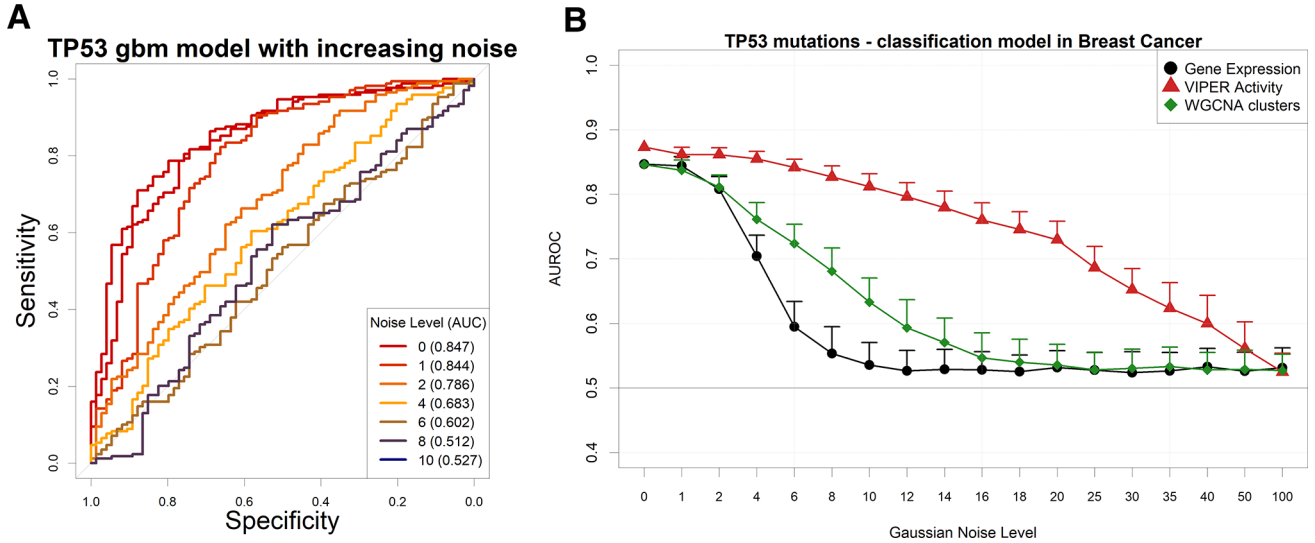


FIGURE 4 | Performance of a TP53 somatic mutation gbm model upon Gaussian noise addiction. **(A)** Receiver operating characteristic (ROC) curves (and area under the curve) upon addition of increasing levels (in terms of SD of a Gaussian distribution with mean = 0) of Gaussian noise. **(B)** AUROCs of the model with increasing noise, calculated using gene expression (black line) or aggregated gene expression using the WGCNA (green line) or VIPER (red line) algorithms. Error bars indicate the standard deviation of AUROC distribution.

Our results show that gene expression, VIPER activity, and WGCNA clusters yield very similar models for predicting TP53 mutations in breast cancer (**Figure S5**). The amount of information contained in the input variables is therefore comparable. Adding noise to the input expression matrix, however, and then aggregating the resulting noise-burdened genes into VIPER or WGCNA clusters (see Materials and Methods), provides robustness to the models (**Figure 4B**). Similar results with higher variances (possibly due to the smaller size of the datasets) can be observed for EGFR amplifications in glioblastoma (**Figure S6**) and LUSC (**Figure S7**), for PVT1 amplifications in ovarian cancer (**Figure S8**) and for PTEN deletions in sarcoma (**Figure S9**). In all these examples, however, the performance of the simple WGCNA/Tukey aggregation is closer (if not worse) to that of simple gene expression.

An alternative way to reduce the information content from an NGS gene expression dataset is to reduce the number of read counts from each sample. This operation reflects either a low-coverage bulk RNA-Seq experiment or an experiment arising from single-cell sequencing (Pollen et al., 2014). In particular, single-cell RNA-Seq (scRNA-Seq) is characterized by the dropout phenomenon (Risso et al., 2018) wherein genes expressed in the cells are sometimes not detected at all. In order to simulate such scenarios, we down-sampled each RNA-Seq gene count profile from the largest TCGA dataset (breast cancer) to a target aligned read number using a beta function, which allows for reduction coupled with random complete gene dropouts (**Figure 5A**). We then modeled again the presence of TP53 mutations using gene expression (**Figure 5B**). We found out that models based on standard unaggregated gene expression experience an accuracy drop at around 30M reads, while aggregating genes using VIPER (but not with WGCNA) allows for better-than-random accuracies even at 3M reads, confirming the benefits of gene aggregation in low-coverage RNA-Seq, as previously found e.g. for sample clustering (Bush et al., 2017).

Mutation Prediction in Single-Cell Data

Based on the results from the pan-cancer analysis, where we predicted sample mutations based on pooled RNA-Seq gene expression patterns, we decided to extend the same approach on single-cell datasets. Recently, the CROP-Seq methodology has been introduced (Datlinger et al., 2017), allowing for the measurement of cell-specific transcriptome-wide gene expression and mutations induced by CRISPR-Cas9 (Ran et al., 2013), thanks to the concurrent sequencing of CRISPR-Cas9 guide RNAs. We therefore tested the capability of gbm models to predict mutations using gene expression variables in two independent single-cell datasets. The first dataset (dubbed "Datlinger") was extracted from the Jurkat cell line derived from human T lymphocytes (Datlinger et al., 2017). The second one (dubbed "Shifrut") derived from primary unstimulated T cells from a human donor (Shifrut et al., 2018). We removed cell unique molecular identifier counts and cell cycle as common confounding effects of single-cell datasets (Tirosh et al., 2016) (**Figure S11**). We generated a regulatory transcription network using ARACNe-AP on the RNA-Seq Cancer Cell Line Encyclopedia dataset (CCLE; Barretina et al., 2012), which comprises 1,021 distinct human cell lines. Using the CCLE network, we aggregated gene expression from the single-cell datasets using the VIPER algorithm and implemented the resulting TF-centered VIPER activity profiles to build prediction models for the Crop-Seq-detected mutations. Parallelly, we built models using un-aggregated variance stabilizing transformation (vst)-normalized gene expression data. Our results show that gbm models based on VIPER activity variables globally

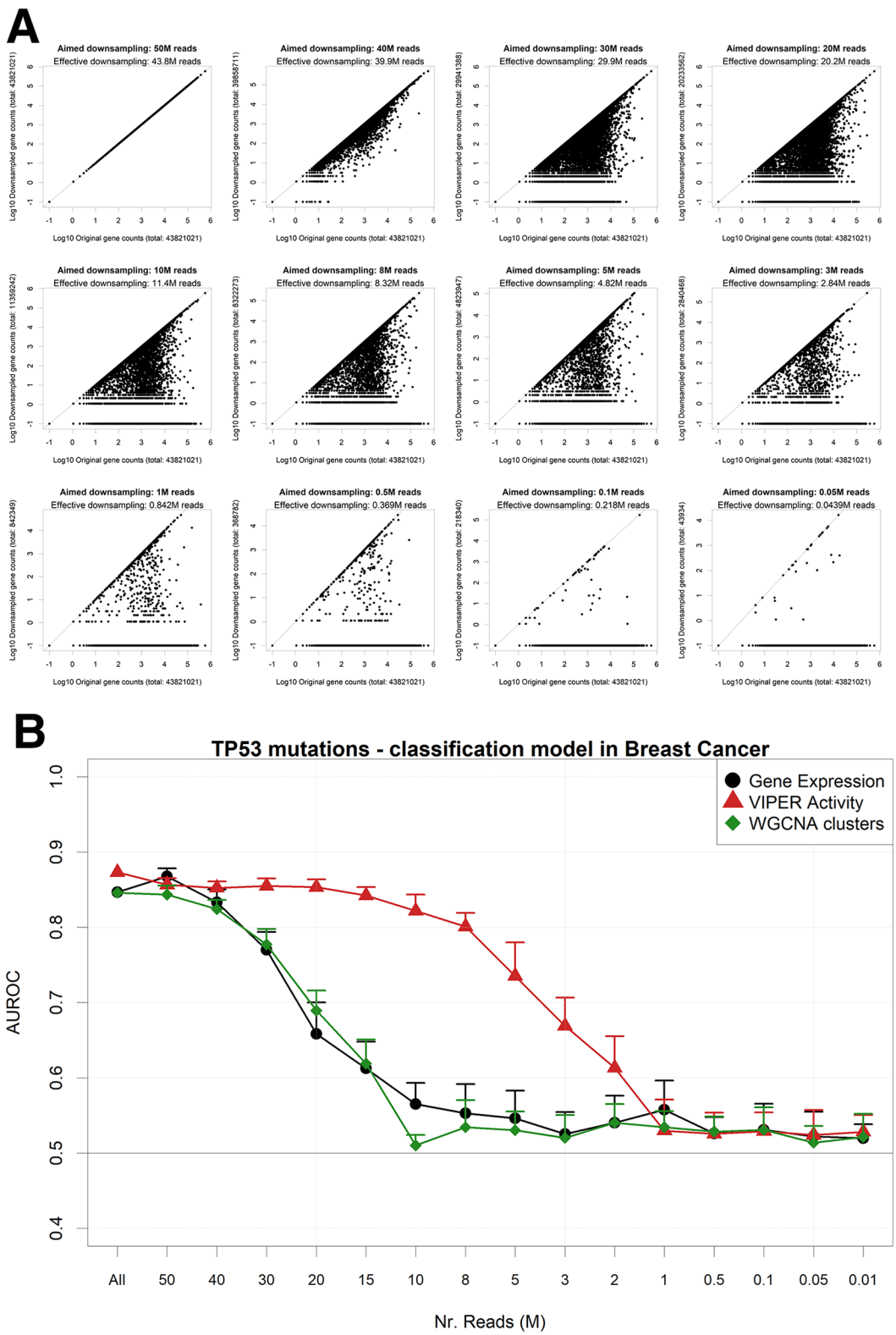


FIGURE 5 | Performance of a TP53 mutation gbm model upon down-sampling of the TCGA breast cancer RNA-Seq dataset. **(A)** for a single TCGA sample (TCGA-A1-A0SB-01) with 43.8 gene mapping reads, the down-sampling algorithm is applied for multiple target read quantities. X-axis shows the count for each gene in the original sample and Y-axis in the down-sampled output. **(B)** AUROCs of the model with decreasing read numbers, calculated using gene expression (black line) or aggregated gene expression using the WGCNA (green line) or VIPER (red line) algorithms. Error bars indicate the standard deviation of AUROC distribution. Pseudocounts of 0.1 are added in order to show zero counts as −1 in log10 scale.

achieve a significantly higher performance in both the Datlinger ($p = 8.0 \times 10^{-85}$) and Shifrut datasets ($p = 2.2 \times 10^{-117}$) when compared with models obtained from gene expression data (**Figure 6**). For specific mutations (TUBB gene, CDKN1B), the VIPER aggregation based on CCLE ARACNe networks seems to be particularly beneficial to increase the performance of mutation prediction models based on gene expression, while for a few mutations, such as RUNX1, the CCLE-based networks significantly decrease the model performance.

DISCUSSION

In this paper, we tested a framework to investigate the complex relationships between genetic events and transcriptional deregulation through machine learning approaches. We demonstrated as a generalized proof-of-principle that genomic alterations can be modeled by gene expression across several human cancers through several machine learning algorithms and, specifically, that a gbm approach seems optimal for the task. In the process, we generated a collection of models for each genomic alteration in each cancer context, showing that the best predicted alterations are not necessarily targeting known oncogenes or tumor suppressors. Interestingly, we show how the aggregation of gene expression profiles in groups of coexpressed genes, *via* the ARACNe/VIPER or WGCNA methods, makes the models more robust and more resistant to perturbations such as Gaussian noise or artificial down-sampling. Finally, we have shown how the same aggregation principle can have beneficial effects in predicting the presence of mutations in intrinsically noisy scenarios, both

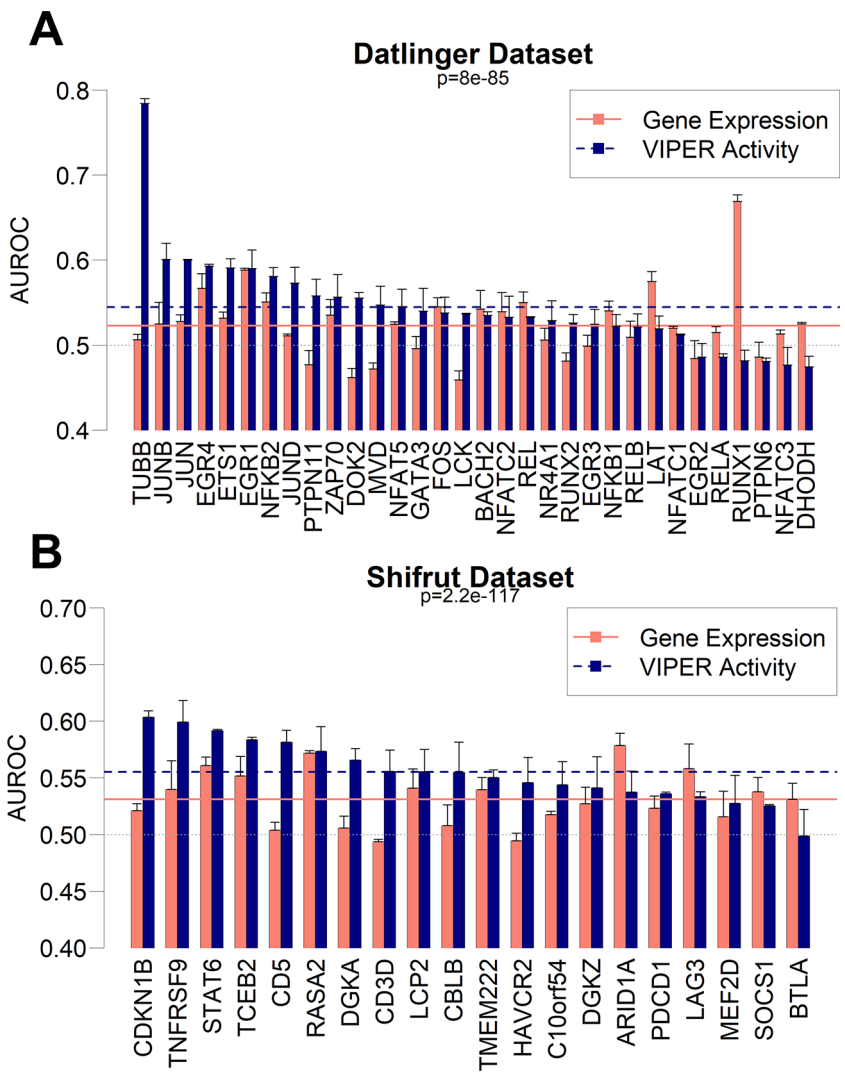


FIGURE 6 | Performance as AUROC of gbm models to predict mutations in CROP-Seq datasets using gene expression (red bars) and VIPER activity (blue bars) derived from CCLE expression data in Datlinger **(A)** and Shifrut **(B)** datasets. The p-value of paired Wilcoxon tests between all VIPER and expression AUROCs in each dataset is reported, as well as the average of all expression models (red solid line) and all VIPER activity models (blue dashed line). Error bars report the standard deviation of 100 AUROCs generated from multiple partitioning of training/test sets. Error bars indicate the standard deviation of AUROC distribution.

with artificial noise introduction and read reduction. At the same time, we have shown that expression-based mutation prediction can be modeled out in single-cell sequencing contexts, which can be considered as real cases of noisy datasets. The capability of predicting mutations based on scRNA-Seq is, however, reduced when compared with datasets derived from pooled cells sequencing, as those provided by the TCGA dataset: the average performances of TCGA models (**Figure 3**) generally rest on a range between 0.6 and 0.9 AUROC, while the performance of CROP-Seq models fall on an average value of 0.55 (**Figure 6**).

As transcriptional and signaling networks themselves gain diagnostic value, particularly for complex, multigenic diseases such as cancer (Alvarez et al., 2016), the network characteristics of coexpressed genes gain similar importance. A growing realization within the field of systems biology is that the activity and characteristic features of a given genomic network stem from the activity of smaller constituent subnetworks, and to this end, aggregated gene coexpression sets can constitute a novel and key focal point in network analysis overall (Wang et al., 2015).

The performance of gene aggregation methods has been tested before for sample clustering in RNA-Seq read reduction scenarios (Alvarez et al., 2016) but never in this specific task nor in a pan-cancer or a single-cell context. As a principle, the usage of robust averages of predefined coexpressed genes can be applied in any context where reliability of gene expression data is necessary, from differential expression to pathway enrichment analyses.

Using transcriptional networks with VIPER has been shown to be beneficial to increase the biological interpretability and reduce experimental noise in low-coverage sequencing setups such as the PLATE-Seq technique (Bush et al., 2017). We expect gene aggregation methods to further complement other RNA-seq noise reduction techniques (Ding et al., 2015), particularly those designed for scRNA-Seq data analysis. These include several recently published methods such as the deep count autoencoder (Eraslan et al., 2019), the factorial single-cell latent variable model (Buettner et al., 2017), the UnifiedRNA-Sequencing Model (Zhu et al., 2018), the single-cell Gene Expression Analysis app (Cai, 2019), the Ordering Effect gene Finder (Leng et al., 2016), and k-nearest neighbor smoothing (Wagner et al., 2017). Results obtained *via* computationally elegant techniques such as these stand to benefit from the inclusion of the types of network interaction features that we outlined previously.

Our analysis, while testing expression-based and network-based models for the entirety of frequent genomic alteration events in the TCGA dataset, is however limited to the presence/absence of single events considered separately. Patient tumor samples are often characterized by the co-occurrence of several mutations, CNVs, or a combination of those (Ciriello et al., 2013). In the future, generating models on a specific combination of genomic alterations will likely require larger clinical datasets, where each combination is represented in enough samples to allow for model training. This combinatorial approach for understanding the relationship between cancer genome and transcriptome will be beneficial in the context of personalized medicine, whereas every patient is considered separately (N-of-1 dataset), as it is characterized by a specific mutational landscape (Kristensen et al., 2014).

A recent study has shown, in agreement to our findings, that the highest part of cancer transcriptional variations are due to genomic alterations (copy number alterations and also somatic mutations) (Sharma et al., 2018) but also to epigenetic features and altered TF and μRNA balances. Those findings can explain why our results (**Figure 3**) highlight a highly variable performance depending on the modeled alterations and rare perfect models (max AUROCs rarely go above 0.9), while at the same time showing a generally better-than-random performance of expression-based prediction of genomic alterations (AUROC median and first quartiles >0.5). The notion that relationships between genomic alterations and gene expression profiles can be modeled across different cancer scenarios, as well as in single-cell and noisy contexts, may have important repercussions in diagnostics and quantification studies of heterogeneous cell populations, where theoretically a single quantitative expression experiment can be used to predict the presence or absence of a mutation.

MATERIALS AND METHODS

Data Processing

We obtained raw expression counts, mutation, and CNV raw data from TCGA using the Firehose portal (gdac.broadinstitute.org). Raw counts were normalized using variance stabilizing transformation as described before (Giorgi et al., 2013). Somatic mutations not changing the amino acid sequence of the protein product were discarded. We flagged genes blacklisted by the MutSig project (Lawrence et al., 2013), such as TTN, ORs, MUCs as false positives, and removed them from further analysis (except the most mutated in the pan-cancer dataset, shown in **Figure S1**). CNV tracks were associated to the targeted gene using the GenomicRanges R package (Lawrence et al., 2013). Gene-centered CNVs were then associated to the expression profile of the gene itself. Genes affected by a CNV in more than 10 samples were used in the rest of the analysis. Samples with more than 0.5% of the genes in the genome somatically amplified, deleted, or mutated were deemed "hypermodified," and the total number was shown in **Figure 1** bottom bars.

Clustering analysis was carried out on the TCGA tumor samples using the expression profiles of 1,172 TFs defined by gene ontology terms "transcription factor activity, sequence-specific DNA binding" (GO:0003700) and "nuclear location" (GO:0005634) (Ashburner et al., 2000).

The dataset expression profiles were visualized after TSNE transformation (van der Maaten and Hinton, 2008) with 1,000 iterations using a 2D kernel density estimate for coloring different tumor types (Duong, 2007). Oncogenes and tumor suppressor genes were obtained from the COSMIC Cancer Gene Census in October 2018 (Futreal et al., 2004).

Modeling

We used the R *caret* package (Kuhn, 2008) v 6.0-81 as the platform to run all our predictive models in a standardized and reproducible way. Default parameters for model training were used. Binary classifiers were built to predict the presence/absence

of mutation, amplification, and deletion events. The CNV value provided by TCGA corresponds to log2(tumor coverage) – genomic median coverage. The threshold for amplification/deletion presence was set to 0.5.

Data partitioning was performed once for each tumor type, with 75% of the samples used for training and 25% for test purposes. Training was performed using 10-fold cross-validation. Technical model robustness was assessed with a bootstrap approach as well (resampling of the patient samples with repetition). This was done in a smaller test scenario (bladder cancer mutation models) using the *caret* implementation of 100 bootstraps per mutation model (**Figure S10**). Bootstrap models have a slightly lower but not significantly different performance (AUROC Wilcoxon test $p = 0.121$) when compared with full dataset models. Recursive feature elimination was carried out by the default *caret* implementation on the 10,000 highest variance gene expression tracks. The algorithms used (and R packages implementing theme) were:

- Bayesian generalized linear model
- Tree models from genetic algorithms
- Gradient boost modeling (gbm)
- Generalized linear model
- k-nearest neighbors
- Linear discriminant analysis
- Neural networks
- Neural networks with feature extraction
- Random forest
- Linear support vector machine
- Radial support vector machine

In order to reduce information from the gene expression profiles, we adopted two strategies. The first, shown e.g. in **Figure 4B**, adds random Gaussian noise to the expression tracks, with a variable standard deviation (indicated as "Gaussian noise level"). Each model run after noise addition was run 100 times to allow for various data partitions. The second strategy (**Figure 5**) reduced the number of reads mapped to each gene in order to obtain expression samples with decreased total gene counts. In order to do so, we applied to each gene in each sample a down-sampling factor from a beta distribution:

$$\frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}$$

where B is the beta function, acting as a normalization constant, x is the raw gene expression count in a particular sample, α is the first shape parameter, and β the second shape parameter. In order to reduce the total sample coverage to the desired level, β is set to 0.1 and α is set to:

$$\alpha = \frac{\beta f / r}{1 - f / r}$$

where f is the desired number of reads and r is the total number of reads in the sample. A real case example of this beta distribution is shown in **Figure S11**.

Aggregation Algorithms

We used ARACNe-AP (Lachmann et al., 2016) to generate TF-centered networks on each of the VST-normalized TCGA expression datasets. TFs were selected *via* gene ontology as described before, with p-value for each network edge set to 10^{-8}. ARACNe networks were then used to obtain an aggregated value of TF activity for each sample using the VIPER algorithm (Alvarez et al., 2016) that reports the collective gene expression level changes of each TF-centered network vs. the mean expression of each gene in the dataset. Only TF networks with at least 10 genes (excluding the TF) were included.

WGCNA clusters of genes were constructed using the WGCNA package (Langfelder and Horvath, 2008) with default parameters and minimum network size set to 10. To obtain a robust median expression value for each WGCNA cluster in each sample, we used Tukey's biweight function as implemented by the R *affy* package (Gautier et al., 2004).

Single-Cell Analysis

We generated TF regulatory networks using ARACNe-AP as described before on the CCLE dataset available at https://portals.broadinstitute.org/ccle/data, raw counts version 2018-09-29, normalized by variance-stabilizing transformation (Pollen et al., 2014).

We downloaded raw RNA-Seq counts and guide RNA mutation data from single-cell CROP-Seq datasets, specifically: 1) the Datlinger dataset available on Gene Expression Omnibus (GEO) series GSE92872 (Datlinger et al., 2017), and 2) the Shifrut dataset was obtained from a healthy donor and is available as raw counts and cell-specific guide RNA from GEO sample GSM3375483 (Shifrut et al., 2018). Both single-cell CROP-Seq datasets were normalized using the R package Seurat with default parameters (Satija et al., 2015), as follows: a global-scaling normalization method ("LogNormalize") was applied on raw gene counts for each cell; then, the values were multiplied by a scale factor (10,000 by default), and the results were log-normalized. These values were then regressed by two variables: unique molecular identifier counts and cell cycle, using cell cycle markers from (Tirosh et al., 2016). As an example of the Seurat regression, the TSNE representation of the Datlinger dataset before and after normalization clearly shows the removal of cell cycle bias effects (**Figure S12**).

Gradient boost modeling (gbm) was applied to each CROP-Seq dataset by aggregating cells carrying mutations on the same genes and using wild-type cells as control. Performance of gbm models using VIPER and expression variables was compared using a two-tailed Wilcoxon test on 100 repetitions of training/test set splits before cross-validation for model testing (Hanley and McNeil, 1982).

Methods Availability

All code used to generate the analysis and the figures of this paper is available in the online materials as Supplementary Code.

DATA AVAILABILITY

Publicly available datasets were analyzed in this study. This data can be found here: https://gdac.broadinstitute.org/

AUTHOR CONTRIBUTIONS

FG conceived the analysis. FG, FR and DM designed the analysis. FG performed the analysis. FG wrote the manuscript. FR provided scientific support on the VIPER algorithm. DM contributed to the single-cell analysis.

ACKNOWLEDGMENTS

We acknowledge the CINECA award (projects HP10CB1R7T and HP10CPQJBV) under the ISCRA initiative, for the support and availability of high-performance computing resources. We also thank Lupo Giorgi, Marco Russo, Luca Pestarino, and Jordan Pflugh Kraft for the fruitful discussions. We further acknowledge the Rita Levi Montalcini Grant (Bando 2015) by the Italian Ministry of University and Research. The manuscript has been released as a pre-print at Mercatelli et al. (2019).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00671/full#supplementary-material.

FIGURE S1 | Table of most somatically mutated genes across TCGA tumor samples, in terms of number of samples where the gene is somatically mutated with altered protein product sequence. This table includes also MutSig-blacklisted genes (in gray) such as Titin (TTN), Obscurin (OBSCN), and Mucin genes.

FIGURE S2 | Performance of 11 machine learning algorithms in binary classification of mutated/nonmutated samples using gene expression predictor variables in the liver hepatocellular carcinoma dataset. Each point corresponds to a specific mutation/model. Performance is indicated as AUROC: area under the receiver operating characteristic curve.

FIGURE S3 | Relationship between alteration models and alteration frequency in the pan-cancer dataset, for mutations (left), amplifications (center), and deletions (right).

FIGURE S4 | Performance of pan-cancer alterations models globally (left) and for MutSig genes, COSMIC oncogenes, and COSMIC tumor suppressors. The y-axis indicates rank-transformed AUROC values. Asterisks indicate a significant (<0.01) difference between a distribution and the global "other genes" distribution according to two-tailed Wilcoxon tests.

FIGURE S5 | ROC curves for gbm TP53 models in breast cancer, using original expression data, VIPER aggregation (TF "activity"), and WGCNA aggregation (robust Tukey biweight average of clusters).

FIGURE S6 | AUROCs of EGFR amplification gbm prediction models in glioblastoma with increasing noise, calculated using gene expression (black line) or aggregated gene expression using the WGCNA (green line) or VIPER (red line) algorithms.

FIGURE S7 | AUROCs of EGFR amplification gbm prediction models in lung squamous carcinoma (LUSC) with increasing noise, calculated using gene expression (black line) or aggregated gene expression using the WGCNA (green line) or VIPER (red line) algorithms.

FIGURE S8 | AUROCs of PVT1 amplification gbm prediction models in ovarian cancer with increasing noise, calculated using gene expression (black line) or aggregated gene expression using the WGCNA (green line) or VIPER (red line) algorithms.

FIGURE S9 | AUROCs of PTEN deletion gbm prediction models in sarcoma with increasing noise, calculated using gene expression (black line) or aggregated gene expression using the WGCNA (green line) or VIPER (red line) algorithms.

FIGURE S10 | Distribution of gbm models AUROCs for predicting bladder cancer mutations. Left: original models shown in the main study (Figures 2 and 3). Right: performance of models with bootstrap. The p-value of a two-tailed Wilcoxon test between the two distributions is indicated.

FIGURE S11 | Beta distribution used to down-sample the 43.8M reads breast cancer sample TCGA-A1-A0SB-01 to 10M reads. The gray line shows the ratio between the target coverage and the original coverage.

FIGURE S12 | TSNE representation of the Datlinger CROP-Seq dataset before (A) and after (B) removal of cell cycle-specific markers. Colors indicated the predicted cell cycle phase according to the Seurat pipeline [79].

SUPPLEMENTARY TABLE S1 | AUROCs for each event in the pan-cancer TCGA dataset (24 tumor types with at least 100 samples with co-measured genomic and expression data. The sheet name indicates the tumor type and genomic alteration type (mut: somatic mutation, amp: amplification, del: deletion).

SUPPLEMENTARY CODE | R and bash code snippets used in this study.

REFERENCES

Alvarez, M. J., Shen, Y., Giorgi, F. M., Lachmann, A., Ding, B. B., Ye, B. H., et al. (2016). Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat. Genet.* 48, 838–847. doi: 10.1038/ng.3593

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556

Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., et al. (2018). Comprehensive characterization of cancer driver genes and mutations. *Cell* 174, 1034–1035. doi: 10.1016/j.cell.2018.07.034

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., et al. (2012). The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483, 603–607. doi: 10.1038/nature11003

Bass, A. J., Watanabe, H., Mermel, C. H., Yu, S., Perner, S., Verhaak, R. G., et al. (2009). SOX2 is an amplified lineage-survival oncogene in lung and esophageal squamous cell carcinomas. *Nat. Genet.* 41, 1238–1242. doi: 10.1038/ng.465

Bell, D. W., Lynch, T. J., Haserlat, S. M., Harris, P. L., Okimoto, R. A., Brannigan, B. W., et al. (2005). Epidermal growth factor receptor mutations and gene amplification in non–small-cell lung cancer: molecular analysis of the IDEAL/INTACT gefitinib trials. *J. Clin. Oncol.* 23, 8081–8092. doi: 10.1200/JCO.2005.02.7078

Bozic, I., Antal, T., Ohtsuki, H., Carter, H., Kim, D., Chen, S., et al. (2010). Accumulation of driver and passenger mutations during tumor progression. *Proc. Natl. Acad. Sci. U.S.A.* 107, 18545–18550. doi: 10.1073/pnas.1010978107

Brennan, C. W., Verhaak, R. G. W., McKenna, A., Campos, B., Noushmehr, H., Salama, S. R., et al. (2013). The somatic genomic landscape of glioblastoma. *Cell* 155, 462–477. doi: 10.1016/j.cell.2013.09.034

Buettner, F., Pratanwanich, N., McCarthy, D. J., Marioni, J. C., and Stegle, O. (2017). f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq. *Genome Biol.* 18, 212. doi: 10.1186/s13059-017-1334-8

Bush, E. C., Ray, F., Alvarez, M. J., Realubit, R., Li, H., Karan, C., et al. (2017). PLATE-Seq for genome-wide regulatory network analysis of high-throughput screens. *Nat. Commun.* 8. doi: 10.1038/s41467-017-00136-z

Cai, J. J. (2019). scGEApp: a Matlab app for feature selection on single-cell RNA sequencing data. *Bioinformatics*. doi: 10.1101/544163

Cancer Genome Atlas Network (2015). Genomic classification of cutaneous melanoma. *Cell* 161, 1681–1696. doi: 10.1016/j.cell.2015.05.044

Chae, Y. K., Anker, J. F., Carneiro, B. A., Chandra, S., Kaplan, J., Kalyan, A., et al. (2016). Genomic landscape of DNA repair genes in cancer. *Oncotarget* 7, 23312–23321. doi: 10.18632/oncotarget.8196

Chang, K., Creighton, C. J., Davis, C., Donehower, L., Drummond, J., Wheeler, D., et al. (2013). The cancer genome atlas Pan-Cancer analysis project. *Nature Genet.* 45, (10), 1113–1120. doi: 10.1038/ng.2764

Chen, J. C., Alvarez, M. J., Talos, F., Dhruv, H., Rieckhof, G. E., Iyer, A., et al. (2014). Identification of causal genetic drivers of human disease through systems-level analysis of regulatory networks. *Cell* 159, 402–414. doi: 10.1016/j.cell.2014.09.021

Chen, H.-I. H., Chiu, Y.-C., Zhang, T., Zhang, S., Huang, Y., and Chen, Y. (2018). GSAE: an autoencoder with embedded gene-set nodes for genomics functional characterization. *BMC Syst. Biol.* 12, 142. doi: 10.1186/s12918-018-0642-2

Cicenas, J., Kvederaviciute, K., Meskinyte, I., Meskinyte-Kausiliene, E., Skeberdyte, A., and Cicenas, J. KRAS (2017). TP53, CDKN2A, SMAD4, BRCA1, and BRCA2 mutations in pancreatic cancer. *Cancers* 9, 42. doi: 10.3390/cancers9050042

Ciriello, G., Miller, M. L., Aksoy, B. A., Senbabaoglu, Y., Schultz, N., and Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* 45, 1127–1133. doi: 10.1038/ng.2762

Cole, A. J., Dwight, T., Gill, A. J., Dickson, K.-A., Zhu, Y., Clarkson, A., et al. (2016). Assessing mutant p53 in primary high-grade serous ovarian cancer using immunohistochemistry and massively parallel sequencing. *Sci. Rep.* 6, 26191. doi: 10.1038/srep26191

Colombo, T., Farina, L., Macino, G., and Paci, P. (2015). PVT1: a rising star among oncogenic long noncoding RNAs. *BioMed Res. Int.* 2015, 304208. doi: 10.1155/2015/304208

Datlinger, P., Rendeiro, A. F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J., et al. (2017). Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* 14, 297–301. doi: 10.1038/nmeth.4177

Ding, B., Zheng, L., Zhu, Y., Li, N., Jia, H., Ai, R., et al. (2015). Normalization and noise reduction for single cell RNA-seq experiments. *Bioinformatics* 31, 2225–2227. doi: 10.1093/bioinformatics/btv122

Duong, T. (2007). ks: Kernel density estimation and kernel discriminant analysis for multivariate data in R. *J. Stat. Softw.* 021. doi: 10.18637/jss.v021.i07

Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S., and Theis, F. J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.* 10, 390. doi: 10.1038/s41467-018-07931-2

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognit. Lett.* 27, 861–874. doi: 10.1016/j.patrec.2005.10.010

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Ann. Stat.* 29, 1189–1232. doi: 10.1214/aos/1013203451

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., et al. (2004). A census of human cancer genes. *Nat. Rev. Cancer* 4, 177–183. doi: 10.1038/nrc1299

Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 20, 307–315. doi: 10.1093/bioinformatics/btg405

Ghazanfar, S., and Yang, J. Y. H. (2016). Characterizing mutation-expression network relationships in multiple cancers. *Comput. Biol. Chem.* 63, 73–82. doi: 10.1016/j.compbiolchem.2016.02.009

Giorgi, F. M., Del Fabbro, C., and Licausi, F. (2013). Comparative study of RNA-seq- and microarray-derived coexpression networks in Arabidopsis thaliana. *Bioinformatics* 29, 717–724. doi: 10.1093/bioinformatics/btt053

Guan, J., Gupta, R., and Filipp, F. V. (2015). Cancer systems biology of TCGA SKCM: efficient detection of genomic drivers in melanoma. *Sci. Rep.* 5, 7857. doi: 10.1038/srep07857

Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell* 144, 646–674. doi: 10.1016/j.cell.2011.02.013

Hanley, J. A., and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143, 29–36. doi: 10.1148/radiology.143.1.7063747

Huang, R. Y., Chen, G. B., Matsumura, N., Lai, H.-C., Mori, S., Li, J., et al. (2012). Histotype-specific copy-number alterations in ovarian cancer. *BMC Med. Genomics* 5, 47. doi: 10.1186/1755-8794-5-47

Irizarry, R. A., Wu, Z., and Jaffee, H. A. (2006). Comparison of Affymetrix GeneChip expression measures. *Bioinformatics* 22, 789–794. doi: 10.1093/bioinformatics/btk046

Kim, H., and Kim, Y.-M. (2018). Pan-cancer analysis of somatic mutations and transcriptomes reveals common functional gene clusters shared by multiple cancer types. *Sci. Rep.* 8, 6041. doi: 10.1038/s41598-018-24379-y

Kimura, E. T., Nikiforova, M. N., Zhu, Z., Knauf, J. A., Nikiforov, Y. E., and Fagin, J. A. (2003). High prevalence of BRAF mutations in thyroid cancer: genetic evidence for constitutive activation of the RET/PTC-RAS-BRAF signaling pathway in Papillary Thyroid Carcinoma. *Cancer Res.* 63, 1454–1457.

Kranenburg, O. (2005). The KRAS oncogene: past, present, and future. *Biochim. Biophys. Acta* 1756, 81–82. doi: 10.1016/j.bbcan.2005.10.001

Kristensen, V. N., Lingjærde, O. C., Russnes, H. G., Vollan, H. K. M., Frigessi, A., and Børresen-Dale, A.-L. (2014). Principles and methods of integrative genomic analyses in cancer. *Nat. Rev. Cancer* 14, 299–313. doi: 10.1038/nrc3721

Kuhn, M. (2008). Building predictive models in R using the caret package. *J. Stat. Softw.* 028. doi: 10.18637/jss.v028.i05

Lachmann, A., Giorgi, F. M., Lopez, G., and Califano, A. (2016). ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics* 32, 2233–2235. doi: 10.1093/bioinformatics/btw216

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9, 559. doi: 10.1186/1471-2105-9-559

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., et al. (2013). Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* 9, e1003118. doi: 10.1371/journal.pcbi.1003118

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499, 214–218. doi: 10.1038/nature12213

Leiserson, M. D. M., Vandin, F., Wu, H.-T., Dobson, J. R., Eldridge, J. V., Thomas, J. L., et al. (2015). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* 47, 106–114. doi: 10.1038/ng.3168

Leng, N., Choi, J., Chu, L.-F., Thomson, J. A., Kendziorski, C., and Stewart, R. (2016). OEFinder: a user interface to identify and visualize ordering effects in single-cell RNA-seq data. *Bioinformatics* 32, 1408–1410. doi: 10.1093/bioinformatics/btw004

Li, X., Chen, W., Wang, H., Wei, Q., Ding, X., and Li, W. (2017). Amplification and the clinical significance of circulating cell-free DNA of PVT1 in breast cancer. *Oncol. Rep.* 38, 465–471. doi: 10.3892/or.2017.5650

Liu, Y., and Zhao, M. (2016). lnCaNet: pan-cancer co-expression network for human lncRNA and cancer genes. *Bioinformatics* 32, 1595–1597. doi: 10.1093/bioinformatics/btw017

Meacham, C. E., and Morrison, S. J. (2013). Tumour heterogeneity and cancer cell plasticity. *Nature* 501, 328–337. doi: 10.1038/nature12624

Mercatelli, D., Ray, F., and Giorgi, F. M. (2019). Pan-Cancer and Single-Cell modelling of genomic alterations through gene expression. *BioRxiv*. doi: 10.1101/492561

Metzeler, K. H., Herold, T., Rothenberg-Thurley, M., Amler, S., Sauerland, M. C., Gorlich, D., et al. (2016). Spectrum and prognostic relevance of driver gene mutations in acute myeloid leukemia. *Blood* 128, 686–698. doi: 10.1182/blood-2016-01-693879

Mistry, M., Zhukova, N., Merico, D., Rakopoulos, P., Krishnatry, R., Shago, M., et al. (2015). BRAF mutation and CDKN2A deletion define a clinically distinct subgroup of childhood secondary high-grade glioma. *J. Clin. Oncol.* 33, 1015–1022. doi: 10.1200/JCO.2014.58.3922

Momand, J., Jung, D., Wilczynski, S., and Niland, J. (1998). The MDM2 gene amplification database. *Nucleic Acids Res.* 26, 3453–3459. doi: 10.1093/nar/26.15.3453

Nawy, T. (2013). Single-cell sequencing. *Nat. Methods* 11, 18. doi: 10.1038/nmeth.2771

Osmanbeyoglu, H. U., Toska, E., Chan, C., Baselga, J., and Leslie, C. S. (2017). Pancancer modelling predicts the context-specific impact of somatic mutations on transcriptional programs. *Nat. Commun.* 8, 14249. doi: 10.1038/ncomms14249

Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., et al. (2000). Molecular portraits of human breast tumours. *Nature* 406, 747–752. doi: 10.1038/35021093

Petitjean, A., Achatz, M. I. W., Borresen-Dale, A. L., Hainaut, P., and Olivier, M. (2007). TP53 mutations in human cancers: functional selection and impact on cancer prognosis and outcomes. *Oncogene* 26, 2157–2165. doi: 10.1038/sj.onc.1210302

Pollen, A. A., Nowakowski, T. J., Shuga, J., Wang, X., Leyrat, A. A., Lui, J. H., et al. (2014). Low-coverage single-cell mRNA sequencing reveals cellular

heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* 32, 1053–1058. doi: 10.1038/nbt.2967

Prior, I. A., Lewis, P. D., and Mattos, C. (2012). A comprehensive survey of Ras mutations in cancer. *Cancer Res.* 72, 2457–2467. doi: 10.1158/0008-5472.CAN-11-2612

Ran, F. A., Hsu, P. D., Wright, J., Agarwala, V., Scott, D. A., and Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* 8, 2281–2308. doi: 10.1038/nprot.2013.143

Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J.-P. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* 9. doi: 10.1038/s41467-017-02554-5

Robertson, A. G., Kim, J., Al-Ahmadie, H., Bellmunt, J., Guo, G., Cherniack, A. D., et al. (2017). Comprehensive molecular characterization of muscle-invasive bladder cancer. *Cell* 171, 540–556.e25. doi: 10.1016/j.cell.2017.09.007

Sager, R. (1989). Tumor suppressor genes: the puzzle and the promise. *Science* 246, 1406–1412. doi: 10.1126/science.2574499

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* 33, 495–502. doi: 10.1038/nbt.3192

Sharma, A., Jiang, C., and De, S. (2018). Dissecting the sources of gene expression variation in a pan-cancer analysis identifies novel regulatory mutations. *Nucleic Acids Res.* 46, 4370–4381. doi: 10.1093/nar/gky271

Shifrut, E., Carnevale, J., Tobin, V., Roth, T. L., Woo, J. M., Bui, C. T., et al. (2018). Genome-wide CRISPR screens in primary human T cells reveal key regulators of immune function. *Cell* 175, 1958–1971.e15. doi: 10.1016/j.cell.2018.10.024

Shlien, A., and Malkin, D. (2009). Copy number variations and cancer. *Genome Med.* 1, 62. doi: 10.1186/gm62

Smith, J. C., and Sheltzer, J. M. (2018). Systematic identification of mutations and copy number alterations associated with cancer patient prognosis. *ELife* 7, e39217. doi: 10.7554/eLife.39217

TCGA network. (2017). Integrated genomic characterization of oesophageal carcinoma. Nature https://www.nature.com/articles/nature20805

Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth, M. H., Treacy, D., Trombetta, J. J., et al. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* 352, 189–196. doi: 10.1126/science.aad0501

Tsuchida, N., Ryder, T., and Ohtsubo, E. (1982). Nucleotide sequence of the oncogene encoding the p21 transforming protein of Kirsten murine sarcoma virus. *Science* 217, 937–939. doi: 10.1126/science.6287573

Usvasalo, A., Savola, S., Räty, R., Vettenranta, K., Harila-Saari, A., Koistinen, P., et al. (2008). CDKN2A deletions in acute lymphoblastic leukemia of adolescents and young adults—An array CGH study. *Leuk. Res.* 32, 1228–1235. doi: 10.1016/j.leukres.2008.01.014

van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn Res.* 9, 2579–2605.

Venneti, S., and Huse, J. T. (2015). The evolving molecular genetics of low-grade glioma. *Adv. Anat. Pathol.* 22, 94–101. doi: 10.1097/PAP.0000000000000049

Wagner, P. L., Stiedl, A.-C., Wilbertz, T., Petersen, K., Scheble, V., Menon, R., et al. (2011). Frequency and clinicopathologic correlates of KRAS amplification in non-small cell lung carcinoma. *Lung Cancer* 74, 118–123. doi: 10.1016/j.lungcan.2011.01.029

Wagner, F., Yan, Y., and Yanai, I. (2017). K-nearest neighbor smoothing for high-throughput single-cell RNA-Seq data. *bioRxiv* 217737. doi: 10.1101/217737

Wang, E., Zaman, N., Mcgee, S., Milanese, J.-S., Masoudi-Nejad, A., and O'Connor-McCourt, M. (2015). Predictive genomics: a cancer hallmark network framework for predicting tumor clinical phenotypes using genome sequencing data. *Semin. Cancer Biol.* 30, 4–12. doi: 10.1016/j.semcancer.2014.04.002

Wang, X., Cao, X., Sun, R., Tang, C., Tzankov, A., Zhang, J., et al. (2018). Clinical significance of PTEN deletion, mutation, and loss of PTEN expression in de novo diffuse large B-cell lymphoma. *Neoplasia* 20, 574–593. doi: 10.1016/j.neo.2018.03.002

Yan, H., Parsons, D. W., Jin, G., McLendon, R., Rasheed, B. A., Yuan, W., et al. (2009). IDH1 and IDH2 mutations in gliomas. *N. Engl. J. Med.* 360, 765–773. doi: 10.1056/NEJMoa0808710

Zhao, D., Lu, X., Wang, G., Lan, Z., Liao, W., Li, J., et al. (2017). Synthetic essentiality of chromatin remodelling factor CHD1 in PTEN-deficient cancer. *Nature* 542, 484–488. doi: 10.1038/nature21357

Zhu, L., Lei, J., Devlin, B., and Roeder, K. (2018). A unified statistical framework for single cell and bulk RNA sequencing data. *Ann. Appl. Stat.* 12, 609–632. doi: 10.1214/17-AOAS1110

Conflict of Interest Statement: 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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 13 August 2019
doi: 10.3389/fgene.2019.00738

Primary Tumor Site Specificity is Preserved in Patient-Derived Tumor Xenograft Models

Lei Chen[1,2,3†], Xiaoyong Pan[4†], Yu-Hang Zhang[1], Xiaohua Hu[5], KaiYan Feng[6], Tao Huang[1] and Yu-Dong Cai[7*]*

[1] Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China, [2] College of Information Engineering, Shanghai Maritime University, Shanghai, China, [3] Shanghai Key Laboratory of PMMP, East China Normal University, Shanghai, China, [4] Department of Medical Informatics, Erasmus Medical Center, Rotterdam, Netherlands, [5] Department of Biostatistics and Computational Biology, School of Life Sciences, Fudan University, Shanghai, China, [6] Department of Computer Science, Guangdong AIB Polytechnic, Guangzhou, China, [7] School of Life Sciences, Shanghai University, Shanghai, China

OPEN ACCESS

Edited by:
Yifei Xu,
University of Oxford, United Kingdom

Reviewed by:
Quan Zou,
University of Electronic Science and Technology of China, China
Guang Wu,
Guangxi Academy of Sciences, China

***Correspondence:**
Tao Huang
tohuangtao@126.com
Yu-Dong Cai
cai_yud@126.com

†These authors have contributed equally to this work.

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *13 April 2019*
Accepted: *15 July 2019*
Published: *13 August 2019*

Citation:
Chen L, Pan X, Zhang Y-H, Hu X, Feng K, Huang T and Cai Y-D (2019) Primary Tumor Site Specificity is Preserved in Patient-Derived Tumor Xenograft Models. Front. Genet. 10:738. doi: 10.3389/fgene.2019.00738

Patient-derived tumor xenograft (PDX) mouse models are widely used for drug screening. The underlying assumption is that PDX tissue is very similar with the original patient tissue, and it has the same response to the drug treatment. To investigate whether the primary tumor site information is well preserved in PDX, we analyzed the gene expression profiles of PDX mouse models originated from different tissues, including breast, kidney, large intestine, lung, ovary, pancreas, skin, and soft tissues. The popular Monte Carlo feature selection method was employed to analyze the expression profile, yielding a feature list. From this list, incremental feature selection and support vector machine (SVM) were adopted to extract distinctively expressed genes in PDXs from different primary tumor sites and build an optimal SVM classifier. In addition, we also set up a group of quantitative rules to identify primary tumor sites. A total of 755 genes were extracted by the feature selection procedures, on which the SVM classifier can provide a high performance with MCC 0.986 on classifying primary tumor sites originated from different tissues. Furthermore, we obtained 16 classification rules, which gave a lower accuracy but clear classification procedures. Such results validated that the primary tumor site specificity was well preserved in PDX as the PDXs from different primary tumor sites were still very different and these PDX differences were similar with the differences observed in patients with tumor. For example, *VIM* and *ABHD17C* were highly expressed in the PDX from breast tissue and also highly expressed in breast cancer patients.

Keywords: Patient-derived tumor xenograft, gene expression profile, Monte Carlo feature selection, support vector machine, rule learning algorithm

INTRODUCTION

Patient-derived tumor xenograft (PDX) mouse models, developed by implanting patients' *in vivo* tumor tissues into immune-deficient mice (Harris et al., 2016), are widely used in tumor biology and drug screening. Compared with cancer cell lines, PDX mouse models can maintain the original tumor development conditions immensely with appropriate tumor microenvironment that mimics similar regulatory factors, which are identified in the primary tumor site *in vivo* (Coats et al., 2017).

Furthermore, with the development of humanized-xenograft models, PDX-humanized mouse models compensate for one of the prominent shortcomings of traditional PDX mouse models—the absence of immune regulation and selection—thereby accomplishing the accurate simulation on tumorigenesis *in vivo* (Jung et al., 2018).

As the PDX mouse model has more advantages in the oncology research field compared with traditional routines, various typical PDX mouse models have been successfully set up with their respective tumor tissues. Early in 2011, *Nature Medicine* published a systematic analysis (DeRose et al., 2011) on the pathological and biological characteristics of tumor tissues implanted into an immune-deficient mouse model as PDX. Such study confirmed that the PDX mouse model can basically reflect the same pathological processes during the initiation and progression of breast cancer, validating the significance of such model in the field of tumor research. Furthermore, PDX mouse models have been applied to various tumor subtypes, including colorectal cancer, pancreatic cancer, and pediatric cancer (Scott et al., 2017). Studies on such tumor subtypes have also confirmed that tumor tissues developed in a PDX mouse model have quite similar pathological and biological characteristics with tumor tissues in situ, though without immune selective pressure. Overall, PDX mouse models have been accepted as one of the most significant methods for tumor research.

In the field of oncology research, wide attention has been paid to gene expression characterizations. Different tumors have different expression pattern of functional tumor-associated genes as tumor-specific expression profile. Given the distinctive microenvironment and environmental selection pressure of human bodies and immune-deficient mice, the expression profile of a PDX mouse model has been confirmed to be different from the expression spectrum of tumor *in situ* (Ben-David et al., 2017). As mentioned above, different tumor subtypes have different tumor-specific expression profiles *in vivo*. However, after the selection and passaging in the mouse microenvironment, it is quite reasonable to speculate that tumor tissues of different subtypes may be differentially selected and lose/gain various differentially expressed genes (DEGs), thus generating a novel tumor subtype-specific expression profile (Ben-David et al., 2017). Although various studies have attempted to identify tumor subtype-specific biomarkers based on the expression profile of tumor tissues in PDX mouse models for years, no direct evidence or studies have revealed whether tumor tissues from different primary tumor subtypes can maintain tumor-specific DEGs during the passaging of PDX mouse models. Moreover, it is not clear whether such identified tumor-specific DEGs are all derived from the primary tumor tissues or from murine microenvironment selection.

To solve the problem, the most convenient way is to explore whether DEGs identified in PDX tumor tissues can still distinguish different tumor subtypes as potential biomarkers. Herein, we selected eight tumor subtypes originating from different tissues, including breast, kidney, large intestine, lung, ovary, pancreas, skin, and soft tissues, for the identification of DEGs in the PDX mouse model based on a study (Gao et al., 2015) on PDX tumor expression profile. Several advanced computational methods were used in this study, including the Monte Carlo feature selection (MCFS) (Draminski et al., 2008), incremental feature selection (IFS) (Liu and Setiono, 1998), and support vector machine (SVM) (Cortes and Vapnik, 1995). As a result, a group of highly related genes was identified, which may be distinctively expressed in different tumor subtypes as PDX tumor tissue. Furthermore, several quantitative rules were set up for the identification of different xenograft tumor subtypes by a specific set of functional distinctive genes. The results reported in this study further validated that PDX mouse models may be a relatively effective and practical mouse model in the field of tumor studies and may be favorable to be applied to indicate DEGs from primary tumor tissues between different tumor subtypes.

MATERIALS AND METHODS

Dataset

We downloaded the expression data of 20,502 genes in eight PDX tumor tissues: (1) kidney, (2) skin, (3) ovary, (4) soft tissue, (5) breast, (6) pancreas, (7) lung, and (8) large intestine. The number of samples in each tissue is shown in **Table 1**. A total of 594 samples were considered in this study. The high-throughput screening data using PDX were obtained from the Gene Expression Omnibus (GEO) with accession number GSE78806 (Gao et al., 2015). To investigate whether the primary site of tumor has great influences on PDX, we compared the gene expression profiles of PDX from different primary sites.

Feature Selection

Many genes are specifically expressed in the tissues; that is, some genes are closely related to certain tissues. To identify highly related genes for different tissues, we first used the MCFS (Draminski et al., 2008) method to analyze the expression data of 20,502 genes, obtaining a feature list and several classification rules. Then, the two-stage IFS (Liu and Setiono, 1998) method was applied to yield optimum features (genes), wherein the SVM (Cortes and Vapnik, 1995) exhibited a strong discriminative power for samples from different tissues.

Monte Carlo Feature Selection

MCFS (Draminski et al., 2008) is a type of feature selection method. As mentioned in the section *Dataset*, 594 samples were

TABLE 1 | Number of samples for each of the eight tissues.

Tissue	Number of samples
Breast	79
Kidney	41
Large intestine	121
Lung	99
Ovary	52
Pancreas	94
Skin	46
Soft tissue	62
Total	594

investigated in this study, and each sample was represented by 20,502 features. Thus, the dataset we studied is a high-dimensional dataset. The MCFS method is ideal in dealing with this type of dataset (Draminski et al., 2008). To date, this method has been applied to deal with several biological problems (Cai et al., 2018; Chen et al., 2018a; Chen et al., 2018c; Pan et al., 2018). In this study, it was also adopted to analyze all features and rank them for supervised classifiers.

MCFS constructs decision tree classifiers for many bootstrap sets that are randomly selected from the original sample set, and each tree is grown from a randomly selected feature subset with m features of original M features, where m is much less than M. During the process, p decision trees are generated on a training set randomly selected from a bootstrapping dataset and a feature subset. The above process is repeated t times to obtain t feature subsets. In total, $p \times t$ decision trees can be constructed.

The relative importance (RI) indicates the importance of each feature, which mainly considers the number of times that the feature is involved in growing the $p \times t$ decision trees. The RI score of a feature g can be calculated using the following formula:

$$RI(g) = \sum_{\tau=1}^{pt} (wAcc)^u IG(n_g(\tau)) \left(\frac{no.in\ n_g(\tau)}{no.in\ \tau}\right)^v, \quad (1)$$

where $wAcc$ is the weighted accuracy across all classes, $n_g(\tau)$ indicates a node using feature g in decision tree τ, $IG(n_g(\tau))$ is the information gain of $n_g(\tau)$, $no.in\ \tau$ is the number of training samples in τ, and $no.in\ n_g(\tau)$ is the number of samples in node $n_g(\tau)$. u and v are two weighting factors, and we used their default setting of $u = v = 1$.

A feature assigning a high MI value means that it is quite important. To extract most important features, the MCFS method adopts a permutation test on class labels. In detail, in a round of permutation test, a permutation of class labels is assigned to samples and the MCFS method is executed on the dataset with new labels, producing a maximal RI value. After several rounds, many maximal RI values are generated. The threshold, indicating high significance level of features, is determined by the one-sided Student's t test. Features receiving the RI value larger than such threshold are selected and termed as informative features. These features are deemed to be essential for the investigated dataset. For a detailed description, please refer to Dramiński et al. (2011).

The informative features are extracted according to the essential properties of the dataset. However, for a given classifier, these features are not always optimal. Thus, we further ranked all features in a list according to their MI values in a way that features with high MI values receive high ranks in the list, whereas those with low MI values are placed at the bottom of the list. Here, we formulated the obtained feature list yielded by MCFS method as

$$F = [f_1, f_2, \ldots, f_N], \quad (2)$$

where N is the total number of features (N = 20,502 in this study). This list was used in the IFS method to select optimal features for a given classifier.

In this study, the program of the MCFS method was retrieved from http://www.ipipan.eu/staff/m.draminski/mcfs.html.

Rule Learning

Aside from analyzing features and ranking them in a list, the program of the MCFS method also integrates a rough set-based rule learning procedure. Based on informative features, the Johnson reducer algorithm (Ohrn, 1999) was used to select some important features that can give competitive classification performance compared with all informative features. After that, Repeated Incremental Pruning to Produce Error Reduction (RIPPER) algorithm (Cohen, 1995) produced the rules with the above-selected features. Each of these rules describes a relation between conditions (the left-hand side of the rule) and the outcome (the right-hand side). For example, a rule can be presented as an IF–THEN relationship based on expression values: IF Gene1 ≥ 6.4 AND Gene2 ≥ 4.8 THEN subtype = "kidney." Following these rules, all samples can be easily classified. In addition, compared with black-box machine learning methods, the classification rules can provide a clearer classification procedure and help in understanding the expression differences among different tissues.

Incremental Feature Selection

The MCFS method only analyzes the importance of each feature and ranks them in a feature list. For a classification problem, it is necessary to extract some optimal features to comprise the feature subspace. Meanwhile, different classifiers require different optimal features. In view of this, the IFS (Liu and Setiono, 1998) method was employed in this study. The IFS method always integrates a supervised classifier to screen optimal features for accurately classifying samples from different groups. In the original IFS method, it first constructs a series of feature subsets according to a feature list in a way that the latter subset is produced by adding one feature to the former one. Then, for each feature subset, the supervised classifier is executed on the dataset, in which samples are represented by features in the subset. Finally, the feature subset yielding the best performance is selected as the optimal feature set. However, this procedure is time-consuming, especially when the number of features is quite large. Accordingly, we adopted a two-stage IFS method to approximately complete the procedure of finding optimal feature set in this study, which are described below.

In the first stage, several feature subsets with a large step (e.g., 10) were constructed. In detail, we constructed the feature subsets, denoted as $F_1^1, F_2^1, \ldots, F_m^1$ where $m = [N/10]$ and $F_i^1 = \{f_1, f_2, \ldots, f_{10\times i}\}$, that is, the ith feature subset contains the top $10 \times i$ features in F. Then, for each of these feature subsets, the selected classifier was trained and evaluated on the samples that were represented by features in this set using 10-fold cross-validation (Kohavi, 1995; Chen et al., 2018b; Chen et al., 2018d; Guo et al., 2018; Pan et al., 2018; Wang et al., 2018; Zhao et al., 2018; Zhao et al., 2019). According to the results of these feature subsets, a feature number interval [min, max], on which the classifier provided satisfied the prediction performance, can be obtained. The size of the optimal feature set was in this interval with a high probability. In the second stage, based on the above feature number interval [min, max], another series of feature

subsets was produced, denoted as $F^2_{\min}, F^2_{\min+1}, \ldots, F^2_{\max}$, in which the latter subset contains one more feature than the former one. Similarly, the classifier was trained and evaluated on these subsets, like the first stage. We can obtain a feature subset with the best performance. For convenience, features in this set were still called optimal features, whereas the corresponding classifier was termed as the optimal classifier.

SVM

As mentioned in the section *Incremental Feature Selection*, the IFS method required a supervised classifier. Here, we selected the classic classifier, SVM (Cortes and Vapnik, 1995). The SVM is a popular supervised learning method that distinguishes samples based on a set of features, and it is widely used to deal with many biological problems (Pan and Shen, 2009; Chen et al., 2017b; Cui and Chen, 2019). The basic principle is to infer a hyperplane with maximum margin between two classes of samples. In reality, most of the data are non-linear in low-dimensional space. In this case, all samples are mapped to a high-dimensional space using kernel function, such as Gaussian kernel. In this space, a linear function can be found to perfectly separate samples of two classes. The original SVM is mainly developed for binary classification. For multi-class classification, the "One Versus the Rest" strategy is adopted. In detail, it constructs m binary SVM classifiers for m classes, where each classifier is trained to separate samples in one class from the rest using the samples of that class as positive samples and other samples as negative ones. For an unseen sample, m probability scores can be yielded by m SVM classifiers, and the label with the highest probability score is assigned to the unseen sample.

Performance Measurement

For a classification problem with multiple classes, the basic measurement is the individual accuracy for each class, which is defined as

$$ACC_i = \frac{M_i}{N_i} \tag{3}$$

where ACC_i represents the individual accuracy of the ith class, M_i represents the number of correctly predicted samples in the ith class, and N_i represents the total number of samples in the ith class. Furthermore, the overall accuracy can completely evaluate the prediction performance, which is formulated by

$$ACC = \frac{\sum_{i=1}^{8} M_i}{\sum_{i=1}^{8} N_i} \tag{4}$$

Although the overall accuracy can completely evaluate the prediction quality, it is not a fair measurement when the class sizes are of great difference. According to **Table 1**, the biggest class ("Large intestine") is about three times as many as the smallest class ("Skin"). In this case, the overall accuracy was not a good choice to assess the prediction quality. Thus, we further employed Matthew's correlation coefficient (MCC) in multi-class (Gorodkin, 2004). It is a generalization version of MCC proposed by Matthew (Matthews, 1975; Chen et al., 2017a; Zhao et al., 2018; Zhao et al., 2019). It is known that the classic MCC is a balanced measurement even if the class sizes vary greatly. The MCC in multi-class keeps such merit. Suppose we have n samples (i = 1, 2,…, n) and C classes (j = 1, 2,…, C). Let $X = (x_{ij})_{n\times C}$ be the predicted classes of samples and $x_{ij} \in \{0,1\}$ be a binary value. x_{ij} is equal to 1 if the sample i is predicted to belong to class j; otherwise, the value x_{ij} is 0. The matrix $Y = (y_{ij})_{n\times C}$ is defined as the true classes of samples, where the binary variable $y_{ij} = 1$ means that the sample i belongs to class j; otherwise, it is set to 0.

According to matrices X and Y, the MCC can be defined as follows:

$$MCC = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{cov}(X,X)\mathrm{cov}(Y,Y)}} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{C}(x_{ij}-\bar{x}_j)(y_{ij}-\bar{y}_j)}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{C}(x_{ij}-\bar{x}_j)^2 \sum_{i=1}^{n}\sum_{j=1}^{C}(y_{ij}-\bar{y}_j)^2}}, \tag{5}$$

where $\bar{x}_j$ and $\bar{y}_j$ are the mean values of members in the j-th column of X and j-th column of Y, respectively.

RESULTS

In this study, a computational investigation on the gene expression data of samples in eight PDX tumor tissues was performed. The entire procedure is illustrated in **Figure 1**.

Results of MCFS Method

To evaluate the investigated features mentioned in the section *Dataset* on discriminating samples from different tissues, the MCFS method was used to analyze and rank them in descending order according to their RI values. The obtained feature list is provided in **Supplementary Table 1**.

Furthermore, the MCFS method produced 530 informative features by determining the threshold of RI value as 0.0155. Based on these features, the Johnson reducer and RIPPER algorithms can generate some classification rules. To evaluate the performance of the rules yielded by these two algorithms, 10-fold cross-validation was performed thrice. The confusion map for such test to classify samples into eight tissues is shown in **Figure 2**. The MCC was 0.794. The individual accuracies for eight tissues and overall accuracy are shown in **Figure 3**. It can be seen that the performance of the rules yielded by Johnson reducer and RIPPER algorithms was acceptable. Thus, we further used Johnson reducer and RIPPER algorithms to generate 16 classification rules with 530 informative features based on all samples, which are listed in **Table 2**. The performance of these rules was evaluated by self-consistency; i.e., these rules were

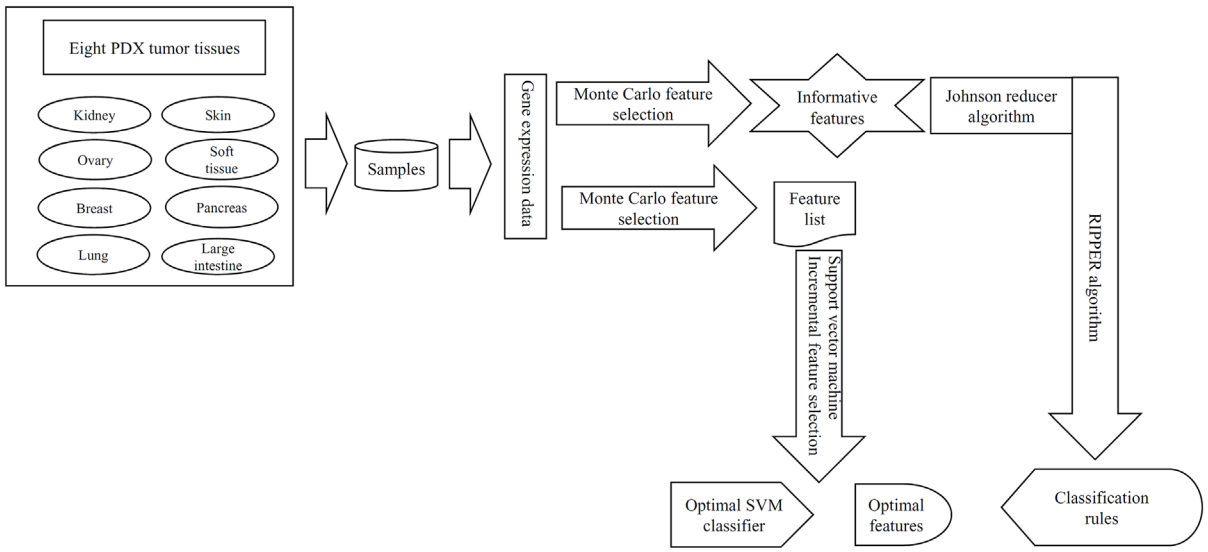


FIGURE 1 | The entire procedures to investigate the gene expression data of samples in eight PDX tumor tissues. These data were first analyzed by the Monte Carlo feature selection method, producing a feature list and informative features. The feature list was used in the incremental feature selection method to extract optimal features for support vector machine (SVM) and construct the optimal SVM classifier. For informative features, the Johnson reducer and Repeated Incremental Pruning to Produce Error Reduction (RIPPER) algorithms were applied on them to generate classification rules.

FIGURE 2 | Confusion map for classifying samples into eight tissues *via* the classification rules yielded by Johnson reducer and Repeated Incremental Pruning to Produce Error Reduction (RIPPER) algorithms, evaluated by 10-fold cross-validation thrice.

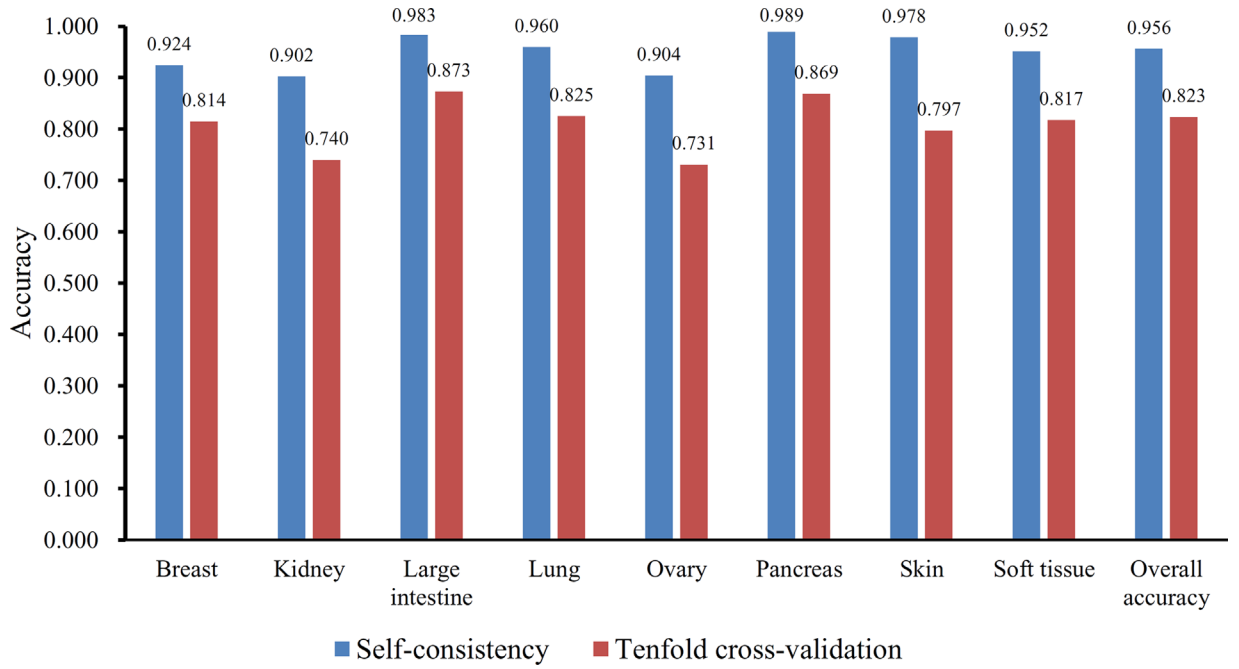


FIGURE 3 | The individual and overall accuracies of the classification rules yielded by Johnson reducer and Repeated Incremental Pruning to Produce Error Reduction (RIPPER) algorithms, evaluated by self-consistency and 10-fold cross-validation.

TABLE 2 | Sixteen produced classification rules for distinguishing samples from different tissues.

Rules	Criteria	Tissues
Rule-1	ANGPTL4 ≥ 6.409 BHMT2 ≥ 4.826	Kidney
Rule-2	UPK1A ≥ 6.474	Kidney
Rule-3	PAX3 ≥ 3.401 MIA ≥ 3.562	Skin
Rule-4	BHMT2 ≥ 5.125 ANXA10 ≥ 3.820	Skin
Rule-5	PAX8 ≥ 3.217 ADAM10 ≥ 5.994	Ovary
Rule-6	TRADD ≤ 3.210 ASRGL1 ≥ 6.703	Ovary
Rule-7	CPVL ≥ 7.240 CDX1 ≤ 2.111	Ovary
Rule-8	F11R ≤ 4.935 VSNL1 ≤ 4.528	Soft tissue
Rule-9	HSD17B11≤5.122 ITGA2 ≤ 6.021	Breast
Rule-10	VIM ≥ 8.697 ABHD17C ≥ 3.622	Breast
Rule-11	ADAM28 ≥ 3.637 BTBD6 ≤ 7.581	Pancreas
Rule-12	CXCL5 ≥3.927 PCDH1 ≥ 4.141	Pancreas
Rule-13	LOC102724689 ≥ 7.396	Pancreas
Rule-14	MSN ≥ 5.037 PDGFC ≥ 1.903 BCL2L15 ≤ 5.317	Lung
Rule-15	TP73-AS1 ≥ 3.462 ADAM10 ≥ 6.134	Lung
Rule-16	Other conditions	Large intestine

applied to samples to make classification. We obtained the MCC of 0.949. The individual and overall accuracies are illustrated in **Figure 3**. It can be observed that the predicted results yielded by self-consistency were much better than those of 10-fold cross-validation. It is reasonable because in self-consistency, samples were classified by the rules generated by themselves.

Results of the IFS Method

Based on the Johnson reducer and RIPPER algorithms, classification rules were generated. However, their performance was not very high. Thus, we further applied SVMs to classify samples from different tissues by integrating the selected features from two-stage IFS method. In the first stage, the feature sets containing multiples of 10 features were constructed, and the SVM was trained on the dataset, in which samples were represented by features in these sets. The 10-fold cross-validation was adopted to evaluate the performance of SVM. The predicted results were counted as individual accuracy for each tissue, overall accuracy, and MCC described in the section *Performance Measurement*, which are provided in **Supplementary Table 2**. For easy observation of the performance of SVM under different feature sets, a curve was plotted in **Figure 4A**, in which the number of used features was termed as *X*-axis and MCC as the *Y*-axis. The curve first follows a sharp increasing trend and eventually becomes stable. To clearly illustrate the increasing trend at the beginning of this curve, we plotted the part of the curve between *X*-axis 10 and 2000 in **Figure 4B**. The highest MCC is 0.986 when the top 780 features were used. Around 780, the MCCs were also very high. Thus, we determined the

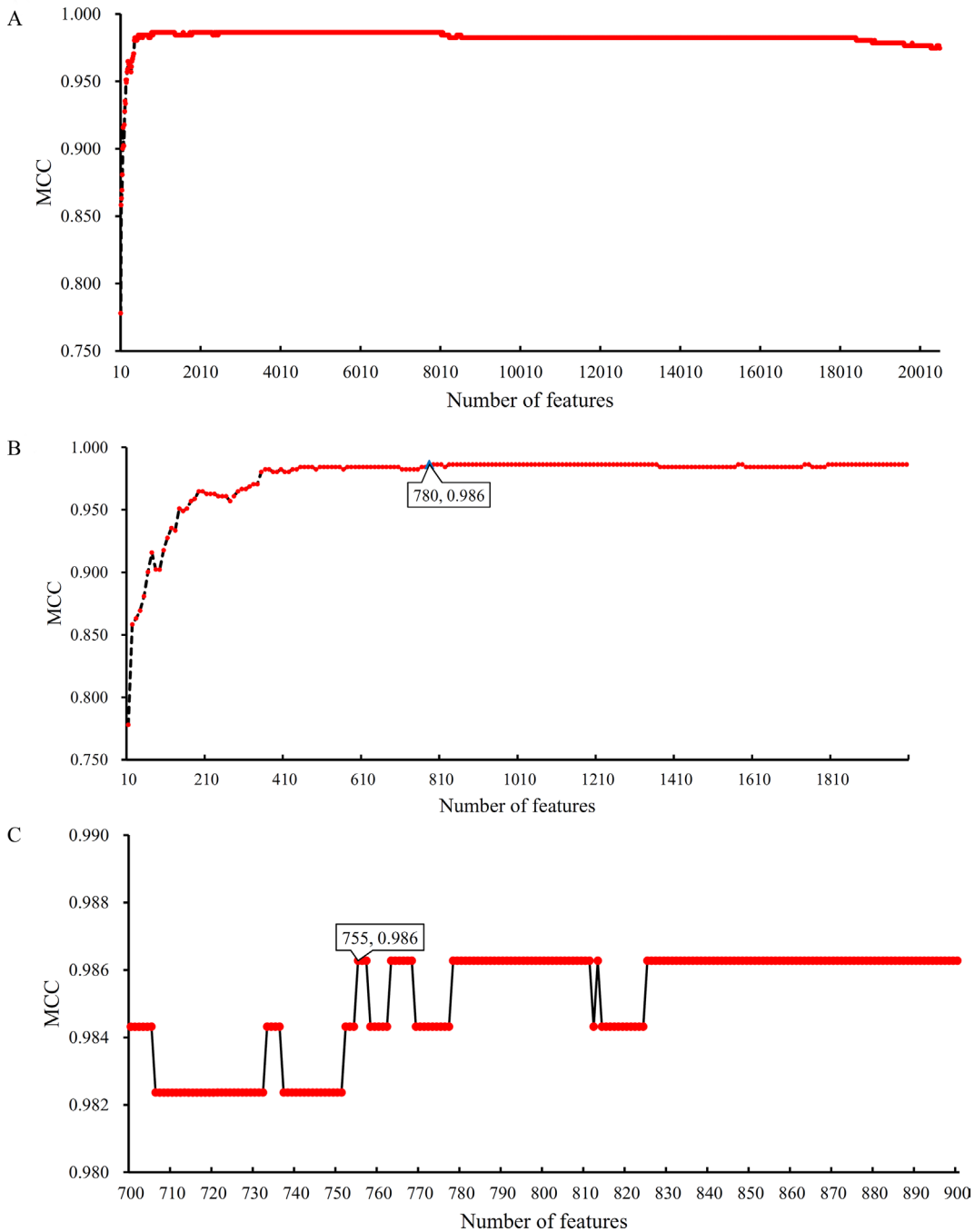


FIGURE 4 | Curves illustrating the performance of SVM on different feature sets. The *X*-axis represents the number of features participating in the classification; the *Y*-axis represents the MCC. **(A)** The whole curve illustrating the performance of SVM on feature sets containing multiples of 10 top features. **(B)** Part of the curve between *X*-axis 10 and 2000. When the top 780 features are used, the MCC reaches the highest (0.986). **(C)** The curve illustrating the performance of SVM on feature sets containing 700–900 top features. When the top 755 features are used, the MCC reaches the highest (0.986).

feature number interval as [700, 900]. The second stage of the IFS method constructed a second set of feature subsets with a step 1 within feature number interval [700, 900]; that is, all feature sets containing 700–900 features were constructed. SVM and 10-fold cross-validation were adopted to test the discriminating ability of each feature set. The obtained measurements, including individual accuracy for each tissue, overall accuracy, and MCC, are listed in **Supplementary Table 3**. Similarly, we also plotted a curve, as shown in **Figure 4C**. The highest MCC is still 0.986; however, it can be achieved only by using the top 755 features. Therefore, these 755 features were termed as optimal features, and the SVM classifier based on these features was the optimal

SVM classifier. The detailed performance of such optimal classifier is illustrated in **Figure 5**, from which we can see that all samples in pancreas and skin were correctly classified, and most samples in other tissues were also predicted correctly, indicating the effectiveness of this classifier.

Superiority of the Optimal Features

The optimal SVM classifier adopted 755 features to represent samples. To further indicate the importance of these features, we randomly produced 1000 feature subsets, each of which contained 755 features. For each subset, an SVM classifier was constructed, and we evaluated its performance *via* 10-fold cross-validation. The obtained 1000 MCCs are illustrated in **Figure 6** (black circles), in which the MCC yielded by the optimal SVM classifier is also listed (red circle). It can be observed that the MCC yielded by the optimal SVM classifier was higher than all other MCCs. In addition, it was also higher than the threshold of high significance level (p value < 0.05), indicating that these 755 features were significant.

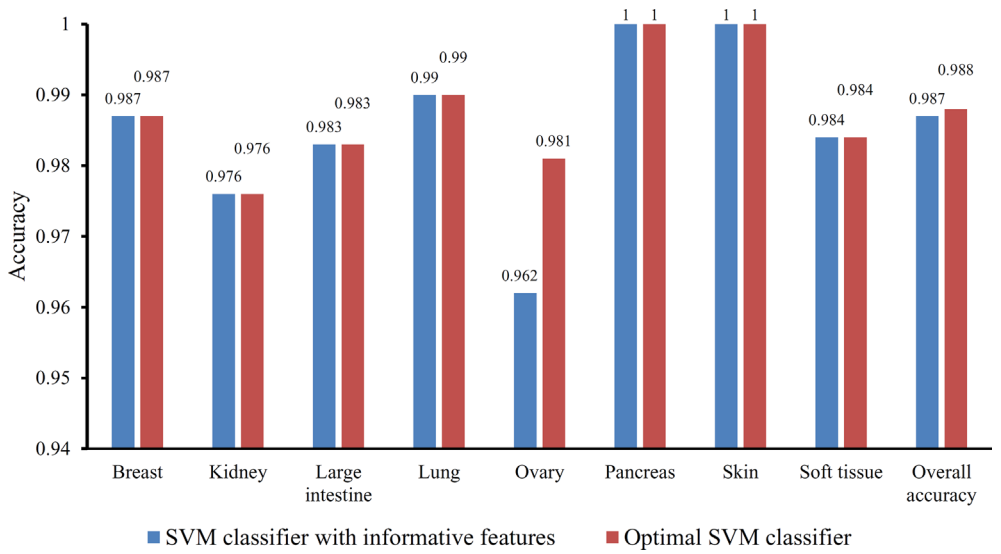


FIGURE 5 | Bar chart illustrating the individual accuracy on each tissue and overall accuracy yielded by the optimal SVM classifier and the classifier with informative features.

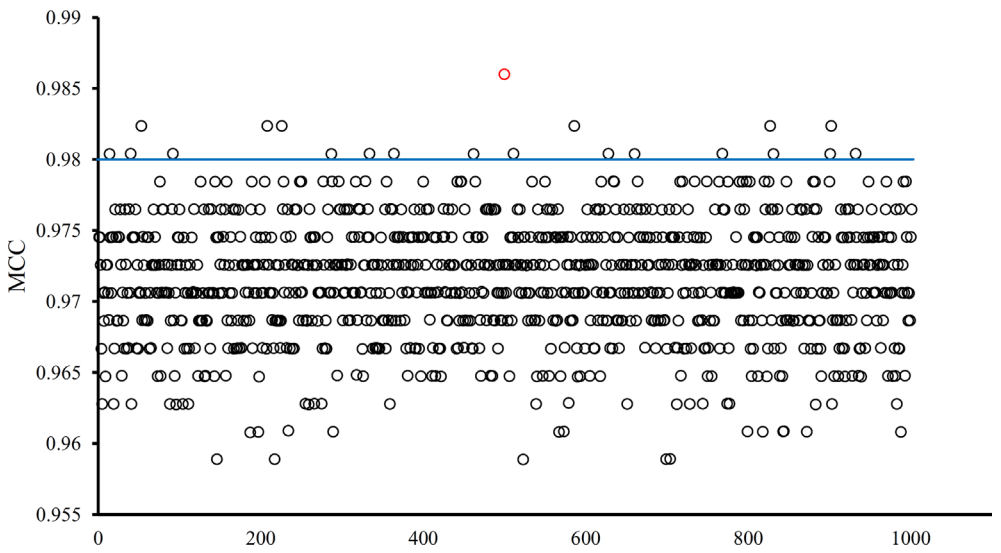


FIGURE 6 | MCCs obtained by the optimal SVM classifier and 1000 SVM classifiers on 1000 randomly generated feature subsets. The red circle represents the MCC yielded by the optimal SVM classifier and black circles represent MCCs produced by SVM classifiers on randomly generated feature subsets. The blue line represents the threshold of high significance level (p value < 0.05).

Besides, the MCFS method can produce informative features for each given dataset. For our dataset, 530 informative features were obtained. An SVM classifier can be constructed on these features. Such classifier was also evaluated by 10-fold cross-validation. The MCC was 0.984, which was lower than that of the optimal SVM classifier (0.986). The individual accuracies for eight tissues and overall accuracy are illustrated in **Figure 5**, from which we can see that each measurement was no higher than that of the optimal SVM classifier. It is implied that the optimal SVM classifier was superior to the classifier with informative features. The IFS method is useful to extract optimal features for a given classifier.

DISCUSSION

Based on a new study (Gao et al., 2015) on the expression profile of various tumor subtypes in PDX models, we deeply analyzed this profile for the accurate identification of eight different candidate tumor subtypes using several advanced computational methods in the present study. On the one hand, a list of effective genes that may directly contribute to the qualitative distinction of different tumor subtypes was screened out. On the other hand, we also identified a group of quantitative rules for the accurate identification of each tumor subtypes. This section provides an extensive analysis on the extracted genes and quantitative rules *via* literature reviewing.

Analysis of Optimal Features (Genes)

For constructing an optimal SVM classifier, the top 755 features (genes) were used to represent samples. However, analyzing them individually is challenging. By carefully checking the performance of SVM classifiers in the first stage of the IFS method, we found that the MCC achieved 0.980 when the top 350 features were used. Thus, we believed that these 350 features were more important than the other 405 features. However, it is still impossible to analyze these 350 features one by one. Here, we selected the most important genes, that is, the top 10 genes, listed in **Table 3**, to provide an extensive analysis.

The top gene is ***IFFO1***, which may have a unique expression pattern in eight tumor tissues. *IFFO1*, encoding a primordial component of the cytoskeleton and nuclear envelope, has been detected with specific methylation patterns and expression profiles in the PDX mouse model of lung cancer (Anglim et al., 2008) and ovarian cancer (Houshdaran et al., 2010), but not in other tumor tissues, indicating that the specific expression pattern of this gene may be a potential biomarker for identifying lung cancer and ovarian cancer.

TABLE 3 | Top 10 features (genes) yielded by the MCFS method.

Rank	Gene symbol	Description	RI
1	IFFO1	Intermediate Filament Family Orphan 1	0.4515
2	CDX1	Caudal Type Homeobox 1	0.4263
3	HSD17B11	Hydroxysteroid 17-Beta Dehydrogenase 11	0.4047
4	CHMP4C	Charged Multivesicular Body Protein 4C	0.4042
5	CLIP4	CAP-Gly Domain Containing Linker Protein Family Member 4	0.4025
6	PAX8	Paired Box 8	0.4024
7	GUCY2C	Guanylate Cyclase 2C	0.4023
8	MLANA	Melan-A	0.3857
9	F11R	F11 Receptor	0.3689
10	NR3C1	Nuclear Receptor Subfamily 3 Group C Member 1	0.3646

The gene ***CDX1*** has also been predicted to contribute to distinguishing different PDX tumor tissues at the expression level. With relatively high expression level in small intestine and colon tissues, *CDX1* plays a role in the differentiation of the intestine (Jones et al., 2015). As for its expression in different PDX tumor tissues, this gene has relatively high expression in large intestine-associated tumor tissues of PDX mouse model, confirming the potential distinguishing effect of such gene (Rankin et al., 2004).

HSD17B11, encoding short-chain alcohol dehydrogenases, has been widely reported to participate in androgen metabolism during steroidogenesis (Rotinen et al., 2011). As for its contribution on tumorigenesis and specific role during PDX implantation, this gene has only been identified in both primary and implanted tumor tissue of the prostate (Hilborn et al., 2017) and breast tumorigenesis (Rotinen et al., 2011), implying that such gene may distinguish different tumor tissues.

CHMP4C is reported to be involved in multi-vesicular body formation and endosomal cargo sorting (Yu et al., 2009). As for its specific expression pattern in different tumor tissues, this gene has a unique pathological expression profile in multiple tumors of the urine system, implying that CHMP4C may be an effective marker for identifying kidney-associated tumor from other tumor subtypes derived from other tissues (Fujita et al., 2017).

CLIP4, encoding one of the components of the cytoplasmic linker protein family, participates in regulating the cellular compartmentalization of the AKT kinase family involved in tumorigenesis (Saber et al., 2016). Such gene has been confirmed to have a unique expression pattern in various tumor PDX mouse models, including clear cell renal cell carcinomas (kidney) (Ahn et al., 2016), lung adenocarcinoma (lung) (Saber et al., 2016), and gastric cancer (stomach) (Chong et al., 2014), implying that this gene may be a biomarker for some tumor subtypes investigated in this study.

PAX8, encoding a transcription factor of the paired box (PAX) family, has been predicted to be a potential identification marker for the distinction of different tumor tissues in PDX mouse models (Narumi et al., 2010). Recent studies (Butler et al., 2017) confirmed that the overexpression of such gene may directly induce the initiation and progression of ovarian cancer in PDX mouse models, distinguishing tumorigenesis of such tissue from the other seven tumor tissues.

GUCY2C, encoding a membrane-associated guanylate kinase, participates in immune regulation, including T-cell receptor-mediated T-cell activation and proliferation (Snook et al., 2012). As for its tissue-specific distribution in the PDX mouse model, recent studies (Witek et al., 2014) confirmed that in the large intestine (especially colon tissue), the high expression level of such gene in the PDX model indicates that such mouse model was implanted with an invasive large intestine-associated tumor subtype.

The next gene ***MLANA*** encodes a GPR143-associated functional protein contributing to the maintenance of expression, stability, trafficking, and processing of melanocyte protein PMEL

(Witek et al., 2014). As for its relationship with different tumor tissues in the PDX mouse model, a recent study (Hollingshead et al., 2014) confirmed that such gene may distinguish melanoma and various skin-derived tumor subtypes in the PDX mouse model from the other seven tumor subtypes.

F11R, as a regulator of cell-to-cell adhesion in epithelial cell sheets, has been reported to encode a multi-functional protein that interacts with reovirus (Birse et al., 2017), integrin LFA1 (Gerhardt and Ley, 2015), and platelets (Kedees et al., 2005). As for its distinctive function for different PDX tumor tissues, recent studies (Jansen et al., 2009) confirmed that in the PDX models of glioblastoma (soft-tissue-derived tumorigenesis), F11R has a unique expression pattern compared with other tumor tissues.

NR3C1, encoding a tissue-specific transcriptional activator, has been widely reported to be involved in chromatin remodeling (Geng et al., 2016) and cell proliferation in tissues *in situ* (Souza et al., 2014). As for its distinctive expression pattern in different tumor tissues, such gene has a relatively high expression pattern in various tumor subtypes, including lung cancer (Lajoie et al., 2014) and kidney cancer (Zaravinos et al., 2014), compared with other tumor subtypes.

Overall, based on advanced computational methods, we screened out a group of effective tumor-associated genes that may distinguish different tumor subtypes from PDX mouse models. From the discussions on the top 10 genes, we confirmed that other optimal features (genes) may also be important biomarkers for distinguishing different tumor subtypes that need further investigation.

Analysis of Classification Rules

Apart from qualitative biomarkers to distinguish different tumor subtypes in the PDX mouse model, we also summarized 16 classification rules for further quantitative analysis. To show the inner relationship between genes involved in these rules, we draw a rule network *via* Ciruvis (Bornelov et al., 2014), which is illustrated in **Figure 7**. Based on the detailed expression

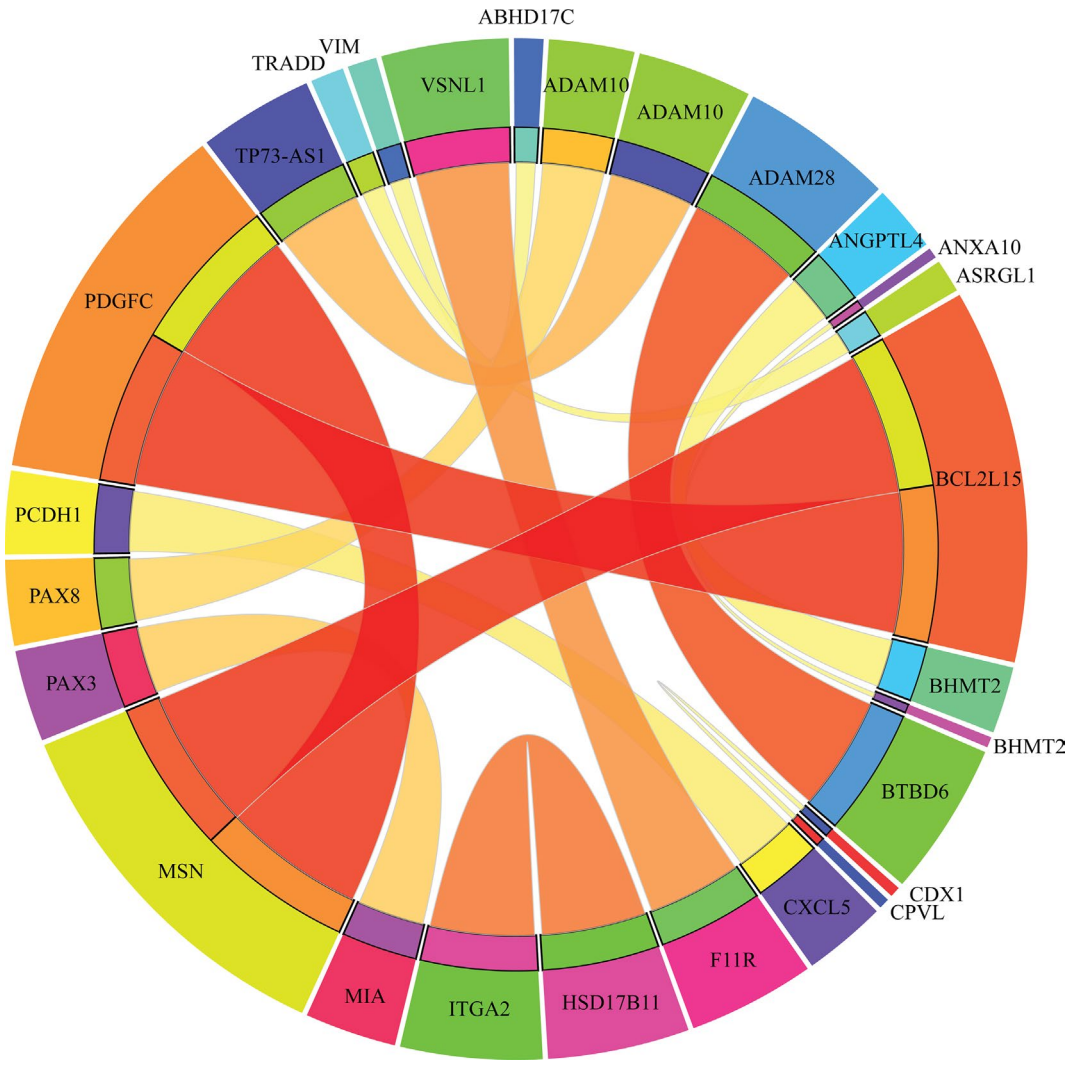


FIGURE 7 | Rule networks for 16 classification rules generated by Ciruvis.

profile data in other similar studies, most of the 16 rules can be confirmed by their rationalities, reflecting the relative expression pattern of such genes involving the rules. The detailed analysis on each rule is shown below.

The first two rules are for the identification of PDX tumor tissues originating from kidney-associated tumor. According to these two quantitative rules, *ANGPTL4* should have higher expression pattern and the expression level of *BHMT2* and *UPK1A* should also be up-regulated. According to recent single-cell RNA sequencing data of the PDX mouse model (Zhu et al., 2017), the expression patterns of the three genes have all been confirmed to have corresponding expression level.

The following two rules are for the identification of skin-derived PDX tumor tissues. Four genes named *PAX3*, *MIA*, *BHMT2*, and *ANXA10* have been screened out as potential parameters for the identification of skin-associated PDX tumors. Based on recent sequencing publications, all four genes have been reported to be upregulated, conforming to these rules (Tso et al., 2014). The combination of such four parameters may improve the efficacy and accuracy for the quantitative identification of skin-derived tumor-implanted PDX mouse model. As for the detailed FPKM value, the dataset provided by similar studies (Wyatt et al., 2014) also corresponds with our rules.

The next three rules describe the expression pattern of ovarian cancer. As we have analyzed above, *PAX8*, encoding a functional transcription factor, has a uniquely high expression pattern in ovarian-cancer-derived PDX tumor tissues, corresponding with Rule-5 (Narumi et al., 2010). As for the other five parameters, a recent study (Dobbin et al., 2014) revealed the specific expression pattern of ovarian cancer after screening the PDX mouse microenvironment. According to recent literature, although the expression profile of *CDX1* (as one of the parameters mentioned above) cannot indicate ovarian cancer alone, the combination of *CDX1* and *CPVL* may be specifically enough to recognize ovarian-tumor-derived PDX mouse tumor tissues (Dobbin et al., 2014). According to the dataset provided by such study, the remaining four parameters (*ADAM10*, *TRADD*, *ASRGL1*, and *CPVL*) have also been validated to basically match our rules.

Only one rule involving two genes may contribute to the identification of soft-tissue-derived PDX tumor tissues. *F11R*, as we have analyzed above, has been confirmed to have a relatively low expression pattern in the PDX tumor tissue derived from soft tissue, which is somewhat different from those derived from other tissues, validating the accuracy and efficacy of this rule (Jansen et al., 2009). A similar expression pattern has also been identified for the remaining soft-tissue-specific expressing gene *VSNL1* (Sarver et al., 2015), corresponding with this rule.

The following two rules contribute to the identification of breast cancer in the PDX mouse model. Four genes, namely, *HSD17B11*, *ITGA2*, *VIM*, and *ABHD17C*, are involved in these rules. The low expression of *HSD17B11* and *ITGA2* and the high expression of *VIM* and *ABHD17C* have all been validated by recent sequencing studies on breast cancer (Rotinen et al., 2011), reflecting the accuracy of these two rules.

The expression levels of five genes (*ADAM28*, *BTBD6*, *CXCL5*, *PCDH1*, and *LOC102724689*) comprise three rules for the identification of pancreatic-tissue-derived PDX tumor tissues. According to another dataset (Martinez-Garcia et al., 2014), the quantitative parameter of such five genes have been basically validated. Among such five genes, *PCDH1* is the most effective tumor-associated gene, contributing to pancreatic cancer with abnormal promoter methylation status and participating in FGFR-associated signaling pathways (Zhang et al., 2014).

The two remaining rules contribute to the identification of lung-tissue-derived PDX tumor tissues. Five genes, namely, *MSN*, *PDGFC*, *BCL2L15*, *TP73-AS1*, and *ADAM10*, were screened out as candidate parameters. Various studies have revealed the expression pattern of lung cancer in PDX mouse model at either the single cell or bullet level (Bradford et al., 2016). By comprehensively analyzing such expression profiles of the five candidate genes, the expression levels of such five genes in lung-cancer-derived PDX tumor tissues correspond to the quantitative rules. Furthermore, if the expression profile of a certain PDX tumor tissue does not satisfy any of the conditions we mentioned above, such PDX tumor tissue may be derived from the large intestine.

Overall, we quantitatively analyzed the 16 rules reported in this study. Several rules can be supported or validated by recent RNA sequencing datasets on PDX tumor tissues, validating the efficacy and accuracy of these rules. Combining the qualitative analysis presented in the section *Analysis of Optimal Features (Genes)*, we not only identified a group of highly related PDX tumor-specific biomarkers at the expression spectrum level but also for the first time attempted to build a systematic distinctive standard for the quantitative identification of PDX tumor originating from different tissue subtypes. The genes and rules that we screened out not only can provide a new tool for the identification of PDX-derived tumors originating from different primary tissues but also reveal the distinctive expression characteristics and expression profile stability of PDX-derived tumor tissues compared with the primary ones, validating the efficacy and practicability of the PDX mouse model in tumor studies.

DATA AVAILABILITY

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE78806

AUTHOR CONTRIBUTIONS

All authors contributed to the research and reviewed the manuscript. TH and YDC designed the study. LC, XP, and KYF performed the experiments. YHZ and XH analyzed the results. LC and XP wrote the manuscript.

FUNDING

This study was funded by the National Natural Science Foundation of China (31701151), the Natural Science Foundation of Shanghai (17ZR1412500), the National Key R&D Program of China (2018YFC0910403), the Shanghai Sailing Program

(16YF1413800), the Youth Innovation Promotion Association of Chinese Academy of Sciences (CAS) (2016245), the fund of the key Laboratory of Stem Cell Biology of Chinese Academy of Sciences (201703), and the Science and Technology Commission of Shanghai Municipality (STCSM) (18dz2271000).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00738/full#supplementary-material

REFERENCES

Ahn, J., Han, K. S., Heo, J. H., Bang, D., Kang, Y. H., Jin, H. A., et al. (2016). FOXC2 and CLIP4: a potential biomarker for synchronous metastasis of </ = 7-cm clear cell renal cell carcinomas. *Oncotarget* 7 (32), 51423–51434. doi: 10.18632/oncotarget.9842

Anglim, P. P., Galler, J. S., Koss, M. N., Hagen, J. A., Turla, S., Campan, M., et al. (2008). Identification of a panel of sensitive and specific DNA methylation markers for squamous cell lung cancer. *Mol. Cancer* 7, 62. doi: 10.1186/1476-4598-7-62

Ben-David, U., Ha, G., Tseng, Y. Y., Greenwald, N. F., Oh, C., Shih, J., et al. (2017). Patient-derived xenografts undergo mouse-specific tumor evolution. *Nat. Genet.* 49 (11), 1567–1575. doi: 10.1038/ng.3967

Birse, K. D., Romas, L. M., Guthrie, B. L., Nilsson, P., Bosire, R., Kiarie, J., et al. (2017). Genital injury signatures and microbiome alterations associated with depot medroxyprogesterone acetate usage and intravaginal drying practices. *J. Infect. Dis.* 215 (4), 590–598. doi: 10.1093/infdis/jiw590

Bornelov, S., Marillet, S., and Komorowski, J. (2014). Ciruvis: a web-based tool for rule networks and interaction detection using rule-based classifiers. *BMC Bioinformatics* 15, 139. doi: 10.1186/1471-2105-15-139

Bradford, J. R., Wappett, M., Beran, G., Logie, A., Delpuech, O., Brown, H., et al. (2016). Whole transcriptome profiling of patient-derived xenograft models as a tool to identify both tumor and stromal specific biomarkers. *Oncotarget* 7 (15), 20773–20787. doi: 10.18632/oncotarget.8014

Butler, K. A., Hou, X., Becker, M. A., Zanfagnin, V., Enderica-Gonzalez, S., Visscher, D., et al. (2017). Prevention of human lymphoproliferative tumor formation in ovarian cancer patient-derived xenografts. *Neoplasia* 19 (8), 628–636. doi: 10.1016/j.neo.2017.04.007

Cai, Y.-D., Zhang, S., Zhang, Y.-H., Pan, X., Feng, K., Chen, L., et al. (2018). Identification of the gene expression rules that define the subtypes in glioma. *J. Clin. Med.* 7 (10), 350. doi: 10.3390/jcm7100350

Chen, L., Chu, C., Zhang, Y.-H., Zheng, M.-Y., Zhu, L., Kong, X., et al. (2017a). Identification of drug–drug interactions using chemical interactions. *Curr. Bioinform.* 12 (6), 526–534. doi: 10.2174/1574893611666160618094219

Chen, L., Li, J., Zhang, Y. H., Feng, K., Wang, S., Zhang, Y., et al. (2018a). Identification of gene expression signatures across different types of neural stem cells with the Monte-Carlo feature selection method. *J. Cell. Biochem.* 119 (4), 3394–3403. doi: 10.1002/jcb.26507

Chen, L., Pan, X., Hu, X., Zhang, Y.-H., Wang, S., Huang, T., et al. (2018b). Gene expression differences among different MSI statuses in colorectal cancer. *Int. J. Cancer* 143 (7), 1731–1740. doi: 10.1002/ijc.31554

Chen, L., Wang, S., Zhang, Y.-H., Li, J., Xing, Z.-H., Yang, J., et al. (2017b). Identify key sequence features to improve CRISPR sgRNA efficacy. *IEEE Access* 5, 26582–26590. doi: 10.1109/ACCESS.2017.2775703

Chen, L., Zhang, S., Pan, X., Hu, X., Zhang, Y.-H., Yuan, F., et al. (2018c). HIV infection alters the human epigenetic landscape. *Gene Ther.* 26, 29–39. doi: 10.1038/s41434-018-0051-6

Chen, L., Zhang, Y.-H., Pan, X., Liu, M., Wang, S., Huang, T., et al. (2018d). Tissue expression difference between mRNAs and lncRNAs. *Int. J. Mol. Sci.* 19 (11), 3416. doi: 10.3390/ijms19113416

Chong, Y., Mia-Jan, K., Ryu, H., Abdul-Ghafar, J., Munkhdelger, J., Lkhagvadorj, S., et al. (2014). DNA methylation status of a distinctively different subset of genes is associated with each histologic Lauren classification subtype in early gastric carcinogenesis. *Oncol. Rep.* 31 (6), 2535–2544. doi: 10.3892/or.2014.3133

Coats, J. S., Baez, I., Stoian, C., Milford, T. M., Zhang, X., Francis, O. L., et al. (2017). Expression of exogenous cytokine in patient-derived xenografts *via* injection with a cytokine-transduced stromal cell line. *J. Vis. Exp.* (123), e55384. doi: 10.3791/55384

Cohen, W. W. (1995). "Fast effective rule induction," in *The twelfth international conference on machine learning*, (Tahoe City, CA, USA: Elsevier), 115–123. doi: 10.1016/B978-1-55860-377-6.50023-2

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20 (3), 273–297. doi: 10.1007/BF00994018

Cui, H., and Chen, L. (2019). A binary classifier for the prediction of EC numbers of enzymes. *Curr. Proteomics* 16 (5), 381–389. doi: 10.2174/1570164616666190126103036

DeRose, Y. S., Wang, G., Lin, Y. C., Bernard, P. S., Buys, S. S., Ebbert, M. T., et al. (2011). Tumor grafts derived from women with breast cancer authentically reflect tumor pathology, growth, metastasis and disease outcomes. *Nat. Med.* 17 (11), 1514–1520. doi: 10.1038/nm.2454

Dobbin, Z. C., Katre, A. A., Steg, A. D., Erickson, B. K., Shah, M. M., Alvarez, R. D., et al. (2014). Using heterogeneity of the patient-derived xenograft model to identify the chemoresistant population in ovarian cancer. *Oncotarget* 5 (18), 8750–8764. doi: 10.18632/oncotarget.2373

Dramiński, M., Kierczak, M., Nowak-Brzezińska, A., Koronecki, J., and Komorowski, J. (2011). The Monte Carlo feature selection and interdependency discovery is unbiased. *Control and Cybernetics*, 40(2), 199–211.

Draminski, M., Rada-Iglesias, A., Enroth, S., Wadelius, C., Koronacki, J., and Komorowski, J. (2008). Monte Carlo feature selection for supervised classification. *Bioinformatics* 24 (1), 110–117. doi: 10.1093/bioinformatics/btm486

Fujita, K., Kume, H., Matsuzaki, K., Kawashima, A., Ujike, T., Nagahara, A., et al. (2017). Proteomic analysis of urinary extracellular vesicles from high Gleason score prostate cancer. *Sci. Rep.* 7, 42961. doi: 10.1038/srep42961

Gao, H., Korn, J. M., Ferretti, S., Monahan, J. E., Wang, Y., Singh, M., et al. (2015). High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nat. Med.* 21 (11), 1318–1325. doi: 10.1038/nm.3954

Geng, L., Zhu, M., Wang, Y., Cheng, Y., Liu, J., Shen, W., et al. (2016). Genetic variants in chromatin-remodeling pathway associated with lung cancer risk in a Chinese population. *Gene* 587 (2), 178–182. doi: 10.1016/j.gene.2016.05.013

Gerhardt, T., and Ley, K. (2015). Monocyte trafficking across the vessel wall. *Cardiovasc. Res.* 107 (3), 321–330. doi: 10.1093/cvr/cvv147

Gorodkin, J. (2004). Comparing two K-category assignments by a K-category correlation coefficient. *Comput. Biol. Chem.* 28 (5), 367–374. doi: 10.1016/j.compbiolchem.2004.09.006

Guo, Z.-H., Chen, L., and Zhao, X. (2018). A network integration method for deciphering the types of metabolic pathway of chemicals with heterogeneous information. *Comb. Chem. High Throughput Screen.* 21 (9), 670–680. doi: 10.2174/1386207322666181206112641

Harris, A. L., Joseph, R. W., and Copland, J. A. (2016). Patient-derived tumor xenograft models for melanoma drug discovery. *Expert. Opin. Drug Discov.* 11 (9), 895–906. doi: 10.1080/17460441.2016.1216968

Hilborn, E., Stal, O., Alexeyenko, A., and Jansson, A. (2017). The regulation of hydroxysteroid 17beta-dehydrogenase type 1 and 2 gene expression in breast cancer cell lines by estradiol, dihydrotestosterone, microRNAs, and genes related to breast cancer. *Oncotarget* 8 (37), 62183–62194. doi: 10.18632/oncotarget.19136

Hollingshead, M. G., Stockwin, L. H., Alcoser, S. Y., Newton, D. L., Orsburn, B. C., Bonomi, C. A., et al. (2014). Gene expression profiling of 49 human tumor xenografts from *in vitro* culture through multiple *in vivo* passages—strategies for data mining in support of therapeutic studies. *BMC Genomics* 15, 393. doi: 10.1186/1471-2164-15-393

Houshdaran, S., Hawley, S., Palmer, C., Campan, M., Olsen, M. N., Ventura, A. P., et al. (2010). DNA methylation profiles of ovarian epithelial carcinoma tumors and cell lines. *PLoS One* 5 (2), e9359. doi: 10.1371/journal.pone.0009359

Jansen, F. H., Krijgsveld, J., van Rijswijk, A., van den Bemd, G. J., van den Berg, M. S., van Weerden, W. M., et al. (2009). Exosomal secretion of cytoplasmic prostate cancer xenograft-derived proteins. *Mol. Cell. Proteomics* 8 (6), 1192–1205. doi: 10.1074/mcp.M800443-MCP200

Jones, M. F., Hara, T., Francis, P., Li, X. L., Bilke, S., Zhu, Y., et al. (2015). The CDX1-microRNA-215 axis regulates colorectal cancer stem cell differentiation. *Proc. Natl. Acad. Sci. U. S. A.* 112 (13), E1550–E1558. doi: 10.1073/pnas.1503370112

Jung, J., Seol, H. S., and Chang, S. (2018). The generation and application of patient derived xenograft (PDX) model for cancer research. *Cancer Res. Treat.* 50(1), 1–10. doi: 10.4143/crt.2017.307

Kedees, M. H., Babinska, A., Swiatkowska, M., Deitch, J., Hussain, M. M., Ehrlich, Y. H., et al. (2005). Expression of a recombinant protein of the platelet F11 receptor (F11R) (JAM-1/JAM-A) in insect cells: F11R is naturally phosphorylated in the extracellular domain. *Platelets* 16 (2), 99–109. doi: 10.1080/09537100400010329

Kohavi, R. (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *International joint Conference on artificial intelligence* (Mahwah, NJ, USA: Lawrence Erlbaum Associates Ltd), 1137–1145.

Lajoie, M., Hsu, Y. C., Gronostajski, R. M., and Bailey, T. L. (2014). An overlapping set of genes is regulated by both NFIB and the glucocorticoid receptor during lung maturation. *BMC Genomics* 15, 231. doi: 10.1186/1471-2164-15-231

Liu, H. A., and Setiono, R. (1998). Incremental feature selection. *App. Intell.* 9 (3), 217–230. doi: 10.1023/A:1008363719778

Martinez-Garcia, R., Juan, D., Rausell, A., Munoz, M., Banos, N., Menendez, C., et al. (2014). Transcriptional dissection of pancreatic tumors engrafted in mice. *Genome Med.* 6 (4), 27. doi: 10.1186/gm544

Matthews, B. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta Protein Struct.* 405 (2), 442–451. doi: 10.1016/0005-2795(75)90109-9

Narumi, S., Muroya, K., Asakura, Y., Adachi, M., and Hasegawa, T. (2010). Transcription factor mutations and congenital hypothyroidism: systematic genetic screening of a population-based cohort of Japanese patients. *J. Clin. Endocrinol. Metab.* 95 (4), 1981–1985. doi: 10.1210/jc.2009-2373

Ohrn, A. (1999). *Discernibility and Rough Sets in Medicine: Tools and Applications*. PhD, Norwegian University of Science and Technology.

Pan, X., Hu, X., Zhang, Y.-H., Feng, K., Wang, S. P., Chen, L., et al. (2018). Identifying patients with atrioventricular septal defect in down syndrome populations by using self-normalizing neural networks and feature selection. *Genes* 9 (4), 208. doi: 10.3390/genes9040208

Pan, X. Y., and Shen, H. B. (2009). Robust prediction of B-factor profile from sequence using two-stage SVR based on random forest feature selection. *Protein Pept. Lett.* 16 (12), 1447–1454. doi: 10.2174/092986609789839250

Rankin, E. B., Xu, W., Silberg, D. G., and Suh, E. (2004). Putative intestine-specific enhancers located in 5' sequence of the CDX1 gene regulate CDX1 expression in the intestine. *Am. J. Physiol. Gastrointest Liver Physiol.* 286 (5), G872–G880. doi: 10.1152/ajpgi.00326.2003

Rotinen, M., Villar, J., Celay, J., Serrano, I., Notario, V., and Encio, I. (2011). Transcriptional regulation of type 11 17beta-hydroxysteroid dehydrogenase expression in prostate cancer cells. *Mol. Cell. Endocrinol.* 339 (1-2), 45–53. doi: 10.1016/j.mce.2011.03.015

Saber, A., van der Wekken, A. J., Kok, K., Terpstra, M. M., Bosman, L. J., Mastik, M. F., et al. (2016). Genomic aberrations in crizotinib resistant lung adenocarcinoma samples identified by transcriptome sequencing. *PLoS One* 11 (4), e0153065. doi: 10.1371/journal.pone.0153065

Sarver, A. E., Sarver, A. L., Thayanithy, V., and Subramanian, S. (2015). Identification, by systematic RNA sequencing, of novel candidate biomarkers and therapeutic targets in human soft tissue tumors. *Lab. Invest.* 95 (9), 1077–1088. doi: 10.1038/labinvest.2015.80

Scott, A. J., Song, E. K., Bagby, S., Purkey, A., McCarter, M., Gajdos, C., et al. (2017). Evaluation of the efficacy of dasatinib, a Src/Abl inhibitor, in colorectal cancer cell lines and explant mouse model. *PLoS One* 12 (11), e0187173. doi: 10.1371/journal.pone.0187173

Snook, A. E., Magee, M. S., Marszalowicz, G. P., Schulz, S., and Waldman, S. A. (2012). Epitope-targeted cytotoxic T cells mediate lineage-specific antitumor efficacy induced by the cancer mucosa antigen GUCY2C. *Cancer Immunol. Immunother.* 61 (5), 713–723. doi: 10.1007/s00262-011-1133-0

Souza, M. C., Martins, C. S., Silva-Junior, I. M., Chriguer, R. S., Bueno, A. C., Antonini, S. R., et al. (2014). NR3C1 polymorphisms in Brazilians of Caucasian, African, and Asian ancestry: glucocorticoid sensitivity and genotype association. *Arq. Bras Endocrinol. Metabol.* 58 (1), 53–61. doi: 10.1590/0004-2730000002868

Tso, K. Y., Lee, S. D., Lo, K. W., and Yip, K. Y. (2014). Are special read alignment strategies necessary and cost-effective when handling sequencing reads from patient-derived tumor xenografts? *BMC Genomics* 15, 1172. doi: 10.1186/1471-2164-15-1172

Wang, T., Chen, L., and Zhao, X. (2018). Prediction of drug combinations with a network embedding method. *Comb. Chem. High Throughput Screen.* 21 (10), 789–797. doi: 10.2174/1386207322666181226170140

Witek, M., Blomain, E. S., Magee, M. S., Xiang, B., Waldman, S. A., and Snook, A. E. (2014). Tumor radiation therapy creates therapeutic vaccine responses to the colorectal cancer antigen GUCY2C. *Int. J. Radiat. Oncol. Biol. Phys.* 88 (5), 1188–1195. doi: 10.1016/j.ijrobp.2013.12.043

Wyatt, A. W., Mo, F., Wang, K., McConeghy, B., Brahmbhatt, S., Jong, L., et al. (2014). Heterogeneity in the inter-tumor transcriptome of high risk prostate cancer. *Genome Biol.* 15 (8), 426. doi: 10.1186/s13059-014-0426-y

Yu, X., Riley, T., and Levine, A. J. (2009). The regulation of the endosomal compartment by p53 the tumor suppressor gene. *FEBS J.* 276 (8), 2201–2212. doi: 10.1111/j.1742-4658.2009.06949.x

Zaravinos, A., Pieri, M., Mourmouras, N., Anastasiadou, N., Zouvani, I., Delakas, D., et al. (2014). Altered metabolic pathways in clear cell renal cell carcinoma: a meta-analysis and validation study focused on the deregulated genes and their associated networks. *Oncoscience* 1 (2), 117–131. doi: 10.18632/oncoscience.13

Zhang, H., Hylander, B. L., LeVea, C., Repasky, E. A., Straubinger, R. M., Adjei, A. A., et al. (2014). Enhanced FGFR signalling predisposes pancreatic cancer to the effect of a potent FGFR inhibitor in preclinical models. *Br. J. Cancer* 110 (2), 320–329. doi: 10.1038/bjc.2013.754

Zhao, X., Chen, L., Guo, Z.-H., and Liu, T. (2019). Predicting drug side effects with compact integration of heterogeneous networks. *Curr. Bioinform*. doi: 10.2174/1574893614666190220114644

Zhao, X., Chen, L., and Lu, J. (2018). A similarity-based method for prediction of drug side effects with heterogeneous information. *Math. Biosci.* 306, 136–144. doi: 10.1016/j.mbs.2018.09.010

Zhu, S., Qing, T., Zheng, Y., Jin, L., and Shi, L. (2017). Advances in single-cell RNA sequencing and its applications in cancer research. *Oncotarget* 8 (32), 53763–53779. doi: 10.18632/oncotarget.17893

Conflict of Interest Statement: 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 reviewer QZ declared a past co-authorship with one of the authors LC to the handling editor.

frontiers in Genetics

ORIGINAL RESEARCH
published: 03 September 2019
doi: 10.3389/fgene.2019.00775

Single-Cell Transcriptomics Reveals Spatial and Temporal Turnover of Keratinocyte Differentiation Regulators

Alex Finnegan[1†], Raymond J. Cho[2†], Alan Luu[1], Paymann Harirchian[2,3], Jerry Lee[2,3], Jeffrey B. Cheng[2,3‡] and Jun S. Song[1*‡]*

[1] Department of Physics, Carl R. Woese Institute of Genomic Biology, University of Illinois at Urbana-Champaign, Champaign, IL, United States, [2] Department of Dermatology, University of California, San Francisco, San Francisco, CA, United States, [3] Veterans Affairs Medical Center, San Francisco, CA, United States

OPEN ACCESS

Edited by:
Tuo Zhang,
Cornell University, United States

Reviewed by:
Hauke Busch,
Universität zu Lübeck, Germany
Yuriy L. Orlov,
Russian Academy of Sciences, Russia

***Correspondence:**
Jeffrey B. Cheng
Jeffrey.Cheng@ucsf.edu
Jun S. Song
songj@illinois.edu

†These authors have contributed equally to this work

‡These authors have contributed equally to this work and share senior authorship

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *17 April 2019*
Accepted: *23 July 2019*
Published: *03 September 2019*

Citation:
Finnegan A, Cho RJ, Luu A, Harirchian P, Lee J, Cheng JB and Song JS (2019) Single-Cell Transcriptomics Reveals Spatial and Temporal Turnover of Keratinocyte Differentiation Regulators. Front. Genet. 10:775. doi: 10.3389/fgene.2019.00775

Keratinocyte differentiation requires intricately coordinated spatiotemporal expression changes that specify epidermis structure and function. This article utilizes single-cell RNA-seq data from 22,338 human foreskin keratinocytes to reconstruct the transcriptional regulation of skin development and homeostasis genes, organizing them by differentiation stage and also into transcription factor (TF)–associated modules. We identify groups of TFs characterized by coordinate expression changes during progression from the undifferentiated basal to the differentiated state and show that these TFs also have concordant differential predicted binding enrichment in the super-enhancers previously reported to turn over between the two states. The identified TFs form a core subset of the regulators controlling gene modules essential for basal and differentiated keratinocyte functions, supporting their nomination as master coordinators of keratinocyte differentiation. Experimental depletion of the TFs ZBED2 and ETV4, both predicted to promote the basal state, induces differentiation. Furthermore, our single-cell RNA expression analysis reveals preferential expression of antioxidant genes in the basal state, suggesting keratinocytes actively suppress reactive oxygen species to maintain the undifferentiated state. Overall, our work demonstrates diverse computational methods to advance our understanding of dynamic gene regulation in development.

Keywords: Single-cell analysis, transcription regulation, keratinocyte, antioxidant, differentiation

INTRODUCTION

Keratinocytes, the predominant cell type of mammalian epidermis, regulate their gene expression programs to fulfill specialized cellular functions within the different epidermal strata. Additionally, they must balance self-renewal against cell loss, given the epidermis' intrinsic replacement rate of ~28 days in normal human skin. How keratinocytes dynamically govern the hierarchy of self-renewal,

Abbreviations: TF, transcription factor; BK, basal keratinocyte; DK, differentiated keratinocyte; SE, super-enhancer; ROS, reactive oxygen species; GO, gene ontology; cpm, counts per million; TSS, transcription start site; CAGE, cap analysis of gene expression.

differentiation, and maturation remains poorly understood. This article reconstructs the dynamic gene regulatory network rearrangements that occur with keratinocyte differentiation by analyzing human foreskin single-cell RNA-seq (scRNA-seq) data.

Basal keratinocytes (BKs) comprise the basal layer, the innermost layer of the epidermis. Basal keratinocytes divide at controlled rates that are thought to be heterogeneous across progenitor cells, ranging from rarely dividing self-renewing stem cells to rapidly cycling transit amplifying cells (Alcolea and Jones, 2014). In addition to replicating, BKs constitute the basement membrane, which is critical for adhesion of the epidermis and dermis and participate in intercellular signaling required for maintaining tissue homeostasis. Upon differentiation, differentiated keratinocytes (DKs) exit the cell cycle and travel from the basal layer through the more superficial spinous and granular layers culminating in cornification/cell death. During the differentiation process, keratinocytes synthesize components necessary for epidermal barrier function, including desmosomes (specialized adhesion structures) in the spinous layer, secretory organelles called lamellar granules that contain lipids and enzymes, and keratohyalin granules, which contain proteins such as loricrin—the latter two providing vital components of the cornified lipid envelope of the epidermis' outer stratum corneum layer.

At the transcriptomic level, the stratum-specific expression patterns of many key Keratinocyte Genes are known, but regulators of these genes are still being identified (Lopez-Pajares et al., 2015). Constructing the dynamic regulatory network of relevant transcription factors (TFs) and their target genes thus remains an active area of investigation. Previous studies have used various genomic and epigenomic data to construct regulatory networks. For example, Lopez-Pajares et al. (2015) analyzed the time-series transcriptome of experimentally differentiated keratinocyte cultures and identified regulatory relations of genes based on temporal coexpression patterns. Joost et al. (2016) advanced this approach to the single-cell level in murine epidermis, identifying TFs varying with differentiation pseudotime and constructing gene modules using correlation-based expression similarity. In *in vitro* keratinocyte epigenomic studies, Cavazza et al. (2016) and Klein et al. (2017) mapped typical enhancers and super-enhancers (SEs)—large clusters of enhancers characterized by strong activating histone modifications, enrichment of cell type–specific TF motifs, and regulation of cell type–specific genes (Hnisz et al., 2013). Both works identified dramatic changes in sets of SEs between the BK and DK states and developed regulatory networks based on patterns of TF binding/motif enrichment in SEs and proximities of SEs to gene loci (Cavazza et al., 2016; Klein et al., 2017). More recently, the single-cell Perturb-ATAC method revealed changes in regulatory element chromatin accessibility during keratinocyte differentiation and targeted genetic perturbation (Rubin et al., 2019); these data permitted the grouping of TFs with correlated binding site accessibility during differentiation, the inference of interactions between TFs, and the detection of synergy in perturbations of chromatin accessibility (Rubin et al., 2019).

While regulation by TFs and epigenetic modifications ultimately determine gene expression, changes in redox state and abundance of reactive oxygen species (ROS) may help guide the transition from basal to differentiated states (Bigarella et al., 2014). For instance, Hamanaka et al. (2013) demonstrated that reducing ROS through inhibition of oxidative phosphorylation impairs epidermal differentiation and increases proliferation of basal cells and that treatment of cultured keratinocytes with antioxidants impairs differentiation. Likewise, Bhaduri et al. (2015) established MPZL3 and FDXR as proteins localizing to the mitochondria and inducing keratinocyte differentiation by increasing ROS levels. These findings demonstrate opposing roles of ROS and antioxidants in regulating differentiation; however, a genome-wide time-course examination of genes potentially modulating differentiation *via* their antioxidant function has not yet been described.

In this article, we use our recently generated scRNA-seq data assaying expression in 22,338 human foreskin keratinocytes (Cheng et al., 2018) to identify regulators of keratinocyte differentiation and computationally infer dynamic TF networks controlling gene expression patterns required for keratinocyte development and function. We find that expression turnover of established and predicted keratinocyte regulators coincides with previously reported change in SE sets between the BK and DK states (Klein et al., 2017). Depletion of two predicted positive regulators of BKs—ZBED2 and ETV4—leads to differentiation of BKs in the absence of external differentiation-inducing queues. The pattern of differential TF binding-motif enrichment between BK- and DK-specific SEs follows the pattern of TF state-specific expression, leading us to develop gene regulatory networks for TFs. These networks recapitulate known and previously predicted regulatory relationships and also identify novel regulators of differentiation stage–specific functions. In particular, our predicted regulation of cadherins by ETV4 suggests that ETV4's established role of controlling cadherin-mediated cell sorting in branches of the neuronal lineage (Livet et al., 2002; Helmbacher, 2018) may extend to keratinocytes. Supporting the role of cellular antioxidants in suppressing ROS levels, we find that genes related to antioxidant function are preferentially expressed in BK cells and also uncover differences in subcellular localization between antioxidant genes exclusively expressed in BK state and those in DK state.

RESULTS

A Subset of Keratinocyte-Specific Transcription Factors Shows Expression and Binding Patterns Coupled to State-Specific Epigenomes

To identify expression patterns of key TFs across distinct keratinocyte transcriptomic states, we examined a set of 49 established and 44 candidate Keratinocyte regulators, to which we refer below as Keratinocyte TFs. Established keratinocyte regulators were obtained from a previous publication (Klein et al., 2017); Candidate TFs were identified based on keratinocyte-specific RNA expression in the FANTOM5 (Functional ANnoTation Of the Mammalian genome) cell atlas (Fantom Consortium et al., 2014) (Methods; **Supplementary File 1: Figure S1**; **Supplementary File 2: Tables S1, Table S2**). Our approach of selecting candidates based on cell type–specific expression

aimed to increase the confidence that changes in TF expression across single-cell transcriptional states reflect rewiring of gene regulatory networks guiding keratinocyte differentiation and to reduce false positives in subsequent identification of TF targets from correlation analysis.

We clustered foreskin keratinocytes into eight stages *via* approximate spectral clustering of imputed scRNA expression values (**Figure 1A**; **Supplementary File 1: Figure S2;** Methods). As observed previously (Cheng et al., 2018), marker gene expression profiles indicated that these stages largely agreed with known keratinocyte states including a BK state (corresponding to stages 1–3), a mitotic state (stage 4), and a DK state (stages 5–7) (**Supplementary File 1: Figure S3**). The mitotic state had markedly increased levels of cyclins as well as the histone H2A isoform *HIST2H2AC* known to be required for proliferation of undifferentiated mammary epithelial cells (Monteiro et al., 2017). Additionally, the mitotic state had high expression of basal markers (*KRT5, KRT14*) and intermediate expression of early differentiation markers (*KRT1, KRT10*), suggesting it is a rapidly cycling subpopulation in transition from the BK to DK states (**Supplementary File 1: Figure S3**). This interpretation is supported by *in situ* hybridization experiments that have identified basal and suprabasal expression of the mitotic marker gene *MKI67* (Cheng et al., 2018). Stage 8 reproduced the "channel" cluster, identified previously as a novel keratinocyte cell state not on the classic differentiation trajectory (Cheng et al., 2018).

Hierarchical clustering of Keratinocyte TFs that exhibit dynamic expression across stages 1 to 7 clearly separated the TFs with peak expression in the BK state from those with peak expression in the DK state (**Figure 1B**), with a sharp transition occurring in the mitotic state (stage 4). This pattern of expression turnover coincided with the dramatic change in distribution of active SEs between the BK and DK states (previously identified from differential histone modification patterns of H3K4

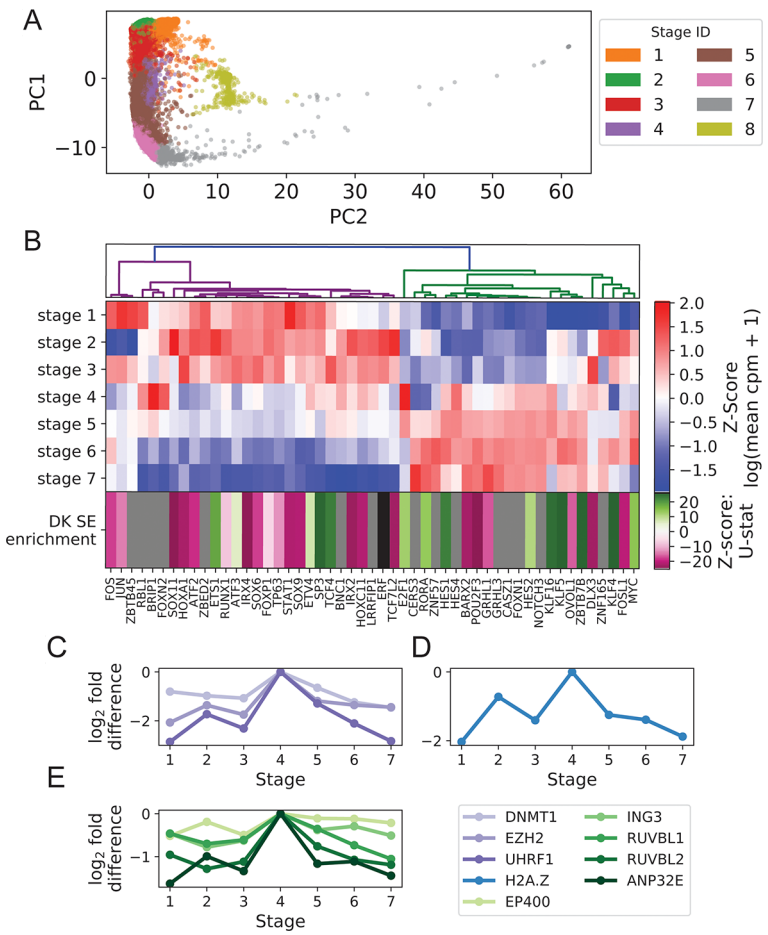


FIGURE 1 | Turnover in Keratinocyte TF expression is temporally and spatially coupled to turnover in SEs. **(A)** Imputed single-cell expression vectors of 22,338 foreskin keratinocytes projected onto first two principal components; stage membership was assigned by k-means–based approximate spectral clustering. **(B)** First seven rows show log-transformed stage-wise mean imputed expression of dynamic Keratinocyte TFs normalized across stages. Bottom row shows the magnitude and direction of differential motif enrichment between BK and DK SEs. Gray and black cells correspond to TFs without a known binding motif and TFs not differentially enriched between SE sets, respectively. Columns are organized by hierarchical clustering on first seven rows (Methods). **(C–E)** Log fold-change in stage-wise mean imputed expression between stage 4 (mitotic state) and other stages for established keratinocyte epigenetic regulators **(C)**, H2A.Z **(D)**, and a subset of components of SWR1 remodeler complex **(E)**. See also **Supplementary File 1**: **Figures S1–4**.

monomethylation, H3K4 trimethylation and H3K27 acetylation) (Cavazza et al., 2016; Klein et al., 2017). We therefore hypothesized that the TFs with peak expression in each state may function through direct binding of state-specific SEs, thereby coupling the transcriptional and epigenetic developmental programs. To test this hypothesis, we first compared the distributions of TF motif occurrence counts (scaled by SE length) between BK and DK SEs and identified 21 and 14 TFs with motifs significantly differentially enriched between BK and DK SEs, respectively. Next, we assigned to each of these TFs a direction and magnitude of differential motif enrichment (**Figure 1B** last row, **Supplementary File 3**: Clustering transcription factor expression trajectories and super-enhancer differential motif enrichment). Grouping the TFs into two expression clusters as shown in **Figure 1B**, we found that the direction and magnitude of TF differential motif enrichment in BK versus DK SEs generally agreed with each cluster's peak expression in BK versus DK state ($p = 0.043$, one-sided Mann-Whitney U test); intuitively, the left (magenta) and right (green) expression branches in **Figure 1B** contained more magenta and more green boxes, respectively, in the last row of **Figure 1B**. This finding, made possible by single-cell analysis, supported the premise that Keratinocyte TF expression and chromatin conformation accessibility are coordinated during transition between keratinocyte cell states.

Next, we identified potential regulators of the switch in state-specific SEs by examining the stage-wise expression of established keratinocyte epigenetic regulators and found several of them, including *EZH2, DNMT1*, and *UHRF1*, to have a strong expression spike in the mitotic state (Ezhkova et al., 2009; Sen et al., 2010) (**Figure 1C**). Additionally, we found that *H2A.Z* and components of the SWR1 remodeling complex, responsible for depositing this enhancer-associated histone subunit, attained peak expression in the mitotic state (**Figures 1D**, **E**). Although the sharp increase in the expression of *H2A.Z* and other histone subunits in this state may be partially explained by the abundance of rapidly dividing cells, the concurrent peak expression of SWR1 components suggested active reorganization of enhancer activities prior to differentiation. Together, these single-cell results highlighted epigenetic remodelers functioning during the mitotic state, potentially to facilitate the turnover of SEs between the BK and DK states.

Knockdown of ETV4 and ZBED2, Predicted Promoters of the BK State, Induces Differentiation

To validate the regulatory function of Candidate Keratinocyte TFs, we ranked the TFs based on their predicted ability to promote the BK state. Candidates were assigned a differentiation-promoting score by first identifying highly correlated keratinocyte-specific regulatory targets and summing their log fold changes between DK and BK states, accounting for the sign of correlation (Methods; **Supplementary File 1: Figure S4**). We filtered out TFs with low expression in undifferentiated keratinocyte cultures (<5 FPKM) and knocked down four of the top five remaining TFs with greatest BK-promoting strength (strong negative differentiation-promoting score) using RNAi in the absence of external differentiation queues.

Depletion of *ETV4* and *ZBED2* transcripts resulted in a significant increase in mRNA expression of the early differentiation marker *KRT10* by 3.84- and 4.17-fold, respectively, compared with control cells transfected with nontargeting siRNA (**Figures 2A**, **B**). Depletion of *ETV4* also showed a significant increase (2.49-fold) in the mRNA expression of the late differentiation marker *FLG*, with *ZBED2* depletion also showing a similar trend (**Figure 2B**). These results confirmed the strong progenitor-promoting function of ETV4 and ZBED2, synthetic reduction of which induced spontaneous differentiation of keratinocytes.

Depletion of *BNC1* and *HOXC11* transcripts did not significantly change the mRNA level of *KRT10* or *FLG* (**Figures 2C**, **D**), suggesting that the regulatory effects of these TFs do not extend to these differentiation markers or that BNC1 and HOXC11 protein expression was not diminished enough to have an effect. Nevertheless, previous knockout of *BNC1* in mouse significantly decreased the number of proliferating keratinocytes in the cornea of the eye (Zhang and Tseng, 2007). Therefore, we conclude that BNC1 likely promotes the BK state in foreskin, although its regulatory targets remain to be experimentally characterized.

Previous reports supported our prediction of the role of SOX9 and IRX4 in keratinocyte differentiation (**Supplementary File 1: Figure S4**). For example, overexpression of SOX9 in keratinocytes has been shown to suppress the late differentiation maker genes IVL and LOR (Shi et al., 2013). Likewise, IRX4 was previously predicted to regulate keratinocyte proliferation and hemidesmosome assembly based on correlation with functionally annotated genes across a large set of publicly available mouse RNA-seq data (Lachmann et al., 2018). Moreover, knockdown of the differentiation-promoting TF GRHL3 in calcium-induced keratinocyte primary cells resulted in a gain of SEs strongly enriched for the IRX4 motif (Klein et al., 2017), suggesting antagonism between IRX4, and this established prodifferentiation TF. Overall, our prioritization of Candidate TFs revealed novel keratinocyte regulators and provided additional candidates for follow-up experiments.

Gene Modules in the Basal Network Promote Tissue Architecture, Control of Hippo Signaling, and Progression to the Mitotic State

We next sought to assign function to Keratinocyte TFs with motifs enriched in state-specific SEs based on their scRNA-seq expression correlation with a set of potential regulatory targets. This set was composed of the Keratinocyte TFs themselves and an additional 747 genes differentially upregulated in FANTOM5 keratinocytes compared with other cell types (Methods; **Supplementary File 2: Table S2**). Focusing first on the regulatory network governing the BK state and its progression to the mitotic state, we clustered the Keratinocyte TFs with enriched motifs in BK SEs based on their expression similarity across single cells in stages 1 to 4. We then clustered the regulatory targets into gene modules based on the similarity of their correlations to the TFs. Organizing the TF/

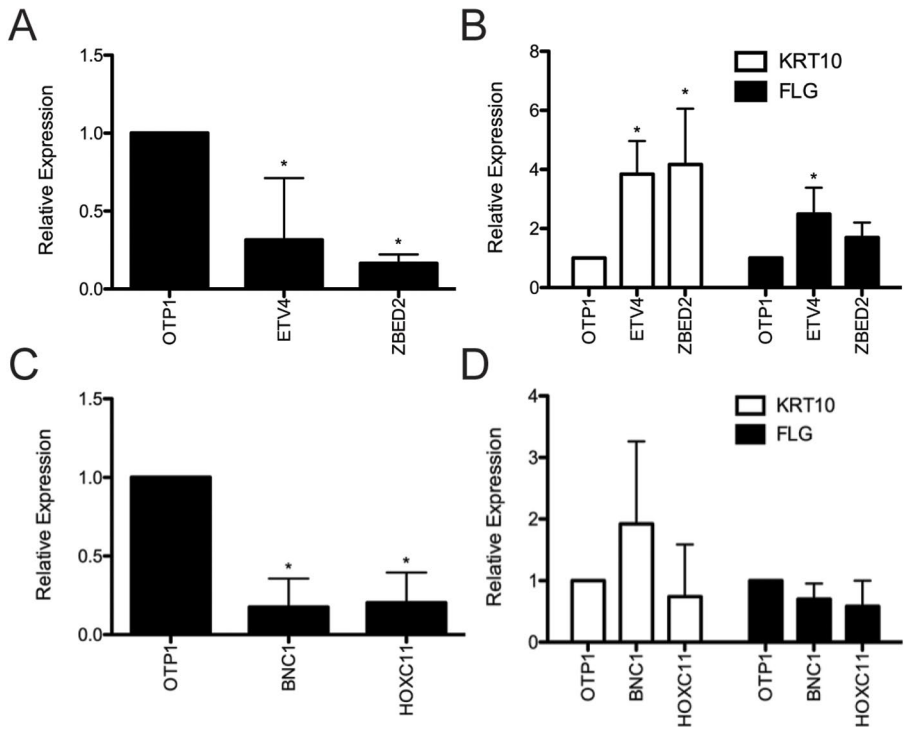


FIGURE 2 | Evaluation of predicted keratinocyte regulators *via* siRNA knockdown. **(A)** RNA was harvested 4 days after transfection from primary human keratinocyte culture treated with *ETV4*, *ZBED2*, or negative control siRNA. Quantitative polymerase chain reaction analysis showed significant ($p < 0.05$, Student *t* test) knockdown of *ETV4* and *ZBED2* mRNA relative to nontargeting siRNA transfected cells. **(B)** Expression of *KRT10* and *FLG* transcript following siRNA knockdown of *ETV4* or *ZBED2*, relative to control. Asterisks indicate $p < 0.05$ (Student *t* test). **(C)** same as **(A)** but for *BNC1* and *HOXC11*. **(D)** Knockdown of *BNC1* and *HOXC11* did not significantly change expression of differentiation marker genes *KRT10* and *FLG*. Error bars indicate 1 standard deviation calculated over four replicates.

target correlation matrix by TF and gene modules (**Supplementary File 1: Figure S5A**) yielded submatrices with strong correlation/anticorrelation delineated by module boundaries. Thresholding on the average correlation strength calculated across gene/TF pairs for each TF and gene module, we identified activating and inhibiting relationships between 13 TF and 23 target gene modules (**Supplementary File 1: Figure S5 (B–D)**; **Supplementary File 2: Table S3; Supplementary File 3:** Regulatory network construction).

Figure 3A shows regulatory relationships for four gene modules enriched in gene ontology (GO) terms (**Figure 3B**) (see **Supplementary File 2**: **Table S4** for full GO output). Gene Module 1 was highly expressed in all BK stages and contained genes important for anchoring cells to the basement membrane and extracellular matrix *via* hemidesmosomes and other cell junctions, genes encoding extracellular signaling molecules, and genes participating in the key Hippo and PI3K intracellular signaling pathways. Transcription factors predicted to activate Module 1 genes recapitulated several established and independently predicted regulatory relationships. For example, TP63 and JUND are known to positively regulate *ITGB4* and *LAMA3A*, respectively (Virolle et al., 1998; Carroll et al., 2006), whereas IRX4 and JUND are both predicted regulators of hemidesmosome assembly (Lachmann et al., 2018).

Notably, four of the six genes in the Hippo pathway (*AJUBA*, *WNT7A*, *WNT7B*, and *WNT3*) and seven of the eight genes in the PI3K pathway (*ITGA3*, *LAMB4*, *LAMB3*, *FGFR2*, *COL4A6*, *ITGB4*, and *LAMA3*) were expressed as extracellular or cell membrane–associated proteins. Given that these pathways involve signaling *via* intracellular posttranslation modification, this result suggested that the primary mechanism for pathway modulation at the transcriptional level might be *via* changing the expression of extracellular signaling molecules and the cell membrane proteins that transduce these signals. Examining the position of Module 1 genes in the Hippo signaling pathway (Kanehisa et al., 2017) illustrated this mechanism and showed that Module 1 genes promoted the pro-proliferative Hippo-OFF signaling state (**Supplementary File 1: Figure S6**). Specifically, the Module 1 cell membrane–associated protein AJUBA and intracellular protein RASSF6 are known to repress MST1/2, allowing nuclear localization of YAP/TAZ, which defines the pro-proliferative Hippo-OFF state (Meng et al., 2016). In the nucleus, TFs activated downstream of Module 1 extracellular WNT signaling proteins (WNT7A, WNT7B, and WNT3) can interact with YAP to promote pro-proliferative genes, including the Module 1 gene *CCND2* (Kanehisa et al., 2017).

Module 2 genes were enriched for keratins and rose sharply in expression at stage 4. Consistent with the strong mitotic signal at

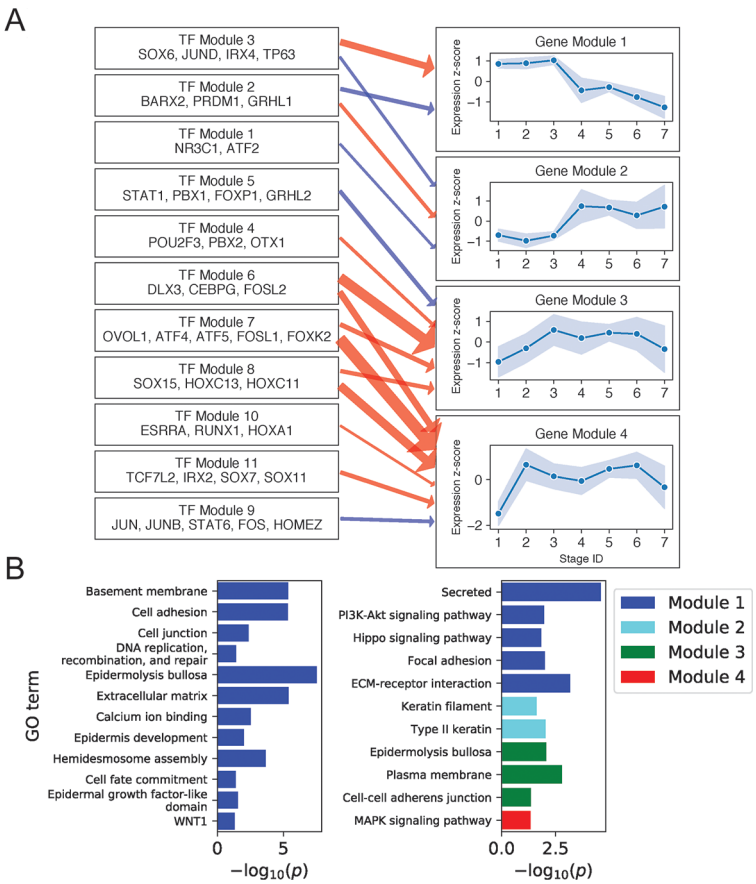


FIGURE 3 | Basal keratinocyte network analysis identifies gene and TF modules specific to basal functions. **(A)** Regulation of four GO-enriched gene modules by TF modules, represented as a directed graph. Gene module nodes show log-transformed stage-wise mean imputed expression normalized across stages 1 to 7 with shading of 1 standard deviation interval. Transcription factor modules list their TF constituents. Arrows indicate regulation with width proportional to predicted strength of activation (red) or inhibition (blue). **(B)** Minus log of adjusted *p* values for selected GO terms enriched in each gene module. See also **Supplementary File 1: Figure S5**.

this stage, two of the three keratins in this module (*KRT6A* and *KRT6B*) were previously implicated in rapid keratinocyte division (Bolognia et al., 2017). Moreover, KRT6A and KRT6B were also shown to suppress keratinocyte migration during wound repair (Rotty and Coulombe, 2012), suggesting that the sharp rise in *KRT6A/B* expression in stage 4 and its fall beyond stage 5 could help inhibit migration of this mitotic cell population from the basal layer (**Supplementary File 1: Figure S7**). The proposed mechanism of impaired migration may explain how this mitotic population remains in or near the basal layer, despite expressing spinous layer markers (e.g., KRT1 and KRT10) at higher levels than BK cells (**Supplementary File 1: Figure S3**).

Previous publications confirmed the function of several transcriptional regulators predicted for Gene Module 2. For example, TP63 knockdown was shown to increase the expression of *KRT6A* in human keratinocyte cell lines (Barbieri et al., 2006). Similarly, conditional knockout of glucocorticoid receptor NR3C1 in mouse keratinocytes was shown to increase the expression of *KRT6A*, *KRT6B*, and *KRT77*, another keratin in the Gene Module (Sevilla et al., 2013).

Gene Module 4 was enriched for MAPK signaling genes (*CRKL*, *FGF11*, *GADD45A*, *FLNB*, *DUSP7*, and *MYC*) and rose sharply in expression at stage 2. The overall effect of Module 4 gene expression on MAPK signaling was complex, with FGF11 and GADD45A activating the ERK and JNK pathways (Kanehisa et al., 2017); DUSP7 inhibiting ERK, JNK, and p38 pathways (Amit et al., 2007; Kanehisa et al., 2017); and CRKL and FLNB serving structural functions. Moreover, different outcomes have been reported for activation of MAPK signaling by Module 4 genes. On the one hand, activation of JNK and P38 pathways by the DNA damage response gene *GADD45A* can promote apoptosis and cell cycle arrest (Hildesheim et al., 2002). On the other hand, activation of ERK signaling by growth factor FGF11 may promote proliferation (Kim et al., 2008). These results, together with our finding of Gene Module 4 regulation by multiple TF modules, including MAPK regulatory targets FOS, JUN (Amit et al., 2007), and FOSL1 (Gillies et al., 2017), suggested complex regulation with multiple feedback mechanisms in controlling proliferation, differentiation, and apoptosis.

Gene Modules in the Differentiated Network Promote Keratinization, Barrier Formation, and Down-Regulation of Basal State Signaling

We next constructed regulatory relationships among gene and TF modules for the DK state using the same method described above, calculating gene correlations across cells in stages 4 to 7 and restricting attention to TFs with motifs enriched in DK-specific SEs (Methods). This analysis identified activating and inhibiting relationships among 21 gene and 9 TF modules (**Supplementary File 1: Figure S8; Supplementary File 2: Table S3**). **Figure 4A** shows regulatory relationships for six gene modules enriched in GO terms (**Figure 4B**) (see **Supplementary File 2**: **Table S4** for full GO output).

Gene Module 1 decreased in expression with differentiation and was enriched for GO terms associated with intercellular signal receptors and intracellular signaling cascades. Many Module genes associated with these terms were also seen to function in basal state signaling pathways. For example, Module genes in the Hippo pathway included cell membrane–associated *AJUBA*, *WNT7B*, and *DLG5* (Elbediwy et al., 2016; Kwan et al., 2016; Kanehisa et al., 2017). Module genes in the MAPK pathway included receptor tyrosine kinases *FGFR3* and *DDR1* (Hilton et al., 2008; Duperret et al., 2014), the kinases *MAPKBP1* and *TNK1* (Hoare et al., 2008; Lecat et al., 2012), the receptor *ADIPOR1* (Shibata et al., 2012), and the phosphoprotein and TF *ATF5*. The decreasing expression of this signaling module thus reflected a shift in the primary cellular function upon differentiation, with basal cells balancing self-renewal and amplification *via* abundant signaling between and within cells, while differentiated cells began suppressing signaling proteins in favor of those needed for barrier function. Several positive regulators of this Module are known to promote cell

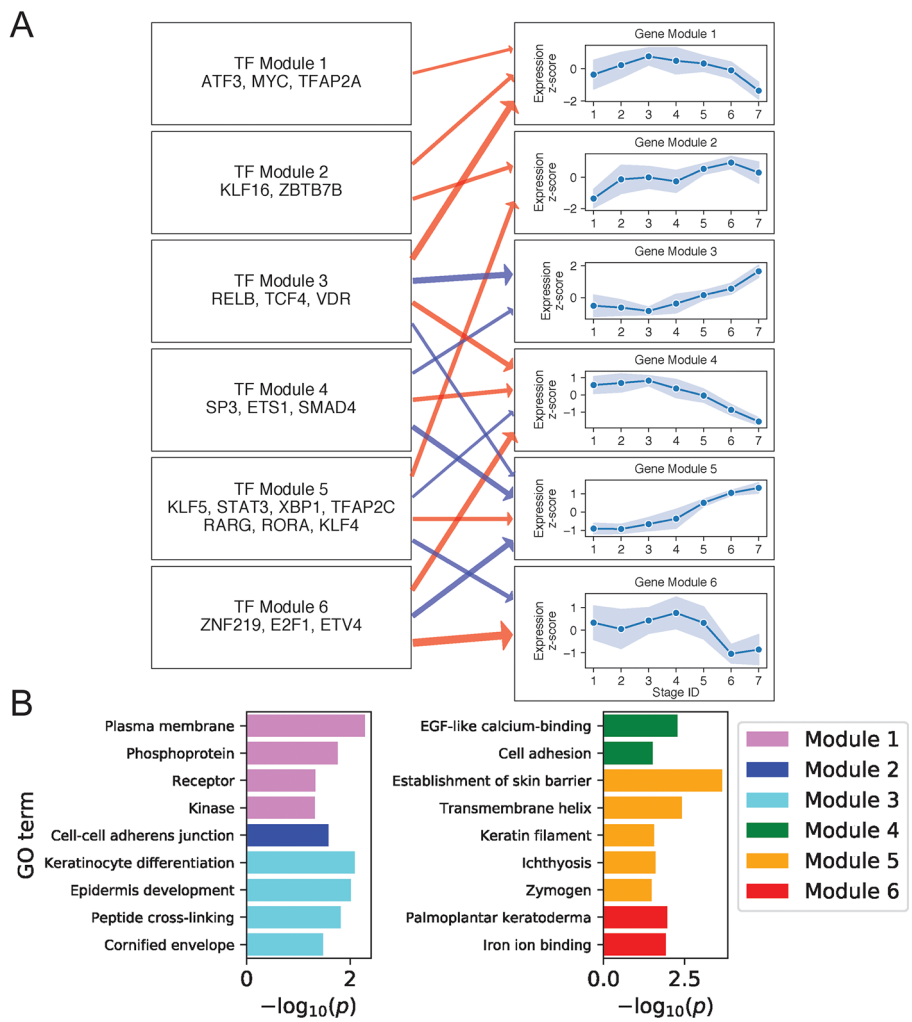


FIGURE 4 | Differentiated keratinocyte network analysis identifies gene and TF modules specific to differentiated functions. **(A)** Regulation of six GO-enriched gene modules by TF modules, represented as a directed graph. Gene module nodes show log-transformed mean stage-wise imputed expression normalized across stages 1 to 7 with shading of 1 standard deviation interval. Transcription factor modules list their TF constituents. Arrows indicate regulation with width proportional to predicted strength of activation (red) or inhibition (blue). **(B)** Minus log of adjusted *p* values for selected GO terms enriched in each gene module. See also **Supplementary File 1: Figure S8**.

cycling, making them plausible regulators of the associated MAPK and Hippo pathways. These regulators included KLF16, which suppresses cyclin-dependent kinase inhibitor CDKN1A (Sakaguchi et al., 2005), and MYC, whose knockdown prevents keratinocyte proliferation (Wu et al., 2012).

Gene Module 4 also decreased with differentiation and was enriched for genes involved in EGF-like calcium binding and cell adhesion. Cell adhesion genes included several members of the cadherin superfamily: *CDH3*, *FAT1*, and *DSG3*. Predicted activators of this Module included our experimentally validated TF ETV4 (**Figure 2B**), which was previously shown to positively regulate cadherins in mouse spinal cord motor neurons, promoting segregation of cells with similar function (Livet et al., 2002; Helmbacher, 2018). Moreover, it was also demonstrated that ETV4 can positively regulate *RUNX1*, another Module 4 gene (Helmbacher, 2018). These findings thus supported that the cadherin regulatory function of ETV4 in the neuronal lineage may extend to keratinocytes.

Gene Module 3 increased its expression with differentiation and was enriched for genes related to the formation of cornified envelope and DK function. For example, the protein products of *LOR*, *SPRR1B*, and *CSTA* in this module are peptides cross-linked in the cornified envelope, while the keratinocyte differentiation protein ACER1 hydrolyzes ceramides, abundant in the granular layer, producing free sphingoid bases with antimicrobial function (Houben et al., 2006). Two other important epidermis development genes in this module were *KLK7* and *CALML5*; KLK7 degrades cellular adhesions of the cornified layer, favoring desquamation (Caubet et al., 2004), and CALML5 is thought to regulate differentiation by mediating cytoplasmic sequestration of YAP1 and initiating the antiproliferative Hippo-ON state (Sun et al., 2015). This gene module did not have positive TF regulators in our network, but had two sets of negative regulators (Modules 3 and 4). Of note, TF Module 4 contained SP3, ETS1, and SMAD4 that were previously shown to interact physically and suppress hematopoiesis (Morikawa et al., 2013; Raz et al., 2014). Our analysis thus indicated that steady reduction of these TFs contributed to the de-repression of Module 3 genes during differentiation.

Gene Module 5, like Module 3, increased its expression with differentiation and was negatively regulated by TF Modules 3 and 4. It contained genes primarily involved in barrier function, with several of these genes (*DEGS2*, *CERS3*, *ABCA12*, *TMEM79*) functioning in lipid synthesis and transport *via* the lamellar granule system. Other module members were involved in cell-cell adhesion (desmosomal proteins DSC1, DSG1, and PERP), tight junctions (CLDN1 and CLDN8), and desquamation (serine-proteases KLK8, KLK11) (Kishibe et al., 2007). Finally, the module also contained the enzymes TGM3 and CASP14 that promote cornification, DK-specific signaling molecules genes *KRTDAP* and *DMKN* (Matsui et al., 2004; Tsuchida et al., 2004), and the antimicrobial gene *DEFB1* (Ali et al., 2001). Apart from negative regulation by TF Modules 3 and 4, Gene Module 5 was positively regulated by TF Module 5. This TF module includes RORA, which is known to positively regulate *ABCA12* and other genes functioning in the granular lipid barrier (Dai et al., 2013). Our analysis thus identified Modules 3 and 5 genes as key components of keratinocyte terminal differentiation coordinately regulated by TFs that may preferentially localize in DK-specific SEs to either suppress or promote terminal differentiation.

Antioxidant Gene Expression Is Enriched in the Basal State and Coupled to the Spatial Organization of Epidermis

Given the documented role of ROS and antioxidants in modulating keratinocyte differentiation (Hamanaka et al., 2013; Bhaduri et al., 2015), we also used our scRNA-seq data to examine coordination between antioxidant gene expression and differentiation state. Clustering of annotated antioxidant genes (Carbon et al., 2009) selected for dynamic expression across stages identified three distinct expression clusters (**Figure 5A**, Methods). The majority of antioxidant genes (20 of 32) belonged to the magenta cluster with peak expression in the basal state. The size of this cluster was significantly larger than expected by chance ($p = 8.5 \times 10^{-4}$, Methods), suggesting that antioxidant genes were preferentially expressed in the basal state to preserve self-renewal capacity by preventing ROS accumulation (Bigarella et al., 2014). In support of this conclusion, the magenta cluster contained the gene *SOD2* whose conditional knockout in mouse keratinocytes has been shown to induce cellular senescence and elevate the expression of differentiation marker genes at wound sites (Velarde et al., 2015).

The remaining two clusters (orange and green) attained peak expression in stages 4 to 5 and stages 5 to 7, respectively. Given the putative role of magenta class genes in preserving the basal state, we sought to identify distinct functions for these late peaking clusters. Gene ontology analysis revealed that magenta cluster proteins were enriched in organelle lumens; by contrast, green cluster gene products were enriched in cytoplasmic vesicles, with a similar trend holding for the group of all genes not in the magenta cluster (**Figure 5B**; **Supplementary File 2: Table S5**). This difference in cellular localization reflected potential differences in function, with magenta cluster proteins localized in key organelles to prevent the initiation of differentiation and green cluster proteins diffused throughout the cytoplasm to mitigate environmental oxidative stress and protect basal cells. Supporting this interpretation, the genes not in the magenta cluster were enriched for the GO term "response to oxidative stress" (**Figure 5B**).

DISCUSSION

Keratinocyte function in the basal and differentiated states depends on complex transcriptional regulation involving TFs, epigenetic modifications, and environmental queues from ROS levels and other stimuli. In this work, we have integrated bulk epigenetic profiles and single-cell expression data to better understand the coordination of these regulatory mechanisms. In particular, by considering known and predicted keratinocyte-specific TFs, we have uncovered that the turnover of this master set of TFs upon differentiation is coupled to the reported transition from BK to DK SEs. We have confirmed that synthetically suppressing

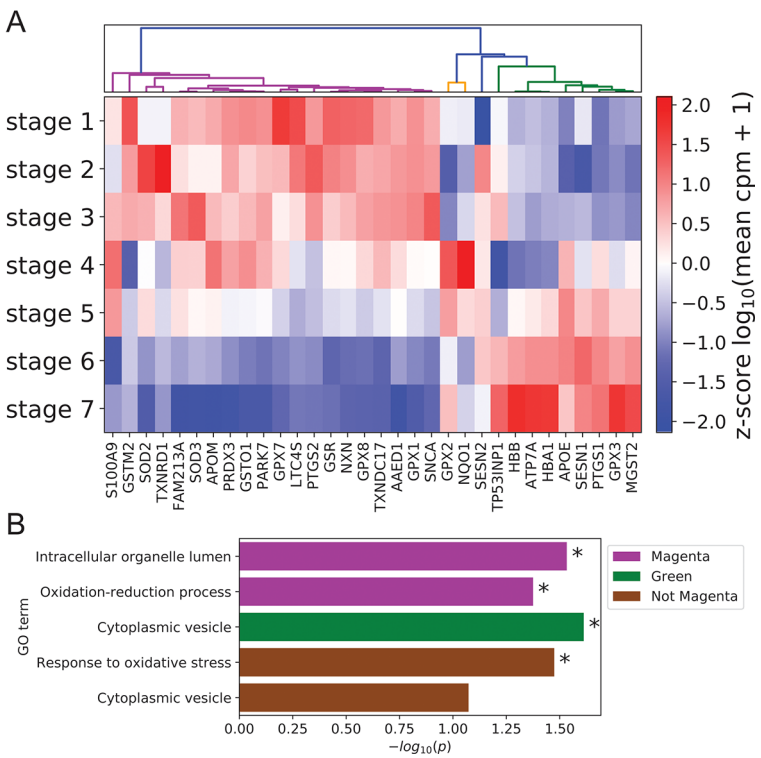


FIGURE 5 | Peak expression of dynamic antioxidant genes is enriched in the BK state. **(A)** Log-transformed stage-wise mean imputed expression of dynamic antioxidant genes normalized across stages. Columns are organized by hierarchical clustering (Methods). **(B)** Minus log of unadjusted *p* values (Methods) for selected GO terms enriched in selected gene sets clustered from **(A)**. Asterisks indicate significance at 0.05 threshold.

the TFs ZBED2 and ETV4, identified in this work as crucial promoters of the basal state, leads to acute differentiation of BKs. We have also prioritized candidate promotors of differentiation that may be studied in subsequent experiments.

The single-cell transcriptomic data have also allowed us to identify a population of mitotic cells containing sharp expression spikes for established keratinocyte epigenetic regulators EZH2, DNMT1, and UHRF1, as well as for the enhancer-associated histone H2A.Z and the SWR1 remodeling complex that deposits the histone variant. The fact that *EZH2*, *DNMT1*, and *UHRF* peak expression coincides with the temporal stage of TF and SE turnover underscores the importance of these genes and helps localize their activity during differentiation pseudotime. Moreover, the co-occurrence of *H2A.Z* and SWR1 complex hypertranscription with this turnover suggests that these genes may have a previously unappreciated role in epigenetic regulation of keratinocyte transition from BK to DK states.

Network analysis has shown that TFs with differential binding in BK versus DK SEs regulate distinct sets of gene modules enriched for important keratinocyte functions. Consistent with previous studies, our BK network analysis has highlighted the role of TP63 in basement membrane adhesion and regulation of intercellular signaling pathways including WNT (Wu et al., 2012), as well as the importance of Hippo signaling in BKs (Elbediwy et al., 2016). Meanwhile, our DK analysis has identified regulators of terminal differentiation gene modules and implicated ETV4 in regulating cadherin superfamily genes, in a manner similar to its established function in motor neurons of the spinal cord (Livet et al., 2002; Helmbacher, 2018). The role of spinal cord cadherins in segregating cells by function suggests that a subset of ETV4 targets may also mediate epidermal cell sorting to assign specific keratinocyte functions to each epidermal layer.

As a proxy for measuring the degree of ROS suppression at each keratinocyte stage, we have demonstrated preferential expression of antioxidant genes in the BK state and uncovered differences in patterns of subcellular localization between BK- and DK-specific antioxidant genes. Notably, BK-specific antioxidant proteins tend to preferentially localize in organelles, such as the mitochondria, where they may control redox levels or the transduction of redox signals, preventing the onset of differentiation. This finding complements previous results that increased expression of select proteins localizing to the mitochondria promotes differentiation by increasing ROS levels (Bhaduri et al., 2015). By contrast, DK-specific antioxidant proteins tend to localize in cytoplasmic vesicles where they may be more important for epidermal barrier function than for regulation of differentiation.

Our integrative models of transcriptional regulation have shown that keratinocyte cell fate determination requires coordinating the expression level of critical TFs with the availability of their binding motifs in differentiation state-specific SEs. The inferred regulatory networks have provided insights into the transcriptional regulation of key genes essential

for skin homeostasis and function. We have thus demonstrated that computational analyses of single-cell transcriptomic profiles in the context of other genomic and epigenomic data provide a powerful method for reconstructing cellular differentiation processes.

MATERIALS AND METHODS

Keratinocyte Isolation and Primary Culture

Primary human keratinocytes were isolated from neonatal foreskin surgical tissue discards obtained with written informed consent using protocols approved by the UCSF institutional review board (#10-00944). Following the method of Lowdon et al. (2014), skin was incubated overnight at 4 °C in 25 U/ml dispase solution (Corning Life Sciences, Corning, NY). Next, epidermis was mechanically separated from the dermis and incubated in 0.05% trypsin for 15 min at 37 °C. Dissociated epidermal cells were filtered with a 100 μm nylon cell strainer (Corning Life Sciences) and then cultured in keratinocyte growth media (KGM; medium 154CF supplemented with 0.07 mM $CaCl_2$ and Human Keratinocyte Growth Supplement; Life Technologies, Waltham, MA).

Data Accession and Cell Selection

Raw counts of scRNA-seq data used in this study were obtained from the European Genome-phenome Archive (EGAS00001002927). The data were generated using Chromium Single Cell 3′ v2 libraries (10X genomics) from three human epidermal samples collected at each of four anatomical locations/disease conditions. Sequence demultiplexing resulted in counts of unique molecular identifiers (UMIs) for genes and noncoding RNA in more than 100,000 cells [see Cheng et al. (2018) for details]. Cell filtering and identification of keratinocytes followed Cheng et al.(2018), with 92,889 passing quality control metrics and 85,345 of these identified as keratinocytes based on average marker gene expression in published cell clusters. This manuscript mainly focuses on the foreskin data from this data set.

RNAi Knockdown of Predicted TFs

ON-TARGETplus siRNA pools targeting *ETV4*, *ZBED2*, *BNC1*, and *HOXC11* as well as the ON-TARGETplus Nontargeting Control siRNA #1 were obtained from Dharmacon (Lafayette, CO). Pooled keratinocytes from five different individuals were seeded at a density of 300,000 cells/ml in 12-well plates. Within 30 min of plating, 10 nM siRNA plus 5 μL/well of Hiperfect transfection reagent (Qiagen, Germantown, MD) was added. Transfections were done in quadruplicates. At 48 hours after transfection, siRNA media was removed and replaced with 1 ml fresh KGM (medium 154CF supplemented with 0.07 mM CaCl2 and Human Keratinocyte Growth Supplement; Life Technologies). Five days after transfection, total RNA was extracted using TRIzol reagent (Life Technologies) following the manufacturer's protocol. cDNA was synthesized using the iScript cDNA Synthesis Kit (Bio-Rad, Hercules, CA) following the manufacturer's protocol. Quantitative polymerase chain reaction was performed with POWER SYBR Green Complete Master Mix (Life Technologies) to measure the expression levels of the housekeeping gene *GUSB*, as well as *ETV4*, *ZBED2*, *BNC1*, *HOXC11*, *KRT10*, and *FLG*. Each sample was measured in triplicate on the Applied Biosystems StepOne System. Melting curves were manually inspected to confirm specificity. When applicable, the results are presented as mean ± standard deviation. Statistical analysis was conducted using GraphPad Prism v5.0f (La Jolla, CA). Student *t* test was used to compare two separate sets of independent and identically distributed samples with $p < 0.05$ considered as significant.

Expression Level of Candidate TFs in Cell Culture

To assess concordance between Candidate TF's differentiation-promoting scores calculated from epidermal scRNA-seq data (Results: Knockdown of ETV4 and ZBED2, predicted promoters of the BK state, induces differentiation; **Supplementary File 3**: Prioritization of knockdown targets; **Figure S4**) and changes in bulk RNA expression of these TFs during *in vitro* differentiation, we generated RNA-seq expression for primary cultured human keratinocytes cultured in basal/proliferating (0.07 mM Ca) or high calcium–induced differentiation (1.2 mM Ca) conditions. Negative control siRNA-treated keratinocytes were used a proxy for normal cultured keratinocytes. Keratinocytes were initially seeded at a density of 100,000 and 150,000 cells in 12-well plates using KGM with 0.07 mM Ca. Within 30 min of plating, 10 nM of either ON-TARGETplus Nontargeting Control siRNA #1 or 2 mixed with 2.5 μl/well of Hiperfect transfection reagent was added. At ~48 h after transfection, subconfluent 100,000-cell wells were harvested using 0.5 ml TRIzol reagent (Life Technologies) for RNA extraction as per manufacturer's protocol. At ~48 h after transfection, the 150,000-cell wells had reached confluency, and the media was replaced with 1 ml fresh KGM with 1.2 mM Ca. After 24 h of exposure to high 1.2 mM calcium, the confluent cells were also harvested using 0.5 ml TRIzol reagent, and RNA-seq was performed. RNA-seq library preparation was performed using KAPA Biosystems Stranded RNA-Seq Kits and RiboErase HMR (Roche, Pleasanton, CA) with 300 to 1,000 ng of total RNA. To minimize batch effects, technical duplicate libraries were generated for each sample. Ribosomal RNA was depleted by hybridization of complementary DNA oligonucleotides plus treatment with RNase H and DNase to remove ribosomal RNA duplexed to DNA and original DNA oligonucleotides, respectively. RNA fragmentation was conducted using heat and magnesium. Using random primers, first-strand complementary DNA (cDNA) synthesis was conducted followed by second-strand synthesis, and A-tailing was added to the 3′ ends using dAMP. Fragments were amplified using appropriate adapter sequences *via* ligation-mediated polymerase chain reaction. Then, the libraries were quantitated with either Quant-iT dsDNA or Qubit dsDNA HS assay kits (Life Technologies). Quality assessment was performed using the LabChip GX Touch HT microfluidics platform (Perkin Elmer, Waltham, MA). 2 × 150 base pair sequencing on a NovaSeq 6000 instrument was performed on libraries

with a PhiX Control v3 (Illumina, San Diego, CA). The RNA-Seq by Expectation Maximization algorithm (Li and Dewey, 2011) was used to quantify gene expression in terms of FPKM for technical replicates in both biological conditions. Change in expression between the differentiation-promoting (1.2 mM Ca) and non–differentiation-promoting (0.07 mM Ca) conditions was quantified as the $\log_2$ ratio of gene expression averaged over technical replicates (**Figure S4**).

Identification of Keratinocyte-Specific Genes and Transcription Factors

Our objective of uncovering regulators and regulatory mechanisms specific to the keratinocyte lineage prompted us to focus analysis on genes and TFs with increased expression in keratinocytes compared with other types of primary cells. On the one hand, focusing on keratinocyte-specific genes and TFs had two benefits: first, it permitted discovery of gene modules particular to keratinocyte functions; and, second, it reduced false positives in our identification of keratinocyte regulators from single-cell data by adding a filter for specificity of expression across primary cells. On the other hand, recognizing that some TFs known to be important for keratinocyte regulation may also function in other cell types, we supplemented the data-driven identification of Keratinocyte TFs with a set of established keratinocyte regulators from the literature.

Identification of genes and TFs with significantly increased expression in keratinocytes used the expression data from the FANTOM consortium (Fantom Consortium et al., 2014). Relative log expression–normalized expression values for transcription start sites identified from cap analysis of gene expression (CAGE) experiments were obtained from http://fantom.gsc.riken.jp/5/datafiles/latest/extra/CAGE_peaks/hg19.cage_peak_phase1and2combined_tpm_ann.osc.txt.gz. Restricting to 495 human primary cell samples not marked for exclusion from expression analysis in Table S2 of (Fantom Consortium et al., 2014), we computed gene-level expression values by associating with each gene's EntrezID the sum of CAGE peak expression values annotated with that ID. We used the Mann-Whitney *U* test to identify genes and TFs differentially expressed in three keratinocyte samples relative to the remaining 491 samples (due to our interest in epidermal keratinocytes, we excluded the oral keratinocyte sample from consideration). A list of annotated TFs (Zhang et al., 2015) was used to distinguish TFs from other protein coding genes and noncoding RNA. Genes and TFs with Benjamini–Hochberg false discovery rate (FDR) less than 0.05 and increased average expression in keratinocytes were selected and filtered to include only those with at least 1 UMI (raw data) in at least 1% of all single-cell keratinocytes (**Supplementary File 2: Table S1**). This differential expression and filtering procedure yielded 793 genes, termed FANTOM genes, and 49 TFs.

The set of differentially expressed TFs, prior to filtering for minimum scRNA-seq expression level, contained several members of the HES superfamily: *HES2*, *HES5*, and *HES7*. Of these, only *HES2* passed the filter. However, we observed that two other superfamily members, *HES1* and *HES4*, were robustly expressed and possessed dynamic expression patterns across our single-cell data (**Figure 1**). For this reason and because HES genes are targets of Notch signaling that has an established function in keratinocyte differentiation (Watt et al., 2008), we elected to add *HES1* and *HES4* to the set of 49 TFs. Below, we refer to the full set of 51 TFs as FANTOM TFs. We supplemented our FANTOM TFs with additional 49 TFs previously shown to regulate keratinocyte differentiation (Klein et al., 2017). Lowly expressed TF were filtered using the threshold on single-cell expression as described above. We refer to this set as Klein TFs.

From these FANTOM genes, FANTOM TFs, and Klein TFs, we constructed the final three sets for further analysis. The set termed Keratinocyte TFs consisted of the union of FANTOM TFs and Klein TFs and was used to study the dynamics of TF expression across single-cell stages, as well as for regulatory network analysis. The set termed Candidate Keratinocyte TFs consisted of FANTOM TFs not in the set of Klein TFs and was the focus of TF prioritization and validation. Finally, the set termed Keratinocyte Genes consisted of the union of Keratinocyte TFs and FANTOM genes and comprised the set of candidate target genes for regulatory network analysis. **Supplementary File 1: Figure S1** illustrates the construction of these sets, and **Supplementary File 2: Table S2** lists the sets' genes.

Summary of scRNA-Seq Data Processing and Analysis

Imputed gene expression was calculated as in Cheng et al. (2018). Briefly, we used the ZINB-WaVE algorithm (Risso et al., 2018) to obtain a low-dimensional, bias-corrected representation of raw singe-cell data, which were then used to construct a distance-based measure of cell similarly and perform imputation with the MAGIC algorithm (version 0.0) (van Dijk et al., 2018). Next, we selected foreskin keratinocytes based on their membership in expression-based clusters previously characterized as keratinocytes in (Cheng et al., 2018). We identified differentiation stages within this cell population by applying principal components analysis followed by k-means–based approximate spectral clustering (Yan et al., 2009) (**Supplementary File 3**: Identification of keratinocyte stages). To reduce false positives in downstream correlation analysis, we removed outlier cells from the eight keratinocyte stages identified by clustering, reduced MAGIC's imputation time parameter, and reimputed (**Supplementary File 1: Figures S9–10**; **Supplementary File 2: Table S6**; **Supplementary File 3**: Calculation of gene correlations).

To construct **Figures 1** and **5**, Keratinocyte TFs and antioxidant genes were filtered for dynamic expression based on stage-wise log fold change and clustered using Pearson correlation distance among vectors of log-transformed stage-wise mean imputed counts per million (cpm) (**Supplementary File 3**: Clustering transcription factor expression trajectories and super-enhancer differential motif enrichment, antioxidant analysis). To prioritize Candidate Keratinocyte TFs for experimental validation, TFs were ranked by the sum of signed log-fold change of their target Keratinocyte Genes during differentiation (positive sign for activation, negative sign for repression). Targets were identified based on strength of TF-gene correlation/anticorrelation (**Supplementary File 3**: Prioritization of knockdown targets).

Regulatory analysis for the BK state used Keratinocyte TFs with motifs enriched in BK-specific SEs compared with DK-specific SEs and Keratinocyte Genes not down-regulated in the BK state compared with the DK state (Methods: Differential expression). Identification of gene and TF modules in the BK state used hierarchical clustering on signed expression similarity scores calculated as soft-thresholded Pearson correlation (Zhang and Horvath, 2005) of log-transformed imputed expression across cells in stages 1 to 4. We identified regulatory relationships between gene and TF modules by considering the distribution of magnitudes of mean similarity scores between all TF-gene module pairs:

$$\left\{\left|\underset{i\in A, j\in B}{\text{mean}}\ S_{i,j}\right| : A \in \text{TF Modules}, \text{B} \in \text{Gene Modules}\right\}$$

where, following the notation of **Supplementary File 3**: Regulatory network construction, $s_{i,j}$ denotes the signed similarity score of TF i and target gene j (**Supplementary File 1: Figure S5B**). Regulatory relationships were assigned for module pairs exceeding the threshold illustrated in **Supplementary File 1: Figure S5(C, D)**. Regulatory analysis for the DK state used an analogous method [**Supplementary File 1: Figures S8(B–D)**]. Further details are given in **Supplementary File 3**: Regulatory network construction.

Source code used to generate results is available at https://github.com/jssong-lab/kcyteReg.

Differential Expression

We used differential expression analysis to identify Keratinocyte Genes specific to the BK (union of stages 1, 2, 3) and DK (union of stages 5, 6, 7) states. First, log (cpm + 1) of nonimputed expression values was calculated for Keratinocyte Genes and for other genes with at least 3 UMIs in 20 foreskin keratinocytes. Next, we used limma-trend version 3.23.9 (Ritchie et al., 2015) to obtain moderated $\log_2$ fold-change values between the two states, as well as adjusted p values for differential expression tests (**Supplementary File 2: Table S7**). Finally, we defined Keratinocyte Genes specific to the BK versus DK states to be those genes differentially expressed at 5% FDR and with magnitudes of moderated $\log_2$ fold change greater than 0.25.

Gene Ontology Analysis

We used the DAVID GO resource (Huang da et al., 2009) to determine functional enrichment in BK and DK gene modules, as well as in clusters of antioxidant genes with similar dynamic gene expression patterns. For BK and DK gene modules, we used the R library RDAVIDWebService (Fresno and Fernandez, 2013) to query DAVID with backgrounds composed of members of each gene module and a common control set of 12,516 expressed genes with at least 1 UMI in at least 1% of all keratinocytes. Bar plots in **Figures 3B** and **4B** show selected GO terms with Benjamini–Hochberg adjusted $p < 0.05$. **Supplementary File 2: Table S4** provides the full DAVID output for all gene modules identified for the BK and DK states. Gene ontology analysis for clusters of dynamically expressed antioxidant genes used the set of 65 antioxidants with at least 1 UMI in at least 1% of all keratinocytes (**Supplementary File 2: Table S2**). Because of the small sizes of gene sets and the large number of enrichment tests performed by DAVID, we did not find any significant enrichment after Benjamini-Hochberg correction for multiple hypothesis testing. We therefore reported uncorrected p values for selected GO terms in **Figure 5B**; **Supplementary File 2: Table S5** provides the full DAVID output.

DATA AVAILABILITY

The data sets analyzed for this study can be found at the European Genome-phenome Archive (https://www.ebi.ac.uk/ega/home) (EGAS00001002927) (single-cell RNAseq data) and at the FANTOM consortium website (http://fantom.gsc.riken.jp/5/datafiles/latest/extra/CAGE_peaks/hg19.cage_peak_phase1and2combined_tpm_ann.osc.txt.gz) (CAGE data). Genomic coordinates of super-enhancers characteristic of basal and differentiated keratinocytes were obtained from Klein et al. (2017). Source code used to generate results is available at https://github.com/jssong-lab/kcyteReg.

ETHICS STATEMENT

Primary human keratinocytes were isolated from neonatal foreskin surgical tissue discards obtained with written informed consent using protocols approved by the UCSF institutional review board (#10-00944).

AUTHOR CONTRIBUTIONS

RC, JC, and JS conceived and supervised the project. AF carried out most of the computational analyses, aided by AL. PH and JL performed the validation experiments. AF, RC, JC, and JS wrote the manuscript with contributions from other authors. All authors read and approved the final manuscript.

FUNDING

This work was supported in part by funds from NIH R01CA163336 and the Grainger Engineering Breakthroughs Initiative to JS, the L.S. Edelheit Family Biological Physics Fellowship to AF, and NIH K08AR067243 to JC.

ACKNOWLEDGEMENTS

We thank Dr. Bogi Andersen and Dr. Rachel Klein for sharing their lists of super-enhancers in keratinocytes.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00775/full#supplementary-material

REFERENCES

Alcolea, M. P., and Jones, P. H. (2014). Lineage analysis of epidermal stem cells. *Cold Spring Harb. Perspect. Med.* 4 (1), a015206. doi: 10.1101/cshperspect.a015206

Ali, R. S., Falconer, A., Ikram, M., Bissett, C. E., Cerio, R., and Quinn, A. G. (2001). Expression of the peptide antibiotics human beta defensin-1 and human beta defensin-2 in normal human skin. *J. Invest. Dermatol.* 117 (1), 106–111. doi: 10.1046/j.0022-202x.2001.01401.x

Amit, I., Citri, A., Shay, T., Lu, Y., Katz, M., Zhang, F., et al. (2007). A module of negative feedback regulators defines growth factor signaling. *Nat. Genet.* 39 (4), 503–512. doi: 10.1038/ng1987

Barbieri, C. E., Tang, L. J., Brown, K. A., and Pietenpol, J. A. (2006). Loss of p63 leads to increased cell migration and up-regulation of genes involved in invasion and metastasis. *Cancer Res.* 66 (15), 7589–7597. doi: 10.1158/0008-5472.CAN-06-2020

Bhaduri, A., Ungewickell, A., Boxer, L. D., Lopez-Pajares, V., Zarnegar, B. J., and Khavari, P. A. (2015). Network analysis identifies mitochondrial regulation of epidermal differentiation by MPZL3 and FDXR. *Dev. Cell.* 35 (4), 444–457. doi: 10.1016/j.devcel.2015.10.023

Bigarella, C. L., Liang, R., and Ghaffari, S. (2014). Stem cells and the impact of ROS signaling. *Development* 141 (22), 4206–4218. doi: 10.1242/dev.107086

Bolognia, J. L., Schaffer, J. V., and Cerroni, L. (2017). *Dermatology E-Book*. Beijing, China: Elsevier Health Sciences.

Carbon, S., Ireland, A., Mungall, C. J., Shu, S., Marshall, B., Lewis, S., et al. (2009). AmiGO: online access to ontology and annotation data. *Bioinformatics* 25 (2), 288–289. doi: 10.1093/bioinformatics/btn615

Carroll, D. K., Carroll, J. S., Leong, C. O., Cheng, F., Brown, M., Mills, A. A., et al. (2006). p63 regulates an adhesion programme and cell survival in epithelial cells. *Nat. Cell. Biol.* 8 (6), 551–561. doi: 10.1038/ncb1420

Caubet, C., Jonca, N., Brattsand, M., Guerrin, M., Bernard, D., Schmidt, R., et al. (2004). Degradation of corneodesmosome proteins by two serine proteases of the kallikrein family, SCTE/KLK5/hK5 and SCCE/KLK7/hK7. *J. Invest. Dermatol.* 122 (5), 1235–1244. doi: 10.1111/j.0022-202X.2004.22512.x

Cavazza, A., Miccio, A., Romano, O., Petiti, L., Malagoli Tagliazucchi, G., Peano, C., et al. (2016). Dynamic transcriptional and epigenetic regulation of human epidermal keratinocyte differentiation. *Stem Cell. Rep.* 6 (4), 618–632. doi: 10.1016/j.stemcr.2016.03.003

Cheng, J. B., Sedgewick, A. J., Finnegan, A. I., Harirchian, P., Lee, J., Kwon, S., et al. (2018). Transcriptional programming of normal and inflamed human epidermis at single-cell resolution. *Cell. Rep.* 25 (4), 871–883. doi: 10.1016/j.celrep.2018.09.006

Dai, J., Brooks, Y., Lefort, K., Getsios, S., and Dotto, G. P. (2013). The retinoid-related orphan receptor RORalpha promotes keratinocyte differentiation *via* FOXN1. *PLoS One* 8 (7), e70392. doi: 10.1371/journal.pone.0070392

Duperret, E. K., Oh, S. J., McNeal, A., Prouty, S. M., and Ridky, T. W. (2014). Activating FGFR3 mutations cause mild hyperplasia in human skin, but are insufficient to drive benign or malignant skin tumors. *Cell. Cycle* 13 (10), 1551–1559. doi: 10.4161/cc.28492

Elbediwy, A., Vincent-Mistiaen, Z. I., Spencer-Dene, B., Stone, R. K., Boeing, S., Wculek, S. K., et al. (2016). Integrin signalling regulates YAP and TAZ to control skin homeostasis. *Development* 143 (10), 1674–1687. doi: 10.1242/dev.133728

Ezhkova, E., Pasolli, H. A., Parker, J. S., Stokes, N., Su, I. H., Hannon, G., et al. (2009). Ezh2 orchestrates gene expression for the stepwise differentiation of tissue-specific stem cells. *Cell* 136 (6), 1122–1135. doi: 10.1016/j.cell.2008.12.043

Fantom Consortium, Forrest, A. R., Kawaji, H., Rehli, M., Baillie, J. K., de Hoon, M. J., et al. (2014). A promoter-level mammalian expression atlas. *Nature* 507 (7493), 462–470. doi: 10.1038/nature13182

Fresno, C., and Fernandez, E. A. (2013). RDAVIDWebService: a versatile R interface to DAVID. *Bioinformatics* 29 (21), 2810–2811. doi: 10.1093/bioinformatics/btt487

Gillies, T. E., Pargett, M., Minguet, M., Davies, A. E., and Albeck, J. G. (2017). Linear integration of ERK activity predominates over persistence detection in Fra-1 regulation. *Cell Syst.* 5 (6), 549–563 e545. doi: 10.1016/j.cels.2017.10.019

Hamanaka, R. B., Glasauer, A., Hoover, P., Yang, S., Blatt, H., Mullen, A. R., et al. (2013). Mitochondrial reactive oxygen species promote epidermal differentiation and hair follicle development. *Sci. Signal.* 6 (261), ra8. doi: 10.1126/scisignal.2003638

Helmbacher, F. (2018). Tissue-specific activities of the Fat1 cadherin cooperate to control neuromuscular morphogenesis. *PLoS Biol.* 16 (5), e2004734. doi: 10.1371/journal.pbio.2004734

Hildesheim, J., Bulavin, D. V., Anver, M. R., Alvord, W. G., Hollander, M. C., Vardanian, L., et al. (2002). Gadd45a protects against UV irradiation–induced skin tumors, and promotes apoptosis and stress signaling *via* MAPK and p53. *Cancer Res.* 62 (24), 7305–7315.

Hilton, H. N., Stanford, P. M., Harris, J., Oakes, S. R., Kaplan, W., Daly, R. J., et al. (2008). KIBRA interacts with discoidin domain receptor 1 to modulate collagen-induced signalling. *Biochim. Biophys. Acta* 1783 (3), 383–393. doi: 10.1016/j.bbamcr.2007.12.007

Hnisz, D., Abraham, B. J., Lee, T. I., Lau, A., Saint-Andre, V., Sigova, A. A., et al. (2013). Super-enhancers in the control of cell identity and disease. *Cell* 155 (4), 934–947. doi: 10.1016/j.cell.2013.09.053

Hoare, S., Hoare, K., Reinhard, M. K., Lee, Y. J., Oh, S. P., and May, W. S., Jr. (2008). Tnk1/Kos1 knockout mice develop spontaneous tumors. *Cancer Res.* 68 (21), 8723–8732. doi: 10.1158/0008-5472.CAN-08-1467

Houben, E., Holleran, W. M., Yaginuma, T., Mao, C., Obeid, L. M., Rogiers, V., et al. (2006). Differentiation-associated expression of ceramidase isoforms in cultured keratinocytes and epidermis. *J. Lipid Res.* 47 (5), 1063–1070. doi: 10.1194/jlr.M600001-JLR200

Huang da, W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4 (1), 44–57. doi: 10.1038/nprot.2008.211

Joost, S., Zeisel, A., Jacob, T., Sun, X., La Manno, G., Lonnerberg, P., et al. (2016). Single-cell transcriptomics reveals that differentiation and spatial signatures shape epidermal and hair follicle heterogeneity. *Cell Syst.* (3) 3221–237, e229. doi: 10.1016/j.cels.2016.08.010

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45 (D1), D353–D361. doi: 10.1093/nar/gkw1092

Kim, D. S., Lee, H. K., Park, S. H., Lee, S., Ryoo, I. J., Kim, W. G., et al. (2008). Terrein inhibits keratinocyte proliferation *via* ERK inactivation and G2/M cell cycle arrest. *Exp. Dermatol.* 17 (4), 312–317. doi: 10.1111/j.1600-0625.2007.00646.x

Kishibe, M., Bando, Y., Terayama, R., Namikawa, K., Takahashi, H., Hashimoto, Y., et al. (2007). Kallikrein 8 is involved in skin desquamation in cooperation with other kallikreins. *J. Biol. Chem.* 282 (8), 5834–5841. doi: 10.1074/jbc.M607998200

Klein, R. H., Lin, Z., Hopkin, A. S., Gordon, W., Tsoi, L. C., Liang, Y., et al. (2017). GRHL3 binding and enhancers rearrange as epidermal keratinocytes transition between functional states. *PLoS Genet.* 13 (4), e1006745. doi: 10.1371/journal.pgen.1006745

Kwan, J., Sczaniecka, A., Heidary Arash, E., Nguyen, L., Chen, C. C., Ratkovic, S., et al. (2016). DLG5 connects cell polarity and Hippo signaling protein networks by linking PAR-1 with MST1/2. *Genes Dev.* 30 (24), 2696–2709. doi: 10.1101/gad.284539.116

Lachmann, A., Torre, D., Keenan, A. B., Jagodnik, K. M., Lee, H. J., Wang, L., et al. (2018). Massive mining of publicly available RNA-seq data from human and mouse. *Nat. Commun.* 9 (1), 1366. doi: 10.1038/s41467-018-03751-6

Lecat, A., Di Valentin, E., Somja, J., Jourdan, S., Fillet, M., Kufer, T. A., et al. (2012). The c-Jun N-terminal kinase (JNK)–binding protein (JNKBP1) acts as a negative regulator of NOD2 protein signaling by inhibiting its oligomerization process. *J. Biol. Chem.* 287 (35), 29213–29226. doi: 10.1074/jbc.M112.355545

Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* 12, 323. doi: 10.1186/1471-2105-12-323

Livet, J., Sigrist, M., Stroebel, S., De Paola, V., Price, S. R., Henderson, C. E., et al. (2002). ETS gene Pea3 controls the central position and terminal arborization of specific motor neuron pools. *Neuron* 35 (5), 877–892. doi: 10.1016/S0896-6273(02)00863-2

Lopez-Pajares, V., Qu, K., Zhang, J., Webster, D. E., Barajas, B. C., Siprashvili, Z., et al. (2015). A LncRNA-MAF: MAFB transcription factor network regulates epidermal differentiation. *Dev. Cell* 32 (6), 693–706. doi: 10.1016/j.devcel.2015.01.028

Lowdon, R. F., Zhang, B., Bilenky, M., Mauro, T., Li, D., Gascard, P., et al. (2014). Regulatory network decoded from epigenomes of surface ectoderm-derived cell types. *Nat. Commun.* 5, 5442. doi: 10.1038/ncomms6442

Matsui, T., Hayashi-Kisumi, F., Kinoshita, Y., Katahira, S., Morita, K., Miyachi, Y., et al. (2004). Identification of novel keratinocyte-secreted peptides dermokine-alpha/-beta and a new stratified epithelium-secreted protein gene complex on human chromosome 19q13.1. *Genomics* 84 (2), 384–397. doi: 10.1016/j.ygeno.2004.03.010

Meng, Z., Moroishi, T., and Guan, K. L. (2016). Mechanisms of Hippo pathway regulation. *Genes Dev.* 30 (1), 1–17. doi: 10.1101/gad.274027.115

Monteiro, F. L., Vitorino, R., Wang, J., Cardoso, H., Laranjeira, H., Simoes, J., et al. (2017). The histone H2A isoform Hist2h2ac is a novel regulator of proliferation and epithelial-mesenchymal transition in mammary epithelial and in breast cancer cells. *Cancer Lett.* 396, 42–52. doi: 10.1016/j.canlet.2017.03.007

Morikawa, M., Koinuma, D., Miyazono, K., and Heldin, C. H. (2013). Genome-wide mechanisms of Smad binding. *Oncogene* 32 (13), 1609–1615. doi: 10.1038/onc.2012.191

Raz, S., Stark, M., and Assaraf, Y. G. (2014). Binding of a Smad4/Ets-1 complex to a novel intragenic regulatory element in exon12 of FPGS underlies decreased gene expression and antifolate resistance in leukemia. *Oncotarget* 5 (19), 9183–9198. doi: 10.18632/oncotarget.2399

Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J. P. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* 9 (1), 284. doi: 10.1038/s41467-017-02554-5

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43 (7), e47. doi: 10.1093/nar/gkv007

Rotty, J. D., and Coulombe, P. A. (2012). A wound-induced keratin inhibits Src activity during keratinocyte migration and tissue repair. *J. Cell Biol.* 197 (3), 381–389. doi: 10.1083/jcb.201107078

Rubin, A. J., Parker, K. R., Satpathy, A. T., Qi, Y., Wu, B., Ong, A. J., et al. (2019). Coupled single-cell CRISPR screening and epigenomic profiling reveals causal gene regulatory networks. *Cell* 176 (1-2), 361–376 e317. doi: 10.1016/j.cell.2018.11.022

Sakaguchi, M., Sonegawa, H., Nukui, T., Sakaguchi, Y., Miyazaki, M., Namba, M., et al. (2005). Bifurcated converging pathways for high Ca2+- and TGFbeta-induced inhibition of growth of normal human keratinocytes. *Proc. Natl. Acad. Sci. U S A* 102 (39), 13921–13926. doi: 10.1073/pnas.0500630102

Sen, G. L., Reuter, J. A., Webster, D. E., Zhu, L., and Khavari, P. A. (2010). DNMT1 maintains progenitor function in self-renewing somatic tissue. *Nature* 463 (7280), 563–567. doi: 10.1038/nature08683

Sevilla, L. M., Latorre, V., Sanchis, A., and Perez, P. (2013). Epidermal inactivation of the glucocorticoid receptor triggers skin barrier defects and cutaneous inflammation. *J. Invest. Dermatol.* 133 (2), 361–370. doi: 10.1038/jid.2012.281

Shi, G., Sohn, K. C., Li, Z., Choi, D. K., Park, Y. M., Kim, J. H., et al. (2013). Expression and functional role of Sox9 in human epidermal keratinocytes. *PLoS One* 8 (1), e54355. doi: 10.1371/journal.pone.0054355

Shibata, S., Tada, Y., Asano, Y., Hau, C. S., Kato, T., Saeki, H., et al. (2012). Adiponectin regulates cutaneous wound healing by promoting keratinocyte proliferation and migration *via the* ERK signaling pathway. *J. Immunol.* 189 (6), 3231–3241. doi: 10.4049/jimmunol.1101739

Sun, B. K., Boxer, L. D., Ransohoff, J. D., Siprashvili, Z., Qu, K., Lopez-Pajares, V., et al. (2015). CALML5 is a ZNF750- and TINCR-induced protein that binds stratifin to regulate epidermal differentiation. *Genes Dev.* 29 (21), 2225–2230. doi: 10.1101/gad.267708.115

Tsuchida, S., Bonkobara, M., McMillan, J. R., Akiyama, M., Yudate, T., Aragane, Y., et al. (2004). Characterization of Kdap, a protein secreted by keratinocytes. *J. Invest. Dermatol.* 122 (5), 1225–1234. doi: 10.1111/j.0022-202X.2004.22511.x

van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A. J., et al. (2018). Recovering gene interactions from single-cell data using data diffusion. *Cell* 174 (3), 716–729 e727. doi: 10.1016/j.cell.2018.05.061

Velarde, M. C., Demaria, M., Melov, S., and Campisi, J. (2015). Pleiotropic age-dependent effects of mitochondrial dysfunction on epidermal stem cells. *Proc. Natl. Acad. Sci. U S A* 112 (33), 10407–10412. doi: 10.1073/pnas.1505675112

Virolle, T., Monthouel, M. N., Djabari, Z., Ortonne, J. P., Meneguzzi, G., and Aberdam, D. (1998). Three activator protein-1-binding sites bound by the Fra-2.JunD complex cooperate for the regulation of murine laminin alpha3A (lama3A) promoter activity by transforming growth factor-beta. *J. Biol. Chem.* 273 (28), 17318–17325. doi: 10.1074/jbc.273.28.17318

Watt, F. M., Estrach, S., and Ambler, C. A. (2008). Epidermal notch signalling: differentiation, cancer and adhesion. *Curr. Opin. Cell Biol.* 20 (2), 171–179. doi: 10.1016/j.ceb.2008.01.010

Wu, N., Rollin, J., Masse, I., Lamartine, J., and Gidrol, X. (2012). p63 regulates human keratinocyte proliferation *via* MYC-regulated gene network and differentiation commitment through cell adhesion-related gene network. *J. Biol. Chem.* 287 (8), 5627–5638. doi: 10.1074/jbc.M111.328120

Yan, D. H., Huang, L., and Jordan, M. I. (2009). Fast approximate spectral clustering. *Kdd-09: 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining,* 907–915. doi: 10.1145/1557019.1557118

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* 4, Article17. doi: 10.2202/1544-6115.1128

Zhang, H. M., Liu, T., Liu, C. J., Song, S., Zhang, X., Liu, W., et al. (2015). AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors. *Nucleic Acids Res.* 43(Database issue), D76–D81. doi: 10.1093/nar/gku887

Zhang, X., and Tseng, H. (2007). Basonuclin-null mutation impairs homeostasis and wound repair in mouse corneal epithelium. *PLoS One* 2 (10), e1087. doi: 10.1371/journal.pone.0001087

Conflict of Interest Statement: 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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 12 September 2019
doi: 10.3389/fgene.2019.00805

Characterization and Expression Analysis of *ERF* Genes in *Fragaria vesca* Suggest Different Divergences of Tandem *ERF* Duplicates

Xiaojing Wang[1†], Shanshan Lin[1†], Decai Liu[1], Quanzhi Wang[2], Richard McAvoy[3], Jing Ding[1] and Yi Li[1,3*‡]*

[1] State Key Laboratory of Crop Genetics and Germplasm Enhancement and College of Horticulture, Nanjing Agricultural University, Nanjing, China, [2] Engineering and Technology Center for Modern Horticulture, Jiangsu Vocational College of Agriculture and Forestry, Zhenjiang, China, [3] Department of Plant Science and Landscape Architecture, University of Connecticut, Storrs, CT, United States

OPEN ACCESS

Edited by:
Jialiang Yang, Geneis (Beijing) Co. Ltd, China

Reviewed by:
Hao Wang, University of Georgia, United States
Bing Wang, Anhui University of Technology, China
Jingyin Yu, University of Georgia, United States

***Correspondence:**
Jing Ding
jding@njau.edu.cn
Yi Li
yi.li@uconn.edu

[†]These authors have contributed equally to this work

[‡]Yi Li is a no pay visiting professor at Nanjing Agricultural University

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *14 January 2019*
Accepted: *31 July 2019*
Published: *12 September 2019*

Citation:
Wang X, Lin S, Liu D, Wang Q, McAvoy R, Ding J and Li Y (2019) Characterization and Expression Analysis of ERF Genes in Fragaria vesca Suggest Different Divergences of Tandem ERF Duplicates. Front. Genet. 10:805. doi: 10.3389/fgene.2019.00805

Ethylene-responsive factors (ERFs) play important roles in plant growth and development and in responses to abiotic stresses. However, little information was available about the *ERF* genes in woodland strawberry (*Fragaria vesca*), a genetic model plant for the *Fragaria* genus and Rosaceae family. In this study, 91 *FveERF* genes were identified, including 35 arrayed in tandem, indicating that tandem duplication is a major mechanism for the expansion of the FveERF family. According to their phylogenetic relationships with *AtERF*s from *Arabidopsis thaliana*, the tandem *FveERF* genes could be grouped into ancestral and lineage-specific tandem ones. The ancestral tandem *FveERF*s are likely derived from tandem duplications that occurred in the common ancestor of *F. vesca* and *A. thaliana*, whereas the lineage-specific ones are specifically present in the *F. vesca* lineage. The lineage-specific tandem *FveERF* duplicates are more conserved than the ancestral ones in sequence and structure. However, their expression in flowers and fruits is similarly diversified, indicating that tandem *FveERF*s have diverged rapidly after duplication in this respect. The lineage-specific tandem *FveERF*s display the same response patterns with only one exception under drought or cold, whereas the ancestral tandem ones are largely differentially expressed, suggesting that divergence of tandem *FveERF* expression under stress may have occurred later in the reproductive development. Our results provide evidence that the retention of tandem *FveERF* duplicates soon after their duplication may be related to their divergence in the regulation of reproductive development. In contrast, their further divergence in expression pattern likely contributes to plant response to abiotic stress.

Keywords: *ERF* genes, tandem duplication, divergence, expression pattern, woodland strawberry

INTRODUCTION

Plants are sessile organisms and cannot escape from environmental stresses, which can negatively impact their survival, development, and productivity. As such, plants have evolved mechanisms to respond and adapt to stress at the physiological and biochemical levels (Figueiredo et al., 2012). Ethylene-responsive factors (ERFs) are transcription factors that have been shown to play critical

roles in stress response and during plant growth and development (Brown et al., 2003; Chakravarthy et al., 2003; Agarwal et al., 2006; Chen G et al., 2008; Chen J. Q et al., 2008; Sun et al., 2014; Tan et al., 2018).

The ERF family belongs to the APETALA2/ERF (AP2/ERF) superfamily, which also contains the AP2 and RAV families (Weigel, 1995). ERF family proteins contain only one AP2/ERF domain, while the AP2 family contains proteins with a double tandem-repeated AP2 domain and the RAV family contains an additional B3 DNA-binding domain along with a single AP2/ERF domain (Matías-Hernández et al., 2014). In *Arabidopsis thaliana*, the ERF family is divided into 10 groups (I to X) based on phylogeny and gene/protein structure analyses (Nakano et al., 2006). ERF family genes have diverse expression patterns during plant growth and development (Wilson et al., 1996; Liu et al., 1998; Banno et al., 2001), as well as in response to abiotic stresses, such as drought, cold, and high salinity (Song et al., 2005; Novillo et al., 2007; Golldack et al., 2011; Licausi et al., 2013).

Tandem gene duplication is one of the main gene-duplication mechanisms in eukaryotes and has contributed to the prevalence of gene family clusters (Fortna et al., 2004; Fan et al., 2008). The number of tandem duplicates in plants varies from 451 (4.6% of gene content) in *Craspedia variabilis* to 16,602 (26.1% of gene content) in apple (*Malus* × *domestica*) (Yu et al., 2015). Genome-wide analysis in *A. thaliana* has revealed that genes that expanded mainly through tandem duplication tend to be involved in plant responses to abiotic and biotic stresses (Hanada et al., 2008). To the contrary, transcription factors including *ERF*s are preferentially retained after whole-genome duplication (WGD) rather than tandem duplication (Maere et al., 2005; Jourda et al., 2014; Charfeddine et al., 2015). Nevertheless, studies in *A. thaliana* and cucumber show that tandem-duplication events have also played an important role in the expansion of the *ERF* gene family (Nakano et al., 2006; Hu and Liu, 2011).

Duplicate genes experience relaxed negative selection following duplication (Carretero-Paulet and Fares 2012). Increased rates of evolution, *via* divergence of gene sequence, structure, and so forth, have been observed in duplicate gene copies (Carretero-Paulet and Fares, 2012; Wang et al., 2013). Divergence in expression patterns of duplicate genes is affected by their functional categories, duplication mechanisms, species, and other factors (Wang et al., 2012a). Studies in *A. thaliana* and rice show that expression divergence among tandem duplicates occurs shortly after duplication (Ganko et al., 2007; Li et al., 2009), and its overall level is similar to that of WGD duplicates but lower than that of duplicates from other mechanisms (Wang et al., 2012b). There is no significant correlation between expression divergence of tandem duplicates and their synonymous substitution rates, a proxy for the time of duplication (Ganko et al., 2007; Panchy et al., 2016). This indicates that young and old tandem duplicates have a similar level of expression divergence. However, this observation is mainly based on expression analysis in developmental tissues/organs; whether it is the case for expression patterns under stressed conditions remains unclear.

Cultivated strawberry (*Fragaria* × *ananassa*) is a popular crop worldwide; however, genetic analysis of cultivated strawberry is extremely complicated due to its octoploid genome ($2n = 8x = 56$), with as many as four diploid ancestors. Nowadays, woodland strawberry (*Fragaria vesca*) is emerging as a model fruit crop plant species. It has a small diploid genome (240 Mb, $2n = 2x = 14$) with a widely available genome sequence (Shulaev et al., 2011) and a short reproductive cycle (14–15 weeks in climate-controlled greenhouses). In this study, we performed a comprehensive analysis of the ERF family in *F. vesca*, including phylogeny, chromosomal localization, gene structure, motif, duplication mechanism, and expression profiling. Tandem *FveERF* genes were grouped into ancestral and lineage-specific tandem ones and subjected to expression pattern analysis during reproductive development and in response to drought or cold stress. The results of this study should be useful towards future analyses of the divergence and functions of *ERF* genes, particularly tandem duplicated *ERF* genes in strawberry.

MATERIALS AND METHODS

Identification of *AP2/ERF* Genes in *F. vesca*

The *F. vesca* genome sequence and corresponding annotations were downloaded from the DOE Joint Genome Institute website (http://genome.jgi.doe.gov/). First, the full alignment file for the AP2 domain (PF00847) obtained from the Pfam database (Finn et al., 2016) was used to build an HMM file using the HMMER3 software package (Eddy, 1998). Second, HMM searches were performed against the local protein databases of *F. vesca* using the HMMER3 package. Moreover, we checked the physical localizations of all candidate genes and rejected redundant sequences with the same chromosome location and short proteins (length < 100 aa). Finally, sequences of all matching proteins were again analyzed in the Pfam database to verify the presence of AP2 domains. AP2 domains were also detected by the SMART (http://smart.embl-heidelberg.de/) database with an *E*-value cutoff of 10^{-10}. After the above four steps, the identified protein sequences that contained the core domains (AP2 domain) of known AP2/ERFs were regarded as putative homologs in the study.

Gene Structure and Chromosomal Localization of *FveERF* Genes

Exon/intron information and chromosomal location of *FveERF* genes were extracted from the *F. vesca* genome annotation database. The data were then plotted using the MapInspect software (http://mapinspect.software.informer.com/). Tandem duplicate *FveERF*s were defined as *FveERF*s in any gene pair that is located within 100 kb of each other and separated by no more than 10 non-homologous intervening genes (Hanada et al., 2008). Fgenesh (http://www.softberry.com) was used to re-annotate the intergenic regions between putative tandem *FveERF* duplicates, to clarify whether there are any unannotated intervening genes. If the number of non-homologous intervening genes based on genome annotation and our re-annotation results is no more than 10, we consider the pair of *FveERF*s as tandem duplicate genes. The tandem *ERF* genes in *Malus* × *domestica*,

Prunus mume, *Populus trichocarpa*, *Brassica rapa*, *Vitis vinifera*, *Solanum tuberosum*, and *Oryza sativa* were identified based on the same criterion without re-annotation of the intergenic regions. Besides, the tandem *AtERF* genes in *A. thaliana* were retrieved from the study by Nakano et al. (2006).

Phylogenetic Analyses of *ERF* Genes from *F. vesca* and *A. thaliana*

The sequences of 146 AP2/ERF proteins from *A. thaliana,* identified by Nakano et al. (2006), were used for comparative analysis in the study. Full-length amino acid sequences of the AP2/ERFs from *F. vesca* and *A. thaliana* were aligned using ClustalX2.0 (Larkin et al., 2007) and MAFFT [version 7, Katoh and Standley (2013)], respectively, with default parameters. A maximum-likelihood (ML) phylogeny based on the ClustalX alignment (**Figure S1A**) and a aBayes phylogeny based on the MAFFT alignment (**Figure S1B**) were constructed, respectively, using the PhyML software (version 3.0, Guindon et al., 2010). Both phylogenies show a same grouping of the FveAP2/ERF superfamily. Next, full-length amino acid sequences of the identified FveERFs were aligned with those of the AtERFs using ClustalX2.0 and MAFFT, respectively. The JTT+G+I substitution model was identified as the optimal model of amino acid sequence evolution using the program MODELGENERATOR (Keane et al., 2006) with four gamma categories (Jones et al., 1992). ML phylogenies based on the ClustalX (**Figure 2**) and MAFFT alignments and an aBayes phylogeny based on the MAFFT alignment (**Figure S2**) were constructed, respectively, using the PhyML software with the model. The reliabilities of the ML phylogenies and the aBayes phylogenies were tested using bootstrapping with 100 replicates and Bayes posterior probabilities, respectively.

Motif Analysis of FveERF Proteins

The MEME5.0.1 online program (http://meme-suite.org/) was used for the identification of motifs in the FveERF protein sequences. The optimized parameters were employed for the analysis as follows: number of repetitions: any; maximum number of motifs: 15; and the optimum width of each motif: between 6 and 50 residues (Bailey et al., 2015).

Synteny Analysis

Synteny analysis of the *F. vesca* genome was conducted locally using a method similar to the one used by the plant genome duplication database (PGDD, http://chibba.agtec.uga.edu/duplication/, Lee et al., 2013). First, BLASTP was performed to search for potential homologous gene pairs ($E < 10^{-5}$, top five matches) in *F. vesca* genome. Then, the homologous pairs were used as input for MCScanX to identify syntenic chains and types of duplication mechanisms (Tang et al., 2008; Wang et al., 2012a).

Calculation of Pi, Ka, Ks, and Ka/Ks Values of *FveERF* Genes

Pairwise nucleotide divergence among paralogs was estimated by Pi using DnaSP v4.0 (Rozas et al. 2003). To analyze evolutionary rates of tandem duplicate *FveERF*s, the coding sequences of *FveERF* genes were aligned on the basis of the corresponding aligned protein sequences using the PAL2NAL software (Suyama et al. 2006). The ratio of nonsynonymous substitutions per nonsynonymous site (Ka) to synonymous substitutions per synonymous site (Ks) in tandem gene pairs was calculated by using the yn00 program of the PAML package (Yang, 1997). Generally, a Ka/Ks ratio >1 indicates positive selection, and a ratio <1 indicates negative or purifying selection, while a ratio of 1 indicates neutral evolution.

Expression Pattern of FveERF Family Genes and Correlation Analysis

Expression data of *FveERF* genes among different stages and tissues of *F. vesca* flowers and early fruits were retrieved from the SGR database (http://bioinformatics.towson.edu/strawberry/). The heat map was created using the log2 "relative RPKM (reads per kilobase per million) values" of individual *FveERF* genes. For a detailed description of the stages and tissues, please see http://bioinformatics.towson.edu/strawberry/newpage/Tissue_Description.aspx. According to Kang et al. (2013), a gene with an RPKM value lower than 0.3 was regarded not to be expressed in a certain stage or tissue. A gene with RPKM values higher than 0.3 in at least two stages or tissues was regarded as an expressed gene during flower or early-fruit development. Statistical tests of differences between expression levels of tandem/clustered and other *FveERF*s, and of ancestral and lineage-specific tandem *FveERF*s were performed using *t*-test. The correlation between expression patterns of tandem duplicate genes was evaluated by calculating correlation coefficients of the expression data, where the RPKM values lower than 0.3 was not included.

Growth Conditions, Plant Material Collection, and Abiotic Treatments

All plant material was collected from a seventh-generation inbred line of *F. vesca* 'Ruegen' (kindly provided by Janet Slovin). Plants were grown in 10 cm × 10 cm pots in a growth chamber on a 16-h light (22 °C)/8-h dark (20 °C) cycle with 65% relative humidity. Light (~160 μmol m^{-2} s^{-1}) was supplied by sodium lamps. Four developmental stages of Ruegen receptacles were collected for quantitative PCR (qPCR) analysis: little white (white flesh with green achenes, ~20 DPA), pre-turning (white flesh with red achenes, ~DPA), pink (light pink flesh with red achenes, ~27 DPA), and red (flesh is all red, ~29 DPA) stages. All samples were collected and immediately put into liquid nitrogen.

Prior to abiotic stress treatments, strawberry seedlings were grown on solid MS media in the growth chamber on a 16-h light (22 °C)/8-h dark (20 °C) cycle for 1 month. Cold stress treatments were carried out as described in Gu et al. (2016). For drought stress treatments, the seedlings were removed from the media, placed on filter paper under dim light and 30% humidity, and collected after 1, 3, and 8 h of dehydration. Following abiotic stress treatment, plant materials were immediately put into liquid nitrogen prior to RNA processing.

RNA Extraction and Quantitative RT-PCR (qRT-PCR) Analysis

The RNA of stress-treated seedlings was isolated using a TaKaRa MiniBEST Plant RNA Extraction Kit. Nine *FveERF*s from all the lineage-specific tandem repeats (all six genes from two tandem repeats plus three genes randomly selected from the six-gene tandem repeat, mrna08071–mrna08075 and mrna08077, **Table S1**) were selected for qRT-PCR analyses. As most lineage-specific tandem *FveERF*s belong to group 9, the nine ancestral tandem *FveERF*s in groups 9 and 10 were selected for comparison. qRT-PCR primers for these genes are listed in **Table S1**. Expression of the four lineage-specific tandem *FveERF*s that are very lowly expressed in early fruits (mrna04911, mrna04913, mrna08873, and mrna08876) was not examined in the fruit-ripening stages. qRT-PCR was performed using SYBR Premix Ex Tag (TaKaRa) using cDNA as the template. Results were analyzed using the $^{-\Delta\Delta}$CT method with *GAPDH* gene expression as an internal reference (Livak and Schmittgen, 2001; Amil-Ruiz et al., 2013). Three biological and three technical replicates were used.

RESULTS

Genome-Wide Identification of *ERF* Genes in *F. vesca*

To identify the ERF family members in *F. vesca*, the full-length alignment of the AP2/ERF domain (PF00847) was downloaded and used to search the *F. vesca* proteome. A total of 115 proteins were considered as AP2/ERF candidates, containing at least one AP2/ERF domain. Maximum-likelihood (ML, **Figure S1A**) and aBayes (**Figure S1B**) phylogenetic trees were created, respectively, based on the ClustalX and MAFFT alignments of these 115 AP2/ERF candidates and 146 AP2/ERF proteins from *A. thaliana*. Both phylogenies show the same grouping of the AP2/ERF superfamily in *F. vesca*. According to these phylogenies, as well as their domain compositions, 91 proteins were classified as *F. vesca* ERFs (FveERFs), and the other 24 proteins were grouped to the AP2, RAV families or soloists (**Table S2**).

Chromosomal location analysis demonstrates that, except 2 *FveERF* genes found within unanchored chromosome sequences, the other 89 *FveERF*s are unevenly distributed among the seven *F. vesca* chromosomes (**Figure 1**). The number of *FveERF* genes on each chromosome has little relationship with chromosome length (correlation coefficient = 0.24), but is positively correlated with the number of tandem-arrayed *FveERF*s (correlation coefficient = 0.90). For example, LG5 and LG7, the two chromosomes with the largest numbers of *FveERF* genes (20 and 17, respectively), also contain the largest numbers of tandem *FveERF*s (13 and 9, respectively), whereas LG1 has the least number of *FveERF* genes (five) and has no tandem ones. This indicates that the uneven distribution of *FveERF*s is mainly due to the location of their tandem members. In total, 38.5% (35/91) of *FveERF* genes are arrayed in tandem repeats, strongly suggesting that a high proportion of *FveERF* genes are derived from tandem duplication events.

Expansion of the FveERF Gene Family

To study the relationships among *FveERF* genes, phylogenetic trees were constructed based on the ClustalX and MAFFT alignments of full-length FveERF and AtERF protein sequences, using ML (**Figure 2**) and aBayes (**Figure S2**) methods, respectively. All the phylogenies display similar grouping of the FveERF gene family,

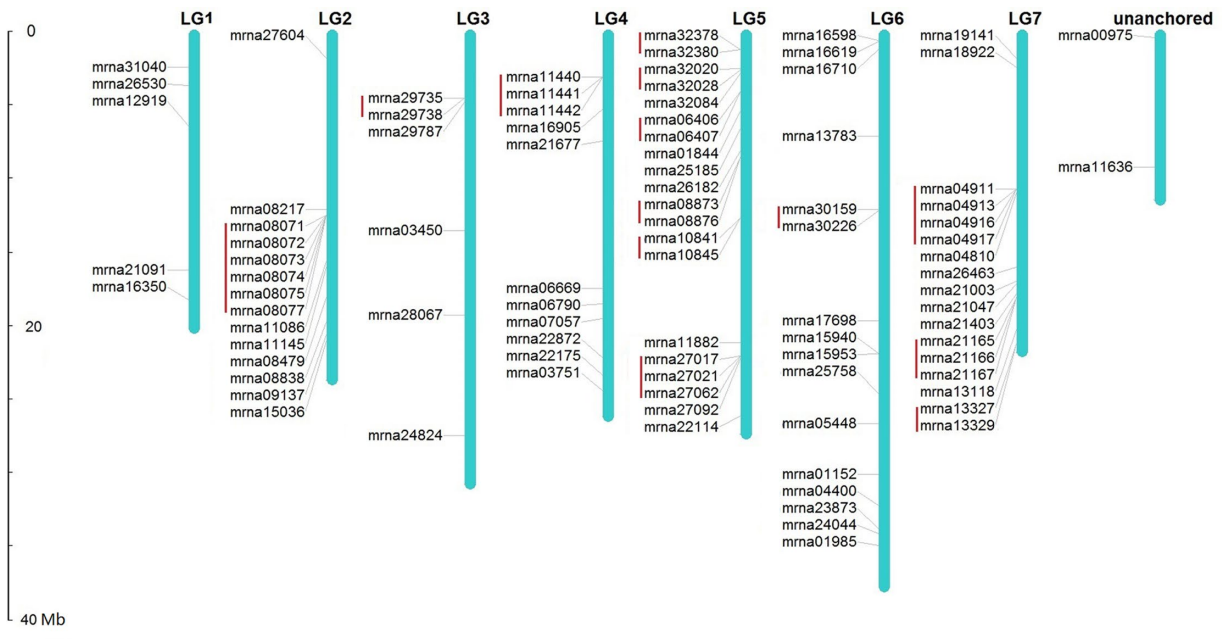


FIGURE 1 | Locations of *FveERF* genes on the *Fragaria vesca* chromosomes. The size of a chromosome is indicated by its relative length. Tandemly duplicated genes are indicated with a red bar.

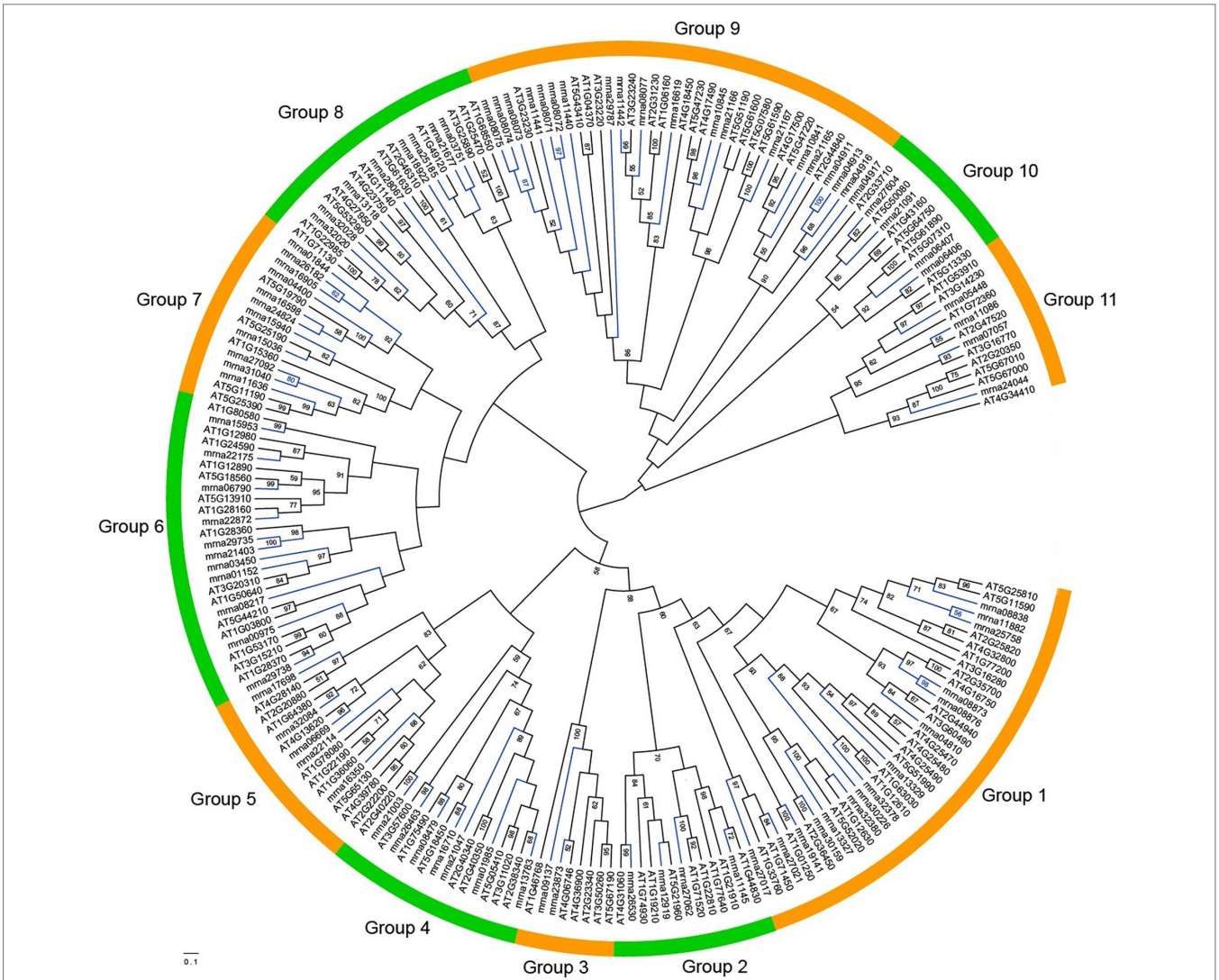


FIGURE 2 | Maximum-likelihood phylogeny of the ERF proteins from *Fragaria vesca* and *Arabidopsis thaliana*. The phylogeny was constructed based on the amino acid sequences of full-length FveERF and AtERF proteins with 100 bootstrapping replicates. Bootstrap values greater than 50 are indicated on the nodes. Green and orange arcs indicate different groups of ERF proteins. Blue and black branches represent FveERF and AtERF proteins, respectively.

which is generally in consistence with the classification of *Arabidopsis ERF* genes (Nakano et al., 2006; **Table S2**). We further classified the *FveERF* genes of the 11 groups (groups 1–11) into two types: I) *FveERF*s that form phylogenetic clusters with other *FveERF*s and II) those that do not form clusters with other *FveERF*s but group with *AtERF* or *AtERF* and *FveERF* gene branch(es) (**Table S3**). The clustering of the type I *FveERF*s is likely a result of lineage-specific expansions of these genes in *F. vesca*. In contrast, type II *FveERF* genes are likely direct descendants of the ancestral genes in the common ancestor of *A. thaliana* and *F. vesca* and remain as single copies in the *F. vesca* genome. Among the 91 *FveERF*s, 24 genes, which form 10 phylogenetic clusters, belong to type I, and the remaining 67 genes belong to type II. This suggests that about one quarter of the *FveERF*s are involved in the expansions specific to the *Fragaria* lineage, while the rest three quarters likely have not expanded following the split of *Arabidopsis* and *Fragaria* lineages.

Chromosome location of the type I *FveERF*s shows that 11 (45.8%) of the 24 lineage-specific expanded *FveERF* genes are arrayed in tandem with their phylogenetically clustered genes. For instance, mrna08071–mrna08075 that form two clusters in group 9 of the phylogeny (mrna08071 and mrna08072 for one cluster and mrna08073–mrna08075 for another, **Figures 2** and **S2**) are located in a six-gene tandem repeat on chromosome 2 (**Figure 1**). These genes are likely derived from tandem duplications, and are hereafter referred to as lineage-specific tandem *FveERF*s. However, not all the type I *FveERF*s located in tandem repeat are lineage-specific tandem *FveERF*s. For instance, the type I gene mrna29735 is phylogenetically clustered with mrna21403 (**Figures 2** and **S2**) but is arrayed in tandem with mrna29738 (**Figure 1**). The relationship among these three genes suggests that a tandem duplication gave rise to the gene pair mrna29735 and mrna29738 rather than the lineage-specific gene pair of mrna29735 and

mrna21403. The MCScanX analysis indicates that, among the twelve non-tandem type I *FveERF*s, seven genes including mrna21403 likely are derived from dispersed duplications, while the rest five are likely from segmental duplications (**Table S3**). Collectively, tandem duplication is the major mechanism for the lineage-specific expansion of the *FveERF* gene family.

In addition to the type I lineage-specific tandem *FveERFs,* 23 (34.3%) of the 67 type II *FveERF* genes that have not undergone lineage-specific expansion also reside in tandem repeats on chromosomes (**Figure 1**, **Table S3**). For example, the type II *FveERF*s mrna10841 and mrna10845 in group 9 are located in a two-gene tandem repeat on chromosome 2. Interestingly, their phylogenetically clustered *AtERF* orthologs (AT5G47220 and AT4G17500 for mrna10841; AT5G47230 and AT4G17490 for mrna10845, **Figures 2** and **S2**) are also arrayed in tandem on *A. thaliana* chromosomes (AT5G47220 and AT5G47230; AT4G17500 and AT4G17490). Therefore, it is very likely that mrna10841 and mrna10845 are derived from ancestral tandem duplications in the most recent common ancestor of *A. thaliana* and *F. vesca* and are maintained in tandem following the split of the two lineages.

There are a total of 15 tandem type II *FveERF* genes having tandem *AtERF* orthologs (**Figures 2** and **S2**, **Table S3**), indicating they are derived from ancestral tandem repeats. Among them, two genes are tandemly arrayed with type I *FveERF* genes, i.e., mrna29738 tandem with mrna29735, and mrna08077 tandem with mrna08071–mrna08075 (**Figure 1**). This suggests that these type I genes are involved in both ancestral and lineage-specific tandem duplications. On the other hand, the rest 10 tandem type II *FveERF*s are phylogenetically clustered with their *AtERF* orthologs which are not arrayed in tandem. We still considered these 10 *FveERF*s to originate from ancestral tandem duplications, because the *A. thaliana* genome has undergone extensive chromosomal rearrangements (del Pozo and Ramirez-Parra, 2015) which would lead to non-tandem arrangements of *AtERF* orthologs. Therefore, at least 34.1% (31 of all 91) *FveERF* genes can be classified into ancestral tandem *FveERF*s.

Taken together, we define the tandem *FveERF* genes that cluster with each other in the phylogenies as lineage-specific tandem *FveERF*s, while the tandem *FveERF*s phylogenetically clustering with their *AtERF* orthologs or retaining in singletons as ancestral tandem *FveERF*s. From the above analyses, the total 35 tandem *FveERF*s include 11 lineage-specific ones and 29 ancestral ones, with 5 belonging to both.

Motif and Gene Structures of *FveERF* Genes

We analyzed motif structures of the FveERF proteins, with 15 conserved motifs (motifs 1–15) identified using MEME suite

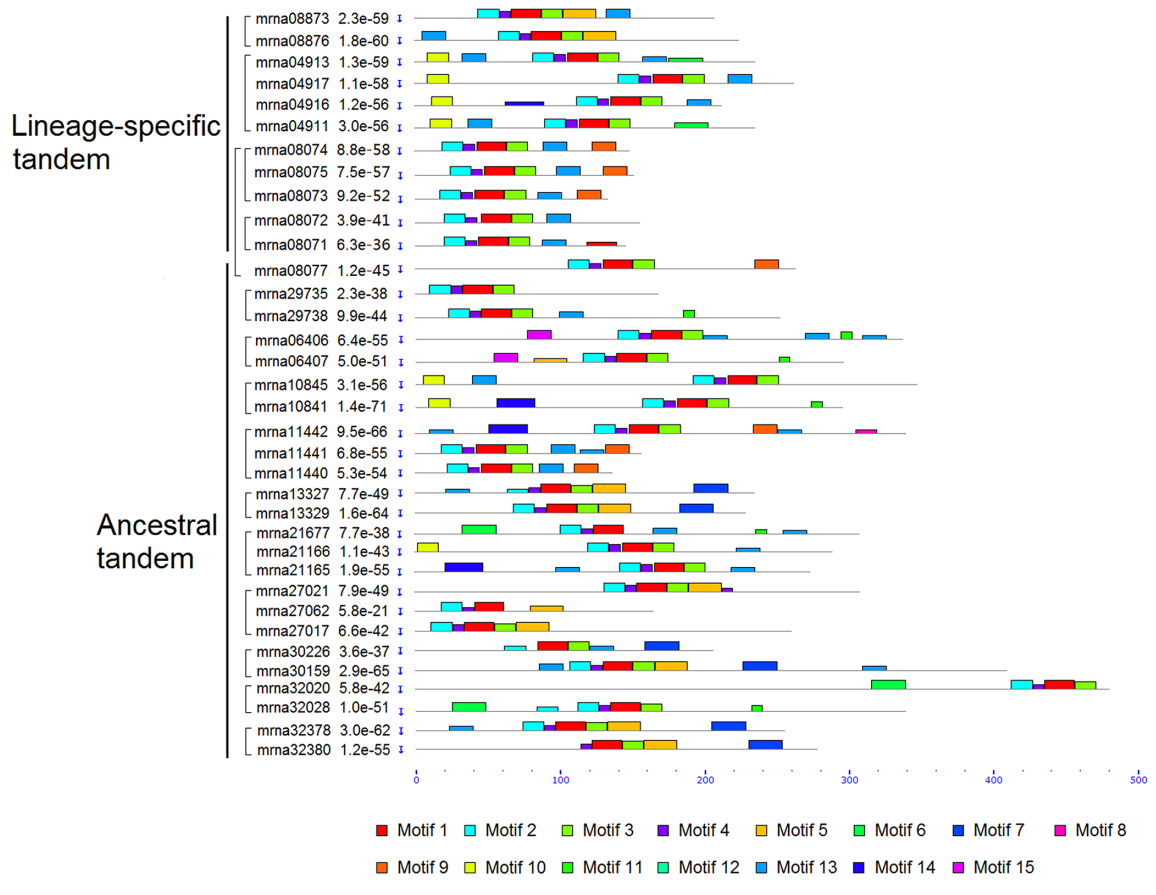


FIGURE 3 | Schematic diagram of amino acid motifs of tandem FveERF proteins. Motif analysis was performed using MEME5.0.1 as described in the *Materials and Methods*. Proteins whose genes located in the same tandem repeat are grouped together. Mrna08077 forms an ancestral tandem repeat with mrna08071–mrna08075.

(**Figures 3**, **S3**, and **S4**). Motifs 1–4 correspond to the AP2/ERF domain and have been identified in nearly all FveERF proteins. The four lineage-specific tandem FveERF pairs show differences in the arrangement of zero to four motifs (an average of 1.75), where totally only four motifs have been differentially identified. In contrast, the ancestral tandem pairs have differences in 1–6 (an average of 3.26) of the 12 differentially distributed motifs, which include motifs 1–4 that are key to the AP2/ERF domain. The average number of *FveERFs* in an ancestral or lineage-specific tandem repeat is similar (2.5 for ancestral vs. 2.75 for lineage-specific tandem repeat). However, the motif analysis demonstrates that the protein structure of the ancestral tandem *FveERFs* is more divergent than that of the lineage-specific tandem ones.

With respect to gene structure, 24 (26.4%) of the 91 *FveERF* genes possess introns (**Table S3**). The average number of introns per intron-containing *FveERF* is 1.83. Around half of these genes (13 of 24) contain a single intron, with others contain two to three except for one that contains eight. These intron-containing *FveERF*s are located on chromosomes 1–6 as well as the unanchored scaffold (**Table S3**). None are found on chromosome 7, which houses the second-most (17) *FveERF* genes. All genes within the four lineage-specific tandem *FveERF* pairs have same numbers of introns with their counterparts, whereas in about half of the 11 ancestral tandem *FveERF* pairs exon/intron structures are different, indicating that the gene structures of ancestral tandem *FveERF*s have diverged.

Expression Profiles of *FveERF* Genes in Flowers and Fruits

To investigate the expression profiles of *FveERF* genes, we downloaded and analyzed the transcriptomic data of *F. vesca* flowers and early fruits (Hollender et al., 2012; Darwish et al., 2013; Kang et al., 2013). All the *FveERF* genes have RPKM values larger than 0.3 in at least two flower-development stages (**Figure 4**); thus, we consider all *FveERF*s to be expressed during flower development in *F. vesca* (see *Materials and Methods*). In contrast, RPKM values for 18 (19.8%) *FveERF*s are lower than 0.3 throughout early-stage fruit development. The expression levels of *FveERF*s in tissues of flowers and early fruits (**Figure S5**) are similar to those in the stages. These results indicate that most, if not all, *FveERF* genes are involved in flower development, whereas ~20% of *FveERF*s may not participate during early-stage fruit development.

The expression levels of tandem or phylogenetically clustered genes are significantly different from those of the non-tandem/clustered *FveERF*s (all $p < 0.001$ from *t*-test). Moreover, among the 33 *FveERF*s with low expression levels (RPKM values <1 in at least two thirds of the 13 stages of reproductive development, **Figure 4**), 81.8% (27) either cluster on the phylogeny or are arrayed in tandem on chromosomes. Meanwhile, 60.1% of the 47 tandem or clustered *FveERFs* have low expression levels, 4.4-fold higher than the percentage of low-expression genes among the other 44 *FveERF*s (13.6%). This percentage increases to 81.8% (9 of 11) for lineage-specific tandem *FveERF*s, 0.8-fold higher than for ancestral tandem *FveERF*s (45.8%). Consistently, the expression levels of lineage-specific tandem *FveERF*s are also significantly lower than

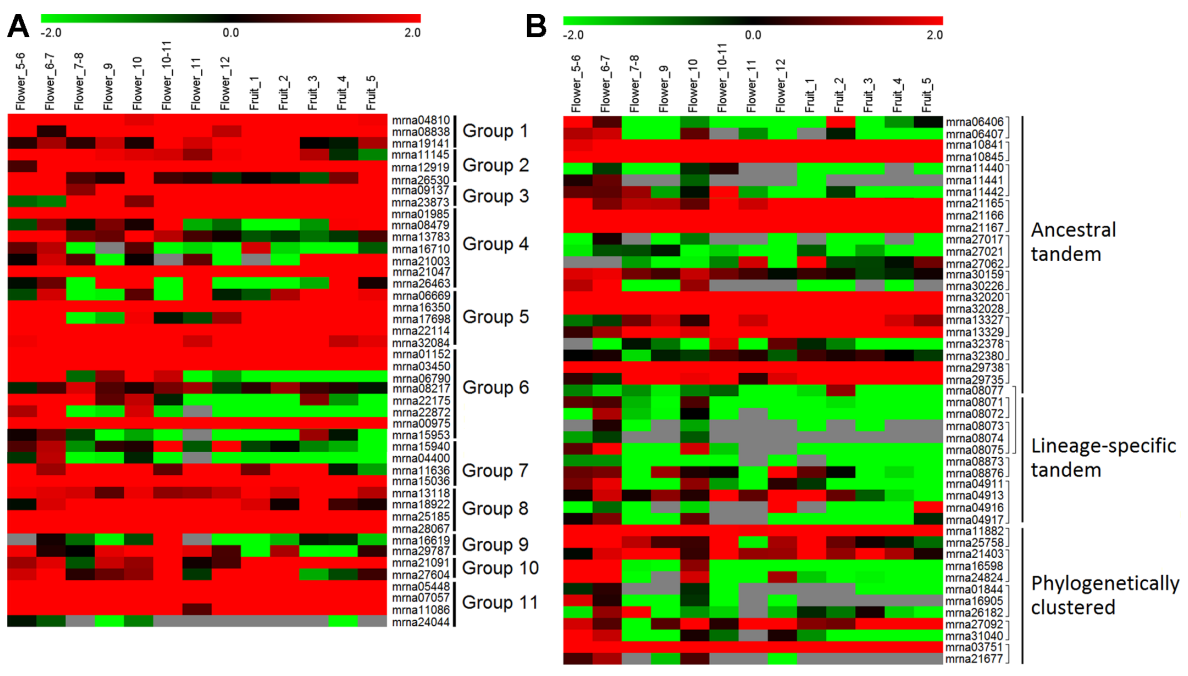


FIGURE 4 | Expression profiles of *FveERF* genes in different stages of *Fragaria vesca* flowers and early-stage fruits. **(A** and **B)** The mRNA levels of the non-tandem **(A)** and tandem/phylogenetically clustered **(B)** *FveERF* genes. Genes located in the same tandem repeat or in a phylogenetic cluster are grouped together. Mrna08077 forms an ancestral tandem repeat with mrna08071–mrna08075. Mrna21403 forms a phylogenetic cluster with mrna29735. Data were retrieved from http://bioinformatics.towson.edu/strawberry/ (Hollender et al., 2012; Darwish et al., 2013; Kang et al., 2013). Expression levels were calculated in the log2 scale. For a detailed description of the stages, please see http://bioinformatics.towson.edu/strawberry/newpage/Tissue_Description.aspx.

those of the ancestral ones ($p < 0.001$). These results demonstrate that the expression levels of tandem or clustered *FveERF*s are lower than those of the other *FveERF*s during reproductive development, with lineage-specific tandem *FveERF*s having the lowest expression.

The expression patterns of tandem *FveERF* pairs are less diversified than those of the non-tandem ones in a same group in flowers and early fruits (**Figures 4** and **S5**). More than 75% non-tandem *FveERF* gene pairs in a group show diversified expression patterns (data not shown), while approximately 50% of tandem *FveERF* pairs have positive correlated expression patterns (correlation coefficient >0.5, **Table S4**). Further, this percentage is nearly the same for both the ancestral tandem *FveERF* gene pairs and the lineage-specific tandem ones. This suggests that the expression patterns of ancestral and lineage-specific tandem *FveERF* duplicates diverge to similar degrees in flowers and early-stage fruits, regardless of the increased age and evolutionary history of ancestral duplicates.

We further investigated expression patterns of the ancestral and lineage-specific tandem *FveERF*s (see *Materials and Methods* for the selection of the tandem *FveERF*s) during the fruit-ripening stages of *F. vesca* using qRT-PCR (**Figure 5A**). The five lineage-specific tandem *FveERF*s have very low expression ($< 1 \times 10^{-4}$ when using *FveGAPDH* as the reference gene) throughout the ripening stages. Five of the nine ancestral tandem *FveERF*s have no detectable expression during these stages, whereas the remaining four (found within two tandem repeats) exhibit much higher expression (**Figure 5B**). These expression patterns are roughly in accordance with the expression patterns for *FveERF*s in early fruits (**Figures 4** and **5B**). Therefore, the tandem *FveERF* genes are most likely consistently expressed throughout fruit development and ripening stages.

Expression of Tandem Duplicated *FveERF* Genes Under Drought/Cold Stress

ERF transcription factors play important roles in abiotic stress response (Lata and Prasad, 2011). We treated the *F. vesca* seedlings with either cold or drought stress, and characterized the expression of nine lineage-specific and nine ancestral tandem *FveERF*s (see *Materials and Methods* for the selection of the tandem *FveERF*s, **Figure 6**). Similar to in fruits, lineage-specific tandem *FveERF*s have very low expression levels in *F.*

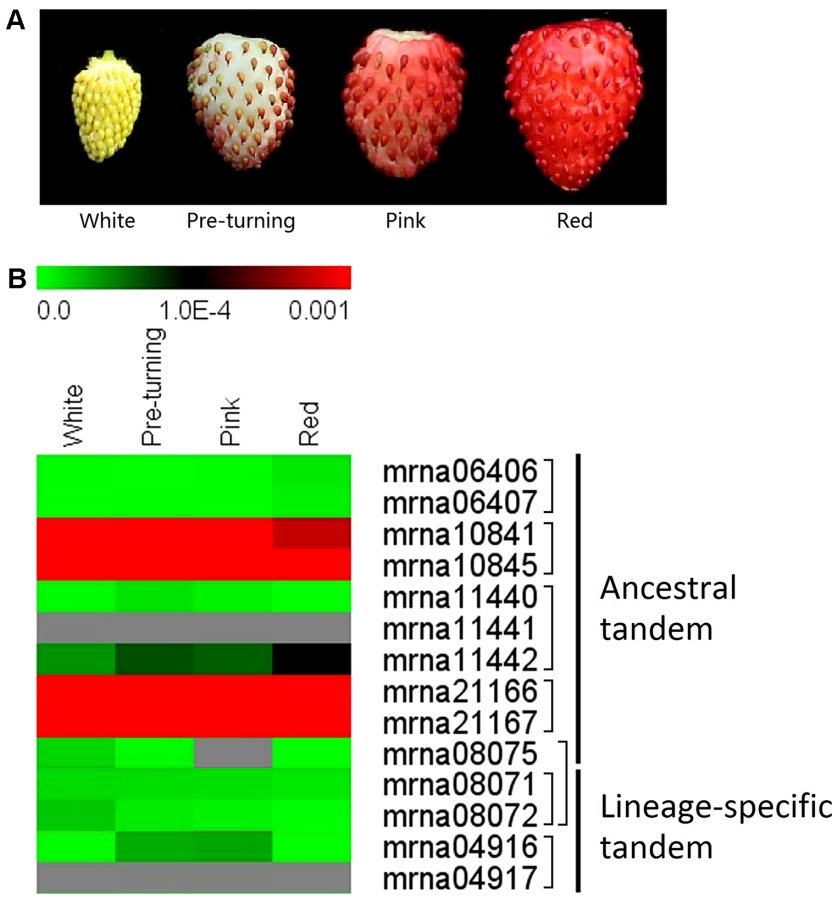


FIGURE 5 | Expression profiles of tandem *FveERF* genes during fruit ripening. **(A)** The schematic diagram for the four stages of fleshy fruits investigated in B. **(B)** The expression levels of tandem *FveERF* genes relative to *GAPDH*, measured by quantitative RT-PCR and displayed in the log2 scale. Genes located in the same tandem repeat are grouped together. Mrna08075 forms an ancestral tandem repeat with mrna08071 and mrna08072. Three biological replicates and three technical replicates were obtained for each data point.

vesca seedlings, regardless of treatment (**Table S5**). Six of the nine lineage-specific tandem *FveERF*s (mrna04911, mrna04913, mrna08071, mrna08072, mrna08075 and mrna08876) have no detectable gene expression under either or both stresses, while only one ancestral tandem *FveERF* (mrna11440) is undetectable under drought stress. Further, among the expressed *FveERF*s, the average expression level of ancestral tandem ones is approximately 100-fold higher than that of the lineage-specific tandem ones (**Figure 5B**). These results suggest that *FveERF* genes generated by recent tandem duplications may generally have low expression levels.

We have observed that the ancestral tandem *FveERF* pairs are differentially expressed following stress treatment, cold or drought (**Figure 6** and **Table S6**). The ancestral tandem pair of mrna11440, mrna11441, and mrna11442 displays divergent expression patterns following both stress treatments, while the other three pairs are only differentially expressed following either cold or drought stress. In contrast, all lineage-specific tandem *FveERF* pairs exhibit similar stress-response expression patterns, except for mrna04913 (compared to mrna04916 or mrna04917) following dehydration (**Figure 6** and **Table S6**). Based on these data, *FveERF* duplicates from ancestral tandem duplications seem to have diverged in their responses to abiotic stress, whereas most lineage-specific tandem genes have not.

DISCUSSION

This is the first study identifying *ERF* genes in woodland strawberry (*F. vesca*). A total of 91 *FveERF*s have been identified and divided into 11 groups based on phylogenetic and motif analyses. The percentage of *ERF* genes in total protein-coding genes in *F. vesca* (0.28%, **Figure 7**) is similar to the percentages found in two other Rosaceae family plants, plum [*Prunus mume*, 0.29% (Du et al., 2013)] and apple [*Malus* × *domestica*, 0.31% (Zhuang et al., 2011)], but lower than those in Brassicaceae family species, such as *A. thaliana* [0.44% (Nakano et al., 2006)] and *Brassica rapa* [0.58% (Song et al., 2013)]. The higher percentage of *AtERF* genes is likely a result of the polyploidization events during the evolution of *A. thaliana*, as 75% of them are proposed to have been preferentially retained after WGDs (Nakano et al., 2006). As being transcription factor genes, *ERF*s would have been retained at a higher than average level after WGD, but not after tandem duplication (Panchy et al., 2016). However, the apple genome that has undergone a recent WGD event does not contain higher percentage of *ERF* genes than *F. vesca*. Our results demonstrate that more *FveERF* genes are involved in tandem duplication than in WGD/segmental duplication, suggesting that tandem duplication is the major mechanism contributing to the expansion of the *FveERF* gene family.

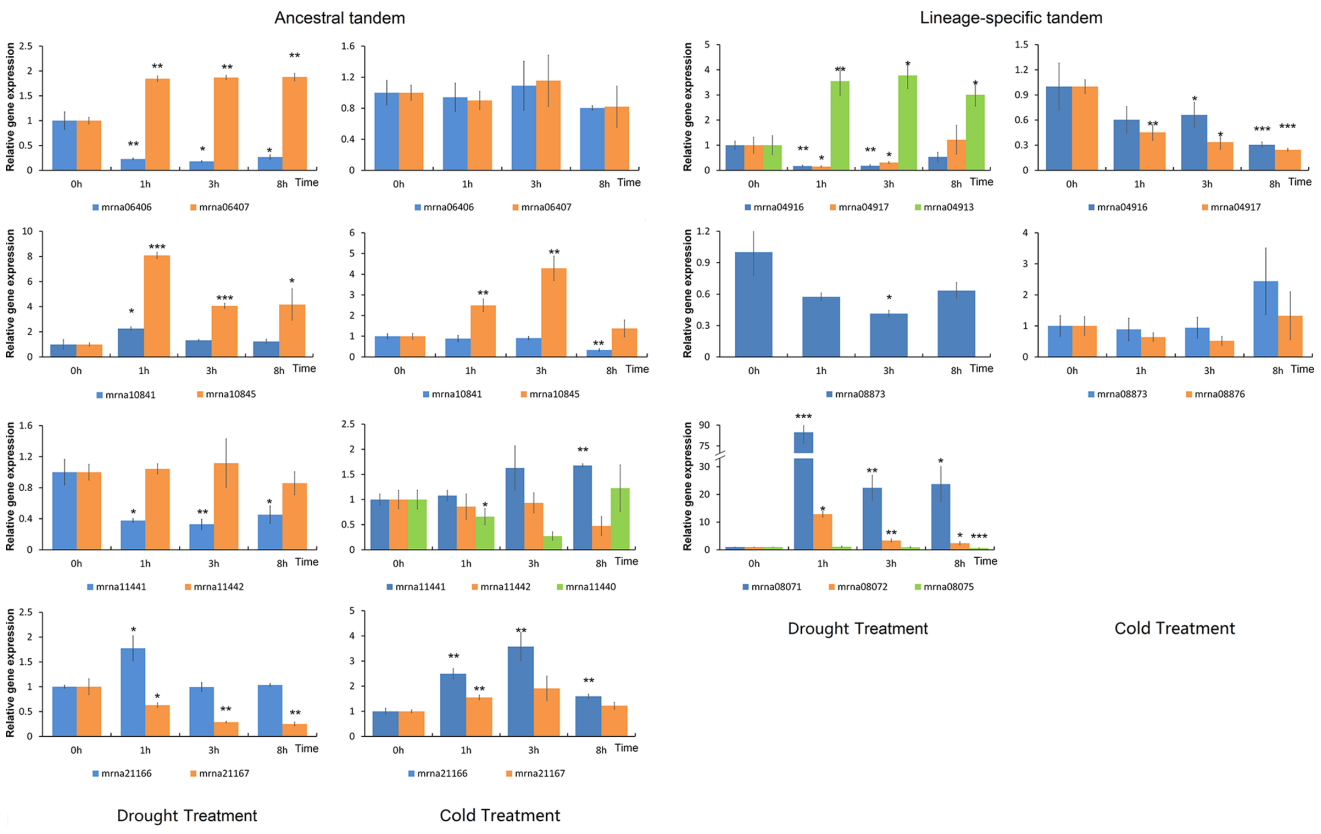
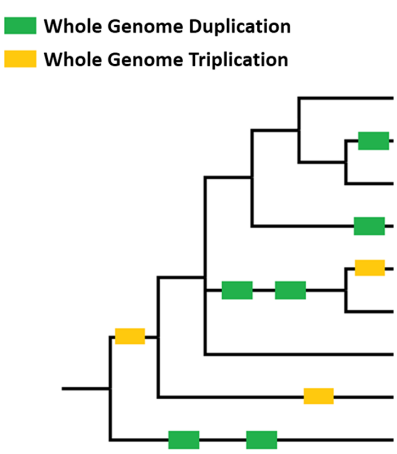


FIGURE 6 | Expression profiles of *FveERF* genes in response to drought and cold. The expression levels relative to *GAPDH* were measured by quantitative RT-PCR. Three biological replicates and three technical replicates were obtained for each data point. Asterisks above the error bars indicate significant differences between the treated and untreated (0h) samples (*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$). Mrna08075 forms an ancestral tandem repeat with mrna08071 and mrna08072. The genes with expression levels lower than 1×10^{-4} at most time points of the treatment are not shown.



Species	Gene number	ERF			ERF%
		Tandem	Total	Tandem%	
Fragaria vesca	32,831	35	91	38%	0.28%
Malus domestica	63,514	67	195	34%	0.31%
Prunus mume	31,390	38	90	42%	0.29%
Populus trichocarpa	41,335	51	168	30%	0.41%
Brassica rapa	40,492	35	236	15%	0.58%
Arabidopsis thaliana	27,416	24	121	20%	0.44%
Vitis vinifera	26,346	47	107	44%	0.41%
Solanum tuberosum	39,031	24	155	15%	0.40%
Oryza sativa	39,049	22	139	16%	0.36%

FIGURE 7 | Percentages of tandem *ERF* genes in the nine species investigated. ERF% shows the percentage of *ERF* genes in the total gene set. The Taxonomy Common Tree constructed online by Taxonomy Browser in the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi) is on the left. The branch length is not proportional to the evolutionary time. Green box, whole-genome duplication; yellow box, whole-genome triplication.

The percentage of tandem *FveERF*s in total *FveERF*s is similar to that of *PmuERF*s in plum and of *VvERF*s in grapevine (*Vitis vinifera*), a little higher than that in apple and poplar (*Populus trichocarpa*), and much higher than that in *A. thaliana* and *B. rapa* (**Figure 7**). *F. vesca*, plum and grapevine have not undergone any WGDs after the triplication event (γ) probably shared by all core eudicots (Bowers et al., 2003; Jaillon et al., 2007; Cenci et al., 2010), while apple and poplar have undergone WGD once (Tuskan et al., 2006; Velasco et al., 2010) and *A. thaliana* and *B. rapa* have undergone WGD at least twice (Bowers et al., 2003). Therefore, the percentage of tandem *ERF* genes retained seems to be negatively correlated with occurrences of the polyploidization events, possibly because of the rearrangement of chromosomal sequences after WGD.

The higher percentage of tandem *ERF* genes in *F. vesca* than in *A. thaliana* is mainly due to a greater number of ancestral tandem *ERF*s (31 vs. 17), rather than lineage-specific tandem ones (11 vs. 11). Further, all ancestral tandem *AtERF*s have tandem *FveERF* orthologs, whereas there are 10 ancestral tandem *FveERF*s whose *AtERF* orthologs are not arrayed in tandem. This number difference of tandem orthologs suggests that the more ancestral tandem *ERF* genes in *F. vesca* than in *A. thaliana* are due to more rearrangements or losses of the ancestral tandem *AtERF*s. Extensive rearrangement and loss of chromosomal segments have occurred in *A. thaliana* during its rediploidization after polyploidization (del Pozo and Ramirez-Parra, 2015). Ancestral tandem *AtERF*s are defined as those derived from tandem duplications in the common ancestor of *F. vesca* and *A. thaliana*, which occurred prior to the twice polyploidization of the *Arabidopsis* lineage. Hence, the ancestral tandem *AtERF*s have experienced at least once rediploidization, leading to the number difference of ancestral tandem *ERF* genes between *F. vesca* and *A. thaliana*. Altogether, genomic rearrangement during rediploidization following polyploidization is an important factor affecting the retention of ancestral tandem *ERF* genes. The higher retention of tandem *FveERF*s than tandem *AtERF*s may be largely attributed to no polyploidization occurred in *F. vesca* after the divergence of core eudicots.

The discrimination of ancestral and lineage-specific tandem *FveERF* genes provides us with a good tool to compare the divergence of tandem *FveERF* duplicates generated at different times. As expected, the average values of pairwise nucleotide divergence, synonymous nucleotide substitutions per synonymous site (Ks), and non-synonymous substitutions per nonsynonymous site (Ka) between lineage-specific tandem *FveERF* pairs are significantly lower than those between ancestral tandem *FveERF* pairs, respectively (**Table S7**). Moreover, lineage-specific tandem *FveERF* genes maintain higher similarities of exon/intron and motif structures than the ancestral tandem ones. These results indicate that sequence and structure divergences of ancestral tandem *FveERF*s are higher than those of lineage-specific tandem *FveERF*s. None of the ancestral tandem *AtERF*s contain an intron (Nakano et al., 2006). In contrast, 35.5% (11 of 37) ancestral tandem *FveERF*s have an average number of 2.36 introns. Particularly, half of ancestral tandem *FveERF* pairs show variable exon/intron structures. Thus, it seems that intron gain/loss has occurred more frequently in the evolutionary histories of *FveERF* genes compared to *AtERF*s, which may play a role in the divergence of *FveERF*s, especially for ancestral tandem ones.

Tandem duplicates are proposed to have higher expression correlation than the duplicates derived from most of the other mechanisms (Wang et al., 2012b). However, our analyses show that the expression correlation of lineage-specific tandem *FveERF*s in flowers and fruits is lower than that of other lineage-specific expanded *FveERF*s, but is similar to that of the ancestral tandem ones (**Table S4**). The studies on expression patterns of tandem duplicates in other families, such as the C_2H_2 zinc-finger gene family in rice (Agarwal et al., 2007) and the phosphatidylethanolamine binding protein (PEBP) family

in soybean (Wang et al., 2015), also demonstrate that ancestral and lineage-specific tandem duplicates have similarly highly diversified expression patterns in developmental tissues. These results support that expression of tandem *FveERF* duplicates in reproductive development has diverged shortly after duplication.

Previous studies have suggested that expression divergence of the tandem duplicates has little relationship with their Ks values (Ganko et al., 2007), mainly based on expression analyses in developmental tissues/organs. Our results with respect to tandem *FveERF* expression in reproductive development are consistent with this suggestion. However, the results under stressed conditions show different patterns. All expressed lineage-specific tandem *FveERF* duplicates exhibit same response patterns upon drought or cold treatment with only one exception, whereas the ancestral ones diverge at a much higher level (**Table S6**). This suggest that expression divergence of tandem *FveERF*s under stress may have occurred later, but evolved faster, than in reproductive development. In addition to growth and development, *ERF*s are also important in the regulation of abiotic stress responses in plants (Lata and Prasad, 2011). Although the roles of the sampled tandem *FveERF*s in abiotic stress responses have not been revealed so far, the *A. thaliana* groups containing their *AtERF* orthologs have been shown with functions in tolerance to abiotic stress. Moreover, the tandem *FveERF*s show induced or reduced expression after drought and cold treatments, supporting that they likely play roles in the responses to these stresses. Therefore, the high expression divergence of the ancestral tandem *FveERF*s under stress conditions could contribute to the responses of *F. vesca* to abiotic stresses.

Besides, with respect to expression levels, no matter under stress conditions or in reproductive development, high proportions of lineage-specific tandem *FveERF* pairs are undetectable. Comparatively, all ancestral tandem *FveERF* pairs, at least one of the members, are expressed at much higher levels. Expression levels of the ancestors of the undetectable lineage-specific tandem *FveERF*s are unknown; analyses on their orthologs in *A. thaliana* and other plants may provide indication that whether recent tandem duplication is a main cause of such low expression levels of these lineage-specific tandem *FveERF* pairs. On the other hand, like in expression patterns, the divergence in expression levels of the expressed lineage-specific tandem *FveERF*s is at similar levels with the ancestral tandem ones in flower and fruit stages, but lower under abiotic stressed conditions (**Table S5**). Thus, the expression divergence of tandem *FveERF* duplicates is probably slower under stress conditions than in reproductive development at early stage after the duplication.

CONCLUSIONS

In this study, the ERF gene family in *F. vesca* was identified and analyzed, especially for their tandem members. Compared with ancestral tandem *FveERF*s, the lineage-specific tandem *FveERF*s are more conserved in sequence, structure, and expression under abiotic stress, whereas are similarly highly diversified in expression during reproductive development. These results suggest that the retention of tandem *FveERF* duplicates soon after their duplication may be related to their divergence in the regulation of reproductive development. On the other hand, their further divergence in response patterns to abiotic stresses likely contributes to stress responses of *F. vesca*. This provides new insights into the expression divergence between tandem duplicates in plants.

AUTHOR CONTRIBUTIONS

YL and JD designed the experiments. XW and SL performed the experiments and data analyses. DL and QW participated in data analyses. JD, XW, and SL wrote the manuscript. YL and RM revised the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China Grants 31471860 (to JD and YL), and A Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (to YL).

ACKNOWLEDGMENTS

We thank Dr. Janet Slovin (United States Department of Agriculture, USDA) for providing seeds of *Fragaria vesca* "Ruegen." We also thank members of the YL laboratory for discussions and comments on the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00805/full#supplementary-material

REFERENCES

Agarwal, P., Arora, R., Ray, S., Singh, A. K., Singh, V. P., Takatsuji, H., et al. (2007). Genome-wide identification of C2H2 zinc-finger gene family in rice and their phylogeny and expression analysis. *Plant Mol. Biol.* 65, 467–485. doi: 10.1007/s11103-007-9199-y

Agarwal, P. K., Agarwal, P., Reddy, M. K., and Sopory, S. K. (2006). Role of DREB transcription factors in abiotic and biotic stress tolerance in plants. *Plant Cell Rep.* 25, 1263–1274. doi: 10.1007/s00299-006-0204-8

Amil-Ruiz, F., Garrido-Gala, J., Blanco-Portales, R., Folta, K. M., Muñoz-Blanco, J., and Caballero, J. L. (2013). Identification and validation of reference genes for transcript normalization in strawberry (*Fragaria* × *ananassa*) defense responses. *PLoS ONE* 8, e70603. doi: 10.1371/journal.pone.0070603

Bailey, T. L., Johnson, J., Grant, C. E., and Noble, W. S. (2015). The MEME suite. *Nucleic Acids Res.* 43, W39–W49. doi: 10.1093/nar/gkv416

Banno, H., Ikeda, Y., Niu, Q. W., and Chua, N. H. (2001). Overexpression of Arabidopsis ESR1 induces initiation of shoot regeneration. *Plant Cell* 13, 2609–2618. doi: 10.1105/tpc.010234

Bowers, J. E., Chapman, B. A., Rong, J., and Paterson, A. H. (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422, 433–438. doi: 10.1038/nature01521

Brown, R. L., Kazan, K., McGrath, K. C., Maclean, D. J., and Manners, J. M. (2003). A role for the GCC-box in jasmonate-mediated activation of the *PDF1.2* gene of Arabidopsis. *Plant Physiol.* 132, 1020–1032. doi: 10.1104/pp.102.017814

Carretero-Paulet, L., and Fares, M. A. (2012). Evolutionary dynamics and functional specialization of plant paralogs formed by whole and small-scale genome duplications. *Mol. Biol. Evol.* 29, 3541–3551. doi: 10.1093/molbev/mss162

Cenci, A., Combes, M. C., and Lashermes, P. (2010). Comparative sequence analyses indicate that Coffea (Asterids) and Vitis (Rosids) derive from the same paleo-hexaploid ancestral genome. *Mol. Genet. Genomics* 283, 493–501. doi: 10.1007/s00438-010-0534-7

Chakravarthy, S., Tuori, R. P., D'Ascenzo, M. D., Fobert, P. R., Despres, C., and Martin, G. B. (2003). The tomato transcription factor Pti4 regulates defense-related gene expression via GCC box and non-GCC box cis elements. *Plant Cell* 15, 3033–3050. doi: 10.1105/tpc.017574

Charfeddine, M., Saïdi, M. N., Charfeddine, S., Hammami, A., and Gargouri Bouzid, R. (2015). Genome-wide analysis and expression profiling of the ERF transcription factor family in potato (*Solanum tuberosum* L.). *Mol. Biotechnol.* 57, 348–358. doi: 10.1007/s12033-014-9828-z

Chen, G., Hu, Z., and Grierson, D. (2008). Differential regulation of tomato ethylene responsive factor LeERF3b, a putative repressor, and the activator Pti4 in ripening mutants and in response to environmental stresses. *J. Plant Physiol.* 165, 662–670. doi: 10.1016/j.jplph.2007.03.006

Chen, J. Q., Meng, X. P., Zhang, Y., Xia, M., and Wang, X. P. (2008). Over-expression of OsDREB genes lead to enhanced drought tolerance in rice. *Biotechnol. Lett.* 30, 2191–2198. doi: 10.1007/s10529-008-9811-5

Darwish, O., Slovin, J. P., Kang, C., Hollender, C. A., Geretz, A., Houston, S., et al. (2013). SGR: an online genomic resource for the woodland strawberry. *BMC Plant Biol.* 13, 223. doi: 10.1186/1471-2229-13-223

del Pozo, J. C., and Ramirez-Parra, E. (2015). Whole genome duplications in plants: an overview from Arabidopsis. *J. Exp. Bot.* 66, 6991–7003. doi: 10.1093/jxb/erv432

Du, D., Hao, R., Cheng, T., Pan, H., Yang, W., Wang, J., et al. (2013). Genome-wide analysis of the AP2/ERF gene family in *Prunus mume*. *Plant Mol. Biol. Rep.* 31, 741–750. doi: 10.1007/s11105-012-0531-6

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics* 14, 755–763. doi: 10.1093/bioinformatics/14.9.755

Fan, C., Chen, Y., and Long, M. (2008). Recurrent tandem gene duplication gave rise to functionally divergent genes in Drosophila. *Mol. Biol. Evol.* 25, 1451–1458. doi: 10.1093/molbev/msn089

Figueiredo, D. D., Barros, P. M., Cordeiro, A. M., Serra, T. S., Lourenço, T., Chander, S., et al. (2012). Seven zinc-finger transcription factors are novel regulators of the stress responsive gene *OsDREB1B*. *J. Exp. Bot.* 63, 3643–3656. doi: 10.1093/jxb/ers035

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., et al. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 44, 279–285. doi: 10.1093/nar/gkv1344

Fortna, A., Kim, Y., MacLaren, E., Marshall, K., Hahn, G., Meltesen, L., et al. (2004). Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS Biol.* 2, E207. doi: 10.1371/journal.pbio.0020207

Ganko, E. W., Meyers, B. C., and Vision, T. J. (2007). Divergence in expression between duplicated genes in Arabidopsis. *Mol. Biol. Evol.* 24, 2298–2309. doi: 10.1093/molbev/msm158

Golldack, D., Luking, I., and Yang, O. (2011). Plant tolerance to drought and salinity: stress regulating transcription factors and their functional significance in the cellular transcriptional network. *Plant Cell Rep.* 30, 1383–1391. doi: 10.1007/s00299-011-1068-0

Gu, T., Ren, S., Wang, Y., Han, Y., and Li, Y. (2016). Characterization of DNA methyltransferase and demethylase genes in *Fragaria vesca*. *Mol. Genet. Genomics* 291, 1333–1345. doi: 10.1007/s00438-016-1187-y

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321. doi: 10.1093/sysbio/syq010

Hanada, K., Zou, C., Lehti-Shiu, M., Shinozaki, K., and Shiu, S. H. (2008). Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. *Plant Physiol.* 148, 993–1003. doi: 10.1104/pp.108.122457

Hollender, C. A., Geretz, A. C., Slovin, J. P., and Liu, Z. (2012). Flower and early fruit development in a diploid strawberry, *Fragaria vesca*. *Planta* 235, 1123–1139. doi: 10.1007/s00425-011-1562-1

Hu, L., and Liu, S. (2011). Genome-wide identification and phylogenetic analysis of the ERF gene family in cucumbers. *Genet. Mol. Biol.* 34, 624–633. doi: 10.1590/S1415-47572011005000054

Jaillon, O., Aury, J. M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449, 463–467. doi: 10.1038/nature06148

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* 8, 275–282. doi: 10.1093/bioinformatics/8.3.275

Jourda, C., Cardi, C., Mbéguié-A-Mbéguié, D., Bocs, S., Garsmeur, O., D'Hont, A., et al. (2014). Expansion of banana (*Musa acuminata*) gene families involved in ethylene biosynthesis and signalling after lineage-specific whole-genome duplications. *New Phytol.* 202, 986–1000. doi: 10.1111/nph.12710

Kang, C., Darwish, O., Geretz, A., Shahan, R., Alkharouf, N., and Liu, Z. (2013). Genome-scale transcriptomic insights into early-stage fruit development in woodland strawberry *Fragaria vesca*. *Plant Cell* 25, 1960–1978. doi: 10.1105/tpc.113.111732

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Keane, T. M., Creevey, C. J., Pentony, M. M., Naughton, T. J., and Mclnerney, J. O. (2006). Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evol. Biol.* 6, 29. doi: 10.1186/1471-2148-6-29

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* 23, 2947–2948. doi: 10.1093/bioinformatics/btm404

Lata, C., and Prasad, M. (2011). Role of DREBs in regulation of abiotic stress responses in plants. *J. Exp. Bot.* 62, 4731–4748. doi: 10.1093/jxb/err210

Lee, T. H., Tang, H., Wang, X., and Paterson, A. H. (2013). PGDD: a database of gene and genome duplication in plants. *Nucleic Acids Res.* 41, D1152–D1158. doi: 10.1093/nar/gks1104

Li, Z., Zhang, H., Ge, S., Gu, X., Gao, G., and Luo, J. (2009). Expression pattern divergence of duplicated genes in rice. *BMC Bioinf.* 10(Suppl 6), S8. doi: 10.1186/1471-2105-10-S6-S8

Licausi, F., Ohme-Takagi, M., and Perata, P. (2013). APETALA2/Ethylene Responsive Factor (AP2/ERF) transcription factors: mediators of stress responses and developmental programs. *New Phytol.* 199, 639–649. doi: 10.1111/nph.12291

Liu, Q., Kasuga, M., Sakuma, Y., Abe, H., Miura, S., Yamaguchi-Shinozaki, K., et al. (1998). Two transcription factors, DREB1 and DREB2, with an EREBP/AP2 DNA binding domain separate two cellular signal transduction pathways in drought- and low-temperature-responsive gene expression, respectively, in Arabidopsis. *Plant Cell* 10, 1391–1406. doi: 10.1105/tpc.10.8.1391

Livak, K. J., and Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(−Delta C(T)) method. *Methods* 25, 402–408. doi: 10.1006/meth.2001.1262

Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M., et al. (2005). Modeling gene and genome duplications in eukaryotes. *Proc. Natl. Acad. Sci. U.S.A.* 102, 5454–5459. doi: 10.1073/pnas.0501102102

Matías-Hernández, L., Aguilar-Jaramillo, A. E., Marín-González, E., Suárez-López, P., and Pelaz, S. (2014). RAV genes: regulation of floral induction and beyond. *Ann Bot.* 114, 1459–1470. doi: 10.1093/aob/mcu069

Nakano, T., Suzuki, K., Fujimura, T., and Shinshi, H. (2006). Genome-wide analysis of the ERF gene family in Arabidopsis and rice. *Plant Physiol.* 140, 411–432. doi: 10.1104/pp.105.073783

Novillo, F., Medina, J., and Salinas, J. (2007). Arabidopsis CBF1 and CBF3 have a different function than CBF2 in cold acclimation and define different gene classes in the CBF regulon. *Proc. Natl. Acad. Sci. U.S.A.* 104, 21002–21007. doi: 10.1073/pnas.0705639105

Panchy, N., Lehti-Shiu, M., and Shiu, S. H. (2016). Evolution of gene duplication in plants. *Plant Physiol.* 171, 2294–2316. doi: 10.1104/pp.16.00523

Rozas, J., Sánchez-DelBarrio, J. C., Messeguer, X., and Rozas, R. (2003). DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 19, 2496–2497. doi: 10.1093/bioinformatics/btg359

Shulaev, V., Sargent, D. J., Crowhurst, R. N., Mockler, T. C., Folkerts, O., Delcher, A. L., et al. (2011). The genome of woodland strawberry (*Fragaria vesca*). *Nat. Genet.* 43, 109–116. doi: 10.1038/ng.740

Song, C. P., Agarwal, M., Ohta, M., Guo, Y., Halfter, U., Wang, P., et al. (2005). Role of an Arabidopsis AP2/EREBP-type transcriptional repressor in abscisic acid and drought stress responses. *Plant Cell* 17, 2384–2396. doi: 10.1105/tpc.105.033043

Song, X., Li, Y., and Hou, X. (2013). Genome-wide analysis of the AP2/ERF transcription factor superfamily in Chinese cabbage (*Brassica rapa ssp. pekinensis*). *BMC Genomics* 14, 573. doi: 10.1186/1471-2164-14-573

Sun, Z. M., Zhou, M. L., Xiao, X. G., Tang, Y. X., and Wu, Y. M. (2014). Genome-wide analysis of AP2/ERF family genes from Lotus corniculatus shows LcERF054 enhances salt tolerance. *Funct. Integr. Genomics* 14, 453–466. doi: 10.1007/s10142-014-0372-5

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* 34, 609–612. doi: 10.1093/nar/gkl315

Tan, X. L., Fan, Z. Q., Shan, W., Yin, X. R., Kuang, J. F., Lu, W. J., et al. (2018). Association of *BrERF72* with methyl jasmonate-induced leaf senescence of Chinese flowering cabbage through activating JA biosynthesis-related genes. *Hortic. Res.* 5, 22. doi: 10.1038/s41438-018-0028-z

Tang, H. B., Wang, X. Y., Bowers, J. E., Ming, R., Alam, M., and Paterson, A. H. (2008). Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.* 18, 1944–1954. doi: 10.1101/gr.080978.108

Tuskan, G. A., DiFazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., et al. (2006). The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science* 313, 1596–1604. doi: 10.1126/science.1128691

Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., et al. (2010). The genome of the domesticated apple (*Malus* × *domestica Borkh.*). *Nat. Genet.* 42, 833–839. doi: 10.1038/ng.654

Wang, Y., Tan, X., and Paterson, A. H. (2013). Different patterns of gene structure divergence following gene duplication in Arabidopsis. *BMC Genomics* 14, 652. doi: 10.1186/1471-2164-14-652

Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., et al. (2012a). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 40, e49. doi: 10.1093/nar/gkr1293

Wang, Y., Wang, X., and Paterson, A. H. (2012b). Genome and gene duplications and gene expression divergence: a view from plants. *Ann. N. Y Acad. Sci.* 1256, 1–14. doi: 10.1111/j.1749-6632.2011.06384.x

Wang, Z., Zhou, Z., Liu, Y., Liu, T., Li, Q., Ji, Y., et al. (2015). Functional evolution of phosphatidylethanolamine binding proteins in soybean and Arabidopsis. *Plant Cell* 27, 323–336. doi: 10.1105/tpc.114.135103

Weigel, D. (1995). The Apetala2 domain is related to a novel type of DNA-binding domain. *Plant Cell* 7, 388–389. doi: 10.1105/tpc.7.4.388

Wilson, K., Long, D., Swinburne, J., and Coupland, G. (1996). A Dissociation insertion causes a semidominant mutation that increases expression of TINY, an Arabidopsis gene related to APETALA2. *Plant Cell* 8, 659–671. doi: 10.1105/tpc.8.4.659

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* 13, 555–556. doi: 10.1093/bioinformatics/13.5.555

Yu, J., Ke, T., Tehrim, S., Sun, F., Liao, B., and Hua, W., (2015). PTGBase: an integrated database to study tandem duplicated genes in plants. *Database* pii, bav017. doi: 10.1093/database/bav017

Zhuang, J., Yao, Q. H., Xiong, A. S., and Jian, Z. (2011). Isolation, phylogeny and expression patterns of AP2-like genes in apple (*Malus* × *domestica* Borkh). *Plant Mol. Biol. Rep.* 29, 209–216. doi: 10.1007/s11105-010-0227-8.

Conflict of Interest Statement: 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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 26 September 2019
doi: 10.3389/fgene.2019.00903

SCDevDB: A Database for Insights Into Single-Cell Gene Expression Profiles During Human Developmental Processes

*Zishuai Wang†, Xikang Feng† and Shuai Cheng Li**

Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong

Single-cell RNA-seq studies profile thousands of cells in developmental processes. Current databases for human single-cell expression atlas only provide search and visualize functions for a selected gene in specific cell types or subpopulations. These databases are limited to technical properties or visualization of single-cell RNA-seq data without considering the biological relations of their collected cell groups. Here, we developed a database to investigate single-cell gene expression profiling during different developmental pathways (SCDevDB). In this database, we collected 10 human single-cell RNA-seq datasets, split these datasets into 176 developmental cell groups, and constructed 24 different developmental pathways. SCDevDB allows users to search the expression profiles of the interested genes across different developmental pathways. It also provides lists of differentially expressed genes during each developmental pathway, T-distributed stochastic neighbor embedding maps showing the relationships between developmental stages based on these differentially expressed genes, Gene Ontology, and Kyoto Encyclopedia of Genes and Genomes analysis results of these differentially expressed genes. This database is freely available at https://scdevdb.deepomics.org

Keywords: single cell, gene expression, development, database, cell type, differential expression

OPEN ACCESS

Edited by:
Jialiang Yang,
Geneis (Beijing) Co. Ltd, China

Reviewed by:
Suman Ghosal,
National Institutes of Health (NIH), United States
Lei Chen,
Shanghai Maritime University, China

****Correspondence:***
Shuai Cheng Li
shuaicli@cityu.edu.hk

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *12 May 2019*
Accepted: *26 August 2019*
Published: *26 September 2019*

Citation:
Wang Z, Feng X and Li SC (2019) SCDevDB: A Database for Insights Into Single-Cell Gene Expression Profiles During Human Developmental Processes. Front. Genet. 10:903. doi: 10.3389/fgene.2019.00903

INTRODUCTION

In developmental biology, gene expression changes during the developmental process is an important feature to understand developmental questions such as cell growth, cell differentiation, cell fate decisions, etc. (Ko, 2001; Merks and Glazier, 2005; Gittes, 2009). Recently, high-throughput RNA sequencing technique has been widely used to study gene expression in developmental processes (Spitz and Furlong, 2006). Bulk RNA sequencing typically uses hundreds to millions of cells and reveals only the average expression level for each gene across a large population of cell populations (Wang and Bodovitz, 2010; Sanchez and Golding, 2013). Single-cell RNA-seq measures the distribution of expression levels for each gene across a population of cells and provides a more accurate representation of cell-to-cell variations instead of the stochastic average (Saliba et al., 2014). Therefore, single-cell RNA-seq is particularly apposite for developmental biology (Liu et al., 2014; Griffiths et al., 2018).

High-resolution single-cell transcriptome analysis has been performed during many developmental processes including preimplantation development from oocyte to morula (Xue et al., 2013; Yan et al., 2013), early forebrain and mid/hindbrain cell differentiation from human embryonic stem

cells (hESCs) (Yao et al., 2017), and digestive tract development from human embryos between 6 and 25 weeks (Gao et al., 2018), etc. These studies not only revealed many biological features, including developmental processes, signaling pathways, cell cycle, and transcription factor networks but also provided resources to investigate the gene expression patterns during different developmental processes. Therefore, there is a strong need for a web resource that curates and provides single-cell gene expression profiles during different developmental processes.

So far, several web resources for human single-cell transcriptome data have been reported. scRNASeqDB contains 38 datasets covering 200 human cell lines or cell types and 13,440 samples (Cao et al., 2017). The single-cell expression atlas, launched by the European Bioinformatics Institute (https://www.ebi.ac.uk/gxa/sc/home), contains 52 single-cell RNA-Seq studies, consisting of 61,073 cells from 9 different species. The single-cell centric database "SCPortalen" covers 23 human single-cell transcriptomics datasets that are publicly available from the International Nucleotide Sequence Database Collaboration sites (Abugessaisa et al., 2017). PanglaoDB integrated 209 human single-cell datasets consisting of gene expression measurements from cells originating from a common biological source or experiment (Franzén et al., 2019). However, users of these databases can only query gene expression in specific cell types or population heterogeneity processed by the authors. Researchers who are interested in gene expression changes during a specific developmental process are not easily able to extract these dynamic features from these databases.

Here, we developed a database to investigate single-cell gene expression profiling during different developmental processes (SCDevDB). In this database, we collected 10 human single-cell RNA-seq datasets, split these datasets into 176 developmental cell groups, and constructed 24 different developmental pathways. Users of SCDevDB are easy to view the expression changes of their interested genes showed with a boxplot. In addition, users can also download differentially expressed (DE) genes during each developmental pathway, the T-distributed stochastic neighbor embedding (t-SNE) map constructed with these genes, Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis results of these differentially expressed (DE) genes. This database is publicly available at https://scdevdb.deepomics.org. It helps researchers within the fields of developmental biology to facilitate gene expression studies in human single cells.

MATERIALS AND METHODS

Transcriptomic Data Collection

We searched the National Center for Biotechnology Information Gene Expression Omnibus database using the successfully utilized keywords, single-cell RNA-seq, single-cell RNA-seq, single-cell transcriptome, and selected the species to humans. In this study, we only focused on the normal human developmental processes; thus, we abnegated experiments using tumor and other samples treated with chemical reagents. After carefully reviewing the resultant papers and datasets, we obtained 10 datasets for human single-cell RNA-seq using normal cell type, tissue, or organs. These datasets including human cell groups related to the nervous system, digestive system, the heart, the brain, hESC, cell lines, and others. Single cells originating from the same cell lines, tissue regions, or organ regions at the same developmental time points are treated as a cell group. Based on this rule, we classified the 18,413 single cells into 176 cell groups (**Supplemental Table S1**). Cell groups originating from the same cell lines, tissue regions, or organ regions but at different developmental time points were regarded as one developmental stage. Therefore, the 176 cell groups were merged into 35 developmental stages (**Supplemental Table S1**).

Data Processing and Gene Expression Profiling Analysis

For the selected RNA-Seq experiments, the gene expression matrices were also retrieved from the Gene Expression Omnibus. For cells in datasets where the fragments per kilobase of exon per million reads mapped (FPKM) were available, we computed the TPM for gene i in cell j, according to:

$$TPM_i = \left(\frac{FPKM_i}{\sum_j FPKM_j} \right) \times 10^6$$

This conversion enables the units to be consistent for dataset-to-dataset comparison. Then, for each dataset, we merged cells originating from the same tissue or organ into one file and performed imputation using the R package single-cell analysis via expression recovery with default parameters. Single-cell analysis via expression recovery takes in a matrix and performs library size normalization during denoising step, which can reduce noise including sequencing depth, the number of cells, and cell composition (Huang et al., 2018). We eventually got 176 different files which are consistent with 176 different cell groups.

Differential Gene Expression and T-SNE Analysis

For each developmental pathway, we merged the expression data of all developmental stages in this pathway into one file. Then, we conducted DE gene analysis between cell groups in the same developmental pathway using Monocle, which will do all needed normalization steps internally, with default parameters (Qiu et al., 2017). We extracted expression data of the DE genes and performed t-SNE analysis with different perplexity for different process (Pedregosa et al., 2011).

GO and KEGG Enrichment Analysis

The symbol names of DE genes were used as the gene list input into R packages "GOstats" (Falcon and Gentleman, 2006) and "KEGG.db" (Carlson et al., 2016) for GO and KEGG analysis, respectively. We selected the "ontology" parameter as "BP," "MF," and "CC" for GO analysis and "pvalueCutoff" parameter as 0.5 for both GO and KEGG analysis. Top 20 significantly enriched GO terms and KEGG terms were selected to show potential functions of DE genes.

Database Construction

The SCDevDB website was built using the Django Python Web framework (https://www.djangoproject.com) coupled

with the MySQL database. The front-end interface was developed based on the Bootstrap open source toolkit (https://getbootstrap.com). The web interactive visualization graphs were developed using Plotly JavaScript Open Source Graphing Library (https://plot.ly/javascript/). SCDevDB was published using the Apache http server and is accessible at https://scdevdb.deepomics.org/.

RESULTS

Datasets Summary and the Developmental Tree Construct

At the time of this publication, the database contains 10 datasets covering 18,413 single cells and 176 cell groups (see Methods). According to the notation of the data resources, we classified these cell groups into 35 developmental stages. Every mammalian individual is developed from the totipotent zygote. Mammalian preimplantation development is a complex process including a series of cell divisions from 1 to 2 cells, 2 to 4 cells, 4 to 8 cells, 8 to 16 cells, and 16 cells to blastocyst (Niakan et al., 2012). After that, nearly all of the human tissues are original from embryoblast (hESC). Then, a developmental tree was constructed based on the development process of the multicellular organism (Hall, 2012) (**Figure 1**). Specifically, we first considered the developmental process from oocyte to hESC as the root process; then, the left 27 developmental stages were classified into 24 different developmental pathways by combining with the root process (**Supplemental Table S1**). The detailed cell number in each stage is shown in **Figure 2**, and the datasets summary is available at https://scdevdb.deepomics.org/data-summary/.

User Interface to the SCDevDB

In order to provide users easy access to the SCDevDB, we designed an interface to allow users to perform basic operations, such as searching, viewing, and downloading data. SCDevDB is composed of two functional pages: "Gene Expression Search" page and "Differential Gene List Collection" page. The web interface of SCDevDB is summarized in **Figures 3** and **4**.

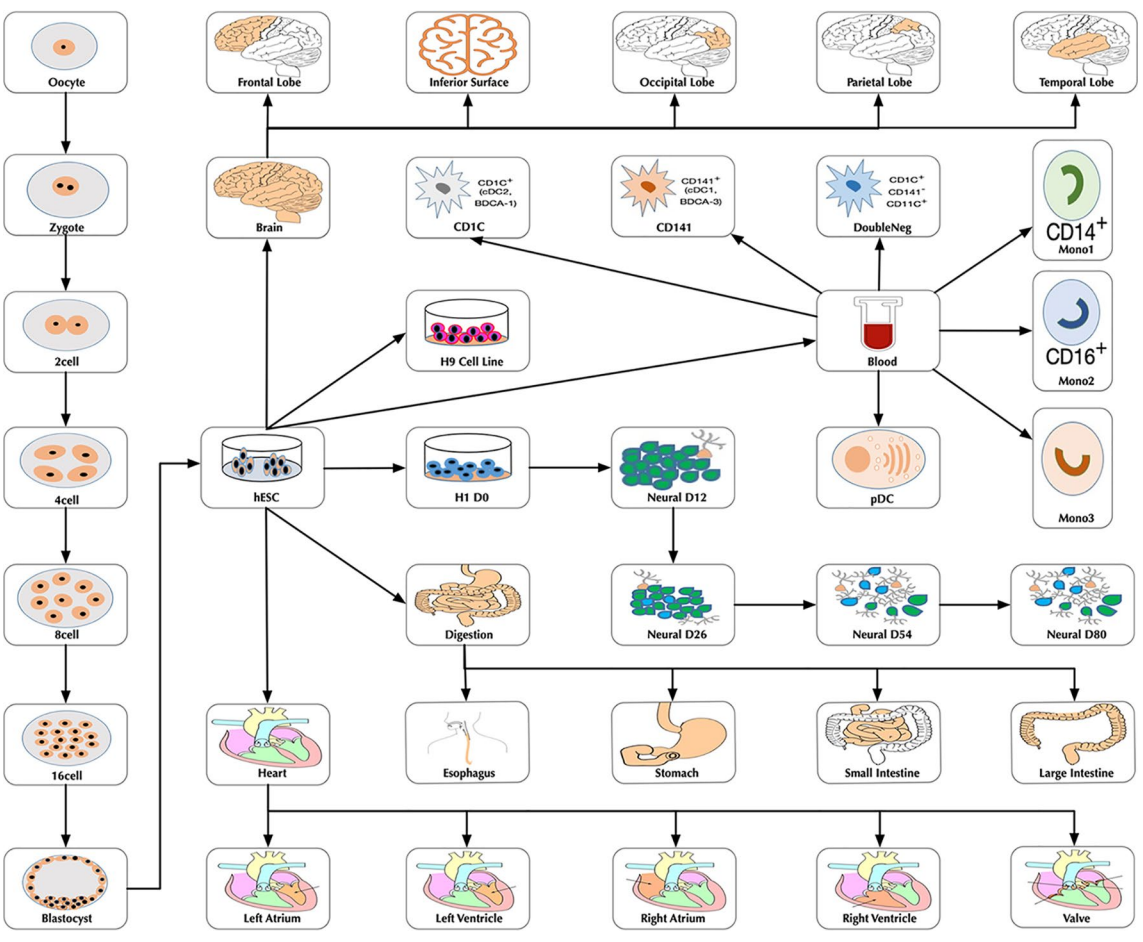


FIGURE 1 | The developmental tree. Figures of the brain, heart and digestion originate from Wikimedia Commons (https://commons.wikimedia.org/wiki/Main_Page).

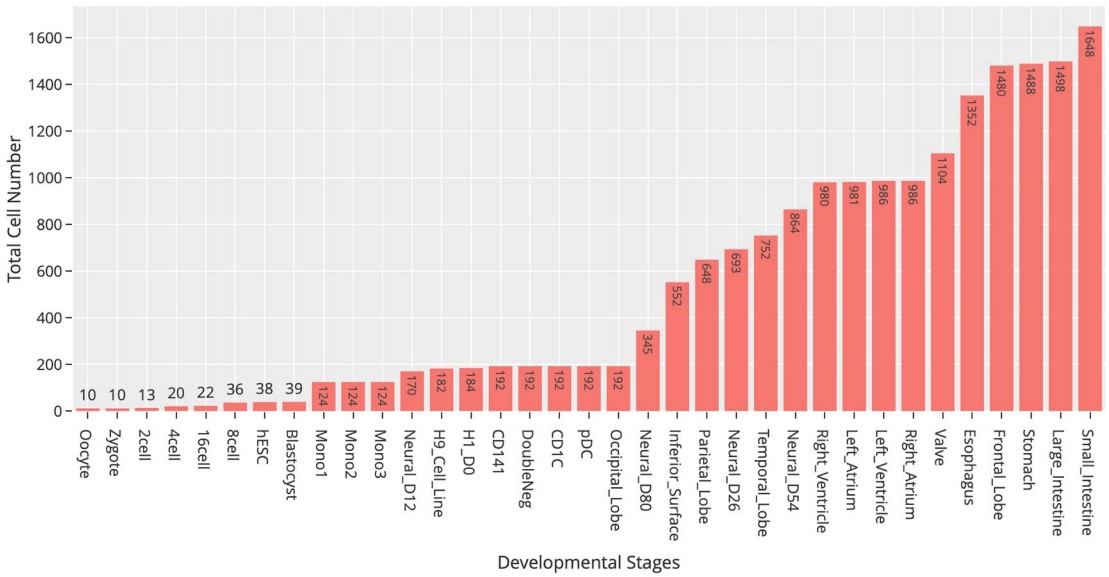


FIGURE 2 | Statistics of cell numbers of 35 developmental stages.

FIGURE 3 | Overview of the gene expression search page. **(A)** Searching result of the gene "MYL2". **(B)** Boxplot shows expression level distribution of MYL2 during developmental process by clicking the image. **(C)** The function of removing uninterested developmental stages by clicking the name of the stage listed in the figure legend. **(D)** An example of double clicking on a stage name.

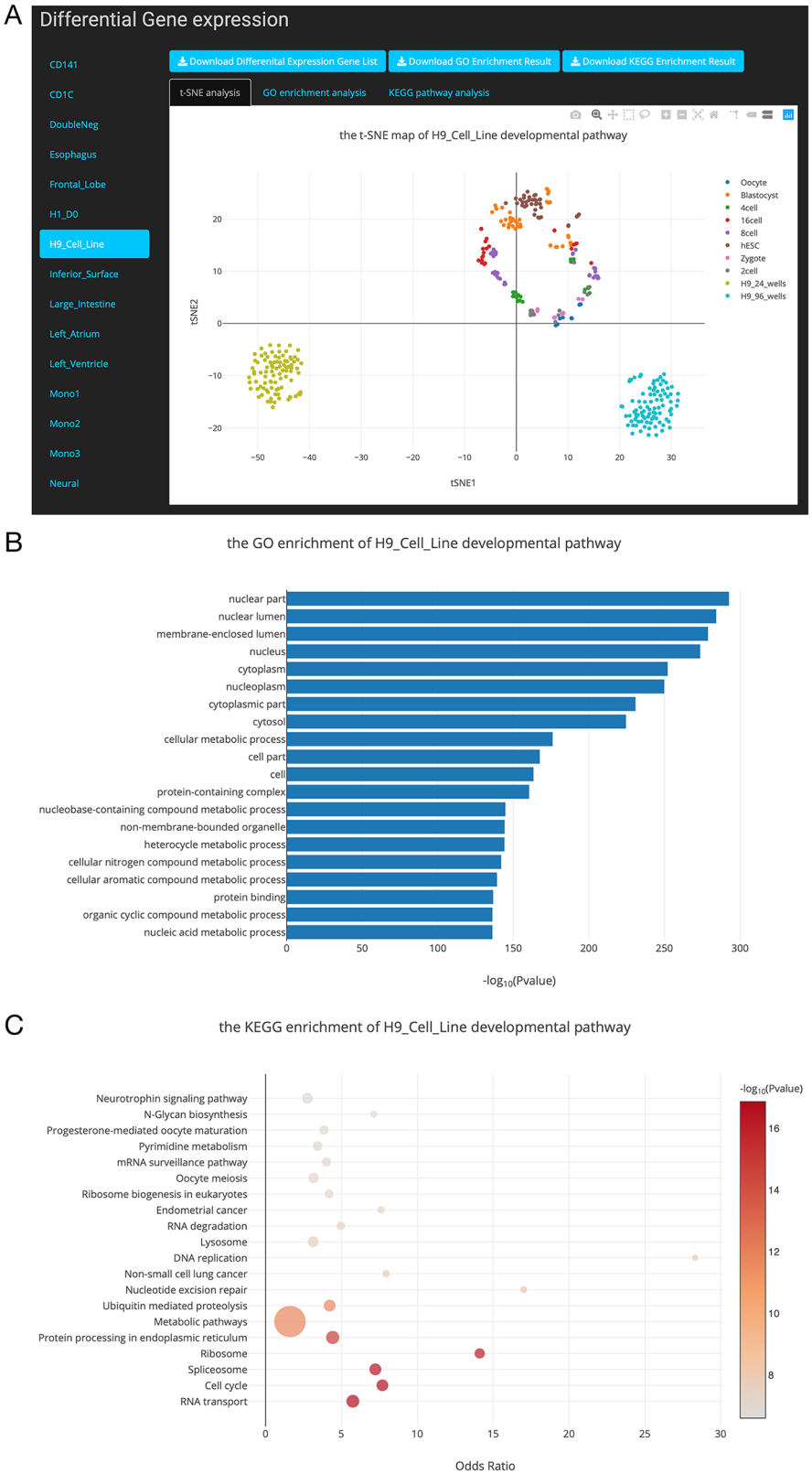


FIGURE 4 | Overview of the differential gene list collection page. **(A)**. T-distributed stochastic neighbor embedding (t-SNE) maps showing the relationships between developmental stages based on these differentially expressed genes. **(B)**. Top 20 Gene Ontology (GO) terms of differential expression genes. **(C)**. Top 20 Kyoto Encyclopedia of Genes and Genomes (KEGG) terms of differential expression genes.

Query Function to Search Gene Expression Across 35 Developmental Stages

In this page, users can view whether an interested gene is expressed in different developmental stages by giving a gene symbol (e.g., APMAP) or an Ensembl ID (e.g., ENSG00000101474) in the searching input box. The searching result will be displayed in the developmental tree. Specifically, if the searched gene ("MYL2" gene as an example) is not expressed at one stage, the stage image will be disabled and cannot be clicked (the light-colored images in **Figure 3A**). Furthermore, the interactive boxplot of gene expression level along with a selected developmental pathway is available by clicking the stage image (**Figure 3B**). To illustrate the interactive function of this boxplot, we took the distribution of the MYL2 expression during left ventricle process as an example. Clicking on the stage name "Left_Ventricle_E7w" listed in the graph legend can remove the boxplot data of this stage (**Figure 3C**). This function allows users to compare their interested stages. Moreover, double clicking on a stage name allows users to view detail gene expression value of this stage (**Figure 3D**). These boxplots can be download in PNG format for further usage.

Differential Gene List Collection for 24 Developmental Pathways

In this study, we performed DE gene analysis for 24 developmental pathways. Finally, 24 differential gene lists were collected into the SCDevDB. Users can download these gene lists by clicking the Download button in Differential Gene List Collection page. Moreover, we performed t-SNE analysis using these differential gene lists, and the result is displayed using an interactive scatterplot (**Figure 4A**). Subsequently, GO and KEGG enrichment analysis of the DE genes were performed using R packages, and top 20 significantly enriched GO or KEGG terms were selected to show potential functions of these DE genes (**Figures 4B**, **C**). In addition, tables showing all of the GO or KEGG terms are also available and free to download on the "Differential Gene List Collection" page. These scatterplots and bar chart can be downloaded in PNG format for further usage.

Case Study

Myosin light chain-2 (MYL2, also called MLC-2) is a protein that belongs to the EF-hand calcium binding protein superfamily and exists as three major isoforms encoded by three distinct genes in mammalian striated muscle (Sheikh et al., 2015). Diseases associated with MYL2 include cardiomyopathy, familial hypertrophic, and congenital fiber-type disproportion (Flavigny et al., 1998; Weterman et al., 2013). Here, we used this gene as an interested example to test the functions of SCDevDB. Previous studies using bulk-seq data have shown that MYL2 is highly expressed in tissue of muscles including skeletal muscle, myocardial, and smooth muscles (Hsu et al., 2012; Lindholm et al., 2014; Renaudin et al., 2018). Searching result of the SCDevDB is consistent with these studies as shown in **Figure 3A**. Moreover, comparing with the expression levels in cells of the atriums, MYL2 has higher levels in cells of the ventricles (**Figure 5**). This result indicated that MYL2 can be used as a marker gene to distinguish ventricle and atrium cells in subpopulation analysis.

hESC lines has been used as a source of cells for regenerative medicine, as well as valuable tools for drug discovery and for understanding human development and disease (Allegrucci and Young, 2006). Notably, H9 is one of the first five lines derived in the University of Wisconsin, i.e. H1, H7, H9, H13 and H14 (Denning et al., 2003), which has been used as an important material in many publications (Amit et al., 2000; Gafni et al., 2013; Kim et al.,

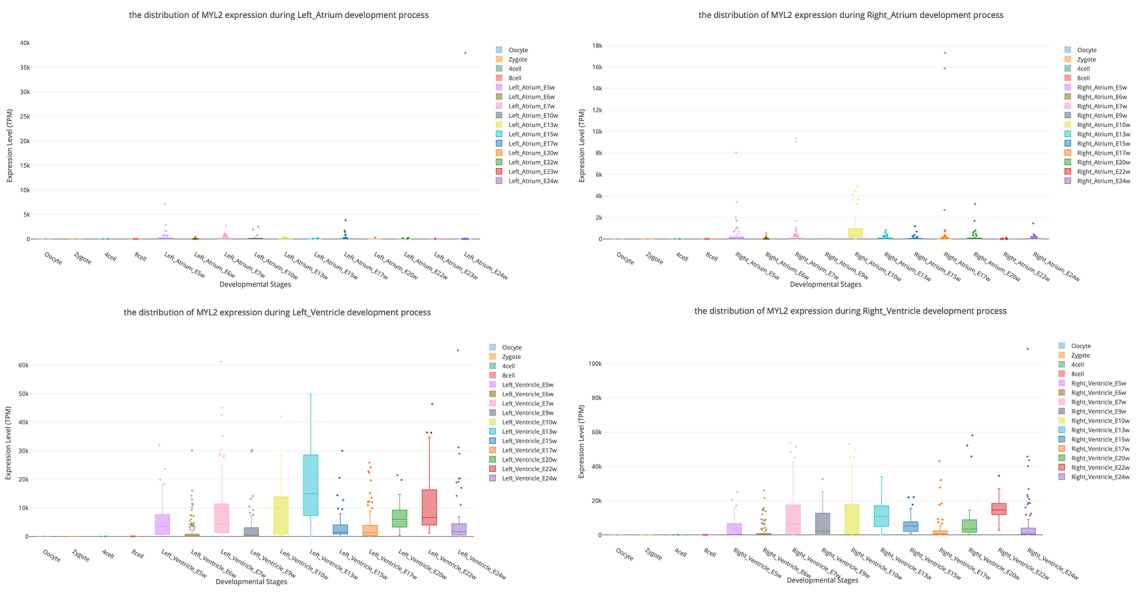


FIGURE 5 | Comparison of MYL2 expression distributions between atrium cell types and ventricle cell types.

2014). In our "Differential Gene List Collection" page, when we selected the "H9_Cell_Line" developmental pathway, the t-SNE map indicates that H9 cell lines are distinct from preimplantation cell types (**Figure 4A**). This result is reasonable as the H9 cell line are different from embryonic stem cells in expression levels of various genes (Telugu et al., 2013). Our GO and KEGG analysis results showed that the potential functions of the DE genes during the H9_Cell_Line developmental pathway were enriched in developmental-related biology processes including cellular metabolic process, nucleobase-containing compound metabolic process, RNA transport, and cell cycle pathways.

CONCLUSION

In summary, unlike previous databases, SCDevDB is an interactive database providing human single cell resources to profiling gene expression distributions in different developmental pathways. This database also provides DE gene lists in each developmental pathway, t-SNE map, and GO and KEGG enrichment analysis based on these differential genes. We believe that this database will facilitate researchers within the fields of developmental biology to investigate gene expression changes during human developmental pathways in the single-cell level.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found here: https://www.ncbi.nlm.nih.gov/geo/.

AUTHOR CONTRIBUTIONS

ZW performed the data collection and analysis, XF developed the database, SL designed and supervised the study, and ZW, XF, and SL wrote the manuscript.

FUNDING

This research was funded by a GRF Project grant from the RGC General Research Fund (9042181; CityU 11203115), the GRF Research Project (9042348; CityU 11257316).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00903/full#supplementary-material

REFERENCES

Abugessaisa, I., Noguchi, S., Böttcher, M., Hasegawa, A., Kouno, T., Kato, S., et al. (2017). SCPortalen: human and mouse single-cell centric database. *Nucleic Acids Res.* 46, D781–D787. doi: 10.1093/nar/gkx949

Allegrucci, C., and Young, L. (2006). Differences between human embryonic stem cell lines. *Hum. Reprod. Update* 13, 103–120. doi: 10.1093/humupd/dml041

Amit, M., Carpenter, M. K., Inokuma, M. S., Chiu, C.-P., Harris, C. P., Waknitz, M. A., et al. (2000). Clonally derived human embryonic stem cell lines maintain pluripotency and proliferative potential for prolonged periods of culture. *Dev. Biol.* 227, 271–278. doi: 10.1006/dbio.2000.9912

Cao, Y., Zhu, J., Jia, P., and Zhao, Z. (2017). scRNASeqDB: a database for RNA-Seq based gene expression profiles in human single cells. *Genes* 8, 368. doi: 10.3390/genes8120368

Carlson, M., Falcon, S., Pages, H., and Li, N. (2016). KEGG. db: A set of annotation maps for KEGG. *R Package Version* 3, 2016. doi: 10.18129/B9.bioc.KEGG.db

Denning, C., Allegrucci, C., Priddle, H., Barbadillo-Muñoz, M. D., Anderson, D., Self, T., et al. (2003). Common culture conditions for maintenance and cardiomyocyte differentiation of the human embryonic stem cell lines, BG01 and HUES-7. *Int. J. Dev. Biol.* 50, 27–37. doi: 10.1387/ijdb.052107cd

Falcon, S., and Gentleman, R. (2006). Using GOstats to test gene lists for GO term association. *Bioinformatics* 23, 257–258. doi: 10.1093/bioinformatics/btl567

Flavigny, J., Richard, P., Isnard, R., Carrier, L., Charron, P., Bonne, G., et al. (1998). Identification of two novel mutations in the ventricular regulatory myosin light chain gene (MYL2) associated with familial and classical forms of hypertrophic cardiomyopathy. *J. Mol. Med.* 76, 208–214. doi: 10.1007/s001090050210

Franzén, O., Gan, L.-M., and Björkegren, J. L. (2019). PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database* 2019 doi: 10.1093/database/baz046

Gafni, O., Weinberger, L., Mansour, A. A., Manor, Y. S., Chomsky, E., Ben-Yosef, D., et al. (2013). Derivation of novel human ground state naive pluripotent stem cells. *Nature* 504, 282. doi: 10.1038/nature12745

Gao, S., Yan, L., Wang, R., Li, J., Yong, J., Zhou, X., et al. (2018). Tracing the temporal-spatial transcriptome landscapes of the human fetal digestive tract using single-cell RNA-sequencing. *Nat. Cell Biol.* 20, 721. doi: 10.1038/s41556-018-0105-4

Gittes, G. K. (2009). Developmental biology of the pancreas: a comprehensive review. *Dev. Biol.* 326, 4–35. doi: 10.1016/j.ydbio.2008.10.024

Griffiths, J. A., Scialdone, A., and Marioni, J. C. (2018). Using single-cell genomics to understand developmental processes and cell fate decisions. *Mol. Syst. Biol.* 14, e8046. doi: 10.15252/msb.20178046

Hall, B. K. (2012). *Evolutionary developmental biology*. Springer Science & Business Media.

Hsu, J., Hanna, P., Van Wagoner, D. R., Barnard, J., Serre, D., Chung, M. K., et al. (2012). Whole genome expression differences in human left and right atria ascertained by RNA sequencing. *Circ. Cardiovasc. Genet.* 5, 327–335. doi: 10.1161/CIRCGENETICS.111.961631

Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., et al. (2018). SAVER: gene expression recovery for single-cell RNA sequencing. *Nat. Methods* 15, 539. doi: 10.1038/s41592-018-0033-z

Kim, J. J., Khalid, O., Namazi, A., Tu, T. G., Elie, O., Lee, C., et al. (2014). Discovery of consensus gene signature and intermodular connectivity defining self-renewal of human embryonic stem cells. *Stem Cells* 32, 1468–1479. doi: 10.1002/stem.1675

Ko, M. S. (2001). Embryogenomics: developmental biology meets genomics. *Trends Biotechnol.* 19, 511–518. doi: 10.1016/S0167-7799(01)01806-6

Lindholm, M. E., Huss, M., Solnestam, B. W., Kjellqvist, S., Lundeberg, J., and Sundberg, C. J. (2014). The human skeletal muscle transcriptome: sex differences, alternative splicing, and tissue homogeneity assessed with RNA sequencing. *FASEB J.* 28, 4571–4581. doi: 10.1096/fj.14-255000

Liu, N., Liu, L., and Pan, X. (2014). Single-cell analysis of the transcriptome and its application in the characterization of stem cells and early embryos. *Cell. Mol. Life Sci.* 71, 2707–2715. doi: 10.1007/s00018-014-1601-8

Merks, R. M., and Glazier, J. A. (2005). A cell-centered approach to developmental biology. *Physica A* 352, 113–130. doi: 10.1016/j.physa.2004.12.028

Niakan, K. K., Han, J., Pedersen, R. A., Simon, C., and Pera, R.A. (2012). Human pre-implantation embryo development. *Development* 139, 829–841. doi: 10.1242/dev.060426

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

Qiu, X., Hill, A., Packer, J., Lin, D., Ma, Y.-A., and Trapnell, C. (2017). Single-cell mRNA quantification and differential analysis with Census. *Nat. Methods* 14, 309. doi: 10.1038/nmeth.4150

Renaudin, P., Janin, A., Millat, G., and Chevalier, P. (2018). A novel missense mutation p. Gly162Glu of the gene MYL2 involved in hypertrophic cardiomyopathy: a pedigree analysis of a proband. *Mol. Diagn. Ther.* 22, 219–223. doi: 10.1007/s40291-018-0324-1

Saliba, A.-E., Westermann, A. J., Gorski, S. A., and Vogel, J. (2014). Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res.* 42, (14) 8845–8860. doi: 10.1093/nar/gku555

Sanchez, A., and Golding, I. (2013). Genetic determinants and cellular constraints in noisy gene expression. *Science* 342, 1188–1193. doi: 10.1126/science.1242975

Sheikh, F., Lyon, R. C., and Chen, J. (2015). Functions of myosin light chain-2 (MYL2) in cardiac muscle and disease. *Gene* 569, 14–20. doi: 10.1016/j.gene.2015.06.027

Spitz, F., and Furlong, E. E. (2006). Genomics and development: taking developmental biology to new heights. *Dev. Cell* 11, 451–457. doi: 10.1016/j.devcel.2006.09.013

Telugu, B., Adachi, K., Schlitt, J., Ezashi, T., Schust, D., Roberts, R., et al. (2013). Comparison of extravillous trophoblast cells derived from human embryonic stem cells and from first trimester human placentas. *Placenta* 34, 536–543. doi: 10.1016/j.placenta.2013.03.016

Wang, D., and Bodovitz, S. (2010). Single cell analysis: the new frontier in 'omics'. *Trends Biotechnol.* 28, 281–290. doi: 10.1016/j.tibtech.2010.03.002

Weterman, M. A., Barth, P. G., Van Spaendonck-Zwarts, K. Y., Aronica, E., Poll-The, B.-T., Brouwer, O. F., et al. (2013). Recessive MYL2 mutations cause infantile type I muscle fibre disease and cardiomyopathy. *Brain* 136, 282–293. doi: 10.1093/brain/aws293

Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C.-Y., Feng, Y., et al. (2013). Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* 500, 593. doi: 10.1038/nature12364

Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., et al. (2013). Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.* 20, 1131. doi: 10.1038/nsmb.2660

Yao, Z., Mich, J. K., Ku, S., Menon, V., Krostag, A.-R., Martinez, R. A., et al. (2017). A single-cell roadmap of lineage bifurcation in human ESC models of embryonic brain development. *Cell Stem Cell* 20, 120–134. doi: 10.1016/j.stem.2016.09.011

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.

frontiers in Genetics

REVIEW
published: 09 October 2019
doi: 10.3389/fgene.2019.00972

Microbiome Big-Data Mining and Applications Using Single-Cell Technologies and Metagenomics Approaches Toward Precision Medicine

*Mingyue Cheng, Le Cao and Kang Ning**

Key Laboratory of Molecular Biophysics of the Ministry of Education, Hubei Key Laboratory of Bioinformatics and Molecular Imaging, Department of Bioinformatics and Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, China

OPEN ACCESS

Edited by:
Jialiang Yang,
Geneis (Beijing) Co. Ltd, China

Reviewed by:
Fengfeng Zhou,
Jilin University, China
Xuefeng Cui,
Tsinghua University, China
Minxian Wang,
Broad Institute, United States

***Correspondence:**
Kang Ning
ningkang@hust.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *12 May 2019*
Accepted: *12 September 2019*
Published: *09 October 2019*

Citation:
Cheng M, Cao L and Ning K (2019) Microbiome Big-Data Mining and Applications Using Single-Cell Technologies and Metagenomics Approaches Toward Precision Medicine. Front. Genet. 10:972. doi: 10.3389/fgene.2019.00972

With the development of high-throughput sequencing technologies as well as various bioinformatics analytic tools, microbiome is not a "microbial dark matter" anymore. In this review, we first summarized the current analytical strategies used for big-data mining such as single-cell sequencing and metagenomics. We then provided insights into the integration of these strategies, showing significant advantages in fully describing microbiome from multiple aspects. Moreover, we discussed the correlation between gut microbiome with host organs and diseases, confirming the importance of big-data mining in clinical practices. We finally proposed new ideas about the trend of big-data mining in microbiome using multi-omics approaches and single-cell sequencing. The integration of multi-omics approaches and single-cell sequencing can provide full understanding of microbiome at both macroscopic level and microscopic level, thus contributing to precision medicine.

Keywords: big data, microbiome, metagenomics, single-cell sequencing, precision medicine

STRATEGIES FOR BIG-DATA MINING

The human gut microbiome has been confirmed to highly correlate with human health and diseases, through influencing human metabolism, nutrition, physiology, and immune function (Hooper and Gordon, 2001; Bäckhed et al., 2005; Manichanh et al., 2012). Hence, the characterization of the human gut microbiome, as well as its correlation with diseases, has fascinated a great number of researchers to explore. However, the human gut microbiome consists of approximately 15,000 to 36,000 species of bacteria (Frank et al., 2007), with the total number of bacterial cells ranging from 10^{13} to 10^{14}, which is of the same order as the number of human cells (3.0×10^{13}) (Sender et al., 2016). The gut microbiome also contains more than 100 times more genes, compared with 25,000 genes in humans (Gill et al., 2006). Considering this big data of the gut microbiome, sequencing would be a promising technology for mining it, rather than the traditional cultural methods. Sequencing is the precondition for obtaining raw genetic materials of the gut microbiome, followed by genetic assembly and taxonomic and functional annotations. Several strategies are currently used for big-data mining in microbial communities from different perspectives as follows (**Table 1**).

TABLE 1 | The overview of pros and cons of current widely used methods for dissecting microbiome.

Methods	Advantages	Disadvantages	Solution
Amplicon sequencing	(1) Relatively low cost; (2) Taxonomic annotations of uncultured microbial communities.	(1) Low resolution: cannot identify microbes at species or stain level; (2) Cannot realize functional annotations of microbial communities.	(1) Combined with metagenomics; (2) Use PICRUSt to obtain predicted metagenomics and functional annotations.
Metagenomic sequencing	(1) Taxonomic and functional annotations of uncultured microbial communities; (2) Obtain the full genetic repertoire of the microbial communities.	(1) Difficulties in metagenome assembly and taxonomically and functionally assign accurately; (2) Lack of high genome coverage; (3) Cannot link all the functional genes of one microbe to its phylogeny.	(1) Long-read sequencing and improved algorithms for assembly; (2) Combined with single-cell sequencing.
Single-cell sequencing	(1) Taxonomic and functional annotations of uncultured microbes at cell level; (2) Generate a high-quality genome for microbes with low abundance; (3) Dissect virus-host interactions of uncultured microbes.	(1) Difficulties in cell sorting; (2) Easily influenced by contaminated DNA; (3) Uneven read coverage, chimeric reads caused by MDA.	(1) Combined with metagenomics; (2) Improved experimental operation and various computational approaches to control DNA contamination and errors caused by MDA.

Amplicon Sequencing

Amplicon sequencing uses specific marker genes of microbes such as 16S ribosomal RNA for bacteria and Internal Transcribed Spacer (ITS) for fungi. This sequencing method mainly answers "who is there" in an uncultured microbial community by assigning reads to reference reads. However, low-resolution level (cannot reach to species or strain level) of amplicon sequencing, as well as its disability in functional annotation, largely limits its application. Therefore, current solution for this problem is to combine the amplicon sequencing and the metagenomic sequencing. Researchers can first use relatively low-cost amplicon sequencing to have a preliminary understanding of the composition of the targeted microbial community, thus determining the hypothesis. Subsequently, they can perform metagenomic sequencing to confirm the hypothesis from a perspective of both phylogeny and functions.

Metagenomic Sequencing

The shotgun metagenomic sequencing process consists of DNA extraction from all cells in a community, DNA fragmentation, DNA sequencing, and sequence analysis such as marker gene analysis, binning, or contig assembly to obtain the taxonomic composition. Metagenomic sequencing not only can shed light on "who is there" at a high resolution to strain level, but also "what are they doing." The metagenomic reads encoding proteins can be predicted for functional annotation, through various ways including gene fragment recruitment, protein family classification, and *de novo* gene prediction (Sharpton, 2014). The disadvantages of metagenomics sequencing are as follows. First, there are limitations of short reads produced by next-generation sequencing and the complexity in sequence assembly, especially when multiple strains are present (Sczyrba et al., 2017). For instance, the closely related genomes in a community might represent genome-sized approximate repeats. Second, metagenomic sequencing cannot obtain high genome coverage and might even lose genomes of low abundant microbes, owing to the high genomic richness and evenness in a community (Mende et al., 2016). Third, functional genes of one microbe cannot be fully linked to its phylogeny. There are two solutions for these problems. First, long-read sequencing can solve the ambiguity in sequence assembly (Bertrand et al., 2019). A recent method named OPERA-MS (Bertrand et al., 2019), which combines nanopore-sequenced long reads and Illumina-sequenced short reads through a hybrid metagenomic assembler, succeeds to promote the accuracy of strain-resolved assembly and obtains genomes with higher coverage. The second solution is to combine metagenomics with single-cell sequencing, which can reconstruct how DNA is compartmentalized into cells and link functions to their corresponding species (Tolonen and Xavier, 2017).

Single-Cell Sequencing

The first step of single-cell sequencing is to isolate the individual cells, using serial dilution, microfluidics, flow cytometry, micromanipulation, or encapsulation in droplets (Bäckhed et al., 2005). The following steps include DNA extraction, whole-genome amplification, DNA sequencing, and sequence analysis such as alignment and assembly. Owing to the fact that minimum requirement of high-throughput sequencing is micrograms, which is more than the femtograms of DNA a bacterial cell generally contains, amplification of the minute amounts of DNA of the cell is necessary (Xu and Zhao, 2018). For this purpose, a non–polymerase chain reaction–based DNA amplification method multiple displacement amplification (MDA) (Dean et al., 2002) uses random hexamer primers annealed to the template and a high-fidelity polymerase of the *Bacillus subtilis* phage phi29 (Blanco et al., 1989). The Phi29 DNA polymerase can work at a moderate isothermal condition, with a high-strand displacement activity and an inherent 3′–5′ proofreading exonuclease activity, thus ensuring enough genome coverage with lower amplification error for the following sequencing analysis.

The major advantage of single-cell sequencing is that it can generate a high-quality genome for species with low abundance, which might be lost by the metagenomic sequencing. Additionally, this method can discriminate and validate the functions of individuals within the community, linking

these functions to specific species. Moreover, the single-cell sequencing can simultaneously recover bacterial genomes and extrachromosomal genetic materials in a cell, dissecting virus–host interactions at cell level (Yoon et al., 2011). Single-cell sequencing has already led to many novel findings such as the discovery of bacteria with an alternative genetic code (Campbell et al., 2013), the ability to observe which gut microbial cells use host-derived compounds (Berry et al., 2013), and the ability to quantify the absolute taxon abundances of the gut microbiome (Props et al., 2017).

However, the single-cell sequencing also has limitations as follows. First, cell sorting is a complicated and time-consuming process. Isolating cells from solid medium such as swabs, biopsies, and tissues remains challenging (Tolonen and Xavier, 2017). Second, the amplification step using MDA might magnify the DNA contamination. DNA contamination is mainly from the tainted specimen at the step of cell sorting, polluted reagents or laboratory apparatuses, and microbes in the environment. The solution for the contamination is to keep strictly clean of the work area with extra precaution. In addition, the reaction volume can be moderately reduced to increase the ratio of targeted DNA to the contaminated DNA. Moreover, contaminated DNA can be partly removed by aligning the reads to the reference of potentially contaminated DNA of human and environment. The third limitation is that the MDA procedure would cause highly uneven read coverage and increased formation of chimera reads that links nonadjacent template sequences; thus, conventional genome-assembly algorithms are not suitable for single-cell data. The solution for uneven read coverage is to normalize the reads by trimming the reads according to their k-mer depth, which has been integrated to several assembly algorithms such as SPAdes (Bankevich et al., 2012). The solution for chimera reads is to identify and remove the chimeras. Owing to the lack of reference genome of a certain number of cells, metagenomic sequencing can provide the contigs as reference for identifying chimeras.

The Integration of Single-Cell Genomics and Metagenomics

The metagenomics represents the whole genome of all microbes in the environment, while single-cell genomics refers to the genomes of individuals cells that may or may not contain the full genetic repertoire in the microbiota. Hence, the integration of these two technologies can make up for each other's shortcomings (**Figure 1**). For instance, reads and contigs of metagenomics can improve the genome assembly of single-cell genomics (Mende et al., 2016). Conversely, single-cell genomics can serve as scaffolds for comparison or recruitment of metagenomics when reference genomes are unavailable (Swan et al., 2013; Roux et al., 2014). Several studies have generated much-improved microbe genome assemblies from a variety of microbial communities, using the integration of single-cell genomics and metagenomics (Dupont et al., 2012; Nobu et al., 2015). The disadvantage of this integration is that the potential errors of both methods would be gathered, thus requiring more sophisticated methods to deal with.

The Integration of Metagenomics and Three-Dimensional Genomics

Metagenomics can quantify the genetic materials of a microbial community, while the Hi-C sequencing can identify all chromatin interactions of the community, producing three-dimensional (3D) genome, reflecting both the genetic content and topological chromatin structures into digital information (Belaghzal et al., 2017). The integration of metagenomics and 3D genomics can fully display the composition and structure of genomes of a microbial community. Moreover, a recent study performed Hi-C for single-cell analysis, to capture 3D genomes of individual cells (Nagano et al., 2017).

Microbial Multi-Omics Analysis

With advances in high-throughput sequencing technologies and bioinformatics approaches, researchers are now able to perform comprehensive analysis in microbial communities, named as "multi-omics analysis." This analysis integrates metagenome, metatranscriptome, metaproteome, and metabolome. The metagenome displays the taxonomic composition in a microbial community and predicted functional expression. The metatranscriptome, metaproteome, and metabolome can confirm the predicted functions, further unveiling how microbes work in a community. These omics can provide significant information about a microbial community from different perspectives. For instance, the microbial communities of twins with Crohn disease have been analyzed at phylogenetic, functional, and metabolic levels, using 16S sequencing (Dicksved et al., 2008; Willing et al., 2009; Willing et al., 2010), metagenomics, proteomics (Erickson et al., 2012), and metabolomics (Jansson et al., 2009). The subjects with Crohn disease contain a microbial community with lower microbial diversity, depletion of *Faecalibacterium prausnitzii*, and lower expression levels of proteins involved in butyrate metabolism (Erickson et al., 2012). At the metabolite level, thousands of metabolites such as the bile acids (BAs) that were detected higher in diseased subjects can distinguish healthy subjects from subjects with Crohn disease (Jansson et al., 2009). Therefore, the integration of these omics is necessary for fully detecting microbial community. In a recent study, researchers succeeded to correlate the process of permafrost thawing with microbial composition and functions, using "multi-omics analysis" (Hultman et al., 2015).

THE CONNECTION BETWEEN MICROBIOTA AND THE HUMAN BODY

The dietary intake (Wu et al., 2011; Liu et al., 2018) and environmental exposure such as administration of antibiotics (Pérez-Cobas et al., 2012; Raymond et al., 2016) can largely influence human gut microbiota. The gut microbiota would then respond to these factors, producing signals adjusting human distal organs including liver (Khalsa et al., 2017), brain (Dinan and Cryan, 2017), and lung (Budden et al., 2017), as described in **Figure 2**. Both of microbes' own structural components and metabolites produced by them can serve as the signal molecules.

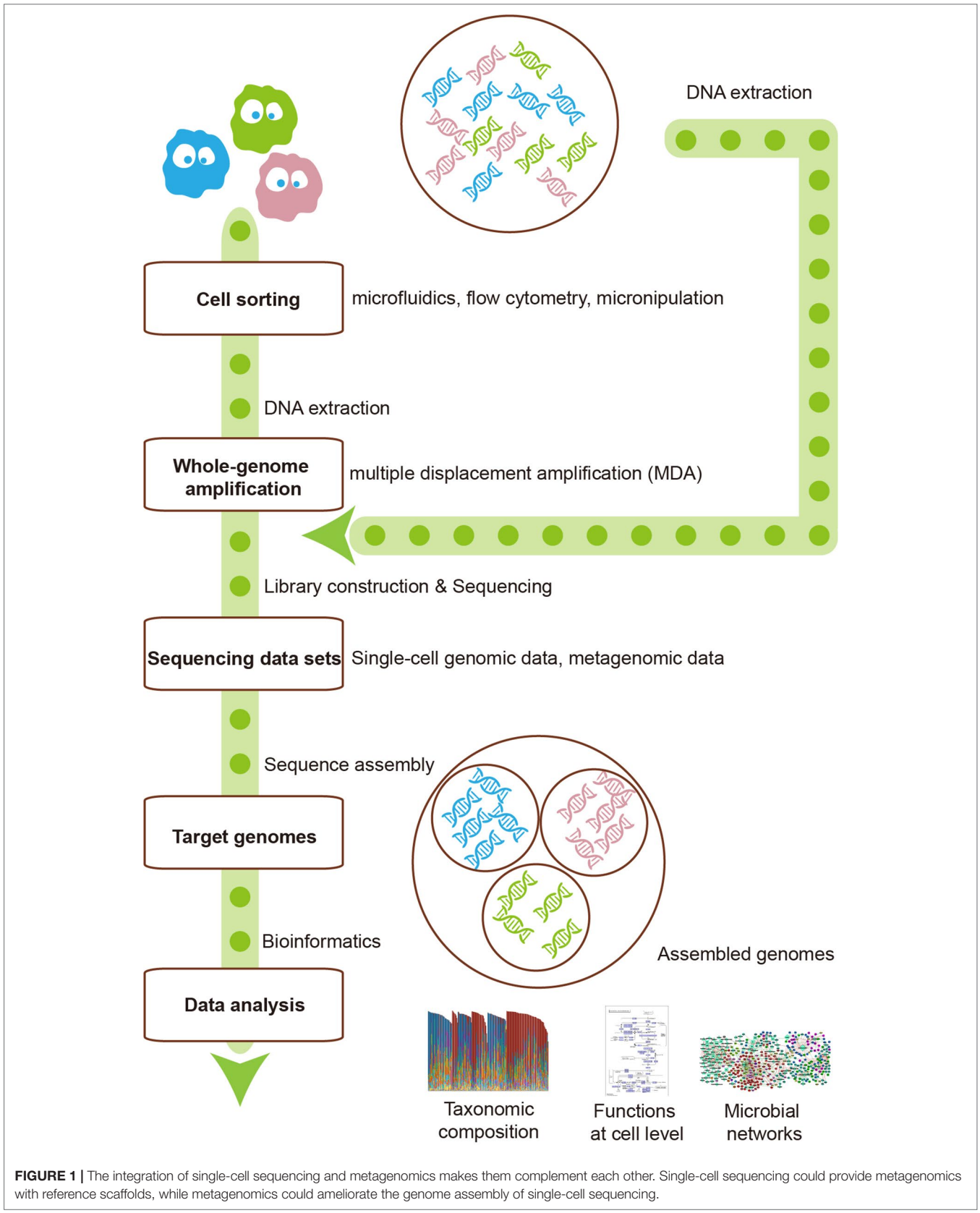


FIGURE 1 | The integration of single-cell sequencing and metagenomics makes them complement each other. Single-cell sequencing could provide metagenomics with reference scaffolds, while metagenomics could ameliorate the genome assembly of single-cell sequencing.

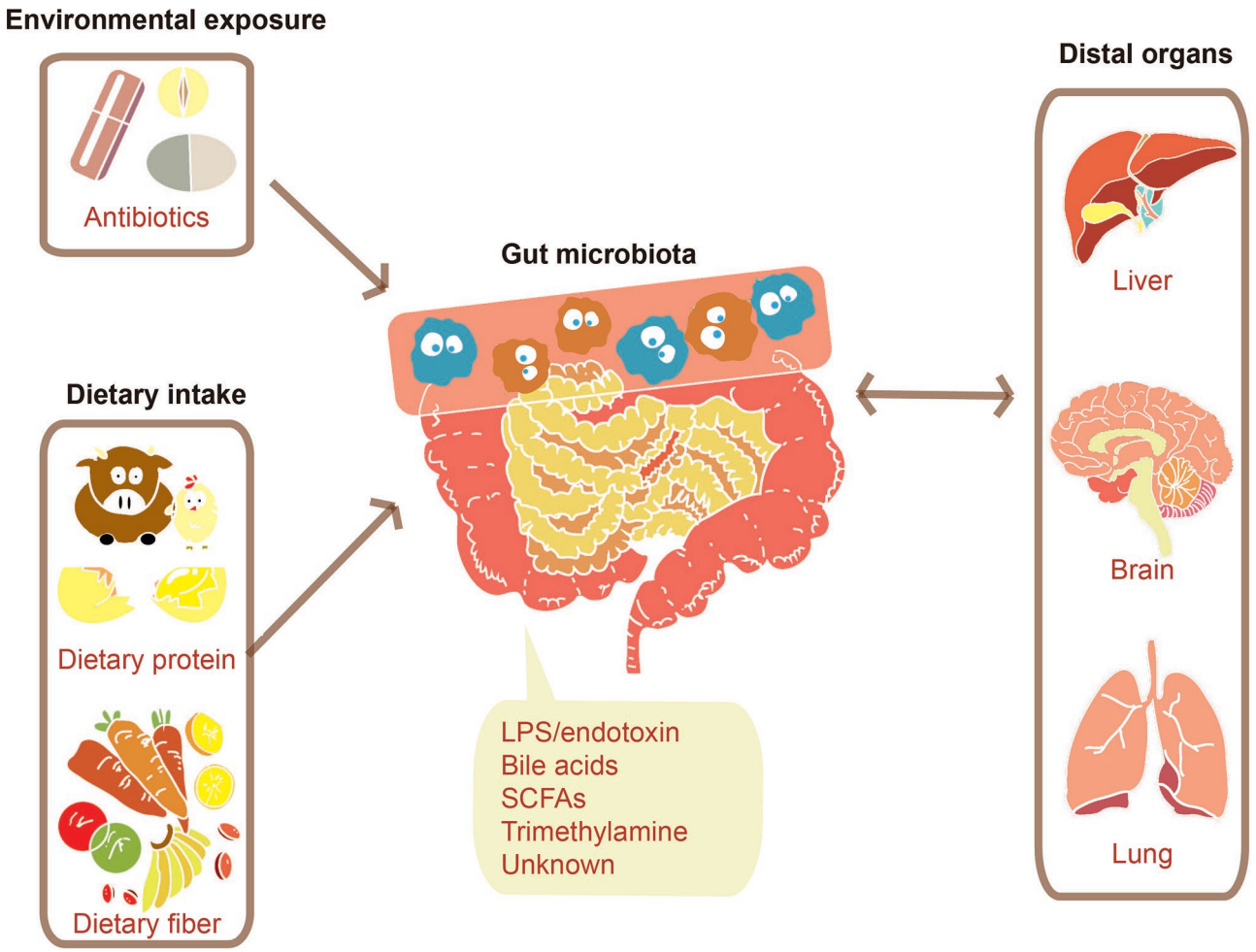


FIGURE 2 | Communications between the gut microbiome and distal organs. Various factors such as environmental exposure and dietary intake can modulate gut microbiota. The change of gut microbiota will bring a certain number of effects on distal organs through signals molecules consisting of their structural components such as lipopolysaccharide (LPS) and their metabolites such as SCFAs.

These signals can affect distal organs metabolism either directly or by signaling through nerves or hormones from the gut (Schroeder and Bäckhed, 2016).

Gut–Liver Axis

The gut microbiota was confirmed to adjust liver metabolism (Kim et al., 2007; Khalsa et al., 2017). BAs, for example, derived from cholesterol in the liver, can be modified by microbiota in the distal small intestine and colon (Schroeder and Bäckhed, 2016). Primary BAs will be deconjugated by the ileal gut microbiota after they are secreted into the small intestine, which makes them manage to escape the reabsorption and then be subjected to further chemical modification by colonic microbiota (Midtvedt, 1974; Swann et al., 2011). BAs are capable of activating nuclear receptors such as farnesoid X receptor (FXR) and G-protein–coupled receptors (GPCRs), which are associated with host metabolism (Fiorucci et al., 2009). The activation of FXR can suppress the rate-limiting step in BA synthesis through a gut microbiota–liver feedback loop, thus controlling the BA levels (Kim et al., 2007). Additionally, TGR5, one of GPCRs, predominately recognizes secondary BAs, which is associated with increased thermogenesis in brown adipose tissue (Broeders et al., 2015). The adjustment of the gut microbiota on the liver is important, while the response of liver cells is important as well, which can be described using single-cell sequencing. A recent study used single-cell RNA sequencing on T cells from hepatocellular carcinoma patients to identify 11 T-cell subsets with special molecular and functional properties, thus contributing to the prediction of their clinical responses in liver cancer (Zheng et al., 2017).

Gut–Brain Axis

The association between the brain and other organs depends on complex pathways consisting of the dual autonomic nervous system and endocrine. The gut–brain axis is defined to encompass afferent and efferent neural, endocrine, and

nutrient signals between the central nervous system and the gastrointestinal system (Romijn et al., 2008). Several studies have shown that the gut microbiota influences our brain morphology and stress response and even causes the stroke (Schroeder and Bäckhed, 2016) via the gut–brain axis. As for brain morphology, most studies were performed using mice due to the challenges in humans. Through the comparison between germ-free mice and colonized mice, the gut microbiota has been found to cause alterations in the structural integrity of the amygdala and hippocampus (Luczynski et al., 2016). Germ-free mice displayed increased hippocampal neurogenesis and hypermyelination of the prefrontal cortex (Hoban et al., 2016). Moreover, a more permeable blood–brain barrier (BBB) in germ-free mice suggests that the gut microbiota is also capable of modulating the BBB (Braniste et al., 2014). In respect to stress response, *Bifidobacterium longum* was observed to activate the vagus nerve to reduce anxiety-like behavior independently of brain-derived neurotrophic factor (Bercik et al., 2011). Moreover, different community members may have distinct influences on the stress response. For instance, when young germ-free mice with originally elevated stress response were colonized with *Bifidobacterium infantis* at an early developing stage, the stress response was then diminished. But when they were colonized with enteropathogenic *Escherichia coli*, their stress responses were observed to aggravate (Sudo et al., 2004). As to the stroke, 87% are ischemic and caused by interruption of the blood supply to the brain. A study displayed that ischemic brain injury in mice can be reduced by antibiotic-induced alterations in the gut microbiota (Benakis et al., 2016), which provided us with a potential therapeutic method in the future. The characterization of brain cells is important for researchers to further explore the gut–brain axis. Recently, a study performed single-cell sequencing, integrated with multi-omics on the human brain, providing new insights into complex processes in the brain (Lake et al., 2018).

Gut–Lung Axis

The conception of the gut–lung axis has emerged these years, which still needs more investigations to excavate mechanisms. First, dietary intake can shape both the gut microbiota and the airway microbiota (Marsland et al., 2015). On the one hand, dietary fiber intake leads to an increased level of short-chain fatty acids (SCFAs), which is associated with shifts in both gut microbiota and airway microbiota (Trompette et al., 2014). On the other hand, a high-fat diet has been confirmed to correlate with compositional changes in intestinal microbiota and elevated allergic airway inflammation (Myles et al., 2013). Second, the gut–lung axis contains several interactions among microbiota, metabolites, immune cells, and the lung. Bacterial metabolites such as SCFAs, with the ability to reach other organs *via* the bloodstream, are able to exert their anti-inflammatory properties. Additionally, the microbial seeding from the intestinal microbiota into the airways makes these bacteria able to act on local immune cells to shape their responses (Marsland et al., 2015). Moreover, migrating immune cells are capable of acquiring information directly from microbiota and the concomitant local cytokine response to adjust inflammatory response, which shapes immune responses at distal sites such as the lung (Trompette et al., 2014; Budden et al., 2017). Scientists have correlated allergic asthma, one of the lung diseases, with the gut microbiota. A study displayed that a fecal transplant from a child at risk of asthma into germ-free mice resulted in severe lung inflammation after challenge with ovalbumin (Arrieta et al., 2015). Moreover, another study showed that the impacts by recurrent antibiotic treatment on the diversity of the microbiota early in life (Fouhy et al., 2012) have been confirmed to strongly correlate with the development of an asthmatic phenotype later in life (Fanaro et al., 2003). There are still a certain number of unknown mechanisms in the gut–lung axis, which provides us with a lot of potential therapeutic methods against lung diseases.

MICROBIOTA AND CLINICAL MEDICINE

Gastrointestinal Disease

The intestine is a critical organ in the human's body, whose functions involve the uptake of nutrients and water. The intestinal barrier (**Figure 3**), as the essential barrier of the intestine, prevents the transfer of harmful substances and pathogens. Pathogenic bacteria may cause the disruption of this barrier resulting in increased intestinal permeability. Enteropathogenic E. coli (EPEC), for instance, causes a loss of enterocyte microvilli and the formation of a raised pedestal structure for firm bacterial attachment (Lapointe et al., 2009). In addition, enterohemorrhagic E. coli also possesses an attaching and effacement locus but with less profound effects on the barrier (Kaper and Nataro, 2004). Moreover, enteroaggregative E. coli and enterotoxigenic E. coli can cause diarrhea through effects on chloride secretion in the intestinal epithelium (Dubreuil, 2012). The single-cell sequencing helps to identify the pathogenic microbes at the intestinal lumen. The main antibody isotype named immunoglobulin A (IgA), which is produced at mucosal surfaces, can bind those pathogenic microbes in the intestinal lumen. The cell sorting then uses a fluorescent anti-IgA antibody, followed by 16S rDNA sequencing to identify the isolated pathogenic microbes (Palm et al., 2014). Furthermore, metagenomic sequencing can also be performed on these isolated microbes to identify the basis of immunogenic differences between and within microbes. Similarly, the elevated IgG coating of gut bacteria has also been observed in patients with sepsis and Crohn disease system (Zeng et al., 2016). Therefore, the single-cell sequencing is a promising method to correlate microbes with host immune response for precision medicine (Tolonen and Xavier, 2017).

Thrombosis

The risk of thrombosis has been observed to be correlated with the plasma levels of trimethylamine (TMA)–N-oxide (TMAO) in humans (Zhu et al., 2016). Especially, the gut microbiome is critically involved in the generation of TMAO (Tang et al., 2013). The gut microbiome can process certain dietary nutrients such as phosphatidylcholine, choline, and carnitine specifically to

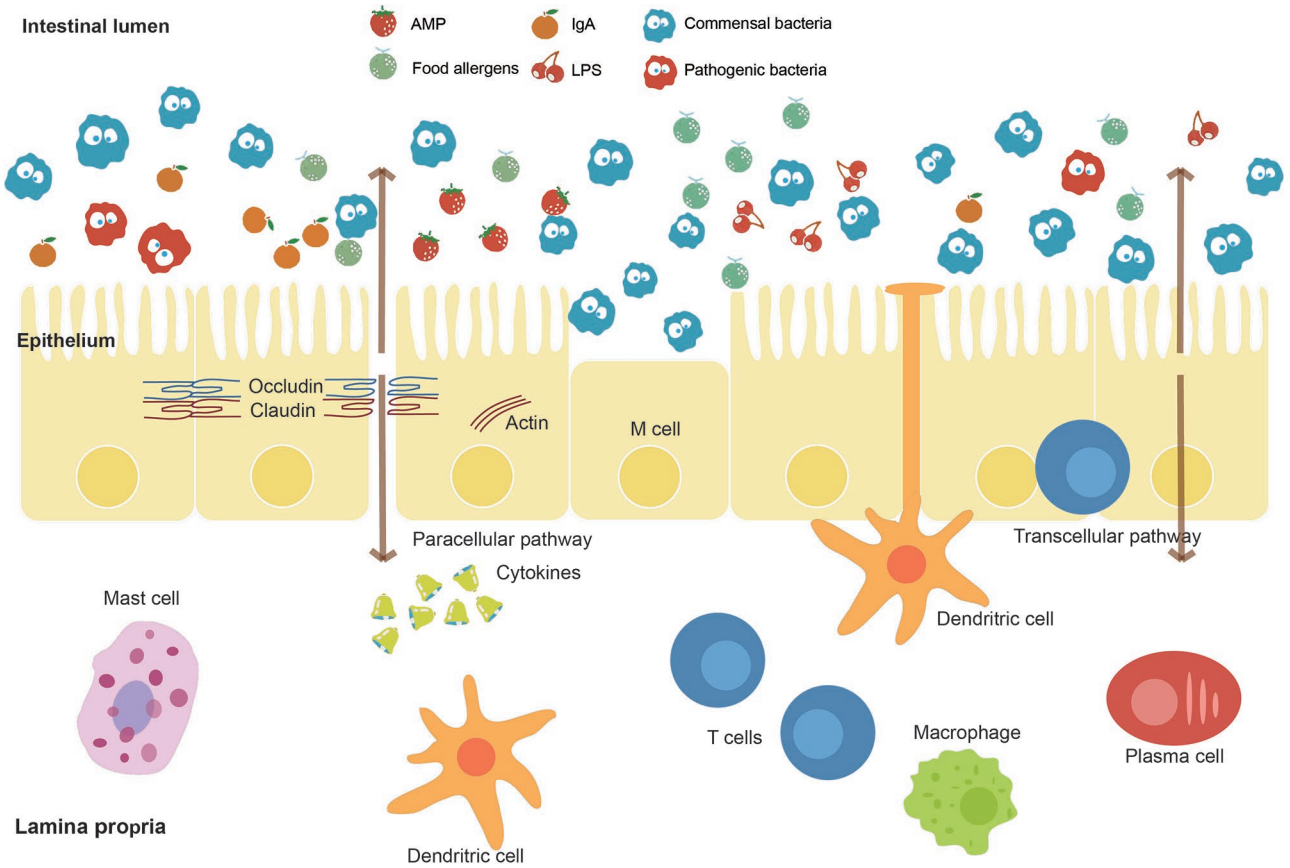


FIGURE 3 | Intestinal barrier and affecting factors. The intestinal barrier, as an essential barrier against harmful pathogens and substances in the intestine, mainly consists of the mucus layer, the epithelial layer, and the underlying lamina propria. The intestinal lumen contains antimicrobial peptides (AMPs), secreted IgA, and commensal bacteria, which prevent the colonization of pathogens. A mucus layer covers the intestinal surfaces as a physical barrier. The epithelium is composed of a single layer of cells sealed by tight junction proteins such as occludin and claudin inhibiting paracellular passage. M cells and intraepithelial lymphocytes are also contained in this layer. The lamina propria harbors lots of immune cells. Factors including food allergens, lipopolysaccharides (LPS), and pathogenic bacteria such as EPEC effect on the intestinal barrier function.

procedure TMA, which is absorbed in the gut and converted in the liver to TMAO by hepatic flavin-containing monooxygenases (Tilg, 2016). In humans, foods such as meat and eggs have been associated with an increased risk of major cardiovascular events in patients with proven coronary heart disease (Tang et al., 2013). In addition, administration of antibiotics can markedly reduce the plasma levels of TMAO.

Hepatitis B Virus

Hepatitis B virus (HBV), as one of the most common infectious agents worldwide, has been associated with the gut microbiome (Chou et al., 2015). Scientists have found that viral clearance heavily depends on the age of exposure. According to the control experiments of adult and young mice, the results showed an immune-tolerating pathway to HBV that prevailed in young mice with immature gut microbiota. After the establishment of gut bacteria, the mature gut microbiota in adult mice stimulated liver immunity, resulting in rapid HBV clearance (Chou et al., 2015). Therefore, full understanding of the interaction of virus–host may help us with the therapy for HBV. The single-cell sequencing can serve as a powerful method to explore the virus–host interaction (Labonte et al., 2015).

Depression

Depressive episodes correlate with dysregulation of the hypothalamic–pituitary–adrenal (HPA) axis (Barden, 2004) and resolution of depressive systems with normalization of the HPA axis (Heuser et al., 1996; Nickel et al., 2003). The gut microbiota has been confirmed to play a part in both the programming of the HPA axis early in life and stress reactivity over the life span (Foster and Neufeld, 2013). The stress response system is functionally immature at birth and then develops throughout the postnatal period, which coincides with the intestinal bacterial colonization. Stress can increase intestinal permeability, providing bacteria with an opportunity to translocate across the intestinal mucosa and directly access both immune cells and neuronal cells of the enteric nervous system (Gareau et al., 2008; Teitelbaum et al., 2008).

AIDS

The gut microbiota has been recently observed to be associated with human immunodeficiency virus (HIV) disease progression (Vujkovic-Cvijin et al., 2013). Scientists identified a dysbiotic mucosal-adherent community enriched in Proteobacteria and depleted of Bacteroidia members that were associated with markers of mucosal immune disruption, T-cell activation, and chronic inflammation in HIV-infected subjects. This dysbiotic community was evident among HIV-infected subjects undergoing highly active antiretroviral therapy (Vujkovic-Cvijin et al., 2013). Furthermore, the extent of dysbiosis correlated with two established markers of disease progression including the activity of the kynurenine pathway of tryptophan catabolism and plasma concentrations of the inflammatory cytokine interleukin 6 (Vujkovic-Cvijin et al., 2013). Hence, a link between mucosal-adherent colonic bacteria and immunopathogenesis during progressive HIV infection deserves better investigations.

Cancer

Gut microbes have been reported to be correlated with a certain number of cancers related to human stomach (*Helicobacter pylori*), liver (*Opisthorchis viverrini*, *Clonorchis sinensis*), and bladder (*Schistosoma haematobium*) (Bhatt et al., 2017). *H. pylori* infections, for instance, can lead to gastritis and gastric ulcers (Marshall et al., 1984), which is considered as the precursor of gastric cancer. Nevertheless, *H. pylori* was also observed to protect against esophageal adenocarcinoma, by influencing stomach pH and ameliorating acid reflux (Vaezi et al., 2000). Hence, owing to the participation of microbes in multiple biological processes, the oncogenicity of microbes should be discussed and determined by multi-omics approaches.

THE TREND OF BIG-DATA MINING FOR MICROBIOME

In the past, owing to limitations in abilities to obtain and process microbial big data, scientists were not able to obtain a full understanding of the microbiota. Neither the sequencing technologies nor the analysis tools can meet the high dimensional complicacy of the intestinal microbiota. Nowadays, the high-throughput sequencing technologies, such as MDA (Dean et al., 2002) for single-cell sequencing, and numerous statistical analysis tools, such as QIIME for 16S sequencing data (Caporaso et al., 2010) and MetaPhlAn (Segata et al., 2012) for metagenomics data, make it possible to unveil the microbiota from various perspectives. The integration of the current sequencing methods would be necessary to conduct a comprehensive study on microbiota in the future. First, the taxonomic information at various levels can be obtained by amplicon sequencing and metagenomic sequencing. Second, the functional annotation can be predicted by metagenomics and confirmed by the multi-omics including metagenome, metatranscriptome, metaproteome, and metabolome. Third, the connection between functions and phylogeny of a single microbe cell can be established by single-cell sequencing. Finally, the interactions between all chromosomes can be detected by Hi-C sequencing. The integration of these methods can answer the questions "who is there," "what are they doing," and "how are they doing" from a macroscopic level of overall microbial composition and microscopic level of single microbe cell and even the single chromosome. The comprehensive analysis of big data, followed by strict *in vivo* and *in vitro* experiments, is required to determine the causality of clinical diseases by microbes for specific medicine. Moreover, a standard pipeline for the integration of these methods proposed in the future can produce a huge amount of data sets. The big-data sets across continents provide the spatial characteristics, and the big-data sets in the long-term investigations provide the characteristics at time scale.

AUTHOR CONTRIBUTIONS

KN conceived the review framework. MC and LC conducted the literature review. MC made the figure illustration. MC and LC wrote the manuscript. KN reviewed and revised the manuscript.

FUNDING

This work was partially supported by the National Key R&D Program of China (grant 2018YFC0910502) and the National Natural Science Foundation of China (grants 61103167, 31271410, and 31671374).

ACKNOWLEDGMENTS

The authors would like to thank Pengshuo Yang, PhD, Maozhen HAN, PhD, Chaoyun Chen, PhD, Chaofang Zhong, PhD, Department of Bioinformatics and Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology, for their discussions about this work.

REFERENCES

Arrieta, M. C., Stiemsma, L. T., Dimitriu, P. A., Thorson, L., Russell, S., Yurist-Doutsch, S., et al. (2015). Early infancy microbial and metabolic alterations affect risk of childhood asthma. *Sci. Transl. Med.* 7 (307), 307ra152–307ra152. doi: 10.1126/scitranslmed.aab2271

Bäckhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. A., and Gordon, J. I. (2005). Host-bacterial mutualism in the human intestine. *Science* 307 (5717), 1915–1920. doi: 10.1126/science.1104816

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19 (5), 455–477. doi: 10.1089/cmb.2012.0021

Barden, N. (2004). Implication of the hypothalamic–pituitary–adrenal axis in the physiopathology of depression. *J. Psychiatry. Neurosci.* 29 (3), 185.

Belaghzal, H., Dekker, J., and Gibcus, J. H. (2017). Hi-C 2.0: an optimized Hi-C procedure for high-resolution genome-wide mapping of chromosome conformation. *Methods* 123, 56–65. doi: 10.1016/j.ymeth.2017.04.004

Benakis, C., Brea, D., Caballero, S., Faraco, G., Moore, J., Murphy, M., et al. (2016). Commensal microbiota affects ischemic stroke outcome by regulating intestinal γδ T cells. *Nat. Med.* 22 (5), 516–523. doi: 10.1038/nm.4068

Bercik, P., Park, A., Sinclair, D., Khoshdel, A., Lu, J., Huang, X., et al. (2011). The anxiolytic effect of *Bifidobacterium longum* NCC3001 involves vagal pathways for gut–brain communication. *Neurogastroenterol. Motil.* 23 (12), 1132–1139. doi: 10.1111/j.1365-2982.2011.01796.x

Berry, D., Stecher, B., Schintlmeister, A., Reichert, J., Brugiroux, S., Wild, B., et al. (2013). Host-compound foraging by intestinal microbiota revealed by single-cell stable isotope probing. *Proc. Natl. Acad. Sci. U.S.A.* 110 (12), 4720–4725. doi: 10.1073/pnas.1219247110

Bertrand, D., Shaw, J., Kalathiyappan, M., Ng, A. H. Q., Kumar, M. S., Li, C., et al. (2019). Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. *Nat. Biotechnol.* 37 (8), 937–944. doi: 10.1038/s41587-019-0191-2

Bhatt A. P., Redinbo M. R., and Bultman S. J. (2017). The role of the microbiome in cancer development and therapy. *CA. Cancer. J. Clin.* 67(4), 326–344. doi: 10.3322/caac.21398

Blanco, L., Bernad, A., Lazaro, J. M., Martin, G., Garmendia, C., and Salas, M. (1989). Highly efficient DNA synthesis by the phage phi 29 DNA polymerase. Symmetrical mode of DNA replication. *J. Biol. Chem.* 264 (15), 8935–8940.

Braniste, V., Al-Asmakh, M., Kowal, C., Anuar, F., Abbaspour, A., Tóth, M., et al. (2014). The gut microbiota influences blood–brain barrier permeability in mice. *Sci. Transl. Med.* 6 (263), 263ra158–263ra158. doi: 10.1126/scitranslmed.3009759

Broeders, E. P., Nascimento, E. B., Havekes, B., Brans, B., Roumans, K. H., Tailleux, A., et al. (2015). The bile acid chenodeoxycholic acid increases human brown adipose tissue activity. *Cell. Metab.* 22 (3), 418–426. doi: 10.1016/j.cmet.2015.07.002

Budden, K. F., Gellatly, S. L., Wood, D. L., Cooper, M. A., Morrison, M., Hugenholtz, P., et al. (2017). Emerging pathogenic links between microbiota and the gut–lung axis. *Nat. Rev. Microbiol.* 15 (1), 55–63. doi: 10.1038/nrmicro.2016.142

Campbell, J. H., O'Donoghue, P., Campbell, A. G., Schwientek, P., Sczyrba, A., Woyke, T., et al. (2013). UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. *Proc. Natl. Acad. Sci. U.S.A.* 110 (14), 5540–5545. doi: 10.1073/pnas.1303090110

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7 (5), 335–336. doi: 10.1038/nmeth.f.303

Chou, H.-H., Chien, W.-H., Wu, L.-L., Cheng, C.-H., Chung, C.-H., Horng, J.-H., et al. (2015). Age-related immune clearance of hepatitis B virus infection requires the establishment of gut microbiota. *Proc. Natl. Acad. Sci. U.S.A.* 112 (7), 2175–2180. doi: 10.1073/pnas.1424775112

Dean, F. B., Hosono, S., Fang, L., Wu, X., Faruqi, A. F., Bray-Ward, P., et al. (2002). Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl. Acad. Sci. U.S.A.* 99 (8), 5261–5266. doi: 10.1073/pnas.082089499

Dicksved, J., Halfvarson, J., Rosenquist, M., Jarnerot, G., Tysk, C., Apajalahti, J., et al. (2008). Molecular analysis of the gut microbiota of identical twins with Crohn's disease. *ISME J.* 2 (7), 716–727. doi: 10.1038/ismej.2008.37

Dinan, T. G., and Cryan, J. F. (2017). The microbiome–gut–brain axis in health and disease. *Gastroenterol. Clin. North. Am.* 46 (1), 77–89. doi: 10.1016/j.gtc.2016.09.007

Dubreuil, J. D. (2012). The whole Shebang: the gastrointestinal tract, *Escherichia coli* enterotoxins and secretion. *Curr. Issues Mol. Biol.* 14 (2), 71–82.

Dupont, C. L., Rusch, D. B., Yooseph, S., Lombardo, M. J., Richter, R. A., Valas, R., et al. (2012). Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *ISME J.* 6 (6), 1186–1199. doi: 10.1038/ismej.2011.189

Erickson, A. R., Cantarel, B. L., Lamendella, R., Darzi, Y., Mongodin, E. F., Pan, C., et al. (2012). Integrated metagenomics/metaproteomics reveals human host–microbiota signatures of Crohn's disease. *PLoS One* 7 (11), e49138. doi: 10.1371/journal.pone.0049138

Fanaro, S., Chierici, R., Guerrini, P., and Vigi, V. (2003). Intestinal microflora in early infancy: composition and development. *Acta Paediatr. Suppl.* 91 (s441), 48–55. doi: 10.1111/j.1651-2227.2003.tb00646.x

Fiorucci, S., Mencarelli, A., Palladino, G., and Cipriani, S. (2009). Bile-acid–activated receptors: targeting TGR5 and farnesoid-X-receptor in lipid and glucose disorders. *Trends Pharmacol. Sci.* 30 (11), 570–580. doi: 10.1016/j.tips.2009.08.001

Foster, J. A., and Neufeld, K.-A. M. (2013). Gut–brain axis: how the microbiome influences anxiety and depression. *Trends. Neurosci.* 36 (5), 305–312. doi: 10.1016/j.tins.2013.01.005

Fouhy, F., Guinane, C. M., Hussey, S., Wall, R., Ryan, C. A., Dempsey, E. M., et al. (2012). High-throughput sequencing reveals the incomplete, short-term recovery of infant gut microbiota following parenteral antibiotic treatment with ampicillin and gentamicin. *Antimicrob. Agents Chemother.* 56 (11), 5811–5820. doi: 10.1128/AAC.00789-12

Frank, D. N., Amand, A. L. S., Feldman, R. A., Boedeker, E. C., Harpaz, N., and Pace, N. R. (2007). Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc. Natl. Acad. Sci. U.S.A.* 104 (34), 13780–13785. doi: 10.1073/pnas.0706625104

Gareau, M. G., Silva, M. A., and Perdue, M. H. (2008). Pathophysiological mechanisms of stress-induced intestinal damage. *Front. Physiol.* 8 (4), 274–281. doi: 10.3389/fphys.2018.00441

Gill, S. R., Pop, M., DeBoy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., et al. (2006). Metagenomic analysis of the human distal gut microbiome. *Science* 312 (5778), 1355–1359. doi: 10.1126/science.1124234

Heuser, I. J., Schweiger, U., Gotthardt, U., and Schmider, J. (1996). Pituitary–adrenal-system regulation and psychopathology during amitriptyline treatment in elderly depressed patients and normal comparison subjects. *Am. J. Psychiatry.* 153 (1), 93. doi: 10.1176/ajp.153.1.93

Hoban, A., Stilling, R., Ryan, F., Shanahan, F., Dinan, T., Claesson, M., et al. (2016). Regulation of prefrontal cortex myelination by the microbiota. *Transl. Psychiatry* 6 (4), e774. doi: 10.1038/tp.2016.42

Hooper, L. V., and Gordon, J. I. (2001). Commensal host–bacterial relationships in the gut. *Science* 292 (5519), 1115–1118. doi: 10.1126/science.1058709

Hultman, J., Waldrop, M. P., Mackelprang, R., David, M. M., McFarland, J., Blazewicz, S. J., et al. (2015). Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. *Nature* 521 (7551), 208–212. doi: 10.1038/nature14238

Jansson, J., Willing, B., Lucio, M., Fekete, A., Dicksved, J., Halfvarson, J., et al. (2009). Metabolomics reveals metabolic biomarkers of Crohn's disease. *PLoS One* 4 (7), e6386. doi: 10.1371/journal.pone.0006386

Kaper, J., and Nataro, J. (2004). Mobley, HLT. Pathogenic *Escherichia coli. Nat. Rev. Microbiol.* 2 (2), 123–140. doi: 10.1038/nrmicro818

Khalsa, J., Duffy, L. C., Riscuta, G., Starke-Reed, P., and Hubbard, V. S. (2017). Omics for understanding the gut–liver–microbiome axis and precision medicine. *Clin. Pharmacol. Drug. Dev.* 6 (2), 176–185. doi: 10.1002/cpdd.310

Kim, I., Ahn, S.-H., Inagaki, T., Choi, M., Ito, S., Guo, G. L., et al. (2007). Differential regulation of bile acid homeostasis by the farnesoid X receptor in liver and intestine. *J. Lipid. Res.* 48 (12), 2664–2672. doi: 10.1194/jlr.M700330-JLR200

Labonte, J. M., Swan, B. K., Poulos, B., Luo, H., Koren, S., Hallam, S. J., et al. (2015). Single-cell genomics–based analysis of virus–host interactions in marine surface bacterioplankton. *ISME J.* 9 (11), 2386–2399. doi: 10.1038/ismej.2015.48

Lake, B. B., Chen, S., Sos, B. C., Fan, J., Kaeser, G. E., Yung, Y. C., et al. (2018). Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat. Biotechnol.* 36 (1), 70–80. doi: 10.1038/nbt.4038

Lapointe, T. K., O'connor, P. M., and Buret, A. G. (2009). The role of epithelial malfunction in the pathogenesis of enteropathogenic *E. coli*–induced diarrhea. *Lab. Invest.* 89 (9), 964. doi: 10.1038/labinvest.2009.69

Liu, H., Han, M., Li, S. C., Tan, G., Sun, S., Hu, Z., et al. (2018). Resilience of human gut microbial communities for the long stay with multiple dietary shifts. *Gut*, 1–2. doi: 10.1136/gutjnl-2018-317298

Luczynski, P., Whelan, S. O., O'sullivan, C., Clarke, G., Shanahan, F., Dinan, T. G., et al. (2016). Adult microbiota-deficient mice have distinct dendritic morphological changes: differential effects in the amygdala and hippocampus. *Eur. J. Neurosci.* 44 (9), 2654–2666. doi: 10.1111/ejn.13291

Manichanh, C., Borruel, N., Casellas, F., and Guarner, F. (2012). The gut microbiota in IBD. *Nat. Rev. Gastroenterol. Hepatol.* 9 (10), 599–608. doi: 10.1038/nrgastro.2012.152

Marshall B. J., and Warren J. R. (1984). Unidentified curved bacilli in the stomach of patients with gastritis and peptic ulceration. *Lancet* 1(8390), 1311–1315. doi: 10.1016/s0140-6736(84)91816-6

Marsland, B. J., Trompette, A., and Gollwitzer, E. S. (2015). The gut–lung axis in respiratory disease. *Ann. Am. Thorac. Soc.* 12 (Supplement 2), S150–S156. doi: 10.1513/AnnalsATS.201503-133AW

Mende, D. R., Aylward, F. O., Eppley, J. M., Nielsen, T. N., and DeLong, E. F. (2016). Improved environmental genomes *via* integration of metagenomic and single-cell assemblies. *Front. Microbiol.* 7, 143. doi: 10.3389/fmicb.2016.00143

Midtvedt, T. (1974). Microbial bile acid transformation. *Am. J. Clin. Nutr.* 27 (11), 1341–1347. doi: 10.1093/ajcn/27.11.1341

Myles, I. A., Fontecilla, N. M., Janelsins, B. M., Vithayathil, P. J., Segre, J. A., and Datta, S. K. (2013). Parental dietary fat intake alters offspring microbiome and immunity. *J. Immunol.* 191 (6), 3200–3209. doi: 10.4049/jimmunol.1301057

Nagano, T., Wingett, S. W., and Fraser, P. (2017). Capturing Three-Dimensional Genome Organization in Individual Cells by Single-Cell Hi-C. *Methods. Mol. Biol.* 1654, 79–97. doi: 10.1007/978-1-4939-7231-9_6

Nickel, T., Sonntag, A., Schill, J., Zobel, A. W., Ackl, N., Brunnauer, A., et al. (2003). Clinical and neurobiological effects of tianeptine and paroxetine in major depression. *J. Clin. Psychopharmacol.* 23 (2), 155–168. doi: 10.1097/00004714-200304000-00008

Nobu, M. K., Narihiro, T., Rinke, C., Kamagata, Y., Tringe, S. G., Woyke, T., et al. (2015). Microbial dark matter ecogenomics reveals complex synergistic networks in a methanogenic bioreactor. *ISME J.* 9 (8), 1710–1722. doi: 10.1038/ismej.2014.256

Palm, N. W., de Zoete, M. R., Cullen, T. W., Barry, N. A., Stefanowski, J., Hao, L., et al. (2014). Immunoglobulin A coating identifies colitogenic bacteria in inflammatory bowel disease. *Cell* 158 (5), 1000–1010. doi: 10.1016/j.cell.2014.08.006

Pérez-Cobas, A. E., Gosalbes, M. J., Friedrichs, A., Knecht, H., Artacho, A., Eismann, K., et al. (2012). Gut microbiota disturbance during antibiotic therapy: a multi-omic approach. *Gut* 62, 1591–1601. doi: 10.1136/gutjnl-2012-303184

Props, R., Kerckhof, F.-M., Rubbens, P., De Vrieze, J., Sanabria, E. H., Waegeman, W., et al. (2017). Absolute quantification of microbial taxon abundances. *ISME J.* 11 (2), 584. doi: 10.1038/ismej.2016.117

Raymond, F., Ouameur, A. A., Déraspe, M., Iqbal, N., Gingras, H., Dridi, B., et al. (2016). The initial state of the human gut microbiome determines its reshaping by antibiotics. *ISME J.* 10 (3), 707–720. doi: 10.1038/ismej.2015.148

Romijn, J. A., Corssmit, E. P., Havekes, L. M., and Pijl, H. (2008). Gut–brain axis. *Curr. Opin. Clin. Nutr. Metab. Care.* 11 (4), 518–521. doi: 10.1097/MCO.0b013e328302c9b0

Roux, S., Hawley, A. K., Torres Beltran, M., Scofield, M., Schwientek, P., Stepanauskas, R., et al. (2014). Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and meta-genomics. *Elife* 3, e03125. doi: 10.7554/eLife.03125

Schroeder, B. O., and Bäckhed, F. (2016). Signals from the gut microbiota to distant organs in physiology and disease. *Nat. Med.* 22 (10), 1079–1089. doi: 10.1038/nm.4185

Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Droge, J., et al. (2017). Critical assessment of metagenome interpretation—a benchmark of metagenomics software. *Nat. Methods* 14 (11), 1063–1071. doi: 10.1038/nmeth.4458

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Med.* 9 (8), 811–814. doi: 10.1038/nmeth.2066

Sender, R., Fuchs, S., and Milo, R. (2016). Revised estimates for the number of human and bacteria cells in the body. *PLoS Biol.* 14 (8), e1002533. doi: 10.1371/journal.pbio.1002533

Sharpton, T. J. (2014). An introduction to the analysis of shotgun metagenomic data. *Front. Plant. Sci.* 5, 209. doi: 10.3389/fpls.2014.00209

Sudo, N., Chida, Y., Aiba, Y., Sonoda, J., Oyama, N., Yu, X. N., et al. (2004). Postnatal microbial colonization programs the hypothalamic–pituitary–adrenal system for stress response in mice. *J. Physiol.* 2558 (1), 263–275. doi: 10.1113/jphysiol.2004.063388

Swan, B. K., Tupper, B., Sczyrba, A., Lauro, F. M., Martinez-Garcia, M., Gonzalez, J. M., et al. (2013). Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proc. Natl. Acad. Sci. U.S.A.* 110 (28), 11463–11468. doi: 10.1073/pnas.1304246110

Swann, J. R., Want, E. J., Geier, F. M., Spagou, K., Wilson, I. D., Sidaway, J. E., et al. (2011). Systemic gut microbial modulation of bile acid metabolism in host tissue compartments. *Proc. Natl. Acad. Sci. U.S.A.* 108 (Supplement 1), 4523–4530. doi: 10.1073/pnas.1006734107

Tang, W. W., Wang, Z., Levison, B. S., Koeth, R. A., Britt, E. B., Fu, X., et al. (2013). Intestinal microbial metabolism of phosphatidylcholine and cardiovascular risk. *N. Engl. J. Med.* 368 (17), 1575–1584. doi: 10.1056/NEJMoa1109400

Teitelbaum, A. A., Gareau, M. G., Jury, J., Yang, P. C., and Perdue, M. H. (2008). Chronic peripheral administration of corticotropin-releasing factor causes colonic barrier dysfunction similar to psychological stress. *Am. J. Physiol. Gastrointest. Liver Physiol.* 295 (3), G452–G459. doi: 10.1152/ajpgi.90210.2008

Tilg, H. (2016). A gut feeling about thrombosis. *N. Engl. J. Med.* 374 (25), 2494–2496. doi: 10.1056/NEJMcibr1604458

Tolonen, A. C., and Xavier, R. J. (2017). Dissecting the human microbiome with single-cell genomics. *Genome Med.* 9 (1), 56. doi: 10.1186/s13073-017-0448-7

Trompette, A., Gollwitzer, E. S., Yadava, K., Sichelstiel, A. K., Sprenger, N., Ngom-Bru, C., et al. (2014). Gut microbiota metabolism of dietary fiber influences allergic airway disease and hematopoiesis. *Nat. Med.* 20 (2), 159–166. doi: 10.1038/nm.3444

Vaezi M. F., Falk G. W., Peek R. M., Vicari J. J, Goldblum J. R. Perez G. I. et al. (2000). CagA-positive strains of Helicobacter pylori may protect against Barrett's esophagus. *Am. J. Gastroenterol.* 95 (9), 2206–2211. doi: 10.1111/j.1572-0241.2000.02305.x

Vujkovic-Cvijin, I., Dunham, R. M., Iwai, S., Maher, M. C., Albright, R. G., Broadhurst, M. J., et al. (2013). Dysbiosis of the gut microbiota is associated with HIV disease progression and tryptophan catabolism. *Sci. Transl. Med.* 5 (193), 193ra191–193ra191. doi: 10.1126/scitranslmed.3006438

Willing, B., Halfvarson, J., Dicksved, J., Rosenquist, M., Jarnerot, G., Engstrand, L., et al. (2009). Twin studies reveal specific imbalances in the mucosa-associated microbiota of patients with ileal Crohn's disease. *Inflamm. Bowel Dis.* 15 (5), 653–660. doi: 10.1002/ibd.20783

Willing, B. P., Dicksved, J., Halfvarson, J., Andersson, A. F., Lucio, M., Zheng, Z., et al. (2010). A pyrosequencing study in twins shows that gastrointestinal microbial profiles vary with inflammatory bowel disease phenotypes. *Gastroenterology* 139 (6), 1844–1854 e1841. doi: 10.1053/j.gastro.2010.08.049

Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y. Y., Keilbaugh, S. A., et al. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science* 334 (6052), 105–108. doi: 10.1126/science.1208344

Xu, Y., and Zhao, F. (2018). Single-cell metagenomics: challenges and applications. *Protein Cell* 9 (5), 501–510. doi: 10.1007/s13238-018-0544-5

Yoon, H. S., Price, D. C., Stepanauskas, R., Rajah, V. D., Sieracki, M. E., Wilson, W. H., et al. (2011). Single-cell genomics reveals organismal interactions in uncultivated marine protists. *Science* 332 (6030), 714–717. doi: 10.1126/science.1203163

Zeng, M. Y., Cisalpino, D., Varadarajan, S., Hellman, J., Warren, H. S., Cascalho, M., et al. (2016). Gut microbiota–induced immunoglobulin G controls systemic infection by symbiotic bacteria and pathogens. *Immunity* 44 (3), 647–658. doi: 10.1016/j.immuni.2016.02.006

Zheng, C., Zheng, L., Yoo, J. K., Guo, H., Zhang, Y., Guo, X., et al. (2017). Landscape of infiltrating T cells in liver cancer revealed by single-cell sequencing. *Cell* 169 (7), 1342–1356.e1316. doi: 10.1016/j.cell.2017.05.035

Zhu, W., Gregory, J. C., Org, E., Buffa, J. A., Gupta, N., Wang, Z., et al. (2016). Gut microbial metabolite TMAO enhances platelet hyperreactivity and thrombosis risk. *Cell* 165 (1), 111–124. doi: 10.1016/j.cell.2016.02.011

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 KN.

frontiers in Genetics

ORIGINAL RESEARCH
published: 24 October 2019
doi: 10.3389/fgene.2019.01031

Identification of Potential Biomarkers in Association With Progression and Prognosis in Epithelial Ovarian Cancer by Integrated Bioinformatics Analysis

Jinhui Liu[1†], Huangyang Meng[1†], Siyue Li[1†], Yujie Shen[2], Hui Wang[1], Wu Shan[1], Jiangnan Qiu[1], Jie Zhang[1] and Wenjun Cheng[1]*

[1] Department of Gynecology, The First Affiliated Hospital of Nanjing Medical University, Nanjing, China, [2] Department of Otorhinolaryngology, The First Affiliated Hospital of Nanjing Medical University, Nanjing, China

OPEN ACCESS

Edited by:
Tuo Zhang,
Cornell University,
United States

Reviewed by:
Hao Zhang,
Jilin University,
China
Andrew Dellinger,
Elon University,
United States

***Correspondence:**
Wenjun Cheng
wenjunchengdoc@163.com

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *13 May 2019*
Accepted: *25 September 2019*
Published: *24 October 2019*

Citation:
Liu J, Meng H, Li S, Shen Y, Wang H, Shan W, Qiu J, Zhang J and Cheng W (2019) Identification of Potential Biomarkers in Association With Progression and Prognosis in Epithelial Ovarian Cancer by Integrated Bioinformatics Analysis. Front. Genet. 10:1031. doi: 10.3389/fgene.2019.01031

Epithelial ovarian cancer (EOC) is one of the malignancies in women, which has the highest mortality. However, the microlevel mechanism has not been discussed in detail. The expression profiles GSE27651, GSE38666, GSE40595, and GSE66957 including 188 tumor and 52 nontumor samples were downloaded from the Gene Expression Omnibus database. The differentially expressed genes (DEGs) were filtered using R software, and we performed functional analysis using the clusterProfiler. Cytoscape software, the molecular complex detection plugin and database STRING analyzed DEGs to construct protein-protein interaction network. We identified 116 DEGs including 81 upregulated and 35 downregulated DEGs. Functional analysis revealed that they were significantly enriched in the extracellular region and biosynthesis of amino acids. We next identified four bioactive compounds (vorinostat, LY-294002,trichostatin A, and tanespimycin) based on ConnectivityMap. Then 114 nodes were obtained in protein–protein interaction. The three most relevant modules were detected. In addition, according to degree ≥ 10, 14 core genes including FOXM1, CXCR4, KPNA2, NANOG, UBE2C, KIF11, ZWINT, CDCA5, DLGAP5, KIF15, MCM2, MELK, SPP1, and TRIP13 were identified. Kaplan–Meier analysis, Oncomine, and Gene Expression Profiling Interactive Analysis showed that overexpression of FOXM1, SPP1, UBE2C, KIF11, ZWINT, CDCA5, UBE2C, and KIF15 was related to bad prognosis of EOC patients. CDCA5, FOXM1, KIF15, MCM2, and ZWINT were associated with stage. Receiver operating characteristic (ROC) curve showed that messenger RNA levels of these five genes exhibited better diagnostic efficiency for normal and tumor tissues. The Human Protein Atlas database was performed. The protein levels of these five genes were significantly higher in tumor tissues compared with normal tissues. Functional enrichment analysis suggested that all the hub genes played crucial roles in citrate cycle tricarboxylic acid cycle. Furthermore, the univariate and multivariate Cox proportional hazards regression showed that ZWINT was independent prognostic indictor among EOC patients. The genes and pathways discovered in the above studies may open a new direction for EOC treatment.

Keywords: epithelial ovarian cancer, bioinformatical analysis, differentially expressed genes, prognosis, Cmap, protein–protein interaction, biomarker

INTRODUCTION

Ovarian cancer is the second most common female malignant tumor in the world and the most common cause of death among female malignant tumors (McAlpine et al., 2014). With the development of the times, although surgery and other treatment methods have been improved, the treatment effect and prognosis of advanced ovarian cancer patients are very poor due to the difficulty in the diagnosis of ovarian cancer (Allemani et al., 2015; La Vecchia, 2017).

Gene expression microarray, as a means of efficient large-scale acquisition of genetic data, has been generally used to collect and study gene chip expression profiling data of many human cancers. New methods are provided by microarrays for studying tumor-associated genes, molecular targeting, molecular prediction, and therapy. The integration of databases where researchers have published their research data containing several gene expression chips allows for a more in-depth study of molecular mechanisms (Nannini et al., 2009; Petryszak et al., 2014).

In this study, we downloaded four gene expression datasets, GSE27651, GSE38666, GSE40595, and GSE66957 from The National Center for Biotechnology Information Gene Expression Omnibus (GEO) database. R software and Bioconductor software package was used to integrate chip data, combined with R package clusterProfiler, to mine gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment pathway. The core genes were screened from the protein–protein interaction (PPI) network of differentially expressed genes (DEGs). Finally, survival analysis was performed using a Kaplan–Meier plotter to further validate core genes. The genes and pathways discovered in the above studies may open a new direction for EOC treatment.

MATERIAL AND METHODS

Data Collection and Data Preprocessing

The raw data for GSE27651, GSE38666, GSE40595, and GSE66957 were integrated for the analysis. The gene chip was obtained from the GEO database (http://www.ncbi.nlm.nih.gov/geo/). GSE27651 included 43 cancer tissues and 6 normal tissues, dataset GSE38666 included 25 cancer tissues and 20 normal tissues, dataset GSE40595 included 63 cancer tissues and 14 normal tissues, and dataset GSE66957 included 57 cancer tissues and 12 normal tissues. They were functioned by Affymetrix Human Genome U133 Plus 2.0 Array [transcript (gene) version] (Affymetrix, Santa Clara, CA, USA) (Harbig et al., 2005). Robust multiarray average approach was performed for background correction and normalization (Irizarry et al., 2003). The original GEO data was then converted into expression measures using affy R package (Gautier et al., 2004).

Differentially Expressed Genes

The "limma" R language package was utilized to detect the DEGs between EOC samples and normal samples in GEO database (Ritchie et al., 2015). We set the adjusted $P < 0.05$ and |log2 fold change (FC)|)1 as the cutoff criteria. Online Wayne diagram was used for identifying the common DEGs among GSE27651, GSE38666, GSE40595, and GSE66957. The drawing of the heatmap was done through the "heatmap" package of R 3.4.4.(Galili et al., 2018)

GO Term and KEGG Pathway Enrichment Analysis

The function and pathway enrichment of the candidate genes were analyzed and annotated using the DAVID database (https://david.ncifcrf.gov/). GO annotations were performed on the screened DEGs using the DAVID online tool and clusterProfiler (Yu et al., 2012). Analysis of KEGG pathway of DEGs was performed using clusterProfiler. We set $p < 0.05$ as a significant criterion.

Comprehensive Analysis of PPI Network and Functional Analysis

STRING (http://www.string-db.org/) was used to assess PPI information (Szklarczyk et al., 2015). In addition, Cytoscape software visualized the results to show the relationship between DEGs. The molecular complex detection (MCODE) plugin was used to search for cluster subnets. We used the following parameters: node score cutoff = 0.2, degree cutoff = 2, max. depth = 100 and k-core = 2. We further used the clusterProfiler to perform functional analysis of the genes in the hub module.

Identification of Potential Drugs

The EOC gene signature was used to query ConnectivityMap (CMap) to find potential drugs for use in patients. The CMap database is a computer simulation method of predicting the potential drugs that may induce or reverse a biological state that encoded by the gene expression signature (Lamb et al., 2006). The different probe components commonly found between EOC tissue samples and normal tissue samples, then used to search the CMap database, are divided into the up- and downregulated groups. An enrichment score representing similarity is finally calculated. The positive connectivity score illustrates that the drug is capable of inducing cancer in human. On the contrary, the negative link score illustrates that the drug is able to reverse the cancer procedure. The negative connectivity score was indicated potential therapeutic value. Tomographs of these candidate molecular drugs were investigated in Pubchem database https://pubchem.ncbi.nlm.nih.gov/).

Validation of Hub Genes

To find key genes that play an important role in EOC, we used Gene Expression Profiling Interactive Analysis (GEPIA) and Kaplan–Meier analysis to analyze the expression and prognosis of 14 hub genes in EOC. GEPIA is based on 9,736 tumors and from cancer genomic map [The Cancer Genome Atlas (TCGA)] and genotype-tissue expression (Tang et al., 2017). We found eight key genes whose expression levels were consistent with the prognosis and was further validated in ONCOMINE database. (www.oncomine.org) (Rhodes et al., 2007) and The Human

Protein Atlas (http://www.proteinatlas.org/) (Lindskog, 2015). Among them, five genes were associated with stage in our study based on GEPIA. Finally, ROC curve analysis was done to distinguish normal and cancer tissues.

Gene Set Enrichment Analysis

In TCGA set validation, EOC samples were divided into two groups according to the median expression level. In order to identify potential function of the hub gene, we conducted a Gene set enrichment analysis (GSEA) (http://software.broadinstitute.org/gsea/index.jsp) (Subramanian et al., 2007) analysis to test whether a series of preferentially defined biological processes were enriched in the gene rank derived from DEGs between the two groups. In addition, we employed "clusterProfiler" package in R to handle the data of gene sets and use "Enrichplot" package to visualize the enriched pathways of the key genes. The adjusted $P < 0.05$ was set as the cutoff criterion.

RESULT

Identification of DEGs in EOC and the Enrichment of These Genes

Four data sets were obtained from the National Center for Biotechnology Information GEO database containing tumor tissue samples and normal ovarian tissue samples: GSE27651, GSE38666, GSE40595, and GSE66957. Then, the R package named "limma" was processed for analysis with adjusted $P < 0.05$ and $|\log2FC| > 1$. All DEGs were displayed in volcano maps (**Figure S1**). Top 200 genes in four databases were displayed in the heatmap (**Figure S2**). A total of 116 genes were finally obtained including 81 upregulated genes and 35 downregulated genes in the EOC tissue samples compared to the normal ovarian tissue samples (**Figure 1**). The data used to create **Figure 1** can be seen in **Table S1**. We also performed clusterProfiler package to do the functional analysis. In GO analysis, the hub upregulated genes were highly enriched in acetylcholine receptor regulator activity, neurotransmitter receptor regulator activity, and vitamin binding (**Figure 2A**); the hub downregulated genes were significantly enriched in peptidase activator activity, collagen binding and transcription factor activity, and RNA polymerase II distal enhancer sequence-specific binding (**Figure 2B**). The data used to create **Figure 2B** can be seen in **Table S2**. In KEGG analysis, the hub upregulated genes were significantly enriched in biosynthesis of amino acids and carbon metabolism (**Figure 2C**); the hub downregulated genes were significantly enriched in retinol metabolism (**Figure 2D**). The above research results can guide us to further study the significance of DEGs in EOC.

GO and Pathway Enrichment Analysis of DEGs

To clarify the major functions of these DEGs, we first explored the associated biological processes and KEGG pathways. The top highly enriched GO terms were divided into three groups: biological process (BP), cellular component (CC), and molecular function (MF) (**Figure 3A**). The most enriched GO terms in biological process was "transcription from RNA polymerase II promoter" ($P < 0.05$), that in cellular component was "extracellular space" and "cell proliferation" ($P < 0.05$), and that in molecular function was "sequence-specific DNA binding" ($P < 0.05$) (**Figures 3B, D**). We further obtained 10 significantly enriched GO terms with a $P < 0.05$. The DEGs included in the top 10 GO terms were shown in the **Figure 3C**. All the GO terms were exhibited in **Table S4**.

In the KEGG analysis, the DEGs were mostly enriched in biosynthesis of amino acids, carbon metabolism, arginine biosynthesis, Wnt signaling pathway, alanine, aspartate, and glutamate metabolism, and glycine, serine, and threonine metabolism (**Figure 3E**).The pathway–protein network is shown in **Figure 3F**.

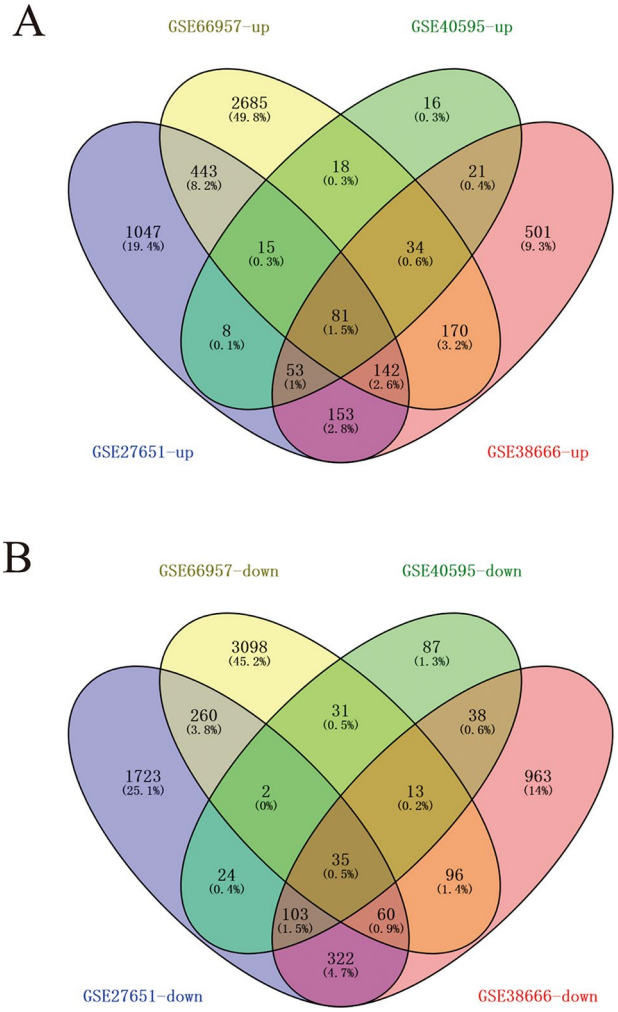


FIGURE 1 | A total of 116 commonly changed differentially expressed genes (DEGs) which were divided into up- and down-regulated groups, were identified from four transcription profile datasets (GSE27651, GSE38666, GSE40595, and GSE66957) using online website. Calculate and draw custom Venn diagrams (available online: http://bioinformatics.psb.ugent.be/webtools/Venn/). Statistically significant DEGs were defined with adjusted $P < 0.05$ and $|\log2FC| > 1$ as the cutoff criterion. **(A)** Up-regulated genes **(B)** Down-regulated genes.

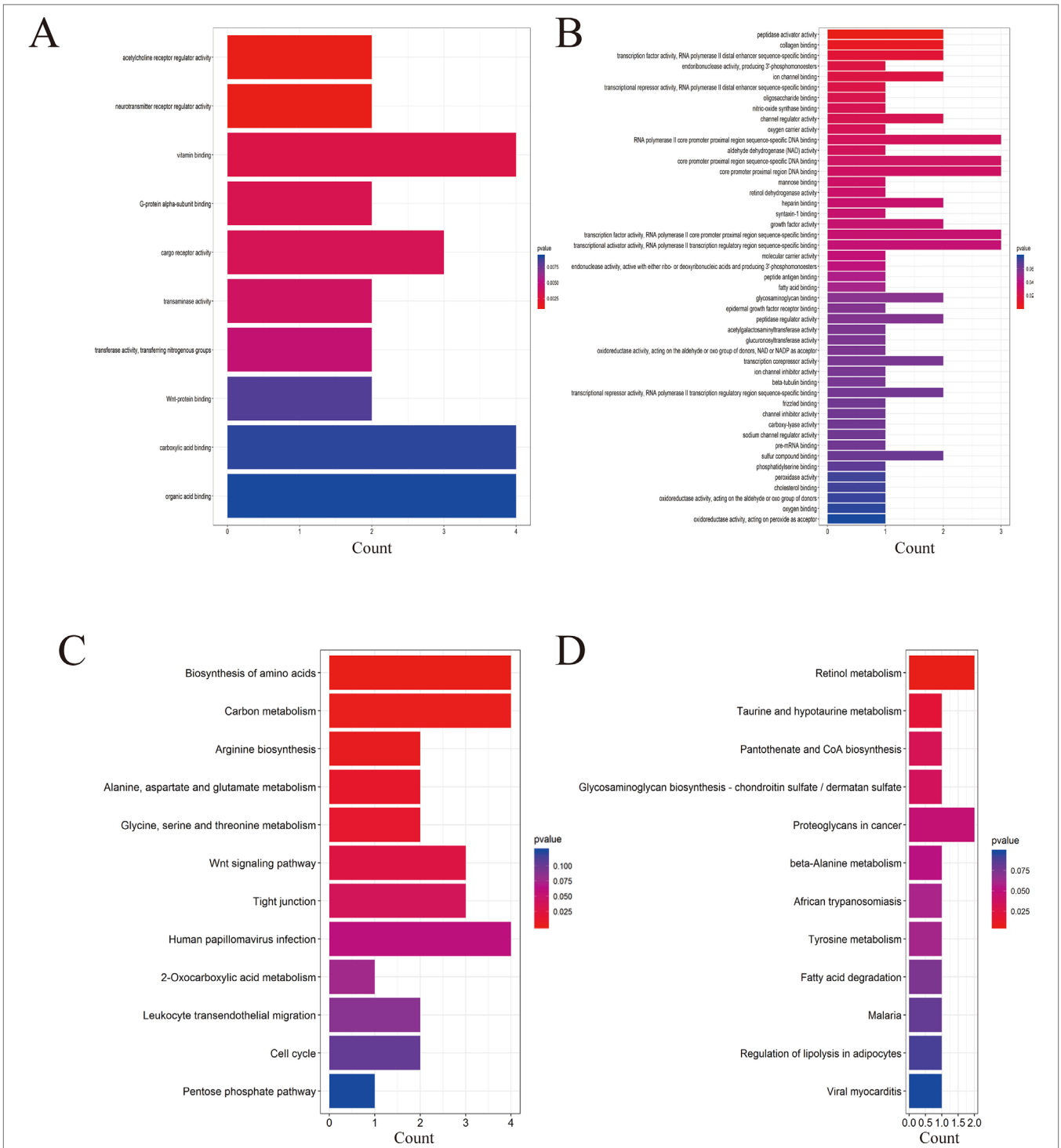


FIGURE 2 | Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis of the up- and downregulated DEGs. **(A)** GO analysis of upregulated genes associated with epithelial ovarian cancer (EOC). **(B)** GO analysis of downregulated genes associated with EOC. **(C)** KEGG analysis of upregulated genes associated with EOC. **(D)** KEGG analysis of downregulated genes associated with EOC.

Related Small Molecule Drugs Screening

We divided the 116 differentially expressed genes consisting of 35 downregulated genes and 81 upregulated genes into two groups, up- and downregulated, which were substituted into the CMap network tool. Among these highly significant correlated molecules, vorinostat, LY-294002, trichostatin A, and tanespimycin showed higher negative correlation with EOC. They all might have the potential therapeutic effects on EOC

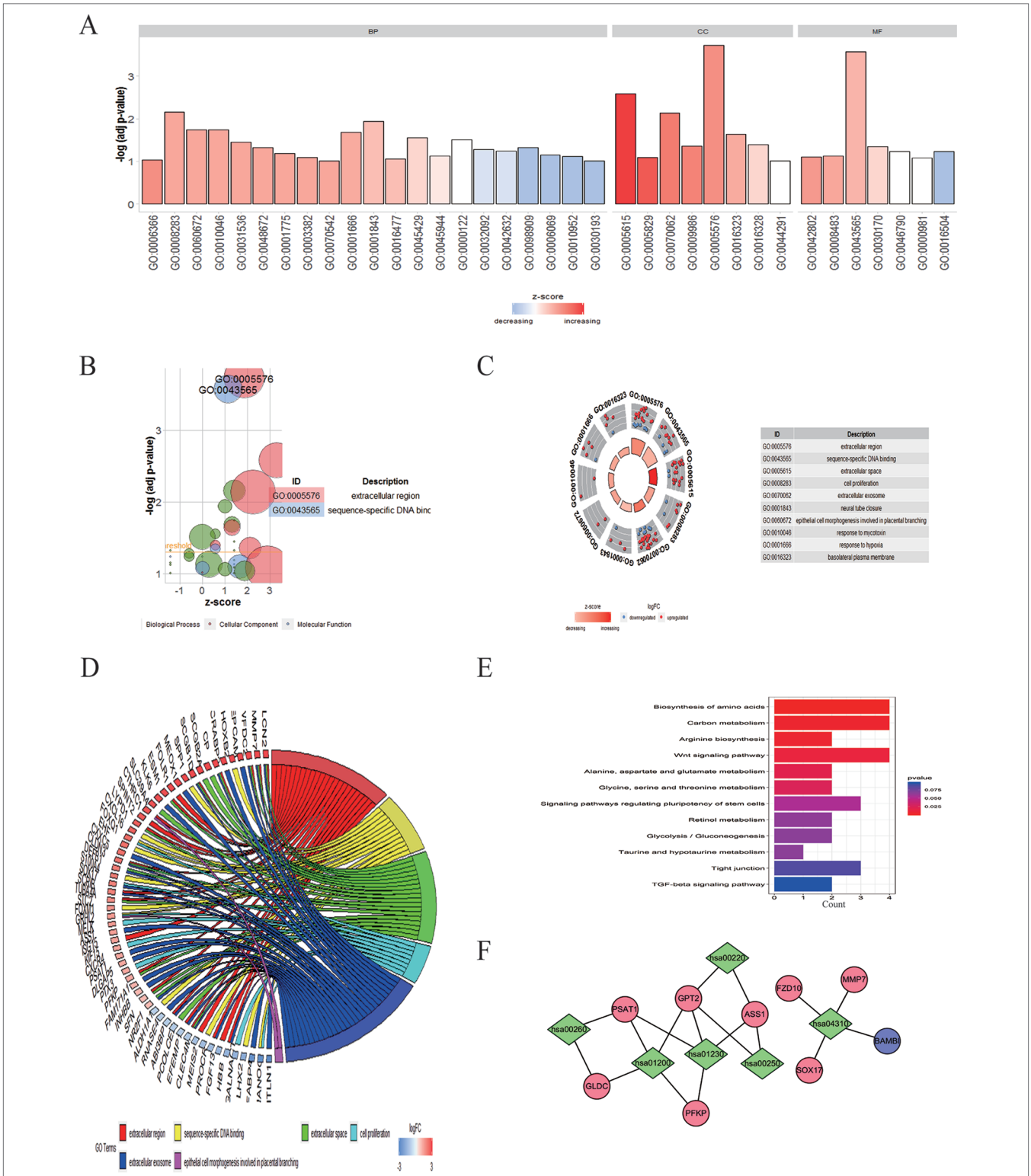


FIGURE 3 | GO enrichment and KEGG analysis of DEGs in EOC. **(A)** GO analysis divided DEGs into three functional groups: molecular function, biological processes, and cell composition. **(B)** The bubble plot of enriched GO terms. The *z*-score is assigned to the *x*-axis, and the negative logarithm of the *P* value to the *y*-axis, as in the barplot (the higher the more significant). The size of the displayed circles is proportional to the number of genes assigned to the term. Green circles correspond to the biological process, red indicates the cellular component, and blue shows the molecular function category. **(C)** The top 10 GO terms of DEGs in EOC. The outer circle shows a scatter plot for each term of the logFC of the assigned genes. Red circles display upregulation and blue ones downregulation. **(D)** Distribution of DEGs in cervical cancer for different GO-enriched functions. **(E)** KEGG analysis of DEGs. **(F)** The pathway–protein network of DEGs.

(**Table 1**). Three-dimensional structure of the top 4 candidate molecule drugs was found in Pubchem database and shown in **Figures 4A–D**.

PPI Network and Cluster Analysis

STRING website screened 114 DEGs into PPI complex, which demonstrated 114 nodes and 157 edges (**Figure 5A**), and 30 important proteins were identified (**Figure 5B**). After that, we applied the MCODE, and three clusters were obtained. Among them, cluster 1 contained 11 core proteins and got the highest score in these clusters (**Figure 6A**), cluster 2 contained 5 proteins (**Figure 6B**), and cluster 3 contained 3 proteins (**Figure 6C**). These results may indicate that the 19 DEGs influence EOC.

We further performed the functional analysis of cluster 1. In GO analysis, the DEGs of cluster 1 were mostly enriched in microtubule motor activity, motor activity, and microtubule binding (**Figure 7A**). In KEGG analysis, the DEGs of cluster 1 were mostly enriched in DNA replication and cell cycle (**Figure 7B**). All pathways of significant molecule in cluster 1 are shown in **Table S3**.

Hub Gene Validation

To further demonstrate the effect of these genes in EOC, we performed survival analysis on all the 114 DEGs first (**Table S5**), and 14 genes were obtained as candidate genes according to degree ≥10. Then, CDCA5, FOXM1, KIF11, KIF15, MCM2, SPP1, UBE2C, and ZWINT, which showed higher expression in CESC samples compared with normal samples (**Figure 8**), were negatively relative to overall survival of EOC patients (**Figure 9**). Patients with higher expression levels were worse than patients with lower expression levels.

TABLE 1 | Results of ConnectivityMap (CMap) analysis.

Rank	CMap name	Mean	*N*	Enrichment	*P* value
1	vorinostat	−0.477	12	−0.639	0
2	LY-294002	−0.25	61	−0.393	0
3	trichostatin A	−0.359	182	−0.347	0
4	tanespimycin	−0.297	62	−0.307	0
5	folic acid	0.66	4	0.889	0.00014
6	gentamicin	0.613	4	0.886	0.00016
7	harmol	0.584	4	0.877	0.00034
8	amantadine	0.518	4	0.837	0.00109
9	trazodone	−0.63	4	−0.911	0.00124
10	hycanthone	0.449	4	0.823	0.00167

Similarly, overexpression of CDCA5, FOXM1, KIF11, KIF15, MCM2, SPP1, UBE2C, and ZWINT in tumors was significantly associated with p*rogression-free survival* in EOC patients (**Figure 10**). Expression analysis of cervical cancer versus normal performed on ONCOMINE also showed that expression of these eight genes were screened higher in EOC

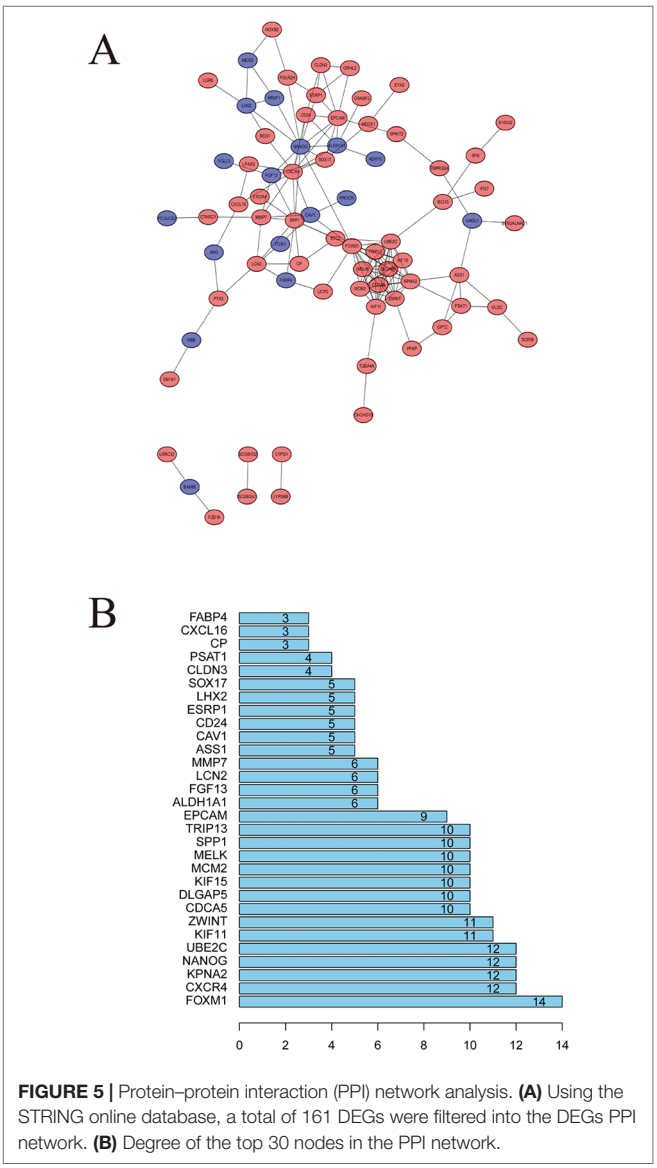


FIGURE 5 | Protein–protein interaction (PPI) network analysis. **(A)** Using the STRING online database, a total of 161 DEGs were filtered into the DEGs PPI network. **(B)** Degree of the top 30 nodes in the PPI network.

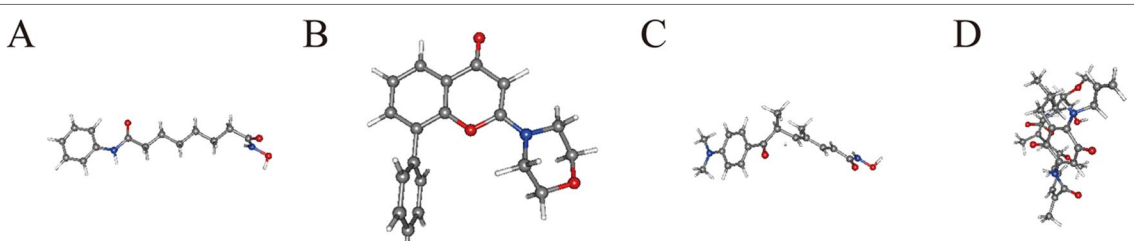


FIGURE 4 | Three-dimensional diagram of the three most significant drugs. **(A)** Vorinostat, **(B)** LY-294002, **(C)** trichostatin A, **(D)** tanespimycin.

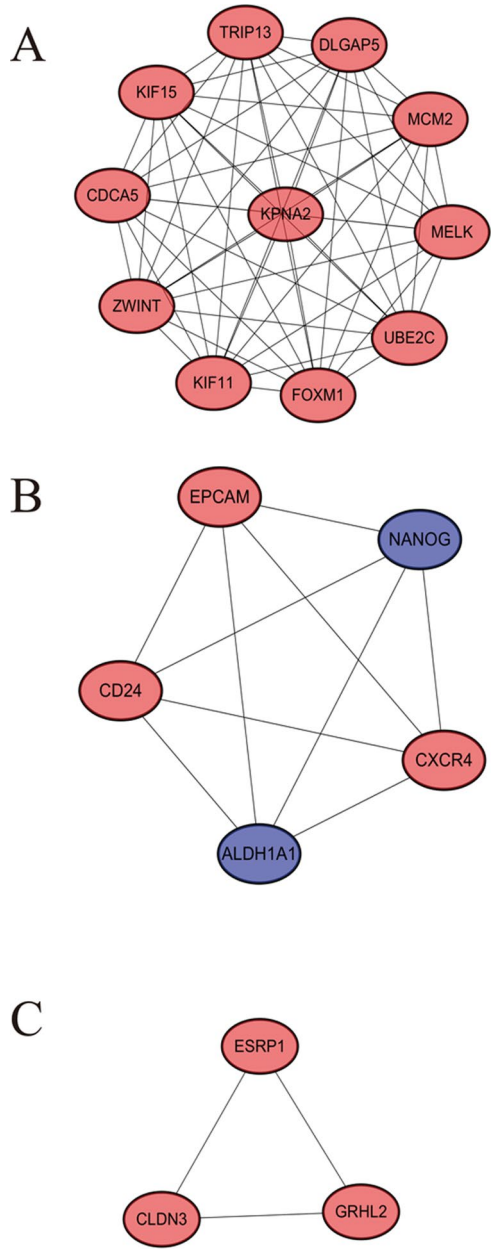


FIGURE 6 | A significant module selected from protein–protein interaction network. **(A)** Cluster 1, **(B)** cluster 2, **(C)** cluster 3.

sample (**Figure 11**). Interestingly, we also found that five genes CDCA5, FOXM1, KIF15, MCM2, and ZWINT were relative to EOC stage by GEPIA analysis (**Figure 12**). In addition, we performed survival analysis based on stage I–II and stage III–IV. The results showed that the high expression of five hub genes was significantly worse than that of low expression in the stage I/II, but there was no statistical significance in stage III/IV (**Figure S3**). Immunohistochemistry also suggested that, compared with normal tissues, the protein expression level of these five genes were obviously higher in tumor tissues (**Figure 13**). In addition, ROC curve analysis was implemented to evaluate the capacity of hub genes to distinguish EOC and normal tissues in GES66957, CDCA5, FOXM1, KIF15 and MCM2, exhibiting better diagnostic efficiency for normal and tumor tissues, and the combined diagnosis of these five genes was more effective. The value of AUC was 0.858 (**Figure 14A**). However, efficiency of the ROC analysis between stage I–II and stage III–IV was weak (**Figure S4**). In addition, the univariate and multivariate Cox proportional hazards regression showed that the ZWINT was an independent prognostic indicator for overall survival among EOC patients (**Table 2**).

Gene Set Enrichment Analysis

To identify the potential function of these five genes in TCGA OV databases, GSEA was conducted to search KEGG pathways enriched in the highly expressed samples. As a result, 10 gene sets "citrate cycle tricarboxylic acid (TCA) cycle," "homologous recombination," "steroid biosynthesis," "pentose phosphate pathway," "glyoxylate and dicarboxylate metabolism," "RNA polymerase," "hypertrophic cardiomyopathy," "dilated cardiomyopathy," and "drug metabolism cytochrome P450" were enriched (**Figure 14B**) (adjusted $P < 0.05$).

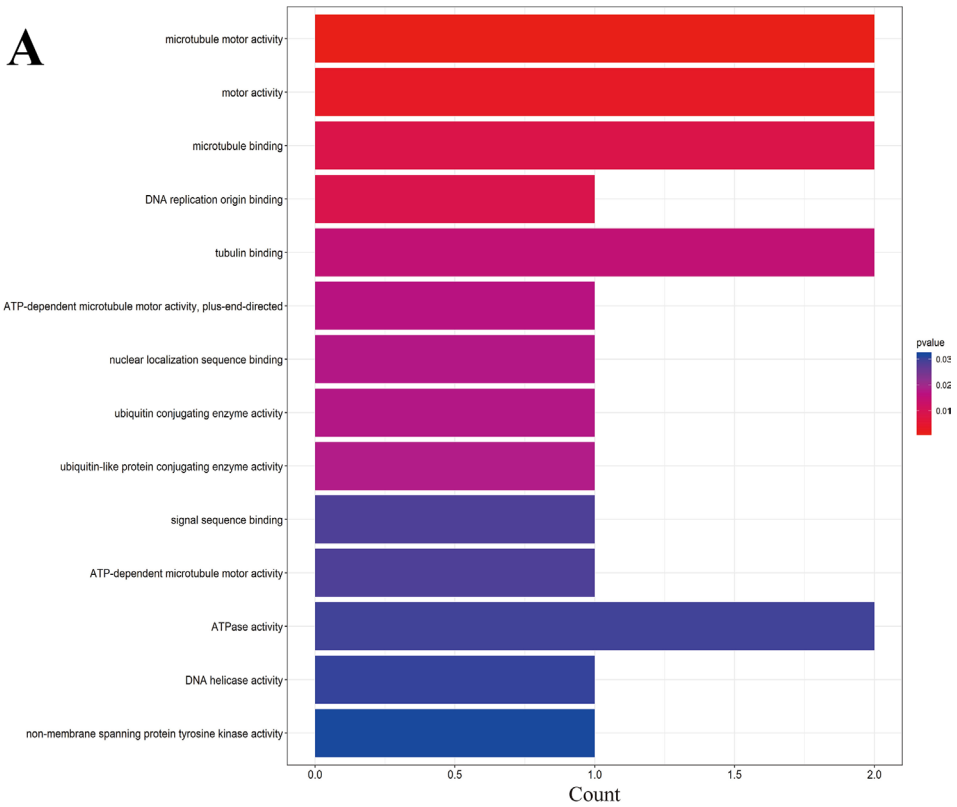
DISCUSSION

Although surgery and other treatment methods have been improved, the treatment effect and prognosis of advanced ovarian cancer patients are very poor due to the difficulty in the diagnosis of ovarian cancer. Although many microlevel studies have been carried out, they are not yet mature. In this study, we integrated four gene chips from GEO databases and selected 116 DEGs between tumor and nontumor samples (81 expression levels were upregulated and 35 expression levels were downregulated), and further functional analysis was performed.

GO analysis displayed that the upregulated DEGs were mainly enriched in acetylcholine receptor regulator activity, and the downregulated genes were highly enriched in peptidase activator activity. The KEGG analysis showed that the upregulated DEGs were highly involved in biosynthesis of amino acids, while the hub downregulated genes were highly enriched in retinol metabolism.

Acetylcholine receptor regulator activity is often mentioned in lung cancer (Wang and Hu, 2018). Peptidase activator activity has been shown to be involved in the regulation of prostate cancer (Fuhrman-Luck et al., 2016). Biosynthesis of amino acids also has a relationship with the treatment of tumors (Manig et al., 2017). Retinol metabolism has been shown to be associated with breast cancer and gallbladder cancer (Chen et al., 1997). These key pathways are related to the occurrence and development of human tumors but have not been studied in detail in EOC, so functional analysis has certain guiding significance.

Several small molecules with potential therapeutic efficacy against EOC were identified. Among them, the most relevant vorinostat, LY-294002, trichostatin A, and tanespimycin had been shown to have different degrees of association with tumors.



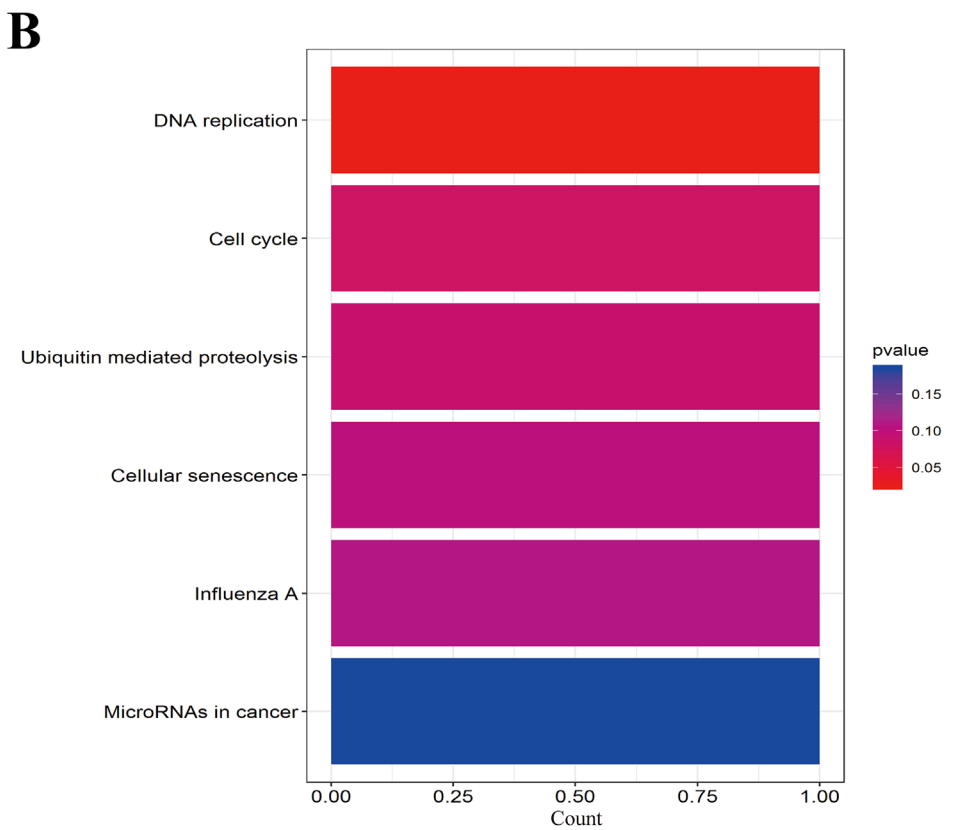


FIGURE 7 | (A) GO analysis of these significant molecule in cluster 1. **(B)** KEGG analysis of these significant molecule in cluster 1.

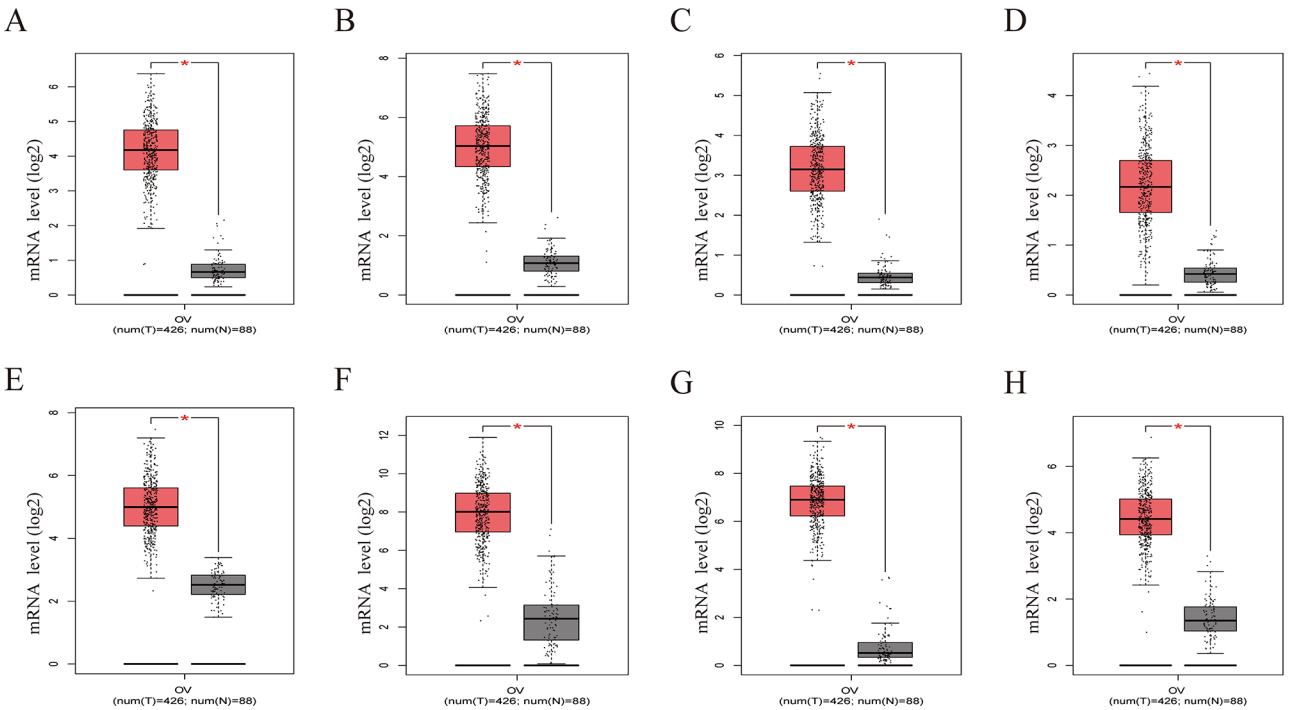


FIGURE 8 | Gene Expression Profiling Interactive Analysis (GEPIA) was performed to validate higher expression of eight hub genes in EOC samples compared with normal samples. Red box was the cancer tissue group, gray was the normal tissue group, and asterisk represented $p < 0.01$. The dots represented expression in each sample. **(A)** CDCA5, **(B)** FOXM1, **(C)** KIF11, **(D)** KIF15, **(E)** MCM2, **(F)** SPP1, **(G)** UBE2C, **(H)** ZWINT.

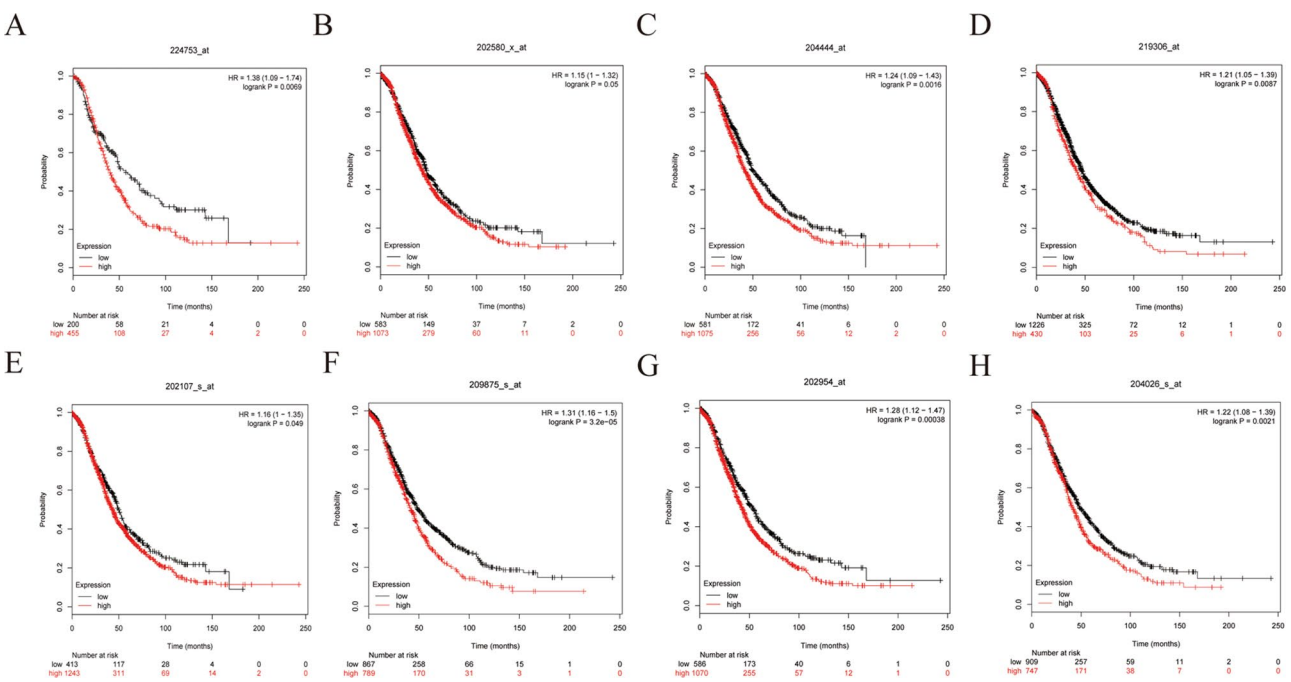


FIGURE 9 | Overall survival (OS) curves of eight hub genes. **(A)** CDCA5, **(B)** FOXM1, **(C)** KIF11, **(D)** KIF15, **(E)** MCM2, **(F)** SPP1, **(G)** UBE2C, **(H)** ZWINT.

Vorinostat is a small molecule inhibitor of both class I and II histone deacetylase enzymes (Munster et al., 2001) that functions by altering acetylation, affecting apoptotic pathways (Finnin et al., 1999). Vorinostat, which has been approved by the US Food and Drug Administration, is used to treat a variety of malignancies including ovarian cancer (Ma et al., 2017). LY-294002 has no

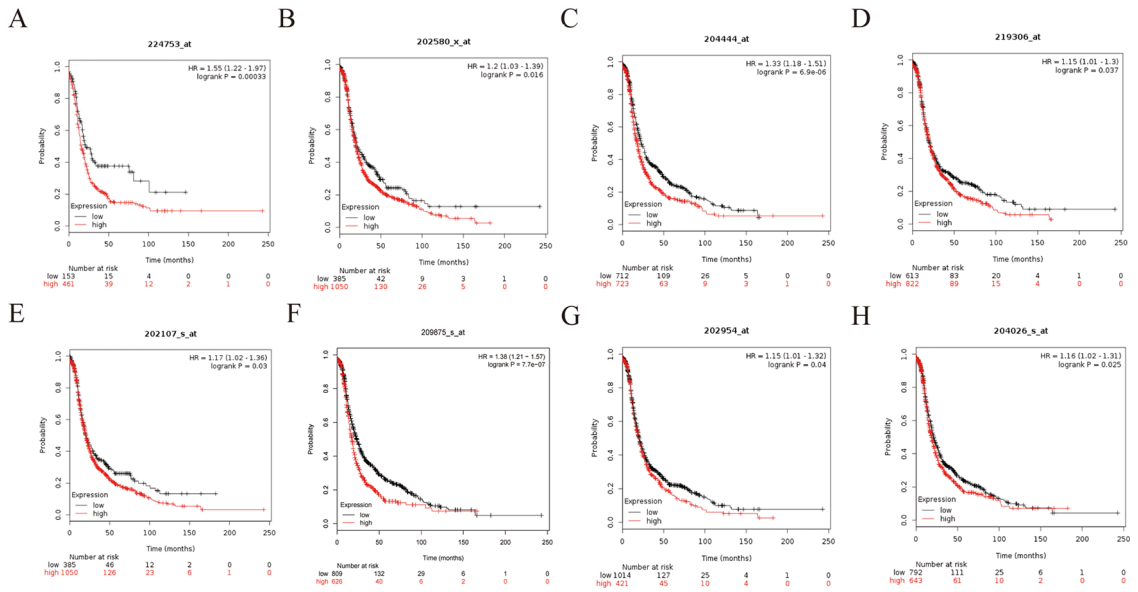


FIGURE 10 | *Progression-free survival* (*PFS*) survival curves of eight hub genes. **(A)** CDCA5, **(B)** FOXM1, **(C)** KIF11, **(D)** KIF15, **(E)** MCM2, **(F)** SPP1, **(G)** UBE2C, **(H)** ZWINT.

clinical applications at present. Trichostatin A, as a histone deacetylase inhibitor, has been shown to exhibit anticancer effects in combination with radiotherapy or chemotherapy (Ranganathan and Rangnekar, 2005; Hajji et al., 2008). In the early 1990s, there were experiments that demonstrated that tanespimycin had antitumor activity against various human-derived tumor cell lines (Erlichman, 2009). The above three drugs have been identified to have antitumor effect in the past. The PPI network analyzed DEGs and displayed 114 nodes. The MCODE plug-in filtered out three related modules. The correlation of module 1 was the most significant. We performed survival analysis on 11 genes which belong to module 1 and found that patients with these DEG disorders had a poor prognosis. Among these genes, CDCA5, FOXM1, KIF15, MCM2, and ZWINT were the most reported genes associated with cancer progression, including EOC. ROC curve analysis demonstrated these genes had better diagnostic efficiency for normal and tumor tissues, and the combination of diagnosis was more effective. Meanwhile, the univariate and multivariate Cox proportional hazards regression showed that ZWINT was an independent prognostic indictor among EOC patients. Besides, GSEA suggested that the five genes were mostly enriched in citrate cycle TCA cycle, homologous recombination, and steroid biosynthesis.

Interestingly, the study by Ren JG et al. found that citrate suppressed tumor growth through inhibition of glycolysis, the TCA cycle and the insulin-like growth factor-1 receptor pathway (Ren et al., 2017). Homologous recombination deficiency was closely related to ovarian cancer and breast cancer (Zhao et al., 2017; Da et al., 2018). These examples show that the results of GESA analysis can be used as a reference for oncogenesis studies to some extent.

Cell-division cycle-associated 5 (CDCA5), also known as sororin, is thought to play a critical role in ensuring the accurate separation of sister chromatids during the S and G2/M phases of the cell cycle through interactions with cohesin and cdk1 (Schmitz et al., 2007; Borton et al., 2016). CDCA5 has also been shown to interact with ERK as well as cyclin E1, a critical regulator of the G1/Smitotic checkpoint (Schmitz et al., 2007; Nguyen et al., 2010; Borton et al., 2016). Recent studies have correlated the expression of CDCA5 with tumorigenesis and tissue invasion in several cancers, including oral squamous cell cancer, nonsmall cell lung cancer, urothelial cell carcinoma, and gastric cancer (Chang et al., 2015; Tokuzen et al., 2016). However, the gene has not been reported in ovarian cancer and deserves further study.

FOXM1 is a member of the forkhead box (Fox) transcription factor family, which is known as an oncogene involved in breast cancer, cervix cancer, prostate cancer, and so on. In agreement with previously published studies (Lok et al., 2011; Wen et al., 2014; Zhao et al., 2014; Zhou et al., 2014; Chiu et al., 2015), our experimental findings demonstrated that FOXM1 was overexpressed in EOC and negatively associated with prognosis of EOC patients.

KIF15 is the breast cancer tumor antigen and is necessary for the maintenance of spindle bipolarity (Scanlan et al., 2001). KIF15 supports K5I resistance in HeLa cells (Sturgill et al., 2016), which is shown to act as target for endocrine therapy-resistant breast cancer (Zou et al., 2014). The same result existed in lung adenocarcinoma and may play a vital role in regulating the cell cycle (Bidkhori et al., 2013). Our study reported for the first time that KIF15 expressed higher in EOC and led to the bad outcome of EOC patients.

Minichromosome maintenance (MCM) 2 is one of six related proteins that comprise the MCM complex (MCM2-7), which has an essential role in DNA replication (Bochman and Schwacha, 2008). Previous studies using human samples have established MCM2 as

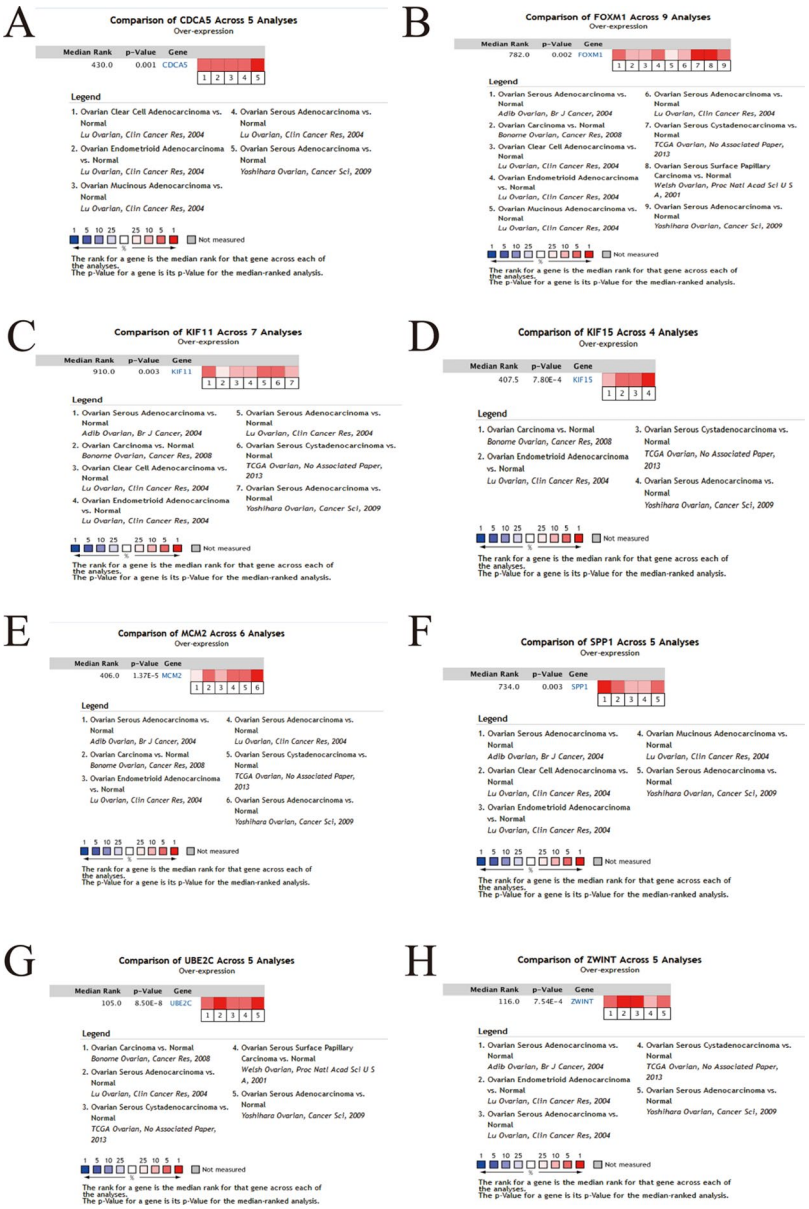


FIGURE 11 | Eight hub genes expression within EOC across multiple datasets by Oncomine. **(A)** CDCA5, **(B)** FOXM1, **(C)** KIF11, **(D)** KIF15, **(E)** MCM2, **(F)** SPP1, **(G)** UBE2C, **(H)** ZWINT.

a proliferation marker of cancer cells. High expression of MCM2 level in malignant tumors, including ovarian cancer, is associated with several clinicopathological parameters such as advanced tumor grade, advanced stage, and poor prognosis (Davies et al., 2002; Going et al., 2002; Dudderidge et al., 2005; Majid et al., 2010; Abe et al., 2015). In our study, we also found that MCM, which had higher expression in EOC, was relative to bad outcome of EOC patient.

ZWINT belongs to the kinetochore complex and is a protein that interacts with ZW10 and participates in chromosome movement (Woo et al., 2015). Endo et al. found that ZWINT promoted cell growth, and targeting KWINT inhibited breast cancer cell growth (Endo et al., 2012). There is also a bioinformation research that reported this gene in OC, which is similar to ours. In summary, CDCA5, FOXM1, KIF15, MCM2, and ZWINT was involved in cell mitosis and supported our research results by affecting the cell cycle regulation of tumor pathogenesis.

There are several limitations in our study as follows. First, there is an urgent need for biological experiments to validate our results because our research is based on data analysis. Second, we lack the molecular mechanisms for these genes, and we will incorporate these for further exploration. In the future, we will further design experiments (including PCR, Western blot, immunohistochemistry, etc.) based on specific mechanisms, conduct in-depth research, and improve the inadequacies.

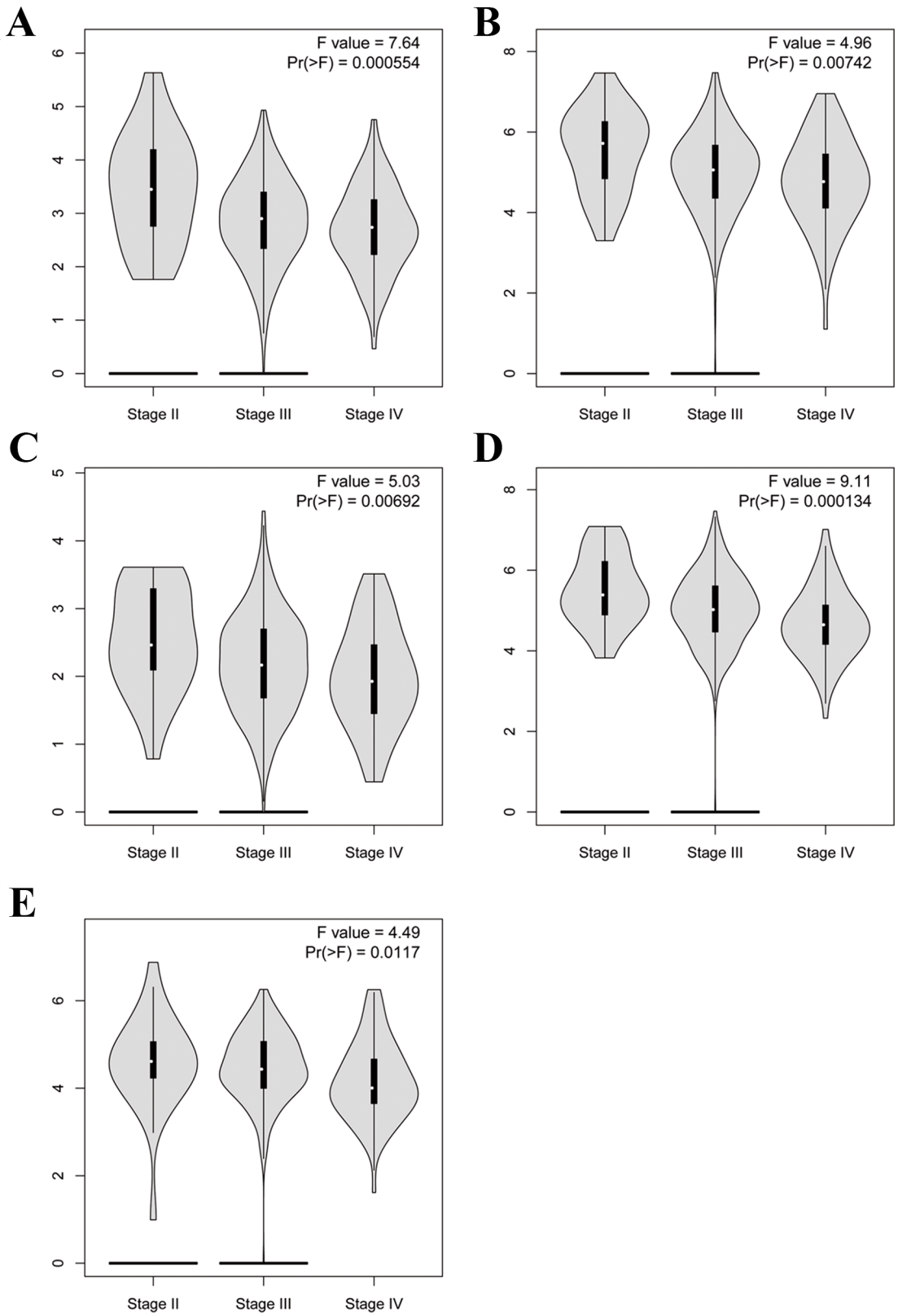


FIGURE 12 | Violin plot showing five key genes expression in different major pathological stages of EOC. The *y*-axis represents log2 (TPM + 1). **(A)** CDCA5, **(B)** FOXM1, **(C)** KIF15, **(D)** MCM2, **(E)** ZWINT.

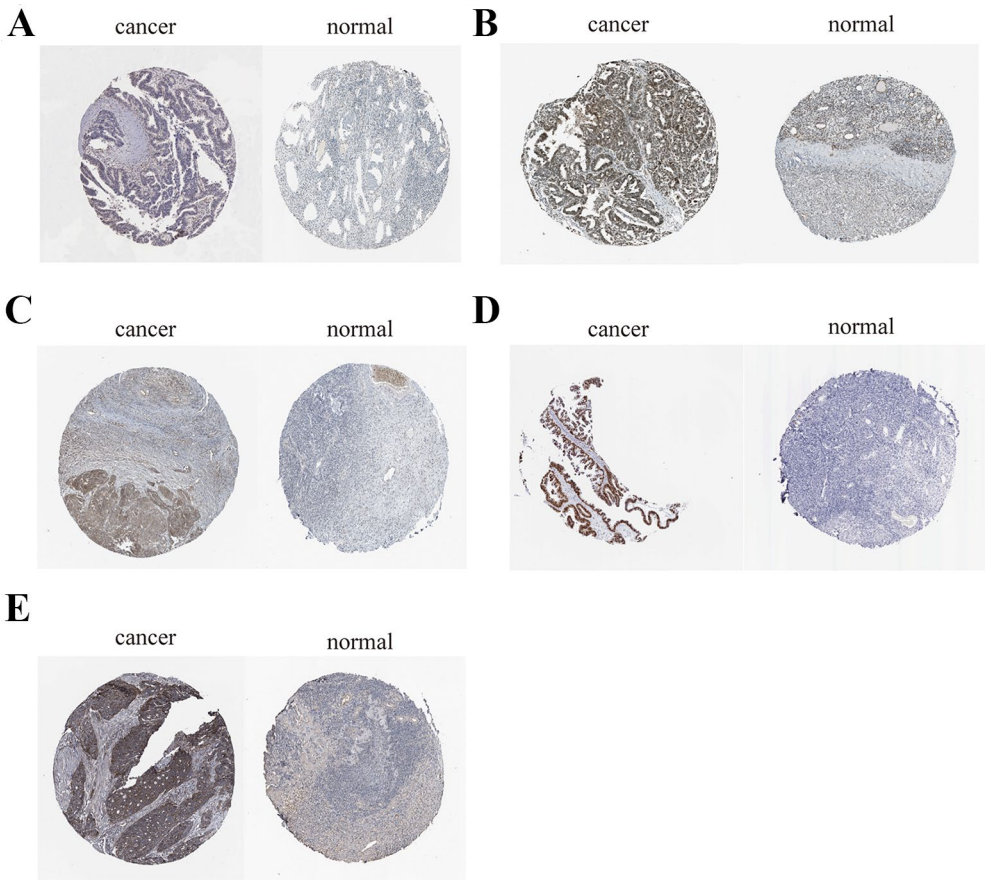


FIGURE 13 | Immunohistochemistry of the five key genes based on the Human Protein Atlas. **(A)** Protein levels of CDCA5 in tumor tissue (staining: low; intensity: weak; quantity: > 75%). Protein levels of CDCA5 in normal tissue (staining: not detected; intensity: weak; quantity: < 25%). **(B)** Protein levels of FOXM1 in tumor tissue (staining: high; intensity: strong; quantity: > 75%). Protein levels of FOXM1 in normal tissue (staining: medium; intensity: moderate; quantity: 75–25%);. **(C)** Protein levels of KIF15 in tumor tissue (staining: medium; intensity: moderate; quantity: > 75%). Protein levels of KIF15 in normal tissue (staining: not detected; intensity: weak; quantity: < 25%). **(D)** Protein levels of MCM2 in tumor tissue (staining: high; intensity: strong; quantity: > 75%). Protein levels of MCM2 in normal tissue (staining: medium; intensity: moderate; quantity: 75–25%). **(E)** Protein levels of ZWINT in tumor tissue (staining: high; intensity: strong; quantity: > 75%). Protein levels of ZWINT in normal tissue (staining: low; intensity: weak; quantity: > 75%).

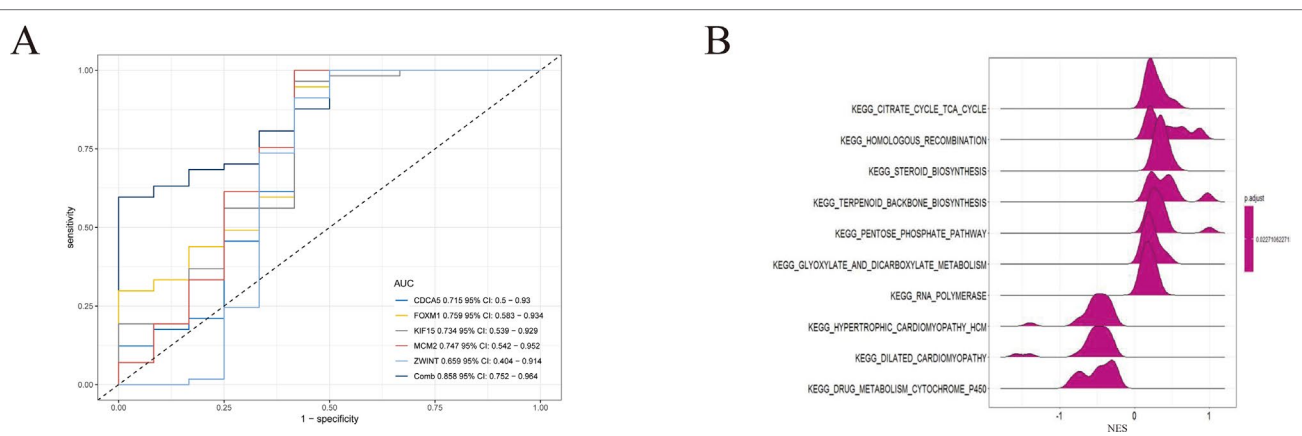


FIGURE 14 | (A) Receiver operating characteristic (ROC) curve analysis and area under the curve (AUC) statistics were implemented on different databases to evaluate the capacity of key genes to distinguish EOC and normal tissues. **(B)** GSEA was applied to obtain biological process enriched in five key genes with highly expressed samples.

TABLE 2 | Univariate and multivariate analyses of the correlation of CDCA5, FOXM1, KIF15, MCM2, and ZWINT expression with overall survival (OS) among epithelial ovarian cancer patients.

Variables	Univariate analysis			Multivariate analysis		
	HR	95%CI	*P*	HR	95%CI	*P*
Age (≤60 vs. >60)	1.358	1.038–1.775	**0.025**	1.275	0.944–1.722	0.113
Stage (stages I and II vs. stages III and IV)	1.095	0.647–1.852	0.735	0.938	0.542–1.622	0.819
Grade (G1 and G2 vs. G3 and G4)	1.373	0.904–2.084	0.137	1.185	0.752–1.868	0.464
CDCA5	0.991	0.965–1.018	0.517	1.003	0.960–1.048	0.888
FOXM1	1.002	0.992–1.013	0.648	1.006	0.990–1.021	0.440
KIF15	0.991	0.910–1.079	0.842	1.002	0.885–1.134	0.977
MCM2	0.998	0.988–1.008	0.675	1.002	0.990–1.014	0.692
ZWINT	0.982	0.965–0.999	**0.038**	0.977	0.956–0.998	**0.035**

Bold values indicate P < 0.05. HR, hazard ratio; CI, confidence interval.

CONCLUSION

In our study, we adopted 4 GEO chips to demonstrate 116 DEGs. Then, we comprehensively analyzed GEPIA, ONCOMINE, and other databases. We identified that CDCA5, FOXM1, KIF15, MCM2, and ZWINT were related to EOC. At the same time, our study also analyzed the potential new drugs for the treatment of ovarian cancer based on the DEGs. In a word, our research proved that bioinformatics analysis might open up new directions for cancer research. More therapeutic targets will be tapped if further clinical trials are combined.

DATA AVAILABILITY STATEMENT

The datasets in this study are available from GEO database (https://www.ncbi.nlm.nih.gov/geo/) using accessions numbers GSE27651, GSE38666, GSE40595 and GSE66957 and TCGA database (https://cancergenome.nih.gov/).

AUTHOR CONTRIBUTIONS

WC and JL designed the project. JL, HM, and SL contributed on data analysis and prepared the main manuscript. All authors reviewed the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (81872119 and 81472442) and the Jiangsu province medical innovation team (CXTDA2017008).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01031/full#supplementary-material

SUPPLEMENTARY FIGURE 1 | Volcano map of DEGs on the basis of |fold change| > 1 and a corrected P-value <0.05. (A) GSE27651 (B) GSE38666 © GSE40595 (D) GSE66957. Red dots represented the upregulated DEGs. Green dots represented the down-regulated DEGs. Black dots represented noDEGs.

SUPPLEMENTARY FIGURE 2 | Hierarchical clustering heatmap of top 200 DEGs. (A) GSE27651 (B) GSE38666 © GSE40595 (D) GSE66957. Red indicates that the expression of genes is relatively upregulated, green indicates that the expression of genes is relatively downregulated, and black indicates no significant changes in gene expression.

SUPPLEMENTARY FIGURE 3 | The survival analysis based on stage I-II and stage III-IV. The one on the left in the figure was survival analysis based on stage I-II, the right one was stage III-IV.

SUPPLEMENTARY FIGURE 4 | ROC analysis between stage I-II and stage III-IV.

REFERENCES

Abe, S., Yamamoto, K., Kurata, M., Abe-Suzuki, S., Horii, R., Akiyama, F., et al. (2015). Targeting MCM2 function as a novel strategy for the treatment of highly malignant breast tumors. *Oncotarget* 6, 34892–34909. doi: 10.18632/oncotarget.5408

Allemani, C., Weir, H. K., Carreira, H., Harewood, R., Spika, D., Wang, X. S., et al. (2015). Global surveillance of cancer survival 1995-2009: analysis of individual data for 25,676,887 patients from 279 population-based registries in 67 countries (CONCORD-2). *Lancet* 385, 977–1010. doi: 10.1016/S0140-6736(14)62038-9

Bidkhori, G., Narimani, Z., Hosseini, A. S., Moeini, A., Nowzari-Dalini, A., and Masoudi-Nejad, A. (2013). Reconstruction of an integrated genome-scale co-expression network reveals key modules involved in lung adenocarcinoma. *PLoS One* 8, e67552. doi: 10.1371/journal.pone.0067552

Bochman, M. L., and Schwacha, A. (2008). The Mcm2-7 complex has *in vitro* helicase activity. *Mol. Cell* 31, 287–293. doi: 10.1016/j.molcel.2008.05.020

Borton, M. T., Rashid, M. S., Dreier, M. R., and Taylor, W. R. (2016). Multiple Levels of Regulation of Sororin by Cdk1 and Aurora B. *J. Cell. Biochem.* 117, 351–360. doi: 10.1002/jcb.25277

Chang, I. W., Lin, V. C., He, H. L., Hsu, C. T., Li, C. C., Wu, W. J., et al. (2015). CDCA5 overexpression is an indicator of poor prognosis in patients with urothelial carcinomas of the upper urinary tract and urinary bladder. *Am. J. Transl. Res.* 7, 710–722. doi: 10.1007/s13277-015-3210-z

Chen, A. C., Guo, X., Derguini, F., and Gudas, L. J. (1997). Human breast cancer cells and normal mammary epithelial cells: retinol metabolism and growth inhibition by the retinol metabolite 4-oxoretinol. *Cancer Res.* 57, 4642–4651.

Chiu, W. T., Huang, Y. F., Tsai, H. Y., Chen, C. C., Chang, C. H., Huang, S. C., et al. (2015). FOXM1 confers to epithelial-mesenchymal transition, stemness and chemoresistance in epithelial ovarian carcinoma cells. *Oncotarget* 6, 2349–2365. doi: 10.18632/oncotarget.2957

Da, C. C. B. R., Fogace, R. N., Miranda, V. C., and Diz, M. (2018). Homologous recombination deficiency in ovarian cancer: a review of its epidemiology

and management. *Clinics (Sao Paulo)* 73, e450s. doi: 10.6061/clinics/2018/e450s

Davies, R. J., Freeman, A., Morris, L. S., Bingham, S., Dilworth, S., Scott, I., et al. (2002). Analysis of minichromosome maintenance proteins as a novel method for detection of colorectal cancer in stool. *Lancet* 359, 1917–1919. doi: 10.1016/S0140-6736(02)08739-1

Dudderidge, T. J., Stoeber, K., Loddo, M., Atkinson, G., Fanshawe, T., Griffiths, D. F., et al. (2005). Mcm2, Geminin, and KI67 define proliferative state and are prognostic markers in renal cell carcinoma. *Clin. Cancer Res.* 11, 2510–2517. doi: 10.1158/1078-0432.CCR-04-1776

Endo, H., Ikeda, K., Urano, T., Horie-Inoue, K., and Inoue, S. (2012). Terf/TRIM17 stimulates degradation of kinetochore protein ZWINT and regulates cell proliferation. *J. Biochem.* 151, 139–144. doi: 10.1093/jb/mvr128

Erlichman, C. (2009). Tanespimycin: the opportunities and challenges of targeting heat shock protein 90. *Expert Opin. Investig. Drugs* 18, 861–868. doi: 10.1517/13543780902953699

Finnin, M. S., Donigian, J. R., Cohen, A., Richon, V. M., Rifkind, R. A., Marks, P. A., et al. (1999). Structures of a histone deacetylase homologue bound to the TSA and SAHA inhibitors. *Nature* 401, 188–193. doi: 10.1038/43710

Fuhrman-Luck, R. A., Stansfield, S. H., Stephens, C. R., Loessner, D., and Clements, J. A. (2016). Prostate cancer-associated kallikrein-related peptidase 4 activates matrix metalloproteinase-1 and thrombospondin-1. *J. Proteome Res.* 15, 2466–2478. doi: 10.1021/acs.jproteome.5b01148

Galili, T., O'Callaghan, A., Sidi, J., and Sievert, C. (2018). heatmaply: an R package for creating interactive cluster heatmaps for online publishing. *Bioinformatics* 34, 1600–1602. doi: 10.1093/bioinformatics/btx657

Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 20, 307–315. doi: 10.1093/bioinformatics/btg405

Going, J. J., Keith, W. N., Neilson, L., Stoeber, K., Stuart, R. C., and Williams, G. H. (2002). Aberrant expression of minichromosome maintenance proteins 2 and 5, and Ki-67 in dysplastic squamous oesophageal epithelium and Barrett's mucosa. *Gut* 50, 373–377. doi: 10.1136/gut.50.3.373

Hajji, N., Wallenborg, K., Vlachos, P., Nyman, U., Hermanson, O., and Joseph, B. (2008). Combinatorial action of the HDAC inhibitor trichostatin A and etoposide induces caspase-mediated AIF-dependent apoptotic cell death in non-small cell lung carcinoma cells. *Oncogene* 27, 3134–3144. doi: 10.1038/sj.onc.1210976

Harbig, J., Sprinkle, R., and Enkemann, S. A. (2005). A sequence-based identification of the genes detected by probesets on the Affymetrix U133 plus 2.0 array. *Nucleic Acids Res.* 33, e31. doi: 10.1093/nar/gni027

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., et al. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4, 249–264. doi: 10.1093/biostatistics/4.2.249

La Vecchia, C. (2017). Ovarian cancer: epidemiology and risk factors. *Eur. J. Cancer Prev.* 26, 55–62. doi: 10.1097/CEJ.0000000000000217

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., et al. (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 313, 1929–1935. doi: 10.1126/science.1132939

Lindskog, C. (2015). The potential clinical impact of the tissue-based map of the human proteome. *Expert Rev. Proteomics* 12, 213–215. doi: 10.1586/14789450.2015.1040771

Lok, G. T., Chan, D. W., Liu, V. W., Hui, W. W., Leung, T. H., Yao, K. M., et al. (2011). Aberrant activation of ERK/FOXM1 signaling cascade triggers the cell migration/invasion in ovarian cancer cells. *PLoS One* 6, e23790. doi: 10.1371/journal.pone.0023790

Ma, X., Wang, J., Liu, J., Mo, Q., Yan, X., Ma, D., et al. (2017). Targeting CD146 in combination with vorinostat for the treatment of ovarian cancer cells. *Oncol. Lett.* 13, 1681–1687. doi: 10.3892/ol.2017.5630

Majid, S., Dar, A. A., Saini, S., Chen, Y., Shahryari, V., Liu, J., et al. (2010). Regulation of minichromosome maintenance gene family by microRNA-1296 and genistein in prostate cancer. *Cancer Res.* 70, 2809–2818. doi: 10.1158/0008-5472.CAN-09-4176

Manig, F., Kuhne, K., von Neubeck, C., Schwarzenbolz, U., Yu, Z., Kessler, B. M., et al. (2017). The why and how of amino acid analytics in cancer diagnostics and therapy. *J. Biotechnol.* 242, 30–54. doi: 10.1016/j.jbiotec.2016.12.001

McAlpine, J. N., Hanley, G. E., Woo, M. M., Tone, A. A., Rozenberg, N., Swenerton, K. D., et al. (2014). Opportunistic salpingectomy: uptake, risks, and complications of a regional initiative for ovarian cancer prevention. *Am. J. Obstet. Gynecol.* 210, 471.e1–471.11. doi: 10.1016/j.ajog.2014.01.003

Munster, P. N., Troso-Sandoval, T., Rosen, N., Rifkind, R., Marks, P. A., and Richon, V. M. (2001). The histone deacetylase inhibitor suberoylanilide hydroxamic acid induces differentiation of human breast cancer cells. *Cancer Res.* 61, 8492–8497.

Nannini, M., Pantaleo, M. A., Maleddu, A., Astolfi, A., Formica, S., and Biasco, G. (2009). Gene expression profiling in colorectal cancer using microarray technologies: results and perspectives. *Cancer Treat. Rev.* 35, 201–209. doi: 10.1016/j.ctrv.2008.10.006

Nguyen, M. H., Koinuma, J., Ueda, K., Ito, T., Tsuchiya, E., Nakamura, Y., et al. (2010). Phosphorylation and activation of cell division cycle associated 5 by mitogen-activated protein kinase play a crucial role in human lung carcinogenesis. *Cancer Res.* 70, 5337–5347. doi: 10.1158/0008-5472.CAN-09-4372

Petryszak, R., Burdett, T., Fiorelli, B., Fonseca, N. A., Gonzalez-Porta, M., Hastings, E., et al. (2014). Expression Atlas update–a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Res.* 42, D926–D932. doi: 10.1093/nar/gkt1270

Ranganathan, P., and Rangnekar, V. M. (2005). Exploiting the TSA connections to overcome apoptosis-resistance. *Cancer Biol. Ther.* 4, 391–392. doi: 10.4161/cbt.4.4.1779

Ren, J. G., Seth, P., Ye, H., Guo, K., Hanai, J. I., Husain, Z., et al. (2017). Citrate suppresses tumor growth in multiple models through inhibition of glycolysis, the tricarboxylic acid cycle and the IGF-1R pathway. *Sci. Rep.* 7, 4537. doi: 10.1038/s41598-017-04626-4

Rhodes, D. R., Kalyana-Sundaram, S., Mahavisno, V., Varambally, R., Yu, J., Briggs, B. B., et al. (2007). Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia* 9, 166–180. doi: 10.1593/neo.07112

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43, e47. doi: 10.1093/nar/gkv007

Scanlan, M. J., Gout, I., Gordon, C. M., Williamson, B., Stockert, E., Gure, A. O., et al. (2001). Humoral immunity to human breast cancer: antigen definition and quantitative analysis of mRNA expression. *Cancer Immun.* 1, 4.

Schmitz, J., Watrin, E., Lenart, P., Mechtler, K., and Peters, J. M. (2007). Sororin is required for stable binding of cohesin to chromatin and for sister chromatid cohesion in interphase. *Curr. Biol.* 17, 630–636. doi: 10.1016/j.cub.2007.02.029

Sturgill, E. G., Norris, S. R., Guo, Y., and Ohi, R. (2016). Kinesin-5 inhibitor resistance is driven by kinesin-12. *J. Cell Biol.* 213, 213–227. doi: 10.1083/jcb.201507036

Subramanian, A., Kuehn, H., Gould, J., Tamayo, P., and Mesirov, J. P. (2007). GSEA-P: a desktop application for Gene Set Enrichment Analysis. *Bioinformatics* 23, 3251–3253. doi: 10.1093/bioinformatics/btm369

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 43, D447–D452. doi: 10.1093/nar/gku1003

Tang, Z., Li, C., Kang, B., Gao, G., Li, C., and Zhang, Z. (2017). GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res.* 45, W98–W102. doi: 10.1093/nar/gkx247

Tokuzen, N., Nakashiro, K., Tanaka, H., Iwamoto, K., and Hamakawa, H. (2016). Therapeutic potential of targeting cell division cycle associated 5 for oral squamous cell carcinoma. *Oncotarget* 7, 2343–2353. doi: 10.18632/oncotarget.6148

Wang, S., and Hu, Y. (2018). alpha7 nicotinic acetylcholine receptors in lung cancer. *Oncol. Lett.* 16, 1375–1382. doi: 10.3892/ol.2018.8841

Wen, N., Wang, Y., Wen, L., Zhao, S. H., Ai, Z. H., Wang, Y., et al. (2014). Overexpression of FOXM1 predicts poor prognosis and promotes cancer cell proliferation, migration and invasion in epithelial ovarian cancer. *J. Transl. Med.* 12, 134. doi: 10.1186/1479-5876-12-134

Woo, S. D., Yeop, Y. S., Chung, W. J., Cho, D. H., Kim, J. S., and Su, O. J. (2015). Zwint-1 is required for spindle assembly checkpoint function and kinetochore-microtubule attachment during oocyte meiosis. *Sci. Rep.* 5, 15431. doi: 10.1038/srep15431

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 16, 284–287. doi: 10.1089/omi.2011.0118

Zhao, E. Y., Shen, Y., Pleasance, E., Kasaian, K., Leelakumari, S., Jones, M., et al. (2017). Homologous Recombination Deficiency and Platinum-Based Therapy Outcomes in Advanced Breast Cancer. *Clin. Cancer Res.* 23, 7521–7530. doi: 10.1158/1078-0432.CCR-17-1941

Zhao, F., Siu, M. K., Jiang, L., Tam, K. F., Ngan, H. Y., Le, X. F., et al. (2014). Overexpression of forkhead box protein M1 (FOXM1) in ovarian cancer correlates with poor patient survival and contributes to paclitaxel resistance. *PLoS One* 9, e113478. doi: 10.1371/journal.pone.0113478

Zhou, J., Wang, Y., Wang, Y., Yin, X., He, Y., Chen, L., et al. (2014). FOXM1 modulates cisplatin sensitivity by regulating EXO1 in ovarian cancer. *PLoS One* 9, e96989. doi: 10.1371/journal.pone.0096989

Zou, J. X., Duan, Z., Wang, J., Sokolov, A., Xu, J., Chen, C. Z., et al. (2014). Kinesin family deregulation coordinated by bromodomain protein ANCCA and histone methyltransferase MLL for breast cancer cell growth, survival, and tamoxifen resistance. *Mol. Cancer Res.* 12, 539–549. doi: 10.1158/1541-7786.MCR-13-0459

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.

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

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