APPLICATION OF NOVEL STATISTICAL AND MACHINE-LEARNING METHODS TO HIGH-DIMENSIONAL CLINICAL CANCER AND (MULTI-)OMICS DATA

EDITED BY: Chao Xu, Md Ashad Alam and Shaolong Cao

PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88971-436-0
DOI 10.3389/978-2-88971-436-0

About Frontiers

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

Frontiers Journal Series

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

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

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

APPLICATION OF NOVEL STATISTICAL AND MACHINE-LEARNING METHODS TO HIGH-DIMENSIONAL CLINICAL CANCER AND (MULTI-)OMICS DATA

Topic Editors:
Chao Xu, University of Oklahoma Health Sciences Center, United States
Md Ashad Alam, Tulane University, United States
Shaolong Cao, University of Texas MD Anderson Cancer Center, United States

Citation: Xu, C., Alam, M. A., Cao, S., eds. (2022). Application of Novel Statistical and Machine-learning Methods to High-dimensional Clinical Cancer and (Multi-)Omics Data. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88971-436-0

Table of Contents

frontiers in Genetics

EDITORIAL
published: 12 August 2021
doi: 10.3389/fgene.2021.739442

Editorial: Application of Novel Statistical and Machine-Learning Methods to High-Dimensional Clinical Cancer and (Multi-)Omics Data

Chao Xu [1]*, *Shaolong Cao* [2] *and Md. Ashad Alam* [3]

[1] Department of Biostatistics and Epidemiology, University of Oklahoma Health Sciences Center, Oklahoma City, OK, United States, [2] University of Texas M. D. Anderson Cancer Center, Houston, TX, United States, [3] Tulane University School of Medicine, New Orleans, LA, United States

Keywords: multi-omics analysis, high-dimensional data analysis, cancer genomics, prediction analysis, integrative analysis

Editorial on the Research Topic

Application of Novel Statistical and Machine-Learning Methods to High-Dimensional Clinical Cancer and (Multi-)Omics Data

The big genomics data from various aspects (e.g., DNA polymorphism, transcriptomics, and proteomics) is now available in cancer research and clinic application. These (multi-)omics data come with a new feature of high dimension: much more features/predictors relative to the available sample size. Meanwhile, researchers are looking beyond individual omics study and exploring integrative analysis of (multi-)omics data. Accordingly, there are novel statistical and machine learning methods designed for the high-dimensional and/or integrative (multi-)omics data analysis. The present Research Topic collects the methodology development and application of statistical and machine-learning methods for high-dimensional clinical (multi-)omics, and integration analysis, mostly, in cancer research.

For multi-omics integration analysis, classical statistical and machine leaning approaches are widely used. Wang et al. did Cox regression analysis on combined immunohistochemical (IHC) markers and synthetic lethal gene pairs. New prognostic markers for Asian oral cancer were reported. Xu et al. used unsupervised cluster-of-clusters analysis to integrate subgroup classification from different omics and identify potential driver genes in cervical cancer. They found four statistically significant expression subtypes by clustering of tumor copy number variation (CNV) and methylation profiles.

New approaches have been developed based on these classical methods as well. For example, the Mimi-Surv Model built on Cox regression was designed to identify miRNA-mRNA integration set associated with survival time (Kim et al.). Ye et al. proposed a new meta-analysis method to integrate multiple transcriptomic studies and categorize biomarkers by concordant patterns with application to Pan-Cancer studies. Jeong et al. presented a kernel canonical correlation analysis (CCA) method to construct condition specific transcriptional networks. CCA with a positive definite kernel is a well-used method for multiple source data analysis. They employed kernel CCA to embed transcription factors (TFs) and target genes (TGs) into a new space where the correlation of TFs and TGs are reflected. Their approach successfully detected novel TF-TG relations in addition to replicated existing regulatory interactions.

OPEN ACCESS

Edited and reviewed by:
Simon Charles Heath,
Center for Genomic Regulation (CRG), Spain

****Correspondence:***
Chao Xu
Chao-Xu@ouhsc.edu

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *10 July 2021*
Accepted: *23 July 2021*
Published: *12 August 2021*

Citation:
Xu C, Cao S and Alam MA (2021) Editorial: Application of Novel Statistical and Machine-Learning Methods to High-Dimensional Clinical Cancer and (Multi-)Omics Data. Front. Genet. 12:739442. doi: 10.3389/fgene.2021.739442

Current methods appropriate for high-dimensional data includes penalized regression models (e.g., LASSO and Ridge regression), kernel-based methods, tree-based methods (e.g., random forest), and latest versatile deep learning models [e.g., Generative Adversarial Networks (GANs)]. In our collection, Ge et al. proposed a modified conditional GANs with new network structures for estimation of individualized treatment effect, which can handle binary and continuous type of treatments. In their framework, LASSO was also used to select biomarkers for optimal treatment selection. Liu and Li proposed a new method for estimation and prediction of heterogeneous restricted mean survival time based on random forest. The application in ovarian cancer showed improved prediction performance vs. existing methods.

With many powerful tools in the field, it is always interesting to evaluate their strengths and appropriate usage. Källberg et al. compared 13 feature selection methods for their ability to identify a subset of genes that can be used to accurately classify cancer subtypes based on gene expression data. Each of the feature selection techniques was applied to four human cancer data sets with known subtypes, enabling accuracy assessment. Their findings demonstrated that the feature selection methods based on modality outperformed the most commonly used approach of selecting the genes with the highest variability.

In addition to the applications in cancer, Zhou et al. used gene expression data and 6 machine learning methods to predict the 3-year survival risk for patients having heart failure with preserved ejection fraction (HFpEF). In their result, the kernel partial least squares with the genetic algorithm (GA-KPLS) outperformed penalized regression, random forest, support vector machine (SVM), and logistic regression. Jiang et al. employed mendelian randomization (MR) and meta-analysis to study the causal relationship between alcohol consumption and risk of autoimmune inflammatory diseases from totaling 1 million individuals' genetic data. With the enormous genetic data and comprehensive analysis, they noted an overall null association between alcohol consumption and common autoimmune inflammatory disorders.

As summarized above, this collection of original research papers presents a significant amount of progress made in the integrative analysis of clinical and (multi-)omics cancer data, prediction in cancer diagnosis/survival/progression, statistical, and machine learning methods for high-dimensional data analysis. While the generation of data has far outpaced our ability to make sense of those data, further development and application of statistical and machine-learning methods are required for the analysis of contemporary genetics in cancer and other diseases.

AUTHOR CONTRIBUTIONS

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

FUNDING

This work was partially supported by the National Institute of Environmental Health Sciences (R21 ES032098-01). We would like to thank the Frontiers editorial staff for providing outstanding assistance in putting together this Research Topic collection.

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

METHODS
published: 11 December 2020
doi: 10.3389/fgene.2020.585804

Conditional Generative Adversarial Networks for Individualized Treatment Effect Estimation and Treatment Selection

Qiyang Ge[1,2], *Xuelin Huang*[3], *Shenying Fang*[4], *Shicheng Guo*[5], *Yuanyuan Liu*[1], *Wei Lin*[2] *and Momiao Xiong*[1]*

[1] Department of Biostatistics and Data Science, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, United States, [2] School of Mathematical Sciences, Fudan University, Shanghai, China, [3] Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, United States, [4] Department of Surgical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, United States, [5] Department of Medical Genetics, University of Wisconsin-Madison, Madison, WI, United States

OPEN ACCESS

Edited by:
Chao Xu,
University of Oklahoma Health Sciences Center, United States

Reviewed by:
Shuangxi Ji,
University of Texas MD Anderson Cancer Center, United States
Chuan Qiu,
Tulane University, United States

***Correspondence:**
Momiao Xiong
momiao.xiong@uth.tmc.edu

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *21 July 2020*
Accepted: *18 November 2020*
Published: *11 December 2020*

Citation:
Ge Q, Huang X, Fang S, Guo S, Liu Y, Lin W and Xiong M (2020) Conditional Generative Adversarial Networks for Individualized Treatment Effect Estimation and Treatment Selection. Front. Genet. 11:585804. doi: 10.3389/fgene.2020.585804

Treatment response is heterogeneous. However, the classical methods treat the treatment response as homogeneous and estimate the average treatment effects. The traditional methods are difficult to apply to precision oncology. Artificial intelligence (AI) is a powerful tool for precision oncology. It can accurately estimate the individualized treatment effects and learn optimal treatment choices. Therefore, the AI approach can substantially improve progress and treatment outcomes of patients. One AI approach, conditional generative adversarial nets for inference of individualized treatment effects (GANITE) has been developed. However, GANITE can only deal with binary treatment and does not provide a tool for optimal treatment selection. To overcome these limitations, we modify conditional generative adversarial networks (MCGANs) to allow estimation of individualized effects of any types of treatments including binary, categorical and continuous treatments. We propose to use sparse techniques for selection of biomarkers that predict the best treatment for each patient. Simulations show that MCGANs outperform seven other state-of-the-art methods: linear regression (LR), Bayesian linear ridge regression (BLR), k-Nearest Neighbor (KNN), random forest classification [RF (C)], random forest regression [RF (R)], logistic regression (LogR), and support vector machine (SVM). To illustrate their applications, the proposed MCGANs were applied to 256 patients with newly diagnosed acute myeloid leukemia (AML) who were treated with high dose ara-C (HDAC), Idarubicin (IDA) and both of these two treatments (HDAC+IDA) at M. D. Anderson Cancer Center. Our results showed that MCGAN can more accurately and robustly estimate the individualized treatment effects than other state-of-the art methods. Several biomarkers such as GSK3, BILIRUBIN, SMAC are identified and a total of 30 biomarkers can explain 36.8% of treatment effect variation.

Keywords: causal inference, generative adversarial networks, counterfactuals, treatment estimation, precision medicine

INTRODUCTION

Traditional clinical management estimates the average treatment effects from observational data, assuming that the complex disease is homogeneous (Rosenbaum and Rubin, 1983; Hansen, 2004; Diamond and Sekhon, 2013; Kennedy et al., 2017; Liu et al., 2018; Luo and Zhu, 2020). Alternatives to traditional clinical management, "precision medicine" or "precision oncology" attempts to match the most accurate and effective treatments with the individual patient (Shin et al., 2017; Ali and Aittokallio, 2019), rather than using monotherapy that treats all patients. In the real world, treatment response is heterogeneous. Therapy should be tailored with the best response possible and highest safety margin to ensure that the right therapy is offered to "the right patient at the right time" (Subbiah and Kurzrock, 2018). Precision oncology can substantially improve progress and treatment outcomes of patients. It plays a central role in revolutionizing cancer research. Consequently, alternative to calculating the average effect of an intervention over a population, many recent methods attempt to estimate individualized treatment effects (ITEs) or conditional average treatment effects from observational data (Makar et al., 2019). To accurately estimate the individualized treatment effects and learn optimal treatment choices are key issues for precision oncology. More accurate estimation of individualized treatment effects, which provides information to guide the individual selection of the target therapies, is essential for the success of precision medicine (Kornblau et al., 2009).

Methods for estimation of individualized treatment effects (ITEs) using observational data largely differ from standard statistical estimation methods. Estimating of ITEs and learning optimal treatment strategies raise a great challenge for the following reasons. First, a common framework for treatment effect estimation is the potential outcomes assumptions (Ray and Szabo, 2019) where every individual has two "potential outcomes" covering the hypothesized individual's outcomes with and without treatment. Estimation of ITEs requires estimation of both factual and counterfactual outcomes for each individual. However, only the factual outcome is actually observed. We never observe the counterfactual outcomes (Rosenbaum and Rubin, 1983; Chen and Paschalidis, 2018; Yoon et al., 2018a).

If the effect of each treatment in the subpopulation which is separately estimated is taken as an individual effect, this can create large biases. The estimated effect of each treatment in the subpopulation is still the average effect of the treatment in that subpopulation and is not an individualized treatment effect in the subpopulation.

Second, clinical data often have many missing values. Simultaneously imputing both counterfactual values and missing values is not easy. Third, the function forms of the treatment effects which are often non-linear functions are unknown (Ray and Szabo, 2019). Statistical methods and computational algorithms that can efficiently deal with unknown forms of non-linear functions are still lacking (Lengerich et al., 2019).

Classical works such as random forest and hierarchical models are adapted to estimate heterogeneous treatment effects (Wager and Athey, 2015). Recently, machine learning and neural network methods are used to move away from average treatment effect estimation to personalized estimation (Johansson et al., 2016; Shalit et al., 2016; Alaa and van der Schaar, 2017). AI and causal inferences are becoming a driving force for innovation in precision oncology (Seyhan and Carini, 2019). A key issue for ITE estimation is to learn unobserved (missing) counterfactuals. The idea of using generative adversarial networks (GANs) for handling missing data is a very promising approach to imputing counterfactual (Goodfellow et al., 2014; Ding and Li, 2017; Yoon et al., 2018a). Using conditional GAN (CGAN) to estimate the individualized treatment effects (GANITE) has been developed (Yoon et al., 2018a,b). The CGANs consist of a generator and a discriminator. The generator (G) observes the factual part of real data and imputes the counterfactuals (missing part) conditioned on observed factual data, and outputs the complete dataset. The discriminator (D) inputs the real dataset and tries to determine which part was actually observed and which part was imputed counterfactuals. The discriminator enforces the generator to learn the desired distribution (hidden data distribution) (Yoon et al., 2018b).

However, the original GANITE was designed for estimation of the effects of binary treatment and cannot be applied to continuous and categorical treatments. The treatment variable in the original GANITE is a binary variable which only represents the presence and absence of treatment. Therefore, the treatment variable in the original GANITE is unable to quantify the dosage of the treatment, and hence the original GANITE cannot be applied to continuous treatment. To overcome this limitation, we introduce a treatment assignment indicator variable and treatment quantity variable. The treatment quantity variable can represent binary treatment, categorical treatment, and continuous treatment. We change mathematical formulations of the generator and discriminator and extend GANITE from binary treatment to all types of treatments including binary, categorical, and continuous treatments. The modified GANITE is abbreviated as MGANITE.

GANITE or in general, CGAN has not systematically investigated the estimation of ITE for chemotherapy and other types of treatments in cancer and compared the results from causal inference using observed data with the results of randomized clinical trials. One of our goals in this manuscript is to examine whether MGANITE still works well in cancer research.

In MGANITE, biomarkers that serve as conditioned variables, will be used to estimate the ITEs of both single and multiple treatments (Mirza and Osindero, 2014; Yoon et al., 2018a). Sparse techniques will be employed to select biomarkers for prediction of treatment effects and to learn optimal treatment choices of patients (Emmert-Streib and Dehmer, 2019).

In summary, The novelty of modified GANITE (MGANITE) is summarized below.

1. The previous conditional generative adversarial network (CGAN)-based causal inference methods (GANITE) only can estimate the individualized effects of binary treatment and cannot estimate the individualized effects of continuous treatments. The proposed MGANITE is the first time to use

modified CGANs for estimation of individualized effects of continuous treatments.

2. We develop new network structures for the generator and discriminator in the CGANs.
3. We combined sparse techniques for selection of biomarkers with MGANITE to predict the best treatment for each patient.

To evaluate its performance for estimating ITEs, simulations are conducted to estimate ITEs using simulated data and MGANITE, and to compare its estimation accuracy with five other state-of-the-art methods (LR, KNN, BLR, random forest, and SVM). To further evaluate its performance, MGANITE is applied to 256 newly diagnosed acute myeloid leukemia (AML) patients, treated with high dose ara-C (HDAC), Idarubicin (IDA), and HDAC+IDA at M. D. Anderson Cancer Center to estimate ITEs and identify the optimal treatment strategy for each patient. Preliminary results from simulations and real data analysis show that MGANITE outperforms five other state-of-the-art methods. A program for implementing the proposed MGANITE for ITE estimation and optimal treatment selection can be downloaded from our website https://sph.uth.edu/research/centers/hgc/software/xiong/.

MATERIALS AND METHODS

Potential Outcome Framework for Estimation of Treatment Effects

We assume the Rubin causal model for estimation of treatment effects (Rubin, 1974) and modifies the approach to the individualized treatment effect estimation in Yoon et al., 2018a). The original GANITE only can estimate ITE of binary treatments, but it cannot be applied to categorical and continuous treatments. We develop MGANITE which can estimate ITE of all types of treatments including binary, categorical, and continuous treatments by introducing a treatment assignment indicator variable and changing the formulation of the generator and discriminator. Consider K treatments. Let T_k be the k^{th} treatment variable that can be binary, categorical or continuous, and $T = [T_1, \ldots, T_K]^T$ be the treatment vector. We assume that there is precisely one non-zero component of the treatment vector T, which is denoted by T_η, where η is the index of this component. Each sample has one and only one assigned treatment T_η. To extend the binary treatment to include categorical and continuous treatments, we define the treatment assignment indicator vector $M = [M_1, \ldots, M_k, \ldots, M_K]^T$ as

$$M_k = \begin{cases} 1 & k = \eta \\ 0 & \text{otherwise} \end{cases}$$

where $\sum_{k=1}^{K} M_k = 1$.
For example, if

$$T = \begin{bmatrix} 0 \\ T_2 \\ 0 \end{bmatrix}$$

then $\eta = 2$ and

$$M = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

If we consider treated and untreated cases, then $K = 2$. Let T_1 denote the treatment and T_2 denote no treatment where $T_2 = 1$. For the sample with the treatment, we have

$$T = \begin{bmatrix} T_1 \\ 0 \end{bmatrix} \text{ and } M = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

For the sample with no treatment, we have

$$T = \begin{bmatrix} 0 \\ T_2 \end{bmatrix} \text{ and } M = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Define the vector of potential outcome $Y(T) = [Y(T_1), \ldots, Y(T_K)]^T$, where $Y(T_k)$ is the potential outcome of the sample under the treatment T_k. When $K = 2$, the potential outcome $Y(T_1)$ corresponds to the widely used notation for one treatment Y^1, the potential outcome of the treated sample, while the potential outcome $Y(T_2)$ corresponds to Y^0, the potential outcome of the untreated sample. Only one of the potential outcomes can be observed. The observed outcome that corresponds to the potential outcome of the individual receiving the treatment T_η is denoted by $Y(T_\eta)$. The observed outcome is called the factual outcome and the unobserved potential outcomes are called counterfactual outcomes, or simply counterfactuals. For the convenience of notation, the factual outcome is also denoted by Y_f and the counterfactuals are denoted by Y_{cf}.

The observed outcome Y_f can be expressed as

$$Y_f = Y_\eta = \sum_{k=1}^{K} M_k Y(T_k)$$

When $K = 2$, we have $M_2 = 1 - M_1$. The above equation becomes

$$Y_f = M_1 Y(T_1) + (1 - M_1) Y(T_2) = M_1 Y^1 + (1 - M_1) Y^0$$

which coincides with the standard expression of the observed outcome for one treatment.

Let $X = [X_1, \ldots, X_q]^T$ be the q-dimensional feature vector. Assume that n individuals are sampled. Let $T^{(i)} = [T_1^{(i)}, \ldots, T_K^{(i)}]^T$, $Y^{(i)} = [Y^{(i)}(T_1^{(i)}), \ldots, Y^{(i)}(T_K^{(i)})]^T$ and $X^{(i)} = [X_1^{(i)}, \ldots, X_q^{(i)}]^T$, $i = 1, \ldots, n$ be the treatment vector, the vector of potential outcomes, and feature vector of the i^{th} individual, respectively.

The most widely used measure of the treatment effect for the multiple treatment is the pair-wise treatment effect. The individual effect $\xi_{jk}^{(i)}$ between the pairwise treatments: T_j and T_k is defined as $\xi_{jk}^{(i)} = Y^{(i)}\left(T_j^{(i)}\right) - Y^{(i)}(T_k^i)$, the average pairwise treatment effect

$\tau_{jk} = E\left[\xi_{jk}^{(i)}\right]$. The average pairwise treatment effect $\tau_{jk|T_j}$ on the patients treated with T_j is defined as $\tau_{jk|T_j} = E\left[\xi_{jk}^{(i)}|T_j\right]$.

The focus of this paper is on the conditional distribution of treatment effect, given the feature vector X. Let $F_{Y|X}(T_k)$ be the conditional distribution of the potential outcome $Y(T_k)$ under the treatment T_k, given the feature vector X, and $F_{Y|X}(T)$ be the conditional joint distribution of the potential outcome vector $Y(T)$ under the K treatment T, given the feature vector X. Assume that n individuals are sampled. For the i^{th} individual, T_η treatment ($M_\eta = 1$) is assigned. Let $X^{(i)}$ and $Y_\eta^{(i)}\left(T_\eta^{(i)}\right) = Y_f^{(i)}$ be the observed feature vector and the observed potential outcome of the i^{th} individual. Therefore, the observed dataset is given by $D = (X^{(i)},\ T^{(i)},\ Y_\eta^{(i)},\ i = 1, \ldots, n)$. The factual and counterfactual outcomes of the i^{th} individual are denoted by $y_f^{(i)}$ and $y_{cf}^{(i)}$, respectively.

To estimate the treatment effects, we often make the following three assumptions (Rubin, 1974; Yoon et al., 2018a):

Assumption 1 (Ignorability Assumption). Conditional on X, the potential outcomes, $Y(T)$ and the treatment T are independent,

$$Y(T) = (Y(T_1), \ldots,\ Y(T_K))T|X \tag{1}$$

This assumption requires no unmeasured confounding variables.

Assumption 2 (Common Support). For the feature vector X and all treatment,

$$0 < P(T_k = t_k|X) < 1 \tag{2}$$

Assumption 3 (Stable Unit Treatment Value Assumption). No interference (units do not interfere with each other).

Conditional Generative Adversarial Networks as a General Framework for Estimation of Individualized Treatment Effects

The key issue for the estimation of individualized treatment effects is unbiased counterfactual estimation. Counterfactuals will never be observed and cannot be tested by data. The true counterfactuals are unknown. Recently developed generative adversarial networks (GANs) started a revolution in deep learning (Luo and Zhu, 2020). GANs are a perfect tool for missing data imputation. An incredible potential of GANs is to accurately generate the hidden (missing) data distribution given some of the features in the data. Therefore, we can use GANs to generate counterfactual outcomes.

GANs consist of two parts: the "generative" part that is called the generator and "adversarial" part that is called the discriminator. Both the generator and discriminator are implemented by neural networks. Typically, a K-dimensional noise vector is input into the generator network that converts the noise vector to a new fake data instance. Then the generated new data instance is input into the discriminator network to evaluate them for authenticity. The generator constantly learns to generate better fake data instances while the discriminator constantly obtains both real data and fake data and improves accuracy of evaluation for authenticity.

Architecture of Conditional Generative Adversarial Networks (CGANs) for Generating Potential Outcomes

Features provide essential information for estimation of counterfactual outcomes. Therefore, we use conditional generative adversarial networks (CGANs) (Mirza and Osindero, 2014) as a general framework for individualized treatment effect (ITE) estimation. The CGANs for ITE estimation consist of two blocks. The first imputation block is to impute the counterfactual outcomes. The second ITE block is to estimate distribution of the treatment effects using the complete dataset that is generated in the imputation block. The architecture of CGANs is shown in **Figure 1**.

Both the generator and discriminator are implemented by feedforward neural networks. The architectures of the neural networks are described as follows. The generator consists of seven layers of feedforward neural network. The first layer is the covariate input layer that input a vector X of covariates. The second and third layers are hidden layers, each layer with 64 nodes. The fourth layer concatenates the output of the third layer, the response vector Y, treatment vector T and treatment assignment indicator vector M and noise vector Z. The fifth and sixth layers are hidden layers, each layer with 64 nodes. Finally, the seventh layer is the output layer. All activation functions of the neurons were sigmoid function. The architecture of the discriminator is similar to the architecture of the generator except for adding one more output layer with sigmoid non-linear activation function.

Imputation Block

To extend GANTITE from binary treatments to all types of treatments, we introduce the treatment assignment vector and change some mathematical formulation of the generator. A counterfactual generator in the imputation block is a non-linear function of the feature vector, treatment vector T, treatment assignment indicator vector M, observed factual outcome y_f and K dimensional random vector z_G with uniform distribution $z_G \sim U((-1, 1)^K)$ where $Y_f = Y_\eta$. The generator is denoted by

$$\tilde{Y} = G\left(X, Y_f, T \odot M,\ (\mathbf{1} - M) \odot z_G, \theta_G\right) \tag{3}$$

where output $\tilde{Y}$ represents a sample of G. It can take binary values, categorical values or continuous values. $\mathbf{1}$ is a vector of 1, $\odot$ denotes element-wise multiplication, and θ_G is the parameters in the generator. We use Y to denote the complete dataset that is obtained by replacing $\tilde{Y}_\eta$with Y_f.

The distribution of $\tilde{Y}$ depends on the determinant of the Jacobian matrix of the transformation function $G\left(X, Y_f, T, M, z_G, \theta_G\right)$. Changing the transformation function can change the distribution of the generated counterfactual outcomes. Let $P_{Y|x,t,m,y_f}(y)$ be the conditional distribution of the potential outcomes, given $X = x, T = t, M = m, Y_f = y_f$. The

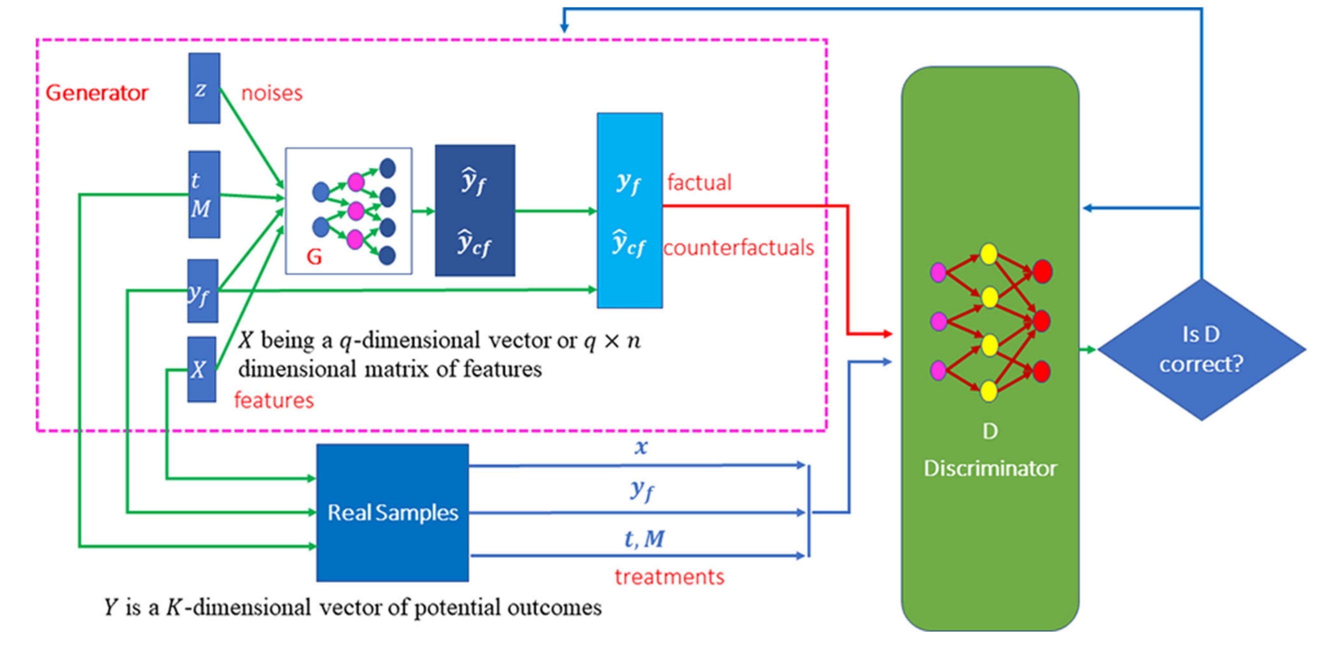


FIGURE 1 | Scheme of MGANITE for the estimation of potential outcomes.

goal of the generator is to learn the neural network G such that $G\left(x, y_f, t, m, z_G, \theta_G\right) \sim P_{Y|x,t,m,y_f}(y)$.

Unlike the discriminator in the standard CGANs where the discriminator evaluates the input data for their authenticity (real or fake data), the counterfactual discriminator D_G that maps pairs (x, y) to vectors in $[0,1]^k$ attempts to distinguish the factual component from the counterfactual components. The output of the counterfactual discriminator D_G is a vector of probabilities that the component represents the factual outcome. Let $D_G(x, \tilde{y},\ t,\ m,\ \theta_d)_i$ represent the probability that the i^{th}component of $\tilde{y}$ is the factual outcome, i.e., $i = \eta$, where θ_d denotes the parameters in the discriminator. The goal of the counterfactual discriminator is to maximize the probability $D_G(x, \tilde{y},\ t, m,\ \ \theta_d)_i$ for correctly identifying the factual component η via changing the parameters in the discriminator neural network D_G.

Loss Function

The imputation block in MGANITE attempts to impute counterfactual outcomes by extending the loss function of the binary treatment in GANITE (Yoon et al., 2018a) to all types of treatments: binary, categorical or continuous treatments. We define the loss function $V(D_G,\ G)$ as

$$E_{(x,t,m,y_f)\sim P_{data}(x,t,m,y_f)} E_{z_G\sim u((-1,1)^K)}$$
$$\left[M^T \log D_G\left(X, \tilde{Y},\ T,\ M\right) + (\mathbf{1}-M)^T \log\left(\mathbf{1} - D_G\left(X, \tilde{Y}, T,\ M\right)\right)\right]$$

where log is an element-wise operation. The goal of the imputation block is to maximize the counterfactual discriminator D_G and then minimize the counterfactual generator G:

$$\min_G \max_{DG} V(D_G,\ G,\ \theta_d) \quad (4)$$

In other words, we train the counterfactual discriminator D_G to maximize the probability of correctly identifying the assigned treatment M_η and the quantity of the treatment T_η or $Y_f(Y_\eta)$, and then train the counterfactual generator G to minimize the probability of correctly identifying M_η and T_η. After the imputation block is performed, the counterfactual generator G produces the complete dataset $\overline{D} = \{x,\ \overline{y}\}$. Next, we use the imputed complete dataset $\overline{D} = \{X,\ \overline{Y}\}$ to generate the distribution of potential outcomes and to estimate the ITE via CGANs which is called the ITE block.

ITE Block

The CGANs consist of three parts: generator, discriminator and loss function which are summarized as follows (Yoon et al., 2018a).

ITE Generator

Unlike the ITE in GANITE where the ITE generator is a non-linear transform function of only X and Z_I, the ITE generator G_I in MGANITE is a non-linear transform function of $X,\ T$ and Z_I:

$$\hat{Y} = G_I(X, T,\ Z_I,\ \theta_{gI}) \quad (5)$$

where $\hat{Y}$ is the generated K-dimensional vector of potential outcomes, X is a feature vector, T is a treatment vector, and Z_I is a K-dimensional vector of random variables and follows the uniform distribution $Z_I \sim u((-1,1)^K)$. The ITE generator attempts to find the transformation $\hat{Y} = G_I(X, T,\ Z_{I,}\ \theta_{gI})$ such that $\hat{Y} \sim\ P_{Y|X,T}(y)$.

ITE Discriminator

Following the CGANs, we define a discriminator D_I as a non-linear classifier with $(X, T, Y^* = Y)$ or $(X, T, Y^* = \hat{Y})$ as input and a scalar that outputs the probability of Y^* being from the complete dataset D.

Loss Function

Again, unlike the loss function in GANITE where the decision function is $D_I(X, Y^*)$, a decision function in MGANITE is defined as $D(X, T, Y^*)$. The loss function for the ITE block in MGANITE is then defined as

$$V_I\,(D_I,\ G_I) = E_{X,T\sim P(x,T)} \left[E_{Y^*\sim P_{Y|X,T}}(y)\left[\log D_I(X,T,Y^*\ )\right] + E_{Z_I\sim u((-1,1)^K)}\left[\log\left(1 - D_I X,T,Y^*\right)\right]\right] \quad (6)$$

where $D_I(X, T,\ Y^*)$ is the non-linear classifier that determines whether Y^* is from the complete dataset $\overline{D}$ or from generator G_I.The goal of the ITE block is to maximize the probability of correctly identifying that Y^* is from the complete dataset $\overline{D}$ and to minimize the probability of a correct classification. Mathematically, the ITE attempts

$$\min{}_{GI}\max{}_{DI} V_I\,(D_I,\ G_I) \quad (7)$$

The algorithms for numerically solving the optimization problems (4) and (7) are summarized in the **Supplementary Note**.

The learning parameters for the feedforward neural networks are given below. We set batch size equal to 16. We assumed that the learning rates for the discriminator and generator were 0.0001 and 0.001, respectively. We further assume that the decay rate was 0.1. The learning rate decayed (exponentially) to 10% of the starting learning rate during 70% of the total batches, and stayed at 10% during the last 30% batches. The total number of batches was 1,000,000. Adam Optimizer was used to perform optimization. We assume that 20% of the nodes were dropped randomly during the training process.

Sparse Techniques for Biomarker Identification

The LASSO (least absolute shrinkage and selection operator) that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the results can be used to select biomarkers for optimal treatment selection (Ali and Aittokallio, 2019). Let Y_k^i and $X^{(i)}$denote the estimated effect of the k^{th} treatment and feature vector of the i^{th} individual, respectively. Let

$$Y_T = \begin{bmatrix} Y_1^1 \cdots Y_K^1 \\ \vdots\vdots\vdots \\ Y_1^n \cdots Y_K^n \end{bmatrix}, X = \begin{bmatrix} x_1^{(1)} \cdots x_q^{(1)} \\ \vdots\vdots\vdots \\ x_1^{(n)} \cdots x_q^{(n)} \end{bmatrix},\ \beta = \begin{bmatrix} \beta_{11} \cdots \beta_{1K} \\ \vdots\vdots\vdots \\ \beta_{q1} \cdots \beta_{qK} \end{bmatrix}$$

The outputs of the neural networks are in general a continuous function even if the potential outcomes are binary. For the convenience of presentation, we assume that the treatment effects are continuous regardless if the potential outcomes are binary, categorical or continuous.

TABLE 1 | Performance of six methods for estimating the potential outcomes.

Methods	MSE	STD	Accuracy
MGANITE	0.062	0.235	0.938
LR	0.104	0.305	0.896
LogR	0.120	0.325	0.880
SVM	0.126	0.332	0.874
KNN	0.148	0.355	0.852
RF (C)	0.098	0.297	0.902

The LASSO estimators for identifying biomarkers that predict treatment effects are given by

$$\hat{\beta}_\lambda = \text{argmin}_\beta ||Y_T - X\beta||_F^2 + \lambda \sum_{j=1}^{q}\sum_{l=1}^{K} |\beta_{jl}| \quad (8)$$

where $\|.\|_F$ is the Frobenius norm of the matrix. Non-zero elements $\beta_{jl} \neq 0$ predict treatment effect variation and hence its correspondence $X_j = \left[X_j^{(1)} \cdots X_j^{(n)}\right]^T$can be used as biomarkers for investigation of the l^{th} treatment. For the continuous treatment, we define the treatment matrix T and its associated coefficient matrix Γ:

$$T = \begin{bmatrix} T_1^{(1)}) \cdots T_K^{(1)} \\ \vdots\vdots\vdots \\ T_1^{(n)} \cdots T_K^{(n)} \end{bmatrix}, \quad \Gamma = \begin{bmatrix} \gamma_{11} \cdots \gamma_{1K} \\ \vdots\vdots\vdots \\ \gamma_{K1} \cdots \gamma_{KK} \end{bmatrix}$$

Equation (8) should be changed to

$$[\ \hat{\gamma}_{\lambda_1},\ \ \hat{\beta}_{\lambda_2}] = \text{argmin}_{\gamma,\beta} ||Y_T - T\Gamma - X\beta||_F^2 + \lambda_1 \sum_{j=1}^{K}\sum_{l=1}^{K} |\gamma_{jl}| +\ \lambda_2 \sum_{j=1}^{q}\sum_{l=1}^{K} |\beta_{jl}| \quad (9)$$

where $\lambda_1,\ \lambda_2$ are penalty parameters.

Biomarker Identification for Optimal Treatment Selection

Consider K treatments. Let $\hat{Y}^i = \left[\hat{Y}_1^i \cdots \hat{Y}_K^i\right]^T$ be the K-dimensional vector of the estimated potential outcomes for the i^{th} individual and $z_i = \text{argmax}_{1,...,k}\{\hat{Y}_1^i, \ldots,\ \hat{Y}_K^i\ \}$ be the index for the optimal potential outcomes of the i^{th} individual. To select biomarkers for optimal treatment selection, we define the following LASSO:

$$\hat{Y}_{z_i}^i = \sum_{j=1}^{q} x_j^{(i)}\alpha_j + \lambda \sum_{j=1}^{q} |\alpha_j|,\ i = 1, \ldots,\ n \quad (10)$$

Solving the above categorical LASSO problem, we obtain a set of non-zero coefficients that are denoted as $\hat{\alpha}_l \neq 0,\ l = 1, \ldots,\ L$. The covariates that correspond to the non-zero coefficients of the

LASSO solution are chosen as biomarkers for optimal treatment selection. Again, for the continuous treatment, Equation (10) needs to be changed to

$$\hat{Y}^{i}_{z_i} = \sum_{l=1}^{K} T_l^{(i)}\delta_l + \sum_{j=1}^{q} x_j^{(i)}\alpha_j + \lambda_1 \sum_{l=1}^{K} |\delta_l| + \lambda_2 \sum_{j=1}^{q} |\alpha_j|, i = 1, \ldots, n. \quad (11)$$

Data Collection

The proposed MGANITE was applied to 256 newly diagnosed acute myeloid leukemia (AML) patients, treated with high dose ara-C (HDAC), Idarubicin (IDA), and HDAC+IDA at M. D. Anderson Cancer Center. There were 212 valid samples and 85 useable features (14 discrete and 71 continuous), including 51 total and phosphoprotein from several biological processes such as apoptosis, cell-cycle, and signal transduction pathways (Kornblau et al., 2009). Among the 212 valid samples, 37 were treated with HDAC, 9 were treated with IDA and 54 were treated with HDAC+IDA, and 112 were treated with other drugs. Data were downloaded from the M. D. Anderson Cancer Center database (http://bioinformatics.mdanderson.org/Supplements/Kornblau-AML-RPPA/aml-rppa.xls) and (https://pubmed.ncbi.nlm.nih.gov/18840713/).

Prediction accuracy was defined as the proportions of correctly predicted potential outcomes. The false positive rate was defined as the proportion of individuals who were wrongly classified as having a positive treatment response. Discriminator accuracy is defined as the proportion of correctly classified real or fake samples. Replication error is defined as cross entropy $-y_f \log \hat{y}_f$ where $\hat{y}_f = G\left(x, t, t_*, y_f, z_G, \theta_g\right), t = t_*$ and separate distance is defined as

$$\frac{1}{n}\sum_{i=1}^{n} |y_{if} - \hat{y}_{if}|$$

where $\hat{y}_{if} = G\left(x, t, t_*, y_f, z_G, \theta_g\right), t \neq t_*$.

RESULTS

Simulations

We first examine the performance of MGANITE in estimating the ITE of binary treatment using simulations. A synthetic dataset is generated as follows. A total of 10,000 individuals with 30-dimentinal feature vectors follow the normal distributions $N(0, I)$. Let

$$\hat{y}_i^0 = 0.05 + 0.4x_{i1}^2 + 0.25x_{i2} + n_{i0}, n_{i0} \sim N(0, 0.05)$$

and

$$\begin{aligned}\hat{y}_i^1 &= 0.15 + 0.5x_{i1}^2 + 0.25x_{i1}x_{i2} + 0.25x_{i2} + n_{i1} \\ i &= 1, 2, \ldots, 10,000, n_{i1} \sim N(0, 0.05),\end{aligned}$$

where i is a sample index.

Then, the potential outcomes are generated as

$$y_i^0 = \begin{cases} 1 & \hat{y}_i^0 \geq 0.5 \\ 0 & \hat{y}_i^0 < 0.5 \end{cases} \text{ and } y_i^1 = \begin{cases} 1 & \hat{y}_i^1 \geq 0.5 \\ 0 & \hat{y}_i^1 < 0.5 \end{cases}$$

Treatment is assigned by the Bernoulli distribution:

$$M = T|X \sim Bern(sigmoid\left(W_t^T X + n_t\right))$$

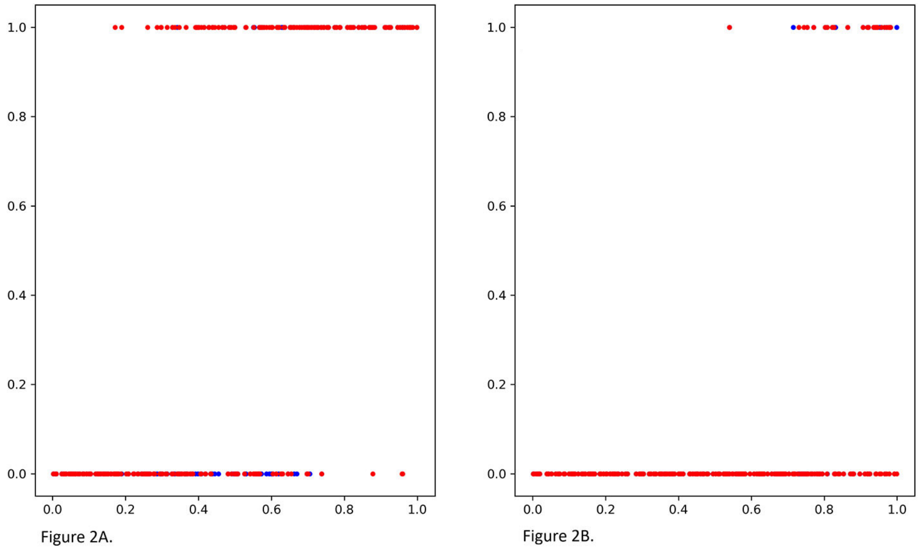


FIGURE 2 | (A) The true potential outcomes with treatment Y^1 and estimated potential outcomes $\hat{y}^1$ using MGANITE, where the x axis denoted a value of covariate X_1, the y axis denoted the potential outcome, a blue color dot represented the true outcome Y^1 and a red color dot represented the estimated outcomes $\hat{y}^1$. **(B)** The true potential outcomes without treatment $Y^0$and estimated potential outcomes $\hat{y}^0$ using MGANITE, where the x axis denoted a value of covariate X_1, the y axis denoted the potential outcome, a blue color dot represented the true outcome Y^0 and a red color dot represented the estimated outcomes $\hat{y}^0$.

where t is a treatment index, $W_t^T \sim u(-0.1, 0.1)^{30\times 1}$, $n_t \sim N(0, 0.1)$, and Bern represents the Bernoulli distribution. When one sample has only one treatment assigned, then $t = i$.

Treatment effect can take three values 1, 0, and −1. In other words,

$$\xi_i = \begin{cases} 1 & y_i^1 = 1,\ y_i^0 = 0 \\ 0 \quad y_i^1 = 1,\ y_i^0 = 1 & or\ y_i^1 = 0,\ y_i^0 = 0 \\ -1 & y_i^1 = 0,\ y_i^0 = 1 \end{cases}$$

We compare MGANITE with linear regression (LR) (Makar et al., 2019), logistic regression (LogR) (Emmert-Streib and Dehmer, 2019; Makar et al., 2019), support vector machine (SVM) (Makar et al., 2019), k- nearest neighbor (k-NN) (Crump et al., 2008), Bayesian linear regression (BLR) (Johansson et al., 2016), causal forest (CForest) (Wager and Athey, 2015), and random forest classification [RF (C)] (Breiman, 2001). We use six methods: MGANITE, LR, LogR, SVM, kNN, and RF (C) to estimate the counterfactual potential outcomes and calculate the mean square error (MSE) between the estimated treatment effect and the true treatment effect, standard deviation (STD) and prediction accuracy. **Table 1** presents MSE, STD, and prediction accuracy of six methods to fit the generated data. We observe that MGANITE more accurately estimate the potential outcomes than the other five state-of-the-art methods. **Figure 2** presents the true counterfactuals and estimates counterfactuals using MGANITE. We observe that MGANITE reaches remarkably high accuracy for estimating counterfactuals.

The treatment effect estimation of eight methods [MGANITE, LR, LogR, SVM, KNN (5,10), BLR, RF (C), RF (R)] are summarized in **Table 2**. **Table 2** shows that MGANITE has the highest accuracy of estimation of all treatment effects: average treatment effect (ATE), average treatment effects on the treated (ATT), and average treatment effect on the control (ATC), followed by RF (R) or RF (C). We observe that the estimations of ATE using all methods are inflated. The inflation rates of ATE using MGANITE and RF (C) are 3.9 and 7.9%, respectively. The SVM reaches the inflation rate of the estimation of ATE as high as 29.8%. All inflation rates of estimation of ATE using LR, LogR, SVM, KNN, and BLR are very high. The simulations also show that the false positive rates using MGANITE, LR, LogR, SVC, KNN (5), KNN (10), BLR, RF (R), and RF (C) are 3.9, 24.7, 28.1, 29.8, 28/1, 19.7, 25.3, 9, and 8.4%, respectively. The results show that false positive rates of LR, LogR, SVM, KNN, and BLR for prediction of positive treatment response are too high to be applied to treatment selection. Even RF (R) reaches the false positive rate as high as 8.4%. **Table 2** also shows that the number of individuals that show positive treatment effects increases while the number of individuals that show no treatment effect decreases from ground truth.

TABLE 2 | Treatment effects estimated for simulation data using nine methods.

Methods	ATT	ATC	ATE	ITE = −1	ITE = 0	ITE = 1
Ground truth	0.391	0.321	0.356	0	322	178
MGANITE	0.399	0.341	0.37	0	315	185
LR	0.52	0.369	0.444	0	278	222
LogR	0.52	0.393	0.456	0	272	228
SVM	0.524	0.401	0.462	0	269	231
KNN (5)	0.508	0.401	0.454	1	271	228
KNN (10)	0.524	0.325	0.424	1	286	213
BLR	0.524	0.369	0.446	0	277	223
RF (C)	0.452	0.325	0.388	0	306	194
RF (R)	0.431	0.337	0.384	1	306	193

Next we examine the performance of MGANITE in estimating the ITE of continuous treatment using simulations. A synthetic dataset is generated as follows.

1. Draw the covariate variable X from the standard normal distribution for 10,000 individuals.
2. The treatment T is exponentially distributed as $P(t) = e^{-(t-1)},\ t \geq 1$. Define $g(t) = 0.1t^2$.
3. Define a non-linear function $f(x) = \frac{1}{2+exp\left(-20\left(x-\frac{1}{3}\right)\right)}$.
4. Define $y_i^0 = 0.3 + f(x) + n_i^0,\ i = 1,.., 10,000$, where n_i^0 is a randomly sampled noise variable from a normal distribution $N(0, 0.01)$.
5. Define $y_i^1 = 0.3 + f(x) + g(t) + n_i^1,\ i = 1, \ldots, 10,000$, where n_i^1 is a randomly sampled noise variable from a normal distribution $N(0, 0.01)$.
6. Treatment assignment indicator variable M_i is drawn from a Bernoulli distribution with $P = 0.5$ for each subject.

The mean square errors (MSE) for MGANITE, Linear Regression, KNN, Bayesian ridge regression, RF (R), and SVM regression are 0.011004916, 0.08500695, 0.012520364, 0.085007192, 0.014281599, 0.013962992, respectively. **Figures 3A,B** plot the true ITE and estimated ITE for in-samples and out-of-samples data, using six methods: MGANITE, LR, KNN, BLR, RF (R), and SVM, respectively, where a dash straight line indicates that the true ITE and the estimated ITE are equal. We observe from **Figures 3A,B** that many green cross points for both in-sample and out-of-sample data are much closer to the dash straight line than other types of points. This shows that the estimated ITE points using MGANITE are much closer to the true ITE point than using the other five methods. In other words, the estimator of the ITE using MGANITE is more accurate than that of using the other five methods. The results clearly demonstrate that MGANITE outperforms the 5 other state-of-the-art treatment effect estimation methods.

To further evaluate the performance of MGANITE, we provide **Figure 4** that plots the receiver operating characteristic (ROC) curve for evaluation of the ability of MGANITE to predict potential outcomes of treatment. Our calculation shows that area under the ROC curve (AUC) for MGANITE reaches 0.98, which is a very high value. The ROC curve and AUC value demonstrate that the power of MGANITE for prediction of the potential outcomes of the treatments is very high.

Real Data Analysis

MGANITE is applied to 256 newly diagnosed acute myeloid leukemia (AML) patients from the clinical trial dataset (Kornblau et al., 2009). We first present the results of treatment using

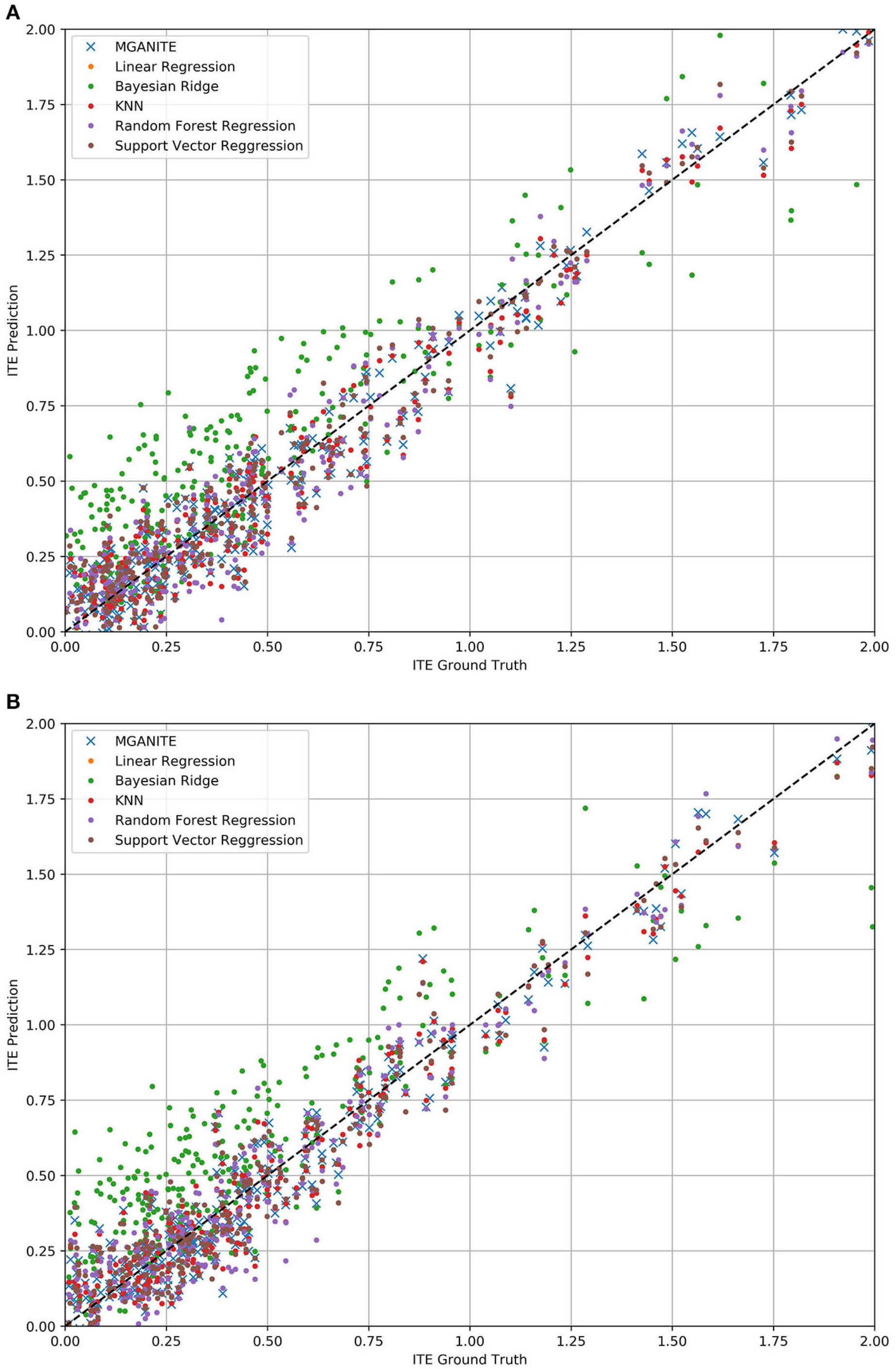


FIGURE 3 | (A) True ITE and estimated ITE for in-sample data using six methods: MGANITE, LR, KNN, BLR, RF (R), and SVM, where MGANTE was denoted by a green cross point, LR was denoted by an orange point, KNN was denoted by a green point, BLR was denoted by a red point, RF (R) was denoted by a purple point and SVM was denoted by a dark red point, the *x* axis denoted the true ITE and the *y* axis denoted the estimated ITE. **(B)** True ITE and estimated ITE for out-of-sample data using six methods: MGANITE, LR, KNN, BLR, RF (R), and SVM, where MGANTE was denoted by a green cross point, LR was denoted by a orange point, KNN was denoted by a green point, BLR was denoted by a red point, RF (R) was denoted by a purple point and SVM was denoted by a dark red point, the *x* axis denoted the true ITE and the *y* axis denoted the estimated ITE.

HDAC, HDAC+IDA (101) vs. all other drugs (111). A key issue for MGANITE is how to train MGANITE. To track the training process of MGANITE, we present **Figure 5** that shows ATE, discriminator accuracy, replication error, and separate distance curves as a function of the number of batches.

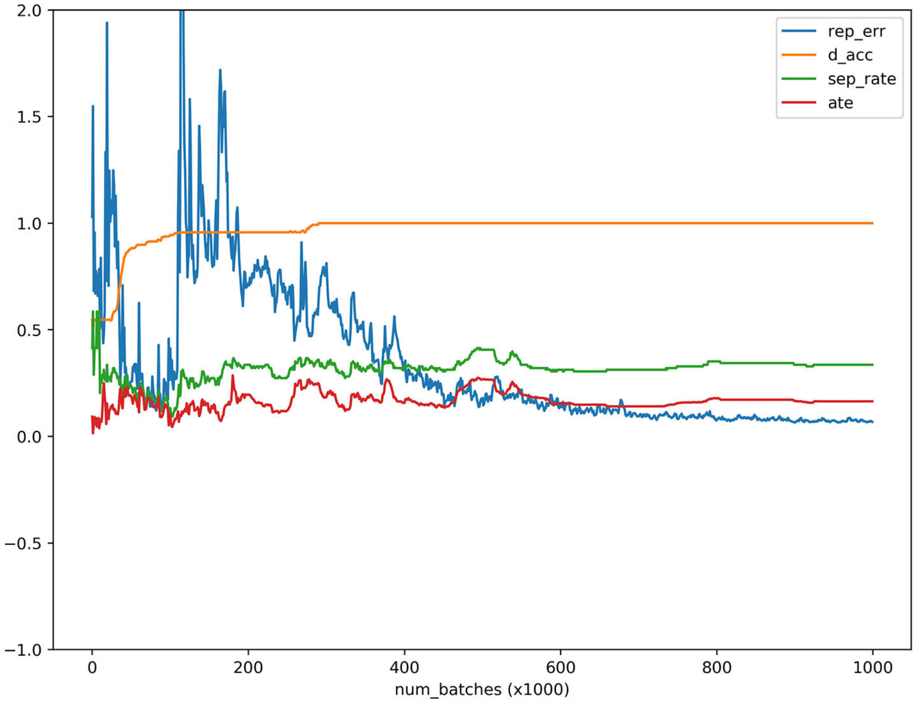


FIGURE 4 | ATE, discriminator accuracy, replication error and separate distance curves as a function of the number of batches where the *x* axis denoted the number of batches, the *y* axis denoted values of the ATE, discriminator accuracy, replication error, and separation distance for ATE, discriminator, replication, and separation curves, respectively, red, orange, blue and green curves were ATE, discriminator, replication and separation curves, respectively.

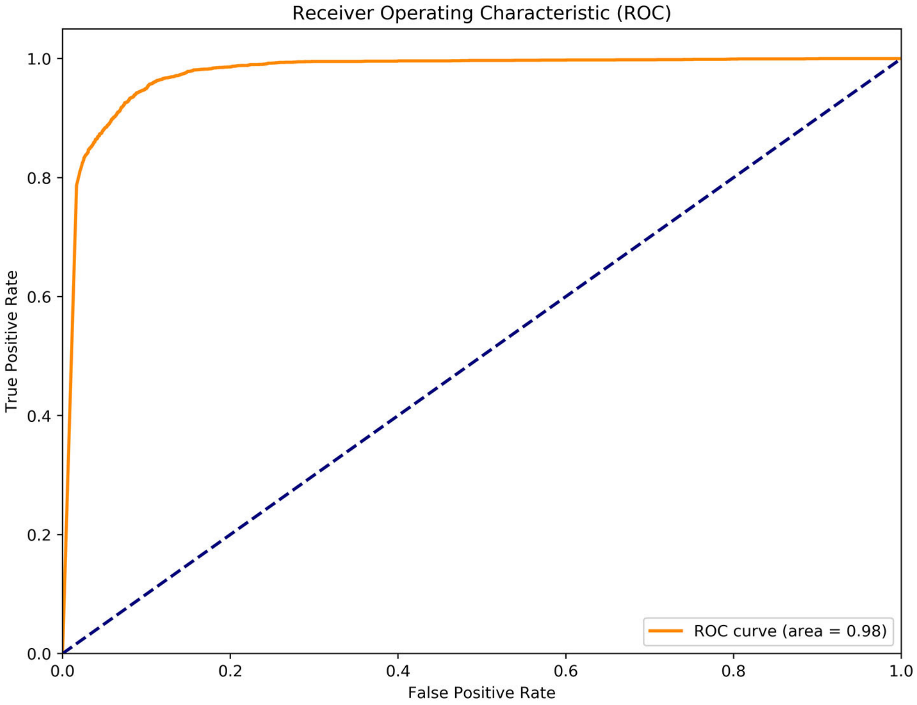


FIGURE 5 | Receiver operating characteristic (ROC) curve for evaluation of performance of MGANITE.

TABLE 3 | Treatment effects estimated for AML dataset using nine methods.

Methods	ATT	ATC	ATE	Number of individuals with positive treatment effect		
				HDAC and HDAC+IDA	No difference	Other drugs
CGANs	0.011	0.356	0.208	59	138	15
LR	0.033	0.207	0.107	62	112	38
LogR	0.083	0.209	0.137	63	115	34
SVM	0.112	0.165	0.135	65	130	17
KNN (5)	0.248	−0.011	0.137	55	131	26
KNN (10)	0.314	0.066	0.208	62	132	18
BLR	0.129	0.139	0.133	57	136	19
RF (C)	0.157	0.286	0.212	70	117	25
RF (R)	0.052	0.099	0.072	37	155	20

We observe from **Figure 5** that discriminator accuracy converges to 1, replication error converges to zero, separation distance converges to a constant, and ATE converges to a stable value. **Figure 4** demonstrates that MGANITE is trained very well.

Next we compared the treatment effect estimations using nine methods: MGANITE, LR, LogR, SVM, KNN (5), KNN (10), BLR, RF (C), and RF (R) where 5 and 10 are the number of neighbors. Treatment with HDAC or HDAC+IDA, and 85 protein expressions and other geographical variables are used as covariates. The response status (response or no response) is used as the outcome.

Table 3 summarizes results of the estimation of HDAC treatment effect using MGANITE and other eight methods where individuals with HDAC or HDAC+IDA are taken as the treated population and individuals with other drugs are taken as the control population. Comparison of treatment effect estimation algorithms on real data analysis is not easy because of the lack of ground truth treatment effects and small sample sizes. In general, using MGANITE, we observe that the majority of individuals who are treated by other drugs do not show any response and that 65% of the individuals who are treated by HDAC or HDAC+IDA respond. Only 13.5% of individuals who are treated by other drugs respond. To illustrate the difference between the estimated treatment effect and treatment response, we present **Figure 6** that shows the histogram of the estimated effects of the treatments HDAC or HDAC+IDA vs. other drugs using MGANITE (**Figure 6A**), and observe the number of responses of the individuals in the population who are treated with HDAC or HDAC+IDA vs. other drugs (**Figure 6B**). ITE is calculated based on both the factual and counterfactual. We observe that $ITE = 0$ consists of two scenarios: (1) no response of the patients to any drugs and (2) response of the patients to both HDAC or HDAC+IDA, and other drugs. A proportion of the patients with response to HDAC or HDAC+IDA on the right side of **Figure 6B** and the patient with response to other drugs on the left side of **Figure 6B** has $ITE = 0$. The observed response of the patients to one drug does not imply that these patients would not respond to other drugs. However, $ITE = 1$ or $ITE = 0$ implies that the patients respond to only one type of drug. To further compare the performance of MGANITE and other methods for evaluation of ITE, we split a given data set into an in-sample dataset (190 samples), used for the initial parameter estimation and model selection, and an out-of-sample dataset (22 samples), used to evaluate performance of ITE estimation. The results are summarized in **Table 4**. We observe that the difference in the estimated ATT, ATC, ATE and proportions of the ITE between in-samples and out-of-samples using MGANITE are much smaller than using other methods. This shows that the ITE estimation using MGANITE is more robust than using other methods. We calculate the Kullback-Leibler (K-L) divergence between the distributions of the ITE using in-sample and out-of-samples, and using nine methods. The results are summarized in **Table 5**. **Table 5** shows that K-L divergence using MGANITE is much smaller than that using other methods, which implies that MGANITE is more robust than the other eight methods.

LASSO is used to identify biomarkers for prediction of treatment effect and treatment selection. **Table 6** lists the top 30 biomarkers identified by LASSO. All top 30 biomarkers explain 36.82% of the variation of HDAC or HDAC+IDA treatment effect. The top Gene *GSK3* accounts for 4.4% of the explanation of treatment effect variation.

Garson's algorithm (Garson, 1991; Siu, 2017; Zhang et al., 2018) that describes the relative magnitude of the importance of input variables (biomarkers) in its connection with outcome variables (ITE) of the neural network can also be used to identify biomarkers for predicting the ITE. The top 30 biomarkers identified by the Garson algorithm are listed in **Supplementary Table 1** where the relative contribution of each biomarker to the ITE variation and cumulative contribution of biomarkers to the ITE variation are also listed in **Supplementary Table 1**. The correlation coefficient between the importance ranking of the markers using the Garson algorithm and LASSO is only −0.05.

Next, we study the joint estimation of effects of the multiple treatments. The number of individuals that are treated with HDAC, HDAC+IDA, and other drugs are 37, 54, and 121, respectively. The widely used treatment estimation methods with multiple treatments are simultaneous estimations of the effects of pairwise treatments. We estimate the effects of the pairwise treatments HDAC vs. HDAC+IDA, HDAC vs. other drugs, and HDAC+IDA vs. other drugs. The results are summarized in **Table 7**. Pairwise comparisons listed in **Table 7** does not present the results of the treatment compared with a placebo (without using any drugs). We compare the effect of one treatment with another treatment. Specifically, we make pairwise comparisons: HDAC vs. other drugs, HDAC+IDA vs. other drugs, and HDAC+IDA vs. HDAC. The average treatment effects (ATE) of these three pairwise treatments: HDAC vs. other drugs, HDAC+IDA vs. other drugs, and HDAC+IDA vs. HDAC using MGANITE, are 0.1001, 0.2311 and 0.1310, respectively. This demonstrates that on the average, the effect of the HDAC+IDA is the largest among the three treatments: HDAC+IDA, HDAC, and other drugs, followed by the treatment HDAC. In other words, the treatment HDAC is better than other drugs, in turn,

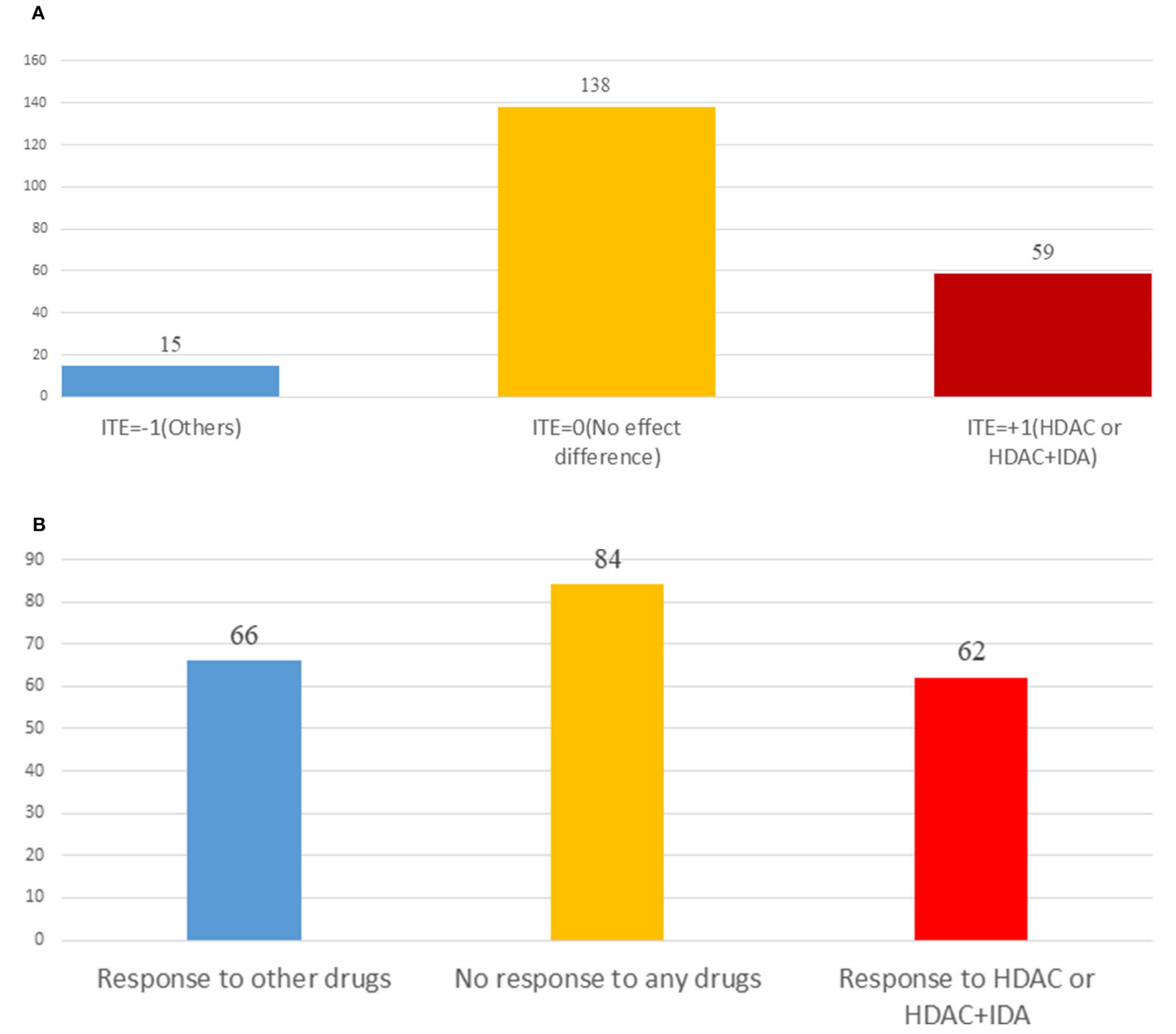


FIGURE 6 | (A) Histogram of estimated drug treatment effect using MGANITE, where the *x* axis denoted the value of ITE and the *y* axis denoted the number of patients, $ITE = +1$ denoted the ITE of patients treated with HDAC or HDAC+IDA, $ITE = -1$ denoted the ITE of patients treated with other drugs, and $ITE = 0$ denoted the ITE of two groups of patients: one group of the patients treated with HDAC or HDAC+IDA and another group of the patients treated with other drugs. **(B)** Histogram of observed drug treatment response where the *x* axis indicated three scenarios as described in **(B)** and the *y* axis denoted the number of patients, the right side in the **(B)** denoted the number of patients only responding to the HDAC or HDAC+IDA, the middle denoted the number of the patients that responds to both (HDAC or HDAC+IDA) and other drugs or did not respond to both (HDAC or HDAC+IDA) and other drugs, and the left side denoted the number of patients only responding to the other drugs.

the combination of HDAC and IDA is better than HDAC. It is also noted that the effect of HDAC+IDA vs. other drugs—effect of HDAC vs. other drugs = 0.2311–0.1001 = 0.1310 = effect of HDAC+IDA vs. HDAC.

However, using LR, LogR, SVM, RF (C), and RF (R), we observe that HDAC is the best treatment. This conclusion violates the biological interpretation. We explain the reasons that causes this incorrect conclusion as follows. The traditional methods for treatment estimation are mainly based on the population average of the treatment responses. The number of observed responses and no responses of the individuals treated with other drugs is 66 and 55, respectively. The average response rate of the other drugs is 0.545. The number of observed responses and no responses of the individuals treated with HDAC is 29 and 8, respectively. The average response rate for HDAC is 0.784. The number of observed response and no response of individuals treated with HDAC + IDA is 33 and 21, respectively. The average response rate for HDAC +IDA is

TABLE 4 | Treatment effects estimated for AML dataset using nine methods.

Method	ATT	ATC	ATE	ITE = −1	ITE = 0	ITE = 1
					Proportion	
In-sample						
MCGAN	0.3152	0.2733	0.2911	0.0842	0.5474	0.3684
LR	0.1077	−0.021	0.0339	0.2474	0.4789	0.2737
BLR	0.0843	0.0817	0.0828	0.1158	0.6684	0.2158
KNN (5)	−0.0247	0.1743	0.0895	0.1474	0.6158	0.2368
KNN (10)	0.0494	0.1835	0.1263	0.1211	0.6316	0.2474
RF (C)	0.2099	0.0826	0.1368	0.1421	0.5789	0.2789
RF (R)	0.0852	0.0459	0.0626	0.1316	0.6737	0.1947
LogR	0.1358	0.1193	0.1263	0.1579	0.5579	0.2842
SVM	0.1081	0.0571	0.0788	0.1158	0.6263	0.2579
Out-of-sample						
MCGAN	0.2000	0.3266	0.2691	0.0455	0.6364	0.3182
LR	0.4974	0.1222	0.2928	0.0909	0.5000	0.4091
BLR	0.3470	0.3129	0.3284	0.0000	0.5909	0.4091
KNN (5)	0.3000	0.5000	0.4091	0.0455	0.5000	0.4545
KNN (10)	0.2000	0.5000	0.3636	0.0000	0.6364	0.3636
RF (C)	0.0000	0.3333	0.1818	0.0455	0.7273	0.2273
RF (R)	0.2600	0.3583	0.3136	0.0000	0.7273	0.2727
LogR	0.6000	0.4167	0.5000	0.0000	0.5000	0.5000
SVM	0.3502	0.2823	0.3132	0.0000	0.5455	0.4545

TABLE 5 | K-L divergence between the distribution of ITEs using in-samples and out-of-samples.

Methods	Kullback–Leibler divergence
MGANITE	0.00920
LR	0.04123
BLR	0.08201
KNN (5)	0.06024
KNN (10)	0.06293
RF (C)	0.02932
RF (R)	0.06407
LogR	0.09887
SVM	0.07913

0.611. Therefore, estimators of ATE for the treatment of HDAC vs. other drugs using LR, LogR, SVM, RF (C), and RF (R) are higher than the estimators of ATE for the HDAC + IDA treatment. However, the individuals treated with HDAC+IDA usually do not respond to HDAC treatment. Therefore, the number of individuals with no response should be adjusted to 62. After adjustment, the response rate of HDAC is changed to 0.319. Therefore, after adjustment, the ATE of HDAC vs. other drugs is smaller than the ATE for HDAC +IDA. Then, the estimators of the pair-wise treatments using MGANITE are consistent with the treatment responses after the data are adjusted. This example shows that these traditional methods that are designed for single treatment effect estimation should be modified when they are applied to multiple treatment effect estimation.

Enrichment analysis to top ranking variables for explanation of treatment effect variation is performed by the hypergeometric test via the Reactome Pathway Database (RPD) (Jassal et al., 2020) to assess whether the number of identified biomarkers associated with the Reactome pathway is over-represented more than expected. The original P-value from the hypergeometric test is then adjusted by FDR for multiple test correction. We find that top ranking biomarkers for the explanation of treatment effect variation are enriched in multiple cancer related pathways (**Figure 7A**), including the intrinsic pathway for apoptosis (R-HSA-109606, $P = 2.86 \times 10^{-14}$), Signaling by Interleukins (R-HSA-449147, $P = 2.86 \times 10^{-14}$), Programmed Cell Death (R-HSA-5357801, $P = 9.7 \times 10^{-11}$), PIP3 activates AKT signaling (R-HSA-1257604, $P = 2.98 \times 10^{-8}$), RUNX3 regulates WNT signaling (R-HSA-8951430, $P = 1.03 \times 10^{-5}$), and RNA Polymerase II Transcription (R-HSA-73857, $P = 9.4 \times 10^{-5}$). In addition, we find that the drug target of idarubicin (TOP2A) and Cytarabine (POLB) form a significant protein-protein interaction network ($P < 1.0 \times 10^{-16}$) (Szklarczyk et al., 2019), indicating that the predictive biomarkers work as the direct interactive proteins of cancer drug targets (**Figure 7B**).

DISCUSSION

In this paper, we present MGANITE coupled with sparse techniques as a framework to estimate the ITEs and select

TABLE 6 | Top ranking variables for explanation of treatment effect variation.

Gene name	R-square (single)	R-square (accumulated)	Gene name	R-square (single)	R-square (accumulated)
GSK3	0.0440	0.0440	CD33	0.0134	0.1984
BILIRUBIN	0.0411	0.0790	TP53	0.0118	0.2376
DIABLO	0.0370	0.1266	STAT3	0.0085	0.2415
SRC	0.0333	0.1329	BIRC5	0.0071	0.2421
MEK	0.0282	0.1373	BAX	0.0070	0.2446
AKT.p308	0.0244	0.1405	DJI	0.0061	0.2591
Age_at_Dx	0.0226	0.1488	CREATININE	0.0057	0.2627
PRIOR_XRT	0.0202	0.1776	BAD	0.0052	0.2646
PSMC4	0.0196	0.1844	ACTB	0.0052	0.2816
PB_Blast	0.0181	0.1858	WBC	0.0045	0.2922
BM_Blast	0.0167	0.1878	PRIOR_MAL	0.0042	0.3190
CD20	0.0167	0.1883	FIBRINOGEN	0.0038	0.3213
NRP1	0.0147	0.1914	STAT6	0.0033	0.3383
TP38.p	0.0143	0.1954	CD13	0.0033	0.3409
PSMC4	0.0135	0.1971	PTEN	0.0030	0.3682

TABLE 7 | Multiple treatment effects estimated for AML dataset using nine methods.

	ATE	Number of individuals with treatment effect		
Method	**HDAC vs. other**	**HDAC**	**No difference**	**Other**
MGANITE	0.1001	59	115	38
LR	0.1149	58	122	32
LogR	0.0896	54	123	35
SVM	0.1463	59	140	13
KNN (5)	0.1887	62	128	22
KNN (10)	0.3538	80	127	5
BLR	0.0860	45	138	29
RF (C)	0.2264	73	114	25
RF (R)	0.2127	58	145	9
Method	**HDAC+IDA vs. other**	**HDAC+IDA**	**No difference**	**Other**
MGANITE	0.2311	79	103	30
LR	0.0965	59	115	38
LogR	0.2123	56	115	41
SVM	0.2453	52	138	22
KNN (5)	0.1012	62	133	17
KNN (10)	0.1604	63	138	11
BLR	0.1307	49	137	26
RF (C)	0.0708	70	106	36
RF (R)	0.0835	43	155	14
Method	**HDAC+IDA vs. HDAC**	**HDAC+IDA**	**No difference**	**HDAC**
MGANITE	0.1310	52	136	24
LR	−0.0184	36	130	46
LogR	−0.0189	45	118	49
SVM	−0.0628	9	181	22
KNN (5)	0.0236	34	149	29
KNN (10)	−0.1085	11	167	34
BLR	0.0152	40	139	33
RF (C)	−0.0660	31	136	45
RF (R)	−0.0821	8	184	20

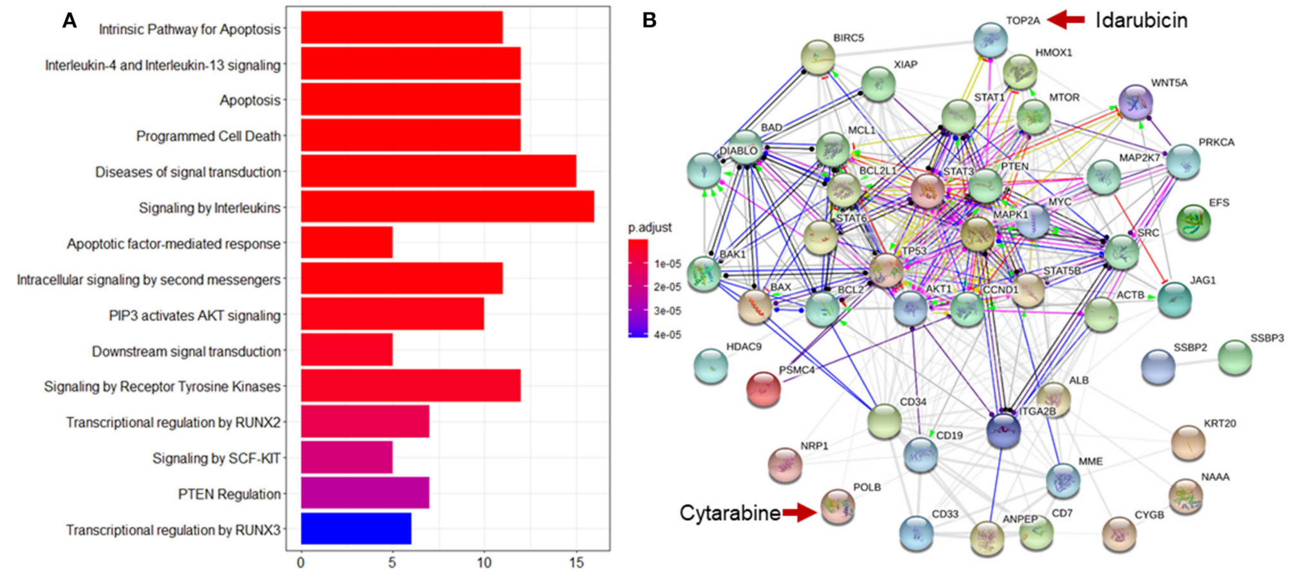


FIGURE 7 | Reactome pathway analysis and protein-protein interaction (PPI) network analysis to top ranking biomarkers for explanation of treatment effect variation. **(A)** Enrichment analysis to the top 44 ranking biomarkers for explanation of treatment effect variation with the Reactome pathway database by hypergeometric test to assess whether the number of identified biomarkers associated with the Reactome pathway was over-represented more than expected. The original *P*-value from the hypergeometric test was then adjusted by FDR for multiple test correction. The top 15 most significantly enriched pathways was shown. **(B)** PPI network analysis was performed by String 11.0 to show the protein-protein interaction among top ranking biomarkers. We found that these proteins were highly interacted which was consistent with pathway enrichment analysis (PPI enrichment *P*-value is 1.0e-16).

the optimal treatments. We demonstrate that the proposed MGANITE has several remarkable features.

First, MGANITE extends GANITE from binary treatment to all types of treatments: binary, categorical, and continuous treatments. We show that MGANITE has a much higher accuracy for estimation of ITE than other state-of-the-art methods.

Second, in-sample and out-of-sample analysis show that the K-L divergence between the distributions of ITE for in-sample and out-of-samples for MGANITE is much smaller than that of other methods, which implies that MGANITE is more robust than other state-of-the art methods.

Third, unlike many popular methods that are usually used to estimate the average effect of the single treatment, MGANITE not only can estimate the ITE of a single treatment, but also can accurately and jointly estimate the ITE of multiple treatments. We also show that the results of the joint estimation of multiple treatments using other classical methods are inconsistent and might violate the biological interpretation.

Fourth, precision oncology is the identification of the right treatment for the right patient. The essential aim is to discover biomarkers that can accurately predict individual treatment effect for each individual. Our results show that MGANITE with sparse techniques can identify a set of biomarkers with significant biological features. The following identified biomarkers are such typical examples.

GSK3 is a kinase so adaptable that it has been recruited evolutionarily to phosphorylate over 100 substrates, and can regulate numerous cellular functions (Beurel et al., 2015). *GSK3* phosphorylates HDAC3 and promotes its activity, including the neurotoxic activity of HDAC3 (Bardai and D'Mello, 2011). *GSK3* also phosphorylates HDAC6 to modify its activity and the link between *GSK3beta* and HDAC6 involved in neurodegenerative disorders (Chen et al., 2010).

Bilirubin is a reddish yellow pigment generated when the normal red blood cells break. Normal levels range from 0.2 to 1.2 mg/dL (Davis, 2020). In adults, indirect hyperbilirubinemia can be due to overproduction, impaired liver uptake or abnormalities of conjugation (Gondal and Aronsohn, 2016). For AML patients,[[Inline Image]][[Inline Image]] enasidenib is an inhibitor of mutant IDH2 proteins used to treat newly diagnosed mutant-IDH2 AML patients (Pollyea et al., 2019). The most common treatment-related adverse events are indirect hyperbilirubinemia (31%), nausea (23%), and fatigue (Steinwascher et al., 2015). Therefore, bilirubin is an important biomarker for monitoring adverse effect in AML patients who receive treatment.

Preclinical studies have discovered that Smac mimetics can directly cause cancer cell death, or make tumor cells become more sensitive to various cytotoxic treatment agents, including conventional chemotherapy, radiotherapy, or new drugs (Fulda, 2015). There is synergistic interaction of Smac mimetic and HDAC inhibitors in AML cell lines, and Smac mimetic and HDAC inhibitors can trigger necroptosis when caspase activation is blocked (Meng et al., 2016).

AKT.p308 and Src.p527 are phosphorylated signal transduction proteins. These two proteins are found to have lower expression in M0, M1, M2, but they have higher levels in the other AML French-American-British (FAB) types. The expression of those two proteins, together with 22 other proteins, can be used to define distinct signatures for each FAB type (Kornblau et al., 2009).

PTEN is a tumor suppressor protein. Promising anti-cancer agents, HDAC inhibitors, particularly trichostatin A (TSA), can promote PTEN membrane translocation. Meng et al. (2016) reveals that non-selective HDAC inhibitors, such as TSA or suberoylanilide hydroxamic acid (SAHA), induces *PTEN* membrane translocation through *PTEN* acetylation at K163 by inhibiting HDAC67. Similarly, treatment with an HDAC6 inhibitor alone promoted *PTEN* membrane translocation and correspondingly dephosphorylated AKT. The combination of celecoxib and an HDAC6 inhibitor synergistically increases *PTEN* membrane translocation and inactivated AKT (Zhang and Gan, 2017).

Our results show that multiple treatments improve efficiency of drugs for curing AML. This can be biologically explained. HDAC inhibitors have emerged as a potent and promising strategy for the treatment of leukemia via inducing differentiation and apoptosis in tumor cells (Jin et al., 2016). A phase II study with 37 refractory acute myelogenous leukemia (AML) patients shows only minimal activity of Vorinostat (HDACi), and Vorinostat fails to control the leukocyte count among most AML patients (Schaefer et al., 2009). A preclinical study reveals that the combination regimen of chidamide (a novel orally active HDAC inhibitor) and IDA could rapidly diminish the tumor burden in patients with refractory or relapsed AML (Li et al., 2017). A Phase II trial of Vorinostat with idarubicin (IDA) and Ara-C for patients with newly diagnosed AML or myelodysplastic syndrome reveals good activity with overall response rates of 85%. No excess toxicity due to Vorinostat is observed (Garcia-Manero et al., 2012). Taken together, HDACs in combination therapy with IDA or other chemotherapeutic drugs show encouraging clinical activity in different hematologic malignancies. This explains that the combination of HDAC and IDA is the best treatment.

Although MGANITE shows remarkable features in ITE for estimation and optimal treatment selection; the results in this paper are very preliminary. Training stable GANs is a challenging task. The training process is inherently unstable, resulting in the inaccurate estimation of ITEs. In this study, we ignore unobserved confounders, unmeasured variables that affect both patients' medical prescription and their outcome. Overlooking the presence of unobserved confounders may lead to biased results. The main purpose of this paper is to stimulate discussion about how to use AI as a powerful tool to improve the estimation of ITEs and optimal treatment selection. We hope that our results will greatly increase the confidence in using AI as a driving force to facilitate the development of precision oncology.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: data were obtained from Department of Biostatistics, The University of Texas MD Anderson Cancer Center. Requests to access these datasets should be directed to Data access need to request from Xuelin Huang, xlhuang@mdanderson.org.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MX and WL: conception and design. QG, MX, and SF: development of methodology. XH: acquisition of data. QG, SG, YL, and SF: analysis and interpretation of data. MX, QG, SG, SF, YL, WL, and XH: writing, review, and/or revision of the manuscript. All authors: contributed to the article and approved the submitted version.

FUNDING

WL was supported by the National Key R&D Program of China (Grant No. 2018YFC0116600), the National Natural Science Foundation of China (Grant No. 11925103), and by the STCSM (Grant Nos. 18DZ1201000, 19511101404).

ACKNOWLEDGMENTS

The authors thank Sara Barton for editing the manuscript and the Texas Advanced Computing Center for computation support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.585804/full#supplementary-material

REFERENCES

Alaa, A. M., and van der Schaar, M. (2017). "Bayesian inference of individualized treatment effects using multi-task gaussian processes," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 3427–3435.

Ali, M., and Aittokallio, T. (2019). Machine learning and feature selection for drug response prediction in precision oncology applications. *Biophys. Rev.* 11, 31–39. doi: 10.1007/s12551-018-0446-z

Bardai, F. H., and D'Mello, S. R. (2011). Selective toxicity by HDAC3 in neurons: regulation by Akt and GSK3beta. *J. Neurosci.* 31, 1746–1751. doi: 10.1523/JNEUROSCI.5704-10.2011

Beurel, E., Grieco, S. F., and Jope, R. S. (2015). Glycogen synthase kinase-3 (GSK3): regulation, actions, and diseases. *Pharmacol. Ther.* 148, 114–131. doi: 10.1016/j.pharmthera.2014.11.016

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Chen, S., Owens, G. C., Makarenkova, H., and Edelman, D. B. (2010). HDAC6 regulates mitochondrial transport in hippocampal neurons. *PLoS ONE* 5:e10848. doi: 10.1371/journal.pone.0010848

Chen, R., and Paschalidis, I. (2018). Learning optimal personalized treatment rules using robust regression informed K-NN. *arXiv Preprint* arXiv:1811.06083.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2008). Nonparametric tests for treatment effect heterogeneity. *Rev. Econ. Stat.* 90, 389–405. doi: 10.1162/rest.90.3.389

Davis, C. P. (2020). *Bilirubin and Bilirubin Blood Test*. MedicineNet. Available online at: https://www.medicinenet.com/bilirubin_and_bilirubin_blood_test/article.htm

Diamond, A., and Sekhon, J. S. (2013). Genetic matching for estimating causal effects: a general multivariate matching method for achieving balance in observational studies. *Rev. Econ. Stat.* 95, 932–945. doi: 10.1162/REST_a_00318

Ding, P., and Li, F. (2017). Causal inference: a missing data perspective. *Statist. Sci.* 33, 214–237.

Emmert-Streib, F., and Dehmer, M. (2019). High-dimensional LASSO-based computational regression models: regularization, shrinkage, and selection. *Mach. Learn. Knowl. Extract.* 1, 359–383. doi: 10.3390/make1010021

Fulda, S. (2015). Promises and challenges of smac mimetics as cancer therapeutics. *Clin. Cancer Res.* 21, 5030–5036. doi: 10.1158/1078-0432.CCR-15-0365

Garcia-Manero, G., Tambaro, F. P., Bekele, N. B., Yang, H., Ravandi, F., Jabbour, E., et al. (2012). Phase II trial of vorinostat with idarubicin and cytarabine for patients with newly diagnosed acute myelogenous leukemia or myelodysplastic syndrome. *J. Clin. Oncol.* 30, 2204–2210. doi: 10.1200/JCO.2011.38.3265

Garson, G. D. (1991). Interpreting neural network connection weights. *Artif. Intell. Expert.* 6, 47–51.

Gondal, B., and Aronsohn, A. (2016). A systematic approach to patients with Jaundice. *Semin. Intervent. Radiol.* 33, 253–258. doi: 10.1055/s-0036-1592331

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). "Generative adversarial networks," in *Proceedings of the 27th International Conference on Neural Information Processing Systems* (Montreal, QC). 2, 2672–2680.

Hansen, B. (2004). Full matching in an observational study of coaching for the SAT. *J. Am. Stat. Assoc.* 99, 609–618. doi: 10.1198/016214504000000647

Jassal, B., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., et al. (2020). The reactome pathway knowledgebase. *Nucl.Acids Res.* 48, 498–503. doi: 10.1093/nar/gkv1351

Jin, Y., Yao, Y., Chen, L., Zhu, X., Jin, B., Shen, Y., et al. (2016). Depletion of γ-catenin by histone deacetylase inhibition confers elimination of CML stem cells in combination with imatinib. *Theranostics* 6, 1947–1962. doi: 10.7150/thno.16139

Johansson, F. D., Shalit, U., and Sontag, D. (2016). "Learning representations for counterfactual inference," in *Proceedings of the 33rd International Conference on Machine Learning*. 48, 3020–3029.

Kennedy, E. H., Ma, Z., McHugh, M. D., and Small, D. S. (2017). Nonparametric methods for doubly robust estimation of continuous treatment effects. *J. R. Stat. Soc. Series B Stat. Methodol.* 79, 1229–1245. doi: 10.1111/rssb.12212

Kornblau, S. M., Tibes, R., Qiu, Y. H., Chen, W., Kantarjian, H. M., Andreeff, M., et al. (2009). Functional proteomic profiling of AML predicts response and survival. *Blood* 113, 154–164. doi: 10.1182/blood-2007-10-119438

Lengerich, B., Aragam, B., and Xing, E. P. (2019). Learning sample-specific models with low-rank personalized regression. *Advances in Neural Information Processing Systems* 3575–3585.

Li, Y., Wang, Y., Zhou, Y., Li, J., Chen, K., Zhang, L., et al. (2017). Cooperative effect of chidamide and chemotherapeutic drugs induce apoptosis by DNA damage accumulation and repair defects in acute myeloid leukemia stem and progenitor cells. *Clin. Epigenetics* 9, 83–83. doi: 10.1186/s13148-017-0377-8

Liu, J., Ma, Y., and Wang, L. (2018). An alternative robust estimator of average treatment effect in causal inference. *Biometrics* 74, 910–923. doi: 10.1111/biom.12859

Luo, W., and Zhu, Y. (2020). Matching using sufficient dimension reduction for causal inference. *J. Business Econ. Stat.* 38, 888–900. doi: 10.1080/07350015.2019.1609974

Makar, M., Swaminathan, A., and Kiciman, E. (2019). A distillation approach to data efficient individual treatment effect estimation. *Proc. AAAI Conf. Artif. Intellig.* 33, 4544–4551. doi: 10.1609/aaai.v33i01.33014544

Meng, Z., Jia, L. F., and Gan, Y. H. (2016). PTEN activation through K163 acetylation by inhibiting HDAC6 contributes to tumour inhibition. *Oncogene* 35, 2333–2344. doi: 10.1038/onc.2015.293

Mirza, M., and Osindero, S. (2014). Conditional generative adversarial nets. *arXiv*.

Pollyea, D. A., Tallman, M. S., de Botton, S., Kantarjian, H. M., Collins, R., Stein, A. S., et al. (2019). Enasidenib, an inhibitor of mutant IDH2 proteins, induces durable remissions in older patients with newly diagnosed acute myeloid leukemia. *Leukemia* 33, 2575–2584. doi: 10.1038/s41375-019-0472-2

Ray, K., and Szabo, B. (2019). Debiased Bayesian inference for average treatment effects. *Advances in Neural Information Processing Systems*, 11952–11962.

Rosenbaum, P. R., and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 41–55. doi: 10.1093/biomet/70.1.41

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* 66, 688. doi: 10.1037/h0037350

Schaefer, E. W., Loaiza-Bonilla, A., Juckett, M., DiPersio, J. F., Roy, V., Slack, J., et al. (2009). A phase 2 study of vorinostat in acute myeloid leukemia. *Haematologica* 94, 1375–1382. doi: 10.3324/haematol.2009.009217

Seyhan, A. A., and Carini, C. (2019). Are innovation and new technologies in precision medicine paving a new era in patients centric care? *J. Transl. Med.* 17:114. doi: 10.1186/s12967-019-1864-9

Shalit, U., Johansson, F. D., and Sontag, D. (2016). "Estimating individual treatment effect: generalization bounds and algorithms," in *Proceedings of the 34th International Conference on Machine Learning.* 70, 3076–3085.

Shin, S. H., Bode, A. M., and Dong, Z. (2017). Addressing the challenges of applying precision oncology. *NPJ Precision Oncol.* 1:28. doi: 10.1038/s41698-017-0032-z

Siu, C. (2017). *Day33: Garson's Algorithm.* Available online at: https://csiu.github.io/blog/update/2017/03/29/day33.html (accessed March 29, 2017).

Steinwascher, S., Nugues, A. L., Schoeneberger, H., and Fulda, S. (2015). Identification of a novel synergistic induction of cell death by Smac mimetic and HDAC inhibitors in acute myeloid leukemia cells. *Cancer Lett.* 366, 32–43. doi: 10.1016/j.canlet.2015.05.020

Subbiah, V., and Kurzrock, R. (2018). Challenging standard-of-care paradigms in the precision oncology era. *Trends Cancer* 4, 101–109. doi: 10.1016/j.trecan.2017.12.004

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucl. Acids Res.* 47, 607–613. doi: 10.1093/nar/gky1131

Wager, S., and Athey, S. (2015). Estimation and inference of heterogeneous treatment effects using random forests. *Stat Med.* 37, 3309–3324.

Yoon, J., Jordon, J., and van der Schaar, M. (2018a). GAIN: missing data imputation using generative adversarial nets. *Proceedings of the 35th International Conference on Machine Learning.* 80, 5689–5698.

Yoon, J., Jordon, J., and van der Schaar, M. (2018b). *GANITE: Estimation of Individualized Treatment Effects Using Generative Adversarial Nets*. ICLR.

Zhang, G., and Gan, Y. H. (2017). Synergistic antitumor effects of the combined treatment with an HDAC6 inhibitor and a COX-2 inhibitor through activation of PTEN. *Oncol. Rep.* 38, 2657–2666. doi: 10.3892/or.2017.5981

Zhang, Z., Beck, M. W., Winkler, D. A., Huang, B., Sibanda, W., et al. (2018). Opening the black box of neural networks: methods for interpreting neural network models in clinical applications. *Ann. Transl. Med.* 6:216. doi: 10.21037/atm.2018.05.32

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 reviewer SJ declared a shared affiliation, with no collaboration, with the authors XH and SF to the handling editor at the time of review.

frontiers in Genetics

ORIGINAL RESEARCH
published: 07 January 2021
doi: 10.3389/fgene.2020.587378

Estimation of Heterogeneous Restricted Mean Survival Time Using Random Forest

*Mingyang Liu and Hongzhe Li**

Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States

Estimation and prediction of heterogeneous restricted mean survival time (hRMST) is of great clinical importance, which can provide an easily interpretable and clinically meaningful summary of the survival function in the presence of censoring and individual covariates. The existing methods for the modeling of hRMST rely on proportional hazards or other parametric assumptions on the survival distribution. In this paper, we propose a random forest based estimation of hRMST for right-censored survival data with covariates and prove a central limit theorem for the resulting estimator. In addition, we present a computationally efficient construction for the confidence interval of hRMST. Our simulations show that the resulting confidence intervals have the correct coverage probability of the hRMST, and the random forest based estimate of hRMST has smaller prediction errors than the parametric models when the models are mis-specified. We apply the method to the ovarian cancer data set from The Cancer Genome Atlas (TCGA) project to predict hRMST and show an improved prediction performance over the existing methods. A software implementation, srf using R and C++, is available at https://github.com/lmy1019/SRF.

Keywords: estimating equation, high dimensional data, non-parametric survival estimation, regression forest, inference

Edited by:
Chao Xu,
University of Oklahoma Health Sciences Center, United States

Reviewed by:
Huaizhen Qin,
University of Florida, United States
Leonidas Bantis,
University of Kansas Medical Center, United States

****Correspondence:***
Hongzhe Li
hongzhe@upenn.edu

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *25 July 2020*
Accepted: *07 December 2020*
Published: *07 January 2021*

Citation:
Liu M and Li H (2021) Estimation of Heterogeneous Restricted Mean Survival Time Using Random Forest. Front. Genet. 11:587378. doi: 10.3389/fgene.2020.587378

1. INTRODUCTION

In epidemiological and biomedical studies, time to an event or survival time T is often the primary outcome of interest. Important quantities related to survival time include hazard rate (HR), t-year survival probability, and the mean survival time. Among these, HR is one of the most commonly used quantity due to its strong connection to the proportional hazards regression model or Cox model. Cox model is a very popular regression model for censored survival data due to its computational feasibility and theoretical properties (Cox, 1972, 1975; Andersen and Gill, 1982; Gill and Gill, 1984; Huang et al., 2013; Fang et al., 2017). However, when there is a departure from the proportional hazards assumption, the connection between HR and survival function is lost and it is difficult to interpret HR (Wang and Schaubel, 2018). The t-year survival probability is the probability of survival time greater than a pre-specified time t. It is not suitable for summarizing the global profile of T over the duration of a study (Tian et al., 2014). In contrast, mean survival time is an alternative quantity since it takes the whole distribution of T into account. However, the mean of T may not always be estimable in the presence of censoring. For example, let C denotes the

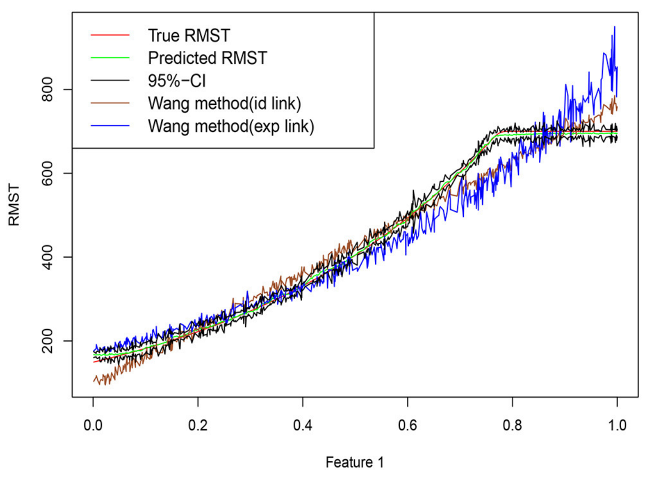


FIGURE 1 | Training data are simulated from Equation (2), with $n = 600$ training points, dimension $p = 20$ and errors $\epsilon \sim N(0, 10^2)$. Random forests are trained based using R package grf. Truth is shown as red curve, with green curve corresponding to the random forest predictions, and upper and lower bounds of the point-wise confidence intervals connected in the black lines. Brown curve and blue curve are based on the approaches of Wang and Schaubel (2018) with Identity and Exp link functions.

censoring time, and $C_{\max} = \inf_c\{P(C \leq c) = 1\}$ be the upper limit of the censoring distribution,

$$E_T[T] = E_T[T|T \leq C_{\max}]P(T \leq C_{\max}) + E_T[T|T > C_{\max}]P(T > C_{\max})$$

If the survival time T satisfies $P(T > C_{\max}) > 0$, then we cannot estimate $E_T[T]$, since we never observe any event after $C_{\max}$.

The restricted mean survival time (RMST) (Royston and Parmar, 2013) summarizes the survival process and provides an attractive alternative to the proportional hazards regression model (Tian et al., 2014). The restricted survival time of T up to a fixed point L is defined as $T \wedge L$, and the restricted mean survival time is defined as the expectation of the restricted survival time. Denote $\mu^L(x) = E[T \wedge L|X = x]$ be the heterogeneous RMST with covariates $X = x$. It can be written as the area under the survival curve on $[0, L]$.

$$\mu^L(x) = \int_0^\infty \left(\int_0^\infty 1_{u<t}1_{u<L}du\right) f_T(t|X = x)dt = \int_0^L S(u|X = x)du. \quad (1)$$

If L is chosen to be less than $C_{\max}$, hRMST is estimable since $P(T \wedge L > C_{\max}) = 0$. RMST also plays a role in the context of inverse probability censoring weighting (IPCW). A key assumption for applying IPCW is $P(T < C_{\max}) = 1$, making $1/(1 - G(T))$ well-defined, where $G(T) = P(C \leq T|T)$. If we set L properly such that $P(T \wedge L < C_{\max}) = 1$, then $G(T \wedge C \wedge L|X) < 1$ and the IPCW is well-defined under the restricted survival time context.

There are two main approaches for hRMST regression. One approach is to estimate hRMST indirectly through hazard regression (Zucker, 1998; Chen and Tsiatis, 2001; Zhang and Schaubel, 2011). This approach starts by estimating the regression parameters and the baseline hazard from a Cox model, calculating the cumulative baseline hazard, transforming it to obtain the survival function and, finally, obtaining the hRMST through Equation (1). Such an indirect hRMST estimation is inconvenient and computationally cumbersome for obtaining a point estimate and its corresponding asymptotic standard error. An alternative approach is to model hRMST with the baseline covariates X directly via some parametric assumptions, eg. $g[\mu^L(X_i)] = \beta_0'X_i$, where g is a strictly monotone link function with a continuous derivative within an open neighborhood (Tian et al., 2014; Wang and Schaubel, 2018). A major weakness of this approach, however, is their inability to choose a proper link function, which may lead to the model misspecification. As an example, we simulate $x_1, \ldots, x_n$ independently from the uniform distribution on $[0, 1]^{20}$ with a survival time model

$$T = \exp(2X_1 + 5) + 1 + \epsilon, \quad \epsilon \sim N(0, 10^2), \quad (2)$$

where we assume that the censoring time C and the restricted time L satisfy $P(C \leq T \wedge L) = 33\%$ and $P(L \leq T \wedge C) = 11\%$. Our goal is to estimate $\mu^L(x)$. **Figure 1** shows a set of predictions on an artificially generated data set from Equation (2). Compared with other methods, the random forest is able to estimate the target function closely, especially when $\mu^L(x)$ approaches L.

For the continuous outcomes without censoring, random forest (Breiman, 2001, 2004) is a popular method of non-parametric regression that has shown effectiveness in many applications (Svetnik et al., 2003; Díaz-Uriarte and Alvarez de Andrés, 2006; Cutler et al., 2007). It is invariant under scaling and various other transformations of feature values, robust to inclusion of irrelevant features (Hastie et al., 2001), and versatile enough to be applied to large-scale problems (Biau and Scornet, 2016). Besides strong empirical results, theoretical results such as consistency (Meinshausen, 2006; Biau et al., 2008; Biau, 2012; Denil et al., 2014) and asymptotic normality (Wager and Athey, 2015; Mentch and Hooker, 2016; Athey et al., 2018; Friedberg et al., 2018) have also been obtained for regression models without censoring. Extending random forest to censored survival data has been proposed in several recent papers (Ishwaran et al., 2008; Steingrimsson et al., 2019), focusing on implementations and algorithms. However, there has been little theoretical work in statistical inference of such random survival forest. Ishwaran and Kogalur (2011) proved the consistency of the random survival forest by showing that the forest ensemble survival function converges uniformly to the true population survival function.

Instead of focusing on predicting the survival function or the survival probability as the algorithms implemented by Ishwaran et al. (2008) and Steingrimsson et al. (2019), we develop in this paper a random forest framework to model the hRMST directly given the baseline covariates in the presence of possibly covariate-dependent censoring. This approach provides a non-parametric estimation of hRMST adjusting for covariates. Due to the complex relationship between the survival time and the covariates, it is desirable to have more flexible methods to estimate the hRMST than the approaches that a certain link function has to be assumed. Our construction of random forest is based on the estimated IPCW. We show that the resulting survival random forest estimates of hRMST has the asymptotic normality property that can be used to obtain the point-wise confidence interval with theoretical guarantees. To the best of our knowledge, it is the first asymptotic normality result for the predictions in the context of censored survival data using random forest.

The remainder of the paper is organized as follows. In section 2, we describe the proposed random forest estimator. Asymptotic properties are given in section 3. In section 4, we conduct simulation studies to evaluate the accuracy of the proposed method in the finite sample settings. In section 5, we apply our method to an ovarian cancer data set of The Cancer Genome Atlas (TCGA) project (http://cancergenome.nih.gov/abouttcga) to evaluate the predictions of the hRMST for ovarian cancer patients using their acylcarnitine measurements and clinical variables. We conclude this chapter with a brief discussion in section 6.

2. RANDOM FOREST FOR ESTIMATING THE hRMST

We begin with some notation. Let X_i be the baseline covariates for subject i from a cohort of sample size n and T_i be the survival time for subject i. Let C_i be the censoring time, which is independent of T_i conditional on the baseline covariates X_i. The observation time for subject i is $Z_i = T_i \wedge C_i$, where $a \wedge b = \min\{a, b\}$. The indicator for censoring is denoted by $\delta_i = 1_{\{T_i \leq C_i\}}$. Our observed *i.i.d.* data are given as $\{(X_i, Z_i, \delta_i) : i = 1, \ldots, n\}$.

Let L be a pre-specified time point of interest, before the maximum follow-up time $\tau = \max\{Z_i : i = 1, \ldots, n\}$. As in Wang and Schaubel (2018), L is normally chosen as a time point of clinical relevance or, at least, of particular interest to the investigators, respecting the bound at the maximum follow-up time. Denote the restricted observation time as $Z_i^L = Z_i \wedge L$ and its corresponding indicator $\delta_i^L = 1_{\{T_i \wedge L \leq C_i\}}$. Our goal is to estimate covariate-adjusted RMST or hRMST $\mu^L(x) = E(Z^L|X = x)$ and to construct its confidence interval.

2.1. Forest-Based Local Estimating Equation for hRMST

Given the observed data $\{(X_i, \delta_i, Z_i)\}_{i=1}^n$, and a restriction threshold L, we first present a random forest method to estimate $\mu^L(x)$. The idea of the approach is to solve a weighted estimating equation for $\mu^L(x)$, where the estimating equation functions of the observations whose covariates closer to x will have larger weights. Specifically, let $w_i = \delta_i^L/(1 - G(Z_i^L|X_i))$ be the IPCW of the ith data point under the true censoring distribution $G(\cdot|X_i)$. The (infeasible) estimating equation function $w_i(Z_i^L - \mu^L(x))$ of $X_i = x$ satisfies $E[w_i(Z_i^L - \mu^L(x))|X_i = x] = E[T_i \wedge L|X_i = x] - \mu^L(x) = 0$. If the local weights $\{\alpha_i(x)\}_{i=1}^n$ are also known, the solution to the empirical estimating equation for $\mu^L(x)$

$$\sum_{i=1}^{n} \alpha_i(x) w_i (Z_i^L - \mu) = 0 \tag{3}$$

is given as

$$\frac{\sum_{i=1}^n \alpha_i(x) w_i Z_i^L}{\sum_{i=1}^n \alpha_i(x) w_i},$$

which provides a good candidate of estimator for $\mu^L(x)$. However we do not know the censoring distribution G and the local weights $\{\alpha_i(x)\}_{i=1}^n$, which need to be estimated from the data. We assume censoring distribution G follows a Cox model, a natural choice for modeling censoring times in the context of IPCW. Let

$$\hat{w}_i = \frac{\delta_i^L}{1 - \hat{G}(Z_i^L|X_i)}$$

be the estimated IPCW for ith observation with $\hat{G}(\cdot|X_i)$ derived from the data through Cox model. We define the estimating equation function for ith observation with its corresponding estimated IPCW as

$$\psi_{\mu^L(x)}(X_i, Z_i^L, \delta_i^L) = \hat{w}_i (Z_i^L - \mu_i^L(x)).$$

Our approach to derive the local weights $\{\alpha_i(x)\}_{i=1}^n$ is through the random forest, which is an ensemble of survival trees constructed by Algorithm 1.

Algorithm 1: Survival tree

```
SurvivalTree (set of observations $J$, domain $X$);
IPCW $\leftarrow$CoxModel($J$);
Root $P_0$ $\leftarrow$CreateNode($J, X$);
Queue $Q$ $\rightarrow$InitializeQueue($P_0$);
while $Q$ is NotNull do
    node $P \leftarrow Pop(Q)$;
    Solve $\hat{\mu}_P^L = \underset{\mu}{\text{argmin}} \,|\sum_{X_i \in P} \psi_\mu(X_i, Z_i^L, \delta_i^L)|$;
    Set $\rho_i = \frac{\hat{w}_i(Z_i^L - \hat{\mu}_P^L)}{(\sum_{X_i \in P} \hat{w}_i)/|\{i : X_i \in P\}|}$;
    Split $P$ by maximizing
    $\tilde{\Delta}(C_1, C_2) = \sum_{j=1}^{2} \frac{1}{|\{i : X_i \in C_j\}|} \left( \sum_{i : X_i \in C_j} \rho_i \right)^2$;
    if split succeeds then
        AddQueue($C_1$);
        AddQueue($C_2$);
    end
end
```

It can be shown that ρ_i is the influence function of the ith observation for $\hat{\mu}_P^L$. Let F_n be the empirical distribution of the observations in node P, and let $F_{n,i} = (1 - \epsilon)F_n + \epsilon \nu_i$, with ν_i be the Dirac delta function at ith observation. Set $\hat{\mu}_{P,i}^L = \hat{\mu}_P^L + \Delta_i$, where $\hat{\mu}_{P,i}^L = \underset{\mu}{\text{argmin}} \,|\int \psi_\mu(X, Z^L, \delta^L) dF_{n,i}|$. By Taylor expansion,

$$0 = \int \psi_{\hat{\mu}_{P,i}^L}(X, Z^L, \delta^L) dF_{n,i}$$

$$= \int [\psi_{\hat{\mu}_P^L}(X, Z^L, \delta^L) + \psi'_{\mu^*}(X, Z^L, \delta^L)\Delta_i] dF_{n,i},$$

where μ^* is a value between $\hat{\mu}_P^L$ and $\hat{\mu}_{P,i}^L$. The above equation implies

$$\Delta_i = -\frac{\epsilon\psi_{\hat{\mu}_P^L}(X_i, Z_i^L, \delta_i^L)}{\int \psi'_{\mu^*}(X, Z^L, \delta^L) dF_{n,i}},$$

and therefore the influence function of ith observation for $\hat{\mu}_P^L$ is

$$\lim_{\epsilon \to 0} \Delta_i/\epsilon = -\frac{\psi_{\hat{\mu}_P^L}(X_i, Z_i^L, \delta_i^L)}{\int \psi'_{\hat{\mu}_P^L}(X, Z^L, \delta^L) dF_n} = \frac{\hat{w}_i(Z_i^L - \hat{\mu}_P^L)}{\sum_{i \in P} \frac{\hat{w}_i}{|\{i : X_i \in P\}|}} = \rho_i.$$

Athey et al. (2018) shows that maximizing the splitting criterion $\tilde{\Delta}(C_1, C_2)$ is approximately equivalent to minimizing the weighted mean squared error $err(C_1, C_2) = \sum_{i=1,2} P(X \in C_i | X \in P) E[(\hat{\mu}_{C_i}^L - \mu^L(X))^2 | X \in C_i]$.

In order to achieve consistency and asymptotic normality, we split the tree and make predictions in an honest way as introduced in Wager and Athey (2015). Specifically, each tree in an honest forest is grown using two non-overlapping subsamples of the training data. For the bth tree, given I_b and J_b, we first choose the tree structure T_b using only the data in J_b, and write $x \leftrightarrow_b x'$ as the boolean indicator for whether the points x and x' fall into the same leaf of T_b. In a second step, we define the set of neighbors of x as $L_b(x) = \{i \in I_b : x \leftrightarrow_b x_i\}$. The weights of point x from a survival forest with B trees can be written as

$$\alpha_i(x) = \frac{1}{B} \sum_{b=1}^{B} \frac{1_{\{X_i \in L_b(x)\}}}{|L_b(x)|}.$$

The empirical locally weighted estimating equation for $\hat{\mu}^L(x)$ is then defined as

$$\sum_{i=1}^{n} \alpha_i(x) \psi_\mu(X_i, Z_i^L, \delta_i^L) = 0, \tag{4}$$

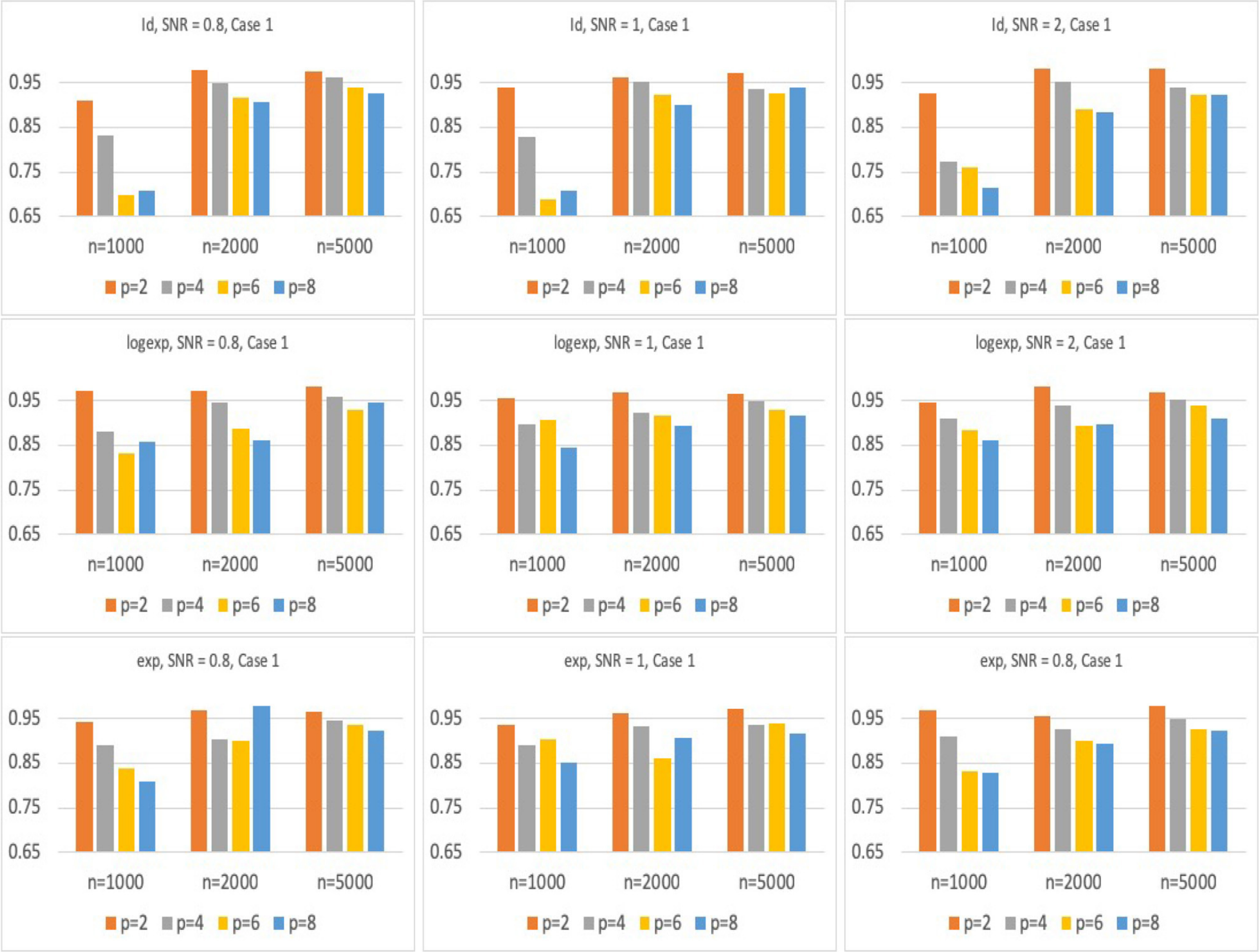


FIGURE 2 | Simulation results of the coverage probability for Model 1 with three different link functions, sample size of $n = 1,000, 2,000, 5,000$, and $p = 2, 4, 6, 8$. For each case, prediction coverage probability is calculated over the samples in the testing data set.

and the random forest estimator for the hRMST is the solution of Equation (4), which is

$$\hat{\mu}^L(x) = \sum_{i=1}^{n} \frac{\alpha_i(x)\hat{w}_i Z_i^L}{\sum_{i=1}^{n} \alpha_i(x)\hat{w}_i}.$$

We emphasize the difference between the IPCW used in building the survival trees and IPCW used to derive $\hat{\mu}^L(x)$. The IPCW used in building survival trees is estimated only by the data points from J_b so that the resulting survival forest is honest. The IPCW used to derive $\hat{\mu}^L(x)$ is estimated from all data points.

3. ASYMPTOTIC DISTRIBUTION OF $\hat{\mu}^L(X)$

3.1. Asymptotic Normality

We derive a central limit theorem for survival forest estimate of hRMST. We first give three common assumptions that required for the most of the theoretical analysis of random forests.

Assumption 1. *$\mu^L(x)$ is Lipschitz continuous w.r.t x.*

Assumption 2. *There exists a restricted time threshold L, such that $P(C > t \wedge L|X = x) \geq \epsilon_L > 0$ for any x, t.*

Assumption 3. *$Var(T \wedge L|X = x) > 0$ for any x.*

As mentioned in the previous section, we model the conditional survival function of censoring distribution G given baseline covariates. Because of its flexibility and popularity in practice, we adopt the proportional hazards model for hazard function of censoring distribution.

Assumption 4. *The hazard function of censoring distribution follows $\lambda_i^C(t) = \lambda_0^C(t)\exp(X_i'\beta_C)$*

We make additional regularity assumptions that are widely used in analysis of estimates from the proportional hazards models. These assumptions are needed in order to quantify the difference between the estimated IPCW and true IPCW.

Assumption 5. $||X||_\infty < M_X < \infty$

Assumption 6. $\lambda_0^C(t) \leq \lambda_0^C < \infty$ *for all t.*

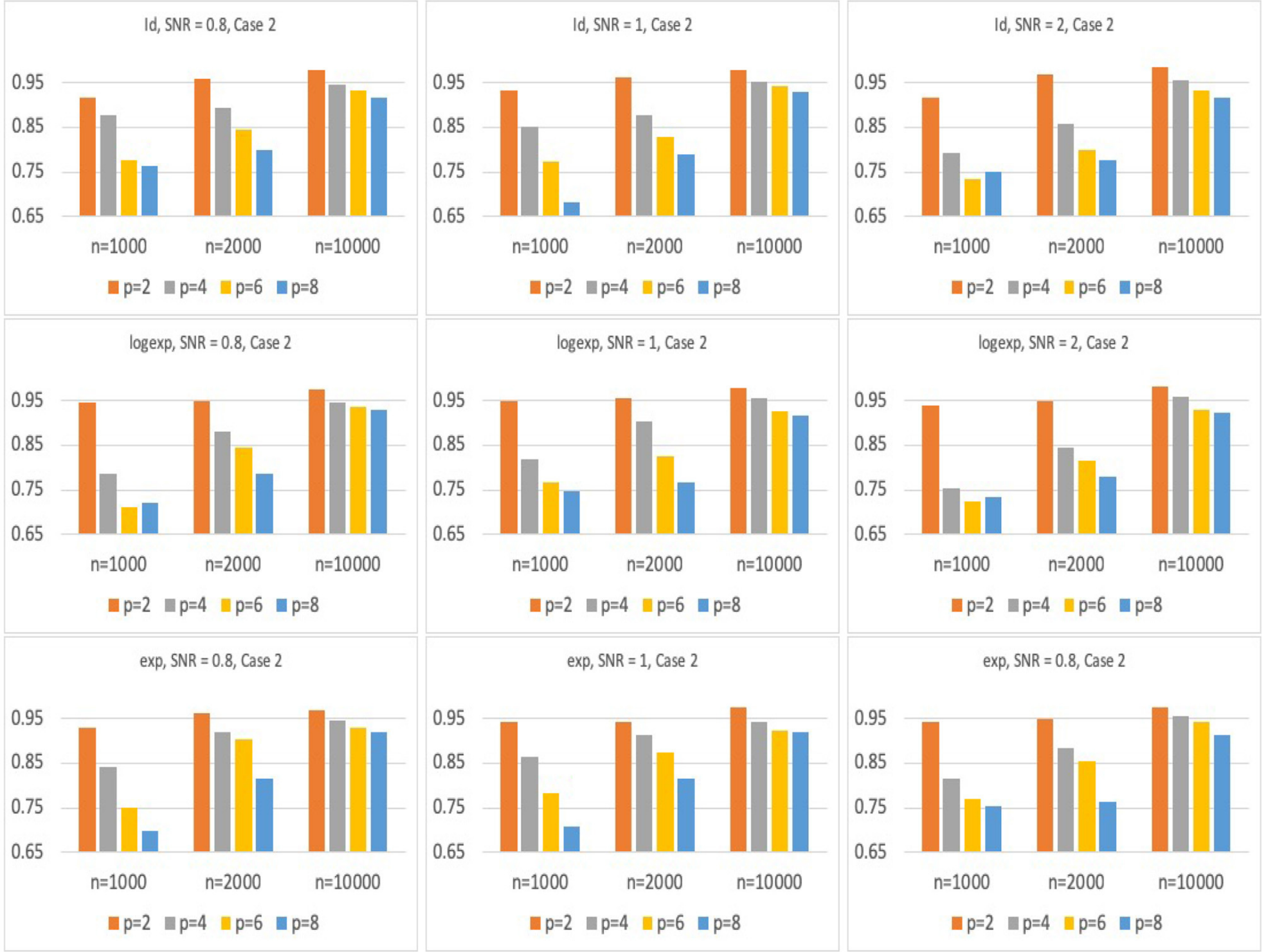


FIGURE 3 | Simulation results of coverage probability for Model 2 with three different link functions, sample size of $n = 1,000, 2,000, 10,000$, and $p = 2, 4, 6, 8$. For each case, prediction coverage probability is calculated over the samples in the testing data set.

Assumption 7. $\Omega_C(\beta) = E\left[\int_0^\tau \frac{r^{(2)}(t,\beta)}{r^{(0)}(t,\beta)} - \bar{x}(t,\beta)^{\otimes 2} dN_i^C(t)\right]$ *is positive definite, where* $R_i(t) = 1(Z_i \geq t)$, $r^{(k)}(t,\beta) = E[\exp(\beta' X_i)R_i(t)X_i^{\otimes k}]$, $\bar{x}(t,\beta) = \frac{r^{(1)}(t,\beta)}{r^{(0)}(t,\beta)}$, $N_i^C(t) = 1_{Z_i \leq t, \delta_i = 0}$.

Assumption 8. $P(R_i(t) = 1|X_i = x) \geq r > 0$ *for some positive constant and for any* t, x. *This assumption implies that*

$$r^{(0)}(t,\beta) = E[\exp(\beta' X_i)R_i(t)] = E[\exp(\beta' X_i)E[R_i(t)|X_i]] \geq r > 0.$$

Following Wager and Athey (2015) and Athey et al. (2018), we assume that all trees are symmetric, in that their output is invariant to permuting the indices of Estimation-Part in training examples (see Corollary 6 of Wager and Athey (2015) for more details about this symmetry). They also require balanced splits in the sense that every split puts at least a fraction ω of the

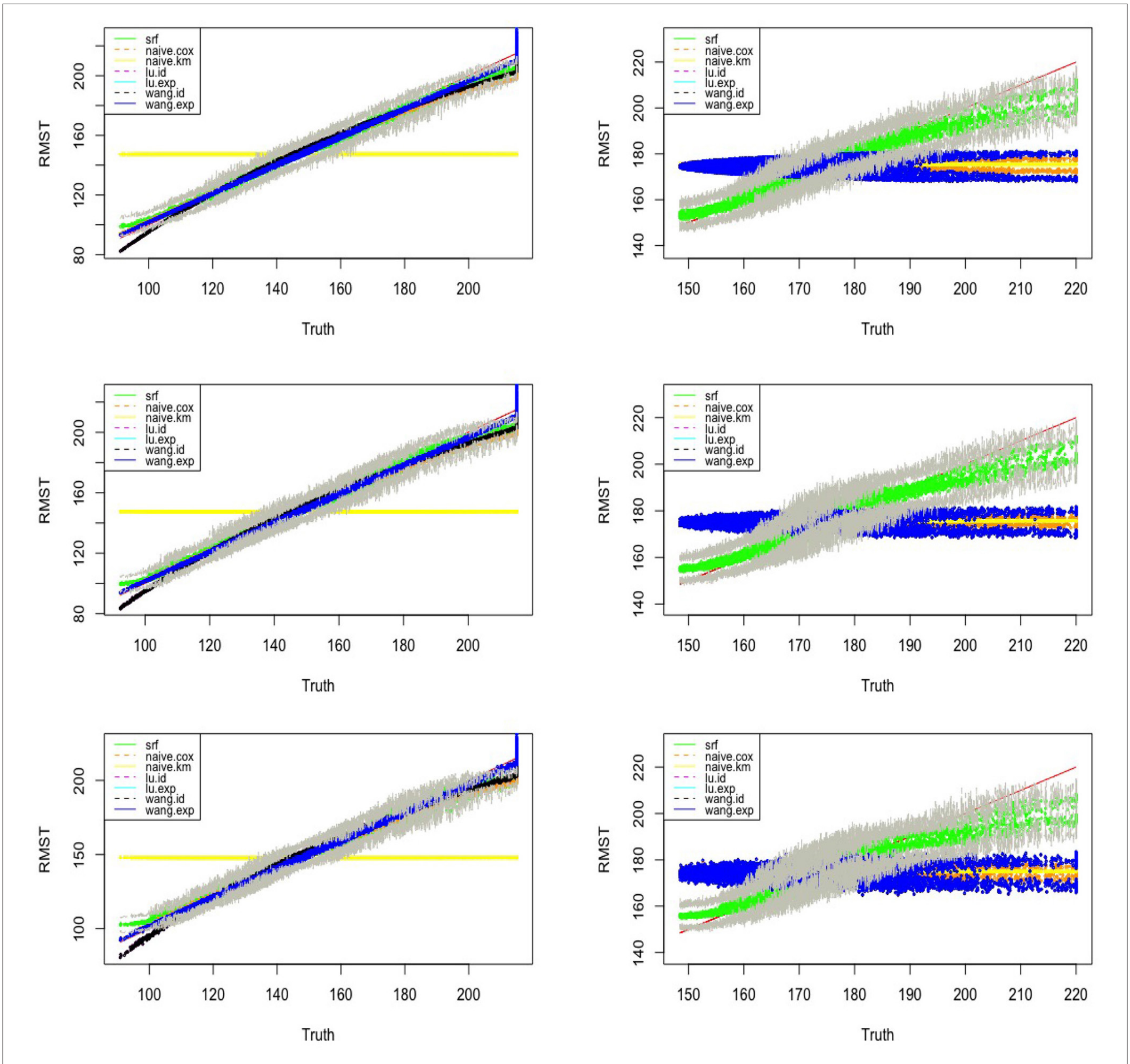


FIGURE 4 | Estimated vs. the true RMST for Model 1 **(left)** and Model 2 **(right)** with exponential link function and the number of covariates $p = 5, 10, 20$ **(top–bottom)**. SRF, proposed random forest-bases estimator, and upper and lower bounds of the point-wise confidence intervals of the proposed random forest estimator are connected in the gray lines; Naive.km, estimate based on Kaplan–Meier estimator without adjusting for the covariates; Naive.Cox, Cox regression based estimator; Lu.id, method of Tian et al. (2014) with identity link; Lu.exp, method of Tian et al. (2014) with exponential link; Wang.id, method of Wang and Schaubel (2018) with identity link; Wang:exp, method of Wang and Schaubel (2018) with exponential link.

observations in the parent node into each child, for some $\omega > 0$. Finally, the trees are randomized in such a way that, at every split, the probability that the tree splits on the jth feature is bounded from below by some $\pi > 0$. The forest is honest and built via subsampling with subsample size s satisfying $s/n \rightarrow 0$ and $s \rightarrow \infty$.

Under the assumptions listed above, we have the following asymptotic distribution result for the random forest-based estimate of the hRMST.

Theorem 1. *Under Assumptions 1, 2, 3, 4, 5, 6, 7, 8, for each fixed test point x, there is a sequence* $\sigma_n^2(x) = Var(\hat{\mu}^L(x)) \rightarrow 0$,

$$\frac{\hat{\mu}^L(x) - \mu^L(x)}{\sigma_n(x)} \rightarrow_d N(0,1)$$

if subsampling size

$$\beta_{\min} = 1 - \left(1 + \frac{\pi^{-1}\left(\log(\omega^{-1})\right)}{\log\left((1-\omega)^{-1}\right)}\right)^{-1},$$

where $\omega > 0$ *is the low-bound fraction for observations in the parent node into each child, and* $\pi > 0$ *is the lower-bound of the probability that the tree splits on any features.*

We give a consistent estimate of $\sigma_n^2(x)$ based on half-sampling (Efron, 1980) and the method of Sexton and Laake (2009).

3.2. Estimation of the Variance

Following Athey et al. (2018), we use the random forest delta method to develop a variance estimate of the survival forest prediction $\hat{\mu}^L(x)$. Athey et al. (2018) provides a consistent estimate of $\sigma_n^2(x)$ using $s_n^2(x)$, where $s_n^2(x) = (V(x)^{-1})H_n(x)(V(x)^{-1})'$ with

$$H_n(x) = Var[\sum_{i=1}^{n} \alpha_i(x)\psi_{\mu^L(x)}(X_i, Z_i^L, \delta_i^L)]$$

$$V(x) = \frac{\partial}{\partial(\mu^L)} E[\psi_{\mu^L}(X, Z^L, \delta^L)|X = x]|_{\mu^L = \mu^L(x)}$$

In our context, $V(x) = -1$, then simply we have $s_n^2(x) = H_n(x)$.

A consistent estimator for $H_n(x)$ can be obtained using half-sampling estimator (Efron, 1980; Athey et al., 2018). Let $\Psi_{\mathcal{H}}$ be the average of the empirical estimating equation functions averaged over the trees that only use the data from the half-sample $\mathcal{H}$, denoted by $S_{\mathcal{H}}$,

$$\Psi_{\mathcal{H}}(x) = \frac{1}{|S_{\mathcal{H}}|} \sum_{b \in S_{\mathcal{H}}} \frac{\sum_{i=1}^{n} 1_{X_i \in L_b(x)} \psi_{\hat{\mu}^L(x)}(X_i, Z_i^L, \delta_i^L)}{\sum_{i=1}^{n} 1_{X_i \in L_b(x)}},$$

where $L_b(x)$ contains neighbors of x in the bth tree. An ideal half-sampling estimator is then defined as

$$\hat{H}_n^{HS}(x) = \binom{n}{n/2}^{-1} \sum_{\mathcal{H}:|\mathcal{H}|=n/2} (E_\Theta[\Psi_{\mathcal{H}}(x)] - E_\Theta \bar{\Psi}(x))^2$$

TABLE 1 | Comparison of Mean-Absolute-Error (MAE) and Rooted-Mean-Squared-Error (RMSE) for Model 1 with different link functions.

p	SRF	Naive.Cox	Naive.km	Lu.id	Lu.exp	Wang.id	Wang.exp
			Model 1: identity link, $n = 3,000$, SNR = 0.3				
5	0.1359	0.1371	0.2067	**0.1341**	0.1346	**0.1341**	0.1346
	0.1699	0.1695	0.2466	0.1687	0.1691	**0.1686**	0.1691
10	0.1396	0.1394	0.2108	**0.1371**	0.1377	**0.1371**	0.1376
	0.1721	0.1710	0.2497	0.1710	0.1715	**0.1709**	0.1714
20	0.1373	0.1372	0.2064	**0.1342**	0.1348	**0.1342**	0.1347
	0.1703	0.1693	0.2464	0.1686	0.1691	**0.1685**	0.1690
			Model 1: log-exp link, $n = 3,000$, SNR = 0.3				
5	0.1347	0.1359	0.2048	**0.1330**	0.1335	**0.1330**	0.1335
	0.1684	0.1680	0.2441	0.1673	0.1677	**0.1672**	0.1677
10	0.1384	0.1382	0.2088	**0.1359**	0.1366	**0.1359**	0.1365
	0.1706	**0.1695**	0.2472	**0.1695**	0.1701	**0.1695**	0.1699
20	0.1361	0.1360	0.2044	0.1331	0.1337	**0.1330**	0.1336
	0.1689	0.1679	0.2439	0.1672	0.1678	**0.1671**	0.1676
			Model 1: exp link, $n = 3,000$, SNR = 0.3				
5	24.724	25.398	33.688	24.496	24.723	**24.436**	24.709
	30.827	30.860	39.296	30.608	30.773	**30.577**	30.749
10	25.254	25.681	34.208	24.843	25.162	**24.812**	25.149
	31.085	31.052	39.621	30.869	31.076	**30.850**	31.048
20	24.878	25.260	33.587	24.390	24.679	**24.325**	24.651
	30.744	30.695	39.181	30.479	30.689	**30.438**	30.646

The number of covariates $p = 5, 10, 20$, for each p, the first row is MAE, the second row is RMSE. SRF, proposed random forest-bases estimator; Naive.km, estimate based on Kaplan–Meier estimator without adjusting for the covariates; Naive.Cox, Cox regression based estimator; Lu.id, method of Tian et al. (2014) with identity link; Lu.exp, method of Tian et al. (2014) with exponential link; Wang.id, method of Wang and Schaubel (2018) with identity link; Wang:exp, method of Wang and Schaubel (2018) with exponential link.

$$\bar{\Psi}(x) = \binom{n}{n/2}^{-1} \sum_{\mathcal{H}: |\mathcal{H}|=n/2} \Psi_{\mathcal{H}}(x)$$

where Θ is the randomness in building honest tree, including splitting data into random halves and randomness in selecting variables to split. $\hat{H}_n^{HS}(x)$ is similar to classic bootstrap estimator for the standard error, except that the sampling distribution for $\hat{H}_n^{HS}(x)$ is the half sampling distribution instead of the bootstrap sampling. Denote E_{ss} and Var_{ss} as the expectation and variance under the half sampling distribution, then $\hat{H}_n^{HS}(x) = Var_{ss}[E_{\Theta}[\Psi_{\mathcal{H}}(x)]]$.

Since carrying out the full half-sampling computation and expectation with respect to Θ are impractical, Sexton and Laake (2009) pointed out that $\hat{H}_n^{HS}(x)$ can be efficiently approximated by the following law of total variance:

$$\begin{aligned} \hat{H}_n^{HS}(x) &= Var_{ss}\left[E_{\Theta}[\frac{1}{M}\sum_{m=1}^{M}\Psi_{\mathcal{H},\Theta_m}(x)]\right] \\ &= Var_{ss}\left[\frac{1}{M}\sum_{m=1}^{M}\Psi_{\mathcal{H},\Theta_m}(x)\right] \\ &\quad -E_{ss}\left[Var_{\Theta}[\frac{1}{M}\sum_{m=1}^{M}\Psi_{\mathcal{H},\Theta_m}(x)]\right] \end{aligned} \tag{5}$$

which leads to a Monte Carlo approximation of $\hat{H}_n^{HS}(x)$ by

$$\begin{aligned} \hat{\sigma}_n^2(x) &= \widehat{Var}_{ss}\left[\frac{1}{M}\sum_{m=1}^{M}\Psi_{\mathcal{H},\Theta_m}(x)\right] \\ &\quad - \hat{E}_{ss}\left[\widehat{Var}_{\Theta}[\frac{1}{M}\sum_{m=1}^{M}\Psi_{\mathcal{H},\Theta_m}(x)]\right]. \end{aligned} \tag{6}$$

In order to approximate random forest randomness quantity $\widehat{Var}_{\Theta}$ and sampling randomness quantities $\widehat{Var}_{ss}, \hat{E}_{ss}$, we split B trees in G groups and each group has l trees, and the trees in the same group have the same half sample. The final consistent estimator $\hat{\sigma}_n^2(x)$ can be written as

$$\begin{aligned} \hat{\sigma}_n^2(x) &= \frac{1}{G-1}\sum_{g=1}^{G}(\bar{\Psi}_g(x) - \bar{\Psi}(x))^2 \\ &\quad - \frac{1}{(l-1)}\frac{1}{B}\sum_{g=1}^{G}\sum_{i=1}^{l}(\Psi_{ig}(x) - \bar{\Psi}_g(x))^2 \end{aligned}$$

where $\bar{\Psi}_g(x) = \frac{1}{l}\sum_{i=1}^{l}\Psi_{ig}(x)$, and $\bar{\Psi}(x) = \frac{1}{G}\sum_{g=1}^{G}\bar{\Psi}_g(x)$.

The following diagram summarizes the procedure of estimating the variance $\sigma_n^2(x)$.

$$\sigma_n^2(x) \xleftarrow{\text{Asym.equivalent}} s_n^2(x) \xleftarrow{\text{Half-Sampling estimator}}$$

TABLE 2 | Comparison of mean-absolute-error (MAE) and rooted-mean-squared-error (RMSE) for Model 2 with different link functions.

p	SRF	Naive.Cox	Naive.km	Lu.id	Lu.exp	Wang.id	Wang.exp
			Model 2: identity link, $n = 3,000$, SNR = 0.3				
5	**0.1218**	0.1386	0.1384	0.1388	0.1388	0.1382	0.1382
	0.1498	0.1658	0.1656	0.1660	0.1660	0.1656	0.1656
10	**0.1257**	0.1414	0.1412	0.1418	0.1418	0.1411	0.1411
	0.1525	0.1682	0.1679	0.1687	0.1687	0.1684	0.1684
20	**0.1239**	0.1390	0.1385	0.1393	0.1393	0.1387	0.1387
	0.1507	0.1662	0.1655	0.1667	0.1667	0.1663	0.1663
			Model 2: log-exp link, $n = 3,000$, SNR = 0.3				
5	**0.1201**	0.1366	0.1364	0.1368	0.1368	0.1362	0.1362
	0.1479	0.1635	0.1633	0.1637	0.1637	0.1634	0.1634
10	**0.1240**	0.1395	0.1393	0.1399	0.1399	0.1392	0.1392
	0.1506	0.1660	0.1657	0.1664	0.1664	0.1661	0.1661
20	**0.1222**	0.1371	0.1366	0.1374	0.1374	0.1368	0.1368
	0.1487	0.1640	0.1633	0.1645	0.1645	0.1641	0.1641
			Model 2: exp link, $n = 3,000$, SNR = 0.3				
5	**21.030**	23.794	23.733	23.915	23.911	23.542	23.541
	25.984	28.185	28.135	28.297	28.292	28.126	28.125
10	**21.641**	24.165	24.127	24.322	24.319	23.928	23.928
	26.357	28.475	28.430	28.618	28.614	28.473	28.472
20	**21.368**	23.802	23.712	23.956	23.952	23.571	23.571
	26.071	28.216	28.102	28.379	28.375	28.208	28.207

The number of covariates $p = 5, 10, 20$, for each p, the first row is MAE, the second row is RMSE. SRF, proposed random forest-bases estimator; Naive.km, estimate based on Kaplan–Meier estimator without adjusting for the covariates; Naive.Cox, Cox regression based estimator; Lu.id, method of Tian et al. (2014) with identity link; Lu.exp, method of Tian et al. (2014) with exponential link; Wang.id, method of Wang and Schaubel (2018) with identity link; Wang:exp, method of Wang and Schaubel (2018) with exponential link.

$$\hat{H}_n^{HS}(x) \xleftarrow{\text{Empirical estimator}} \hat{\sigma}_n^2(x)$$

where from left to right, the first arrow is based on Theorem 5 of Athey et al. (2018), the second arrow is based on half-sampling of Efron (1980), and the third arrow is supported by Equations (5) and (6) and the method of Sexton and Laake (2009).

4. SIMULATION STUDIES

We present simulations to evaluate the performance of the proposed method in finite sample setting. Two different models for the survival time are considered

- Model 1: $T = g^{-1}(\alpha_0 + \sum_{i=1}^{p} \alpha_i X_i) + \epsilon$
- Model 2: $T = g^{-1}(\alpha_0 + \sum_{i=1}^{p} \alpha_i X_i^2) + \epsilon$

where $X_{i1}, \ldots, X_{ip}$ are independently generated from $Unif(-1, 1)$, $\alpha_0 = 5$, $\alpha_1 = \alpha_2 = 0.25$ and $\alpha_i = 0$ for $i > 2$, and $\epsilon \sim N(0, \sigma^2)$. The variance σ^2 is chosen to have proper signal-noise ratio (SNR),

$$\text{SNR} = \frac{Var(g^{-1}(\alpha_0 + \sum_{i=1}^{p} \alpha_i X_i))}{Var(\epsilon)}.$$

We generate the independent censoring time C_i from a Cox model with the following hazard $\lambda = \lambda_C \exp(X_1 \log 2)$ and λ_C is chosen to have a proper un-censoring rate. The link function g can have the following form

- Identity link: $g^{-1}(x) = x$;
- Exp link: $g^{-1}(x) = \exp(x)$;
- Log-exp link: $g^{-1}(x) = \log(\exp(x) + 1)$.

4.1. Evaluation of Coverage Probability of Predictions

To evaluate the asymptotic results in Theorem 1, we generate five training data sets and one testing data set with the same sample size. The coverage probability performance is evaluated on the testing data set with predictions and confidence intervals derived from 5 independent training data sets. More specifically, for each observation in the testing sample, we obtain the 95% confidence intervals and record how many times a hRMST observation in test sample is within five estimated 95% confidence intervals. The coverage probability of an observation is defined by the its proportion of being covered, and the overall coverage probability of the testing sample is defined by the average of coverage probability of each of its observation. We present the coverage probability results with sample size $n = 1,000, 2,000, 5,000$ for Model 1, and $n = 1,000, 2,000, 10,000$ for Model 2. By choosing the proper λ_C, we control the un-censoring rate around 60–70% for different link functions: $\lambda_C \sim 0.08$ for Identity link and Log-exp link, and $\lambda_C \sim 0.003$ for Exp link. The truncation time L is

TABLE 3 | Comparison of Mean-Absolute-Error (MAE) and Rooted-Mean-Squared-Error (RMSE) for Model 1 with different link functions and the censoring distribution is mis-specified with $\alpha = 0.5$.

p	SRF	Naive.Cox	Naive.km	Lu.id	Lu.exp	Wang.id	Wang.exp
			Model 1: identity link, $n = 3,000$, SNR = 0.3				
5	0.1361	0.1353	0.2051	0.1337	0.1344	**0.1336**	0.1342
	0.1706	**0.1681**	0.2457	0.1687	0.1693	0.1685	0.1690
10	0.1444	0.1430	0.2160	**0.1402**	0.1408	0.1403	0.1408
	0.1755	0.1732	0.2523	0.1726	0.1731	**0.1725**	0.1730
20	0.1392	0.1372	0.2078	**0.1345**	0.1351	**0.1345**	0.1351
	0.1723	0.1699	0.2484	0.1694	0.1700	**0.1692**	0.1698
			Model 1: log-exp link, $n = 3,000$, SNR = 0.3				
5	0.1348	0.1341	0.2032	0.1325	0.1333	**0.1324**	0.1330
	0.1691	**0.1667**	0.2432	0.1673	0.1679	0.1671	0.1676
10	0.1431	0.1418	0.2139	**0.1390**	0.1396	0.1391	0.1396
	0.1740	0.1718	0.2497	0.1712	0.1717	**0.1711**	0.1716
20	0.1380	0.1360	0.2060	0.1335	0.1341	**0.1334**	0.1340
	0.1708	0.1685	0.2460	0.1681	0.1687	**0.1679**	0.1685
			Model 1: exp link, $n = 3,000$, SNR = 0.3				
5	24.906	25.157	33.628	24.471	24.826	**24.427**	24.784
	30.984	30.687	39.205	30.609	30.852	**30.591**	30.800
10	26.381	26.553	35.410	25.738	26.015	**25.678**	25.996
	31.799	31.593	40.265	31.403	31.607	**31.373**	31.574
20	25.096	25.145	33.418	24.461	24.741	**24.365**	24.680
	30.940	30.746	39.152	30.609	30.831	**30.551**	30.759

The number of covariates $p = 5, 10, 20$, for each p, the first row is MAE, the second row is RMSE. SRF, proposed random forest-bases estimator; Naive.km, estimate based on Kaplan–Meier estimator without adjusting for the covariates; Naive.Cox, Cox regression based estimator; Lu.id, method of Tian et al. (2014) with identity link; Lu.exp, method of Tian et al. (2014) with exponential link; Wang.id, method of Wang and Schaubel (2018) with identity link; Wang:exp, method of Wang and Schaubel (2018) with exponential link.

chosen to make the truncation rate fall into 2% − 5%. Specifically, $L \sim 5.4$ for Identity link and Log-exp link, and $L \sim 220$ for Exp link.

Figures 2, **3** present the results for Model 1 and Model 2 under three different link functions. We see that the coverage probability approaches to nominal level 95% when the sample size gets larger. If p is smaller, the coverage probability is closer to 95%. This corresponds to the result of Theorem 3 in Wager and Athey (2015), which states that the rate of convergence of the bias of random forest estimator is $O(n^{\frac{K}{p}})$ for some constant K. When the sample size n is fixed, bigger p leads to larger bias in the estimates of hRMST, and under-coverage of the confidence interval. On the other hand, when p is fixed, bigger n results in a smaller bias and leads to a better coverage of the confidence interval.

4.2. Comparison of Prediction Performance With Existing Methods

We compare our proposed method with several existing methods for hRMST estimation, including

- *Naive.km*: using Kaplan–Meier estimator for survival function and computing hRMST by Equation (1). Covariates are not adjusted.
- *Naive.Cox*: using proportational hazards estimator for the survival function and computing hRMST by Equation (1). The censoring distribution is assumed to follow the proportional hazards assumption.
- *Lu.method*: using some parametric forms of hRMST and computing hRMST by solving a weighted estimating equation. The censoring distribution is assumed to be independent of the covariates (Tian et al., 2014). We consider Identity link and Exp link in the simulations.
- *Wang.method*: using some parametric forms of hRMST and computing hRMST by solving a weighted estimating equation. The censoring distribution is assumed to follow the proportional hazards assumption. We consider Identity link and Exp link in the simulations (Wang and Schaubel, 2018).

We compare all these methods under Model 1 and Model 2, and use the Mean-Absolute-Error (MAE) and Rooted-Mean-Squared-Error (RMSE), introduced in Davison and Hinkley (1997), Tian et al. (2007), and Wang and Schaubel (2018), to measure the performance of these methods.

$$\begin{aligned} \text{MAE} &= \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i^L}{1-\hat{G}(Z_i^L|X_i=x)}\left|Z_i^L-\hat{\mu}^L(X_i)\right|, \\ \text{RMSE} &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i^L}{1-\hat{G}(Z_i^L|X_i=x)}\left[Z_i^L-\hat{\mu}^L(X_i)\right]^2}. \end{aligned} \tag{7}$$

TABLE 4 | Comparison of Mean-Absolute-Error (MAE) and Rooted-Mean-Squared-Error (RMSE) for Model 2 with different link functions and the censoring distribution is mis-specificed with $\alpha = 0.5$.

p	SRF	Naive.Cox	Naive.km	Lu.id	Lu.exp	Wang.id	Wang.exp
			Model 1: identity link, $n = 3,000$, SNR = 0.3				
5	**0.1230**	0.1378	0.1374	0.1385	0.1385	0.1377	0.1377
	0.1514	0.1657	0.1653	0.1663	0.1663	0.1658	0.1658
10	**0.1310**	0.1450	0.1442	0.1457	0.1457	0.1447	0.1447
	0.1562	0.1704	0.1695	0.1712	0.1712	0.1704	0.1704
20	**0.1262**	0.1394	0.1384	0.1403	0.1403	0.1392	0.1392
	0.1533	0.1668	0.1657	0.1681	0.1681	0.1673	0.1673
			Model 1: log-exp link, $n = 3,000$, SNR = 0.3				
5	**0.1213**	0.1359	0.1355	0.1365	0.1365	0.1358	0.1358
	0.1494	0.1634	0.1630	0.1640	0.1640	0.1636	0.1636
10	**0.1292**	0.1430	0.1422	0.1437	0.1437	0.1427	0.1427
	0.1543	0.1681	0.1673	0.1689	0.1689	0.1681	0.1681
20	**0.1244**	0.1374	0.1364	0.1383	0.1383	0.1372	0.1372
	0.1512	0.1645	0.1634	0.1658	0.1658	0.1650	0.1650
			Model 1: exp link, $n = 3,000$, SNR = 0.3				
5	**21.270**	23.793	23.697	24.016	24.009	23.535	23.534
	26.187	28.147	28.075	28.329	28.322	28.133	28.132
10	**22.824**	25.159	24.946	25.408	25.399	24.843	24.842
	27.067	29.009	28.823	29.239	29.227	28.945	28.943
20	**21.832**	23.896	23.708	24.188	24.177	23.698	23.697
	26.635	28.417	28.221	28.753	28.740	28.499	28.499

The number of covariates $p = 5, 10, 20$, for each p, the first row is MAE, the second row is RMSE. SRF, proposed random forest-bases estimator; Naive.km, estimate based on Kaplan–Meier estimator without adjusting for the covariates; Naive.Cox, Cox regression based estimator; Lu.id, method of Tian et al. (2014) with identity link; Lu.exp, method of Tian et al. (2014) with exponential link; Wang.id, method of Wang and Schaubel (2018) with identity link; Wang:exp, method of Wang and Schaubel (2018) with exponential link.

TABLE 5 | Comparison of Mean-Absolute-Error (MAE) and Rooted-Mean-Squared-Error (RMSE) for Model 1 with different link functions and the censoring distribution is mis-specificed with $\alpha = 1.5$.

p	SRF	Naive.Cox	Naive.km	Lu.id	Lu.exp	Wang.id	Wang.exp
			Model 1: identity link, $n = 3,000$, SNR = 0.3				
5	0.1363	0.1378	0.2067	**0.1352**	0.1357	**0.1352**	0.1357
	0.1701	0.1702	0.2467	**0.1697**	0.1702	**0.1697**	0.1702
10	0.1376	0.1385	0.2073	**0.1358**	0.1363	**0.1358**	0.1363
	0.1709	0.1706	0.2472	**0.1699**	0.1704	**0.1699**	0.1704
20	0.1371	0.1371	0.2062	**0.1341**	0.1347	0.1342	0.1347
	0.1698	0.1691	0.2464	**0.1682**	0.1688	**0.1682**	0.1688
			Model 1: log-exp link, $n = 3,000$, SNR = 0.3				
5	0.1350	0.1366	0.2046	**0.1340**	0.1345	**0.1340**	0.1345
	0.1686	0.1687	0.2441	**0.1683**	0.1688	**0.1683**	0.1688
10	0.1363	0.1373	0.2053	**0.1346**	0.1352	0.1347	0.1352
	0.1695	0.1692	0.2447	**0.1685**	0.1690	**0.1685**	0.1690
20	0.1359	0.1359	0.2043	**0.1330**	0.1335	**0.1330**	0.1336
	0.1683	0.1677	0.2439	**0.1669**	0.1674	**0.1669**	0.1674
			Model 1: exp link, $n = 3,000$, SNR = 0.3				
5	24.537	25.171	33.190	24.322	24.601	**24.304**	24.600
	30.701	30.750	38.999	30.549	30.735	**30.532**	30.715
10	24.802	25.317	33.359	24.468	24.743	**24.445**	24.744
	30.798	30.832	39.142	30.577	30.757	**30.560**	30.742
20	24.852	25.188	33.406	24.300	24.567	**24.272**	24.570
	30.732	30.654	39.103	30.384	30.583	**30.371**	30.576

The number of covariates $p = 5, 10, 20$, for each p, the first row is MAE, the second row is RMSE. SRF, proposed random forest-bases estimator; Naive.km, estimate based on Kaplan–Meier estimator without adjusting for the covariates; Naive.Cox, Cox regression based estimator; Lu.id, method of Tian et al. (2014) with identity link; Lu.exp, method of Tian et al. (2014) with exponential link; Wang.id, method of Wang and Schaubel (2018) with identity link; Wang:exp, method of Wang and Schaubel (2018) with exponential link.

We set $n = 3,000, \text{SNR} = 0.3$. For Identity link and Log-exp link, $\lambda_C = 0.08, L = 5.3$. For Exp link $\lambda_C = 0.0026, L = 190$. We calculate the MAE and RMSE for our method and four existing methods(both Lu.method and Wang.method have two link functions) under Model 1 and Model 2 and $p = 5, 10, 20$. Among all the considered models, our method in general has a better performance. As an example, **Figure 4** visualizes the observed hRMST generated from Log-exp link and predicted hRMST from our method and Wang.method, showing that the random forest can give better predictions.

Tables 1, **2** show the MAE and RMSE for Model 1 and Model 2, respectively. For Model 1, the parametric models are correctly specified using the methods of Tian et al. (2014), Wang and Schaubel (2018), we expect that both methods perform well, and our method can have a comparable performance. For Model 2, our proposed method dominates all other methods. Increasing the number of non-predictive covariates does not have a big impact on the performance of our method.

When the censoring distribution does not follow PH assumption, we may expect a difference in the prediction performance because of the bias of IPCW from mis-specification. To check whether our method can still outperform the existing methods, we conduct additional numerical studies. In particular, we simulate the censoring time from the following gamma distributions

$$C \sim \Gamma(\alpha, \beta), \beta = \frac{1}{\lambda_C \exp(X_1 \log 2)}, \text{ and } \alpha \in \{0.5, 1.5\}$$

When $\alpha = 1$, the gamma distribution degenerates to the exponential distribution we used for **Tables 1**, **2**. **Tables 3**, **4** show the MAE and RMSE for Model 1 and Model 2 when $\alpha = 0.5$, and **Tables 5**, **6** show the MAE and RMSE for Model 1 and Model 2 when $\alpha = 1.5$. Results of $\alpha \in \{0.5, 1.5\}$ are not very different from the results of $\alpha = 1$. Under Model 1, our method performs comparably well as methods of Tian et al. (2014), Wang and Schaubel (2018), and it dominates the others under Model 2. When feature dimension is low($p = 5$), the error metrics of our method when $\alpha = 1$ are in general lower than the error metrics when $\alpha = 0.5, 1.5$ for both Model 1 and Model 2. The additional errors can be regarded as the bias induced from the violation of PH assumption of the censoring distribution. When feature dimension is high($p = 10, 20$), bias from large p may dominate the bias from the violation of PH assumption of the censoring distribution.

5. APPLICATION TO THE TCGA OVARIAN CANCER DATA SET

We apply the proposed method to The Cancer Genome Atlas (TCGA) ovarian cancer functional proteomics data set (Akbani et al., 2015) that is publicly available (http://gdac.broadinstitute.org). The data sets include proteomic characterization of tumors using reverse-phase protein arrays (RPPA). Specifically, Akbani et al. (2015) reported an RPPA-based proteomic analysis using 195 high-quality antibodies that target total, cleaved, acetylated and phosphorylated forms of proteins in 412 high-grade serous ovarian cystadenocarcinoma (OVCA) samples. The function space covered by the antibodies used in the RPPA analysis emcompasses major functional and signaling pathways of relevance to human cancer, including proliferation, DNA damage, polarity, vesicle function, EMT, invasiveness, hormone signaling, apoptosis, metabolism, immunological, and stromal function as well as transmembrane receptors, integrin, TGFβ, LKB1/AMPK, TSC/mTOR, PI3K/Akt, Ras/MAPK, Hippo, Notch, and Wnt/beta-catenin signaling (Akbani et al., 2015).

After removing a few samples with missing data, the final data set includes 407 OVCA samples with a mean/median follow-up of 3.20/2.79 years and a total of 242 deaths and 40% censoring. To assess how different methods predict the hRMST, we performed the following cross-validation analysis. For a given L, we did 10-fold cross-validation on the data set. For each training data set in the cross-validation, we perform a univariate analysis to select top 5 most significant features based on univariate Cox regression analysis. We then estimate the hRMST on the test set using the training data sets with these 5 features as the predictors. We apply 7 different methods, including estimate based on the KM estimator, estimate based on the Cox model, the method of Tian et al. (2014) and the method of Wang and Schaubel (2018). We report the average

TABLE 6 | Comparison of Mean-Absolute-Error (MAE) and Rooted-Mean-Squared-Error (RMSE) for Model 2 with different link functions and the censoring distribution is mis-specificed with $\alpha = 1.5$.

p	SRF	Naive.Cox	Naive.km	Lu.id	Lu.exp	Wang.id	Wang.exp
Model 1: identity link, $n = 3,000$, SNR = 0.3							
5	**0.1227**	0.1396	0.1395	0.1397	0.1397	0.1394	0.1394
	0.1507	0.1666	0.1664	0.1668	0.1668	0.1666	0.1666
10	**0.1241**	0.1391	0.1389	0.1393	0.1393	0.1390	0.1390
	0.1514	0.1667	0.1664	0.1669	0.1669	0.1668	0.1668
20	**0.1232**	0.1390	0.1386	0.1393	0.1393	0.1389	0.1389
	0.1499	0.1659	0.1654	0.1663	0.1663	0.1661	0.1661
Model 1: log-exp link, $n = 3,000$, SNR = 0.3							
5	**0.1210**	0.1376	0.1375	0.1378	0.1378	0.1374	0.1374
	0.1487	0.1643	0.1642	0.1645	0.1645	0.1643	0.1643
10	**0.1224**	0.1372	0.1370	0.1374	0.1374	0.1371	0.1371
	0.1494	0.1644	0.1642	0.1646	0.1646	0.1645	0.1645
20	**0.1215**	0.1371	0.1368	0.1374	0.1374	0.1370	0.1370
	0.1480	0.1637	0.1632	0.1641	0.1641	0.1638	0.1638
Model 1: exp link, $n = 3,000$, SNR = 0.3							
5	**21.071**	23.719	23.699	23.787	23.785	23.581	23.580
	26.092	28.241	28.217	28.313	28.311	28.238	28.238
10	**21.334**	23.649	23.612	23.711	23.710	23.524	23.524
	26.159	28.231	28.186	28.283	28.281	28.224	28.224
20	**21.176**	23.629	23.571	23.748	23.745	23.492	23.492
	25.893	28.077	27.993	28.208	28.204	28.085	28.085

The number of covariates $p = 5, 10, 20$, for each p, the first row is MAE, the second row is RMSE. SRF, proposed random forest-bases estimator; Naive.km, estimate based on Kaplan–Meier estimator without adjusting for the covariates; Naive.Cox, Cox regression based estimator; Lu.id, method of Tian et al. (2014) with identity link; Lu.exp, method of Tian et al. (2014) with exponential link; Wang.id, method of Wang and Schaubel (2018) with identity link; Wang:exp, method of Wang and Schaubel (2018) with exponential link.

TABLE 7 | Performance of the proposed random forest estimator compared with other methods for $L = 3, 4, 5$.

L	SRF	Naive.Cox	Naive.km	Lu.id	Lu.exp	Wang.id	Wang.exp
3	**0.6879**	0.9247	0.9463	0.9266	0.9355	0.7630	0.7721
	0.8258	0.8925	0.8967	0.8966	0.8983	0.8438	0.8455
4	**1.2033**	1.5450	1.5686	1.5704	1.5777	1.2862	1.3044
	1.2403	1.3597	1.3648	1.3830	1.3817	1.2719	1.2752
5	**1.7479**	2.2107	2.2395	2.2467	2.2306	1.8251	1.8540
	1.6761	1.8594	1.8655	1.8989	1.8858	1.7168	1.7193

The first row is MAE, the second row is RMSE. SRF, proposed random forest estimator; Naive.km, estimate based on Kaplan–Meier estimator without adjusting for the covariates; Naive.Cox, Cox regression based estimator; Lu.id, method of Tian et al. (2014) with identity link; Lu.exp, method of Tian et al. (2014) with exponential link; Wang.id, method of Wang and Schaubel (2018) with identity link; Wang:exp, method of Wang and Schaubel (2018) with exponential link.

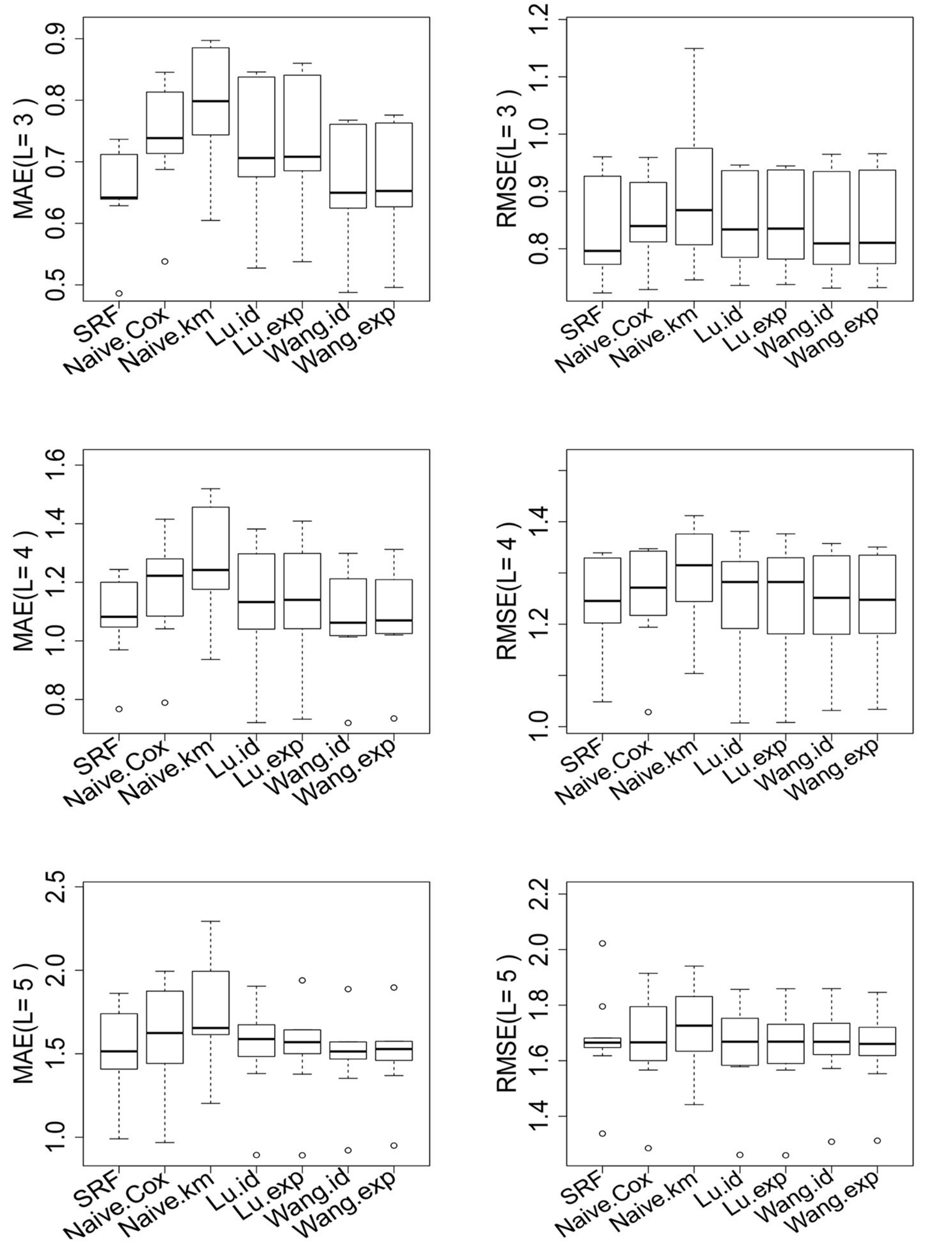


FIGURE 5 | Performance of the proposed random forest estimator compared with other methods for $L = 3, 4, 5$. The left penal is the MAE across of 10-fold cross-validation. The right panel is the RMSE across of 10-fold cross-validation. SRF, proposed random forest estimator; Naive.km, estimate based on Kaplan–Meier estimator without adjusting for the covariates; Naive.Cox, Cox regression based estimator; Lu.id, method of Tian et al. (2014) with identity link; Lu.exp, method of Tian et al. (2014) with exponential link; Wang.id method of Wang and Schaubel (2018) with identity link; Wang:exp, method of Wang and Schaubel (2018) with exponential link.

of MAE and RMSE on the samples in the testing sets over the 10-fold cross-validation.

The results are shown in **Table 7** and **Figure 5** for $L = 3, 4, 5$ (see **Supplementary Material** for $L = 6, 7, 8$). There are $45.9, 31.2, 19.4, 11.8, 8.1, 4.4\%$ of the observations larger than L for $L = 3, 4, 5, 6, 7, 8$ correspondingly. For different choices of L, our proposed random forest based method dominates the other methods in MAE and RMSE. The methods of Tian et al. (2014) and Wang and Schaubel (2018) are based on parametric form of hRMST. Cox model is heavily dependent on the proportional hazard assumption, and the Kaplan–Meier approach does not take the covariates into account. We also notice that the method of Wang and Schaubel (2018) always performs better than the method of Tian et al. (2014), possibly due to the fact that the censoring mechanism in the data depends on the covariates.

6. DISCUSSION

In this paper, we have developed a non-parametric random forest-based method for estimation of hRMST. Compared with traditional Cox model, which gets hRMST estimates by transforming the estimated hazard functions, directly modeling hRMST would be more preferable for computation and feature importance analysis. The proposed estimator can relax the parametric assumptions imposed on the survival time used in Tian et al. (2014) and Wang and Schaubel (2018), and can achieve better prediction performance. We have derived the asymptotic distribution of the random forest estimator using IPCW approach, and presented a procedure based on bags of little bootstraps to obtain the variance of the estimator. Our simulation results and analysis of TCGA data sets have shown promising performance in predicting hRMST as compared to the other available methods, even when the dimension is high and the covariates include irrelevant variables. The method is implemented by R and C++, and is available at https://github.com/lmy1019/SRF.

The proposed method can be used to estimate the heterogeneous treatment effects in randomized clinical trials when the outcome is censored. One can simply apply the method separately to the treated group and the placebo group and take the difference. However, for the observational studies, one needs to account for the fact that the treatment assignments might not be completely at random. Wager and Athey (2015) developed a non-parametric causal forest for estimating heterogeneous treatment effects that extends Breiman's random forest algorithm. In the potential outcomes framework with non-confounding, they showed that causal forest are pointwise consistent for the true treatment effect and have an asymptotically Gaussian and centered sampling distribution. For the observational studies with censored survival outcomes, it is also possible to combine the methods proposed here and the method of Wager and Athey (2015) in order to estimate the treatment effect on the restricted mean survival time.

The proposed methods can also be extended to take into account possible competing risk. This can be done by introducing an additional inverse probability weight (IPCW) to differentiate the non-informative censoring and competing risk censoring. In this case, the estimation equation ψ function with covariates history $\tilde{X} = \tilde{x}$ under true G_C and G_R becomes

$$\tilde{\psi}_\mu(\tilde{x}, Z^L, \delta^L) = \frac{1}{1 - G_C(Z^L|X = x)} \frac{1}{1 - G_R(Z^L|\tilde{X} = \tilde{x})} \delta^L \left(Z^L - \mu \right), \tag{8}$$

where under competing risk scenario, $\delta^L = 1_{\{T \wedge L \leq C \wedge R\}}$. The method proposed in this paper can be automatically adapted to the competing risk case and the asymptotic normality result can be derived similarly.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

ML and HL developed the ideas and the methods together, analyzed the real data sets, and wrote the manuscript. ML implemented the methods and performed the numerical analysis. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded by NIH grants GM123056 and GM129781.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.587378/full#supplementary-material

REFERENCES

Akbani, R., Ng, P. K. S., Werner, H. M., Shahmoradgoli, M., Zhang, F., Ju, Z., et al. (2015). Corrigendum: a pan-cancer proteomic perspective on the Cancer Genome Atlas. *Nat. Commun*. 6:5852. doi: 10.1038/ncomms5852

Andersen, P. K., and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Stat*. 10, 1100–1120. doi: 10.1214/aos/1176345976

Athey, S., Tibshirani, J., and Wager, S. (2018). *Generalized Random Forests*. Technical report. Stanford, CA: Stanford University.

Biau, G. (2012). Analysis of a random forests model. *J. Mach. Learn. Res*. 13, 1063–1095.

Biau, G., Devroye, L., and Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *J. Mach. Learn. Res*. 9, 2015–2033.

Biau, G., and Scornet, E. (2016). A random forest guided tour. *Test* 25, 197–227. doi: 10.1007/s11749-016-0481-7

Breiman, L. (2001). Random forests. *Mach. Learn*. 45, 5–32. doi: 10.1023/A:1010933404324

Breiman, L. (2004). *Consistency for a Simple Model of Random Forests*. Technical report 670. Statistics Department, University of California at Berkeley.

Chen, P. Y., and Tsiatis, A. A. (2001). Causal inference on the difference of the restricted mean lifetime between two groups. *Biometrics* 57, 1030–1038. doi: 10.1111/j.0006-341X.2001.01030.x

Cox, D. (1972). Regression models and life-tables. *J. R. Stat. Soc. Ser, B* 34, 187–220. doi: 10.1111/j.2517-6161.1972.tb00899.x

Cox, D. (1975). Partial likelihood. *Biometrika* 62, 269–276. doi: 10.1093/biomet/62.2.269

Cutler, D., Edwards, T. C., Beard, K., Cutler, A., Hess, K., Gibson, J., and Lawler, J. (2007). Random forests for classification in ecology. *Ecology* 8811, 2783–2792.

Davison, A., and Hinkley, D. V. (1997). *Bootstrap Methods and Their* Application. Cambridge University Press. Available online at: https://www.cambridge.org/core/books/bootstrap-methods-and-their-application/

ED2FD043579F27952363566DC09CBD6A. doi: 10.1017/CBO9780511802843

Denil, M., Matheson, D., and De Freitas, N. (2014). "Narrowing the gap: random forests in theory and in practice," in *Proceedings of The 31st International Conference on Machine Learning*, 665–673. Available online at: http://proceedings.mlr.press/v32/denil14.html

Díaz-Uriarte, R., and Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* 7:3. doi: 10.1186/1471-2105-7-3

Erfon, B. (1980). *The Jackknife, the Bootstrap, and Other Resampling Plans.* Available online at: https://statistics.stanford.edu/research/jackknife-bootstrap-and-other-resampling-plans

Fang, E. X., Ning, Y., and Liu, H. (2017). Testing and confidence intervals for high dimensional proportional hazards model. *J. R. Stat. Soc. Ser. B*. 79, 1415–1437. doi: 10.1111/rssb.12224

Friedberg, R., Tibshirani, J., Athey, S., and Wager, S. (2018). Local linear forests. *J. Comput. Graph. Stat*. 1–25. doi: 10.1080/10618600.2020.1831930

Gill, R. D., and Gill, R. D. (1984). Understanding Cox's regression model: a martingale approach. *J. Am. Stat. Assoc*. 79, 441–447. doi: 10.1080/01621459.1984.10478069

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning.* New York, NY: Springer New York Inc. Available online at: https://www.bibsonomy.org/bibtex/2f58afc5c9793fcc8ad8389824e57984c/sb3000. doi: 10.1007/978-0-387-84858-7

Huang, J., Sun, T., Ying, Z., Yu, Y., and Zhang, C. H. (2013). Oracle inequalities for the lasso in the cox model. *Ann. Stat*. 41, 1142–1165. doi: 10.1214/13-AOS1098

Ishwaran, H., and Kogalur, U. B. (2011). Consistency of random survival forests. *Stat. Probab. Lett*. 80, 1056–1064. doi: 10.1016/j.spl.2010.02.020

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *Ann. Appl. Stat*. 2, 841–860. doi: 10.1214/08-AOAS169

Meinshausen, N. (2006). Quantile regression forests. *J. Mach. Learn. Res.* 7, 983–999.

Mentch, L., and Hooker, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *J. Mach. Learn. Res.* 17, 26:1–26:41.

Royston, P., and Parmar, M. K. B. (2013). Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Med. Res. Methodol.* 13:152. doi: 10.1186/1471-2288-13-152

Sexton, J., and Laake, P. (2009). Standard errors for bagged and random forest estimators. *Comput. Stat. Data Anal.* 53, 801–811. doi: 10.1016/j.csda.2008.08.007

Steingrimsson, J. A., Diao, L., and Strawderman, R. L. (2019). Censoring unbiased regression trees and ensembles. *J. Am. Stat. Assoc*. 114, 370–383. doi: 10.1080/01621459.2017.1407775

Svetnik, V., Culberson, J. C., Tong, C., Cullberson, J. C., Sheridan, R. P., and Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inform. Comput. Sci.* 43, 1947–1958. doi: 10.1021/ci034160g

Tian, L., Tianxi, C., Goetghebeur, E., and Wei, L. J. (2007). Model evaluation based on the sampling distribution of estimated absolute prediction error. *Biometrika* 2, 297–311. doi: 10.1093/biomet/asm036

Tian, L., Zhao, L., and Wei, L. J. (2014). Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. *Biostatistics* 15, 222–233. doi: 10.1093/biostatistics/kxt050

Wager, S., and Athey, S. (2015). Estimation and inference of heterogeneous treatment effects using random forests. *J. Am. Stat. Assoc.* 113, 1228–1242. doi: 10.1080/01621459.2017.1319839

Wang, X., and Schaubel, D. E. (2018). Modeling restricted mean survival time under general censoring mechanisms. *Lifetime Data Anal*. 24, 176–199. doi: 10.1007/s10985-017-9391-6

Zhang, M., and Schaubel, D. E. (2011). Estimating differences in restricted mean lifetime using observational data subject to dependent censoring. *Biometrics* 67, 740–749. doi: 10.1111/j.1541-0420.2010.01503.x

Zucker, D. M. (1998). Restricted mean life with covariates: modification and extension of a useful survival analysis method. *J. Am. Stat. Assoc*. 93, 702–709. doi: 10.1080/01621459.1998.10473722

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.

ORIGINAL RESEARCH
published: 16 February 2021
doi: 10.3389/fgene.2021.598304

Identification of Potential Driver Genes Based on Multi-Genomic Data in Cervical Cancer

Yuexun Xu[1†], *Hui Luo*[2†], *Qunchao Hu*[2*] *and Haiyan Zhu*[2*]

[1] *Department of Gynecology and Obstetrics, Henan Provincial People's Hospital, People's Hospital of Zhengzhou University, Zhengzhou, China,* [2] *Department of Gynecology, Shanghai First Maternity and Infant Hospital, Tongji University School of Medicine, Shanghai, China*

Background: Cervical cancer became the third most common cancer among women, and genome characterization of cervical cancer patients has revealed the extensive complexity of molecular alterations. However, identifying driver mutation and depicting molecular classification in cervical cancer remain a challenge.

Methods: We performed an integrative multi-platform analysis of a cervical cancer cohort from The Cancer Genome Atlas (TCGA) based on 284 clinical cases and identified the driver genes and possible molecular classification of cervical cancer.

Results: Multi-platform integration showed that cervical cancer exhibited a wide range of mutation. The top 10 mutated genes were TTN, PIK3CA, MUC4, KMT2C, MUC16, KMT2D, SYNE1, FLG, DST, and EP300, with a mutation rate from 12 to 33%. Applying GISTIC to detect copy number variation (CNV), the most frequent chromosome arm-level CNVs included losses in 4p, 11p, and 11q and gains in 20q, 3q, and 1q. Then, we performed unsupervised consensus clustering of tumor CNV profiles and methylation profiles and detected four statistically significant expression subtypes. Finally, by combining the multidimensional datasets, we identified 10 potential driver genes, including GPR107, CHRNA5, ZBTB20, Rb1, NCAPH2, SCA1, SLC25A5, RBPMS, DDX3X, and H2BFM.

Conclusions: This comprehensive analysis described the genetic characteristic of cervical cancer and identified novel driver genes in cervical cancer. These results provide insight into developing precision treatment in cervical cancer.

Keywords: cervical cancer, TCGA, multi-platform analysis, molecular classification, driver mutation

OPEN ACCESS

Edited by:
Md. Ashad Alam,
Tulane University, United States

Reviewed by:
Md. Mehedi Hasan,
Kyushu Institute of Technology, Japan
Ye Zhu,
Yangzhou University, China

****Correspondence:***
Haiyan Zhu
zhuhaiyandoc@sina.com
Qunchao Hu
hqc82@163.com

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *24 August 2020*
Accepted: *07 January 2021*
Published: *16 February 2021*

Citation:
Xu Y, Luo H, Hu Q and Zhu H (2021) Identification of Potential Driver Genes Based on Multi-Genomic Data in Cervical Cancer. Front. Genet. 12:598304. doi: 10.3389/fgene.2021.598304

INTRODUCTION

As the most common gynecological malignancy, cervical cancer has been reported to have about 570,000 new cases and 311,365 deaths in 2018 worldwide and has become the third most common cancer among women (Bray et al., 2018). Persistent infection with oncogenic types of human papillomavirus (HPV) is now considered the principal etiological agent in cervical cancer (Moody and Laimins, 2010; Litwin et al., 2017). In fact, the majority of HPV infections are transient and do not result in malignant transformation. Only a small percentage of women experience persistent infection, which leads to genomic instability and accumulation of somatic mutations, thus developing malignant cancers finally (Litwin et al., 2017).

Although major achievements have been made in surgery, chemotherapy, and radiotherapy in current decades, the molecular biomarkers and potential treatment targets remain necessarily.

Appreciable evidence implicates specific genomic alterations involved in the initiation and progression of cervical cancer. The genome characterization of a large number of cervical patients has revealed the extensive complexity of molecular alterations, such as somatic aberrations (Ojesina et al., 2014), copy number alterations (CNAs) (Rao et al., 2004), DNA methylation (Verlaat et al., 2017), and dysfunctional microRNA (miRNA) (Cheung et al., 2012). Chen et al. (2013) performed the first genome-wide association study (GWAS) of cervical cancer and identified three independently acting loci (DAP, NR5A2, and MIR365-2 gene regions) within the major histocompatibility complex (MHC) region contributing to the risk of developing cervical cancer, which support its role in high-risk HPV infection and persistence. Ojesina et al. (2014) reported 115 cervical carcinoma–normal paired samples' whole-exome sequence analysis, 79 cases' transcriptome sequence, and 14 tumor–normal pairs' whole genome sequence and detected significantly recurrent somatic mutations in the mitogen-activated protein kinase 1 (MAPK1) gene among squamous cell cervical cancers and provided evidence of potential ERBB2 (also means HER2/neu) activation by somatic mutation, amplification, and HPV integration to combat cervical carcinoma. Despite these discoveries, attempts to apply molecular-targeted agents for treatment of cervical cancer have met with limited success thus far.

During the development of cancer, a large number of somatic mutations occur; however, only a handful of somatic mutations are expected to initiate and promote tumor growth, so-called driver mutations (Nehrt et al., 2012). Several driver mutations have been identified as a subtype for specific cancer type or as a target in therapy. Li et al. (2018a) identified 11 novel driver genes through integrative analysis of 1,061 hepatocellular carcinoma genomes and employed three MutSig algorithms, non-negative matrix factorization, Kaplan–Meier survival and Cox regression analyses, as well as logistic regression model and discovered 11 novel driver genes and further validated AURKA, a small molecule inhibitor, as a druggable target in this disease. Ganly et al. (2018) identified the genomic characterization of 56 primary Hurthle cell carcinoma and elucidate the mutational profile and driver mutations of these tumors. They also identified the disease pathogenesis signaling pathway and the importance of the receptor tyrosine kinase (RTK)/(It is encoded by ras gene which acts as a oncogene) RAS/(it has Ser/Thr protein kinase activity) RAF/MAPK and phosphoinositide 3-kinase (PIK3)/AKT/mammalian target of rapamycin (mTOR) pathways in Hurthle cell carcinoma, and further clinical trial demonstrated multiple tyrosine kinase inhibitor sorafenib and the mTOR inhibitor everolimus showed a significant response rate for these agents (Ganly et al., 2018).

However, driver genes in cervical cancer remain to be identified. In the current study, we integrated somatic mutation, copy number variation (CNV), DNA methylation, and miRNA profile; depicted a comprehensive genomic landscape of cervical cancer; performed molecular classification; and finally identified driver genes. Thus, developing novel targeted therapy against specific somatic alterations finally improves current strategies to combat cervical carcinomas.

MATERIALS AND METHODS

Data Resource

The mutant MAF file of cervical cancer was downloaded using the R package TCGA biolinks (Colaprico et al., 2016), which contains the mutation results of 297 samples. Screening the various cancer type, single-nucleotide polymorphism (SNP)6 copy number segment 287 datasets, and 299 methylation chip data of cervical cancer samples were downloaded from FireBrowse (http://firebrowse.org/) with Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma (platform for Illumina 450K chip). Besides, 304 messenger RNA (mRNA) expression profile data and 307 miRNA expression profile data of cervical cancer samples were downloaded from the National Cancer Institute Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/). Overall, we integrated 284 samples of multiple data features for further analysis, including mutation location, CNV information, methylation data, and mRNA and miRNA expression profile datasets. In addition, cervical cancer fusion genes were downloaded from the Tumor Fusion Gene Data Portal (https://tumorfusions.org/PanCanFusV2/database).

Single-Nucleotide Polymorphism Correlation and Copy Number Variation Analysis

Driver gene analysis was performed by GenePattern (https://cloud.genepattern.org/gp/pages/index.jsf) with corresponding MutSigCV module (Reich et al., 2006). Maftools of R package was used for mutation spectrum to identify mutations in tumor samples. SomaticSignatures was applied for mutation detection and plots the mutation spectrum and mutation characteristics (Gehring et al., 2015; Mayakonda et al., 2018). The GISTIC algorithm was used to detect the common CNV regions in all samples with q-value <0.05, including chromosome arm horizontal CNV and the smallest common region between samples. For chromosomal mutation, a region ratio higher than 0.98 was recognized as a chromosomal arm alternative site. Tumor purity and ploidy analysis were performed based on CNV results using R-package Absolute (https://software.broadinstitute.org/cancer/cga/absolute_download).

Subgroup Identification and Molecular Characteristics Analysis

Unsupervised clustering algorithm was applied to cluster the data from four different platforms (DNA copy, DNA methylation, mRNA expression profile, miRNA expression profile), and subpopulations were identified based on each data platform analysis. The cluster-of-clusters analysis (CoCA) was used to recluster the obtained classification results and integrated the subgroup classification results from different data platforms (Hoadley et al., 2014; Chen et al., 2016).

Chi-statistical tests were performed on each subgroup and clinical features, including tumor stage, differentiation grade, HPV infection, and the association relationship between each subgroup and clinical features. Furthermore, we applied the R package Seurat (https://satijalab.org/seurat/) FindAllMarkers to preform characteristic marker screening of subpopulations including mRNA, miRNA, and methylation profiles. Subpopulation gene mutation characterization: Maftools was applied for each subgroup mutation type (C > T, T > C, C > A, T > G, C > G, T > A, converting Ti, translating Tv). Statistical analysis was performed to compare the differences in the types of mutations between subgroups and used for the identification of co-mutation/exclusion mutation genes and mutation signature analysis. In addition, comparing the difference features between subgroups, APOBEC (apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like) enrichment analysis was performed to count the TCW (W refers to G or T) and non-TCW mutation ratio. Genes with significant differences in the proportion of mutations in each subpopulation were screened for further analysis.

Subgroup CNV characteristics: each subgroup was checked for the copy number changes of all chromosomes, counting the samples with copy number changes for each chromosome segment in each subgroup and performing chi-square test. The identified region is filtered by significantly different copy number changes for chromosome segments in each subgroup.

Statistical Analysis

Two-tailed Student's *t*-test was used to compare the means of two groups. One-way ANOVA analysis of variance with Tukey–Kramer *post-hoc* test was used for analyzing data when means from more than two groups were compared. *P* < 0.05 was considered to be statistically significant. All the statistical analysis was performed with SPSS 17.0 statistical software.

RESULTS

Patient Cohort and Molecular Analysis Strategy

To identify and characterize cervical cancer genome alterations, tissue specimens were analyzed by multiple genomic assays, including whole-exome sequencing for mutations, SNP arrays for

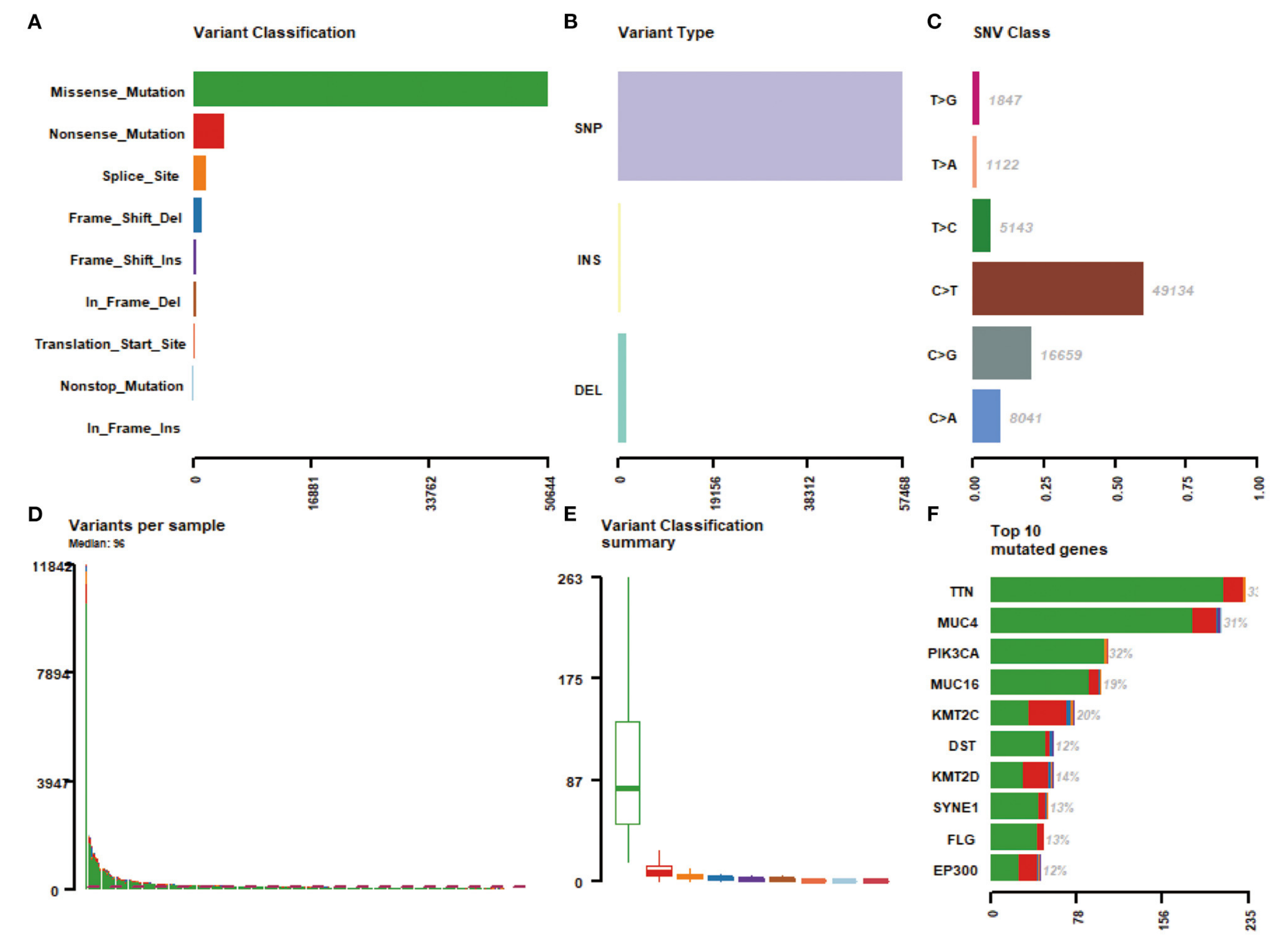


FIGURE 1 | A summary of the genes mutated in 284 cervical cancer samples. **(A)** Variant classification; **(B)** Varian type; **(C)** SNV class; **(D)** Variants per sample; **(E)** Variant classification summary; **(F)** Top 10 mutated genes.

copy number analysis, mRNA sequencing, miRNA sequencing, and DNA methylation arrays (**Supplementary Table 1**). Totally, 284 cases were available for the multiplatform, and the clinical characteristics of the included patients are presented in **Supplementary Table 2**. The mean age at initial diagnosis of cervical cancer was 46 years, with a range of 20–88 years. Among them, 233 patients (81.7%) were squamous cervical cancer, 46 were adenocarcinoma, and five were adenosquamous carcinoma. After a median follow-up period of 636 days, 221 patients suffered death.

Mutation Landscape of Cervical Cancer

Massively parallel sequencing was performed to detect somatic mutations on tumor samples from the cohort of cervical cancer patients. Here, 233 patient samples (82.04%) have been detected to have somatic mutations, and a total number of 83,386 somatic mutations were obtained, including 50,644 missense mutations. **S**NV occurs predominantly in cervical cancer, with C > T being the most common type of mutation. **Figure 1** showed a summary of the genes mutated in cervical cancer. The top 10 mutated genes were TTN, PIK3CA, MUC4, KMT2C, MUC16, KMT2D, SYNE1, FLG, DST, and EP300, with a mutation rate from 12 to 33% (**Figures 1F**, **2A,B**).

We then described the mutation spectrum and mutational signatures among cervical cancers and identified 96 types of mutation signatures (**Figure 2C**). Mutational signatures of cervical cancer were enriched in deficiency of DNA mismatch repair (COSMIC Signature 6; cosine similarity: 0.895), APOBEC-cytidine deaminase (COSMIC Signature 2; cosine similarity: 0.846), and spontaneous deamination of 5-methyl cytosine (COSMIC Signature 1; cosine similarity: 0.951) (**Figure 2D**).

Copy Number Variation of Cervical Cancer

Applying GISTIC to detect CNV, the most frequent chromosome arm-level CNVs included losses in 4p, 11p, and 11q and gains in 20q, 3q, and 1q (**Figure 3A**). Besides, 25 focal deletion peaks and 21 focal amplification peaks were detected (**Figure 3B**). Among them, the most significant amplification region was 3q26.31 and 11q22.1, while the most marked deletion region was 11q24.2 and 2q37.2 (**Figure 3B**). We used ABSOLUTE to estimate tumor purity and tumor ploidy. As described in **Figure 3C**, tumor purity was in range in 0.21–1 and the ploidy was 1.70–9.87, suggesting genomic disorder was a common phenomenon in the development of cervical cancer.

Molecular Classification

To derive a molecular classification for cervical cancer, we performed unsupervised consensus clustering of tumor CNV profiles, methylation profile, mRNA profile, and miRNA profile, respectively, finally detecting four statistically significant expression subtypes. Firstly, hierarchical clustering was performed according to CNV profile, resulting in 284 samples

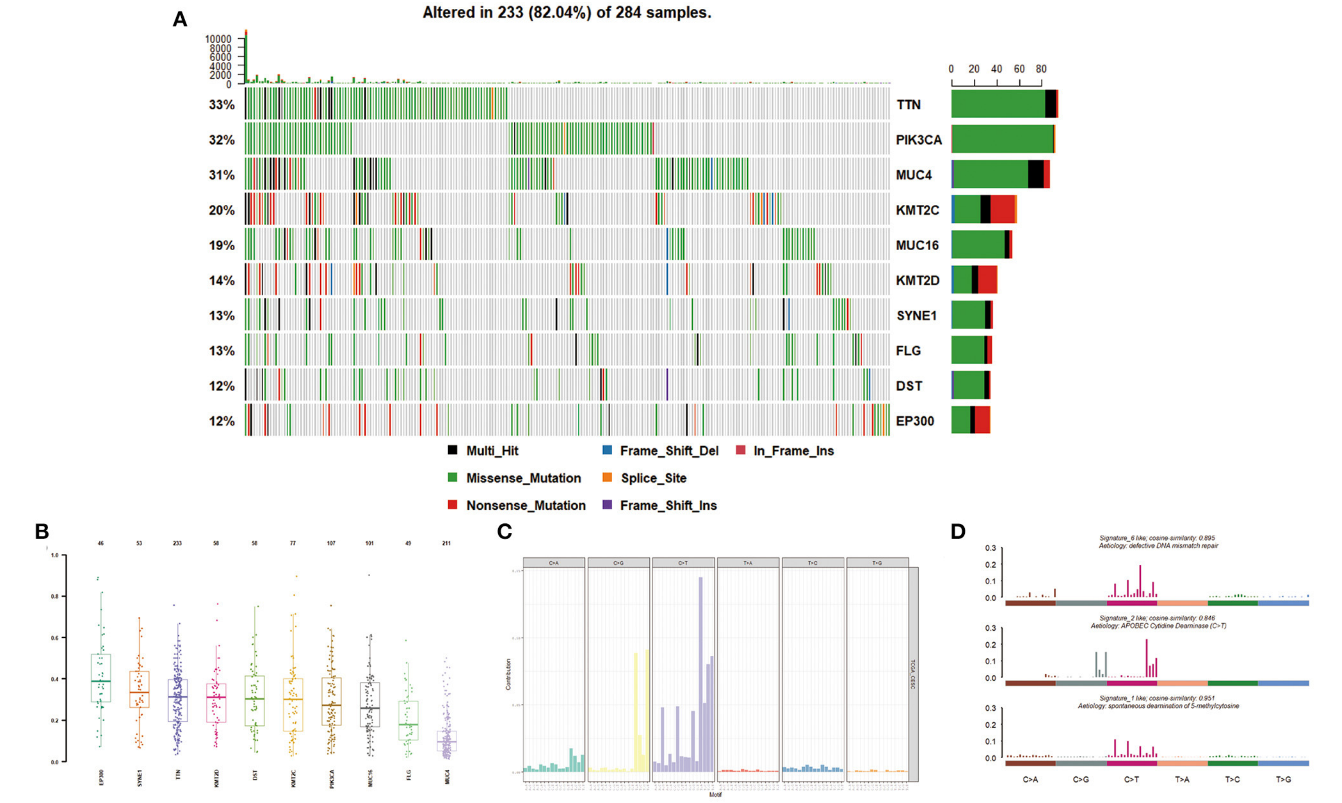


FIGURE 2 | Mutation distribution in cervical cancer patients. **(A,B)** Frequency of specific mutation genes. **(C,D)** Mutation signature analysis.

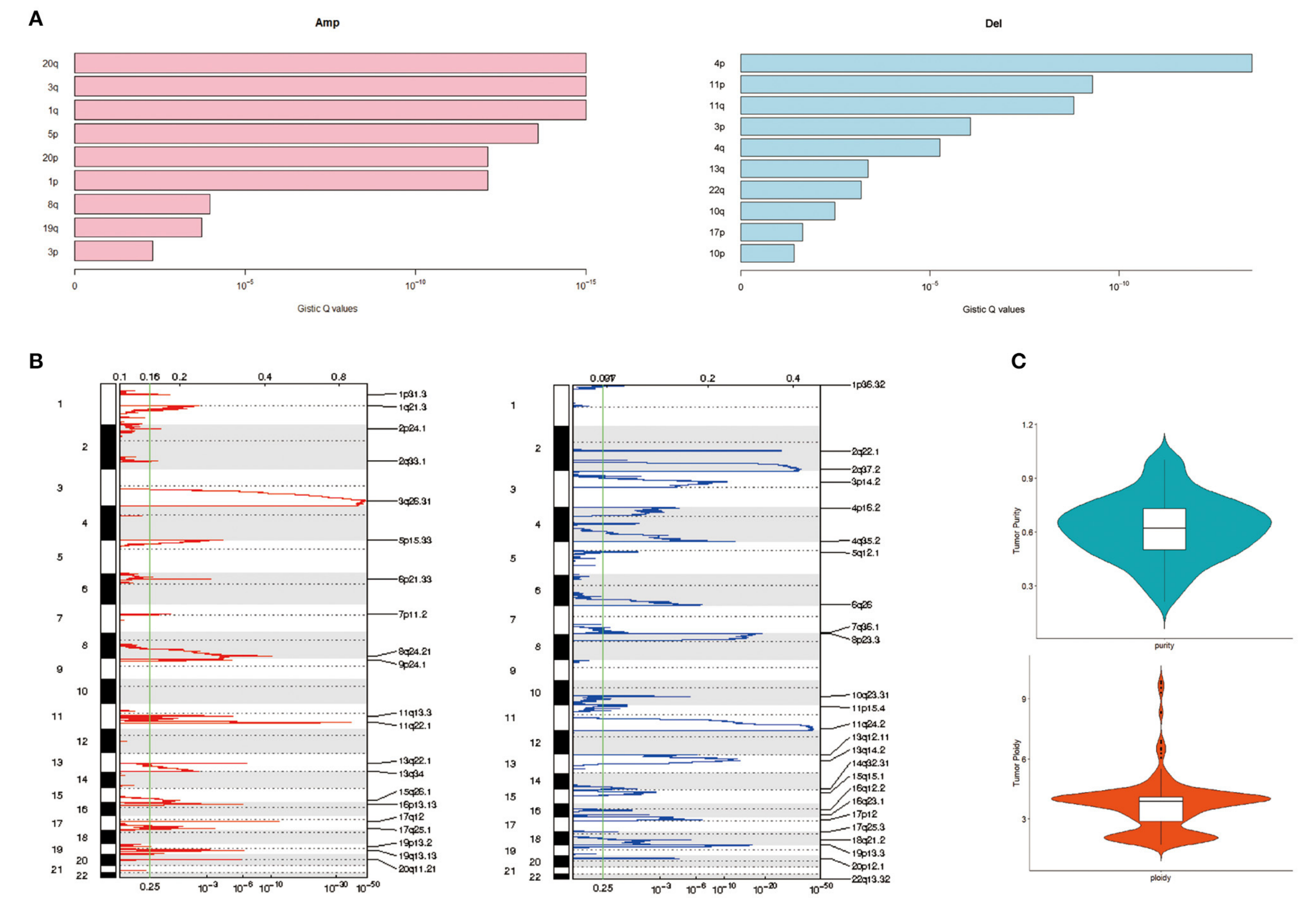


FIGURE 3 | Copy number variation (CNV) of cervical cancer. **(A)** Chromatin amplification and deletion. **(B)** Genome-wide distribution of chromatin amp and del. **(C)** Purity and ploidy of cervical cancer.

divided into two subtypes (**Figure 4A**). Then, gene methylation data of 284 cervical tumor tissues were clustered, and cases were divided into a higher cluster and lower cluster based on the clustering results (**Figure 4B**). However, the effect of clustering was not obvious based on mRNA or miRNA profile. Therefore, unsupervised clustering of all samples based on CNV profile and methylation profile was further performed for molecular classification. Finally, unsupervised clustering defined four subtypes that had diverse CNV and methylation events using COCA approach. Cluster 1 was enriched for CNV and poor in methylation. Cluster 2 was enriched for methylation and poor in CNV. Cluster 3 was poor in both CNV and methylation. Cluster 4 was enriched for both CNV and methylation (**Figures 4C,D**).

We then analyzed the correlation between each subgroup and clinical characteristic, including pathology, differentiation, TNM stage, HPV integration, and survival status. As shown in **Figure 5**, with respect to pathology, squamous cell carcinoma, adenosquamous carcinoma, and adenocarcinoma had significant differences in the distribution of four subpopulations, especially, Cluster 3 is almost squamous cell carcinoma. In addition, comparing the distribution of HPV integration samples, HPV integration was significantly different among the four subpopulations, with the highest proportion of HPV integration samples in Cluster 2. We then analyzed gene mutation in these four clusters (**Figure 5**), 81 gene mutations showed differences across clusters. Of note, mutation samples were more frequent in Cluster 3 than in other clusters, further suggesting Cluster 3 has special molecular mutation characteristics. Distinguishing the characteristic genes of each subgroup, we calculated the differentially expressed genes, miRNAs, and methylation of each subgroup. Several specific high expression genes were identified in cluster 2, and one specific high expression gene (MAL) was identified in cluster 1. However, there was no specific high expression gene in clusters 3 and 4. These results indicated that cluster 2 was significantly different from other subgroups in gene expression and had its special molecular features. In clusters 2 and 3, 182 and 138 special methylation sites were detected, commenting on 130 and 96 genes, respectively. Functional enrichment analysis showed these genes were involved in bone morphogenesis and skeletal development (**Figure 5**). In Cluster 4, 104 special methylation sites were detected, commenting on

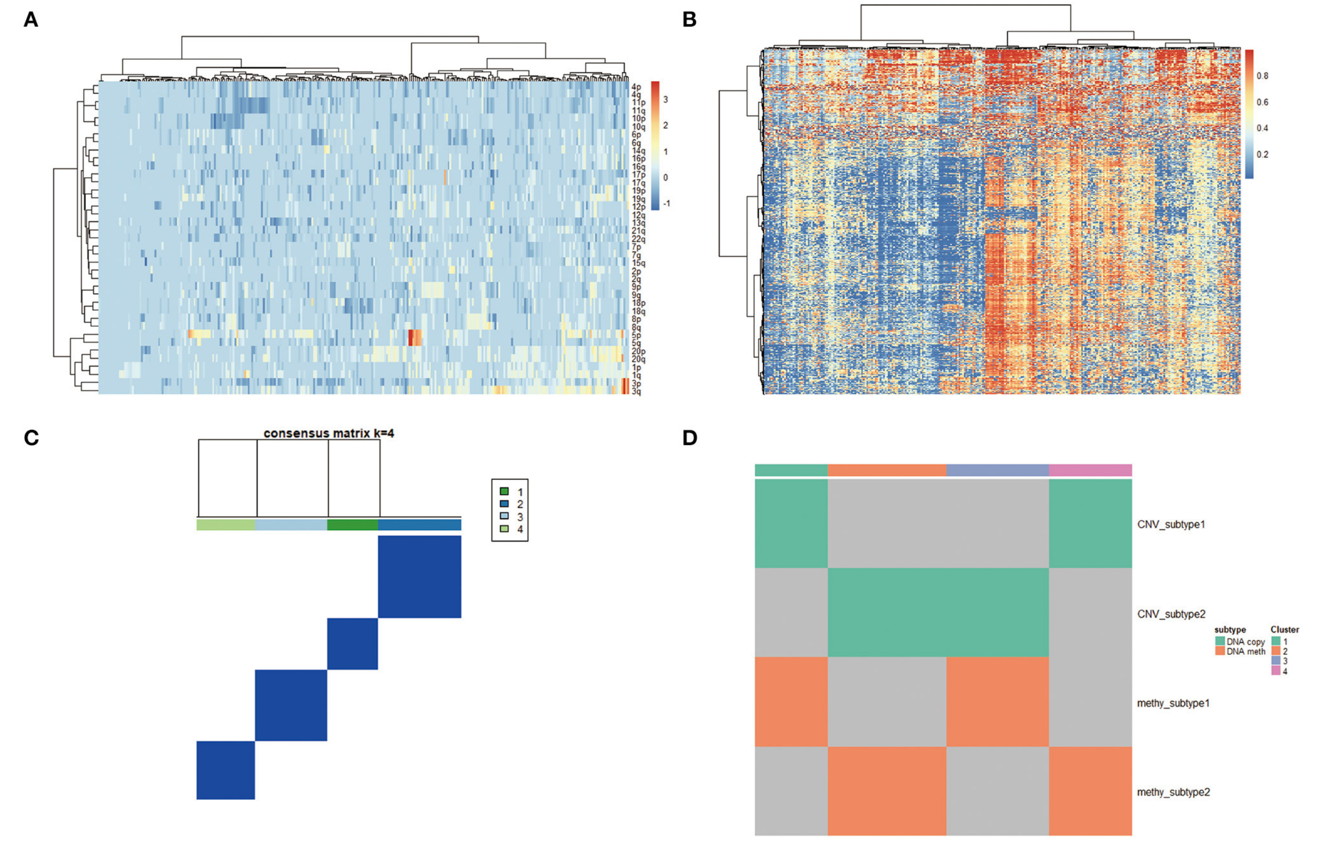


FIGURE 4 | (A) Copy number variation (CNV) landscape in cervical cancer. Hierarchical clustering of CNV data, with the heatmap showing beta values ordered by CNV clusters. **(B)** DNA methylation landscape in cervical cancer. Unsupervised clustering of DNA methylation data, with the heatmap showing beta values ordered by DNA methylation clusters. **(C,D)** Cluster-of-clusters analysis (CoCA) clustering for subgroup identification.

92 genes. Functional enrichment analysis showed these genes were involved in Rap1 pathway, hypoxia-inducible factor (HIF)-1 pathway, and cell adhesion (**Figure 5**).

Moreover, after analyzing the mutation types among the four subtypes, the results showed that all these four subtypes were mainly C > T mutation and the conversion ratio was generally higher than the transversion ratio (**Figure 6A**). Mutually exclusive or co-occurring events were determined by Fisher exact test, and there were more co-mutated genes in cluster 3 and no exclusive mutations were detected in all subpopulations (**Figure 6B**). APOBEC enrichment analysis showed that the majority samples were APOBEC enriched samples (**Figure 6C**). Further signature analysis showed that signatures 1, 2, and 13 were involved in clusters 1, 2, and 4, and signatures 6 and 10 were involved in cluster 3 (**Figure 6D**).

With respect to CNV, seven deletion regions and 22 amplification regions were identified, showing significant differences across clusters. Both CNV samples and CNV values in clusters 2 and 3 were less compared with those of clusters 1 and 4 (**Figure 7A**), suggesting that the main factor promoting tumor in clusters 2 and 3 was not CNV but mutation. Tumor purity and tumor ploidy were analyzed by using ABSOLUTE. As described in **Figure 7B**, tumor purity showed no difference among subgroups, whereas tumor ploidy showed a difference between cluster 1 (mean = 3.75) and cluster 3 (mean = 3.32) and between cluster 2 (mean = 3.80) and cluster 3 (mean = 3.32). With respect to fusion gene detection, 5UTR-3UTR was only in cluster 2 (**Figure 7C**), and CDS-3UTR was only in clusters 1 and 4. Thus, fusion genes varied in different clusters.

Identification of Drive Mutation

As it is both clinically important and challenging to distinguish high-risk cervical cancer patients with poor progression and prognosis, we sought to identify molecular features associated with poor prognosis. Combining the above multidimensional datasets, a series of genes associated with poor prognosis was identified, including 77 genes in cluster 1, 17 genes in cluster 2, 92 genes in cluster 3, and 20 genes in cluster 4. Further Mut2sigC analysis finally identified a total of 10 unique driver genes, including GPR107, CHRNA5, ZBTB20, Rb1, NCAPH2, SCA1, SLC25A5, RBPMS, DDX3X, and H2BFM.

DISCUSSION

Previous studies have implicated somatic mutations in PIK3CA, TP53, STK11, EP300, FBXW7, and HLA-B in the pathogenesis of

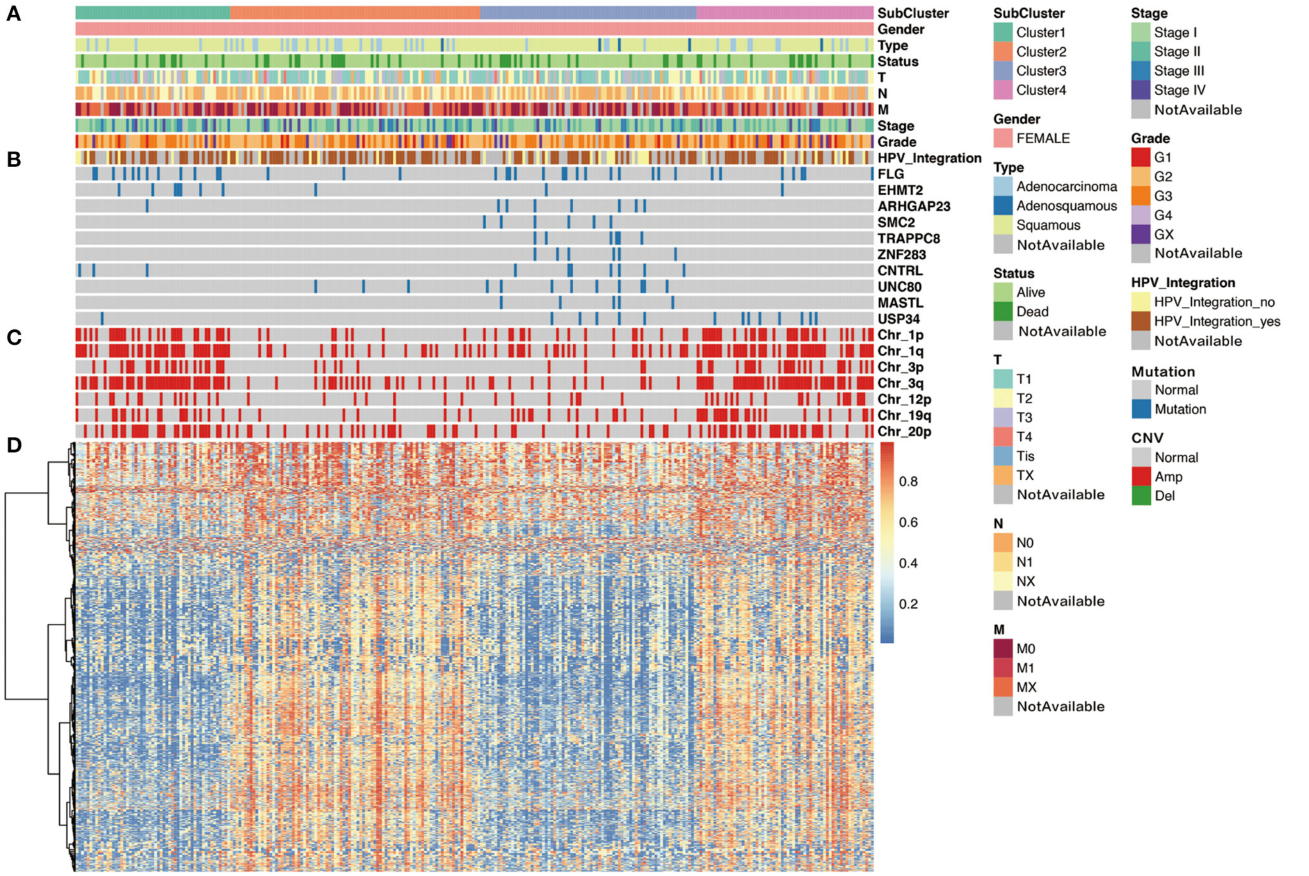


FIGURE 5 | The cluster-of-clusters analysis separated 276 cervical cancers into four clusters. Upper covariate tracks show **(A)** clinical characteristics; **(B)** mutations in top 10 different mutated genes across four clusters; and **(C)** copy number variation (CNV) in 1p, 1q, 3q, 3p, 12p, 19q, and 20p. **(D)** The heatmap shows methylation in cervical cancers.

cervical carcinomas (Ojesina et al., 2014; Bager et al., 2015). As expected, in the current study, recurrent mutations in PIK3CA, EP300, and FBXW7 were presented in 32, 12, and 7% cervical patients, respectively, consistent with similar findings in previous reports (Ojesina et al., 2014). In addition, we found significantly recurrent mutations in TTN (33%), MUC4 (31%), and MUC16 (19%), here reported for the first time, to our knowledge, in cervical carcinomas. The most frequently mutated gene in the current study is titin (TTN). The 364 exon TTN gene encodes TTN, the largest known protein, playing key structural, developmental, mechanical, and regulatory roles in cardiac and skeletal muscles (Gerull et al., 2002; Chauveau et al., 2014). Missense mutation of TTN was detected in 85% lung squamous cell carcinoma and predicted a favorable prognosis of these diseases (Cheng et al., 2019). More recently, TTN mutation was reported to predict an increased tumor mutational burden, a beneficial response to immune checkpoint blockade treatment, and a long survival among pan-solid tumors, including cervical cancer (Jia et al., 2019). MUC4, a transmembrane glycoprotein, was involved in many different biological processes such as cell proliferation, cell death, invasion, and metastasis (Singh et al., 2007). MUC4 was activated during the process of cervical squamous dysplastic transformation (Lopez-Ferrer et al., 2001), aberrantly expressed in cervical cancer (Munro et al., 2009), and associated with lymph node metastasis (Munro et al., 2009). Abrogation of MUC4 expression reduces invasion and the mesenchymal properties of cervical cancer cells (Xu et al., 2017). We observed MUC16 mutation in our dataset, similar to recent reports in gastric cancers (Li et al., 2018b). Therefore, the recurrent site-specific TTN and MUC4 mutations and the known role of these genes in cancer suggest the possibility that mutant TTN and MUC4 may exert oncogenic activity in cervical cancer. Further validation of these results is required in the future, especially the predictive role of TTN in cervical cancer immunotherapy response.

Pathway analyses revealed that the most significantly mutated gene set in cervical cancer involved a deficiency of DNA mismatch repair, APOBEC-cytidine deaminase, and spontaneous deamination of 5-methyl cytosine. Previous study has described deficient DNA mismatch repair as a common phenomenon in the process of cervical cancer development (Nijhuis et al., 2007; Feng et al., 2018). APOBEC-cytosine deaminase activity has recently

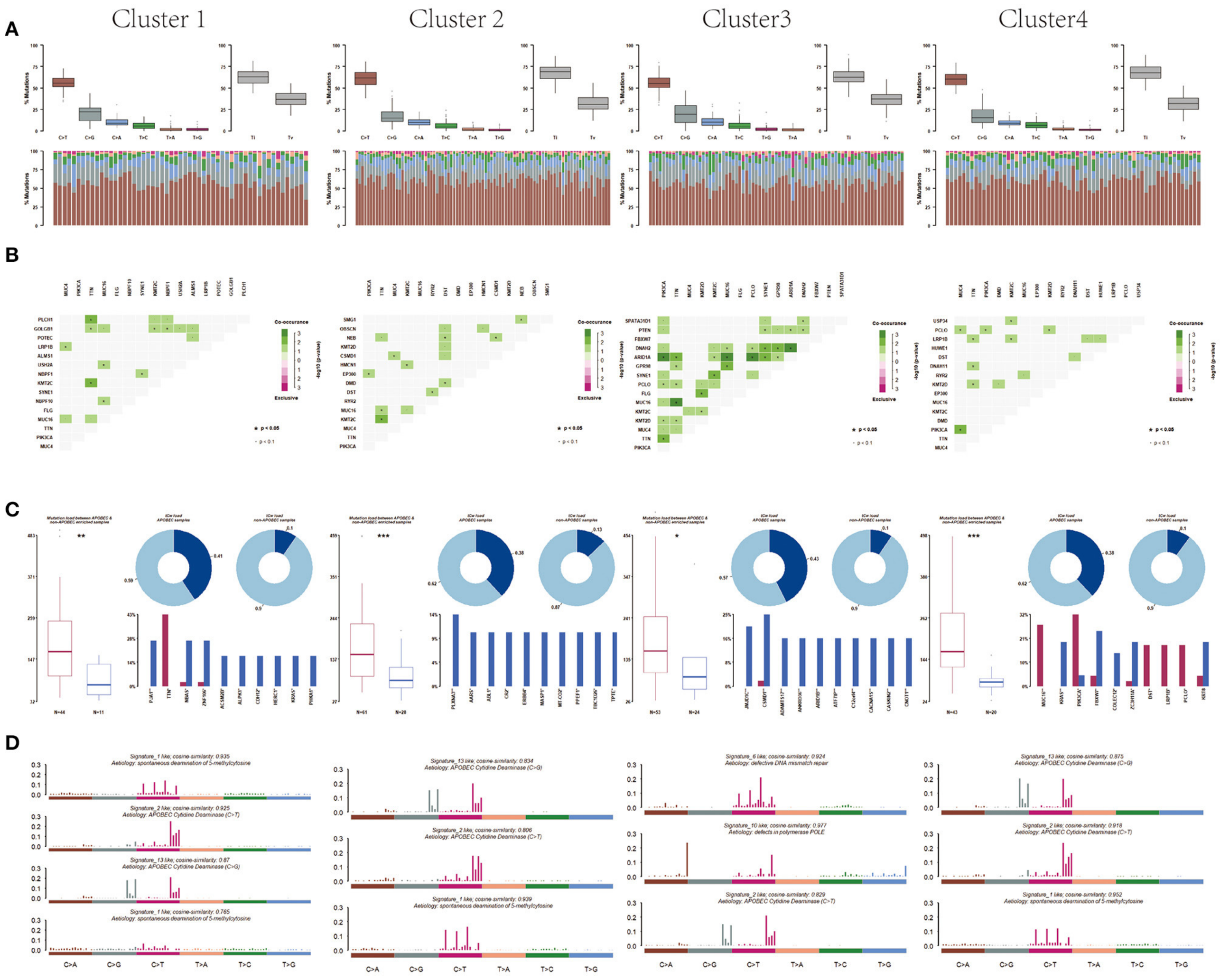


FIGURE 6 | Differentiation between clusters including **(A)** mutation, **(B)** co-mutated genes, **(C)** APOBEC enriched samples, and **(D)** mutation signature analysis.

emerged as a significant mutagenic factor in human cancer. APOBEC activity served as a key driver of PIK3CA mutagenesis and HPV-induced transformation in head and neck squamous cell carcinomas (Henderson et al., 2014). Moreover, APOBEC cytidine deaminase mutagenesis pattern has been detected in human cervical cancer (Roberts et al., 2013). Our current results further support the concept that deficient DNA mismatch repair and APOBEC-mediated mutagenesis were carcinogenic in the cervix.

CNV is a very common phenomenon and contributes to gene transcript expression in cervical cancer (Dellas et al., 2003; Narayan et al., 2007; Yan et al., 2017). In our genome-wide CNV analysis, the most prevalent gains are detected at the 3q26.31 and 11q22.1, while the most frequent deletions are at 11q24.2 and 2q37.2, consistent with previous reports (Rao et al., 2004; Narayan et al., 2007). These observations further suggest genomic disorder was a common phenomenon in the development of cervical cancer.

Molecular classification may prove more clinically impactful compared to traditional histopathological classifications in terms of treatment predictions and predicting patient prognosis. Based on the above comprehensive genetic alterations, using a "cluster-of-clusters" analytic approach, we identified four major genomic subtypes of cervical cancer. Cluster 2 was enriched in methylation and poor in CNV. HPV integration was most enriched in cluster 2 with lots of overexpressed genes. Rb-1 was detected as the driver mutation in this subgroup, suggesting that HPV integration unregulated lots of genes *via* methylation, especially the driver gene Rb-1, abrogated cell cycle arrest, and stimulated proliferation in cervical cancer. More recently, cervical cancer with Rb1 mutation is reported to be more sensitive to cisplatin through PI3K/AKT pathway. Cluster 3 was characterized by poor CNV and poor methylation, most of which were squamous carcinoma. In this subgroup, co-mutations were common events. NCAPH2, SCA1, and SLC25A5 were identified as driver mutations.

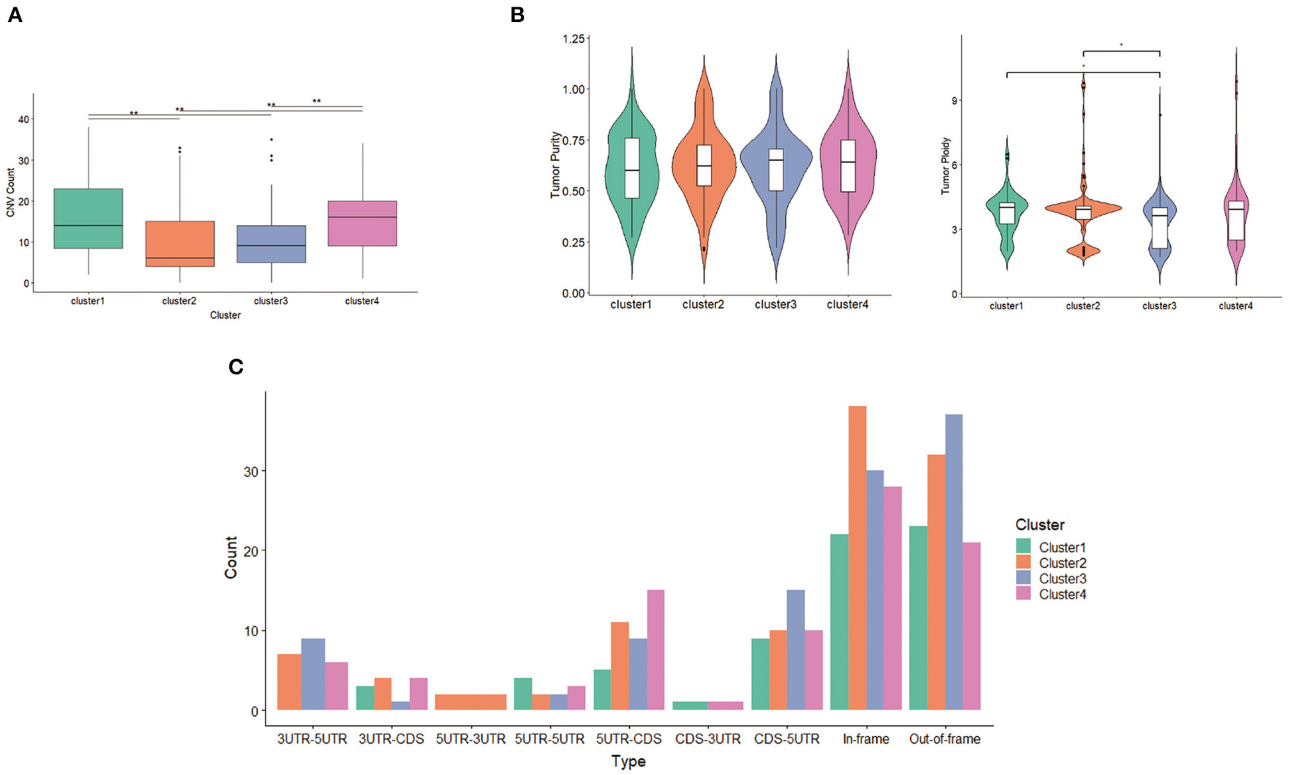


FIGURE 7 | Differentiation between clusters including **(A)** copy number variation (CNV) counts, **(B)** tumor purity and tumor ploidy, and **(C)** fusion gene types.

Cluster 4 was enriched both for CNV and methylation. In this subgroup, RBPMS,DDX3X和H2BFM were identified as driver mutations.

Our study represents the first integrated multidimensional molecular and computational investigation of somatic mutations in cervical cancer, which strongly complements previous gene- and pathway-focused studies. Cervical cancer is a heterogenous disease likely driven by multiple genomic disorders. We tried to elucidate the driver gene(s) and potential molecular subtypes of cervical cancer by using a public database. In the current study, we integrated multi-omics data including somatic mutation, CNV, DNA methylation, and miRNA profile, depicted a comprehensive genomic landscape of cervical cancer, and then performed molecular classification, finally identifying driver genes, such as GPR107, ZBTB20, NCAPH2, and SLC25A5. These results contribute to the identification of clinically important biomarkers and potential treatment targets. However, this paper also has some limitations. Firstly, majority samples of selected cohorts were confirmed as squamous cancers, limited numbers of different histologic types and para-cancer tissues working as control, which might bring bias into the classification process. As for the unsupervised classification, we used COCA, a two-step approach, to build binary matrix from multiple omics, and then returned a global clustering structure. The algorithm COCA was first introduced in TCGA network (2012), combining and summarizing the clustering structures, even if the original datasets (level 1/2) are unavailable to the public. Yet, we should notice that the first step combination of such clustering structures from each dataset is unweighted, which might make the output of the algorithm sensitive to the inclusion of poor-quality datasets. Therefore, biologic functions of these driver genes in cervical cancer remain to be verified, which is now under further exploration.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

HZ and QH conceived and designed the experiments and revised the manuscript. YX and HL performed the experiments. HZ analyzed the data and wrote the paper. QH contributed reagents/materials/analysis tools and revised the manuscript. All authors have read and approved the final manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.598304/full#supplementary-material

REFERENCES

Bager, P., Wohlfahrt, J., Sorensen, E., Ullum, H., Hogdall, C. K., Palle, C., et al. (2015). Common filaggrin gene mutations and risk of cervical cancer. *Acta Oncol*. 54, 217–223. doi: 10.3109/0284186X.2014.973613

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin*. 68, 394–424. doi: 10.3322/caac.21492

Chauveau, C., Rowell, J., and Ferreiro, A. (2014). A rising titan: TTN review and mutation update. *Hum. Mutat*. 35, 1046–1059. doi: 10.1002/humu.22611

Chen, D., Juko-Pecirep, I., Hammer, J., Ivansson, E., Enroth, S., Gustavsson, I., et al. (2013). Genome-wide association study of susceptibility loci for cervical cancer. *J. Natl. Cancer Inst*. 105, 624–633. doi: 10.1093/jnci/djt051

Chen, F., Zhang, Y., Senbabaoglu, Y., Ciriello, G., Yang, L., Reznik, E., et al. (2016). Multilevel genomics-based taxonomy of renal cell carcinoma. *Cell Rep*. 14, 2476–2489. doi: 10.1016/j.celrep.2016.02.024

Cheng, X., Yin, H., Fu, J., Chen, C., An, J., Guan, J., et al. (2019). Aggregate analysis based on TCGA: TTN missense mutation correlates with favorable prognosis in lung squamous cell carcinoma. *J. Cancer Res. Clin. Oncol*. 145, 1027–1035. doi: 10.1007/s00432-019-02861-y

Cheung, T. H., Man, K. N., Yu, M. Y., Yim, S. F., Siu, N. S., Lo, K. W., et al. (2012). Dysregulated microRNAs in the pathogenesis and progression of cervical neoplasm. *Cell Cycle* 11, 2876–2884. doi: 10.4161/cc.21278

Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., et al. (2016). TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res*. 44:e71. doi: 10.1093/nar/gkv1507

Dellas, A., Torhorst, J., Gaudenz, R., Mihatsch, M. J., and Moch, H. (2003). DNA copy number changes in cervical adenocarcinoma. *Clin. Cancer Res*. 9, 2985–2991.

Feng, Y. C., Ji, W. L., Yue, N., Huang, Y. C., and Ma, X. M. (2018). The relationship between the PD-1/PD-L1 pathway and DNA mismatch repair in cervical cancer and its clinical significance. *Cancer Manag. Res*. 10, 105–113. doi: 10.2147/CMAR.S152232

Ganly, I., Makarov, V., Deraje, S., Dong, Y., Reznik, E., Seshan, V., et al. (2018). Integrated genomic analysis of hurthle cell cancer reveals oncogenic drivers, recurrent mitochondrial mutations, and unique chromosomal landscapes. *Cancer Cell* 34, 256–270. doi: 10.1016/j.ccell.2018.07.002

Gehring, J. S., Fischer, B., Lawrence, M., and Huber, W. (2015). SomaticSignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics* 31, 3673–3675. doi: 10.1093/bioinformatics/btv408

Gerull, B., Gramlich, M., Atherton, J., McNabb, M., Trombitas, K., Sasse-Klaassen, S., et al. (2002). Mutations of TTN, encoding the giant muscle filament titin, cause familial dilated cardiomyopathy. *Nat. Genet*. 30, 201–204. doi: 10.1038/ng815

Henderson, S., Chakravarthy, A., Su, X., Boshoff, C., and Fenton, T. R. (2014). APOBEC-mediated cytosine deamination links PIK3CA helical domain mutations to human papillomavirus-driven tumor development. *Cell Rep*. 7, 1833–1841. doi: 10.1016/j.celrep.2014.05.012

Hoadley, K. A., Yau, C., Wolf, D. M., Cherniack, A. D., Tamborero, D., Ng, S., et al. (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* 158, 929–944. doi: 10.1016/j.cell.2014.06.049

Jia, Q., Wang, J., He, N., He, J., and Zhu, B. (2019). Titin mutation associated with responsiveness to checkpoint blockades in solid tumors. *JCI Insight* 4:127901. doi: 10.1172/jci.insight.127901

Li, X., Pasche, B., Zhang, W., and Chen, K. (2018b). Association of MUC16 mutation with tumor mutation load and outcomes in patients with gastric cancer. *JAMA Oncol*. 4, 1691–1698. doi: 10.1001/jamaoncol.2018.2805

Li, X., Xu, W., Kang, W., Wong, S. H., Wang, M., Zhou, Y., et al. (2018a). Genomic analysis of liver cancer unveils novel driver genes and distinct prognostic features. *Theranostics* 8, 1740–1751. doi: 10.7150/thno.22010

Litwin, T. R., Clarke, M. A., Dean, M., and Wentzensen, N. (2017). Somatic host cell alterations in HPV carcinogenesis. *Viruses* 9:206. doi: 10.3390/v9080206

Lopez-Ferrer, A., Alameda, F., Barranco, C., Garrido, M., and de Bolos, C. (2001). MUC4 expression is increased in dysplastic cervical disorders. *Hum. Pathol*. 32, 1197–1202. doi: 10.1053/hupa.2001.28938

Mayakonda, A., Lin, D. C., Assenov, Y., Plass, C., and Koeffler, H. P. (2018). Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res*. 28, 1747–1756. doi: 10.1101/gr.239244.118

Moody, C. A., and Laimins, L. A. (2010). Human papillomavirus oncoproteins: pathways to transformation. *Nat. Rev. Cancer* 10, 550–560. doi: 10.1038/nrc2886

Munro, E. G., Jain, M., Oliva, E., Kamal, N., Lele, S. M., Lynch, M. P., et al. (2009). Upregulation of MUC4 in cervical squamous cell carcinoma: pathologic significance. *Int. J. Gynecol. Pathol*. 28, 127–133. doi: 10.1097/PGP.0b013e318184f3e0

Narayan, G., Bourdon, V., Chaganti, S., Arias-Pulido, H., Nandula, S. V., Rao, P. H., et al. (2007). Gene dosage alterations revealed by cDNA microarray analysis in cervical cancer: identification of candidate amplified and overexpressed genes. *Genes Chromosomes Cancer* 46, 373–384. doi: 10.1002/gcc.20418

Nehrt, N. L., Peterson, T. A., Park, D., and Kann, M. G. (2012). Domain landscapes of somatic mutations in cancer. *BMC Genomics* 13:S9. doi: 10.1186/1471-2164-13-S4-S9

Nijhuis, E. R., Nijman, H. W., Oien, K. A., Bell, A., ten Hoor, K. A., Reesink-Peters, N., et al. (2007). Loss of MSH2 protein expression is a risk factor in early stage cervical cancer. *J. Clin. Pathol*. 60, 824–830. doi: 10.1136/jcp.2005.036038

Ojesina, A. I., Lichtenstein, L., Freeman, S. S., Pedamallu, C. S., Imaz-Rosshandler, I., Pugh, T. J., et al. (2014). Landscape of genomic alterations in cervical carcinomas. *Nature* 506, 371–375. doi: 10.1038/nature12881

Rao, P. H., Arias-Pulido, H., Lu, X. Y., Harris, C. P., Vargas, H., Zhang, F. F., et al. (2004). Chromosomal amplifications, 3q gain and deletions of 2q33-q37 are the frequent genetic changes in cervical carcinoma. *BMC Cancer* 4:5. doi: 10.1186/1471-2407-4-5

Reich, M., Liefeld, T., Gould, J., Lerner, J., Tamayo, P., and Mesirov, J. P. (2006). GenePattern 2.0. *Nat. Genet*. 38, 500–501. doi: 10.1038/ng0506-500

Roberts, S. A., Lawrence, M. S., Klimczak, L. J., Grimm, S. A., Fargo, D., Stojanov, P., et al. (2013). An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat. Genet*. 45, 970–976. doi: 10.1038/ng.2702

Singh, A. P., Chaturvedi, P., and Batra, S. K. (2007). Emerging roles of MUC4 in cancer: a novel target for diagnosis and therapy. *Cancer Res*. 67, 433–436. doi: 10.1158/0008-5472.CAN-06-3114

Verlaat, W., Snijders, P. J. F., Novianti, P. W., Wilting, S. M., De Strooper, L. M. A., Trooskens, G., et al. (2017). Genome-wide DNA methylation profiling reveals methylation markers associated with 3q gain for detection of cervical precancer and cancer. *Clin. Cancer Res*. 23, 3813–3822. doi: 10.1158/1078-0432.CCR-16-2641

Xu, D., Liu, S., Zhang, L., and Song, L. (2017). MiR-211 inhibits invasion and epithelial-to-mesenchymal transition (EMT) of cervical cancer cells via targeting MUC4. *Biochem. Biophys. Res. Commun*. 485, 556–562. doi: 10.1016/j.bbrc.2016.12.020

Yan, D., Yi, S., Chiu, W. C., Qin, L. G., Kin, W. H., Kwok Hung, C. T., et al. (2017). Integrated analysis of chromosome copy number variation and gene expression in cervical carcinoma. *Oncotarget* 8, 108912–108922. doi: 10.18632/oncotarget.22403

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

ORIGINAL RESEARCH
published: 24 February 2021
doi: 10.3389/fgene.2021.632620

Comparison of Methods for Feature Selection in Clustering of High-Dimensional RNA-Sequencing Data to Identify Cancer Subtypes

David Källberg[1,2†], *Linda Vidman*[2,3†] *and Patrik Rydén*[2*]

[1] Department of Statistics, USBE, Umeå University, Umeå, Sweden, [2] Department of Mathematics and Mathematical Statistics, Umeå University, Umeå, Sweden, [3] Department of Radiation Sciences, Oncology, Umeå University, Umeå, Sweden

Cancer subtype identification is important to facilitate cancer diagnosis and select effective treatments. Clustering of cancer patients based on high-dimensional RNA-sequencing data can be used to detect novel subtypes, but only a subset of the features (e.g., genes) contains information related to the cancer subtype. Therefore, it is reasonable to assume that the clustering should be based on a set of carefully selected features rather than all features. Several feature selection methods have been proposed, but how and when to use these methods are still poorly understood. Thirteen feature selection methods were evaluated on four human cancer data sets, all with known subtypes (gold standards), which were only used for evaluation. The methods were characterized by considering mean expression and standard deviation (SD) of the selected genes, the overlap with other methods and their clustering performance, obtained comparing the clustering result with the gold standard using the adjusted Rand index (ARI). The results were compared to a supervised approach as a positive control and two negative controls in which either a random selection of genes or all genes were included. For all data sets, the best feature selection approach outperformed the negative control and for two data sets the gain was substantial with ARI increasing from (−0.01, 0.39) to (0.66, 0.72), respectively. No feature selection method completely outperformed the others but using the dip-rest statistic to select 1000 genes was overall a good choice. The commonly used approach, where genes with the highest SDs are selected, did not perform well in our study.

Keywords: feature selection, gene selection, RNA-seq, cancer subtypes, high-dimensional

OPEN ACCESS

Edited by:
Shaolong Cao,
University of Texas MD Anderson Cancer Center, United States

Reviewed by:
Jonathan Greenbaum,
Tulane University, United States
Fangtang Yu,
Hinda and Arthur Marcus Institute for Aging Research, United States

Correspondence:
Patrik Rydén
patrik.ryden@umu.se

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *23 November 2020*
Accepted: *03 February 2021*
Published: *24 February 2021*

Citation:
Källberg D, Vidman L and Rydén P (2021) Comparison of Methods for Feature Selection in Clustering of High-Dimensional RNA-Sequencing Data to Identify Cancer Subtypes. Front. Genet. 12:632620. doi: 10.3389/fgene.2021.632620

INTRODUCTION

The human genome consists of around 21,000 protein coding genes (Pertea et al., 2018). By analyzing genes using high-throughput technologies (e.g., sequence and microarray technologies), researchers get access to huge amount of data that can be of relevance for prognosis of a disease (classification), identification of novel disease subtypes (cluster analysis) and detection of differentially expressed genes. Aside from the fact that the number of genes often far exceeds the number of samples, most features (i.e., genes) contain no information related to the trait of interest. Whether the aim is to distinguish between different tumor stages or identifying new disease subtypes, the identification of discriminating features is key.

Diseases like cancer arise by various causes and there is reason to believe that today's cancer diseases can be divided further into several subtypes, which potentially should be treated differently. Cluster analysis applied on gene expression data from samples (e.g., tumor samples or blood samples) taken from cancer patients has successfully been used to detect novel cancer subtypes (Eisen et al., 1998; Sotiriou et al., 2003; Lapointe et al., 2004; Bertucci et al., 2005; Fujikado et al., 2006; Ren et al., 2016). However, the problem of detecting new subtypes is challenging since most of the genes' expressions are not affected by disease subtype and some genes are influenced by other factors such as gender, age, diet, presence of infections and previous treatments. Ideally, a cluster analysis aimed at detecting novel disease subtypes should only utilize genes that are informative for the task, i.e., genes that have their expression mainly governed by which disease subtype the patient has. Hence, it is of interest to apply some sort of gene selection procedure prior to the cluster analysis. This task would be relatively easy if it was known which subtypes (i.e., labels) the patients have, but for unsupervised classification problems, the labels are unknown making gene selection a true challenge. When the labels are unknown, statistical tests such as t-tests, Wilcoxon rank sum tests or one-way ANOVA cannot be used to identify differentially expressed genes. Instead, other data characteristics need to be considered. For example, a common approach to discover subgroups in high-dimensional genomic data is to apply clustering on a subset of features that are selected based on their standard deviation (SD) across samples (Bentink et al., 2012; Kim et al., 2020; Shen et al., 2020). Thus, the SD is used as a score that measures how informative a gene is for the underlying subgroups. Here we also consider a set of alternative scores for selecting informative genes, i.e., genes affected by the subtype. Aside SD, other examples within the category of variability scores include, e.g., the interquartile range (IQR) and measures based on entropy (Liu et al., 2005; Seal et al., 2016). If instead it is assumed that informative genes are likely to be expressed at a relatively high level it makes sense to select highly expressed genes. Another class of measures is based on quantifying the extent to which the gene expression distribution can be described by two or more relatively distinct peaks, or modes, which represent different subtypes. In the simplest case, we assume that the tumor samples can be divided into two subtypes. Given that this assumption is true, the gene expression of an informative gene may have a bimodal distribution. By ranking genes according to some bimodality measure and including only the top scoring genes (i.e., the genes with the highest bimodality measures), it is possible to remove uninformative and redundant genes before performing clustering. Several gene selection procedures based on bimodality have been proposed (Moody et al., 2019), including the bimodality index (BI; Wang et al., 2009), the bimodality coefficient (BC; SAS Institute, 1990) and various variants of the variance reduction score (VRS; Bezdek, 1981; Hellwig et al., 2010). A more general approach is to search for genes with an apparent multimodal distribution. The dip-test suggested by Hartigan and Hartigan (1985) addresses this problem.

It may be argued that genes that are involved in the same biological processes should have similar expression profiles across samples (Wang et al., 2014). Under the assumption that a fair number of genes are affected by the disease subtype, it is natural to search for a large set of genes that are highly correlated. In the established taxonomy for feature selection approaches, the methods studied here are *filtering* methods, other important classes are wrapper, embedded, and hybrid methods thereof (Ang et al., 2016).

It is evident that fundamentally different selection procedures will identify different sets of genes. Moreover, several of the approaches are likely to include not only informative genes but also genes affected by other factors and genes that have general inclusion properties (e.g., genes with highly variable gene-expressions). In the worst-case scenario, a gene selection may fail to identify genes associated with the subtype partition of interest. This risk is particularly relevant if the influence of the disease subtype is weak compared to other factors. Hence, gene selection can have a negative influence on the clustering performance.

Multiple studies have compared feature selection methods where the ultimate goal is to classify patients according to some disease status. Arun Kumar et al. (2017) compared feature selection algorithms based on execution time, number of selected features and classification accuracy in two microarray gene expression data sets. Abusamra (2013) compared eight feature selection methods on two publicly available microarray gene expression data sets of glioma and found that no single method outperformed the others. Cilia et al. (2019) concluded that the feature selection process plays a key role in disease classification and that a reduced feature set significantly improved classification, but no selection method had a superior performance in all data sets. Much fewer studies have compared feature selection methods where the objective is detection of novel subgroups using clustering (Freyhult et al., 2010).

Here we focus on evaluating and comparing means of selecting informative genes in high-dimensional RNA-seq data from human cancers before performing cluster analysis for identification of subtypes. The study is extensive and evaluates 13 gene selection procedures on four human cancer tumor types, each with two known subtypes. The approaches are compared to two negative controls (including all genes or a set of randomly selected genes) and a positive control (genes selected using label information). We study the performance of the methods, properties of the selected genes and overlap between sets of selected genes. We also investigate how the performance changes when the relative distribution of the subtypes is altered.

MATERIALS AND METHODS

Data

Experimental RNA-sequencing raw count data from the TCGA-database were obtained through Broad institute GDAC Firehose[1]. Four different cancer types with known subgroups were used in the analyses: breast (BRCA), kidney (KIRP), stomach (STAD),

[1] https://gdac.broadinstitute.org

and brain (LGG) cancer. In all evaluations we treated the defined cancer subtypes as gold standard partitions, although there exist several ways of grouping the data.

The *Brain data* (denoted LGG by TCGA) consists of data from 226 tumor samples from patients with lower grade glioma, where 85 patients had the IDH mutation and 1p/19q co-deletion (IDHmut-codel) while the remaining 141 patients had the IDH mutation without the 1p/19q co-deletion (IDHMut-NOcodel)(Brat et al., 2015). The *Breast data* (BRCA by TCGA) consists of data from 929 tumor samples from patients with breast invasive carcinoma (BRCA), where 216 patients had negative Estrogen Receptor status (ER−) while the remaining 713 patients had positive ER status (ER+). The *Kidney data* (KIRP by TCGA) includes data from tumor samples from 150 patients with kidney renal papillary cell carcinoma (KIRP), where 73 patients were histologically determined as subtype 1 and the remaining 77 samples were determined as subtype 2 (The Cancer Genome Atlas Research Network, 2016). The *Stomach data* was obtained from tumors in 178 patients with stomach adenocarcinoma (STAD), where 55 patients had microsatellite instability (MSI) tumors and the remaining 123 patients had tumors with chromosomal instability (CIN) (Cancer Genome Atlas Research Network, 2014).

Clustering of Samples

Raw gene level count data were obtained from the TCGA-database, i.e., an integer value was observed for each sample and gene. First, the raw data were pre-processed, including initial filtration, between sample normalization and applying a variance stabilizing transformation, see section "Pre-processing" for further details. A variety of gene selection approaches were applied to the pre-processed data, see section "Selection of Informative Genes". Hierarchical clustering using Ward's linkage and the Euclidean distance was performed on the selected genes. In addition, k-means ($k = 2$) clustering (Hartigan and Wong, 1979) and hierarchical clustering using Ward's linkage and a correlation-based distance (i.e., 1-$|\rho|$, where ρ is the Spearmans correlation coefficient) were performed in some selected cases. The two major groups identified by the clustering algorithm defined a binary sample partition that was compared to our gold standard partition (i.e., the partition defined by the considered subgroups), see section "Evaluations".

Simulation Study

Prior to analyzing the cancer data, a small simulation study was conducted to understand if inclusion of non-informative features (here defined as features with identically distributed feature values) has a negative effect on the clustering performance. Data from 100 samples (50 labeled A and 50 labeled B) with 10,000 features were simulated. Here, 100 features were informative such that the A-values were simulated from a normal distribution with mean 0 and variance 1 [i.e., N(0,1)] and the B-values were simulated from N(1,1). All the non-informative values were simulated from N(0,1). Hierarchical clustering using Ward's linkage and the Euclidean distance was performed on: all features, only the 100 informative features and the k features with the highest SD, $k = 100, 200, \ldots, 10{,}000$. The simulations were repeated 40 times. For each clustering, the performance was measured using the adjusted Rand index (ARI) (Hubert and Arabie, 1985), where the clustering result was compared to the AB-partition.

Pre-processing

All four data sets originally contained gene expression for 20,531 genes. As a first step in finding informative genes, we excluded genes expressed at low levels. A score was constructed for each gene by counting the number of samples with expression values below the 25th gene percentile (i.e., the expression value below which 25% of the genes in a sample can be found). The 25% of the genes with highest score were filtered out. Next, the R-package *DESeq2* (Love et al., 2014) was used for between sample normalization using the standard settings. Finally, the normalized data was transformed using a variance-stabilizing transform (VST), which conceptually takes a given variance-mean relation $\sigma^2 = var(x) = h(\mu)$ and transforms the data according to

$$y(x) = \int^x \frac{1}{\sqrt{h(\mu)}} d\mu.$$

We used the VST implemented in the R-package *DESeq2*, a model-based approach that relies on the variance-mean relation implied by a negative binomial distribution for the gene expression count data. The choice of transformation approach was motivated by properties of the clustering method, which often yields best results for (approximately) homoscedastic data, meaning that the variance of the variable, such as gene expression, does not depend on the mean. For RNA-seq count data, however, the variance typically increases with the mean. The commonly used procedure to handle this is to apply a logarithmic transform to the normalized count values after adding a small pseudo count. Unfortunately, now genes with low counts have a tendency to dominate the clustering result since they give the strongest signals in terms of relative difference between samples.

Selection of Informative Genes

Our focus was to study how gene selection affects the clustering performance. For the considered data sets there are two "true" clusters defined by our gold standards. For supervised problems, where the class labels of the samples are known, feature selection is done by identifying a set of informative genes, in the simplest case, by applying a two-sample t-test to each gene and select the genes with the lowest p-values (Önskog et al., 2011). For cluster analysis problems, it can be argued that removing "non-informative" genes prior to the clustering will increase the clustering performance (Freyhult et al., 2010). Feature selection for cluster analysis is difficult for two reasons: (a) the sample labels are unknown and cannot be used to select informative genes, (b) in contrast to supervised classification it is not possible to use performance measures (e.g., error rates in classification) to compare and choose the best feature selection approach for the considered clustering problem.

We evaluated 13 different methods used for gene selection, where some are commonly used while others were included because they constitute principally different approaches. The methods ranked all genes based on how informative they were predicted to be, and the top ranked genes (100, 1000, or 3000 genes) were used in the downstream clustering. Hence, altogether 39 gene selection approaches were applied to the four data sets and evaluated against the gold standard.

The considered feature selection methods are motivated by fundamentally different ideas, which were used to group the methods into four categories. The four principles include selecting highly expressed genes, highly variably genes, highly correlated genes and genes with bi- or multimodal profiles. Below we give a general motivation behind the selection procedures within each group and a detailed description of the included methods.

One idea is to select genes with overall high expression values. Discriminating between disease subtypes can be difficult when the level of noise is high compared to the mean expression values, which makes it easier to detect differentially expressed genes among highly expressed genes. In this category of methods, we included the methods mean value (M) and third quartile (Q3).

Another group of methods is based on the spread of gene expression values across samples. Genes with large variability can contain interesting variations caused by disease subtype. In this category, we included the SD, the IQR, and the quadratic Rényi entropy (ENT).

A category involving correlation of genes includes a technique called co-expression (CoEx1) and a modified version (CoEx2). Tumor cells are under constant attack by the immune system and to survive, the genes must coordinate against the threat. Genes that are highly correlated to other genes may be involved in the same exposed networks and is therefore of interest as potential biomarkers. Co-expression among informative genes has been used for variable selection in clustering problems for high-dimensional microarray data (Wang et al., 2014).

Six of the methods studied in this article are based on the idea of modality. For informative genes, the distribution of gene expression among patients with different cancer subtypes can be expected to differ. It is therefore of interest to identify genes that have an expression distribution with more than one peak. We included the so-called dip-test (DIP), a method that identifies genes with multimodal distributions. In addition, we considered five methods that identify genes with bimodal distributions. We included the parametric method called the BI and four non-parametric methods: VRS, weighted variance reduction score (wVRS), modified variance reduction score (mVRS), and BC. The relationship between the considered gene selection methods is summarized in **Figure 1**.

The Mean Value Selection (M)

The mean value was calculated for each gene over all samples and the highest expressed genes were included in the analyses.

Third Quartile Selection (Q3)

Genes were arranged according to decreasing values of the third quartile and the genes with the highest Q3 values were selected.

The Standard Deviation Selection (SD)

The SD was calculated for each gene and the genes with the highest SDs were selected.

The Interquartile Range Selection (IQR)

The distance between the first and third quartile was calculated for each gene and genes with large distances were selected.

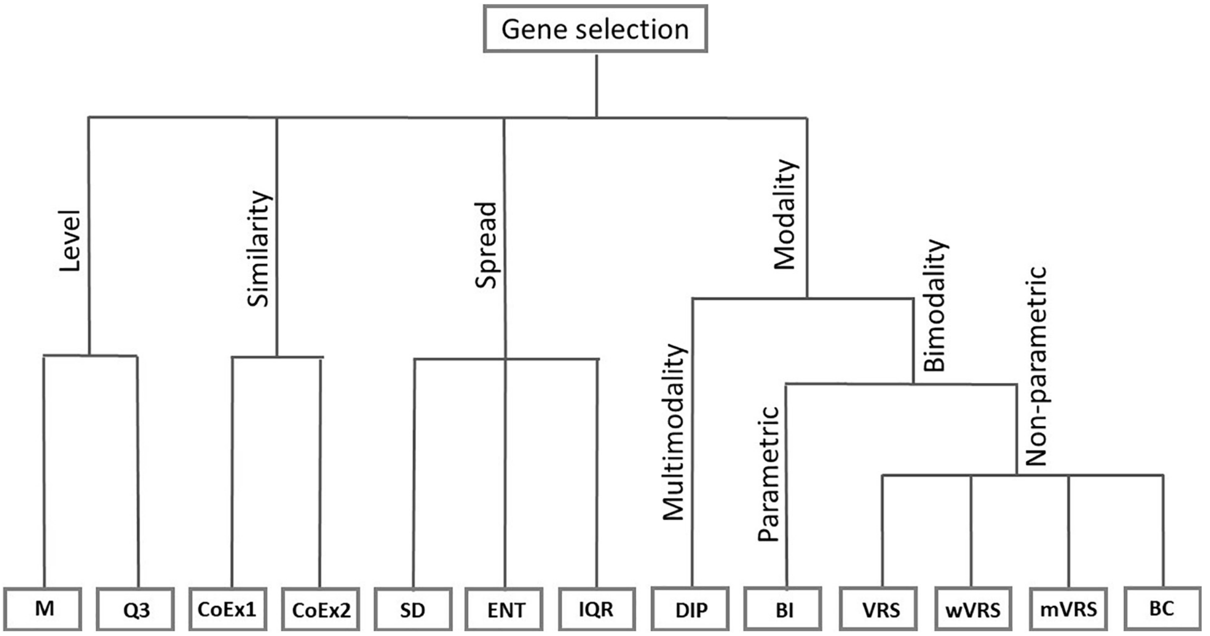


FIGURE 1 | Feature selection methods divided into groups based on their properties. The included methods are mean value (M), third quartile (Q3), co-expression (CoEx1), modified co-expression (CoEx2), standard deviation (SD), interquartile range (IQR), entropy estimator (ENT), dip-test statistic (DIP), bimodality index (BI), variance reduction score (VRS), weighted variance reduction score (wVRS), modified variance reduction score (mVRS). and bimodality coefficient (BC).

The Entropy Estimator Selection (ENT)

Entropy is an alternative to SD for measuring variability in gene expression across samples. Assuming the observed values $x_{1,}x_{2,\dots,}x_n$ for a gene can be described by a distribution with density $f(x)$, its quadratic Rényi entropy is defined as:

$$H_2(X) = -\log\left(\int f(x)^2 dx\right).$$

To estimate this parameter, we use a non-parametric kernel-estimator (Gine and Nickl, 2008), obtained as

$$\text{ENT} = -\log\left(\frac{2}{n(n-1)h}\sum_{i<j} K\left(\frac{x_i - x_j}{h}\right)\right).$$

The user specifies the kernel function $K(\cdot)$ and the bandwidth h. Here we employed the rectangular kernel, and for h we applied Silverman's rule-of-thumb for kernel-density estimators and put $h = 1.06 \times \widehat{\sigma} \times n^{-1/5}$, where $\widehat{\sigma}$ is the sample SD. Genes with high entropy values were selected for the cluster analysis.

The Co-Expression Selection (CoEx1)

For each gene, the co-expressions to all other genes were calculated using Spearman correlation. Let $s_{ij} = \left|\rho_{ij}\right|$ denote the absolute value of the Spearman rank correlation ρ_{ij} between expression profiles for genes i and j. The matrix S with elements s_{ij} is considered as a similarity matrix for the genes with respect to co-expression. In the original article, the authors use Pearson correlation, but we applied Spearman correlation instead, which in earlier studies have proven to be more efficient in identifying co-expressed genes (Kumari et al., 2012; Wang et al., 2014). To rank genes according to their co-expression we define the CoEx1 score for gene i as the median of the s_{ij} values, i.e.,

$$\text{CoEx1}_i = \text{median}_{j_{j\neq i}}\left\{s_{ij}\right\}.$$

The genes with highest median correlations were selected.

The Modified Co-Expression Selection (CoEx2)

The co-expression network analysis was developed for variable selection in cluster analysis of microarray data. Since microarray data tend to be noisy, the authors argue that directly using the similarity matrix for co-expression analysis may be inappropriate and therefore suggests a transformation of the similarity matrix. The modified version uses a power transformation of the elements in the similarity matrix (Wang et al., 2014). The CoEx2 score for gene i is defined as:

$$\text{CoEx2}_i = \sum_{j\neq i} s_{ij}^3,$$

where s_{ij} are the elements of the similarity matrix. The genes with highest scores were selected for analysis.

The Dip-Test Statistic Selection (DIP)

The dip-test was used to test unimodality and is based on the maximum difference between the empirical distribution and the unimodal distribution that minimizes that maximum difference (Hartigan and Hartigan, 1985). Genes with low p-values were selected for analysis. The R-package *diptest* was used for calculations (Maechler, 2013).

The Bimodality Index Selection (BI)

For each gene, it is assumed that the density $f(x)$ of the expression value can be described by a normal-mixture model with two components, i.e.,

$$f(x) = pN(\mu_A, \sigma) + \left(1-p\right)N(\mu_B, \sigma),$$

where μ_A and μ_B denote the mean in the two subgroups and p is the proportion of samples in one group (Wang et al., 2009). The BI is defined as

$$BI = \sqrt{p(1-p)}\frac{|\mu_A - \mu_B|}{\sigma}.$$

The expectation-maximization (EM) algorithm was used to estimate the BI using the R package *mixtools* (Benaglia et al., 2009). Ten different starting values were used for the EM-algorithm, generated from a grid with 10 values for the fraction parameter p, evenly spaced between 0 and 1, for more details, see Karlis and Xekalaki (2003). Genes with high BI were selected for analysis.

The Variance Reduction Score Selection (VRS)

The VRS is used for measuring the reduction of variance when splitting the data into two clusters (A and B) and is defined as the ratio of the within sum of squares (WSS) and the total sum of squares (TSS):

$$\text{VRS} = \frac{\text{WSS}}{\text{TSS}} = \frac{\sum_A (x_i - \bar{x}_A)^2 + \sum_B (x_i - \bar{x}_B)^2}{\sum_i (x_i - \bar{x})^2},$$

where $\bar{x}_A$ and $\bar{x}_B$ denotes the mean values within group A and B. These values lie between zero and one, where a low score indicates an informative split (Hellwig et al., 2010). Hence, genes with a low score were selected for cluster analysis. The clusters were obtained using k-means clustering with $k = 2$.

The Weighted Variance Reduction Score Selection (wVRS)

The wVRS is a weighted version of VRS that takes sample size into account, i.e.,

$$\text{wVRS} = \frac{\frac{1}{2}(\frac{1}{n_A}\sum_A (x_i - \bar{x}_A)^2 + \frac{1}{n_B}\sum_B (x_i - \bar{x}_B)^2)}{\frac{1}{n}\sum_i (x_i - \bar{x})^2},$$

where n_A and n_B are the sample sizes in group A and B (Hellwig et al., 2010). The grouping of the data was obtained by the k-means algorithm, $k = 2$. Again, genes with a low score were selected.

The Modified Variance Reduction Score Selection (mVRS)

The mVRS considers the proportion of variance reduction when splitting data into two cluster by using the fuzzy c-means algorithm, also known as soft k-means clustering (Bezdek, 1981). Genes with a low score were selected for further analysis. The R-package *cluster* was used for calculations (Maechler et al., 2019).

The Bimodality Coefficient Selection (BC)

The BC yields a value between 0 and 1 (for large samples) and is calculated by

$$\mathrm{BC} = \frac{\gamma^2 + 1}{\kappa + 3\frac{(n-1)^2}{(n-2)(n-3)}},$$

where γ is the sample skewness, κ is the sample excess kurtosis and n is the sample size. The genes with largest BCs were selected for cluster analysis. The R-package *modes* was used for calculating the coefficient (Sathish and 4D Strategies, 2016).

Evaluations

The considered gene selection approaches (13 methods times three levels of number of selected genes) were evaluated and compared to two negative controls (random selection and no selection) and a positive control (supervised selection).

Random Selection (RAND)

Here we randomly selected k genes, k = 100, 1000, or 3000. The performance of the random selection (RAND) was highly variable, therefore, the procedure was repeated 1000 times, resulting in 1000 performance measures. The evaluated gene selection methods were compared to the 25th, 50th, and 75th percentile and the mean value (RAND) of the random selection performance measures.

Supervised Selection (PVAL)

The gold standard partitions were used to rank genes according to how well they separated the two subtypes. A standard test for comparing two groups is the t-test, but for identification of differentially expressed genes it is common to use a generalized linear model (GLM). To describe the read count K_{ij} for gene i observed in sample j, we used a GLM from the negative-binomial (NB) family with a logarithmic link, given as:

$$K_{ij} \sim \mathrm{NB}\,(\mathrm{mean} = \mu_{ij},\ \mathrm{dispersion} = \alpha_i),$$
$$\mu_{ij} = s_{ij} q_{ij},$$
$$\log_2 q_{ij} = \beta_{i0} + \beta_{i1} x_j.$$

The normalizing factors s_{ij} compensate for differences in sequencing depth between samples and for eventual gene-related technical biases such as gene length. We used the default procedure where these factors are considered as fixed within each sample, $s_{ij} = s_j$ and then only accounts for differences in sequencing depth between samples. These so-called size factors were estimated by the median-of-ratios method:

$$s_j = \mathrm{median}_{i:K_i^R \neq 0}\left(\frac{K_{ij}}{K_i^R}\right), K_i^R = \left(\prod_{j=1}^{n} K_{ij}\right)^{1/n}$$

The linear part $\beta_{i0} + \beta_{i1} x_{j1}$ contains a categorical variable x_{j1} with two levels, corresponding to the cancer subgroups. The coefficient β_{i1} quantifies the extent to which gene i is differentially expressed between the groups. The intercept term β_{i0} models the base mean, which is allowed to differ between genes. The dispersion α_i was regarded as a gene-specific parameter in the model.

To fit the model (i.e., estimation of the parameters α_i, β_{i0}, β_{i1} for each gene i) we applied the R-package *DESeq2*, which implements the empirical Bayes shrinkage method (Love et al., 2014). The p-value for the test that gene i is differently expressed (i.e., $H_0 : \beta_{i1} = 0$) was then used to rank genes, so that genes with lowest p-values were used for clustering. The method was applied to data that had been filtered for low expressed genes, but not normalized using the variance stabilizing transform.

No Selection (ALL)

Gene selection is performed to remove non-informative and irrelevant genes. An alternative is to base the clustering on all genes, and we included the case of no selection as a reference point.

Similarity Between Feature Selection Methods

Each selection procedure was characterized by calculating the mean value and SD of the selected genes. Procedures with similar characteristics may also make similar selections. In addition, we carried out a more direct analysis by measuring the overlap between the approaches, i.e., for each pair of approaches we measured the percentage of genes selected by both methods.

Performance of the Feature Selection Methods

The clustering performance was measured using the ARI based on the clustering result compared to the gold standard partition. An ARI-value of 1 indicates a complete match to the gold standard partition, whereas a value of 0 indicates an agreement as good as a random clustering.

Detailed Evaluation of Top-Performing Feature Selection Methods

The evaluations described above utilize four data sets and were used to identify a set of interesting feature selection methods. To deeper understand our findings, we used these data sets to simulate two types of data sets using stratified subsampling with replacement from the original data: balanced data sets where 50 samples were drawn from each subtype and skewed data sets were 25 (75) of the samples were drawn from the least (most) common subtype. Hundred data sets were sampled for each type. The chosen selection methods were applied to each of the simulated data sets, hierarchical clustering with Euclidean distance, was performed on the top 1000 ranked genes to cluster the samples in two groups and ARI was used to measure the performance. For each pair of methods, the pairwise ARI-observations were used to construct differences and the one sample t-test was used to test if the expected value of the difference deviated from zero.

In addition to the ARI-values we also observed the number of samples in the smallest of the two groups generated by the clustering, i.e., a number between 1 and 50. Again, the one sample t-test was used to investigate differences between the considered selection methods. Since BI is computationally heavy, the EM-algorithm was used with only one initial value of the parameter vector, obtained as follows: first the data was divided into two groups using the k-means ($k = 2$) clustering algorithm, and then the means, SDs and size fraction in the two subsamples were used as starting values for the mixing parameters.

For the top-performing feature selection methods, we also investigated the change in ARI-values when increasing the number of selected genes in the cluster analysis. The number of selected genes was increased gradually in 1000 steps between two selected genes up to all genes remaining after initial filtration. Since the clustering result is highly variable, we applied a running mean over 100 values to get a smoother curve.

The aim of the feature selection is to exclude genes that are non-informative for distinguishing between the disease subtypes. As a way of measuring the relevance of the selected features, the list of 1000 top scoring genes were compared to 299 known cancer driver genes (Bailey et al., 2018). Enrichment of genes relevant to cancer etiology were tested using a one sided Fisher's exact test.

RESULTS

We tested the performance of 13 feature selection methods when identifying subgroups using cluster analysis on four human cancer data sets. For each method the k top ranked genes were selected, $k = 100$, 1000, and 3000. Three references were considered: a negative control where all genes were selected, a negative control where k genes were randomly selected and a positive control where genes were selected using a supervised approach. The selection methods were applied on the 15,298, 15,388, 15,397, and 15,397 genes that remained after filtering low expressed genes in KIRP, STAD, LGG, and BRCA, respectively. In addition, a small simulation study was performed with the objective to investigate how clustering is affected when non-informative features are included in the analysis.

Simulation Study

In the case when the clustering was based on only the informative features all the clustering results were identical to the desired AB-partition with an average ARI equal to 1. In the case when all features were included, the average ARI was 0.34. For the case when the clustering was based on the features with the highest SDs the clustering performance peaked when around 600 features were included and declined when more features were added, see **Supplementary Figure 1**. Although the negative effect of including non-informative features is likely to be general, it should be stressed that the magnitude of the effect depends on the effect size, the sample size, and the percentage of informative features. Moreover, in real problems we may in addition to informative and non-informative features have features that are informative to secondary factors, e.g., gender, age, and prior treatments.

Characteristics of Top Ranked Genes

As an initial investigation, we studied the mean expression and SD of top ranked genes obtained for the considered feature selection approaches.

The Mean Value of Selected Genes

As expected, genes selected using the median (M) or the third quartile (Q3) were highly expressed compared to the other methods. The BC approach selected genes expressed at a very low level. The supervised approach (PVAL) selected genes at an intermediate gene expression level, which was comparable to the expression level seen in the whole data (ALL). Approaches using SD, IQR, ENT, CoEx1, and CoEx2 selected genes with mean gene expression similar to that obtained by the supervised approach. The remaining methods, BI, the dip-test statistic (DIP) and the variance reduction scores (VRS, wVRS, and mVRS), selected genes expressed at a relatively low level. Interestingly, the same relative patterns were observed for all data sets and independently of the number of selected genes, see **Figure 2** and **Supplementary Figures 2, 3**.

The Standard Deviation of Selected Genes

It is natural to assume that informative genes should have relatively high SD, compared to most other genes. As expected, genes selected using SD, ENT, and IQR, had high SDs. The M and Q3 methods selected genes with relatively low variation, which was close to the SD observed in the whole data sets (ALL). Intermediate values of SD were observed among genes selected using the variance reduction scores (VRS, wVRS, and mVRS), and BI. For BC, DIP, CoEx1, CoEx2, and the supervised approach (PVAL), the level of SD varied between low and intermediate depending on data set and number of selected genes, see **Figure 3** and **Supplementary Figures 4, 5**.

Overlap of Selected Genes

The above results show that methods based on similar selection principles also have similar properties with respect to the mean and SD of the selected genes, see **Figures 2**, **3**. Next, we investigated to what degree the methods selected the same genes, by studying the overlap when 1000 genes were selected. The overlap between M and Q3 was high (>90%) in all data sets, but both methods showed very limited resemblance to the other methods (<6% overlap in average). High agreement was also observed between SD, ENT and IQR (85% in average), as well as between CoEx1 and CoEx2 (77% in average). CoEx1 and CoEx2 showed low overlap with the remaining methods (<10%). The intersection between VRS, mVRS and wVRS was in average 73% in all data sets, and the group showed a greater resemblance to BI than to BC (in average 67 vs 40%), see **Figure 4**. The positive control (PVAL), that is expected to be a good selection procedure, had a very small overlap with the methods M, Q3, CoEx1, and CoEx2. For detailed results, see **Figure 4**.

Feature Selection Methods

The performance was measured using the adjusted Rand index (ARI) comparing the obtained clustering result with the gold standard. Each of the 13 selection methods were used to cluster the four cancer data sets by selecting the 100, 1000, or 3000 top ranked genes. Hence, each method was used to perform 12 cluster analyses and generated 12 ARI-values. The results were compared to two negative controls (randomly selected genes and a selection including all genes) and a positive control (PVAL). As expected, the supervised selection approach (PVAL) had the highest combined performance (considering the median value of the 12 ARI-values) followed in decreasing order by DIP, BI, IQR, ENT, RAND, Q3, mVRS, M, VRS, SD, BC, wVRS, CoEx1, and

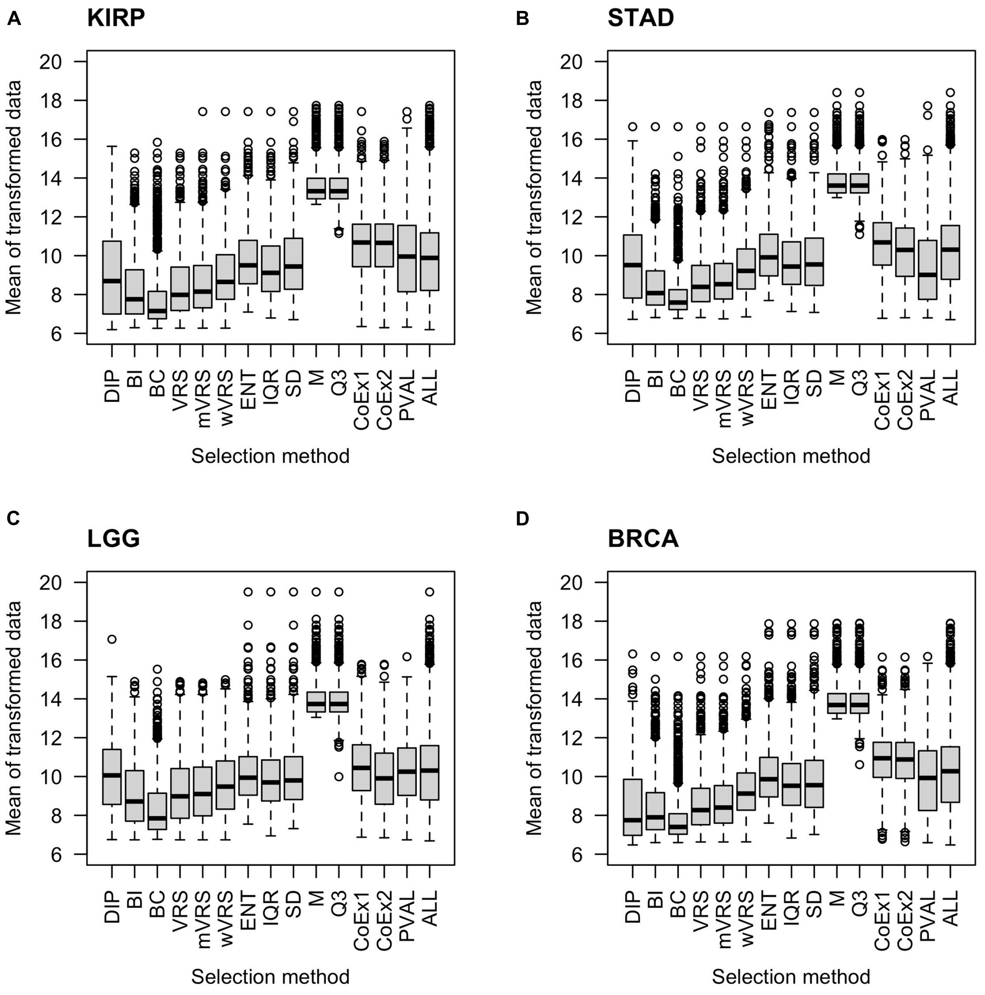


FIGURE 2 | Boxplots of mean expression values over all samples for 1000 selected genes. The figure shows the result for the data sets KIRP **(A)**, STAD **(B)**, LGG **(C)**, and BRCA **(D)**. Each plot displays expression values of preprocessed data for the 13 feature selection methods, the positive control (PVAL) and the negative control (ALL) including all genes. The gene selection methods are: dip-test statistic (DIP), bimodality index (BI), bimodality coefficient (BC), variance reduction score (VRS), modified variance reduction score (mVRS), weighted variance reduction score (wVRS), entropy estimator (ENT), interquartile range (IQR), standard deviation (SD), mean value (M), third quartile (Q3), co-expression (CoEx1), and modified co-expression (CoEx2).

CoEx2, see **Figure 5**. However, the relative performance of the methods varied between the four data sets and was also affected by the number of selected genes. Evaluating the approaches based on their mean ranking taken over all 12 analyses revealed that the supervised approach performed best followed by BI, mVRS, DIP, VRS, RAND, IQR, ENT, Q3, M, SD, wVRS, BC, CoEx1, and CoEx2, see **Table 1**.

Ranking genes according to Q3 or M is a simple way of selecting highly expressed genes. The performance for Q3 and M varied from being top performing (BRCA 3000 genes) to be at the very bottom (STAD 3000 genes). In KIRP, Q3 was always ranked higher than M and for STAD it was the other way around. For LGG and BRCA it varied depending on number of features included, see **Figure 6** and **Supplementary Figures 6, 7**. Altogether, Q3 performed slightly better than M.

Of the three methods relying on variability across samples, IQR and ENT generally performed better than the commonly used SD procedure. IQR outperformed ENT in the LGG data, while ENT performed better on the BRCA data. In KIRP and STAD it depended on the number of included features, see **Figure 6** and **Supplementary Figures 6, 7**. Within this category, IQR performed best and should be considered as a simple alternative to SD.

The methods relying on gene correlation (CoEx1 and CoEx2), performed worst of all considered methods, with CoEx2 slightly worse than CoEx1, see **Table 1**.

Among methods based on modality (i.e., DIP, BI, VRS, mWRS, wVRS, and BC), BI and DIP where the methods with the overall highest performance. The relative performance of DIP was particular good when more genes were selected (1000 or 3000), while BI performed particularly well on the KIRP and STAD data, see **Table 1**, **Figure 6**, and **Supplementary Figures 6, 7**. When 1000 genes were selected, the overlap between DIP and BI varied between 21 and 37%, see **Figure 4**. Furthermore,

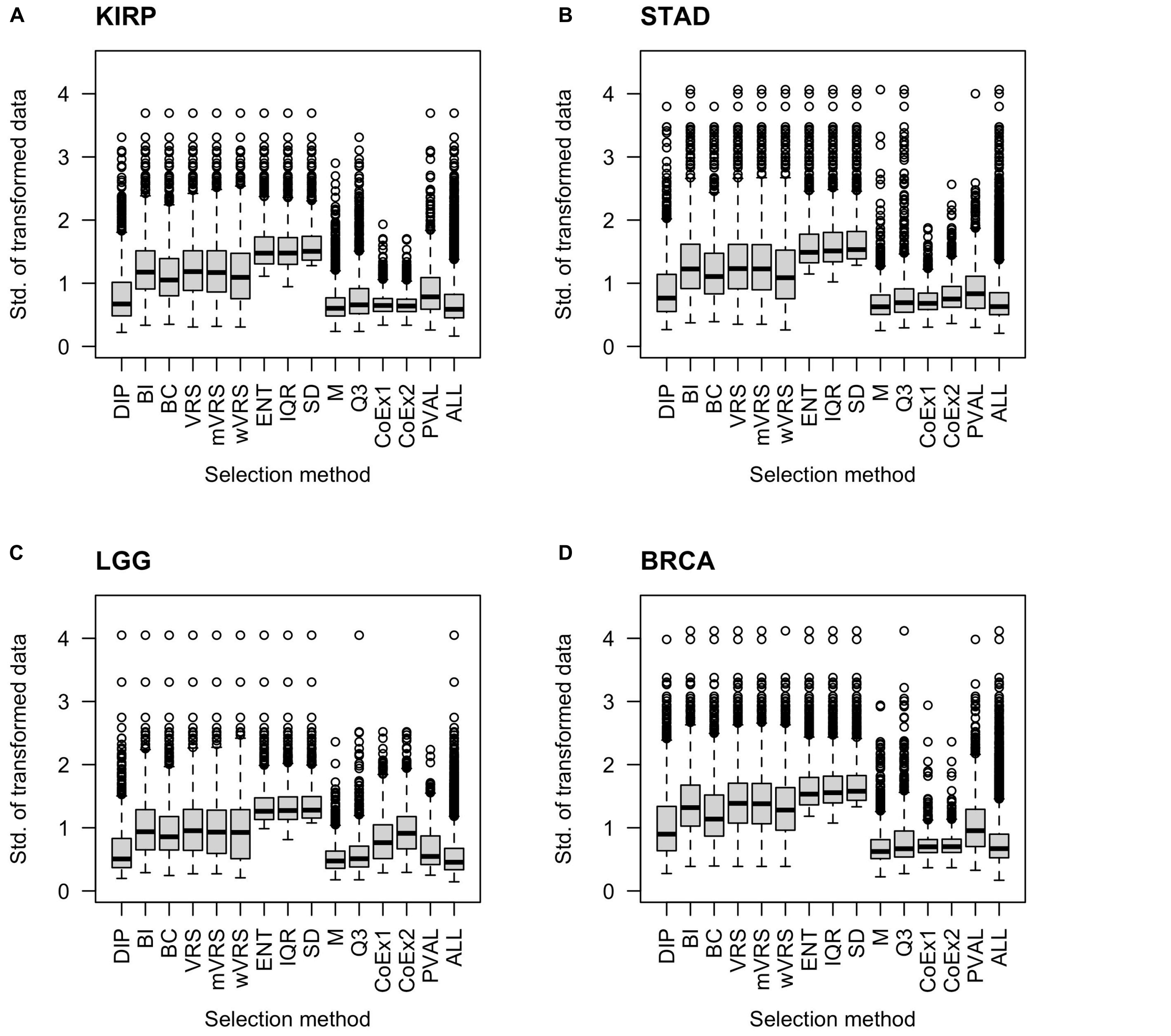


FIGURE 3 | Boxplots of standard deviation across samples for 1000 selected genes. Each plot displays standard deviation based on preprocessed data for the 13 feature selection methods, the positive control (PVAL) and the negative control (ALL) including all genes. The figure shows the result for the data sets KIRP **(A)**, STAD **(B)**, LGG **(C)**, and BRCA **(D)**. The gene selection methods are: dip-test statistic (DIP), bimodality index (BI), bimodality coefficient (BC), variance reduction score (VRS), modified variance reduction score (mVRS), weighted variance reduction score (wVRS), entropy estimator (ENT), interquartile range (IQR), standard deviation (SD), mean value (M), third quartile (Q3), co-expression (CoEx1), and modified co-expression (CoEx2).

DIP tended to select genes that were slightly higher expressed than BI, see **Figure 2** and **Supplementary Figures 2, 3**. More evident, BI selected genes with higher SD than DIP, see **Figure 3** and **Supplementary Figures 4, 5**. Altogether, this suggests that although performing similar, and relatively well, DIP and BI select rather different genes with different characteristics.

Comparisons to Positive and Negative Controls

Intuitively, selecting the *k* top scoring genes using a good feature selection method should in average result in a better clustering performance than obtained when randomly selecting *k* genes, but worse performance than using a supervised approach. However, if the gene expressions are highly influenced by a *secondary factor* (i.e., a factor that is not informative for predicting the subgroups) applying feature selection may result in a performance worse than the random selection.

As expected, the supervised approach PVAL was commonly superior to the unsupervised selection approaches, although occasionally performed slightly worse than some other methods, see **Table 1**. For the LGG data randomly selecting *k* genes outperformed most of the selection methods, see **Figure 5** and **Supplementary Figures 6, 7**. This may indicate that the RNA-expression of the individuals is influenced by secondary factors or that the binary partitions defined by the gold standard are heterogeneous and preferably should be divided further.

An alternative to applying feature selection is to include all genes in the cluster analysis for which the ARI-values 0.28, −0.01, 0.39, and 0.73 were observed for KIRP, STAD, LGG, and

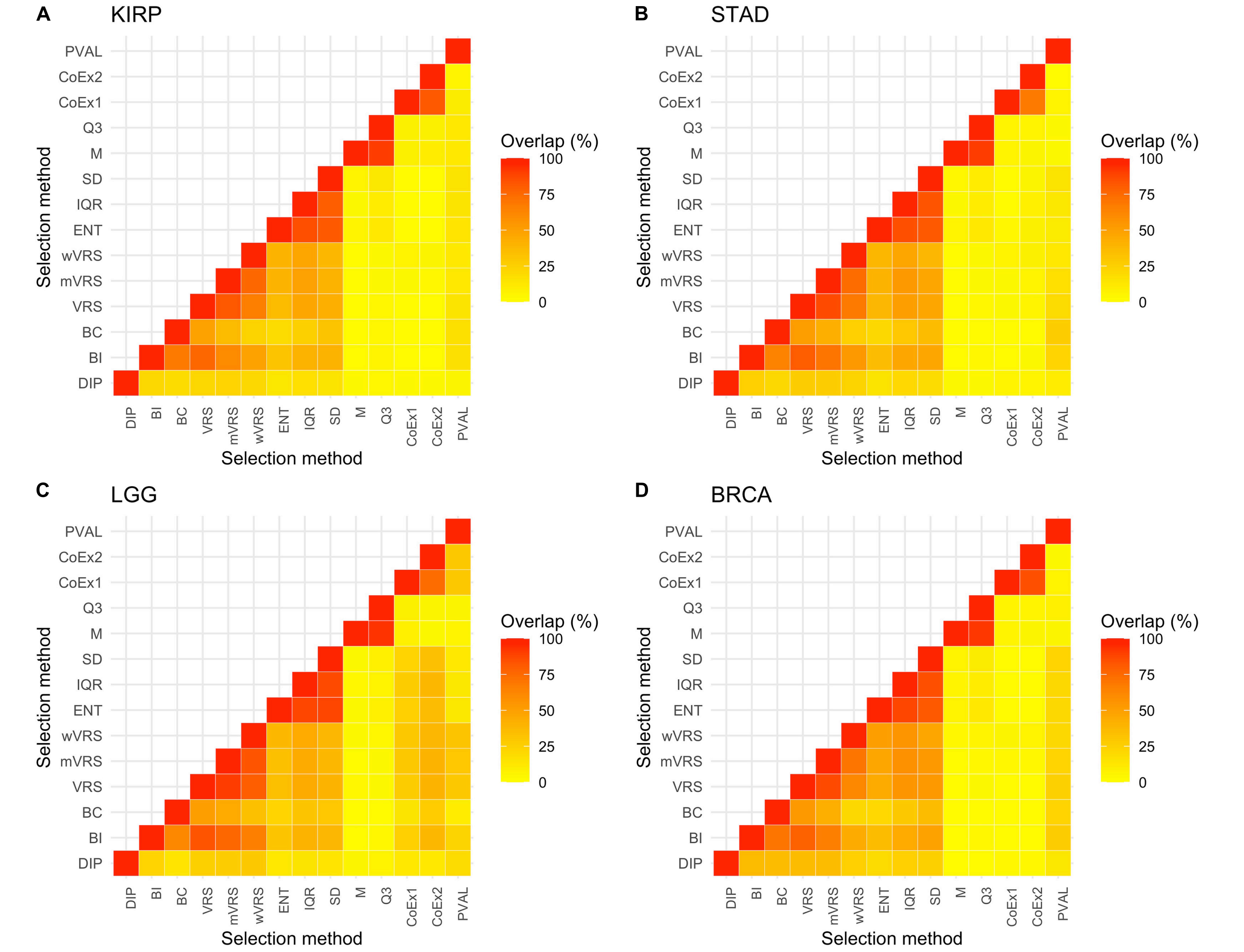


FIGURE 4 | Percentage overlap between 1000 selected genes for the 13 different feature selection methods and the four data sets KIRP **(A)**, STAD **(B)**, LGG **(C)**, and BRCA **(D)**. The feature selection methods are: dip-test statistic (DIP), bimodality index (BI), bimodality coefficient (BC), variance reduction score (VRS), modified variance reduction score (mVRS), weighted variance reduction score (wVRS), entropy estimator (ENT), interquartile range (IQR), standard deviation (SD), mean value (M), third quartile (Q3), co-expression (CoEx1), modified co-expression (CoEx2), and the positive control (PVAL).

BRCA, respectively. For KIRP and BRCA, including all genes was as good as the best performing selection methods, but for STAD and LGG the best selection methods yielded considerably higher ARI-values, 0.66 and 0.72, respectively, see **Table 2**. On the other hand, variable selection often resulted in lower ARI-values compared to including all genes, in particular when just 100 genes were selected, see **Table 2**. This suggests that variable selection has potential to improve the clustering, but that the choice of methods and the number of selected genes are crucial for the performance.

Detailed Evaluation of Top-Performing Feature Selection Methods

Based on our findings we conclude that DIP, BI, and mVRS are the most promising methods and that good performance is usually obtained when 1000 genes are selected. These methods also ranked high when *k*-means ($k = 2$) clustering and hierarchical clustering with a correlation-based distance measure were used, see **Supplementary Tables 1–4**. DIP, BI, and mVRS together with the commonly used SD method were therefore selected for a deeper study based on hundreds of simulated balanced and skewed data sets, see section "Detailed Evaluation of Top-Performing Feature Selection Methods."

Pairwise comparisons with respect to ARI between DIP, BI, mVRS, and SD showed that DIP was as good or better than the other methods, with the exception that BI was slightly better than DIP for the skewed KIRP data set, see **Table 3** and **Figure 7**. Furthermore, SD did not perform well and was the worst performing method for most of the simulations, **Table 3**. For LGG, STAD, and BRCA the difference in average ARI between DIP and SD ranged between 0.03 and 0.35 and five out of six findings were significant, see **Table 3**.

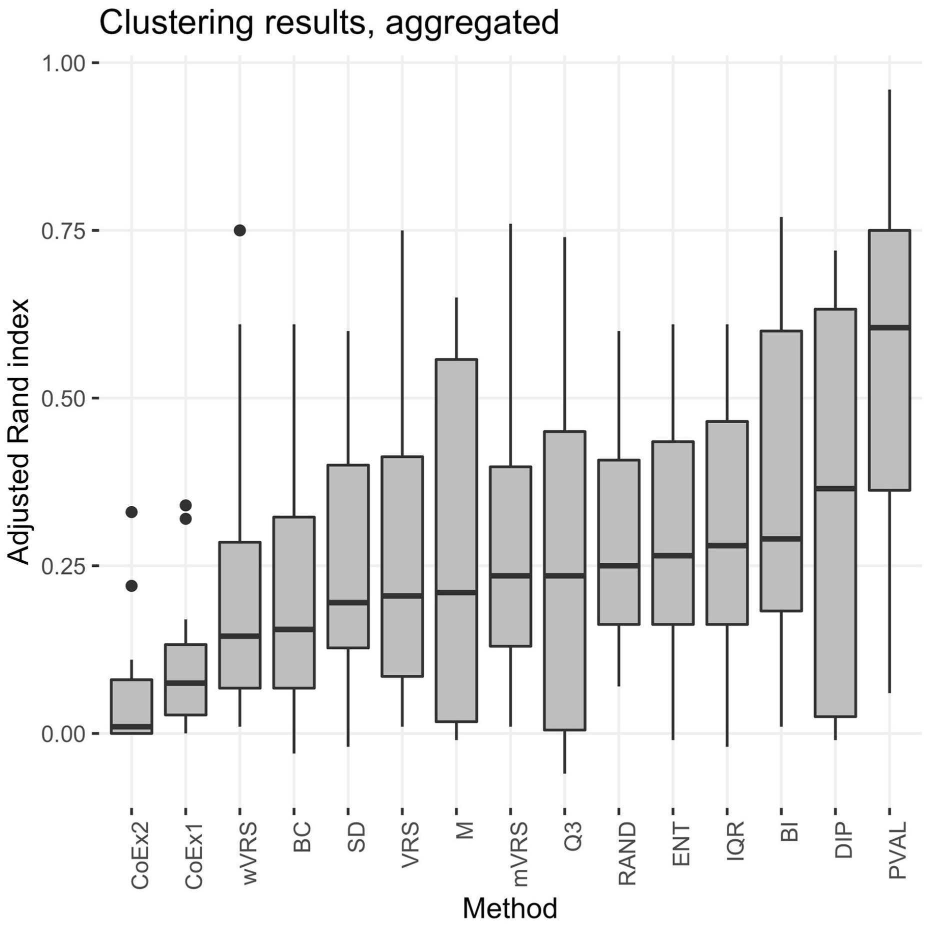


FIGURE 5 | Boxplot of aggregated clustering performance over the four data sets KIRP, STAD, LGG, and BRCA. Performance is measured using adjusted Rand index and the feature selection methods are ordered according to increasing median values. The selection methods are: dip-test statistic (DIP), bimodality index (BI), bimodality coefficient (BC), variance reduction score (VRS), modified variance reduction score (mVRS), weighted variance reduction score (wVRS), entropy estimator (ENT), interquartile range (IQR), standard deviation (SD), mean value (M), third quartile (Q3), co-expression (CoEx1), modified co-expression (CoEx2), the positive control (PVAL), and the negative control (RAND).

TABLE 1 | Rank of feature selection methods for data sets KIRP, STAD, LGG, and BRCA based on adjusted Rand index.

	DIP	BI	BC	VRS	mVRS	wVRS	ENT	IQR	SD	M	Q3	CoEx1	CoEx2	PVAL	RAND
KIRP100	14	2	3.5	6	3.5	12	7.5	9	10	7.5	5	13	15	1	11
STAD100	7	4.5	8	4.5	6	1	13	15	14	10	12	11	9	3	2
LGG100	14	10	10	10	10	10	7	5	6	15	13	4	3	1	2
BRCA100	6	2	8.5	5	3	4	7	8.5	10	11	13	14	15	1	12
KIRP1000	6.5	1	10.5	8	6.5	13	4.5	2.5	9	4.5	2.5	14	15	10.5	12
STAD1000	2	3	14	11	6	9	7	4	5	12	15	10	13	1	8
LGG1000	2	10	13	14	9	15	11	7	8	3	4	6	12	1	5
BRCA1000	5	8.5	6.5	1.5	3	6.5	11.5	11.5	13	10	4	14	15	1.5	8.5
KIRP3000	7	1	12	6	11	9.5	3.5	3.5	3.5	13	9.5	14	15	3.5	8
STAD3000	13.5	1	6	5	8	9	4	3	12	13.5	15	10	11	2	7
LGG3000	2	9	15	13	8	14	11.5	10	11.5	6	3	7	5	1	4
BRCA3000	6	11.5	11.5	3.5	3.5	11.5	6	9	6	2	1	14	15	11.5	8
Mean rank	7.1	5.3	9.9	7.3	6.5	9.5	7.8	7.3	9.0	9.0	8.1	10.9	11.9	3.2	7.3

The table shows results for selection of top ranked genes at three levels: 100, 1000, and 3000 genes. The gene selection methods are: dip-test statistic (DIP), bimodality index (BI), bimodality coefficient (BC), variance reduction score (VRS), modified variance reduction score (mVRS), weighted variance reduction score (wVRS), entropy estimator (ENT), interquartile range (IQR), standard deviation (SD), mean value (M), third quartile (Q3), co-expression (CoEx1), modified co-expression (CoEx2), and the positive control (PVAL).

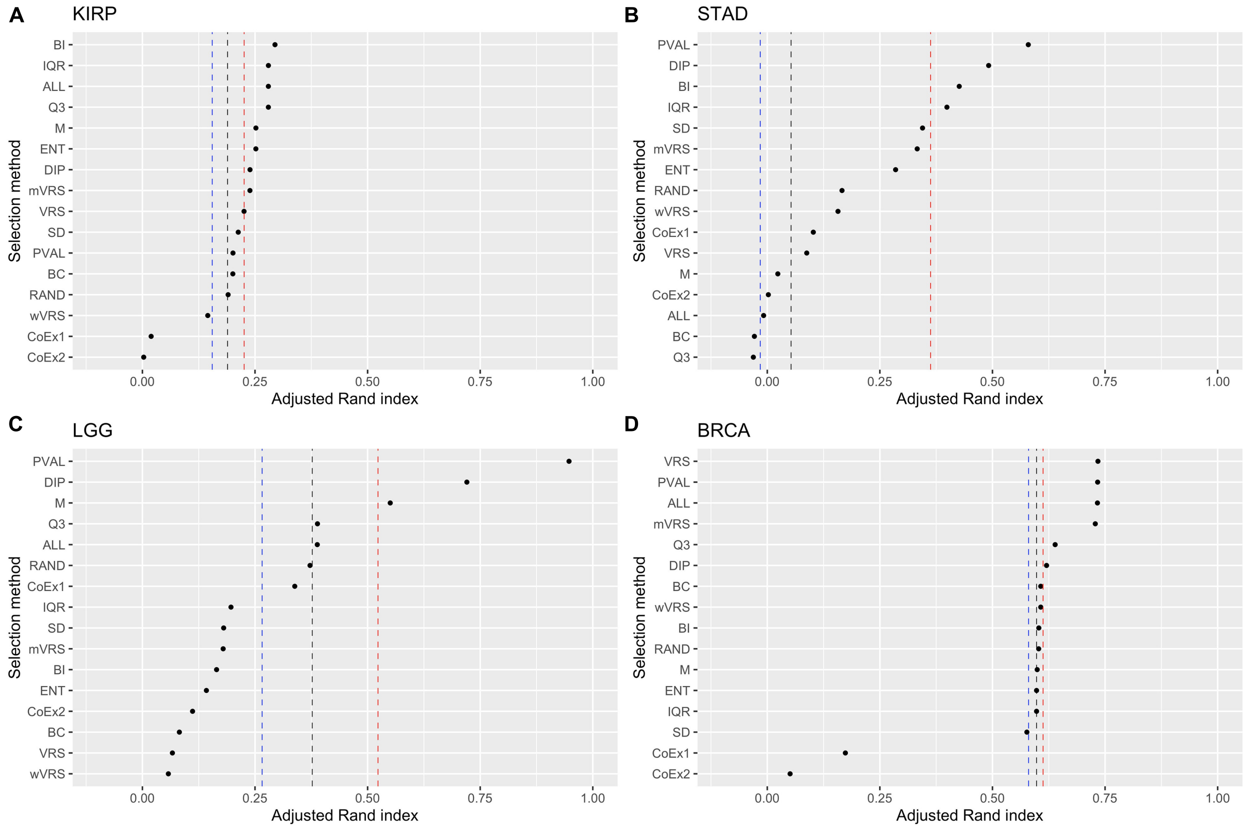


FIGURE 6 | Adjusted Rand index for hierarchical clustering using the top 1000 ranked genes on the datasets KIRP **(A)**, STAD **(B)**, LGG **(C)**, and BRCA **(D)**. The vertical lines represent the first, second, and third quartiles of a random selection. The feature selection methods on the *y*-axis are: dip-test statistic (DIP), bimodality index (BI), bimodality coefficient (BC), variance reduction score (VRS), modified variance reduction score (mVRS), weighted variance reduction score (wVRS), entropy estimator (ENT), interquartile range (IQR), standard deviation (SD), mean value (M), third quartile (Q3), co-expression (CoEx1), modified co-expression (CoEx2), the positive control (PVAL), and the two negative controls ALL and RAND. The selection methods are ordered according to increasing performance.

TABLE 2 | Adjusted Rand index for 13 feature selection methods, a negative (RAND) and positive control (PVAL) for data sets KIRP, STAD, LGG, and BRCA.

	DIP	BI	BC	VRS	mVRS	wVRS	ENT	IQR	SD	M	Q3	CoEx1	CoEx2	PVAL	RAND
KIRP100	0.01	0.25	0.23	0.18	0.23	0.10	0.17	0.14	0.14	0.17	0.19	0.05	0.00	0.39	0.11
STAD100	0.03	0.05	0.03	0.05	0.04	0.07	−0.01	−0.02	−0.02	0.01	−0.01	0.01	0.01	0.06	0.07
LGG100	0.00	0.01	0.01	0.01	0.01	0.01	0.08	0.10	0.09	0.00	0.01	0.12	0.22	0.96	0.29
BRCA100	0.67	0.77	0.61	0.75	0.76	0.75	0.61	0.61	0.60	0.58	0.37	0.11	0.07	0.78	0.49
KIRP1000	0.24	0.29	0.20	0.23	0.24	0.15	0.25	0.28	0.21	0.25	0.28	0.02	0.00	0.20	0.19
STAD1000	0.49	0.43	−0.03	0.09	0.33	0.16	0.28	0.40	0.34	0.02	−0.03	0.10	0.00	0.58	0.17
LGG1000	0.72	0.16	0.08	0.07	0.18	0.06	0.14	0.20	0.18	0.55	0.39	0.34	0.11	0.95	0.37
BRCA1000	0.62	0.60	0.61	0.73	0.73	0.61	0.60	0.60	0.58	0.60	0.64	0.17	0.05	0.73	0.60
KIRP3000	0.21	0.29	0.12	0.27	0.16	0.18	0.28	0.28	0.28	0.11	0.18	0.03	0.00	0.28	0.21
STAD3000	−0.01	0.66	0.19	0.35	0.04	0.02	0.38	0.42	−0.01	−0.01	−0.06	0.00	0.00	0.61	0.14
LGG3000	0.71	0.19	0.08	0.15	0.27	0.14	0.17	0.17	0.17	0.32	0.63	0.32	0.33	0.74	0.38
BRCA3000	0.60	0.60	0.60	0.60	0.60	0.60	0.60	0.60	0.60	0.65	0.74	0.04	0.01	0.60	0.60

The table shows results for selection of top ranked genes at three levels: 100, 1000, and 3000 genes. Adjusted Rand index when including all genes was obtained as 0.28, −0.01, 0.39, and 0.73 for KIRP, STAD, LGG, and BRCA, respectively. The gene selection methods are dip-test statistic (DIP), bimodality index (BI), bimodality coefficient (BC), variance reduction score (VRS), modified variance reduction score (mVRS), weighted variance reduction score (wVRS), entropy estimator (ENT), interquartile range (IQR), standard deviation (SD), mean value (M), third quartile (Q3), co-expression (CoEx1), and modified co-expression (CoEx2).

In order to better understand the results, we investigated the *number of samples in the smallest group* (NSSG) obtained doing a cluster analysis resulting in two groups. This number should be close to 50 for the balanced data set and close to 25 for the skewed data sets. Moreover, if this number is very small it indicates that the clustering is governed by just

TABLE 3 | The mean value of 100 pairwise adjusted Rand index-differences (row method – column method) for different pairs of feature selection methods: the dip-test (DIP), bimodality index (BI), modified variance reduction score (mVRS) and standard deviation (SD).

		25%			50%		
		BI	**mVRS**	**SD**	**BI**	**mVRS**	**SD**
KIRP	DIP	−0.01	0.00	0.01	0.00	0.00	0.01
	BI		0.01	0.02*		0.00	0.01
	mVRS			0.01			0.01
STAD	DIP	0.05	0.07**	0.10***	0.00	0.02	0.05**
	BI		0.02	0.06**		0.02	0.06**
	mVRS			0.03			0.03
LGG	DIP	0.06**	0.04	0.04	0.37***	0.37***	0.35***
	BI		−0.03*	−0.02*		0.00	−0.02
	mVRS			0.00			−0.02
BRCA	DIP	0.01	0.00	0.03**	0.06***	0.02	0.07***
	BI		−0.01	0.03***		−0.03*	0.01
	mVRS			0.03**			0.05***

*Simulations were made for the data sets KIRP, STAD, LGG, and BRCA, and two types of data sets were simulated: unbalanced data where 25% of the individuals belonged to the minor class and a balanced data set were 50% of the individuals belonged to each of the two classes. The number of samples was 100 in each simulation. The one sample t-test was used to test if the mean difference deviated from zero. Positive (negative) differences indicate that the row-method was better (worse) than the column method. Here *, **, ** denote a significant result at the 0.05, 0.01, and 0.001 significance level, respectively.*

a few samples (outliers). Interestingly, the ARI-differences and NSSG-differences were correlated, so that methods with relatively high ARI also had a relatively high NSSG, see **Table 3** and **Supplementary Table 5**. In particular SD had considerably lower NSSG than the other methods, especially for the balanced data, see **Supplementary Table 5**.

It is not trivial to select how many genes to include in the cluster analysis. The results from the analysis of ARI in relation to the number of selected genes showed a highly variable performance, especially in STAD and LGG, see **Figure 8**. The most noticeable result was the gradual decrease in performance in the LGG data for DIP and PVAL when including more genes, indicating that it is possible to increase the ability to identify disease subgroups substantially when choosing features wisely. For BI, mVRS, and SD in the LGG data, the general trend was an increase in performance when including more genes. In STAD, the general trend was a decreasing performance when including more genes. In both KIRP and BRCA the performance was relatively stable when changing the number of included genes. At least for BRCA, this might be explained by the high number of informative genes.

The overlap between the 299 cancer driver genes and all genes remaining after initial filtration were 279, 285, 279, and 286 for KIRP, STAD, LGG, and BRCA, respectively. No enrichment of cancer driver genes was observed for the 1000 top ranked genes for DIP, BI, mVRS, and SD, except for in BRCA where genes selected using SD had a significantly higher proportion of cancer driver genes ($p < 0.001$), see **Table 4**. When extending the comparison to include all 13 selection methods, a significant overrepresentation of cancer driver genes was observed for both M and Q3 (data not shown) in all four data sets, suggesting that the detected cancer driver genes are generally high expressed.

DISCUSSION

Feature selection prior to clustering RNA-seq is common and is often done by selecting the genes with the highest SD (i.e., the SD method). However, this problem has not been well studied and there is little evidence that selecting genes with high SD is the best approach. Before we discuss our findings, it is worth pointing out that measuring the performance of feature selection methods is difficult. The clustering performance, in our case measured using ARI-values, does in addition to the feature selection algorithm also depend on clustering method, the nature of the data and how the gold standard is defined. Samples from a cancer cohort can be divided in several logical partitions, e.g., partitions defined by gender, age or disease subtype. For these partitions, it is likely that a set of genes will be differentially expressed between the groups. Hence, a low ARI-value does not automatically mean that the cluster analysis failed, it can also be a consequence of secondary factors affecting the data or that the groups defined by the gold standards are heterogeneous and should be further divided.

The general idea behind feature selection prior to cluster analysis is to remove genes that do not contain information about the "true partition" of the samples, e.g., genes that are identically distributed among all samples and therefore only contribute with noise, making the analysis harder.

In the considered simulation study only a small set of the genes were informative and including all genes in the analysis had a negative effect on the clustering result. This negative effect will be reduced when the number of informative genes increases (data not shown). RNA-seq cancer data are much more complex than the simulated data and the informative genes are unknown although they in our case can be predicted using a supervised test. For the LGG and STAD data considerably better clustering results were obtained when genes predicted to be informative were used compared to when all genes were included, which suggests that feature selection has the potential to improve the clustering performance. For BRCA and KIRP, the gain of using supervised selection was limited, which suggests that feature selection is unlikely to have a positive effect on the clustering result. Arguably, a feature selection approach should identify informative genes related to the factor of interest.

For the considered four data sets, there were feature selection approaches that either were equally good or considerably better than including all genes, which again suggests that feature selection has potential. For example, for the STAD data, including all genes resulted in a partition no better than expected by *chance* (i.e., ARI close to 0) while the best feature selection approach resulted in a partition highly correlated with the partition defined by the gold standard (ARI = 0.66). On the other hand, applying feature selection often resulted in a lower performance than including all genes in the analysis, which suggests that the

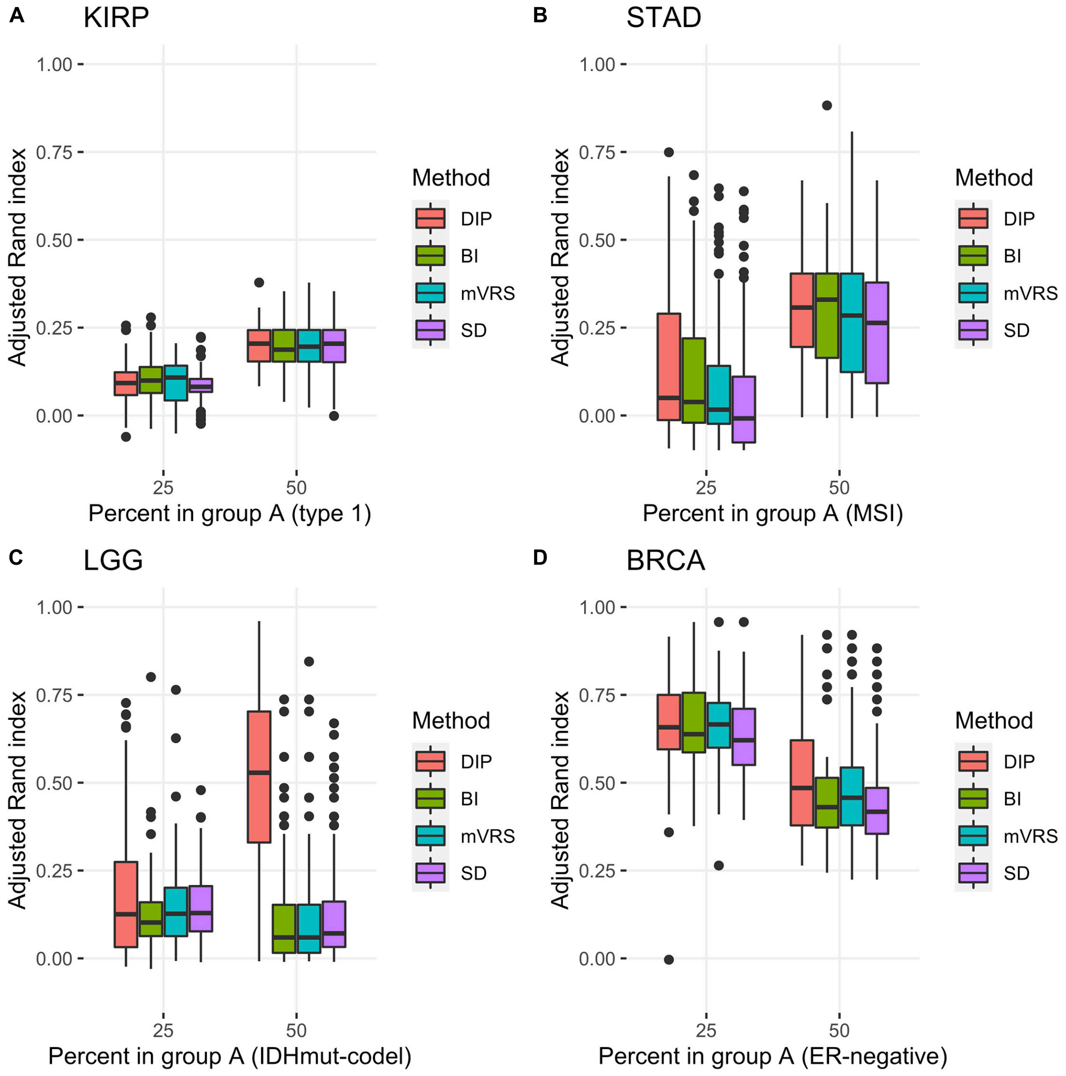


FIGURE 7 | Adjusted Rand index for clustering based on 1000 selected genes with different proportions of the subtypes based on 100 random samplings of the KIRP **(A)**, STAD **(B)**, LGG **(C)**, and BRCA **(D)** data. The figure shows results for unbalanced data with 25 (75) samples in the smaller (larger) subgroup and for balanced data sets with 50 samples in each subgroup. The selection methods on the *x*-axis are the dip-test statistic (DIP), bimodality index (BI), the modified variance reduction score (mVRS), and standard deviation (SD).

choice of feature selection method and the number of selected genes are important.

We included 13 variable selection methods that theoretically and methodologically can be grouped in four fundamentally different groups: methods that select highly expressed genes (M and Q3), methods that select highly variable genes (ENT, IQR, and SD), methods that select highly correlated genes (CoEx1 and CoEx2) and methods that select genes with respect to modality (BI, BC, DIP, mVRS, VRS, and wVRS). The correlation-based methods had surprisingly low performance, often worse than by selecting genes randomly. These methods were developed for variable selection in microarray data, which might explain the poor performance and suggest that these methods need to be modified for RNA-sequence data. Since CoEx1 and CoEx2 select genes with a relatively low SD across samples (**Figure 3**), a hybrid method that combines correlation and a bimodality score or measure of spread could be worth to investigate further. Selecting highly expressed genes is motivated by the fact that the signal to noise ratio is believed to be relatively high for highly expressed genes. Hence, by including highly expressed genes we get less noisy data and thereby better results. These methods, in particular Q3, worked surprisingly well and commonly better than selecting genes randomly. Although these methods did not perform as well as the best selection methods, the results suggest that they may work well in combined approaches as discussed at the end of this section.

An important finding is that the commonly used SD method did not perform well. One reason for this may be that SD compared to other methods is more likely to include genes with

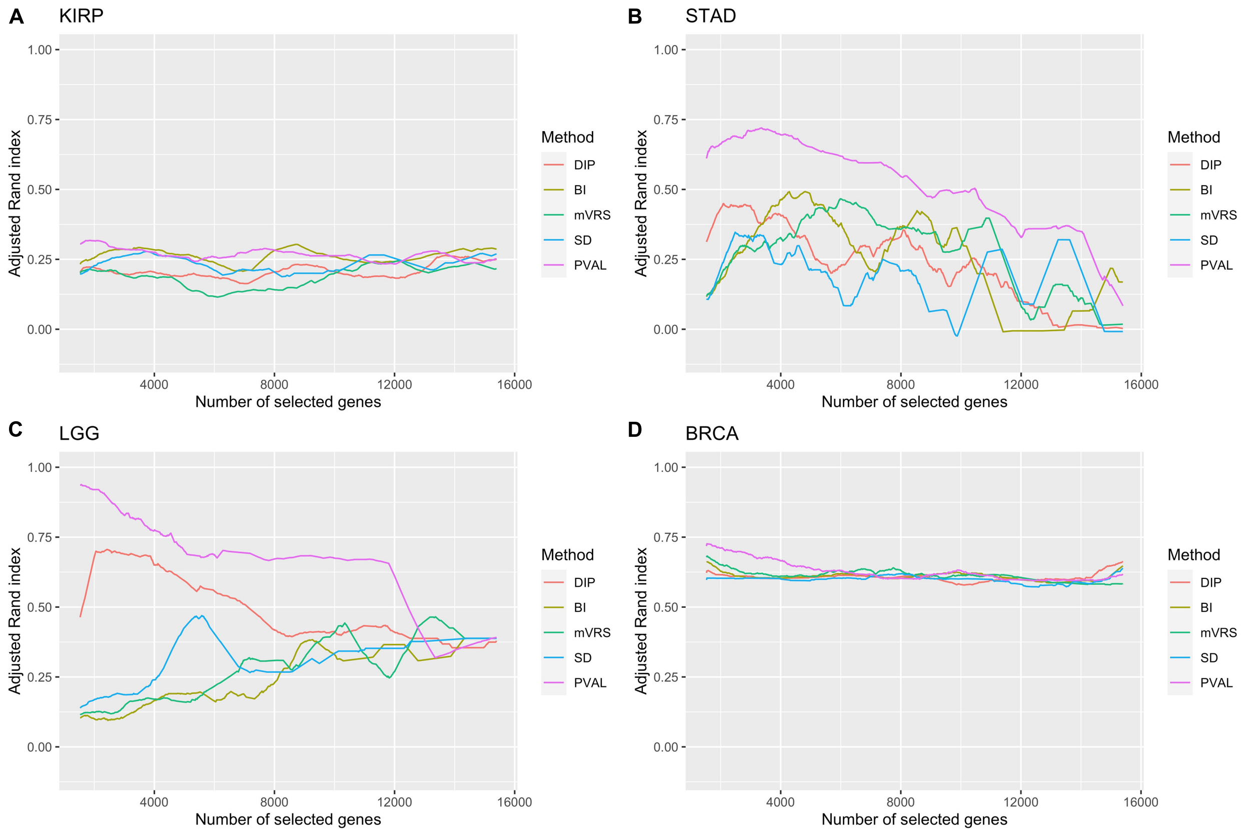


FIGURE 8 | Adjusted Rand index for a gradual increase in number of selected genes. The number of included genes was increased over a 1000 steps from 2 to up to all genes remaining after filtration. Of low expressed genes. The feature selection methods are: dip-test statistic (DIP), bimodality index (BI), modified variance reduction score (mVRS), standard deviation (SD), and the positive control (PVAL). The performance is shown for the KIRP **(A)**, STAD **(B)**, LGG **(C)**, and BRCA **(D)** data. A running mean over 100 values was applied to get a smoother curve.

outliers and extreme values. Samples with extreme values can govern the clustering and incorrectly result in a binary clustering where the smaller of the two groups contains a low number of individuals. The results showed that SD indeed had fewer samples in the minority group compared to DIP, BI, and mVRS, which may explain the ARI-results. Furthermore, the IQR that is a robust alternative to SD performed better than SD.

The best performing selection methods BI, DIP, and mVRS all aim to identify genes based on modality. With the exception of DIP, these methods strive to detect genes with a clear bi-modality pattern, while DIP is more general a search for multimodality patterns. For the case where 1000 genes were selected, DIP achieved the best performance and worked well for both balanced and skewed data sets. Interestingly, DIP had compared to BI and mVRS often more samples in the minority group obtained from the binary clustering.

For the LGG data with 1000/3000 selected genes, both the original data and the simulated data sets, DIP performed considerably better than BI and mVRS, which in turn performed worse than a strategy including all genes. Furthermore, the overlap between the genes selected by DIP and the two other methods was small, much smaller than observed for the other data sets, see **Figure 4**. This may indicate that the partition defined by the gold standard should be further divided, which in turn explain why methods searching for genes with a bimodal pattern fails. In general, all methods aiming to identify bimodality will suffer if the partitioning of interest consists of more than two groups, in particular if there exist one or more secondary factors that define two groups, e.g., gender.

TABLE 4 | Proportion of cancer driver genes among the 1000 top ranked genes.

	DIP	BI	mVRS	SD	PVAL
KIRP	0.013 (0.92)	0.011 (0.98)	0.011 (0.98)	0.013 (0.92)	0.009 (0.99)
STAD	0.018 (0.59)	0.015 (0.84)	0.015 (0.84)	0.023 (0.17)	0.021 (0.31)
LGG	0.021 (0.27)	0.016 (0.73)	0.015 (0.81)	0.013 (0.92)	0.021 (0.27)
BRCA	0.017 (0.68)	0.018 (0.59)	0.016 (0.77)	0.032 (0)	0.024 (0.12)

The proportions in the entire data setswere obtained as 0.018, 0.019, 0.018, and 0.019 for KIRP, STAD, LGG, and BRCA, respectively. The p-value from a one sided Fisher's exact test is given within paranthesis.

Some potential secondary factors and their partitions are sometimes known, e.g., the age and gender of the patients, prior treatments and technical design questions, e.g., which hospital analyzed the samples. This information can in principle be used when selecting the genes, e.g., by omitting genes that are highly correlated to any of the known secondary factors. Another way to improve feature selection may be to combine two or more approaches, e.g., demand that the selected genes are both highly expressed and have high DIP scores. How to include additional meta information and combine different selection methods are open questions that requires more research.

For most of the feature selection methods, the overlap between selected genes and previously identified cancer driver genes was relatively low. Lists of candidate cancer driver genes are continually updated as new discoveries are made and there are several published lists of genes that are important for cancer development. Comparing against alternative gene lists may affect the results. Moreover, many of the considered cancer driver genes are affected by cancer in general but are not necessarily informative for the partition of interest. We did observe an enrichment of cancer driver genes among the set of genes selected using M and Q3, suggesting that the confirmed cancer driver genes are in general expressed at higher levels.

Here k-means ($k = 2$) and hierarchical clustering with Ward's linkage and either the Euclidean distance or a correlation-based distance were used to cluster the samples. It should be stressed that the choice of clustering method may affect the relative performance of the considered feature selection methods. The choice was motivated by prior findings and since these approaches are widely used. The number of selected genes were 100, 1000, or 3000 and these choices were based on our prior experience (Vidman et al., 2019 and Freyhult et al., 2010). However, how to determine the optimal number of genes to include is an open question that needs more research. These choices affect the ARI-values and may also have an effect on the relative performance of the considered feature selection methods. Moreover, the performance of any clustering approach, including pre-processing, standardization, feature selection, and the clustering, is highly dependent on the data making it difficult to give general advices. Nevertheless, the results presented in this article suggest that variable selection using DIP with 1000 selected genes is a good choice and considerably better than selecting genes based on the observed SD.

The study focuses on the relative merits of feature selection strategies commonly categorized as filtering methods in the literature, and a direction of future research with great potential would be to investigate other classes of methods that have been developed in the field, for example, wrapper, ensemble, and hybrid methods (Ang et al., 2016).

CONCLUSION

Partitioning cancer patients based on RNA-seq data with the objective to identify subgroups is an important but also very challenging problem. The main difficulty is that only some genes are differentially expressed between the subgroups of interest and that several secondary factors affect gene expressions. Therefore, it is reasonable to assume that the clustering should be based on a set of carefully selected genes rather than all genes. The commonly used SD-approach, where genes with the highest SDs are selected, did not perform well in our study. We argue that SD is more likely to select genes affected by outliers, which in turn has a negative effect on the downstream cluster analysis. Although the performance in general is highly data-dependent, our study shows that selecting 1000 genes using the dip-test is a sensible selection approach, which performs considerably better than the SD-selection.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: the gene expression data in this paper are available at https://gdac.broadinstitute.org/ under cohorts BRCA, LGG, KIRP, and STAD.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

DK was responsible for the data collection, pre-processing, and implementation of methods. All authors were involved in the evaluation, interpreting results, and preparing the final manuscript.

FUNDING

This work was supported by grants from the Swedish Research Council, Dnr 340-2013-5185 (LV and PR), the Kempe Foundations, Dnr JCK-1315 (DK and PR), and the Faculty of Science and Technology, Umeå University (DK and PR).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.632620/full#supplementary-material

REFERENCES

Abusamra, H. (2013). A comparative study of feature selection and classification methods for gene expression data of glioma. *Procedia Comput. Sci.* 23, 5–14. doi: 10.1016/j.procs.2013.10.003

Ang, J. C., Mirzal, A., Haron, H., and Hamed, H. N. A. (2016). Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection.

IEEE/ACM Trans. Comput. Biol. Bioinform. 13, 971–989. doi: 10.1109/TCBB.2015.2478454

Arun Kumar, C., Sooraj, M. P., and Ramakrishnan, S. (2017). A comparative performance evaluation of supervised feature selection algorithms on microarray datasets. *Procedia Comput. Sci.* 115, 209–217. doi: 10.1016/j.procs.2017.09.127

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* 173, 371–385.e18. doi: 10.1016/j.cell.2018.02.060

Benaglia, T., Chauveau, D., Hunter, D., and Young, D. (2009). mixtools: an R package for analyzing finite mixture models. *J. Stat. Softw.* 32, 1–29. doi: 10.18637/jss.v032.i06

Bentink, S., Haibe-Kains, B., Risch, T., Fan, J.-B., Hirsch, M. S., Holton, K., et al. (2012). Angiogenic mRNA and microRNA gene expression signature predicts a novel subtype of serous ovarian cancer. *PLoS One* 7:e30269. doi: 10.1371/journal.pone.0030269

Bertucci, F., Finetti, P., Rougemont, J., Charafe-Jauffret, E., Cervera, N., Tarpin, C., et al. (2005). Gene expression profiling identifies molecular subtypes of inflammatory breast cancer. *Cancer Res.* 65, 2170–2178. doi: 10.1158/0008-5472.Can-04-4115

Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Norwell, MA: Kluwer Academic Publishers.

Brat, D. J., Verhaak, R. G., Aldape, K. D., Yung, W. K., Salama, S. R., Cooper, L. A., et al. (2015). Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *N. Engl. J. Med.* 372, 2481–2498. doi: 10.1056/NEJMoa1402121

Cancer Genome Atlas Research Network (2014). Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* 513, 202–209. doi: 10.1038/nature13480

Cilia, N. D., Stefano, C. D., Fontanella, F., Raimondo, S., and Scotto di Freca, A. (2019). An experimental comparison of feature-selection and classification methods for microarray datasets. *Information* 10:109. doi: 10.3390/info10030109

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.* 95, 14863–14868. doi: 10.1073/pnas.95.25.14863

Freyhult, E., Landfors, M., Önskog, J., Hvidsten, T. R., and Rydén, P. (2010). Challenges in microarray class discovery: a comprehensive examination of normalization, gene selection and clustering. *BMC Bioinformatics* 11:503. doi: 10.1186/1471-2105-11-503

Fujikado, N., Saijo, S., and Iwakura, Y. (2006). Identification of arthritis-related gene clusters by microarray analysis of two independent mouse models for rheumatoid arthritis. *Arthritis Res. Ther.* 8:R100. doi: 10.1186/ar1985

Gine, E., and Nickl, R. (2008). A simple adaptive estimator of the integrated square of a density. *Bernoulli* 14, 47–61. doi: 10.3150/07-BEJ110

Hartigan, J. A., and Hartigan, P. M. (1985). The dip test of unimodality. *Ann. Stat.* 13, 70–84. doi: 10.1214/aos/1176346577

Hartigan, J. A., and Wong, M. A. (1979). Algorithm AS 136: a K-means clustering algorithm. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* 28, 100–108. doi: 10.2307/2346830

Hellwig, B., Hengstler, J. G., Schmidt, M., Gehrmann, M. C., Schormann, W., and Rahnenführer, J. (2010). Comparison of scores for bimodality of gene expression distributions and genome-wide evaluation of the prognostic relevance of high-scoring genes. *BMC Bioinformatics* 11:276. doi: 10.1186/1471-2105-11-276

Hubert, L., and Arabie, P. (1985). Comparing partitions. *J. Classif.* 2, 193–218. doi: 10.1007/BF01908075

Karlis, D., and Xekalaki, E. (2003). Choosing initial values for the EM algorithm for finite mixtures. *Comput. Stat. Data Anal.* 41, 577–590. doi: 10.1016/S0167-9473(02)00177-9

Kim, S., Kim, A., Shin, J.-Y., and Seo, J.-S. (2020). The tumor immune microenvironmental analysis of 2,033 transcriptomes across 7 cancer types. *Sci. Rep.* 10:9536. doi: 10.1038/s41598-020-66449-0

Kumari, S., Nie, J., Chen, H.-S., Ma, H., Stewart, R., Li, X., et al. (2012). Evaluation of gene association methods for coexpression network construction and biological knowledge discovery. *PLoS One* 7:e50411. doi: 10.1371/journal.pone.0050411

Lapointe, J., Li, C., Higgins, J. P., van de Rijn, M., Bair, E., Montgomery, K., et al. (2004). Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc. Natl. Acad. Sci. U.S.A.* 101, 811–816. doi: 10.1073/pnas.0304146101

Liu, X., Krishnan, A., and Mondry, A. (2005). An entropy-based gene selection method for cancer classification using microarray data. *BMC Bioinformatics* 6:76. doi: 10.1186/1471-2105-6-76

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550. doi: 10.1186/s13059-014-0550-8

Maechler, M. (2013). *diptest: Hartigan's Test Statistic for Unimodality – Corrected Code R* Package Version 0.75-5.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2019). *cluster: Cluster Analysis Basics and Extensions.* R Package Version 2.1.0.

Moody, L., Mantha, S., Chen, H., and Pan, Y.-X. (2019). Computational methods to identify bimodal gene expression and facilitate personalized treatment in cancer patients. *J. Biomed. Inform. X* 1:100001. doi: 10.1016/j.yjbinx.2018.100001

Önskog, J., Freyhult, E., Landfors, M., Rydén, P., and Hvidsten, T. R. (2011). Classification of microarrays; synergistic effects between normalization, gene selection and machine learning. *BMC Bioinformatics* 12:390. doi: 10.1186/1471-2105-12-390

Pertea, M., Shumate, A., Pertea, G., Varabyou, A., Chang, Y.-C., Madugundu, A. K., et al. (2018). Thousands of large-scale RNA sequencing experiments yield a comprehensive new human gene list and reveal extensive transcriptional noise. *bioRxiv* [preprint] doi: 10.1101/332825

Ren, Z., Wang, W., and Li, J. (2016). Identifying molecular subtypes in human colon cancer using gene expression and DNA methylation microarray data. *Int. J. Oncol.* 48, 690–702. doi: 10.3892/ijo.2015.3263

SAS Institute (1990). *SAS/STAT User's Guide: Version 6 4:th*. Cary, NC: SAS Institute Inc.

Sathish, D., and 4D Strategies (2016). *modes: Find the Modes and Assess the Modality of Complex and Mixture Distributions, Especially with Big Datasets R package version 0.7.0*.

Seal, D. B., Saha, S., Mukherjee, P., Chatterjee, M., Mukherjee, A., and Dey, K. N. (2016). "Gene ranking: an entropy & decision tree based approach," in *Proceedings of the 2016 IEEE 7th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, New York, NY, 1–5.

Shen, R., Li, P., Li, B., Zhang, B., Feng, L., and Cheng, S. (2020). Identification of distinct immune subtypes in colorectal cancer based on the stromal compartment. *Front. Oncol.* 9:1497. doi: 10.3389/fonc.2019.01497

Sotiriou, C., Neo, S.-Y., McShane, L. M., Korn, E. L., Long, P. M., Jazaeri, A., et al. (2003). Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc. Natl. Acad. Sci. U.S.A.* 100, 10393–10398. doi: 10.1073/pnas.1732912100

The Cancer Genome Atlas Research Network (2016). Comprehensive molecular characterization of papillary renal-cell carcinoma. *N. Engl. J. Med.* 374, 135–145. doi: 10.1056/NEJMoa1505917

Vidman, L., Källberg, D., and Rydén, P. (2019). Cluster analysis on high dimensional RNA-seq data with applications to cancer research – An evaluation study. *PLoS One* 14:e0219102. doi: 10.1371/journal.pone.0219102

Wang, J., Wen, S., Symmans, W. F., Pusztai, L., and Coombes, K. R. (2009). The bimodality index: a criterion for discovering and ranking bimodal signatures from cancer gene expression profiling data. *Cancer Inform.* 7, 199–216.

Wang, Z., Lucas, F. A. S., Qiu, P., and Liu, Y. (2014). Improving the sensitivity of sample clustering by leveraging gene co-expression networks in variable selection. *BMC Bioinformatics* 15:153. doi: 10.1186/1471-2105-15-153

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

ORIGINAL RESEARCH
published: 22 March 2021
doi: 10.3389/fgene.2021.652315

Risk Prediction in Patients With Heart Failure With Preserved Ejection Fraction Using Gene Expression Data and Machine Learning

Liye Zhou[1†], *Zhifei Guo*[1†], *Bijue Wang*[1], *Yongqing Wu*[2], *Zhi Li*[3], *Hongmei Yao*[4], *Ruiling Fang*[2], *Haitao Yang*[5], *Hongyan Cao*[2,6]* and *Yuehua Cui*[7]*

[1] Division of Health Management, School of Management, Shanxi Medical University, Taiyuan, China, [2] Division of Health Statistics, School of Public Health, Shanxi Medical University, Taiyuan, China, [3] Department of Hematology, Taiyuan Central Hospital of Shanxi Medical University, Taiyuan, China, [4] Department of Cardiology, First Hospital of Shanxi Medical University, Taiyuan, China, [5] Division of Health Statistics, School of Public Health, Hebei Medical University, Shijiazhuang, China, [6] Key Laboratory of Major Disease Risk Assessment, Shanxi Medical University, Taiyuan, China, [7] Department of Statistics and Probability, Michigan State University, East Lansing, MI, United States

OPEN ACCESS

Edited by:
Chao Xu,
University of Oklahoma Health
Sciences Center, United States

Reviewed by:
Zhongshang Yuan,
Shandong University, China
Chuan Qiu,
Tulane University, United States

***Correspondence:**
Hongyan Cao
caohy@sxmu.edu.cn
Yuehua Cui
cuiy@msu.edu

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to
Statistical Genetics and Methodology,
a section of the journal
Frontiers in Genetics

Received: 12 January 2021
Accepted: 02 March 2021
Published: 22 March 2021

Citation:
Zhou L, Guo Z, Wang B, Wu Y, Li Z, Yao H, Fang R, Yang H, Cao H and Cui Y (2021) Risk Prediction in Patients With Heart Failure With Preserved Ejection Fraction Using Gene Expression Data and Machine Learning. Front. Genet. 12:652315. doi: 10.3389/fgene.2021.652315

Heart failure with preserved ejection fraction (HFpEF) has become a major health issue because of its high mortality, high heterogeneity, and poor prognosis. Using genomic data to classify patients into different risk groups is a promising method to facilitate the identification of high-risk groups for further precision treatment. Here, we applied six machine learning models, namely kernel partial least squares with the genetic algorithm (GA-KPLS), the least absolute shrinkage and selection operator (LASSO), random forest, ridge regression, support vector machine, and the conventional logistic regression model, to predict HFpEF risk and to identify subgroups at high risk of death based on gene expression data. The model performance was evaluated using various criteria. Our analysis was focused on 149 HFpEF patients from the Framingham Heart Study cohort who were classified into good-outcome and poor-outcome groups based on their 3-year survival outcome. The results showed that the GA-KPLS model exhibited the best performance in predicting patient risk. We further identified 116 differentially expressed genes (DEGs) between the two groups, thus providing novel therapeutic targets for HFpEF. Additionally, the DEGs were enriched in Gene Ontology terms and Kyoto Encyclopedia of Genes and Genomes pathways related to HFpEF. The GA-KPLS-based HFpEF model is a powerful method for risk stratification of 3-year mortality in HFpEF patients.

Keywords: risk prediction, kernel partial least squares, genetic algorithm, heart failure with preserved ejection fraction, machine learning

INTRODUCTION

Heart failure (HF) is the leading cause of death and disability worldwide among older adults (Manolis et al., 2019). Over 50% of patients with HF exhibit heart failure with preserved ejection fraction (HFpEF; Komajda et al., 2011; Rich et al., 2018), and the prevalence of HFpEF is increasing relative to heart failure with reduced ejection fraction (HFrEF) at an alarming rate of 1% per year (Monika et al., 2018). HFpEF is a heterogeneous syndrome that contributes to abnormal cardiac structure or function, seriously endangering human health (Antlanger et al., 2017; Garg et al., 2017). HFpEF patients have a poor prognosis, and the 5-year mortality rate of HFpEF is as high as 50% (Shah et al., 2017). While the mortality rate of HFrEF has significantly decreased over the past few years because of specific HFrEF treatments (Loh et al., 2013), no effective treatment has been identified for HFpEF patients (Shah et al., 2014). Arguably, with an aging population worldwide, the emerging epidemic of HFpEF requires urgent attention to determine methods for faster disease risk assessment and to predict clinical outcomes to guide therapy, monitoring, and patient management.

While numerous risk assessment models have been developed in cohorts with HFrEF or a mixture of HFrEF and HFpEF, risk prediction in HFpEF patients has been less studied (Thorvaldsen et al., 2017; Angraal et al., 2020). This may be associated with the poor prognostic factors used to predict HFpEF patients (Kanda et al., 2018). The existing risk assessment models for HFpEF are predominantly based on clinical phenotype data, such as baseline demographic and clinical data and electrocardiographic, echocardiographic, and laboratory testing data (Komajda et al., 2011; Thorvaldsen et al., 2017; Rich et al., 2018; Angraal et al., 2020). Unfortunately, these models constructed using clinical phenotypic data have low sensitivity or specificity, and patients are likely to be misdiagnosed. No model has gained widespread acceptance to date. The estimate of an HFpEF patient's prognosis in daily practice is still mainly based on the experience of clinicians (Ferrero et al., 2015; Thorvaldsen et al., 2017; Manolis et al., 2019). A great need exists to develop an effective risk model for HFpEF to aid in the design of future clinical trials.

With advances in sequencing and computer technology, high throughput expression data can be extracted without limits. Genomic measures of gene expression offer rich information about the underlying disease mechanism and have provided new possibilities of using these molecular data to understand the disease gene function and further predict disease outcomes (Haring and Wallaschofski, 2012). Based on the expression data, great efforts have been devoted to disease classification, clinical outcome prediction, and the identification of genes with potential therapeutic molecular signatures (Penney et al., 2011; Khan et al., 2012; Vargas and Lima, 2013; Wang et al., 2019). HFpEF is a complicated clinical syndrome with high molecular heterogeneity and diverse manifestations (Shah et al., 2015) and is further complicated with a potentially nonlinear relationship between genes and the clinical outcome. Thus, conventional generalized linear models (e.g., logistic regression) are poor choices for risk prediction. Advanced statistical techniques and machine learning methods show great potential in improving the classification performance over conventional statistical tools through the nonlinear effects of variables to achieve accurate prediction (Angraal et al., 2020) and should be studied for HFpEF prediction.

The purpose of this work is to evaluate six different risk stratification models and to predict the survival risk of HFpEF patients based on gene expression profiles using data from a high-quality epidemiologic study, the Framingham Heart Study (FHS). We applied five advanced machine learning methods [i.e., kernel partial least squares based on the genetic algorithm (GA-KPLS), random forest (RF), the least absolute shrinkage and selection operator (LASSO), ridge regression (RR), support vector machine (SVM), and a conventional logistic regression model (Logit)] to build an optimal risk stratification model. Identification of patients with a high risk of HFpEF will be helpful for targeted interventions and clinical trials to further improve the survival of HFpEF patients.

MATERIALS AND METHODS

Data

Framingham Heart Study

The FHS data used in this study included clinical, survival, and expression data downloaded from dbGAP (study accession: phs000007, http://dbgap.ncbi.nlm.nih.gov). The FHS has recruited participants from Framingham, MA, United States, to undergo biennial examinations to investigate cardiovascular disease and its risk factors since 1948 (Oppenheimer, 2005). Offspring (and their spouses) and adult grandchildren of the original cohort of participants were recruited into the second- and third-generation cohorts in 1971 and 2002, respectively (Yao et al., 2015). In this study, the clinical and gene expression data were obtained from the offspring cohort who (i) attended the eighth examination cycle conducted between 2005 and 2008 and (ii) had both clinical and gene expression profiles.

HFpEF Patients

According to the guidelines of the European Society of Cardiology (McMurray et al., 2018), patients were diagnosed with HFpEF using the following four conditions: (1) typical signs or symptoms of HF, (2) B-type natriuretic peptide >35 pg/ml and/or N-terminal-pro hormone B-type natriuretic peptide >125 pg/ml, (3) left ventricular ejection fraction >50%; and (4) structural HF (left ventricular hypertrophy/left atrial enlargement) and/or diastolic dysfunction. We excluded patients with valvular stroma and/or hypertrophic cardiomyopathy, resulting in inclusion of 172 HFpEF patients (103 males and 69 females). Patients whose 3-year survival status was unknown were filtered out by design (Fransen et al., 2011). Finally, 149 individuals (91 males and 58 females) who had full survival information after 3 years were included in the study.

Gene Expression Data

The expression data contained 17,873 gene expression probes. We mapped these probes to genes following the annotation from the Affymetrix Human Exon 1.0 ST GeneChip platform, which yielded 17,358 genes. The gene expression data were $log_2\ (x + 1)$ transformed and then standardized (Cheerla and Gevaert, 2017). A variable screening procedure called as sure independence screening was applied to reduce the gene expression dimensionality from an ultra-high to a moderate scale, with a binary response defined as a "good outcome" or "poor outcome" for each individual. Following the sure independence screening criterion {i.e., keeping $d = [2n/log(n)]$ features; Fan and Lv, 2008}, the top 137 features were retained for further analysis.

Clinical Outcome

The clinical outcome was defined as a good or poor outcome based on patients' survival status. The good-outcome group had event-free survival for at least 3 years [survival time was measured from the time of admission for HFpEF diagnosis to the time of last follow-up (2011) or time of death from cardiovascular disease]. The poor-outcome group included patients who died because of cardiovascular disease during the 3-year period. We further explored the differentially expressed genes (DEGs) between the good-outcome and poor outcome groups using significance analysis of microarrays (Tusher et al., 2001) and then conducted Gene Ontology (GO) enrichment analysis and the Kyoto Encyclopedia of the Genes and Genomes (KEGG) pathway analysis based on the DEGs using KOBAS software[1] (Ai and Kong, 2018).

Statistical Analysis

KPLS Prediction Model Optimized With the Genetic Algorithm

The kernel partial least squares method can map the original data points from the original input space R^N into a high-dimensional feature space F, and therefore, original data that cannot be linearly separated in R^N can be separated in F (Rosipal and Trejo, 2002), which improves the classification performance to achieve accurate prediction. A genetic algorithm (GA) is an optimization method based on the genetic mechanism of "survival of the fittest." In this study, we used a Gaussian kernel function to construct the kernel matrix for gene expression data and then used the genetic algorithm to optimize the Gaussian kernel function parameter σ. The Gaussian kernel function is given as $K\left(x_i, x_j\right) = \exp\left(-\left\|x_i - x_j\right\|^2 / 2\sigma^2\right)$. For the details of the method, readers are referred to Yang et al. (2020). Because we only used gene expression data for prediction, the only parameter that needed to be optimized was the kernel bandwidth σ.

Other Prediction Models

Ridge regression and LASSO fit prediction models by shrinkage or regularization of the regression coefficients (Frank and Friedman, 1993; Tibshirani, 1996). The LASSO method can shrink some coefficients to exactly zero. Both models were developed to minimize prediction errors. For the LASSO and RR methods, the optimal tuning parameter λ was chosen by 10-fold cross-validation over a grid of 100 λ values. The RR and LASSO methods were performed using the R glmnet package.

The SVM method was developed to solve high-dimensional classification problems (Furey et al., 2000) and was performed using the R e1071 package. The radial basis kernel function was used in the SVM.

An RF uses the bootstrap method to extract n samples from the original data and generate B classification trees. These B trees constitute a random forest. Each observation's predictive result is determined by a majority vote; the overall prediction is the most commonly occurring class among the B classification trees (Austin et al., 2013). The RF method was performed using the randomForest package in R. All parameter values were set using the default.

Model Training and Testing

In our study, the original data were divided into two non-overlapping data sets: modeling data and external testing data. We randomly selected modeling data and external testing data at a ratio of 80:20. The modeling set was used to train the prediction model, and the testing set was used to evaluate the prediction performance. The entire process of randomly selecting the modeling and testing data was repeated 1,000 times to increase the stability and repeatability of the results.

Model Performance

We used multiple evaluation criteria to evaluate the predictive performances of the six models, including the area under the curve (AUC), sensitivity (Se), specificity (Sp), accuracy (ACC), Youden index, G-means, and Matthews correlation coefficient (MCC). The MCC and AUC were mainly used to evaluate the model performance because they are more comprehensive evaluation criteria. We employed one-way ANOVA, followed by Dunnett's multiple-comparison test, to compare the performance of the GA-KPLS and the five other models (RF, LASSO, RR, Logit, and SVM). Statistical significance was indicated by a value of $p < 0.05$.

RESULTS

Characteristics of HFpEF Patients in the FHS

At the end of the 3-year period, 42 patients (28.19%) met the study endpoint of cardiovascular disease-related death, and 107 patients (71.81%) had survived. There were 91 males (61.07%) and 58 females (38.93%). The average age was 75.02 (±8.02) years old. **Table 1** shows the baseline condition of both groups, patients with good outcomes, and those with poor outcomes. There was no significant difference in age, gender, comorbidities, vital signs, or laboratory data (except for systolic blood pressure) between the two groups.

Model Performance Comparison

We compared the classification performance of the six models: GA-KPLS, RF, LASSO, RR, SVM, and Logit. The evaluation index of the six models was summarized as the average value obtained by repeating the data partition 1,000 times. **Table 2** shows the prediction results of the six models. As shown in the table, the GA-KPLS model exhibited the best performance

[1] http://kobas.cbi.pku.edu.cn

TABLE 1 | Clinical characteristics of the study population (*N* = 149).

Characteristic	Good-outcome group (107)	Poor-outcome group (42)	χ^2/t	p-value
Age, years	74.44 ± 8.23	76.50 ± 7.46	0.572	0.568
Female, *n* (%)	40(37.4)	18(42.9)	0.380	0.538
Comorbidities, *n* (%)				
Hypertension	84(78.5)	33(78.6)	<0.001	0.993
Hyperlipidemia	70(65.4)	26(61.9)	0.163	0.687
Diabetes	27(25.2)	11(26.2)	0.015	0.904
Vital signs and laboratory data				
Systolic blood pressure, mmHg*	127.74 ± 18.44	138.88 ± 22.71	−3.102	0.002
Diastolic blood pressure, mmHg	65.64 ± 11.58	67.83 ± 9.55	−1.08	0.279
Body mass index, kg/m^2	29.84 ± 5.47	29.21 ± 5.68	0.633	0.528
Serum creatinine, mg/dl	1.24 ± 0.86	1.29 ± 0.88	0.288	0.774
Total cholesterol, mg/dl	162.12 ± 36.70	167.74 ± 41.31	−0.811	0.419
Heart rate, bpm	62.50 ± 10.90	64.45 ± 12.97	−0.929	0.354

**Shows the statistical significance at the α = 0.05 level.*

TABLE 2 | Model performance.

Model	Se	Sp	AUC	ACC	Youden	F-measure	MCC	G-means
GA-KPLS	0.925	0.984	0.955	0.968	0.909	0.939	0.921	0.953
RF	0.319	0.974	0.646	0.793	0.293	0.445	0.427	0.535
LASSO	0.605	0.943	0.774	0.850	0.548	0.678	0.608	0.745
RR	0.469	1.000	0.734	0.853	0.469	0.618	0.620	0.669
Logit	0.549	0.574	0.591	0.567	0.122	0.410	0.112	0.548
SVM	0.870	0.989	0.929	0.956	0.859	0.913	0.891	0.926

in nearly all the criteria except for specificity. This finding clearly demonstrates the superior performance of the GA-KPLS model. To further display the prediction results, we chose the evaluation criterion AUC to demonstrate the performance obtained by 1,000 random splits (see **Figure 1**). The AUC of the GA-KPLS model was significantly different from those of the RF, LASSO, RR, Logit, and SVM models, indicating the superior performance of the GA-KPLS model over the other models. It is interesting to note that the performance of the SVM model was quite similar to that of the GA-KPLS model. Based on the results, we concluded that the risk prediction model constructed by the GA-KPLS method had the best performance and can provide a methodological reference to assess the risk of HFpEF.

Prediction Result of HFpEF Using the GA-KPLS Method

To demonstrate the clinical significance of identifying high-risk patients, we selected the prediction result of one random split with 120 training samples and 29 testing samples, which gave an MCC = 0.920 (close to MCC_{mean} = 0.921). The Kaplan-Meier curves based on the original and predicted data yielded significantly different survival probabilities ($p < 0.0001$). **Figure 2** shows the survival curves of the two groups. The left panel shows the survival curve from the original data, and the right panel shows the survival curve based on the newly predicted risk group with the GA-KPLS method. The prediction method exhibited good performance because the survival curves using the original and predicted values were very similar. To predict a future event, all the data can be used as the training set, and then the risk group status can be predicted based on measured gene expression data.

DEGs Between the Good-Outcome and Poor-Outcome Patients

We treated the good-outcome group as the control group to identify DEGs. Of a total of 137 top genes, 116 DEGs were identified based on a threshold value of $q < 0.05$, among which 70 genes were upregulated and 46 were downregulated. The significant features of gene expression are shown in a heat map (see **Figure 3**). A block-like structure can be observed between the good-outcome and poor-outcome groups.

Among the 116 DEGs, the *TRAF3IP2*, *C1QTNF9*, *TECRL*, and *Eph* genes have been reported to be associated with HF. *TRAF3IP2* is an upstream regulator of multiple proinflammatory pathways. *TRAF3IP2* overexpression may activate IKK/NF-B, p38 MAPK, and JNK/AP-1 and induce proinflammatory cytokines, leading to cardiac fibrosis and contractile dysfunction (Yariswamy et al., 2016). *C1QTNF9* (*CTRP9*) is an important member of the *CTRP* protein family. Appari et al. (2016) found that *C1QTNF9* knock-out mice were protected from left ventricular dilatation and contractile dysfunction; however, *C1QTNF9* overexpression promoted ventricular remodeling and systolic dysfunction. *TECRL* was recently suggested to play a key role in the electrical activity of the heart. *TECRL* affects the electrical conduction system of the heart by causing mutations in a calcium-processing protein, which eventually leads to arrhythmia (Perry and Vandenberg, 2016). The Eph/ephrin receptor ligand comprises the largest family of receptor tyrosine kinases and affects the behavior of cells mainly by activating signal transduction pathways. Eph/ephrin expression may lead to phenotypic changes in the vascular endothelium during inflammation,

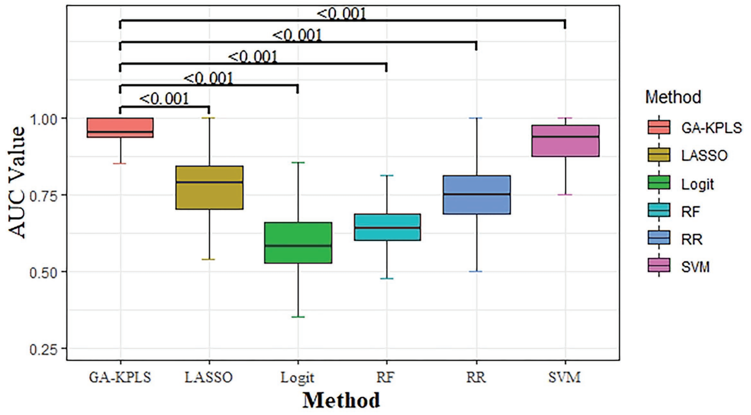


FIGURE 1 | Boxplot of the area under the curve (AUC) values for the six different models (based on 1,000 random splits). The *y*-axis represents the AUC value. Values of *p* were obtained using Dunnett's multiple-comparison test.

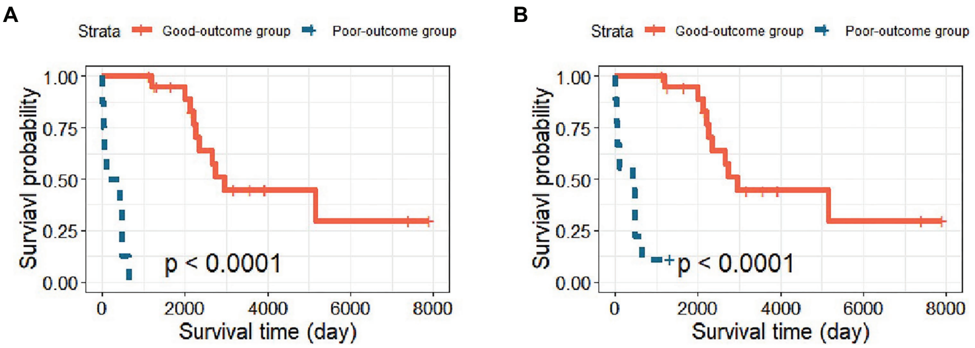


FIGURE 2 | Kaplan-Meier survival curves of the good-outcome and poor-outcome groups. **(A)** The survival curve including the original 29 patients in the testing cohort and **(B)** the survival curve based on the predicted survival outcomes using the GA-KPLS method.

causing inflammatory cells to enter the interstitial tissue from the vascular space (Coulthard et al., 2012).

The role of *DUSP1* is controversial, as both anti-inflammatory and pro-atherosclerotic actions have been suggested (Hahn et al., 2014). Auger-Messier et al. (2013) suggested that the disruption of *DUSP1* promoted p38 MAPK activity, which could reduce cardiac contractility and calcium handling; thus, *DUSP1* could be a target gene for prevention of HF. In addition, *LHFPL2* and *SNX24* are associated with coronary artery disease (Lin et al., 2013; Shendre et al., 2017). *HIST1H4B* is associated with the immune process (Zhang et al., 2019). *OXER1* is involved in the inflammatory response of the disease (Dattilo et al., 2015). The empirical evidence suggests the importance of the identified DEGs associated with HFpEF.

Functional Analysis of DEGs

To further investigate the functional relevance of the DEGs, we performed GO enrichment and KEGG pathway analyses. The DEGs were significantly enriched in 12 GO terms, with a *corrected value of p* < 0.05. GO terms comprised three categories: biological process, cell component, and molecular function. **Figure 4** shows all significant GO terms. The most significantly enriched GO terms were plasma membrane (*corrected value of p* = 2.67E–07), G protein-coupled receptor signaling pathway (*corrected value of p* = 3.06E−04), and protein binding (*corrected value of p* = 3.06E−04). The plasma membrane plays important roles in maintaining homeostasis, cell material exchange, and information transmission (Lutz et al., 2003; Wang et al., 2017). The G protein-coupled receptor signaling pathway mediates cardiac functions, such as those of inotropy and vasodilation in peripheral vessels, participates in the occurrence and development of HF and may serve as the molecular underpinning for future HF therapeutics (Wang et al., 2018; Altamish et al., 2020). Protein binding, including fatty acid-binding proteins, has been related to cardiac alterations, e.g., systolic and diastolic cardiac dysfunction (Rodriguez-Calvo et al., 2017). In the KEGG analysis, the olfactory transduction pathway was identified, with a *corrected value of p* < 0.05. The olfactory system uses G protein-coupled receptors to accomplish its vital task (Ronnett and Moon, 2002).

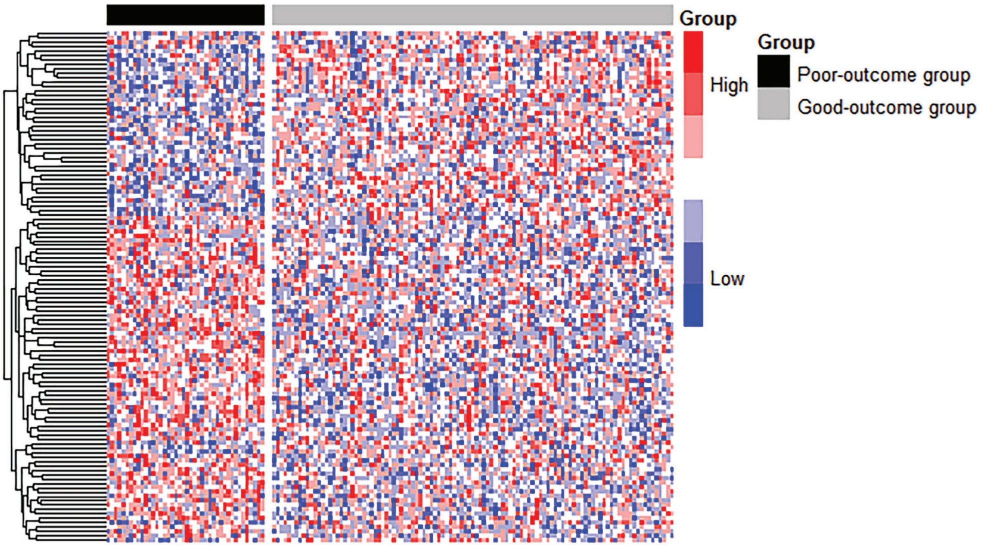


FIGURE 3 | The heatmap of DEGs between the good-outcome and poor-outcome groups. Each column represents a patient, and each row represents a gene. Patients labeled with the black bar are poor-outcome samples, and those with the gray bar are good-outcome samples.

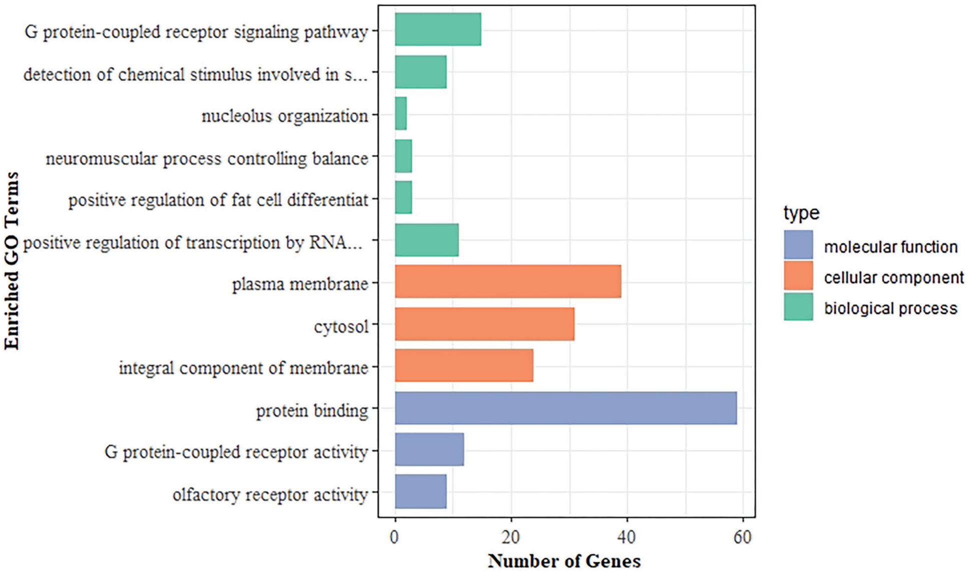


FIGURE 4 | Gene Ontology (GO) enrichment analysis of DEGs. The x-axis shows the number of genes, and the y-axis indicates the GO terms. Bars with different colors correspond to different GO categories, with green representing biological process, orange representing cellular component, and blue representing molecular function.

DISCUSSION

Accurately predicting disease outcomes are essential for patient-centered care, both for making treatment decisions and monitoring the quality of health care (Angraal et al., 2020). Using the gene expression data of HFpEF patients, this study explored five machine learning methods and one conventional logistic regression model to predict the survival status of patients with HFpEF. The GA-KPLS based HFpEF model could predict patient survival status with high accuracy. Furthermore, the identification of molecular markers (i.e., DEGs) of HFpEF may lead to the development of novel targeted therapies.

The ability to assess survival outcomes of patients with cardiovascular diseases has great clinical value in an era with multiple treatment options. Although previous studies have devoted great effort to predicting clinical outcomes of

HF patients, the current study has several unique merits. There are many studies being conducted to predict HF. However, few studies are focused on HFpEF. By evaluating six models, we showed that the GA-KPLS model using gene expression data may be a powerful and highly accurate prediction model of survival status in HFpEF patients. A prediction model using gene expression data can be an alternative means to the currently used models based on clinical data, such as the Enhanced Feedback for Effective Cardiac (EFFECT) study risk scores (Thorvaldsen et al., 2017) and Meta-Analysis Global Group in Chronic Heart Failure (MAGGIC) scores (Pocock et al., 2013).

Second, because of the highly heterogenous nature of HFpEF, a consensus has not been reached on which predictors can be used to reliably predict HFpEF. We demonstrated that gene expression can be used to predict HFpEF survival status with high accuracy using the GA-KPLS prediction model. With the availability of increasing types of omics data (e.g., copy number variants, microRNAs, and epigenetic data), we can further improve the prediction accuracy by integrating different data sources with the GA-KPLS model. Our study illustrates the development of new machine learning methods for HFpEF risk prediction by integrating different omics data types.

Current studies have focused on single or multiple clinical indicators to identify patients at high risk for HFpEF. However, most methods can only achieve an AUC of 0.7, which is unrealistic for application in clinical practice (Kanda et al., 2018; Shen et al., 2020). Many researchers have also used statistical methods to construct stratification models such as Cox proportional hazards models and logistic regression models. However, these methods fail to capture the nonlinear relationship between predictors and the disease outcome (Komajda et al., 2011; Rich et al., 2018; Angraal et al., 2020). In contrast, the GA-KPLS model uses the advantage of kernel functions to extract nonlinear relationships between genomic features and survival outcomes, hence achieving more accurate predictions than its counterparts.

Risk prediction in HFpEF patients using the GA-KPLS model may (1) serve to motivate patients to adhere to recommended treatments and lifestyle modifications (Oktay et al., 2013); (2) help clinicians to make treatment decisions, especially for high-risk groups of patients who may progress to circulatory failure when administered routine clinical therapeutics, and these patients may have the opportunity to undergo active therapeutic interventions such as mechanical circulatory assistance, heart transplantation, or new trials (Wang et al., 2019); and (3) help to inform the design of future HFpEF clinical trials.

However, our study had some limitations. First, because of the lack of additional external data on HFpEF, we cannot validate our findings in another data set. Second, we focused on gene expression data in our study. As lifestyle is an important risk factor for HF, further research should be performed to predict HFpEF risk by integrating both clinical and genomic data to improve the prediction performance because potential interactions may exist between these factors. Third, the HFpEF data set is imbalanced, with a ratio of 28:72 between the poor-outcome and good-outcome groups. However, the GA-KPLS and SVM methods performed well, with high sensitivity and specificity. If either low sensitivity or specificity becomes a concern, the SMOTE algorithm can be applied (Chawla et al., 2002), which is designed to handle prediction with imbalanced data.

In conclusion, the GA-KPLS-based HFpEF prediction model using gene expression data represents a valuable tool to improve the prognosis of HFpEF patients with different risk levels. The discovered transcriptional biomarkers of HFpEF provide new insight to the understanding the complex mechanism of HFpEF, leading to the development of novel targeted therapies for HFpEF. It is expected that integrating multi-omics and clinical data can further improve HFpEF outcome prediction, leading to the development of targeted, adaptive, and precision treatment of HFpEF patients with different risk levels.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: The data analyzed in this study require NIH approval through the dbGap website. Requests to access these datasets should be directed to http://dbgap.ncbi.nlm.nih.gov.

AUTHOR CONTRIBUTIONS

LZ and ZG performed the study and drafted the manuscript. BW, YW, ZL, RF, and HtY participated in the data processing and analysis. HY provided the clinical interpretation. HC and YC conceived of the idea and revised the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by grants from NSFC (71403156 and 81872717), the Applied Basic Research Program of Shanxi Province (201901D111204), the China Scholarship Council (201908140151), a fund from the Provincial Education Department of Hebei Province (ZD2018022), and a fund from Michigan State University.

ACKNOWLEDGMENTS

We are grateful to the two reviewers for their insightful comments that helped us to greatly improve our manuscript. The Framingham Heart Study was conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University (contract no. N01-HC-25195, HHSN268201500001I, and 75N92019D00031). This manuscript was not prepared in collaboration with the investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham

Heart Study, Boston University, or the NHLBI. We thank Lisa Kreiner, PhD, from Liwen Bianji, Edanz Editing China (www.liwenbianji.cn/ac), for editing the English text of a draft of this manuscript.

REFERENCES

Ai, C., and Kong, L. (2018). CGPS: a machine learning-based approach integrating multiple gene set analysis tools for better prioritization of biologically relevant pathways. *J. Genet. Genom.* 45, 489–504. doi: 10.1016/j.jgg.2018.08.002

Altamish, M., Samuel, V. P., Dahiya, R., Singh, Y., Deb, P. K., Bakshi, H. A., et al. (2020). Molecular signaling of G-protein-coupled receptor in chronic heart failure and associated complications. *Drug Dev. Res.* 81, 23–31. doi: 10.1002/ddr.21627

Angraal, S., Mortazavi, B. J., Gupta, A., Khera, R., Ahmad, T., Desai, N. R., et al. (2020). Machine learning prediction of mortality and hospitalization in heart failure with preserved ejection fraction. *JACC Heart Fail.* 8, 12–21. doi: 10.1016/j.jchf.2019.06.013

Antlanger, M., Aschauer, S., Kopecky, C., Hecking, M., Kovarik, J. J., Werzowa, J., et al. (2017). Heart failure with preserved and reduced ejection fraction in hemodialysis patients: prevalence, disease prediction and prognosis. *Kidney Blood Press. Res.* 42, 165–176. doi: 10.1159/000473868

Appari, M., Breitbart, A., Brandes, F., Szaroszyk, M., and Heineke, J. (2016). C1q-TNF-related protein-9 promotes cardiac hypertrophy and failure. *Circ. Res.* 120, 66–77. doi: 10.1161/CIRCRESAHA.116.309398

Auger-Messier, M., Accornero, F., Goonasekera, S. A., Bueno, O. F., Lorenz, J. N., Van Berlo, J. H., et al. (2013). Unrestrained p38 MAPK activation in Dusp1/4 double-null mice induces cardiomyopathy. *Circ. Res.* 112, 48–56. doi: 10.1161/CIRCRESAHA.112.272963

Austin, P. C., Tu, J. V., Ho, J. E., Levy, D., and Lee, D. S. (2013). Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. *J. Clin. Epidemiol.* 66, 398–407. doi: 10.1016/j.jclinepi.2012.11.008

Chawla, N., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16, 321–357. doi: 10.1613/jair.953

Cheerla, N., and Gevaert, O. (2017). MicroRNA based pan-Cancer diagnosis and treatment recommendation. *BMC Bioinform.* 18:32. doi: 10.1186/s12859-016-1421-y

Coulthard, M. G., Morgan, M., Woodruff, T. M., Arumugam, T. V., Taylor, S. M., Carpenter, T. C., et al. (2012). Eph/Ephrin signaling in injury and inflammation. *Am. J. Pathol.* 181, 1493–1503. doi: 10.1016/j.ajpath.2012.06.043

Dattilo, M., Neuman, I., Muñoz, M., Maloberti, P., and Cornejo Maciel, F. (2015). OxeR1 regulates angiotensin II and cAMP-stimulated steroid production in human H295R adrenocortical cells. *Mol. Cell. Endocrinol.* 408, 38–44. doi: 10.1016/j.mce.2015.01.040

Fan, J., and Lv, J. (2008). Rejoinder: sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Series B Stat. Methodol.* 70, 849–911. doi: 10.1111/j.1467-9868.2008.00674.x

Ferrero, P., Iacovoni, A., D'Elia, E., Vaduganathan, M., Gavazzi, A., and Senni, M. (2015). Prognostic scores in heart failure - critical appraisal and practical use. *Int. J. Cardiol.* 188, 1–9. doi: 10.1016/j.ijcard.2015.03.154

Frank, L. E., and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* 35, 109–135. doi: 10.1080/00401706.1993.10485033

Fransen, J., Popa-Diaconu, D., Hesselstrand, R., Carreira, P., Valentini, G., and Beretta, L. (2011). Clinical prediction of 5-year survival in systemic sclerosis: validation of a simple prognostic model in EUSTAR centres. *Ann. Rheum. Dis.* 70, 1788–1792. doi: 10.1136/ard.2010.144360

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., and Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* 16, 906–914. doi: 10.1093/bioinformatics/16.10.906

Garg, A., Virmani, D., Agrawal, S., Agarwal, C., Sharma, A., Stefanini, G., et al. (2017). Clinical application of biomarkers in heart failure with a preserved ejection fraction: a review. *Cardiology* 136, 192–203. doi: 10.1159/000450573

Hahn, R. T., Hoppstädter, J., Hirschfelder, K., Hachenthal, N., Diesel, B., Kessler, S. M., et al. (2014). Downregulation of the glucocorticoid-induced leucine zipper (GILZ) promotes vascular inflammation. *Atherosclerosis* 234, 391–400. doi: 10.1016/j.atherosclerosis.2014.03.028

Haring, R., and Wallaschofski, H. (2012). Diving through the "-omics": the case for deep Phenotyping and systems epidemiology. *OMICS* 16, 231–234. doi: 10.1089/omi.2011.0108

Kanda, T., Uematsu, M., Fujita, M., Iida, O., Masuda, M., Okamoto, S., et al. (2018). A novel predictor of clinical outcomes in patients with heart failure with preserved left-ventricular ejection fraction: a pilot study. *Heart Vessel.* 33, 1490–1495. doi: 10.1007/s00380-018-1211-8

Khan, J., Wei, J. S., and Greer, B. T. (2012). Prediction of clinical outcome using gene expression profiling and artificial neural networks for patients with neuroblastoma. *Cancer Res.* 64, 6883–6891. doi: 10.1158/0008-5472.CAN-04-0695

Komajda, M., Carson, P. E., Hetzel, S., McKelvie, R., McMurray, J., Ptaszynska, A., et al. (2011). Factors associated with outcome in heart failure with preserved ejection fraction: findings from the Irbesartan in heart failure with preserved ejection fraction study (I-PRESERVE). *Circ. Heart Fail.* 4, 27–35. doi: 10.1161/CIRCHEARTFAILURE.109.932996

Lin, Y. J., Chang, J. S., Liu, X., Lin, T. H., Huang, S. M., Liao, C. C., et al. (2013). Sorting nexin 24 genetic variation associates with coronary artery aneurysm severity in Kawasaki disease patients. *Cell Biosci.* 3:44. doi: 10.1186/2045-3701-3-44

Loh, J. C., Creaser, J., Rourke, D. A., Livingston, N., Harrison, T. K., Vandenbogaart, E., et al. (2013). Temporal trends in treatment and outcomes for advanced heart failure with reduced ejection fraction from 1993-2010: findings from a university referral center. *Circ. Heart Fail.* 6, 411–419. doi: 10.1161/CIRCHEARTFAILURE.112.000178

Lutz, S., Mura, R. A., Hippe, H. J., Tiefenbacher, C., and Niroomand, F. (2003). Plasma membrane-associated nucleoside diphosphate kinase (nm23) in the heart is regulated by beta-adrenergic signaling. *Br. J. Pharmacol.* 140:1019. doi: 10.1038/sj.bjp.0705527

Manolis, A. S., Manolis, A. A., Manolis, T. A., and Melita, H. (2019). Sudden death in heart failure with preserved ejection fraction and beyond: an elusive target. *Heart Fail. Rev.* 24, 847–866. doi: 10.1007/s10741-019-09804-2

McMurray, J. J., Adamopoulos, S., Anker, S. D., Auricchio, A., Böhm, M., Dickstein, K., et al. (2018). ESC guidelines for the diagnosis and treatment of acute and chronic heart failure 2012: the task force for the diagnosis and treatment of acute and chronic heart failure 2012 of the European Society of Cardiology. Developed in collaboration with the heart. *Eur. Heart J.* 33, 1787–1847. doi: 10.1093/eurheartj/ehs104

Monika, R., Arantxa, B. A., Vanessa, V. E., Marc, V. B., and Blanche, S. (2018). Pathophysiological understanding of HFpEF: microRNAs as part of the puzzle. *Cardiovasc. Res.* 114, 782–793. doi: 10.1093/cvr/cvy049

Oktay, A. A., Rich, J. D., and Shah, S. J. (2013). The emerging epidemic of heart failure with preserved ejection fraction. *Curr. Heart Fail. Rep.* 10, 401–410. doi: 10.1007/s11897-013-0155-7

Oppenheimer, G. M. (2005). Becoming the Framingham study 1947-1950. *Am. J. Public Health* 95, 602–610. doi: 10.2105/AJPH.2003.026419

Penney, K. L., Sinnott, J. A., Fall, K., Pawitan, Y., Hoshida, Y., Kraft, P., et al. (2011). mRNA expression signature of Gleason grade predicts lethal prostate cancer. *J. Clin. Oncol.* 29, 2391–2396. doi: 10.1200/JCO.2010.32.6421

Perry, M. D., and Vandenberg, J. I. (2016). TECRL: connecting sequence to consequence for a new sudden cardiac death gene. *EMBO Mol. Med.* 8, 1364–1365. doi: 10.15252/emmm.201606967

Pocock, S. J., Ariti, C. A., McMurray, J. J., Maggioni, A., Køber, L., Squire, I. B., et al. (2013). Predicting survival in heart failure: a risk score based on 39 372 patients from 30 studies. *Eur. Heart J.* 34, 1404–1413. doi: 10.1093/eurheartj/ehs337

Rich, J. D., Burns, J., Freed, B. H., Maurer, M. S., Burkhoff, D., and Shah, S. J. (2018). Meta-analysis Global Group in Chronic (MAGGIC) heart failure risk score: validation of a simple tool for the prediction of morbidity and mortality in heart failure with preserved ejection fraction. *J. Am. Heart Assoc.* 7:e009594. doi: 10.1161/JAHA.118.009594

Rodriguez-Calvo, R., Girona, J., Alegret, J. M., Bosquet, A., Ibarretxe, D., and Masana, L. (2017). Role of the fatty acid binding protein 4 in heart failure and cardiovascular disease. *J. Endocrinol.* 233, R173–R184. doi: 10.1530/JOE-17-0031

Ronnett, G. V., and Moon, C. (2002). G proteins and olfactory signal transduction. *Annu. Rev. Physiol.* 64, 189–222. doi: 10.1146/annurev.physiol.64.082701.102219

Rosipal, R., and Trejo, L. J. (2002). Kernel partial least squares regression in reproducing kernel Hilbert space. *J. Mach. Learn. Res.* 2, 97–123. doi: 10.1162/15324430260185556

Shah, S. J., Katz, D. H., and Deo, R. C. (2014). Phenotypic spectrum of heart failure with preserved ejection fraction. *Heart Fail. Clin.* 10, 407–418. doi: 10.1016/j.hfc.2014.04.008

Shah, S. J., Katz, D. H., Selvaraj, S., Burke, M. A., Yancy, C. W., Gheorghiade, M., et al. (2015). Phenomapping for novel classification of heart failure with preserved ejection fraction. *Circulation* 131, 269–279. doi: 10.1161/CIRCULATIONAHA.114.010637

Shah, K. S., Xu, H., Matsouaka, R. A., Bhatt, D. L., Heidenreich, P. A., Hernandez, A. F., et al. (2017). Heart failure with preserved, borderline, and reduced ejection fraction: 5-year outcomes. *J. Am. Coll. Cardiol.* 70, 2476–2486. doi: 10.1016/j.jacc.2017.08.074

Shen, L., Jhund, P. S., Anand, I. S., Carson, P. E., Desai, A. S., Granger, C. B., et al. (2020). Developing and validating models to predict sudden death and pump failure death in patients with heart failure and preserved ejection fraction. *Clin. Res. Cardiol.* doi: 10.1007/s00392-020-01786-8 [Epub ahead of print]

Shendre, A., Wiener, H., Irvin, M. R., Zhi, D., Limdi, N. A., Overton, E. T., et al. (2017). Admixture mapping of subclinical atherosclerosis and subsequent clinical events among African Americans in 2 large cohort studies. *Circ. Cardiovasc. Genet.* 10:e001569. doi: 10.1161/CIRCGENETICS.116.001569

Thorvaldsen, T., Claggett, B. L., Shah, A., Cheng, S., Agarwal, S. K., Wruck, L. M., et al. (2017). Predicting risk in patients hospitalized for acute decompensated heart failure and preserved ejection fraction. *Circ. Heart Fail.* 10:e003992. doi: 10.1161/CIRCHEARTFAILURE.117.003992

Tibshirani, R. J. (1996). Regression shrinkage and selection via the LASSO. *J. R. Stat. Soc. Ser. B Methodol.* 73, 273–282.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U. S. A.* 98, 5116–5121. doi: 10.1073/pnas.091062498

Vargas, J. D., and Lima, J. A. (2013). Coronary artery disease: a gene-expression score to predict obstructive CAD. *Nat. Rev. Cardiol.* 10, 243–244. doi: 10.1038/nrcardio.2013.50

Wang, J., Gareri, C., and Rockman, H. A. (2018). G-protein–coupled receptors in heart disease. *Circ. Res.* 123, 716–735. doi: 10.1161/CIRCRESAHA.118.311403

Wang, Y., Wilson, C., Cartwright, E. J., and Lei, M. (2017). Plasma membrane Ca^{2+} -ATPase 1 is required for maintaining atrial Ca^{2+} homeostasis and electrophysiological stability in the mouse. *J. Physiol.* 595, 7383–7398. doi: 10.1113/JP274110

Wang, Q., Xu, M., Sun, Y., Chen, J., and Yang, W. (2019). Gene expression profiling for diagnosis of triple-negative breast cancer: a Multicenter, retrospective cohort study. *Front. Oncol.* 9:1576. doi: 10.3389/fonc.2019.01576

Yang, H., Cao, H., He, T., Wang, T., and Cui, Y. (2020). Multilevel heterogeneous omics data integration with kernel fusion. *Brief. Bioinform.* 21, 156–170. doi: 10.1093/bib/bby115

Yao, C., Chen, B. H., Joehanes, R., Otlu, B., Zhang, X., Liu, C., et al. (2015). Integromic analysis of genetic variation and gene expression identifies networks for cardiovascular disease phenotypes. *Circulation* 131, 536–549. doi: 10.1161/CIRCULATIONAHA.114.010696

Yariswamy, M., Yoshida, T., Valente, A. J., Kandikattu, H. K., Sakamuri, S. S. V. P., Siddesha, J. M., et al. (2016). Cardiac-restricted overexpression of TRAF3 interacting protein 2 (TRAF3IP2) results in spontaneous development of myocardial hypertrophy, fibrosis, and dysfunction. *J. Biol. Chem.* 291:19425. doi: 10.1074/jbc.M116.724138

Zhang, Q., Hu, H., Chen, S. Y., Liu, C. J., Hu, F. F., Yu, J., et al. (2019). Transcriptome and regulatory network analyses of CD19-CAR-T immunotherapy for B-ALL. *Geno. Prot. Bioinfo.* 17, 190–200. doi: 10.1016/j.gpb.2018.12.008

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

ORIGINAL RESEARCH
published: 15 April 2021
doi: 10.3389/fgene.2021.643461

Immunohistochemical Expression of Five Protein Combinations Revealed as Prognostic Markers in Asian Oral Cancer

Hui-Ching Wang[1,2], *Chien-Jung Chiang*[3], *Ta-Chih Liu*[4]*, *Chun-Chieh Wu*[5], *Yi-Ting Chen*[5], *Jan-Gowth Chang*[6,7,8,9,10] *and Grace S. Shieh*[3,11,12,13]*

[1] *Graduate Institute of Clinical Medicine, College of Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan,* [2] *Division of Hematology and Oncology, Department of Internal Medicine, Kaohsiung Medical University Hospital, Kaohsiung Medical University, Kaohsiung, Taiwan,* [3] *Institute of Statistical Science, Academia Sinica, Taipei, Taiwan,* [4] *Department of Hematology-Oncology, Chang Bing Show Chwan Memorial Hospital, Changhua, Taiwan,* [5] *Department of Pathology, Kaohsiung Medical University Hospital, Kaohsiung Medical University, Kaohsiung, Taiwan,* [6] *Epigenome Research Center, China Medical University Hospital, Taichung, Taiwan,* [7] *Department of Laboratory Medicine, China Medical University Hospital, Taichung, Taiwan,* [8] *Center for Precision Medicine, China Medical University Hospital, Taichung, Taiwan,* [9] *School of Medicine, China Medical University, Taichung, Taiwan,* [10] *Department of Bioinformatics and Medical Engineering, Asia University, Taichung, Taiwan,* [11] *Bioinformatics Program, Taiwan International Graduate Program, Academia Sinica, Taipei, Taiwan,* [12] *Genome and Systems Biology Degree Program, Academia Sinica and National Taiwan University, Taipei, Taiwan,* [13] *Data Science Degree Program, Academia Sinica and National Taiwan University, Taipei, Taiwan*

Oral squamous cell carcinoma (OSCC) has a high mortality rate (∼50%), and the 5-year overall survival rate is not optimal. Cyto- and histopathological examination of cancer tissues is the main strategy for diagnosis and treatment. In the present study, we aimed to uncover *immunohistochemical* (IHC) markers for prognosis in Asian OSCC. From the collected 742 synthetic lethal gene pairs (of various cancer types), we first filtered genes relevant to OSCC, performed 29 IHC stains at different cellular portions and combined these IHC stains into 398 distinct pairs. Next, we identified novel *IHC* prognostic markers in OSCC among Taiwanese population, from the single and paired IHC staining by univariate Cox regression analysis. Increased nuclear expression of RB1 [RB1(N)↑], CDH3(C)↑-STK17A(N)↑ and FLNA(C)↑-KRAS(C)↑were associated with survival, but not independent of tumor stage, where C and N denote cytoplasm and nucleus, respectively. Furthermore, multivariate Cox regression analyses revealed that CSNK1E(C)↓-SHC1(N)↓ ($P = 5.9 \times 10^{-5}$; recommended for clinical use), BRCA1(N)↓-SHC1(N)↓ ($P = 0.030$), CSNK1E(C)↓-RB1(N)↑ ($P = 0.045$), [CSNK1E(C)-SHC1(N), FLNA(C)-KRAS(C)] ($P = 0.000$, rounded to three decimal places) and [BRCA1(N)-SHC1(N), FLNA(C)-KRAS(C)] ($P = 0.020$) were significant factors of poor prognosis, independent of lymph node metastasis, stage and alcohol consumption. An external dataset from The Cancer Genome Atlas HNSCC cohort confirmed that *CDH3*↑-*STK17A*↑ was a significant predictor of poor survival. Our approach identified prognostic markers with components involved in different pathways and revealed IHC marker pairs while neither single IHC was a marker, thus it improved the current state-of-the-art for identification of IHC markers.

Keywords: biomarker, cox regression, immunohistochemistry, oral cancer, overall survival, prognosis, gene expression data

OPEN ACCESS

Edited by:
Chao Xu,
University of Oklahoma Health Sciences Center, United States

Reviewed by:
Shankargouda Patil,
Jazan University, Saudi Arabia
Marco Mascitti,
Marche Polytechnic University, Italy
Vui King Vincent-Chong,
University at Buffalo, United States

****Correspondence:***
Ta-Chih Liu
touchyou3636@gmail.com
Grace S. Shieh
gshieh@stat.sinica.edu.tw

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *18 December 2020*
Accepted: *01 March 2021*
Published: *15 April 2021*

Citation:
Wang H-C, Chiang C-J, Liu T-C, Wu C-C, Chen Y-T, Chang J-G and Shieh GS (2021) Immunohistochemical Expression of Five Protein Combinations Revealed as Prognostic Markers in Asian Oral Cancer. Front. Genet. 12:643461. doi: 10.3389/fgene.2021.643461

INTRODUCTION

Head and neck squamous cell carcinoma (HNSCC) is the sixth most common cancer globally (Bray et al., 2018). Every year, more than 700,000 new cases of HNSCC are diagnosed and 350,000 related deaths are reported worldwide. Oral squamous cell carcinoma (OSCC) is the most common cancer of the head and neck region and has a high mortality rate. However, little improvement has been made in the five-year overall survival rate over the years (Bray et al., 2018). Identifying reliable prognostic factors remains challenging. OSCC is believed to originate from the multistep accumulation of heterogeneous genetic changes in squamous cells. These changes progressively enable transformed cells to proliferate and invade (Oliveira and Ribeiro-Silva, 2011). These accumulated changes may explain why tumors at the same clinical stage and localization often show significant differences in clinical outcome.

The main causes of OSCC in Taiwan and some South Asian countries (Belcher et al., 2014) are the consumption of alcohol, tobacco, and betel nut. This contrasts with human papillomavirus (HPV)-positive oropharyngeal SCC which is associated with HPV infection, with higher proportions in western populations than Asian populations (Gillison et al., 2000).

Unlike other malignancies, the relationships between mutations of genes and clinical morphological characteristics such as tumor grade in OSCC are obscure, which has impeded the development of personalized medicine. Cytopathological and histopathological examination of cancer tissues remains the main diagnostic and treatment strategy for OSCC. Although immunohistochemical (IHC) staining may be limited by small volumes taken from samples, varying expression with selected antibodies, and partial reliance on subjective perception, IHC staining provides morphological information about protein expression, and it is simple and cost-effective. Moreover, the procedures and guidelines (Wolff et al., 2007; Hammond et al., 2010; Dowsett et al., 2011) for IHC staining are well established, and widely used in clinics.

The primary aim is to identify a panel of IHC prognostic markers for Asian OSCC, to enable the selection of patients best suited for intensive adjuvant therapy in clinics. Most of previous results on IHC prognostic markers in OSCC were mainly based on one protein, few on two proteins or on one pathway and are reviewed briefly as follows. IHC of cyclin D1, MDM2, and γ-catenin were shown to be potential prognostic markers in a study of 55 patients with buccal SCC who regularly chewed betel nut (Peng et al., 2011). In 2005, IHC of cyclin D1 and Rb overexpression combined with p16 underexpression (denoted by cyclin D1↑-Rb↑p16↓) (Jayasurya et al., 2005) and Rb↓–p53↑ (Soni et al., 2005) were shown to be associated with poor prognosis in a cohort of 348 and a cohort of 98 Indian patients with OSCC, respectively. Moreover, simultaneous coexpression of p53, cyclin D1, and EGFR was a significant prognostic factor in a cohort of 140 Japanese patients with oral cancer (Shiraki et al., 2005). P-cadherin was reported to be marginally significantly associated with poor survival in a small cohort (Muzio et al., 2005). About a decade later, CK1ϵ nonexpression (Lin et al., 2014) and expression of BRCA1 and γH2AX (Oliveira-Costa et al., 2014) were shown to be associated with poor overall and disease-specific survival, respectively.

We started with a list of 742 synthetic lethal (SL) gene pairs collected from the literature, which consisted of several oncogenes, tumor-suppressor genes, genome stability and other cancer genes with important functions. Two genes are termed SL genes if a single mutation of either is not lethal, but their simultaneous mutation leads to cell death (Chang et al., 2016). The SL interactions of these collected pairs in various cancer types are validated either with human cancer cell lines (Bryant et al., 2005; Farmer et al., 2005) or by genome-wide RNA interference (RNAi) knockdown (Barbie et al., 2009; Luo et al., 2009). The list of SL gene pairs can be accessed at[1]. SL pairs are shown to be correlated to survival of cancer cells (Kaelin, 2005). In general, the more cancer cells killed, the better cancer patients' survival. Thus, we speculated that SL pairs are relevant to prognosis assessment. We hypothesized that IHC (protein) expression is concordant to its gene expression, and we used gene expression data of Asian OSCC to select an initial panel of SL pairs which are relevant to OSCC, from the collected SL pairs (of all types of cancer). Next, we adopted the rule of frequently co-expressed gene pairs along with prior knowledge of OSCC to select ∼20 genes for IHC staining. IHC staining is conducted because protein is more stable than mRNA, ultimately functions in cells, and IHC is usable in the clinic. We also combined single IHC into 398 distinct IHC pairs. To identify prognostic markers, we applied Cox regression analysis to each single IHC (each combined IHC pair) and the overall survival of patients with OSCC. Previously, we applied this approach to colorectal cancer (Chang et al., 2016) and lung adenocarcinoma (Liu et al., 2004), and both studies successfully uncovered IHC prognostic markers independent of tumor stage, in addition to revealing novel IHC marker pairs where neither single IHC was a marker. Our approach starts with the collected SL pairs, which allows components participating in different pathways to be identified as prognostic marker pairs. This improves the current state of the art for IHC marker discovery, which hitherto has mainly relied on one protein, few on two proteins or one pathway (Oliveira-Costa et al., 2014). After the prognostic markers were revealed, we validated them using OSCC data with HPV(−) from The Cancer Genome Atlas (TCGA) HNSCC cohort. A schematic graph of our method is presented in **Figure 1**.

MATERIALS AND METHODS

Study Population

A total of 163 cases of oral cavity cancers were identified in the Kaohsiung Medical University Hospital. Although this sample size was moderate, it was less than only four of the 20 and more previous studies. Furthermore, we conducted a large scale of IHC study and the sample size was sufficiently large for multivariate Cox regression analysis. The inclusion criteria for this study were as follows: (1) age at diagnosis of 20 years or older; (2) tumor histology of squamous cell carcinoma with

[1] http://www.stat.sinica.edu.tw/~gshieh/OC/SL_pairs.html

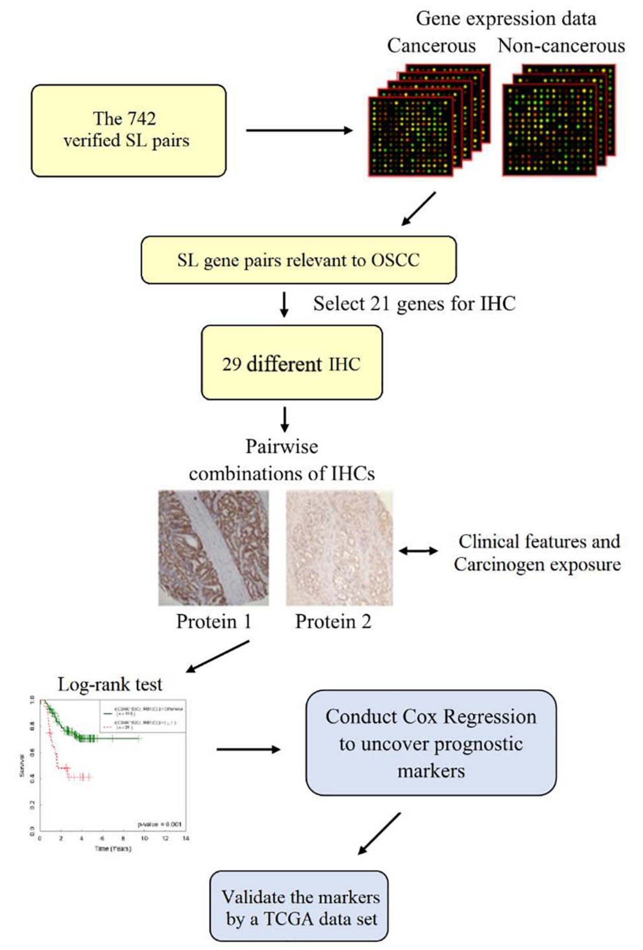

Gene expression data
Cancerous
Non-cancerous
The 742 verified SL pairs
SL gene pairs relevant to OSCC
Select 21 genes for IHC
29 different IHC
Pairwise combinations of IHCs
Protein 1
Protein 2
Clinical features and Carcinogen exposure
Log-rank test
Conduct Cox Regression to uncover prognostic markers
Validate the markers by a TCGA data set

FIGURE 1 | Schematic graph of the study approach. Microarray gene expression of 57 oral cancerous and 22 noncancerous tissues selected OSCC-relevant gene pairs from 742 verified synthetic lethal pairs. Twenty-one genes were marked for immunohistochemistry staining. Pairwise combinations of the 29 IHCs followed by a log-rank test and Cox regression models revealed single/paired and combined prognostic markers.

grade 1 to grade 3; (3) ICD-9 site code specific for the oral cavity; (4) patients underwent surgical interventions,; and (5) disease was diagnosed between 2012 and 2014. The exclusion criteria were: (1) patients who underwent biopsy without surgery; (2) patients with secondary malignancy; (3) tumor histology of carcinoma *in situ*; and (4) SCC from nasopharynx, oropharynx, hypopharynx, and larynx.

Statistical Analyses, Tissue Arrays and IHC Staining

In the following, all statistical tests were two-sided except where otherwise specified, and all analyses were conducted in R software (R Core Team, 2019).

Preprocessing of Gene Expression Profiles for Oral Cavity Cancerous Versus Non-cancerous Tissues

Gene expression datasets were selected based on the following parameters: cancerous and noncancerous tissues, no treatments, no metastasis, and Affymetrix chips (up to November 2010). The OSCC gene profiles conforming to the aforementioned criteria were downloaded from GEO. Mutated genes associated with oncogenesis may differ among various ethnic groups (Ding et al., 2008). Therefore, we collected gene expression data from patients of Han Chinese origin [tissues from patients in Taiwan, GSE 25099 (Peng et al., 2011)], which was the same ethnicity as that of IHC and clinicopathological data used here. Gene expression profiles of the 57 OSCC and 22 noncancerous tissues in the dataset were quantile-normalized using “expresso” in R, then for a given gene the log ratio of its expression in each cancer tissue versus that of the averaged non-cancerous tissues was computed.

Inference of the Initial Panel of Relevant SL Gene Pairs (Table 2) Using Microarray Gene Expression Data

For each SL gene pair, the fractions of (up, up), (up, down), (down, up), and (down, down) patterns were computed, where the cutoff value for up and downregulation was 1.5-fold. The pattern fractions were computed using the log ratios of the microarray gene expression data for the 57 patients with OSCC (GSE 25099).

Permutation Test and False-Positive Rates of the Fractions of Paired Gene Expression

To evaluate the statistical significance (P value) of the fractions of (up, up) and (down, up) patterns of each gene pair in **Table 2**, for each fraction we conducted a permutation test to generate its nonparametric distribution. The total rearrangements of the labels of (57) cancer and (22) noncancerous tissues was equal to $\binom{79}{22}$, from which we randomly chose 10,000 rearrangements. For each rearrangement, we computed the fraction of a pattern to form its distribution, from which we assessed the P value of an observed fraction. Moreover, we applied the q-value (Storey and Tibshirani, 2003) (“*q* value” in R) to estimate the false discovery rate (FDR) of the significance of the gene pairs in **Table 2**.

Selection of Genes From the Initial Panel for IHC Staining

We first selected genes whose fractions of the (up, up) and (down, up) patterns were ≥15%, except ≥25% (more stringent) for the (down, up) pattern of *KRAS* SL pair, because the mutation rate of *KRAS* was only ∼2% in OC and there were 200 and more *KRAS* pairs. Next, we applied prior knowledge to (i) select *CDH3* (the top-3) from the top three partners of *EGFR* and the top-1 [*STK17A*, relevant to OSCC (Pickering et al., 2013)] and top-3 (*CDK6*, a tumor suppressor gene) partners of *KRAS* from the *KRAS* SL pairs satisfying the above fraction cutoffs, and to (ii) include genes whose fractions were on the borderline of 15%; this included *FEN1-RAD54B* [involved in nonhomologous DNA end joining repair pathway (Storey and Tibshirani, 2003); 14%], *RB1* [relevant to OSCC (Liu et al., 2004; Presson et al., 2011); 12%] and *MSH2-POLB* (Kang et al., 2009; Tiong et al., 2014; Chang et al., 2016).

Tissue Microarray Preparation

Clinicopathological features of 163 OSCC patients were collected (**Table 1**), and their representative cancer specimens were randomly selected from H&E-stained sections and confirmed by pathologists (Chun-Chieh Wu and Yi-Ting Chen). Three cancerous and one noncancerous tissue cores (diameter 2 mm) were longitudinally cut from each paraffin block. The tissue cores were mounted with fine steel needles in new paraffin blocks to produce tissue microarrays. This study was approved by the Institutional Review Board and Ethics Committee of Kaohsiung Medical University Hospital and the Institutional Review Board of Academia Sinica [Nos. KMUHIRB-E(I)-20170034 and AS-IRB-BM-16075]. The data was analyzed anonymously, and therefore no additional informed consent was required. All methods were performed in accordance with the approved guidelines and regulations and the waiver for the informed consent had been obtained from the approving committee.

Immunohistochemistry Staining

Patients cancer samples were cut into 4-μm-thick sections and deparaffinized in xylene as previously described (Chang et al., 2016). Endogenous peroxidase activity was quenched with 3% (v/v) H_2O_2. The sections were boiled in 10 mM citrate buffer for 20 min to revive the antigens. The tissues were incubated with 21 primary antibodies at room temperature for 30 min then rinsed three times with phosphate-buffered saline (PBS) (**Supplementary Table 1**) according to the manufacturer's protocol. The tissues were then incubated at 25°C for 30 min with secondary antibodies and a horseradish peroxidase/Fab polymer conjugate [EnVision detection systems peroxidase/DAB, rabbit/mouse (K5007 HRP; DaKo)] then rinsed three times with PBS. Finally, chromogen was developed using 3,3′-diaminobenzidine tetrahydrochloride as the substrate, and counterstained with hematoxylin and viewed under a microscope. Staining intensity in the cancer tissue was independently examined by two pathologists (Chun-Chieh Wu and Yi-Ting Chen).

The scoring criteria used here were the same as those of previous studies (Su et al., 2004; Tiong et al., 2014; Chang et al., 2016) (**Supplementary Table 2**). Stain intensity is graded as negative (0), indeterminate (±), weakly positive (1+), moderately positive (2+), or strongly positive (3+). The criterion is exactly based on the strongest intensity followed by the % expression of the detected protein. Negative (0) indicates no expression of the detected protein, indeterminate means that the staining is weak and its percentage cannot be accurately counted, weakly positive indicates <5% expression of the detected protein, moderately positive is focal expression in 5–20% of the cancer cells, and strongly positive indicates diffuse expression in >20% of the cancer cells. The mean staining intensity of three cancerous tissues was compared with that of noncancerous oral mucosa and was categorized as either over- or underexpressing, determining the criteria for IHC analyses. The cutoffs for the 29 IHC stains are listed in **Table 3**.

TABLE 1 | Clinicopathological characteristics of the OC patients in the study population.

Study population		
	KMU (*N* = 163)	
Characteristic	***N***	**%**
Age at diagnosis, year		
≦55	84	51.5
>55	71	43.6
NA[a]	8	4.9
Grade		
Low	67	41.1
Intermediate	85	52.1
High	2	1.2
NA	9	5.5
Stage		
I	66	40.5
II	26	16.0
III	19	11.7
IV	41	25.2
NA	11	6.7
Morphology		
Squamous	163	100.0
T[b]		
T1	74	45.4
T2	42	25.8
T3	11	6.7
T4	28	17.2
NA	8	4.9
N[b]		
N0	115	70.6
N1	24	14.7
N2	16	9.8
NA	8	4.9

[a]NA denotes missing data; [b]T and N denote tumor size and lymph node status of "TNM" (AJCC version 7), respectively.

Log-Rank Test

For each individual and paired IHC staining, a log-rank test of the "high-" and "low-" risk patients was conducted. The high- and low-risk groups consisted of patients classified according to the IHC amounts shown in **Table 3**. If the log-rank test was significant ($P < 0.05$), which indicated the survival curves of the two groups significantly different, then a Kaplan-Meier survival curve was plotted by the R software.

Univariate Cox Proportional Hazard (PH) Regression Models

In the univariate Cox regression models, the associations between the 29 individual IHC or 398 combined IHC staining pairs and the 10-year overall survival of the OSCC patients were analyzed in the study cohort. The associations between clinical factors such as age (>55 vs. ≤55 years), sex (male vs. female), tumor grade

TABLE 2 | The initial panel of SL gene pairs relevant to oral cancer.

Relevant SL gene pair		Fractions of SL gene pairs computed from 79 Asian OSCC versus non-cancerous tissues that were expressed 1.5-fold or higher					
Gene1	Gene2	(up, up)	(up, down)	(down, up)	(down,down)	Permutation	*q*-value
Gene1	Gene2	pattern	pattern	pattern	Pattern	*p*-value	*q*-value
EGFR	DUSP6	0.26[a]	0.00	0.12	0.00	0.0001	0.0004
EGFR	PLSCR1	0.26	0.00	0.12	0.00	0.0001	0.0004
EGFR	CDH3	0.26	0.00	0.05	0.00	0.0001	0.0004
BRCA1	PARP1	0.26	0.00	0.00	0.00	0.0001	0.0004
BRCA2	PARP1	0.26	0.00	0.00	0.00	0.0001	0.0004
EGFR	FLNA	0.18	0.00	0.04	0.00	0.0001	0.0004
EGFR	SHC1	0.18	0.00	0.02	0.00	0.0001	0.0004
FEN1	RAD54B	0.14	0.00	0.00	0.00	0.0001	0.0004
EGFR	SL_{EGFR}[b]	0.11~0.12	0.00	0.00~0.12	0.00	0.0001~0.3250	0.0004~0.4220
PIMI	PLK1	0.00	0.00	0.68	0.02	0.0001[c]	0.0004
TP53	MET	0.00	0.00	0.63	0.00	0.0001	0.0004
TP53	PLK1	0.00	0.00	0.54	0.02	0.0001	0.0004
TP53	CDKN2A	0.00	0.00	0.39	0.05	0.0001	0.0004
TP53	BRCA1	0.00	0.00	0.37	0.02	0.0001	0.0004
KRAS	SL_{KRAS}[d]	0.00	0.00	0.18~0.28	0.00	0.0001~0.0011	0.0004~0.0185
TP53	CSNK1E	0.00	0.00	0.18	0.00	0.0001	0.0004
TP53	PARP1	0.00	0.00	0.18	0.00	0.0001	0.0004
TP53	RB1	0.00	0.00	0.12	0.00	0.0024	0.0063

[a] *The four fractions were computed from gene pairs that were 1.5-fold differentially expressed, thus they might not sum up to 100%.*
[b] *Four verified EGFR SL pairs were identified in the (up, up) pattern.*
[c] *The p-value shows the significance of the (down, up) pattern.*
[d] *26 verified KRAS SL pairs were identified in the (down, up) pattern.*
The p-value for the highest fraction four patterns was computed by permutation test with 10,000 repeats, and the false discovery rate was estimated by q-value.
Fractions of the four differentially expressed patterns based on the 1.5-fold threshold and filtered from 742 synthetic lethal gene pairs.

(medium and high vs. low), lymph node metastasis (yes vs. no), stage (III, VI vs. I, II), and habits alcohol use (yes/no), betel nut chewing (yes/no), and cigarette smoking (yes/no) with 10-year Taiwanese OSCC overall survival were also assessed.

Multivariate Cox PH Regression Model

When fitting the multivariate Cox regression models, the clinical factor stage significantly associated with overall survival in the univariate Cox regression models was adjusted, because the stage had stronger significance than that of the grade. Likelihood ratio test (LR_{χ^2}) and the statistical significance values generated (*P* values) were used to compare model fit between the uncovered prognostic IHC markers.

Determination of the Cutoff for Differential Expression of the TCGA Data

We first used 1.5-fold as the threshold for differential TCGA OSCC gene expression, but there were too few patients (less than 5) (Vittinghoff and McCulloch, 2007) in the poor/good overall survival subsets to perform univariate Cox regression for most of the six prognostic markers (**Table 4A**). Thus, the cutoff was relaxed to 1.4-fold, and there were ensure adequate numbers of patients in the poor/good overall survival subsets of two pairs *CDH3-STK17A* and *FLNA-KRAS*, respectively, for the univariate Cox regression analysis.

RESULTS

Description of Study Population

As shown in **Table 1**, about half of the patients in our study cohort were <55 years old at the time of diagnosis. The histologic grades were defined as low grade: well differentiated, intermediate grade: moderately differentiated, and high grade: poorly differentiated. Most of the cancers (98%) were intermediate- or low histological grade, only 1.3% were high grade. About 60% of the patients were stage I and II, and 26.5% were stage IV (most of them were stage IVA). All cancers were squamous cell carcinoma. According to the stratification of 7th version of the American Joint Committee on Cancer (AJCC), 71.2% of the tumor sizes belonged to T1 and T2, and 17.2% belonged to T4. Most of the lymph node statuses were N0 and N1 (85.3%).

Initial Panel of Relevant SL Gene Pairs for OSCC

In general, tumor cells show aberrant expression of oncogenes and tumor suppressor genes. Validated SL pairs comprised oncogenes and tumor suppressor- and genome stability genes. Therefore, we first selected gene pairs relevant to OSCC from the 742 SL pairs, using the microarray gene expression data of 57 Asian OSCC and 22 non-cancerous tissues {GSE 25099 from the gene expression omnibus database [GEO (Srivastava and

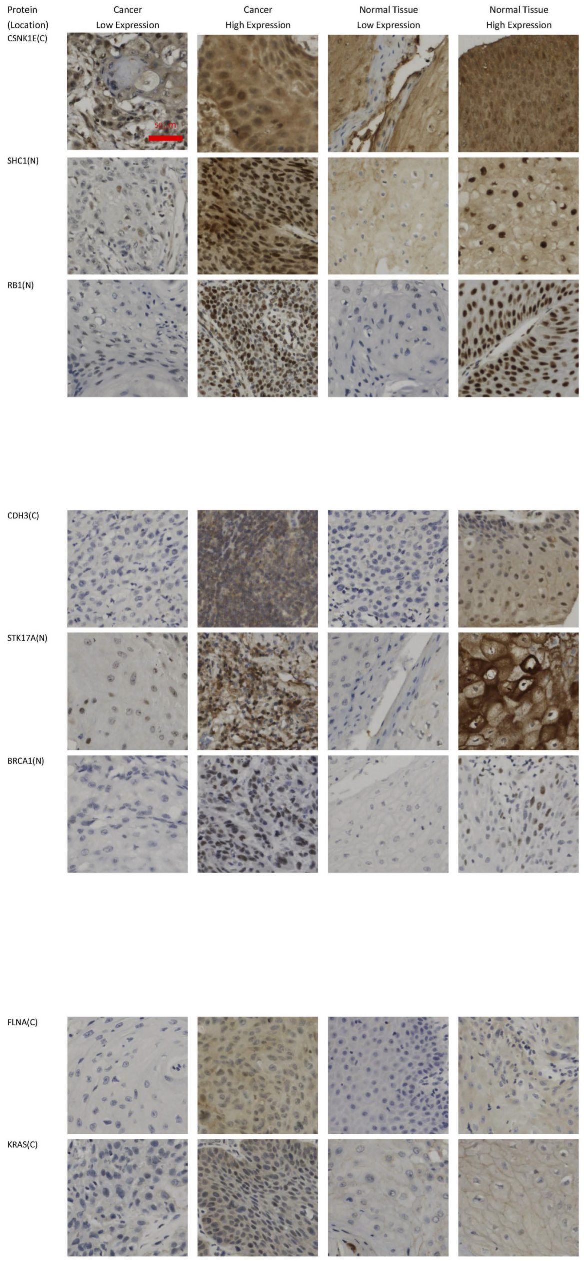


FIGURE 2 | Representative IHC images. Over- and underexpression of IHCs involved in the revealed markers [CSNK1E(C), SHC1(N), RB1(N), CDH3(C), STK17A(N), BRCA(N), FLNA(C), and KRAS(C)] are shown for cancer and normal tissues from OSCC patients (original magnification: × 400).

Raghavan, 2015)]} (Peng et al., 2011). The selected SL gene pairs were further sorted by the fractions of the (up, up), (down, up), (up, down), and (down, down) patterns (**Table 2**), where up and down denoted upregulation and downregulation with the cutoff 1.5-fold; this less stringent cutoff was set to include important OSCC onco- and tumor suppressor genes not expressed at twofold level, e.g., *TP53*, *EFGR,* and *CDKN2A*, but that were frequently mutated in Asian OSCC (Liu et al., 2004; Presson et al., 2011). Overexpression of tumor suppressor- and genome stability gene pairs associated with DNA repair such as *BRCA1* and *FEN1* was unexpectedly noted (**Table 2**). However, this finding was consistent with the dramatic increase in genomic instability and DNA replication caused by mutant oncogenes such as *MYC*.

Twenty-One Genes Were Selected for IHC Staining

We selected 21 genes from **Table 2** to conduct IHC staining, and some of them were stained at two cellular portions. Most of the genes were selected according to relatively high fractions of the (up, up) and (down, up) patterns (≥15%) in **Table 2**. For an extended list of the sorted (up, up) and (down, up) gene pairs, please see[2]. Next, we applied prior knowledge to (i) select *CDH3* from the *EGFR* SL pairs and *STK17A* and *CDK6* from the *KRAS* SL pairs, which satisfied the above fraction cutoffs, and to (ii) include genes whose fractions were on the borderline of 15%; this included *FEN1-RAD54B* (Srivastava and Raghavan, 2015) (14%), *RB1* (Liu et al., 2004; Presson et al., 2011) (12%) and *MSH2-POLB* (Kang et al., 2009; Tiong et al., 2014; Chang et al., 2016). Please see the section "Materials and Methods" for details of the selection method.

Eight out of these 21 genes were stained at two cellular portions, such as *CDH3* and *EGFR*, the remaining 13 genes were stained at one cellular portion. **Table 3** lists these 29 different IHC stains, the cutoffs for over- and underexpression of IHC staining and the corresponding fractions of OSCC patients satisfying the cutoffs. See section "Materials and Methods" for the basis determining the cutoff values. Some representative IHC figures are shown in **Figure 2**, including CSNK1E(C), SHC1(N), RB1(N), CDH3(C), STK17A(N), BRCA(N), FLNA(C), and KRAS(C). The IHC figures of all proteins are in **Supplementary Figure 1**.

We next explored if the results of IHC stains are suitable for use as prognostic markers. For each of the 29 IHC results in **Table 3** and all of the combined IHC pairs, we applied log-rank tests to the 153 Taiwanese patients with OSCC for whom overall survival was recorded. We first observed that the patients with overexpressed RB in nucleus (denoted as RB1(N)↑) had significantly poorer overall survival than patients with underexpressed RB1 (P = 0.027, **Figure 3A**). Additionally, underexpressed FLNA in cytoplasm [FLNA(C)↓] was also associated with poor clinical outcomes (P = 0.047, **Figure 3B**).

RB1↑ AND FLNA(C)↓ WERE ASSOCIATED WITH POOR OVERALL SURVIVAL

IHC of Eight Protein Pairs Were Associated With Overall Survival

Furthermore, we combined the 29 IHC stains into all the possible distinct IHC pairs (398 in total), which allowed novel paired IHC markers to be uncovered, excluding those of the same protein stained at different cellular

[2] https://www.stat.sinica.edu.tw/gshieh/OC/UU-DU_list

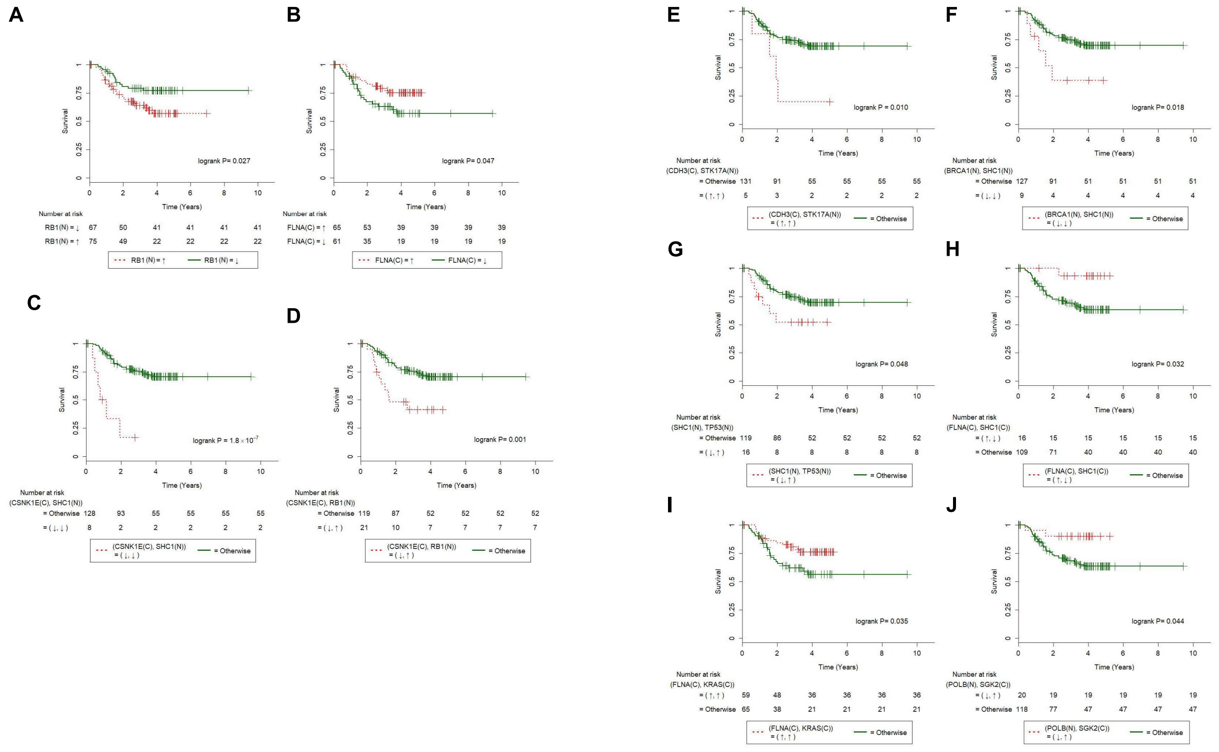


FIGURE 3 | Immunohistochemistry of individual and paired proteins correlated with overall survival of 153 Taiwanese oral squamous cell carcinoma patients. Kaplan-Meier survival curves were significantly different in terms of **(A)** RB1(N), **(B)** FLNA(C), **(C)** CSNK1E(C)-SHC1(N), **(D)** CSNK1E(C)-RB1(N), **(E)** CDH3(C)-STK17A(N), **(F)** BRCA1(N)-SHC1(N), **(G)** SHC1(N)-TP53(N), **(H)** FLNA(C)-SHC1(C), **(I)** FLNA(C)-KRAS(C), and **(J)** POLB(N)-SGK2(C). Curves for patients with paired abnormal IHCs (according to **Table 3**) are plotted with dashed lines. Curves for the other patients are plotted with solid lines. The symbols ↑ and ↓ denote overexpression and underexpression of the corresponding IHCs, respectively.

portions. Univariate Cox regression procedure revealed that CSNK1E(C)↓-SHC1(N)↓, CSNK1E(C)↓-RB1(N)↑, CDH3(C)↑-STK17A(N)↑, BRCA1(N)↓-SHC1(N)↓, and SHC1(N)↓-TP53↑ were associated with poorer overall survival (**Figures 3C–G**; P = 1.8×10^{-7}, 0.001, 0.010, 0.018, and 0.048, respectively; log-rank test). On the other hand, FLNA(C)↑-SHC1(N)↓, FLNA(C)↑-KRAS↑, and POLB↓-SGK2↑ were correlated with better overall survival (**Figures 3H–J**; P = 0.032, 0.035, and 0.044, respectively; log-rank test).

Multivariate Cox Regression Analysis Revealed That CSNK1E↓-SHC1(N)↓, CSNK1E↓-RB1↑, and BRCA1(N)↓-SHC1(N)↓Were Independent Prognostic Markers

As reported previously, biomarkers can be identified from gene- or protein expression data (Presson et al., 2011; Ha et al., 2015). For the 29 IHC results, univariate Cox regression models (**Table 4A**) confirmed that RB1(N) [hazard ratio (95% confidence interval) = 2.03 (1.07–3.86); $P < 0.05$] was a prognostic marker. The univariate Cox regression analysis was also applied to the combined IHC pairs. The results suggested that CSNK1E↓-SHC1(N)↓ [hazard ratio (95% confidence interval) = 7.54 (3.08–18.43); $P < 0.001$], CSNK1E↓-RB1↑ [hazard ratio (95% confidence interval) = 2.92 (1.46–5.83); $P = 0.002$], CDH3(C)↑-STK17A(N)↑ [hazard ratio (95% confidence interval) = 3.58 (1.27–10.10); $P = 0.016$], BRCA1(N)↓-SHC1(N)↓ [hazard ratio (95% confidence interval) = 2.96 (1.15–7.59); $P = 0.024$], and FLNA(C)↑-KRAS↑[hazard ratio (95% confidence interval) = 0.49 (0.25–0.96); $P = 0.039$] were significant predictors of the risk of death in Asian patients with OSCC (**Table 4**). In addition, the paired markers CSNK1E↓-SHC1(N)↓ (LR_x^2 = 12.8) and CSNK1E(C)↓-RB1(N)↑ (LR_x^2 = 7.6) provided more powerful prognostic information than the individual marker RB1(N) (LR_x^2 = 4.7). There were too few patients in the MSH2↓-TP53↑ and MSH2↓-SHC1↓ subsets to perform univariate Cox regression analysis.

Of the clinical variables [age, sex, tumor grade, lymph node (LN) metastasis and stage], grade, lymph node metastasis and stage were significantly associated with the patients' overall survival (**Table 4A**). The univariate model based on stage (LR_x^2= 13.1) fit better than that based on grade (LR_x^2= 4.8). Therefore, we used stage as the adjustment factor in the multivariate Cox regression models.

Because the high incidence of oral cancer in Asian OSCC is related to alcohol use, betel nut chewing, and cigarette smoking, we investigated whether these habits were associated with overall survival in this population. As shown in **Table 4A**, only alcohol use [hazard ratio (95% confidence interval) = 2.01 (1.01–3.97); *P* = 0.045] was a significant predictor of overall survival in these Taiwanese patients with OSCC. Betel nut chewing [hazard ratio (95% confidence interval) = 0.72 (0.38–1.38); *P* = 0.329] and smoking [hazard ratio (95% confidence interval) = 1.79 (0.64–5.00); *P* = 0.267] were not significant predictors for the risk of death in these patients. Furthermore, the correlation between alcohol use and each IHC marker was tested (Fisher's exact test) and none was significant at *P* = 0.05. Similarly, the correlation between "the combined habits" and each IHC marker was tested, and again none was significant at *P* = 0.05; please see **Supplementary Table 3** for details.

We then evaluated the associations of the novel IHC prognostic markers with overall survival after adjusting for alcohol use, age and tumor stage, *via* multivariate Cox regression analysis. The paired prognostic markers CSNK1E↓-SHC1(N)↓ [hazard ratio (95% confidence interval) = 7.75 (2.85–21.07); $P = 5.9 \times 10^{-5}$], CSNK1E↓-RB1(N)↑ [hazard ratio (95% confidence interval) = 2.16 (1.02–4.58); *P* = 0.045], and BRCA1(N)↓-SHC1(N)↓ [hazard ratio (95% confidence interval) = 2.87 (1.11–7.42); *P* = 0.030] were significant predictors of the overall survival of the patients with OSCC. For patients with CSNK1E↓-SHC1(N)↓, CSNK1E↓-RB1(N)↑, and BRCA1(N)↓-SHC1(N)↓, the risk of death was 7.8, 2.2, and 2.9 times higher, respectively, than that for the other patients in this population. However, RB1(N) [hazard ratio (95% confidence interval) = 1.71 (0.89–3.30); *P* = 0.108] was no longer a significant predictor (**Table 4B**). After alcohol use, age, and stage were entered into the multivariate Cox regression models along with each of the markers, neither alcohol use nor age was selected by stepwise selection or Akaike information criterion (AIC). Thus, neither appeared in the final models (**Table 4B**). Following a reviewer's suggestion, we further adjusted the effect of lymph node metastasis and tumor stage in the multivariate Cox regression models, because lymph node density and metastasis were shown to be significant prognosis predictors in OSCC (Zanaruddin et al., 2013; Chang et al., 2018). Excluding the effect of LN metastasis and stage that are used in clinical practice conventionally, the revealed five combined markers are still significant (**Supplementary Table 4**). This highlights the potential of these markers being targeted for cancer treatments.

Combinations of Significant Markers Were Studied; CSNK1E↓-SHC1(N) Was Suggested for Clinical Practice

We then combined any two of the significant markers in **Table 4A** and selected eight combinations whose good/poor OS subsets consisted of a sufficient number of (≥5) patients. Note that the (↑,↑) subset of FLNA(C)-KRAS(C) was correlated with a good OS, so we combined its complementary subsets (↓, ↑), (↑, ↓), and (↓, ↓) with the poor OS subsets of the remaining five markers in **Table 4A**. We fitted multivariate Cox regression to these combinations, and found that [CSNK1E-SHC1(N), FLNA(C)-KRAS(C)] (hazard ratio = 8.71; *P* = 0.000 rounded to three decimal places) and [BRCA1(N)-SHC1(N), FLNA(C)-KRAS(C)] (hazard ratio = 3.14; *P* = 0.020) were significant prognostic markers (**Table 4C**). For combinations of three or more significant markers, the (good/poor OS) subsets had too few patients to fit any multivariate Cox regression model. Taking **Table 4A–C** together, we suggest using the combination CSNK1E↓-SHC1(N)↓, which has the most significant *P* value (from likelihood-ratio test) among all markers, to identify Asian OSCC patients with worst survival in clinical practice.

TABLE 3 | Immunohistochemistry (IHC) proteins derived from cancerous tissues which were sampled from 163 local oral cancer patients, the cutoff values for over- and under-expression of IHC.

No. No.	Protein name	Criterion for under-expression	Criterion for over- expression
1	BRCA1(N)[a]	<1+	≧1+
2	CDH3(C)[a]	<1+	≧1+
3	CDH3(N)	<1+	≧1+
4	CDK6(C)	≤1+	>1+
5	CSNK1E(C)	≤1+	>1+
6	EGFR(C)	≤1+	>1+
7	EGFR(M)[a]	<1+ and < 10%[b]	≧1+ and ≧10%[b]
8	FEN1(C)	<1+	≧1+
9	FLNA(C)	<1+	≧1+
10	FLNA(N)	≤1+	>1+
11	KRAS(C)	<1+	≧1+
12	MET (C)[c]	≤1+	>1+
13	MSH2(N)	<1+	≧1+
14	P16(C)	≤1+	>1+
15	P16(N)	<1+	≧1+
16	PARP1(N)	<1+	≧1+
17	PIM1(C)	<3+	≧3+
18	PIM1(N)	<3+	≧3+
19	PLK1(C)	<3+	≧3+
20	POLB(C)	≤ 1+	>1+
21	POLB(N)	≤ 1+	>1+
22	RAD54B(N)	<1+	≧1+
23	RB1(N)	<1+	≧1+
24	SGK2(C)	≤1+	>1+
25	SHC1(C)	<1+	≧1+
26	SHC1(N)	≤1+	>1+
27	STK17A(C)	≤1+	>1+
28	STK17A(N)	≤1+	>1+
29	TP53(N)	≤0%	>0%

[a] The notation (C), (N) and (M) represent cytoplasm, nucleus and membrane, respectively.

[b] Both strength of staining ≥ 1+ and stained area ≥ 10% are required.

[c] Phosphorylated MET was stained.

[d] The stained area > 0% is required.

External Validation of the Association of CDH3-STK17A With Overall Survival

Ethnicity and geography may play a role in the etiology of cancer. If the newly discovered markers are confirmed by independent

TABLE 4 | Overall survival of 153 oral squamous cell carcinoma patients relative to clinical covariates, IHC prognostic markers, and habits.

A. Univariate Cox regression

Variable	Subset	Hazard ratio (95% CI)	*p*-value	LR_{x^2}
Lymph node metastasis	yes/no	3.47 (1.92–6.28)	0.000	15.6
Stage	III–IV/I–II	3.15 (1.67–5.95)	0.000	13.1
Grade	low, moderate and high	1.94 (1.06–3.54)	0.031	4.8
RB1(N)	↑/↓[a]	2.03 (1.07–3.86)	0.031	4.7
[CSNK1E(C), SHC1(N)]	(↓, ↓)/otherwise	7.54 (3.08–18.43)	0.000	12.8
[CSNK1E(C), RB1(N)]	(↓, ↑)/otherwise	2.92 (1.46–5.83)	0.002	7.6
[CDH3(C), STK17A(N)]	(↑, ↑)/otherwise	3.58 (1.27–10.10)	0.016	5.4
[BRCA1(N), SHC1(N)]	(↓, ↓)/otherwise	2.96 (1.15–7.59)	0.024	4.2
[FLNA(C), KRAS(C)]	(↑, ↑)/otherwise	0.49 (0.25–0.96)	0.039	3.9

Habit	Subset	Hazard ratio (95% CI)	*p*-value	LR_{x^2}
Alcohol use	Yes/No	2.01 (1.01–3.97)	0.045	4.4

B. Multivariate Cox Regression[b]

Variable	Subset	Hazard ratio (95% CI)	*p*-value	LR_{x^2}
RB1(N)	↑/↓	1.71(0.89–3.30)	0.108	16.2
Stage	III–IV/I–II	3.18(1.65–6.14)	0.001	
[CSNK1E(C), SHC1(N)]	(↓, ↓)/otherwise	7.75(2.85–21.07)	5.9×10^{-5}	23.3
Stage	III–IV/I–II	3.45(1.74–6.85)	4.1×10^{-4}	
[CSNK1E(C), RB1(N)]	(↓, ↑)/otherwise	2.16(1.02–4.58)	0.045	16.4
Stage	III–IV/I–II	3.04(1.57–5.87)	0.001	
[BRCA1(N), SHC1(N)]	(↓, ↓)/otherwise	2.87(1.11–7.42)	0.030	17.1
Stage	III–IV/I–II	3.34 (1.68–6.61)	0.001	

C. Combination of two gene pairs

Variable	Hazard ratio (95% CI)	*p*-value	LR_{x^2}
CSNK1E(C)-SHC1(N) (↓, ↓) and FLNA(C)-KRAS(C) (↑, ↑)[c]*	8.71(2.88–26.36)	0.000	1.98
Stage	2.95(1.45–6.02)	0.003	
BRCA1(N)-SHC1(N) (↓, ↓) and FLNA(C)-KRAS(C) (↑, ↑)	3.14 (1.2–8.24)	0.020	14.94
Stage	2.91 (1.42–5.95)	0.004	

[a] *The symbols "↑" and "↓" denote over- and under-expression of IHC, respectively of the corresponding protein.*

[b] *Variables were selected by stepwise selection and AIC.*

**The symbol (↑, ↑);[c]denotes the complementary set of (↑, ↑), namely (↓, ↑), (↑, ↓) and (↓, ↓), in which FLNA(C)-KRAS(C) is in the same direction (poor OS) as that of BRCA1(N)-SHC1(N).*

datasets of patients with OSCC from different geographic regions and ethnicities, they may be useful tools in clinical medicine. OSCC with HPV(−) from the TCGA head and neck SCC cohort (henceforth, TCGA) (The Cancer Genome Atlas Network, 2015) more closely resembled Asian OSCC than those with HPV(+). Therefore, we analyzed microarray gene expression data of 160 OSCC cases with HPV (−) to validate the novel prognostic markers in **Table 4**.

We used 1.4-fold as the cutoff for differential expression of TCGA OSCC RNA-seq data (see section "Materials and Methods" for details), such that of all markers in **Table 4A**, *CDH3-STK17A* and *FLNA-KRAS* had a sufficient number of (five or more) (Vittinghoff and McCulloch, 2007) patients in the (good/poor OS) subsets for the univariate Cox regression analysis. Of these, *CDH3*↑*-STK17A*↑ was a significant gene predictor of good survival [hazard ratio (95% confidence interval) = 0.55(0.35–0.87); *P* = 0.011], while *FLNA*↑*-KRAS*↑ was not significant (*P* = 0.117). The former finding was not consistent with ours, which showed that CDH3(C)↑-STK17A(N)↑ was a significant predictor of poorer overall survival in Taiwanese patients with OSCC (**Table 4A**). This discrepancy may be explained by the different genetic backgrounds in the two populations, since the significant downregulation of the *CDH3* gene has been reported in metastatic OSCC (Méndez et al., 2009). The estimated survival curves of *CDH3-STK17A* are shown in **Figure 4**. This external validation demonstrates that if IHC or gene expression data are available, CDH3-STK17A can be used to stratify patients with OSCC in the future.

DISCUSSION

Here, we established a cost-effective approach for the identification of prognostic IHC markers of OSCC. This approach is also efficient, as merely 29 IHC stains were performed, but five clinically beneficial prognostic markers were identified through extensive statistical analysis. Our technique rapidly uncovered the prognostic markers without any prerequisite knowledge of the molecular pathways. In contrast, previous studies relied on pathway information

(Sadanandam et al., 2013; Kosari et al., 2014) to reveal prognostic markers, such as cellular phenotypes and protein expression levels. Moreover, our approach was able to reveal IHC prognostic markers with components from different pathways. This improves the current state of art, as most of methods to uncover IHC markers to date have been mainly based on one or two proteins (Lin et al., 2014), or one pathway (Oliveira-Costa et al., 2014).

Of the single IHC results, RB1(N) was a predictor of poorer survival in the Taiwanese patients with OSCC, however, it was not independent of tumor stage. This finding was consistent with earlier studies wherein RB1 was a biomarker in HPV(−) head and neck cancers (The Cancer Genome Atlas Network, 2015; Beck et al., 2016). Previous studies showed that expression of Rb increased in the development and/or with disease progression of OSCC (Pavelic et al., 1996; Schoelch et al., 1999; Thomas et al., 2015), and the latter study reported high expression of Rb in patients with combined habits (alcohol use, betel nut chewing and smoking), suggesting Rb pathway altered. However, in our study, the over-expression of Rb was confounded with stage (**Tables 4A,B**), but not associated with the combined habits ($P = 0.19$, Fisher's exact test; **Supplementary Table 3**). The over-expression of RB1(N) in our study may be due to over-expression of cyclin D1 or under-expression of p16^{INK4A}, as cyclin D1 and p16^{INK4A} are related to Rb through an autoregulatory loop (Gimenez-Conti et al., 1996; Andl et al., 1998). Although expression of RB1(N) was high in our study, its function was likely inactivated which may be due to regulation of cyclin D1, HPV infection (Gimenez-Conti et al., 1996; Andl et al., 1998), loss of heterozygosity (Maestro et al., 1996; Yokoyama et al., 1996) or Rb hyperphosporylation (Chatterjee et al., 2004), but further studies are required to elucidate this.

Of the 398 IHC pairs, multivariate Cox regression analyses showed that CSNK1E↓-SHC1(N)↓, CSNK1E↓-RB1(N)↑, and BRCA1(N)↓-SHC1(N)↓ were significant predictors of the risk of death in this Taiwanese OSCC population, independent of tumor stage. Of all combinations of two significant markers in **Table 4A**, [CSNK1E-SHC1(N), FLNA(C)-KRAS(C)] was the most significant poor prognostic factor. Nevertheless, this marker was less significant than CSNK1E↓-SHC1(N)↓ statistically. Thus, in clinical practice we recommend using CSNK1E↓-SHC1(N)↓ to identify patients with severe and/or advanced Asian OSCC, who should be suggested for alternate or more intense treatment strategies in clinical practice. CK1 ϵ could be an oncoprotein or a tumor suppressor (Lin et al., 2014), but phosphorylation of CK1 ϵ can stabilize and activate tumor suppressor p53 (Knippschild et al., 2005). SHC1 is a known downstream target of p53, which involves in stress-induced signal transduction pathway (Trinei et al., 2002). Moreover, SHC1 was downregulated by miR-5582-5p, thus led a tumor suppressive activity with GAB1 (An et al., 2016). In our study, the mean survival rate of OSCC patients with CSNK1E↓-SHC1↓ is 13.8 months compared to 37.8 months of the remaining group. Collectively, CSNK1E-SHC1 might be a tumor suppressor, but this requires further studies for elucidation. As phosphorylation of CK1 ϵ can stabilize and activate tumor suppressor p53, moreover, expression of p53 was lower in OSCC lesions than in malignant lesions, and Rb expression was observed in OSCC lesions (Oliveira and Ribeiro-Silva, 2011). Thus, we speculate p53 may indirectly interfere with RB1, after p53 been regulated by phosphorylated CK1, which supports the finding CSNK1E↓-RB1(N)↑ is a poor prognostic marker.

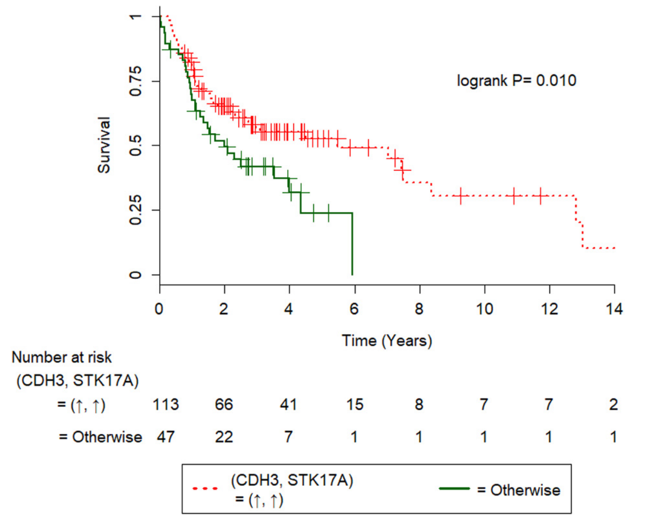


FIGURE 4 | Kaplan-Meier survival curves of 160 HPV(−) oral squamous cell carcinoma patients from the TCGA cohort. Kaplan-Meier survival curves were significantly different in terms of gene expression for CDH3-STK17A, where the symbols ↑ and ↓ denote overexpression and underexpression of the corresponding genes at the 1.4-fold cutoff.

External gene expression data of HPV(−) OSCC from the TCGA cohort (98.7% non-Asian patients) validated that *CDH3*↑-*STK17A*↑as a significant predictor of good survival in 160 patients with HPV(−) OSCC, where the cutoff was set at 1.4-fold. Nevertheless, when we set the cutoff at 1.5-fold, this gene pair was no longer significant ($P = 0.124$). Thus, this gene pair may not be a robust prognosis marker. Our result showed that CDH3(C)↑-STK17A(N)↑ was correlated with poor survival of 163 Taiwanese patients with OSCC. The difference in the aforementioned results may be because overexpression of P-cadherin (coded by CDH3) in membrane was associated with good survival of 67 OSCC patients, however, cytoplasmic expression of P-cadherin was correlated with poor survival (Muzio et al., 2005). Furthermore, high cytoplasmic expression of STK17A was reported to increase tumorigenic potential through inhibition of TGF-beta1-mediated tumor suppressor activity in HNSCC cells (Park et al., 2015).

Some of the prognostic markers our approach revealed are well-known and reported in the literature, thus we performed a comparative analysis as follows. Among all biomarkers, CSNK1E↓-SHC1(N)↓ was the most significant in terms of $P < 0.0001$ (**Table 5**). Consistently, the loss of CK1ε expression was shown to be a poor prognostic marker in Taiwanese patients with oral cancer (Lin et al., 2014). Next, overexpression of cyclin D1 and Rb and low expression of p16 was significantly associated with reduced disease-free survival in 348 Indian patients with OSCC (Jayasurya et al., 2005), consistent with our findings that the overexpression of cytoplasmic Rb was a poor prognostic marker in Taiwanese patients. However, Soni et al. (2005) reported that 105 Indian patients with OSCC

TABLE 5 | A comparison of our prognostic markers to those reported in literature.

Previous studies[a]		**This study**	
IHC marker	**Sample size/*P* value**	**IHC marker**	**Sample size/*P* value**
Cyclin D↑-Rb↑-p16↓$^{\text{Jayasurya}}$	348/0.002	CSNK1E↓- SHC1(N)↓	163/5.9×10^{-5}
Rb↑	348/0.062	CSNK1E↓- Rb↑	163/0.002
Rb↓$^{\text{Soni}}$	98/0.036	Rb↑	163/0.031
Rb↓-p53↑$^{\text{Soni}}$	98/0.004		
CSNK1E↓$^{\text{Lin}}$	195/0.024	CSNK1E↓	163/insignificant[b]
p53-Cyclin D1-EGFR$^{\text{Shiraki}}$	140/0.0019		
EGFR	140/insignificant	EGFR↑	163/insignificant
p53	140/insignificant	p53↑	163/insignificant
BRCA1↑$^{\text{Oliveira}*}$	150/0.030	BRCA1(N)↓	163/insignificant
		BRCA1(N)↓- SHC1(N)↓	163/0.024
P-cadherin↓$^{\text{Muzio}}$	67/0.056	CDH3(C)↑-STK17A (N)↑	163/0.016
		CDH3(C)↑	163/insignificant

[a] *The first author's last name was indicated in the upper right corner of each study's first marker.*
[b] *P value ≥ 0.05.*
**BRCA1↑ was associated with disease-specific survival.*

with loss of Rb expression had poor prognosis. This discrepancy may be explained by a recent finding (Sanidas et al., 2019) that there are many different forms of active Rb, and they have distinct functional properties. Both Shiraki et al. (2005) and our team found that overexpression of p53 (EGFR) was not a significant prognostic marker in OSCC, but the former study revealed that p53-Cyclin D1-EGFR was significantly associated with poor overall survival (P = 0.019). Moreover, BRCA1 overexpression was shown to be associated with reduced overall survival of 150 Brazilian patients with OSCC (Oliveira-Costa et al., 2014), whereas we did not find prognostic significance of BRCA1 underexpression, but BRCA1(N) ↓-SHC1(N)↓ was an independent prognostic marker (P = 0.024; **Table 4B**). This discrepancy may be due to the different genetic backgrounds of the populations.

In conclusion, our study revealed that the combined evaluation of CSNK1E↓-SHC1(N)↓ in OSCC identified a group of patients with the poorest survival, who should be suggested to undergo alternate or more intense treatment strategies. CK1ε combined with SHC adaptor protein 1 emerged as the most promising IHC prognostic marker in Asian OSCC. Of the 398 combined IHC pairs, genes of ten pairs are known to be SL, out of which only FLNA-KRAS was revealed to be a good OS marker, but not independent of stage. Excluding the effect of tumor stage and LN metastasis (Zanaruddin et al., 2013; Chang et al., 2018) that are used in clinical practice conventionally, the revealed markers of our study are still significant (**Supplementary Table 4**). This highlights the potential of these markers being targeted for cancer treatments.

Despite that we conducted a large scale of IHC study, the present study is limited by the moderate sample size and no genomic data profiled. Further studies based on larger sample sizes of patients with OSCC and on DNA sequencing data will reveal whether the expression of the uncovered IHC markers are due to their mutations. With ready availability of gene expression and tissue array data and resources to match clinicopathological features in the public and commercial domains, our approach can immediately be applied to other types of cancers. Moreover, additional IHC stain of cyclin D1 will enable us to evaluate the prognostic significance of protein triplets such as cyclin D1-Rb-p16 and p53-cyclin D1-EGFR. This is interesting, as the component of our most significant marker SHC adaptor protein 1-CK1↑ is involved in the EGFR pathway and is SL to both TP53 and EGFR, respectively. Given that the triplet IHC cyclin D1-Rb-p16 is a promising marker, future studies will extend to the prognostic effect of triplets of IHC in OSCC.

DATA AVAILABILITY STATEMENT

Part of the datasets in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the articles. However, the clinical data that support the findings of this study are available from Kaohsiung Medical University Hospital with restrictions applying to the availability of these data, which were used under license for the current study, are not publicly available. Data are however available from the authors upon reasonable request and with permission from Kaohsiung Medical University Hospital.

ETHICS STATEMENT

This study involving human participants was approved by the Institutional Review Board and Ethics Committee of Kaohsiung Medical University Hospital (KMUHIRB-E(I)-20170034). The data were analyzed anonymously, and therefore, no informed consent was required. All methods were performed under approved guidelines and regulations.

AUTHOR CONTRIBUTIONS

GS conceived the study. H-CW, C-CW, and Y-TC did IHC staining. C-JC implemented the methods, wrote the algorithm, and performed data analysis. H-CW, J-GC, and GS interpreted the data. H-CW and GS wrote the manuscript; T-CL modified H-CW's writing in an earlier version. GS and T-CL designed and supervised the study. All authors read and approved the final manuscript.

FUNDING

This work was supported in part by Grant No. S10518-3 and KMUH109-9T02 from Kaohsiung Medical University Hospital, Kaohsiung, Taiwan to H-CW and T-CL, and the Ministry of Science and Technology, Republic of China [MOST 106-2118-M-001-017 and MOST 107-2118-M-001-009-MY2] to GS.

ACKNOWLEDGMENTS

We are grateful to the editor and three reviewers for constructive comments which improved the manuscript much. We are also grateful to Sam Tsai for computational work of the revisions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.643461/full#supplementary-material

REFERENCES

An, H. J., Kwak, S. Y., Yoo, J. O., Kim, J. S., Bae, I. H., Park, M. J., et al. (2016). Novel miR-5582-5p functions as a tumor suppressor by inducing apoptosis and cell cycle arrest in cancer cells through direct targeting of GAB1, SHC1, and CDK2. *Biochim. Biophys. Acta* 1862, 1926–1937. doi: 10.1016/j.bbadis.2016.07.017

Andl, T., Kahn, T., Pfuhl, A., Nicola, T., Erber, R., Conradt, C., et al. (1998). Etiological involvement of oncogenic human papillomavirus in tonsillar squamous cell carcinomas lacking retinoblastoma cell cycle control. *Cancer Res.* 58, 5–12.

Barbie, D. A., Tamayo, P., Boehm, J. S., Kim, S. Y., Moody, S. E., Dunn, I. F., et al. (2009). Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* 462:108. doi: 10.1038/nature08460

Beck, T. N., Georgopoulos, R., Shagisultanova, E. I., Sarcu, D., Handorf, E. A., Dubyk, C., et al. (2016). EGFR and RB1 as dual biomarkers in HPV-negative head and neck cancer. *Mol. Cancer Therap.* 15, 2486–2497. doi: 10.1158/1535-7163.mct-16-0243

Belcher, R., Hayes, K., Fedewa, S., and Chen, A. Y. (2014). Current treatment of head and neck squamous cell cancer. *J. Surg. Oncol.* 110, 551–574.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," in *CA: A Cancer Journal for Clinicians*, 68, 394–424. doi: 10.3322/caac.21492

Bryant, H. E., Schultz, N., Thomas, H. D., Parker, K. M., Flower, D., Lopez, E., et al. (2005). Specific killing of BRCA2-deficient tumours with inhibitors of poly (ADP-ribose) polymerase. *Nature* 434, 913. doi: 10.1038/nature03443

Chang, J.-G., Chen, C.-C., Wu, Y.-Y., Che, T.-F., Huang, Y.-S., Yeh, K.-T., et al. (2016). Uncovering synthetic lethal interactions for therapeutic targets and predictive markers in lung adenocarcinoma. *Oncotarget* 7, 73664. doi: 10.18632/oncotarget.12046

Chang, W. C., Lin, C. S., Yang, C. Y., Lin, C. K., and Chen, Y. W. (2018). Lymph node density as a prognostic predictor in patients with betel nut-related oral squamous cell carcinoma. *Clin. Oral Invest.* 22, 1513–1521. doi: 10.1007/s00784-017-2247-3

Chatterjee, S. J., Datar, R., Youssefzadeh, D., George, B., Goebell, P. J., Stein, J. P., et al. (2004). Combined effects of p53, p21, and pRb expression in the progression of bladder transitional cell carcinoma. *J. Clin. Oncol.* 22, 1007–1013. doi: 10.1200/JCO.2004.05.174

Ding, L., Getz, G., Wheeler, D. A., Mardis, E. R., Mclellan, M. D., Cibulskis, K., et al. (2008). Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* 455:1069.

Dowsett, M., Nielsen, T. O., A'hern, R., Bartlett, J., Coombes, R. C., Cuzick, J., et al. (2011). Assessment of Ki67 in breast cancer: recommendations from the International Ki67 in Breast Cancer working group. *J. Natl. Cancer Inst.* 103, 1656–1664. doi: 10.1093/jnci/djr393

Farmer, H., Mccabe, N., Lord, C. J., Tutt, A. N., Johnson, D. A., Richardson, T. B., et al. (2005). Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature* 434:917. doi: 10.1038/nature03445

Gillison, M. L., Koch, W. M., Capone, R. B., Spafford, M., Westra, W. H., Wu, L., et al. (2000). Evidence for a causal association between human papillomavirus and a subset of head and neck cancers. *J. Natl. Cancer Inst.* 92, 709–720. doi: 10.1093/jnci/92.9.709

Gimenez-Conti, I. B., Collet, A. M., Lanfranchi, H., Itoiz, M. E., Luna, M., Xu, H. J., et al. (1996). p53, Rb, and cyclin D1 expression in human oral verrucous carcinomas. *Cancer* 78, 17–23. doi: 10.1002/(sici)1097-0142(19960701)78:1<17::aid-cncr4>3.0.co;2-e

Ha, M. J., Baladandayuthapani, V., and Do, K.-A. (2015). Prognostic gene signature identification using causal structure learning: applications in kidney cancer: supplementary issue: sequencing platform modeling and analysis. *Cancer inform.* 14:S14873.

Hammond, M. E. H., Hayes, D. F., Dowsett, M., Allred, D. C., Hagerty, K. L., Badve, S., et al. (2010). American Society of Clinical Oncology/College of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer (unabridged version). *Arch. Pathol. Lab. Med.* 134, e48–e72.

Jayasurya, R., Sathyan, K., Lakshminarayanan, K., Abraham, T., Nalinakumari, K., Abraham, E. K., et al. (2005). Phenotypic alterations in Rb pathway have more prognostic influence than p53 pathway proteins in oral carcinoma. *Mod. Pathol.* 18, 1056–1066. doi: 10.1038/modpathol.3800387

Kaelin, W. G. (2005). The concept of synthetic lethality in the context of anticancer therapy. *Nat. Rev. Cancer* 5, 689–698. doi: 10.1038/nrc1691

Kang, X., Chen, W., Kim, R. H., Kang, M. K., and Park, N.-H. (2009). Regulation of the hTERT promoter activity by MSH2, the hnRNPs K and D, and GRHL2 in human oral squamous cell carcinoma cells. *Oncogene* 28, 565–574. doi: 10.1038/onc.2008.404

Knippschild, U., Wolff, S., Giamas, G., Brockschmidt, C., Wittau, M., Würl, P. U., et al. (2005). The role of the casein kinase 1 (CK1) family in different signaling pathways linked to cancer development. *Oncol. Res. Treat.* 28, 508–514. doi: 10.1159/000087137

Kosari, F., Ida, C., Aubry, M., Yang, L., Kovtun, I. V., Klein, J., et al. (2014). ASCL1 and RET expression defines a clinically relevant subgroup of lung adenocarcinoma characterized by neuroendocrine differentiation. *Oncogene* 33:3776. doi: 10.1038/onc.2013.359

Lin, S.-H., Lin, Y.-M., Yeh, C.-M., Chen, C.-J., Chen, M.-W., Hung, H.-F., et al. (2014). Casein kinase 1 epsilon expression predicts poorer prognosis in low

T-stage oral cancer patients. *Int. J. Mol. Sci.* 15, 2876–2891. doi: 10.3390/ijms15022876

Liu, C. J., Chang, K. W., Chao, S. Y., Kwan, P. C., Chang, S. M., Yen, R. Y., et al. (2004). The molecular markers for prognostic evaluation of areca-associated buccal squamous cell carcinoma. *J. Oral Pathol. Med.* 33, 327–334. doi: 10.1111/j.1600-0714.2004.00092.x

Luo, J., Emanuele, M. J., Li, D., Creighton, C. J., Schlabach, M. R., Westbrook, T. F., et al. (2009). A genome-wide RNAi screen identifies multiple synthetic lethal interactions with the Ras oncogene. *Cell* 137, 835–848. doi: 10.1016/j.cell.2009.05.006

Maestro, R., Piccinin, S., Doglioni, C., Gasparotto, D., Vukosavljevic, T., Sulfaro, S., et al. (1996). Chromosome 13q deletion mapping in head and neck squamous cell carcinomas: identification of two distinct regions of preferential loss. *Cancer Res.* 56, 1146–1150.

Méndez, E., Houck, J. R., Doody, D. R., Fan, W., Lohavanichbutr, P., Rue, T. C., et al. (2009). A genetic expression profile associated with oral cancer identifies a group of patients at high risk of poor survival. *Clin. Cancer Res.* 15, 1353–1361. doi: 10.1158/1078-0432.ccr-08-1816

Muzio, L. L., Campisi, G., Farina, A., Rubini, C., Pannone, G., Serpico, R., et al. (2005). P-cadherin expression and survival rate in oral squamous cell carcinoma: an immunohistochemical study. *BMC Cancer* 5:63. doi: 10.1186/1471-2407-5-63

Oliveira, L., and Ribeiro-Silva, A. (2011). Prognostic significance of immunohistochemical biomarkers in oral squamous cell carcinoma. *Int. J. Oral Maxillofac. Surg.* 40, 298–307. doi: 10.1016/j.ijom.2010.12.003

Oliveira-Costa, J. P., Oliveira, L. R., Zanetti, R., Zanetti, J. S., Da Silveira, G. G., Buim, M. E. C., et al. (2014). BRCA1 and γH2AX as independent prognostic markers in oral squamous cell carcinoma. *Oncoscience* 1:383. doi: 10.18632/oncoscience.47

Park, Y., Kim, W., Lee, J. M., Park, J., Cho, J. K., Pang, K., et al. (2015). Cytoplasmic DRAK1 overexpressed in head and neck cancers inhibits TGF-β1 tumor suppressor activity by binding to Smad3 to interrupt its complex formation with Smad4. *Oncogene* 34, 5037–5045. doi: 10.1038/onc.2014.423

Pavelic, Z. P., Lasmar, M., Pavelic, L. J., Sorensen, C., Stambrook, P. J., Zimmermann, N., et al. (1996). Absence of Retinoblastoma gene product in human primary oral cavity carcinoma. *Eur. J. Cancer B Oral Oncol.* 32, 347–351. doi: 10.1016/0964-1955(96)00025-5

Peng, C.-H., Liao, C.-T., Peng, S.-C., Chen, Y.-J., Cheng, A.-J., Juang, J.-L., et al. (2011). A novel molecular signature identified by systems genetics approach predicts prognosis in oral squamous cell carcinoma. *PLoS One* 6:e23452. doi: 10.1371/journal.pone.0023452

Pickering, C. R., Zhang, J., Yoo, S. Y., Bengtsson, L., Moorthy, S., Neskey, D. M., et al. (2013). Integrative genomic characterization of oral squamous cell carcinoma identifies frequent somatic drivers. *Cancer Discov.* 3, 770–781. doi: 10.1158/2159-8290.cd-12-0537

Presson, A. P., Yoon, N. K., Bagryanova, L., Mah, V., Alavi, M., Maresh, E. L., et al. (2011). Protein expression based multimarker analysis of breast cancer samples. *BMC cancer* 11:230. doi: 10.1186/1471-2407-11-230

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Sadanandam, A., Lyssiotis, C. A., Homicsko, K., Collisson, E. A., Gibb, W. J., Wullschleger, S., et al. (2013). A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nat. Med.* 19, 619. doi: 10.1038/nm.3175

Sanidas, I., Morris, R., Fella, K. A., Rumde, P. H., Boukhali, M., Tai, E. C., et al. (2019). A code of mono-phosphorylation modulates the function of RB. *Mol. Cell* 73, 985.–1000.

Schoelch, M. L., Regezi, J. A., Dekker, N. P., Ng, I. O., McMillan, A., Ziober, B. L., et al. (1999). Cell cycle proteins and the development of oral squamous cell carcinoma. *Oral Oncol.* 35, 333–342.

Shiraki, M., Odajima, T., Ikeda, T., Sasaki, A., Satoh, M., Yamaguchi, A., et al. (2005). Combined expression of p53, cyclin D1 and epidermal growth factor receptor improves estimation of prognosis in curatively resected oral cancer. *Mod. Pathol.* 18, 1482–1489. doi: 10.1038/modpathol.3800455

Soni, S., Kaur, J., Kumar, A., Chakravarti, N., Mathur, M., Bahadur, S., et al. (2005). Alterations of rb pathway components are frequent events in patients with oral epithelial dysplasia and predict clinical outcome in patients with squamous cell carcinoma. *Oncology* 68, 314–325. doi: 10.1159/000086970

Srivastava, M., and Raghavan, S. C. (2015). DNA double-strand break repair inhibitors as cancer therapeutics. *Chem. Biol.* 22, 17–29. doi: 10.1016/j.chembiol.2014.11.013

Storey, J. D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A.* 100, 9440–9445. doi: 10.1073/pnas.1530509100

Su, J.-L., Shih, J.-Y., Yen, M.-L., Jeng, Y.-M., Chang, C.-C., Hsieh, C.-Y., et al. (2004). Cyclooxygenase-2 induces EP1-and HER-2/Neu-dependent vascular endothelial growth factor-C up-regulation: a novel mechanism of lymphangiogenesis in lung adenocarcinoma. *Cancer Res.* 64, 554–564. doi: 10.1158/0008-5472.can-03-1301

The Cancer Genome Atlas Network. (2015). Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* 517, 576–581. doi: 10.1038/nature14129

Thomas, S., Balan, A., and Balaram, P. (2015). The expression of retinoblastoma tumor suppressor protein in oral cancers and precancers: a clinicopathological study. *Dent. Res. J.* 12, 307–314. doi: 10.4103/1735-3327.161427

Tiong, K.-L., Chang, K.-C., Yeh, K.-T., Liu, T.-Y., Wu, J.-H., Hsieh, P.-H., et al. (2014). CSNK1E/CTNNB1 are synthetic lethal to TP53 in colorectal cancer and are markers for prognosis. *Neoplasia* 16, 441–450. doi: 10.1016/j.neo.2014.04.007

Trinei, M., Giorgio, M., Cicalese, A., Barozzi, S., Ventura, A., Migliaccio, E., et al. (2002). A p53-p66Shc signalling pathway controls intracellular redox status, levels of oxidation-damaged DNA and oxidative stress-induced apoptosis. *Oncogene* 21, 3872–3878. doi: 10.1038/sj.onc.1205513

Vittinghoff, E., and McCulloch, C. E. (2007). Relaxing the rule of ten events per variable in logistic and Cox regression. *Am. J. Epidemiol.* 165, 710–718. doi: 10.1093/aje/kwk052

Wolff, A. C., Hammond, M. E. H., Schwartz, J. N., Hagerty, K. L., Allred, D. C., Cote, R. J., et al. (2007). American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. *Arch. Pathol. Lab. Med.* 131, 18–43.

Yokoyama, J., Shiga, K., Sasano, H., Suzuki, M., and Takasaka, T. (1996). Abnormalities and the implication of retinoblastoma locus and its protein product in head and neck cancers. *Anticancer Res.* 16, 641–644.

Zanaruddin, S. N. S., Saleh, A., Yang, Y. H., Hamid, S., Mustafa, W. M. W., Bariah, A. K., et al. (2013). Four-protein signature accurately predicts lymph node metastasis and survival in oral squamous cell carcinoma. *Hum. Pathol.* 44, 417–426. doi: 10.1016/j.humpath.2012.06.007

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

ORIGINAL RESEARCH
published: 20 May 2021
doi: 10.3389/fgene.2021.652623

Construction of Condition-Specific Gene Regulatory Network Using Kernel Canonical Correlation Analysis

*Dabin Jeong [1], Sangsoo Lim [2], Sangseon Lee [3], Minsik Oh [4], Changyun Cho [1], Hyeju Seong [5], Woosuk Jung [5] and Sun Kim [1,2,6]**

[1] Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, South Korea, [2] Bioinformatics Institute, Seoul National University, Seoul, South Korea, [3] BK21 FOUR Intelligence Computing, Seoul National University, Seoul, South Korea, [4] Department of Computer Science and Engineering, Seoul National University, Seoul, South Korea, [5] Department of Crop Science, Konkuk University, Seoul, South Korea, [6] Department of Computer Science and Engineering, Institute of Engineering Research, Seoul National University, Seoul, South Korea

OPEN ACCESS

Edited by:
Shaolong Cao,
University of Texas MD Anderson Cancer Center, United States

Reviewed by:
Md. Ashad Alam,
Tulane University, United States
Wenxing Hu,
University of Pittsburgh, United States

***Correspondence:**
Sun Kim
sunkim.bioinfo@snu.ac.kr

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *12 January 2021*
Accepted: *26 March 2021*
Published: *20 May 2021*

Citation:
Jeong D, Lim S, Lee S, Oh M, Cho C, Seong H, Jung W and Kim S (2021) Construction of Condition-Specific Gene Regulatory Network Using Kernel Canonical Correlation Analysis. Front. Genet. 12:652623. doi: 10.3389/fgene.2021.652623

Gene expression profile or transcriptome can represent cellular states, thus understanding gene regulation mechanisms can help understand how cells respond to external stress. Interaction between transcription factor (TF) and target gene (TG) is one of the representative regulatory mechanisms in cells. In this paper, we present a novel computational method to construct condition-specific transcriptional networks from transcriptome data. Regulatory interaction between TFs and TGs is very complex, specifically multiple-to-multiple relations. Experimental data from TF Chromatin Immunoprecipitation sequencing is useful but produces one-to-multiple relations between TF and TGs. On the other hand, co-expression networks of genes can be useful for constructing condition transcriptional networks, but there are many false positive relations in co-expression networks. In this paper, we propose a novel method to construct a condition-specific and combinatorial transcriptional network, applying kernel canonical correlation analysis (kernel CCA) to identify multiple-to-multiple TF–TG relations in certain biological condition. Kernel CCA is a well-established statistical method for computing the correlation of a group of features vs. another group of features. We, therefore, employed kernel CCA to embed TFs and TGs into a new space where the correlation of TFs and TGs are reflected. To demonstrate the usefulness of our network construction method, we used the blood transcriptome data for the investigation on the response to high fat diet in a human and an arabidopsis data set for the investigation on the response to cold/heat stress. Our method detected not only important regulatory interactions reported in previous studies but also novel TF–TG relations where a module of TF is regulating a module of TGs upon specific stress.

Keywords: kernel canonical correlation analysis, gene regulatory network, network dynamics, transcription factor, TF cooperation, condition specific network

1. INTRODUCTION

In a living cell, rewiring of interactions among proteins, genes, and RNA molecules orchestrates how cells respond to external stimuli. One of the most fundamental regulatory relationships arise from transcription factors (TFs) that bound to the promoter of target genes (TGs) resulting in changing transcriptional dynamics. Since TF–TG interactions can be represented as a network, dynamics of gene regulatory mechanisms upon stimuli can be modeled and analyzed as gene regulatory network (GRN). High-throughput experimental techniques, such as Chromatin Immunoprecipitation sequencing (ChIP-seq), have been widely utilized to construct GRNs detecting one-to-multiple relationships of TF and TGs (i.e., relations of a TF and the promoters of TGs where the TF binds to). Such experimental techniques are powerful but provide only partial snapshot of condition-specific GRN. TF ChIP-seq can measure only one TF at a time and it is not practical to perform ChIP-seq experiments for all TFs under various conditions. More importantly, multiple TFs work together to regulate multiple TGs in a condition-specific way, thus data from TF ChIP-seq needs to be combined for constructing networks of multiple TFs and multiple TGs simultaneously. Thus, it is necessary to develop computational methods for elucidating multiple-to-multiple relations of TFs and TGs in a specific condition. There have been several studies to identify multiple-to-multiple interactions. A study by Jolma et al. (2015) tried to identify TF–TG regulations using a tailored experimental technique in a multiple-to-multiple fashion. Their work is still limited in identifying only 315 TF–TF interactions from ∼2,000 putative TFs.

There have been growing attention in *in silico* reverse engineering methods that infer GRNs from gene expression data. Correlation-based network inference methods—the most straightforward approach—detect regulatory relations if two genes are linearly correlated (Eisen et al., 1998). However, the correlation-based methods are prone to produce many false-positive relations, i.e., the relations predicted by computational methods but not detected in experimental validations, because the methods consider solely a linearly correlated expression pattern between a pair of genes. For example, if two genes *B* and *C* are regulated by a common gene *A*, expression patterns of *B* and *C* are correlated thus detected as regulatory relations even though there are no direct regulatory relationships between *B* and *C*. A number of computational methods with different strategies have been developed over two decades. Methods based on mutual-information (MI) is a generalization of correlation-based model that can detect non-linear dependencies, taking into account the effect of third-party genes in addition to two correlating genes. ARACNe (Margolin et al., 2006) and ARACNe-AP, one of the most popular reverse engineering methods, use the data-processing inequality to prune the indirect regulations if a pair of genes interact only through a third gene in every possible gene triplets. Likewise, the three-way mutual information (MI3) and conditional mutual information (CMI)-based models consider the effect of co-regulators in order to remove false-positive interactions (Luo et al., 2008; Zhang et al., 2012). Besides, regression-based methods considers multiple-to-one relations of TFs and a TG as a feature selection problem, where the expression of TGs is predicted from the expression of all other TF genes (Xiong and Zhou, 2012; Hill et al., 2016). GENIE3, one of the most best-performing methods, utilized an ensemble of regression trees to select putative TFs for each TG. Although MI-based approaches showed lower false-positive rate than correlation-based methods, they do not consider the biological nature of TFs—combinatorial and cooperative nature of TFs—when regulating TGs are disregarded.

Then, how TFs work in order to coordinate certain biological functions? First, TFs regulate a biological function through interacting with protein complexes rather than simply elevating mRNA concentration (Sutherland and Bickmore, 2009; Rieder et al., 2012; Duren et al., 2019). Therefore, to detect important TFs that are related to a certain biological function, TF interaction network should be utilized rather than simply detecting TFs with the highest mRNA concentration. Second, combinatorial interaction of TFs regulates TGs to control certain biological functions. That is, given alternative stimuli, different combinations of TFs may regulate expression of different sets of TGs to certain cellular response involving multiple-to-multiple relations of TFs and TGs. Several studies have suggested an atlas of combinatorial TF module interactions (Ravasi et al., 2010; Wise and Bar-Joseph, 2015; Guo and Gifford, 2017) and inferred their associated regulators using probabilistic graph models (Segal et al., 2003).

In this paper, we present a new computational method that reconstructs GRN from gene expression data incorporating the aforementioned biological nature of TFs. We detected cooperating TFs that coordinate common biological functions utilizing public protein–protein interaction (PPI) network. For detection of combinatorial relations of TFs and TGs specific to the dataset, i.e., condition-specific combinatorial relations, we utilized kernel canonical correlation analysis (kernel CCA). Kernel CCA is a well-established statistical method for learning coefficients of two groups of features that maximize the correlation of a group of features vs. another group of features (Kuss and Graepel, 2003; Akaho, 2006; Rhee et al., 2009; Ashad Alam and Fukumizu, 2015; Richfield et al., 2016; Tang et al., 2019). A high value of coefficients or weights of features implies that the features from different groups are relevant. For example, applying kernel CCA in motif data and gene expression data, features (e.g., motif) with high weights are deduced as relevant motifs in regulating gene expression (Rhee et al., 2009). Therefore, conducting kernel CCA on gene expression data consisting of groups of features—one feature set composed of TFs and another feature set composed of TGs—can detect TF–TG regulatory relations. Specifically, we employed kernel CCA to embed TFs and TGs into a new space where the correlation of TFs and TGs are reflected to detect context-specific, i.e., response to external stimulus, TF–TG relations. This enables the construction of GRN that models responses to stimuli shows dynamics of GRN over time, applying our method in time-series data. Since we utilized PPI network to detect co-working TFs, we can modularize a GRN into sub-networks of manageable size, which resulted in the improved interpretability of GRN.

2. METHOD

The method proposed in this paper aimed at constructing condition-specific GRNs considering the cooperative and combinatorial nature of TFs. To detect cooperative TFs that share common biological process, we utilized PPI network as a prior knowledge. Then, to detect combinatorial multiple-to-multiple regulatory relations between TFs and TGs, we utilized kernel CCA in inferring regulatory interactions. Our approach uses gene expression profile data in multiple conditions (e.g., time points) as input and produces a network of gene–gene regulatory relations. Public PPI network and GRN network were utilized as a prior knowledge to guide the detection of correct TF–TG relations. Specifically, our approach consists of three steps—**Step 1:** Identification of TFs and TGs modules. **Step 2:** Construction of regulator relationships among the TF/TG modules. **Step 3:** Inference of condition-specific GRN—as described in **Figure 1**.

2.1. STEP 1: Identification of TF and TG Modules

Since TFs work as a protein complex or as a group to direct common biological functions (Sutherland and Bickmore, 2009; Rieder et al., 2012; Duren et al., 2019), we aimed at identifying a group of TFs that work together and TGs that are regulated by the TFs. Genes were classified as TFs referring to the public TF catalogs (Jin et al., 2016; Lambert et al., 2018), otherwise as TGs. We used PPI network—STRING (v10.5) (Szklarczyk et al., 2016) and BioGrid (v.3.5.179) (Stark et al., 2006) database—as putative interactions of genes. STRING database compiled interaction based on experimental data or from the literature. Some interactions in STRING are made by using computational prediction methods, which may contain many false-positive interactions. On the other hand, BioGrid primarily compiled experimentally validated interactions. Thus, interactions in BioGrid may be more reliable but inference using BioGrid may suffer a high level of false-negatives. We concatenated both of the databases to complement each other's limitations. Then, we filtered the network with TFs to build TF–TF interaction network and with TGs to build non-TF–non-TF interaction network (i.e., TG–TG interaction network). In our study, these two networks are used as template networks of co-working or interacting genes. To detect condition-specific network of TFs and TGs for a given context, we instantiated the TF–TF interaction network with expression data of TFs and the TG–TG interaction network with expression data of non-TFs (Ahn et al., 2017). In particular, among gene–gene interactions in template networks, interactions whose Pearson's correlation coefficient between expression vector of corresponding genes below 0.5 are discarded. Using condition-specific networks, respectively, we detected clusters of TFs ad TGs with a multi-level community detection algorithm to detect condition-specific TFs and TG modules. We utilized `multilevel.community` function in R `igraph` package that implemented the Louvain algorithm for community detection (Csardi and Nepusz, 2006).

2.2. STEP 2: Construction of Preliminary GRN Between TF and TG Modules

A very large search space of TF–TG relationships is one of the challenges in reverse engineering of GRN. Given n genes, n^2 combinations of interactions should be considered. In particular, it is not computationally feasible to perform kernel CCA analysis on a very large network. Even if it is feasible, no computational methods can produce correct results when there are many unknown factors, true relations in this case. To reduce search space, we used publicly reported gene regulatory relationships as a guide to navigate TF–TG relationships. Specifically, we merged public GRNs: TRRUST (Han et al., 2018) and HTRIdb (Bovolenta et al., 2012), computationally predicted TF-DNA-binding sites data (Ernst et al., 2010) for Human dataset; PlantRegMap (Tian et al., 2020) and ATRM (Jin et al., 2015) for Arabidopsis dataset. Then, we pruned the network with genes with signature genes—for example, differentially expressed genes (DEGs) or genes with high variance across samples—to navigate the GRN in condition-specific perspectives. A subgraph of GRN that contained signature genes and their first nearest neighbors in public GRNs is utilized as condition-specific gene regulation candidates. For every combination of TF modules and TG modules, projection from a TF cluster to a TG cluster through all shortest paths in the GRN yields a sub-network of GRN and we utilized these edges from the sub-network as a TF–TG relationship candidate.

2.3. STEP 3: Inference of Condition-Specific GRN With Kernel CCA

For each of preliminary sub-network of GRN determined in section 2.2, our goal is to construct condition-specific sub-networks considering multiple-to-multiple relationships of TFs and TGs. Specifically, we utilized kernel CCA to embed TFs and TGs in canonical dimensions. Then, we measured cosine similarity between TFs and TGs in the embedding space to discover TF–TG pairs that contribute to the correlation between the groups of TFs and TGs. Since TFs can also regulate expression of other TFs, which in turn generate TF cascading network, we iteratively conducted kernel CCA embedding and TF–TG relation detection for every possible relationship in each GRN sub-network.

2.3.1. Kernel Canonical Correlation Analysis

A common biological phenomenon shared by groups of genes tends to yield a high correlation detected between expression vectors of the genes (Yamanishi et al., 2003; Rhee et al., 2009). CCA is a method to detect shared correlation across variables from heterogeneous datasets and yield canonical vectors, which are weight coefficients for linear combination of variables in each dataset. These canonical vectors represent how much contribution or weights each variable has in correlation. Kernel CCA is a generalized version of CCA that can detect non-linear relationships between variables. Therefore, we utilized regularized kernel CCA (Bilenko and Gallant, 2016) to retrieve

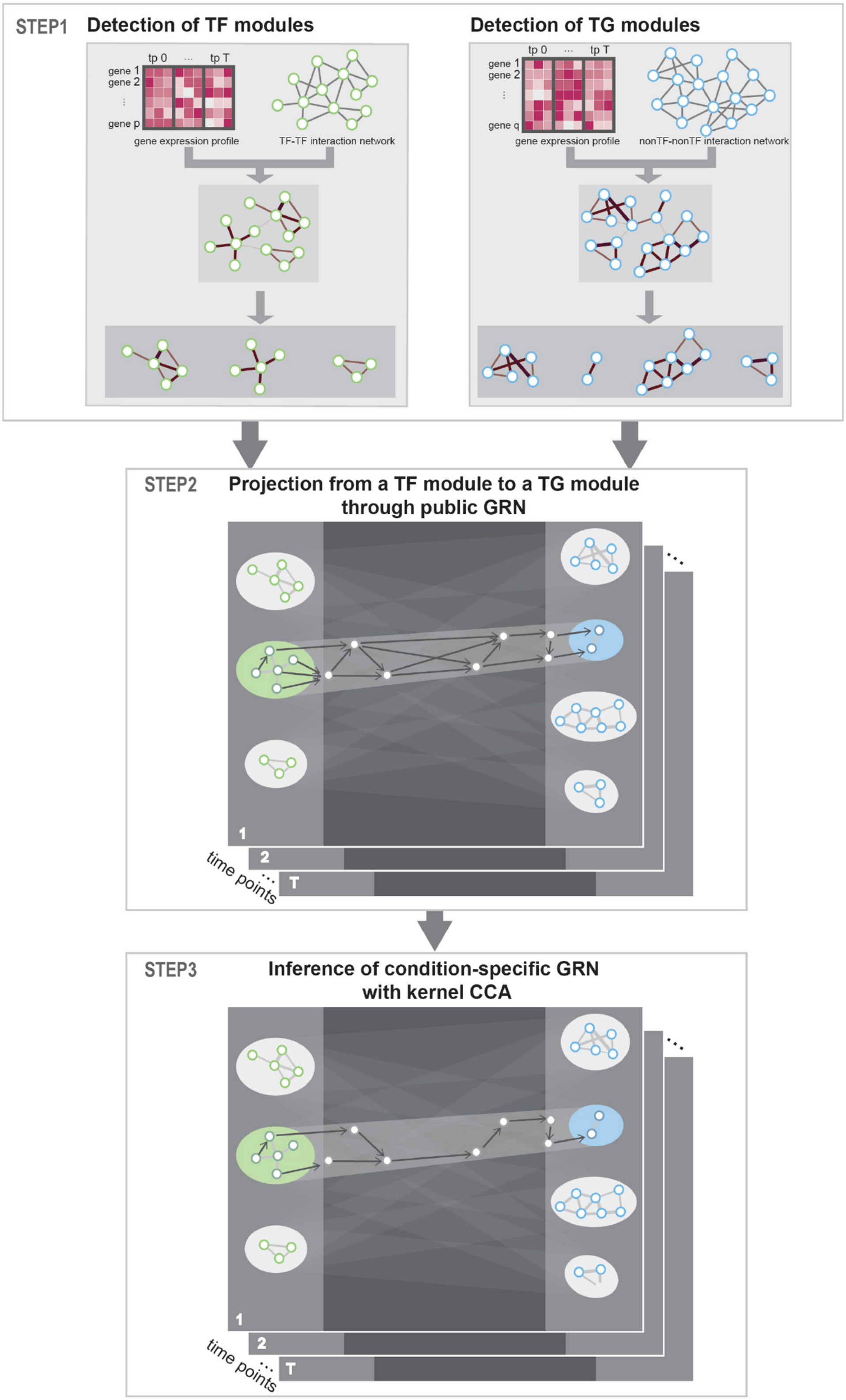

STEP1
Detection of TF modules
tp 0
...
tp T
gene 1
gene 2
gene p
gene expression profile
TF-TF interaction network
Detection of TG modules
gene q
nonTF-nonTF interaction network
STEP2
Projection from a TF module to a TG module through public GRN
1
2
T
time points
STEP3
Inference of condition-specific GRN with kernel CCA

FIGURE 1 | Workflow. STEP 1: To detect interacting transcription factor (TF) and target gene (TG) modules, respectively, prior protein–protein interaction (PPI) network was instantiated with gene expression data and community detection algorithm was used to detect condition-specific TF and TG modules. STEP 2: To get putative TF–TG relations, we conducted projection from a TF module to a TG module through public gene regulatory network (GRN). This process is conducted for every possible TF–TG module pair. STEP 3: Utilizing kernel canonical correlation analysis (CCA), we constructed condition-specific GRN that detects multiple-to-multiple regulatory relationships between TFs and TGs.

new embedding of TFs and TGs that reflects contribution of genes in correlation between expression level of TFs and TGs. Highly scored TFs and TGs in canonical vectors are considered as genes that contribute to correlation of shared biological phenomenon between TFs and TGs.

Let $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) \in \mathbb{R}^{n\times p}$ and $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n) \in \mathbb{R}^{n\times q}$ be the gene expression matrices of TFs and TGs with n samples and p genes and with n samples and q genes, respectively. The original gene expression profiles are mapped to high-dimensional feature space, reproducing kernel Hilbert space (RKHS), through feature maps $\phi_x : \mathbf{x} \in \mathbb{R}^p \mapsto \mathcal{H}_x$ and $\phi_y : \mathbf{y} \in \mathbb{R}^q \mapsto \mathcal{H}_y$. Feature vector $\phi_x(\mathbf{x})$ is the projection of a data point $\mathbf{x} \in \mathbf{X}$ and likewise $\phi_x(\mathbf{x})$ is the projection of a data point $\mathbf{y} \in \mathbf{Y}$. We represent the datasets projected in feature space as $\Phi_x = \big(\phi_x(\mathbf{x}_1), \phi_x(\mathbf{x}_2), \cdots, \phi_x(\mathbf{x}_n)\big)$ and $\Phi_y = \big(\phi_y(\mathbf{y}_1), \phi_y(\mathbf{y}_2), \cdots, \phi_y(\mathbf{y}_n)\big)$, respectively. Applying kernel trick, the similarities of feature vectors can be defined as a positive definite kernel $k_x(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi_x(\mathbf{x}_i), \phi_x(\mathbf{x}_j)\rangle_{\mathcal{H}_x}$ and $k_y(\mathbf{y}_i, \mathbf{y}_j) = \langle \phi_y(\mathbf{y}_i), \phi_y(\mathbf{y}_j)\rangle_{\mathcal{H}_y}$, where $i, j = 1, 2, \ldots, n$. Specifically, we applied Gaussian RBF kernel (Equation 1)

$$\begin{aligned} k_x(\mathbf{x}_i, \mathbf{x}_j) &= \exp\Big[-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\Big] \\ k_y(\mathbf{y}_i, \mathbf{y}_j) &= \exp\Big[-\frac{\|\mathbf{y}_i - \mathbf{y}_j\|^2}{2\sigma^2}\Big] \end{aligned} \tag{1}$$

We define kernel projection of data or kernel Gram matrices as $\mathbf{K}_x = (k_x(\mathbf{x}_i, \mathbf{x}_j))^n_{i,j=1} = \Phi_x^T \Phi_x$ and $\mathbf{K}_y = (k_y(\mathbf{y}_i, \mathbf{y}_j))^n_{i,j=1} = \Phi_y^T \Phi_y$.

The aim of kernel CCA is to find projection vectors f_x and f_y that maximize the correlation of canonical components $\mathbf{u} = \langle f_x, \phi_x(\mathbf{x})\rangle_{\mathcal{H}_x}$ and $\mathbf{v} = \langle f_y, \phi_y(\mathbf{y})\rangle_{\mathcal{H}_y}$. Since canonical vectors f_x and f_y lie in space spanned by the feature space mapped objects, we can represent canonical vectors as linear combinations of Φ_x and Φ_y, where $f_x = \Phi_x^T \boldsymbol{\alpha}$ and $f_y = \Phi_y^T \boldsymbol{\beta}$. Therefore, canonical components u and v are represented with kernel matrix, $u = \Phi_x^T \Phi_x \boldsymbol{\alpha} = \mathbf{K}_x \boldsymbol{\alpha}$ and $v = \Phi_y^T \Phi_y \boldsymbol{\beta} = \mathbf{K}_y \boldsymbol{\beta}$. The objective function of the kernel CCA is restated with kernel projections as follows:

$$\operatorname*{argmax}_{\boldsymbol{\alpha},\boldsymbol{\beta}} \operatorname{corr}(\boldsymbol{u}, \boldsymbol{v}) = \operatorname*{argmax}_{\boldsymbol{\alpha},\boldsymbol{\beta}} \boldsymbol{\alpha}\ \mathbf{K}_x\ \mathbf{K}_y\ \boldsymbol{\beta} \tag{2}$$

where $\alpha \in \mathbb{R}^n, \beta \in \mathbb{R}^n$ are expansion coefficients. The problem can be reformulated as a generalized eigenvalue problem with regularization as follows:

$$\begin{pmatrix} 0 & \mathbf{K}_x\mathbf{K}_y \\ \mathbf{K}_y\mathbf{K}_x & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \rho^2 \begin{pmatrix} \mathbf{K}_x^2 + \lambda I & 0 \\ 0 & \mathbf{K}_y^2 + \lambda I \end{pmatrix} \tag{3}$$

where $\mathbf{I}$ denotes the identity matrix, λ is regularization parameter, and $\rho = \max\langle u, v\rangle / (\|u\| \|v\|)$. Once we obtain solutions for the above equations that represent the amount of contribution of each sample, we multiplied the transpose of gene expression matrices $\mathbf{X}^T \in \mathbb{R}^{p\times n}$ and $\mathbf{Y}^T \in \mathbb{R}^{q\times n}$ with canonical weight vectors $\alpha \in \mathbb{R}^n$ and $\beta \in \mathbb{R}^n$ to get the TF and TG embeddings, $w_x \in \mathbb{R}^p$ and $w_y \in \mathbb{R}^q$ that represents the amount of contribution of each gene (Equation 4).

$$\begin{aligned} w_x &= \mathbf{X}^T\ \alpha \\ w_y &= \mathbf{Y}^T\ \beta \end{aligned} \tag{4}$$

We can now compute k canonical components orthogonal to each other, so that we can get TF and TG embeddings matrix $\mathbf{W}_x \in \mathbb{R}^{p\times k}$ and $W_y \in \mathbb{R}^{q\times k}$ where each row in matrix stands for new embeddings of TGs and TGs in k canonical dimensions.

2.3.2. Detection of Multiple-To-Multiple Relations of TFs and TGs

Using kernel CCA, genes that greatly contribute to the correlations of TFs and TGs gain greater weights in canonical embeddings and TF–TG pair that both TF and TG show high weights should be remarked as valid pair. Inspired by Seo and Kim (2013), we weighted every k dimension with the corresponding eigenvalue so that the eigenvalue-weighted embeddings is dominated by the leading eigenvectors. For every possible TF–TG pair retrieved from public GRN, we next computed dot-product similarity of TF and TG embeddings to define an edge weight of the pair. We then filtered out edges that have weights below 0.5. This process is iteratively performed until there are no TFs left in candidate TG lists.

3. DATA AND PERFORMANCE EVALUATION SCHEME

3.1. Data

We analyzed public time-series gene expression data from NCBI GEO datasets (GSE127530, GSE5621, and GSE5628).

- GSE127530 is an RNA-seq data that measure human blood transcriptome after high-fat meal (HFM) measured in three time points (Fast, +3, and +6 h after stimulus) with 15 samples for each time point, where each time point denoted as tp0, tp1, and tp3. Raw counts are normalized in terms of gene length with TPM (transcripts per million). For our method, we applied MinMaxscaler in Python sklearn library in order not to make correlations dominated by highly expressed genes
- GSE5621 is an microarray data that measures transcriptome from shoots in *Arabidopsis thaliana* in response to cold stress at seven time points (0, +0.5, +1, +3, +6, +12, and +24 h) with two replicates for each time point, where each time point denoted as tp0, tp1, tp2, tp3, tp4, tp5, and tp6. GSE5628 is an microarray data responsive to heat stress, which consists of heat-shocked samples at 38 Centigrade and recovered samples after heat-shock treatment prolongs to 21 h at 25 Centigrade measured at five time points (0, +0.25, +0.5, +1, and +3 h) with two replicates for each time point, where each time point denoted as tp0, tp1, tp2, tp3, and tp4. We applied MinMaxscaler in Python sklearn library for normalization.

3.2. Evaluation 1: Performance Comparison With Existing Methods

We compared our method with the existing methods: ARACNe-AP (Lachmann et al., 2016) and GENIE3 (Irrthum et al., 2010). ARACNe-AP is a representative reverse engineering method based on information theoretic approach for GRN construction while GINIE3 uses a regression tree method. We then compared how much condition-specific signature each method can capture, utilizing GSE127530 dataset. ARACNe-AP does not yield valid edges from the datasets, thus GRN constructed with GSE5621 and GSE5628 datasets were excluded.

- **Construction of Ground Truth GRN**: To assess the network inference performance, we constructed condition-specific GRN as a ground truth gene set. A comprehensive biomedical entity search tool, BEST (Lee et al., 2016), was utilized to retrieve condition-specific gene sets using four keywords from literature search related to HFM: "lipid metabolism (Ming et al., 2009)," "obesity (Golay and Bobbioni, 1997)," "diabetes (Salmeron et al., 2001; Marshall and Bessesen, 2002)," and "innate immunity (McLaughlin et al., 2017; Childs et al., 2019)." Among the four keywords, "innate immunity" is the term reported as a related biological term in the paper that reported the GSE127530 dataset (Lemay et al., 2019). The literature search identified 1,131 HFM-associated genes. These genes were mapped according to the public GRN described in section 2.2. As a result, we constructed a ground truth network of 738 nodes and 1,991 edges that connect 2 HFM-associated genes.
- **Metrics for Performance Measurements**: Given the nodes and edges of an inferred GRN by our method, we measured the overlap of the nodes and edges between the inferred GRN and the ground truth GRN.

  - specificity = TN/(TN+FP)
  - precision = TP/(TP+FP)
  - recall = TP/(TP+FN)

  For the node-level comparison, we measured specificity and recall. True-positive (TP) are a set of genes that are both in the ground truth network and reported by our method. False-positive (FP) is a set of genes that are reported by our method but do not exist in the ground truth network. True-negative (TN) is a set of genes that are not in the ground truth network and not reported by our method. False negative (FN) is a set of genes that are not reported by our method but exist in the ground truth network. For the edge-level comparison, we measured precision and recall. TP are a set of edges that are both in the ground truth network and reported by our method. FP is a set of edges that are reported by our method but do not exist in the ground truth network. FN is a set of edges that are not reported by our method but exist in the ground truth network.

3.3. Evaluation 2: Investigation of TF Cooperation

A sub-network constructed by our method contains multiple TFs that cooperate with each other for regulating TGs in the sub-network. One way to evaluate the power of TF cooperation is to compare metrics from all TFs in the sub-network vs. metrics from a set of individual TFs in the sub-network. That is, we constructed sub-networks using individual TFs in TF modules without considering the cooperativeness of TFs. The original sub-network (denoted as G_{all}) that was constructed using all cooperating TFs in TF modules was compared to the sub-networks (each sub-network denoted as G_i) that was constructed using individual TFs in TF modules. We used two metrics for the evaluation of TF cooperation: the biological significance and the cooperative potential.

3.3.1. Biological Significance

Biological significance (B_p) of TF cooperation in terms of each pathway was calculated using Equation (5). Pathway enrichment with nodes in G_{all} and all $G_i's$ were was calculated using Enrichr (FDR < 0.05) (Chen et al., 2013) in `gseapy` library. For each pathway p, the p-value obtained from G_{all} is denoted as p_a^p and the p-value obtained from G_i is denoted as p_i^p. Since multiple G_is are constructed, aggregating pathway p-values from G_i was performed by Fisher's combined probability test (Fisher, 1992). Specifically, a set of p-values from k independent tests to calculate a test statistic $\chi_F^2 = -2\sum_{i=1}^{k}[\ln[p_i^p]$ that follows χ^2 distribution with $2k$ degrees of freedom under the null hypotheses of the k tests. The p-value combined with the Fisher's combined probability test as denoted as p_c^p.

For each pathway p, R_p value was calculated to compare the relative significance of G_a and G_is dividing p_a^p with p_c^p).

$$B_p = log_2\left[\frac{log_{10}(p_a^P)}{log_{10}(p_c^P))}\right] \tag{5}$$

3.3.2. Cooperative Potential

The cooperative property of TFs was measured by comparing network centrality values between G_{all} and G_is (Equation 6). We used *betweenness centrality* of a node in a given sub-network that measures the proportion of the shortest paths present in the sub-network that pass through the corresponding node. Gene-level network centrality values were calculated on the G_{all} and the set of G_is, which are denoted as c_{all}^g and c_i^gs. Then, the centrality value of the G_{all} (c_{all}^g) was divided by the square-rooted squared sum of c_i^gs. The cooperative potential of a pathway (C_p) was calculated by summing up the cooperative potential of the overlap genes.

$$C_p = \sum_{g \in P} \log_2\left[\frac{c_{all}^g}{\sqrt{\sum_{i=1}^{k}(c_i^g)^2)}}\right] \tag{6}$$

3.4. Evaluation 3: Dynamics of GRNs Across Time

One of the advantages of our method is that the whole GRN is divided into small sub-networks. We suggest two approaches to choose sub-networks for detailed inspection.

- To emphasize on the dynamics of network over time, we chose sub-networks where regulatory relations vary significantly over time. For assessing the amount of variance across time, we measured the fraction of time-point exclusive nodes and edges to the size of a sub-network for each time point and then averaged across time. We applied this approach to the human dataset.
- To investigate how combinations of co-working TFs vary over time, we chose a Differentially Expressed Gene (DEG)-enriched TG module and inspected the DEG-enriched sub-networks connected to the TG module. We applied this approach to the *Arabidopsis thaliana* datasets. There were too

TABLE 1 | Comparison of our method to ARACNe-AP and GENIE3 in terms of specificity, precision, and recall with respect to the ground truth network from a literature search tool, BEST (Lee et al., 2016).

			ARACNe-AP	GENIE3	Linear CCA	Kernel CCA
+3 h	Node comparison	Specificity	0.841	0.270	0.692	0.961
		Recall	0.230	0.829	0.533	0.483
	Edge comparison	Precision	0	8.05×10^{-6}	9.01×10^{-3}	3.12×10^{-2}
		Recall	0	9.04×10^{-3}	0.413	0.591
+6 h	Node comparison	Specificity	0.869	0.277	0.741	0.957
		Recall	0.197	0.830	0.451	0.389
	Edge comparison	Precision	6.30×10^{-6}	8.20×10^{-6}	6.51×10^{-4}	2.89×10^{-2}
		Recall	8.84×10^{-4}	9.04×10^{-3}	0.188	0.340

many genes in the *Arabidopsis thaliana* network, thus we used only DEGs to reduce the number of genes.

4. RESULTS

Given gene expression profiles, our method produces GRN that consists of multiple sub-networks where condition-specific interacting TFs regulate a set of TGs through intermediate genes. Utilizing the public GRN and signature genes as a guide, our method selects edges of the network with kernel CCA to model cooperative and combinatorial natures TFs and TGS. Another strength is that our method decomposes the whole GRN in to sub-networks to improve interpretability. When an organism is exposed to an environmental stimulus, it orchestrates multiple biological process as a response and what our method determines is the intermingled regulatory interactions. Therefore, decomposition of the whole GRN into sub-networks helps us to interpret the result better.

4.1. Comparative Analysis

We compared our method with the existing methods: ARACNe-AP (Lachmann et al., 2016) and GENIE3 (Irrthum et al., 2010). We compared how well each method can capture condition-specific network using GSE127530 dataset. Our method produced a set of TF–TG modules, i.e., a set of sub-networks, but existing methods produced a single network of large size. To compare the results, we combined a set of sub-networks from our method into a large single network. GENIE3 produced a set of million edges with importance score, and top 0.5% edges in terms of importance score were used for comparative analyses. To assess the performance of network inference, we retrieved 1131 HFM-related gene sets using a comprehensive biomedical entity search tool, BEST (Lee et al., 2016), as a condition-specific gene set (see section 3.2 for details). Both specificity and recall were used as metrics to compare the three methods for quantitative evaluation (**Table 1**). In node-level comparison, our method showed the best performance in terms of specificity and the second best in terms of recall in all time points. In edge-level comparison, our method showed the best performance in terms of both precision and recall in all time points. Additionally, in order to demonstrate that the non-linear technique for the construction of canonical components is necessary, we compared the performance of network inference by the regularized linear CCA with the performance of the regularized non-linear kernel CCA. In a majority of cases of performance comparisons, except recall of node comparison, utilizing kernel CCA exceeds in inferring the ground truth network.

4.2. Case Study 1: GRN in Response to HFM in Human

4.2.1. Dynamics of GRN Over Time in Response to HFM

Dynamics of GRN over time in response to HFM was investigated. We executed our method on GSE127530 dataset obtaining a GRN for each time point (tp1 and tp2) with respect to tp0 as a baseline; a GRN with 7,021 nodes and 99,455 edges in tp1, and with 5,985 nodes and 61,646 edges in tp2. One challenge that arises in inspection of GRN is that regulatory relations are too complex to interpret in which multiple biological processes are intermingled together. One of the strengths of our method is that we can decompose the giant network into a feasible size of sub-network consisting of GRN projection from a TF module to a TG module. The resulting GRN from our method consisted of 31 TF modules and 76 TG modules in tp1 and 26 TF modules and 52 TG module in tp2 which means that 31 × 76 sub-networks and 26 × 52 sub-networks consists of a GRN of each time point.

To investigate the regulatory mechanism over time, a network dynamics score of a TF–TG sub-network between two adjacent time points was measured. Basically, the score represents an average proportion of exclusiveness of genes at each time point. Detailed description of the score is given in section 3.4. With the score, we now can sort out TF–TG networks that show bigger change in network dynamics over time. By sorting TF–TG networks in terms of the score, we selected top 100 TF–TG networks. Each TF–TG network is a pair of a TF module and a TG module. Interestingly, many sub-networks shared common TF modules. Among 100 TF–TG networks, i.e., 100 pairs of a TF module and a TG module, 95 pairs of TF and TG modules share a TF module. With this observation, we can merge multiple TF–TG networks into single sub-networks. One sub-network that include 17 TG modules was used to investigate network dynamics over time after HFM—denoted as G_{3h} for tp1 and G_{6h} for tp2. We then compared how much biological pathways were enriched in these networks over time (**Figure 2**). As a result, immune system related pathways—Th17 differentiation, Th1 and Th2 cell differentiation, and inflammatory bowel disease (IBD)—were high ranked both in G_{3h} and G_{6h}. Specifically, AGE-RAGE signaling pathway in diabetic complications was enriched in both time points. Advanced glycation end products (AGEs) and their receptor, RAGE, are known to deal with the accumulation of metabolite end product in diabetes (Ramasamy et al., 2011). The amount of soluble RAGE is also reported to play an important role in post-prandial response to HFM (Fuller et al., 2018).

Here are detailed discussions on dynamics of a TF–TG sub-network with the highest dynamics score (**Figure 2**). FOXO3 and FOXO4, which are the interacting TFs and are at the top hierarchy in TF cascading network, are isoforms of well-known nuclear TFs—FOXO family—that are involved in metabolic regulation (Barthel et al., 2005) and promoting inflammatory response in T cell (Kerdiles et al., 2010; Hedrick et al., 2012) implying the regulatory link between immune response and metabolic process. After 3 h after HFM, tumor necrosis factor α (TNF-α) and interleukin-6 (IL-6) are pro-inflammatory cytokines whose concentration reaches peak around 2–3 h after HFM (Herieka and Erridge, 2014). One of the TGs in the sub-network, S1P phosphatase 2 (SPP2) is known to play a pro-inflammatory role in induction of TNF-α and IL-6 (Mechtcheriakova et al., 2007). A differentially expressed gene, ETS1, encodes a TF involved in production of cytokine and chemokine in T helper cells (Russell and Garrett-Sinha, 2010; Garrett-Sinha, 2013) where one of the early responses of HFM is pro-inflammatory cytokine production. GATA3 is a family of GATA TF family that is an important regulator of T-cell development. According to Ibarra et al. (2020), FOXO1-ETS1 is reported as a potential cooperative TFs. FOXO1 and FOXO3 are the most dominant isotypes of Forkhead box family TF that coordinate common biological function—regulatory T cell development (Ohkura and Sakaguchi, 2010), implying that cooperative potential of FOXO regulation with ETS1 genes which is detected in our network. After 6 h after HFM, TF–TG relations that are regulating SPP2—one of the acute post-prandial responses—is diminished in the sub-network. However, other immune-responsive genes (i.e., POU2F1, RUNX1, NFKB1, and LEF1) are still enriched that are promoting other immune responses.

4.2.2. Investigation of TF Cooperation in HFM

We next investigated how much cooperation occurs in sub-networks (**Figure 3**). To demonstrate this, we analyzed the level of disruption if a single TF were considered—there are n simulations for n TFs in a given sub-network. To demonstrate this, we analyzed the level of disruption in pathways comparing sub-networks using multiple TFs (denoted as G'_{all}) vs. simulated networks using individual TFs (denoted as G'_i). The level of cooperation was measured at two perspectives: biological significance (B_p) and cooperative potential (C_p) between the G'_{all} and G'_i.

The greater the R_P value is in a certain pathway p, the more genes exist in the p utilizing multiple TFs compared to the simulation with individual TFs. Heatmap in the left panel of **Figure 3** depicts B_p value of the enriched pathways in G'_{all}. Pathways including inflammatory bowel disease, hepatitits B, and estrogen signaling pathways showed greater TF cooperation at tp1. While at tp2, FOXO signaling pathway showed greater B_p at tp2 compared to that of the previous time point despite most of the pathways showed subtle enrichment changes against simulations. Such temporal changes indicate that there are regulatory dynamics in multiple pathways co-regulated by multiple TFs. C_p value in Equation (6) was developed here to investigate the degree of TF cooperation at G'_{all} in comparison to G'_i by summing up the individual contribution to cooperative potential of the genes in a sub-network. The greater betweenness centrality of a node is, the more shortest paths go through the node. As the whole network topology is more likely to be disrupted, the genes with high centrality are removed, and the node would play an important role in maintaining the given network topology.

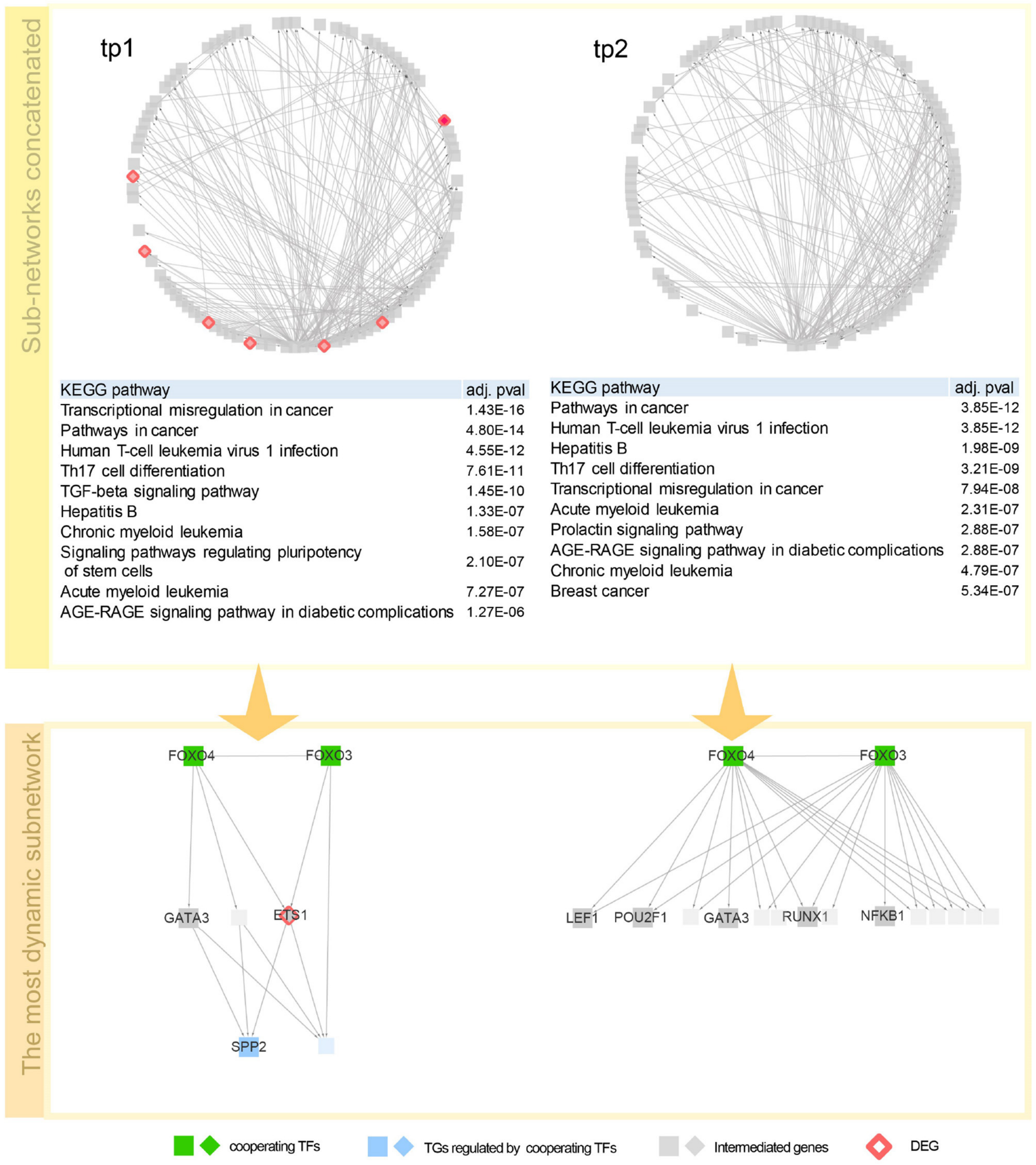
C_p showed that tighter TF regulations were made by multiple TFs compared to the simulations throughout the enriched pathways. Compared to the subtle changes in B_p value, the ability of kernelCCA to construct a sensitive regulatory sub-network to temporal dynamics reflected greater C_p value across the pathways. Specifically, pathways related to cellular signaling were consistently co-regulated by two TFs (FOXO3 and FOXO4). This was also supported by a previous study that suggests greater co-regulation by multiple TFs stay invariant to perturbations as well as play a central role in controlling pivotal dynamics in response to external stimuli (Kim et al., 2012).

4.3. Case Study 2: GRN in Response to Heat and Cold Stress in *Arabidopsis thaliana*

4.3.1. Dynamics of GRN Over Time in Response to Heat and Cold Stress

To investigate how combinations of co-working TFs vary over time, the sub-networks connected to the most DEG-enriched TG module are scope of our inspection. Since GRN of *Arabidopsis thaliana* is denser and more DEGs are detected than in human dataset, we, therefore, used DEG-centric approach that paths to DEGs from co-working TFs are inspected. All paths detected are listed in **Appendix A**.

It has been a long question how plants detect the lower and higher temperature and how they are sensing differences in the temperature. Usually, plants complete their whole life in one place where they germinated. Their growth undergoes diurnal rhythm and seasonal periodicity, which means the temperature condition is changing all the time. Plants can recognize the small change of temperature, such as 2–3 centigrade, called ambient temperature. The effects of these small changes are cumulative, having retention time to appear certain consequences even though the ranges of results vary depending on the stage of growth, other environmental conditions, and their genetic backgrounds. All of these processes occur in plants started from very minute changes at the molecular level. So, it has been an important task to undercover how plants recognize



KEGG pathway	adj. pval
Transcriptional misregulation in cancer	1.43E-16
Pathways in cancer	4.80E-14
Human T-cell leukemia virus 1 infection	4.55E-12
Th17 cell differentiation	7.61E-11
TGF-beta signaling pathway	1.45E-10
Hepatitis B	1.33E-07
Chronic myeloid leukemia	1.58E-07
Signaling pathways regulating pluripotency of stem cells	2.10E-07
Acute myeloid leukemia	7.27E-07
AGE-RAGE signaling pathway in diabetic complications	1.27E-06

KEGG pathway	adj. pval
Pathways in cancer	3.85E-12
Human T-cell leukemia virus 1 infection	3.85E-12
Hepatitis B	1.98E-09
Th17 cell differentiation	3.21E-09
Transcriptional misregulation in cancer	7.94E-08
Acute myeloid leukemia	2.31E-07
Prolactin signaling pathway	2.88E-07
AGE-RAGE signaling pathway in diabetic complications	2.88E-07
Chronic myeloid leukemia	4.79E-07
Breast cancer	5.34E-07

FIGURE 2 | Network dynamics of a gene regulatory network (GRN) sub-network after high-fat meal (HFM) over time. Two circular diagrams in the upper panel show the change in gene–gene relationship, in particular, TF–TG regulation. Two tables in the middle summarize top 10 most enriched biological pathways with *p*-value corrected by false discovery rate (FDR) < 0.05. Two networks are dynamics of a TF–TG sub-network with the highest dynamics score. In the TF–TG sub-network, two transcription factors (TFs), FOXO3 and FOXO4, regulates different sets of target genes (TGs) over time. TFs were denoted with diamond-shaped nodes. Square nodes denotes TGs and circle nodes denote genes connected to TFs and TGs. Nodes colored pink denote genes that consist of GRN in each time point. Nodes colored red denote shared TFs among concatenated sub-networks and nodes colored green DEGs that are detected by DESeq with FDR < 0.05.

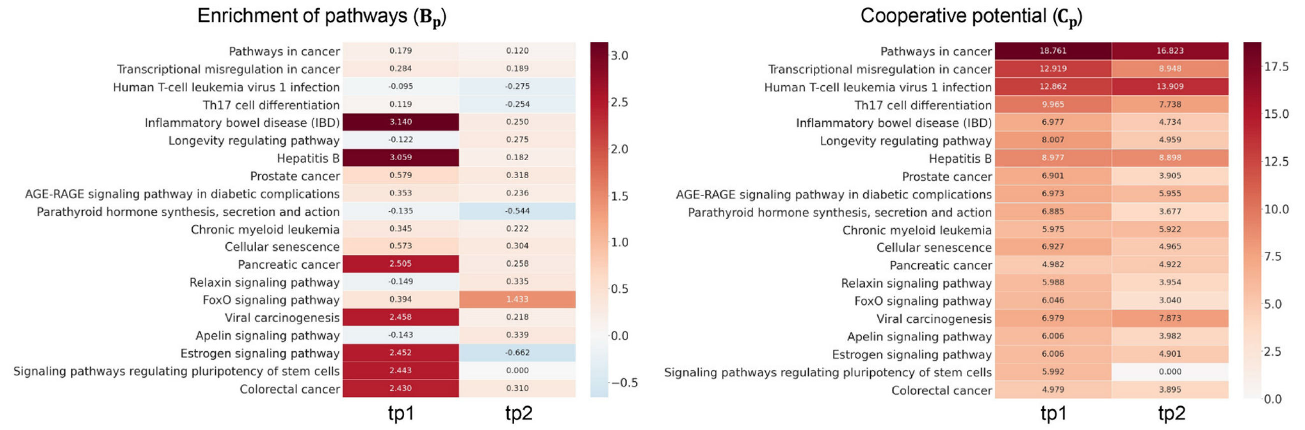


FIGURE 3 | Examination of transcription factor (TF) cooperation. High score of B_p and C_p represents the amount of cooperativity of co-working TFs in the pathways. Heatmap in the left panel shows the cooperation in terms of pathway enrichment over time in high-fat meal (HFM). Pathway enrichment in the G_{all} was compared to the simulations given each TF (G_i) and measured using Equation (5). Heatmap in the right panel shows the cooperative potential using enriched pathway genes. Betweenness centrality was compared between G_{all} and G_i using Equation (6).

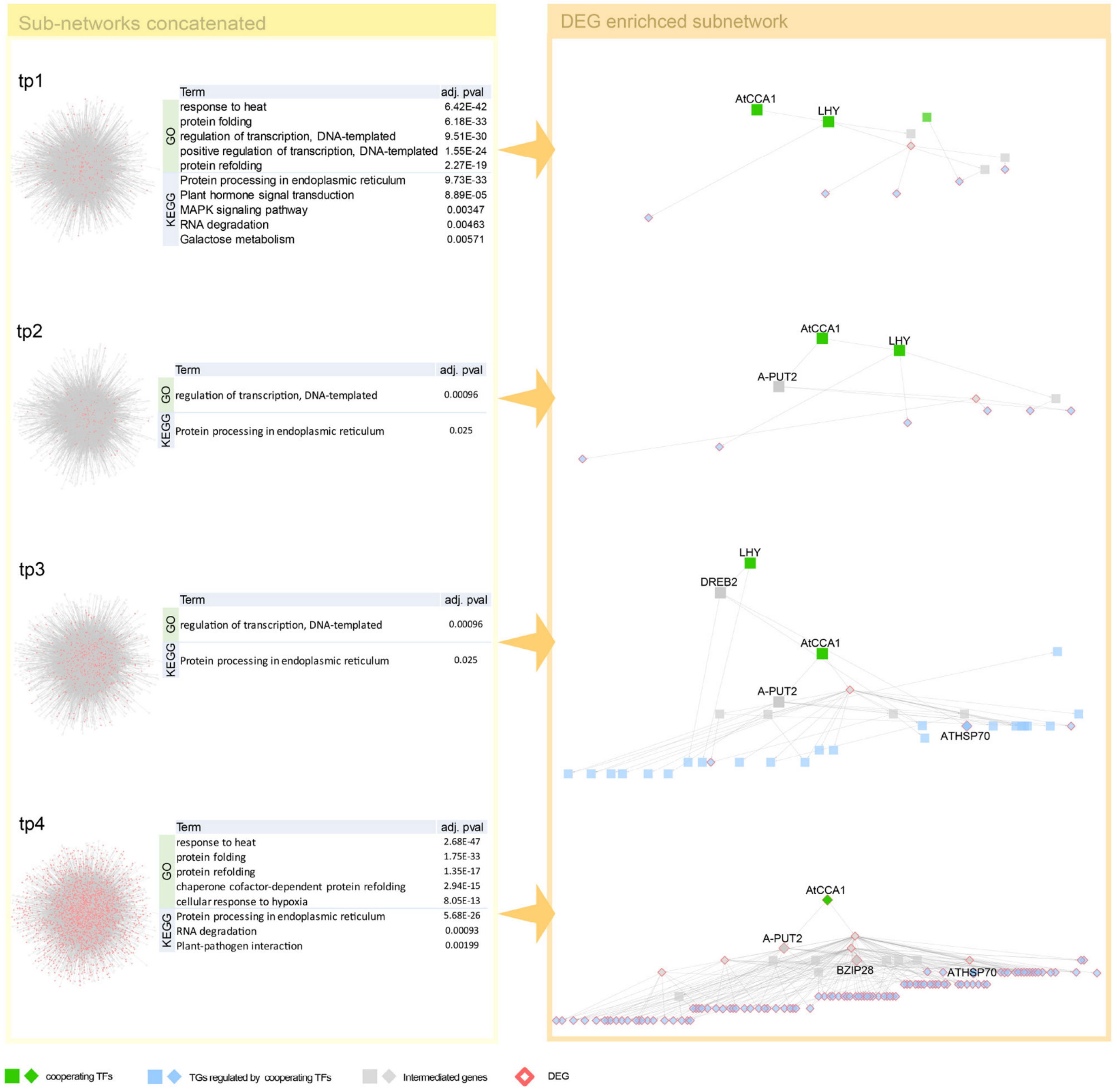
and trigger the serial and reversible and sometimes irreversible responses. Still, it is challenging to find out the group of genes in the thermal physiology of plants. We used two different Arabidopsis datasets, GSE5628 and GSE5621. GSE5628 represents heat stress that consists of heat-shocked samples up to 3 h at 38 centigrade and recovered samples after heat-shock treatment prolongs to 21 h at 25 centigrade. When outranged thermal changes have occurred, all responses of plants go for stabilizing homeostasis.

Interestingly, we detected both genes of circadian clock associated (CCA1) and late elongated hypocotyl (LHY), a short period after high-temperature treatment. These two genes, detected as a co-working TFs in our proposed method (**Figure 4**), involve in common biological pathway—a central role in the phytochrome-medicated circadian clock (Alabadi et al., 2002; Dong et al., 2011). After that, we observed in one of our early-stage tp1 and tp2 of heat-path various TCP genes, PIF5 (PUT2) and CAT genes, those involved in thermosensory (Michael et al., 2003; Zhou et al., 2019; Balcerowicz, 2020). A path of phytochrome-mediated thermo-response appears tp3 stage. Mainly PIF4 and many of its downstream genes include directly related genes, such as TCPs and BZIP28 and indirectly related genes that mediate heat shock responses (Che et al., 2010).

Unlike a higher temperature treatment for several hours that increases physiological reactions and results in less severe consequences, lower temperature treatment over hours is life threatening. This characteristic difference of temperature treatment is why we found a relatively broad range of gene regulatory paths from cold treatment. We found well-defined cold response genes, such as CBF, DREB, COR, ERF, ZAT, RVE, and ABF1 (Vogel et al., 2005; Lee and Thomashow, 2012; Meissner et al., 2013; Wang et al., 2017; Dubois et al., 2018) and many cold stress-related genes from the early stage of cold treatment (**Figure 5**, **Appendix A**). Co-working TFs, such as RVE1, CPD45, and ATCBF2, detected in our GRN are involved in common cold related pathways implying cooperative functions of the TFs (Eremina et al., 2016; Chen et al., 2020). We found CCA1, LHY, and PIF4 gene from DEGs of cold temperature treated samples (**Figure 5**). It might have resulted from the thermosensory networkś change even though the treatmentś degree was far beyond the ambient temperature to the lower direction. It might be noteworthy that we observed the genes of developmental processes like RVE and cold acclimation related COR and CBF. There are several reports on CBF gene regulation. We found most of CBF promoter binding TFs, such as PIFs, CCA1, and LHY (Dong et al., 2011; Jiang et al., 2017). Several genes reported as intermediate genes—connecting the co-working TFs and the DEGs regulated by the TFs—are involved in common cold responsive pathways, implying that cooperative action of regulating downstream DEGs (**Appendix A**).

4.4. Discussion and Conclusion

In this paper, we proposed a kernel CCA based condition-specific GRN inference method that models combinatorial and cooperative nature of TF–TG relations. The traditional approach is to start with the whole network and test validity of edges, which lead to a condition-specific network based on gene expression data. One major issue with this approach is to deal with a single large network as a whole, which is challenging. However, we know that each TF regulates a relatively small number of genes, typically several hundred genes. So, it is possible to limit the scope of TGs that are regulated by a single TF. In fact, experimental techniques, such as TF ChIP-seq provide condition-specific comprehensive snapshot of genes that are targeted by a TF. Although these experimental data provides condition-specific targets of a TF, there are two major issues for utilizing such TF ChIP-seq data. First, a TF ChIP-seq experiment provides TGs of the TF only. Since TF may target different genes under different conditions, reconstruction of condition-specific networks requires TF ChIP-seq experiments for "all" relevant TFs, which is infeasible due to the time and budget constraints.


Sub-networks concatenated
DEG enrichched subnetwork
tp1
tp2
tp3
tp4
AtCCA1
LHY
A-PUT2
DREB2
ATHSP70
BZIP28
cooperating TFs
TGs regulated by cooperating TFs
Intermediated genes
DEG

tp1

	Term	adj. pval
GO	response to heat	6.42E-42
	protein folding	6.18E-33
	regulation of transcription, DNA-templated	9.51E-30
	positive regulation of transcription, DNA-templated	1.55E-24
	protein refolding	2.27E-19
KEGG	Protein processing in endoplasmic reticulum	9.73E-33
	Plant hormone signal transduction	8.89E-05
	MAPK signaling pathway	0.00347
	RNA degradation	0.00463
	Galactose metabolism	0.00571

tp2

	Term	adj. pval
GO	regulation of transcription, DNA-templated	0.00096
KEGG	Protein processing in endoplasmic reticulum	0.025

tp3

	Term	adj. pval
GO	regulation of transcription, DNA-templated	0.00096
KEGG	Protein processing in endoplasmic reticulum	0.025

tp4

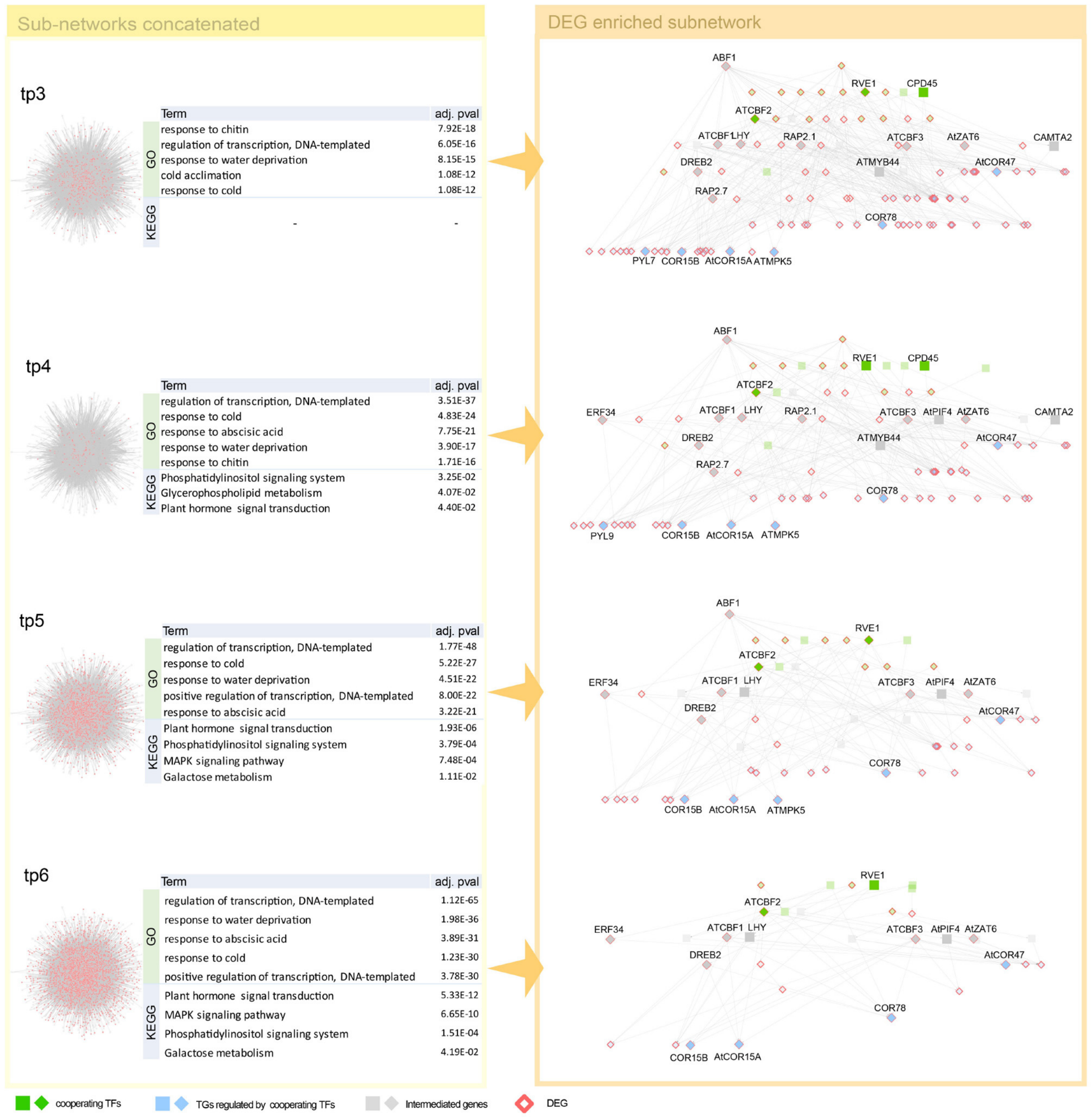
	Term	adj. pval
GO	response to heat	2.68E-47
	protein folding	1.75E-33
	protein refolding	1.35E-17
	chaperone cofactor-dependent protein refolding	2.94E-15
	cellular response to hypoxia	8.05E-13
KEGG	Protein processing in endoplasmic reticulum	5.68E-26
	RNA degradation	0.00093
	Plant-pathogen interaction	0.00199

FIGURE 4 | Network dynamics of a gene regulatory network (GRN) sub-network after heat stress over time. Network in the left panel shows the change in gene-gene relationship, in particular, TF–TG regulation. DEGs are denoted as pink. Four tables in the middle summarize top five most enriched biological pathways with *p*-value corrected by false discovery rate (FDR) < 0.05. Heat stress related Gene Ontology (GO) terms are enriched in GO enrichment tests with DEGs. The networks in the right panel are dynamics of a TF–TG sub-network that are DEG enriched. Transcription factors (TFs) were denoted with green nodes. Blue nodes denote target genes (TGs) and gray nodes denote genes connected to TFs and TGs. Square nodes denote DEGs that are detected by Limma with FDR < 0.05.

Second, even if we can perform such expensive experiments, we need to combine many TF-networks into large networks. One major issue for this task is to identify co-operating TFs in a specific condition, but this is largely unknown.

4.4.1. Advantages and Limitations

The novelty of our approach is to address the two issues in a single computational framework. First, we used clustering approach to reduce the search space by generating a set of TF clusters and a set of TG clusters. This approach allows us handle much smaller networks. Specifically, a TF set vs. a TG set is considered one at a time. Second, use of kernel CCA allows us to investigate on the complex relationship of multiple TFs vs. multiple TGs. In the final step of our computational framework, all TG sets that are related to a single TF set are merged, which generates condition-specific sub-networks. By



tp3

	Term	adj. pval
GO	response to chitin	7.92E-18
	regulation of transcription, DNA-templated	6.05E-16
	response to water deprivation	8.15E-15
	cold acclimation	1.08E-12
	response to cold	1.08E-12
KEGG	-	-

tp4

	Term	adj. pval
GO	regulation of transcription, DNA-templated	3.51E-37
	response to cold	4.83E-24
	response to abscisic acid	7.75E-21
	response to water deprivation	3.90E-17
	response to chitin	1.71E-16
KEGG	Phosphatidylinositol signaling system	3.25E-02
	Glycerophospholipid metabolism	4.07E-02
	Plant hormone signal transduction	4.40E-02

tp5

	Term	adj. pval
GO	regulation of transcription, DNA-templated	1.77E-48
	response to cold	5.22E-27
	response to water deprivation	4.51E-22
	positive regulation of transcription, DNA-templated	8.00E-22
	response to abscisic acid	3.22E-21
KEGG	Plant hormone signal transduction	1.93E-06
	Phosphatidylinositol signaling system	3.79E-04
	MAPK signaling pathway	7.48E-04
	Galactose metabolism	1.11E-02

tp6

	Term	adj. pval
GO	regulation of transcription, DNA-templated	1.12E-65
	response to water deprivation	1.98E-36
	response to abscisic acid	3.89E-31
	response to cold	1.23E-30
	positive regulation of transcription, DNA-templated	3.78E-30
KEGG	Plant hormone signal transduction	5.33E-12
	MAPK signaling pathway	6.65E-10
	Phosphatidylinositol signaling system	1.51E-04
	Galactose metabolism	4.19E-02

FIGURE 5 | Network dynamics of a gene regulatory network (GRN) sub-network after cold stress over time. Networks in the left panel shows the change in gene–gene relationship, in particular, TF–TG regulation. DEGs are denoted as pink. Four tables in the middle summarize top five most enriched biological pathways with *p*-value corrected by false discovery rate (FDR) < 0.05. Cold stress related Gene Ontology (GO) terms are enriched in GO enrichment tests with DEGs. The networks in the right panel are dynamics of a TF–TG sub-network that are DEG enriched. TFs were denoted with green nodes. Blue nodes denote TGs and gray nodes denote genes connected to transcription factors (TFs) and target genes (TGs). Square nodes denote DEGs that are detected by limma with FDR < 0.05. We showed GRN from tp3 to tp6, since GRN constructed in tp1 and tp2 is too small because the number of DEGs are too small in tp1 and tp2—41 and 23, respectively.

performing analysis on transcriptome of human high-fat data and of arabidopsis cold and heat data at each time point, temporal dynamics of TF–TG networks was constructed by explaining condition-specific biological mechanisms successfully.

Although our method was successful in constructing dynamics of condition-specific TF–TG networks over time in both data sets, there are several issues remaining as further study. In the current framework, clustering of TF and TG modules need more rigorous definitions. The size of TF and TG modules vary greatly—some clusters consist of few genes while others consists of hundred genes. Merging TF–TG sub-networks in the final step of our method also need more rigorous guideline. We suggested two approaches in selecting sub-networks that show condition-specific response, which is meaningful, but there is still room for improvement to consider the non-responsive gene regulatory interactions that are required for fundamental cellular functions.

In terms of biological perspectives, there are also several issues that requires further study. First, our method does not discriminate stimulative or repressive gene regulation. Another issue is with kernel CCA. Kernel CCA can detect multiple-to-multiple relations of TFs and TGs, it does not discriminate whether correlations are positive or negative. In addition, our method assume that TF is a major regulator. However, there are other regulatory mechanisms, such as mutations, copy number variations, and epigenetic mechanisms, that can affect transcription level of genes. This requires a comprehensive model, e.g., ensemble of deep learning (Kang et al., 2020). Combining network analysis techniques and deep learning technologies is a major current research topic.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. RNA-seq data of human blood transcriptome analyzed in this study can be found in the GEO (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE127530) and microarray data of Arabidopsis thaliana can be found in the GEO (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5621; https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5628). Code for GRN construction with kernel CCA can be found at: https://github.com/DabinJeong/GRN_construction_with_kernelCCA (Mölder et al., 2021).

AUTHOR CONTRIBUTIONS

SK designed and directed the whole project. DJ designed and implemented the GRN construction algorithm. SLi, MO, and SLe involved in the discussion for building thesis. SLi designed the demonstration strategy and visualized the results. CC conducted the comparison analysis. WJ, SLi, DJ, SLe, MO, and CC biologically interpreted the analysis results. SK, DJ, and SLi wrote and revised the paper. All authors contributed to the article and approved the submitted version.

FUNDING

This research was supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) funded by the Ministry of Science & ICT (NRF-2019M3E5D307337511, NRF-2019M3E5D4065965), the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI) funded by the Ministry of Health & Welfare (HI15C3224), and the Collaborative Genome Program for Fostering New Post-Genome Industry of the National Research Foundation (NRF) funded by the Ministry of Science and ICT (MSIT) (NRF-2014M3C9A3063541).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.652623/full#supplementary-material

REFERENCES

Ahn, H., Jung, I., Shin, S. J., Park, J., Rhee, S., Kim, J. K., et al. (2017). Transcriptional network analysis reveals drought resistance mechanisms of AP2/ERF transgenic rice. *Front. Plant Sci.* 8:1044. doi: 10.3389/fpls.2017.01044

Akaho, S. (2006). A kernel method for canonical correlation analysis. *arXiv* cs/0609071.

Alabadi, D., Yanovsky, M. J., Mas, P., Harmer, S. L., and Kay, S. A. (2002). Critical role for CCA1 and LHY in maintaining circadian rhythmicity in arabidopsis. *Curr. Biol.* 12, 757–761. doi: 10.1016/S0960-9822(02)00815-1

Ashad Alam, M., and Fukumizu, K. (2015). Higher-order regularized kernel canonical correlation analysis. *Int. J. Pattern Recogn. Artif. Intell.* 29:1551005. doi: 10.1142/S0218001415510052

Balcerowicz, M. (2020). Phytochrome-interacting factors at the interface of light and temperature signalling. *Physiol. Plant.* 169, 347–356. doi: 10.1111/ppl.13092

Barthel, A., Schmoll, D., and Unterman, T. G. (2005). FOXO proteins in insulin action and metabolism. *Trends Endocrinol. Metab.* 16, 183–189. doi: 10.1016/j.tem.2005.03.010

Bilenko, N. Y., and Gallant, J. L. (2016). Pyrcca: regularized kernel canonical correlation analysis in python and its applications to neuroimaging. *Front. Neuroinform.* 10:49. doi: 10.3389/fninf.2016.00049

Bovolenta, L. A., Acencio, M. L., and Lemke, N. (2012). HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions. *BMC Genomics* 13:405. doi: 10.1186/1471-2164-13-405

Che, P., Bussell, J. D., Zhou, W., Estavillo, G. M., Pogson, B. J., and Smith, S. M. (2010). Signaling from the endoplasmic reticulum activates brassinosteroid signaling and promotes acclimation to stress in arabidopsis. *Sci. Signal.* 3:ra69. doi: 10.1126/scisignal.2001140

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., et al. (2013). EnrichR: interactive and collaborative html5 gene list enrichment analysis tool. *BMC Bioinformatics* 14:128. doi: 10.1186/1471-2105-14-128

Chen, S., Huang, H. A., Chen, J. H., Fu, C. C., Zhan, P. L., Ke, S. W., et al. (2020). SgRVE6, a LHY-CCA1-like transcription factor from fine-stem stylo, upregulates NB-LRR gene expression and enhances cold tolerance in tobacco. *Front. Plant Sci.* 11:1276. doi: 10.3389/fpls.2020.01276

Childs, C. E., Calder, P. C., and Miles, E. A. (2019). *Diet and Immune Function*. Basel: Nutrients.

Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. *InterJ. Complex Syst.* 1695, 1–9.

Dong, M. A., Farr, E. M., and Thomashow, M. F. (2011). Circadian clock-associated 1 and late elongated hypocotyl regulate expression of the C-repeat binding factor (CBF) pathway in arabidopsis. *Proc. Natl. Acad. Sci. U.S.A.* 108, 7241–7246. doi: 10.1073/pnas.1103741108

Dubois, M., Van den Broeck, L., and Inzé, D. (2018). The pivotal role of ethylene in plant growth. *Trends Plant Sci.* 23, 311–323. doi: 10.1016/j.tplants.2018.01.003

Duren, Z., Wang, Y., Wang, J., Zhao, X. M., Lv, L., Li, X., et al. (2019). Hierarchical graphical model reveals HFR1 bridging circadian rhythm and flower development in arabidopsis thaliana. *NPJ Syst. Biol. Appl.* 5, 1–11. doi: 10.1038/s41540-019-0106-3

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.* 95, 14863–14868. doi: 10.1073/pnas.95.25.14863

Eremina, M., Rozhon, W., and Poppenberger, B. (2016). Hormonal control of cold stress responses in plants. *Cell. Mol. Life Sci.* 73, 797–810. doi: 10.1007/s00018-015-2089-6

Ernst, J., Plasterer, H. L., Simon, I., and Bar-Joseph, Z. (2010). Integrating multiple evidence sources to predict transcription factor binding in the human genome. *Genome Res.* 20, 526–536. doi: 10.1101/gr.096305.109

Fisher, R. A. (1992). "Statistical methods for research workers," in. *Breakthroughs in Statistics* (New York, NY: Springer), 66–70. doi: 10.1007/978-1-4612-4380-9_6

Fuller, K. N., Valentine, R. J., Miranda, E. R., Kumar, P., Prabhakar, B. S., and Haus, J. M. (2018). A single high-fat meal alters human soluble rage profiles and pbmc rage expression with no effect of prior aerobic exercise. *Physiol. Rep.* 6:e13811. doi: 10.14814/phy2.13811

Garrett-Sinha, L. A. (2013). Review of ETS1 structure, function, and roles in immunity. *Cell. Mol. Life Sci.* 70, 3375–3390. doi: 10.1007/s00018-012-1243-7

Golay, A., and Bobbioni, E. (1997). The role of dietary fat in obesity. *Int. J. Obes. Relat. Metab. Disord.* 21:S2.

Guo, Y., and Gifford, D. K. (2017). Modular combinatorial binding among human trans-acting factors reveals direct and indirect factor binding. *BMC Genomics* 18:45. doi: 10.1186/s12864-016-3434-3

Han, H., Cho, J. W., Lee, S., Yun, A., Kim, H., Bae, D., et al. (2018). TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res.* 46, D380–D386. doi: 10.1093/nar/gkx1013

Hedrick, S. M., Michelini, R. H., Doedens, A. L., Goldrath, A. W., and Stone, E. L. (2012). FOXO transcription factors throughout T cell biology. *Nat. Rev. Immunol.* 12, 649–661. doi: 10.1038/nri3278

Herieka, M., and Erridge, C. (2014). High-fat meal induced postprandial inflammation. *Mol. Nutr. Food Res.* 58, 136–146. doi: 10.1002/mnfr.201300104

Hill, S. M., Heiser, L. M., Cokelaer, T., Unger, M., Nesser, N. K., Carlin, D. E., et al. (2016). Inferring causal molecular networks: empirical assessment through a community-based effort. *Nat. Methods* 13, 310–318. doi: 10.1038/nmeth.3773

Ibarra, I. L., Hollmann, N. M., Klaus, B., Augsten, S., Velten, B., Hennig, J., et al. (2020). Mechanistic insights into transcription factor cooperativity and its impact on protein-phenotype interactions. *Nat. Commun.* 11:124. doi: 10.1038/s41467-019-13888-7

Irrthum, A., Wehenkel, L., Geurts, P., et al. (2010). Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE* 5:e12776. doi: 10.1371/journal.pone.0012776

Jiang, B., Shi, Y., Zhang, X., Xin, X., Qi, L., Guo, H., et al. (2017). PIF3 is a negative regulator of the cbf pathway and freezing tolerance in arabidopsis. *Proc. Natl. Acad. Sci. U.S.A.* 114, E6695–E6702. doi: 10.1073/pnas.1706226114

Jin, J., He, K., Tang, X., Li, Z., Lv, L., Zhao, Y., et al. (2015). An arabidopsis transcriptional regulatory map reveals distinct functional and evolutionary features of novel transcription factors. *Mol. Biol. Evol.* 32, 1767–1773. doi: 10.1093/molbev/msv058

Jin, J., Tian, F., Yang, D. C., Meng, Y. Q., Kong, L., Luo, J., et al. (2016). PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res.* 45, D1040–D1045. doi: 10.1093/nar/gkw982

Jolma, A., Yin, Y., Nitta, K. R., Dave, K., Popov, A., Taipale, M., et al. (2015). DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* 527, 384–388. doi: 10.1038/nature15518

Kang, M., Lee, S., Lee, D., and Kim, S. (2020). Learning cell-type-specific gene regulation mechanisms by multi-attention based deep learning with regulatory latent space. *Front. Genet.* 11:869. doi: 10.3389/fgene.2020.00869

Kerdiles, Y. M., Stone, E. L., Beisner, D. L., McGargill, M. A., Ch'en, I. L., Stockmann, C., et al. (2010). FOXO transcription factors control regulatory T cell development and function. *Immunity* 33, 890–904. doi: 10.1016/j.immuni.2010.12.002

Kim, J., Choi, M., Kim, J. R., Jin, H., Kim, V. N., and Cho, K. H. (2012). The co-regulation mechanism of transcription factors in the human gene regulatory network. *Nucleic Acids Res.* 40, 8849–8861. doi: 10.1093/nar/gks664

Kuss, M., and Graepel, T. (2003). *The Geometry of Kernel Canonical Correlation Analysis*. Technical Report, Tübingen: Max Planck Institute for Biological Cybernetics.

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

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., et al. (2018). The human transcription factors. *Cell* 172, 650–665. doi: 10.1016/j.cell.2018.01.029

Lee, C. M., and Thomashow, M. F. (2012). Photoperiodic regulation of the C-repeat binding factor (CBF) cold acclimation pathway and freezing tolerance in arabidopsis thaliana. *Proc. Natl. Acad. Sci. U.S.A.* 109, 15054–15059. doi: 10.1073/pnas.1211295109

Lee, S., Kim, D., Lee, K., Choi, J., Kim, S., Jeon, M., et al. (2016). Best: next-generation biomedical entity search tool for knowledge discovery from biomedical literature. *PLoS ONE* 11:e0164680. doi: 10.1371/journal.pone.0164680

Lemay, D. G., Huang, S., Huang, L., Alkan, Z., Kirschke, C., Burnett, D. J., et al. (2019). Temporal changes in postprandial blood transcriptomes reveal subject-specific pattern of expression of innate immunity genes after a high-fat meal. *J. Nutr. Biochem.* 72:108209. doi: 10.1016/j.jnutbio.2019.06.007

Luo, W., Hankenson, K. D., and Woolf, P. J. (2008). Learning transcriptional regulatory networks from high throughput gene expression data using continuous three-way mutual information. *BMC Bioinformatics* 9:467. doi: 10.1186/1471-2105-9-467

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., et al. (2006). ARACNe: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 7:S7. doi: 10.1186/1471-2105-7-S1-S7

Marshall, J. A., and Bessesen, D. H. (2002). *Dietary Fat and the Development of Type 2 Diabetes*. Arlington, VA: Diabetes Cares. doi: 10.2337/diacare.25.3.620

McLaughlin, T., Ackerman, S. E., Shen, L., Engleman, E., et al. (2017). Role of innate and adaptive immunity in obesity-associated metabolic disease. *J. Clin. Invest.* 127, 5–13. doi: 10.1172/JCI88876

Mechtcheriakova, D., Wlachos, A., Sobanov, J., Kopp, T., Reuschel, R., Bornancin, F., et al. (2007). Sphingosine 1-phosphate phosphatase 2 is induced during inflammatory responses. *Cell. Signal.* 19, 748–760. doi: 10.1016/j.cellsig.2006.09.004

Meissner, M., Orsini, E., Ruschhaupt, M., Melchinger, A. E., Hincha, D. K., and Heyer, A. G. (2013). Mapping quantitative trait loci for freezing tolerance in a recombinant inbred line population of a *Rabidopsis thaliana* accessions tenela and C24 reveals reveille1 as negative regulator of cold acclimation. *Plant Cell Environ.* 36, 1256–1267. doi: 10.1111/pce.12054

Michael, T. P., Salomé, P. A., and McClung, C. R. (2003). Two arabidopsis circadian oscillators can be distinguished by differential temperature sensitivity. *Proc. Natl. Acad. Sci. U.S.A.* 100, 6878–6883. doi: 10.1073/pnas.1131995100

Ming, M., Guanhua, L., Zhanhai, Y., Guang, C., and Xuan, Z. (2009). Effect of the *Lycium barbarum* polysaccharides administration on blood lipid metabolism and oxidative stress of mice fed high-fat diet *in vivo*. *Food Chem.* 113, 872–877. doi: 10.1016/j.foodchem.2008.03.064

Mölder, F., Jablonski K. P., Letcher B., Hall M. B., Tomkins-Tinch C. H., Socha V., et al. (2021). Sustainable data analysis with Snakemake. *F1000Res* 10, 33. doi: 10.12688/f1000research.29032.1

Ohkura, N., and Sakaguchi, S. (2010). FOXO1 and FOXO3 help FOXP3. *Immunity* 33, 835–837. doi: 10.1016/j.immuni.2010.12.004

Ramasamy, R., Yan, S. F., and Schmidt, A. M. (2011). Receptor for age (RAGE): signaling mechanisms in the pathogenesis of diabetes and its complications. *Ann. N. Y. Acad. Sci.* 1243:88. doi: 10.1111/j.1749-6632.2011.06320.x

Ravasi, T., Suzuki, H., Cannistraci, C. V., Katayama, S., Bajic, V. B., Tan, K., et al. (2010). An atlas of combinatorial transcriptional regulation in mouse and man. *Cell* 140, 744–752. doi: 10.1016/j.cell.2010.01.044

Rhee, J. K., Joung, J. G., Chang, J. H., Fei, Z., and Zhang, B. T. (2009). Identification of cell cycle-related regulatory motifs using a kernel canonical correlation analysis. *BMC Genomics* 10:S29. doi: 10.1186/1471-2164-10-S3-S29

Richfield, O., Alam, M. A., Calhoun, V., and Wang, Y. P. (2016). "Learning schizophrenia imaging genetics data via multiple kernel canonical correlation analysis," in *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (Shenzhen: IEEE), 507–511. doi: 10.1109/BIBM.2016.7822570

Rieder, D., Trajanoski, Z., and McNally, J. (2012). Transcription factories. *Front. Genet.* 3:221. doi: 10.3389/fgene.2012.00221

Russell, L., and Garrett-Sinha, L. A. (2010). Transcription factor ETS-1 in cytokine and chemokine gene regulation. *Cytokine* 51, 217–226. doi: 10.1016/j.cyto.2010.03.006

Salmeron, J., Hu, F. B., Manson, J. E., Stampfer, M. J., Colditz, G. A., Rimm, E. B., et al. (2001). Dietary fat intake and risk of type 2 diabetes in women. *Am. J. Clin. Nutr*. 73, 1019–1026. doi: 10.1093/ajcn/73.6.1019

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., et al. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet*. 34, 166–176. doi: 10.1038/ng1165

Seo, J. W., and Kim, S. D. (2013). "Novel PCA-based color-to-gray image conversion," in *2013 IEEE International Conference on Image Processing* (Melbourne, VIC: IEEE), 2279–2283. doi: 10.1109/ICIP.2013.6738470

Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). Biogrid: a general repository for interaction datasets. *Nucleic Acids Res*. 34, D535–D539. doi: 10.1093/nar/gkj109

Sutherland, H., and Bickmore, W. A. (2009). Transcription factories: gene expression in unions? *Nat. Rev. Genet*. 10, 457–466. doi: 10.1038/nrg 2592

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., et al. (2016). The string database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res*. 45, D362–D368. doi: 10.1093/nar/gkw937

Tang, M., Marin, D., Ayed, I. B., and Boykov, Y. (2019). Kernel cuts: Kernel and spectral clustering meet regularization. *Int. J. Comput. Vision* 127, 477–511. doi: 10.1007/s11263-018-1115-1

Tian, F., Yang, D. C., Meng, Y. Q., Jin, J., and Gao, G. (2020). PlantRegMap: charting functional regulatory maps in plants. *Nucleic Acids Res*. 48, D1104–D1113. doi: 10.1093/nar/gkz1020

Vogel, J. T., Zarka, D. G., Van Buskirk, H. A., Fowler, S. G., and Thomashow, M. F. (2005). Roles of the CBF2 and ZAT12 transcription factors in configuring the low temperature transcriptome of arabidopsis. *Plant J*. 41, 195–211. doi: 10.1111/j.1365-313X.2004.02288.x

Wang, D. Z., Jin, Y. N., Ding, X. H., Wang, W. J., Zhai, S. S., Bai, L. P., et al. (2017). Gene regulation and signal transduction in the ICE-CBF-COR signaling pathway during cold stress in plants. *Biochemistry* 82, 1103–1117. doi: 10.1134/S0006297917100030

Wise, A., and Bar-Joseph, Z. (2015). CDREM: inferring dynamic combinatorial gene regulation. *J. Comput. Biol*. 22, 324–333. doi: 10.1089/cmb.2015.0010

Xiong, J., and Zhou, T. (2012). Gene regulatory network inference from multifactorial perturbation data using both regression and correlation analyses. *PLoS ONE* 7:e43819. doi: 10.1371/journal.pone.0043819

Yamanishi, Y., Vert, J. P., Nakaya, A., and Kanehisa, M. (2003). Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis. *Bioinformatics* 19, i323–i330. doi: 10.1093/bioinformatics/btg1045

Zhang, X., Zhao, X. M., He, K., Lu, L., Cao, Y., Liu, J., et al. (2012). Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics* 28, 98–104. doi: 10.1093/bioinformatics/btr626

Zhou, Y., Xun, Q., Zhang, D., Lv, M., Ou, Y., and Li, J. (2019). TCP transcription factors associate with phytochrome interacting factor 4 and cryptochrome 1 to regulate thermomorphogenesis in arabidopsis thaliana. *iScience* 15, 600–610. doi: 10.1016/j.isci.2019.04.002

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

ORIGINAL RESEARCH
published: 22 June 2021
doi: 10.3389/fgene.2021.687745

Alcohol Consumption and Risk of Common Autoimmune Inflammatory Diseases—Evidence From a Large-Scale Genetic Analysis Totaling 1 Million Individuals

Xia Jiang[1,2], Zhaozhong Zhu[2], Ali Manouchehrinia[1], Tomas Olsson[1], Lars Alfredsson[1] and Ingrid Kockum[1]*

[1] Department of Clinical Neuroscience, Center for Molecular Medicine, Karolinska Institute, Stockholm, Sweden,
[2] Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, United States

OPEN ACCESS

Edited by:
Chao Xu,
University of Oklahoma Health Sciences Center, United States

Reviewed by:
Chikashi Terao,
RIKEN, Japan
Tianzhong Yang,
University of Minnesota Twin Cities, United States

***Correspondence:**
Xia Jiang
xia.jiang@ki.se

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: 30 March 2021
Accepted: 26 April 2021
Published: 22 June 2021

Citation:
Jiang X, Zhu Z, Manouchehrinia A, Olsson T, Alfredsson L and Kockum I (2021) Alcohol Consumption and Risk of Common Autoimmune Inflammatory Diseases—Evidence From a Large-Scale Genetic Analysis Totaling 1 Million Individuals. Front. Genet. 12:687745. doi: 10.3389/fgene.2021.687745

Purpose: Observational studies have suggested a protective effect of alcohol intake with autoimmune disorders, which was not supported by Mendelian randomization (MR) analyses that used only a few (<20) instrumental variables.

Methods: We systemically interrogated a putative causal relationship between alcohol consumption and four common autoimmune disorders, using summary-level data from the largest genome-wide association study (GWAS) conducted on inflammatory bowel disease (IBD), rheumatoid arthritis (RA), multiple sclerosis (MS), and systemic lupus erythematosus (SLE). We quantified the genetic correlation to examine a shared genetic similarity. We constructed a strong instrument using 99 genetic variants associated with drinks per week and applied several two-sample MR methods. We additionally incorporated excessive drinking as reflected by alcohol use disorder identification test score.

Results: We observed a negatively shared genetic basis between alcohol intake and autoimmune disorders, although none was significant (r_g = −0.07 to −0.02). For most disorders, genetically predicted alcohol consumption was associated with a slightly (10–25%) decreased risk of onset, yet these associations were not significant. Meta-analyzing across RA, MS, and IBD, the three Th1-related disorders yielded to a marginally significantly reduced effect [OR = 0.70 (0.51–0.95), P = 0.02]. Excessive drinking did not appear to reduce the risk of autoimmune disorders.

Conclusions: With its greatly augmented sample size and substantially improved statistical power, our MR study does not convincingly support a beneficial role of alcohol consumption in each individual autoimmune disorder. Future studies may be designed to replicate our findings and to understand a causal effect on disease prognosis.

Keywords: Mendelian Randomization (MR), alcohol consumption amount, excessive drinking, autoimmune disease, genetic correlation, large-scale genetic analysis

INTRODUCTION

Alcohol contains components such as ethanol and antioxidants and is considered as a complex modulator to the immune system (Barr et al., 2016). Several *in vitro* and *in vivo* studies have demonstrated that ethanol modulates the function of monocytes and dendritic cells (innate immune cells) in a dose- and time-dependent manner. For example, while acute high-level exposure to ethanol inhibits proinflammatory cytokine production, long-term moderate administration of ethanol stimulates the process. In addition, *in vivo* consumption of moderate doses of alcohol enhances phagocytosis and reduces inflammatory cytokine production whereas chronic consumption of large doses inhibits phagocytosis and production of growth factors. For cell-mediated and humoral immunity (adaptive immunity), chronic alcohol abuse significantly reduces both the number and frequency of T lymphocytes, resulting in an increased proportion of memory T cells relative to naïve T cells, which interferes the development of efficacious responses to infection and vaccination. In contrast, moderate alcohol intake increases the frequency of lymphocytes. Moreover, alcohol also modulates the hypothalamic–pituitary–adrenal axis and influences the function of immune cells residing in the central nervous system (CNS) particularly astrocytes and microglia, which tightly regulates the stress response, neuronal function, and CNS homeostasis, in turn affecting immunity (Barr et al., 2016).

While it appears that high doses of alcohol directly suppress a wide range of immune responses and moderate doses of alcohol play a beneficial role in the immune system, the complex interplay among alcohol intake, immune response, and inflammatory processes remains to be understood (Romeo et al., 2007). The relationship between alcohol consumption and a number of chronic autoimmune inflammatory disorders has been investigated through conventional epidemiological studies, of which results remain inconclusive (Wang et al., 2008, 2015; Jin et al., 2014; Linneberg and Gonzalez-Quintela, 2016). It has been argued that the validity of findings from observational studies could be plagued by measurement error, confounding, and/or reverse causality.

Mendelian randomization (MR) is a novel statistical approach that uses genetic variants (instrumental variables, IVs; usually single-nucleotide polymorphisms, SNPs) as proxies to make causal inference between exposure(s) and outcome(s). Since genotypes are randomly assigned at conception and always precede disease onset, MR mirrors the randomization process in controlled trials and is less susceptible to confounding and reverse causality (Smith and Ebrahim, 2003). Nevertheless, application of MR in the field of autoimmune diseases remains limited—so far, only two MR(s) have been conducted to investigate the effect of alcohol with the risk of rheumatoid arthritis (RA; Bae and Lee, 2019b) and systemic lupus erythematosus (SLE; Bae and Lee, 2019a), each involving less than 20 genetic instruments.

A recent genome-wide association study (GWAS) conducted in alcohol drinking behavior (defined as drinks per week) has identified 99 significant independent loci (Liu et al., 2019), and the GWAS summary statistics for most autoimmune diseases have been made publicly available. Taking advantage of these enormous progresses made in genetic discoveries for complex traits, we aim to perform a large-scale comprehensive study to systemically interrogate the effect of alcohol consumption on a range of common autoimmune inflammatory disorders, leveraging the genetic information available for 1 million individuals of European ancestry. We will explore both a *shared genetic basis* as reflected by genetic correlation analysis and a *causal relationship* as reflected by MR analysis.

MATERIALS AND METHODS

We performed the current study employing a standard framework, that is, a genetic correlation analysis defined as the proportion of variance that two traits share due to genetic causes, and a two-sample MR analysis, where instrument–exposure (or IV–exposure, SNP–exposure) and instrument–outcome (or IV–outcome, SNP–outcome) associations were extracted from two independent non-overlapping sets of participants. For a conceptual framework of our MR (a flowchart of current study), please see **Supplementary Figure 1**; for characteristics of exposure and outcome genetic data, please see **Supplementary Table 1**.

IV–Exposure

The hitherto largest GWAS of alcohol consumption was conducted using an imputation-accuracy-aware meta-analysis totaling 941,280 individuals of European ancestry recruited from 34 participating studies (Liu et al., 2019). The exposure, drinks per week, was defined as the average number of drinks a participant reported drinking each week, aggregated across all types of alcohol. If a participating study recorded binned response ranges (e.g., one to four drinks per week, 5–10 drinks per week), the midpoint of the range was used. The phenotype was left-anchored at 1 and log-transformed prior to analysis. This large-scale meta-GWAS has identified 99 genome-wide significant variants associated with drinks per week after conditional and joint analyses. We used these 99 independent SNPs as our instruments and extracted IV–exposure associations (beta-coefficients, standard errors) and relevant information (rsID, effect allele, allele frequency, genomic coordinates) from the abovementioned alcohol GWAS. Details on characteristics of the 99 IVs are presented in **Supplementary Table 1**. We also obtained full-set GWAS summary data for genetic correlation analysis.

While drinks per week reflect normal or general drinking behavior, we included one additional exposure, alcohol use disorder identification test consumption score (AUDIT), which reflects excessive or harmful drinking behavior. The GWAS of AUDIT was conducted in a multi-ancestry Million Veteran Program sample of 274,424 individuals, and 13 GWAS-significant independent loci were identified among Europeans to be associated with alcohol use disorder (Kranzler et al., 2019). We used these 13 SNPs as IVs to perform additional analysis and to complement with our main findings (**Supplementary Table 2**).

IV-Outcome

We systemically examined the role of alcohol consumption in four autoimmune diseases. We collected the hitherto largest full-set GWAS summary data of inflammatory bowel disease (IBD; Liu et al., 2015) and its subsets [Crohn's disease (CD) and ulcerative colitis (UC)], RA (Okada et al., 2014), SLE, (Bentham et al., 2015) and multiple sclerosis (MS; International Multiple Sclerosis Genetics Consortium, 2019), all of European ancestry. We selected these four autoimmune disorders due to two reasons: (1) they are common and (2) they had GWAS with decent sample size and SNP coverage (>5,000 cases and >10,000 controls and >1,000,000 genetic markers) to ensure statistical power. From these GWAS summary data, we extracted IV–outcome associations (beta-coefficients and standard errors) and relevant information (rsID, effect allele, allele frequency, genomic coordinates).

The abundant available samples make our study so far the largest of its kind, leveraging on the genetic information from 49,336 cases of autoimmune disorders and 108,387 controls (number of cases/controls for each outcome, IBD: 12,882/21,770; UC: 6,968/20,464; CD: 5,956/14,927; RA: 14,361/43,923; MS: 14,802/26,703; SLE: 7,291/15,991). Details of the outcome GWAS(s) are shown in **Supplementary Table 3**.

Statistical Analysis

Genetic Correlation Analysis

The correlation between the genetic influences on a trait and the genetic influences on a different trait estimates the degree of *causal overlap or pleiotropy*. We quantified the genome-wide genetic correlation between alcohol consumption and each disorder, using an algorithm implemented in statistical software linkage disequilibrium score regression (LDSC). LDSC leverages the relationship between association statistics and linkage disequilibrium patterns across the genome and estimates the genetic correlation using only GWAS summary-level data (Bulik-Sullivan et al., 2015).

Mendelian Randomization Analysis

We next evaluated a *causal relationship* between alcohol consumption and autoimmune disorders. MR yields an unbiased causal estimate based on observational data only when three model assumptions are satisfied. Namely, IVs should be robustly associated with the exposure (relevance), affect outcome only through the exposure (exclusion restriction), and should not be associated with confounders in the exposure–outcome relationship (exchangeability). To guarantee model assumption, we applied several MR approaches including a random-effect inverse variance-weighted method (IVW; Burgess et al., 2015), a maximum likelihood approach (Burgess et al., 2013), a weighted median approach (Bowden et al., 2016), and an MR–Egger regression (Bowden et al., 2015).

Briefly, the random-effect IVW pools estimate from each IV and provide causal estimation, assuming that all IVs are valid or are invalid in a way that the overall pleiotropy is balanced to be zero (Burgess et al., 2015). When there is considerable imprecision in the estimates, causal effect estimates from the IVW are overprecise, whereas the likelihood method gives appropriately sized confidence intervals (Burgess et al., 2013). In addition, we performed MR–Egger regression to test for bias due to directional pleiotropy, where the average of direct effects of the tested genetic variants on outcome is non-zero (Bowden et al., 2015). We employed a weighted median to provide consistent estimates even when up to 50% of the analyzed genetic variants are invalid (Bowden et al., 2016).

In addition, we performed several important sensitivity analyses to further validify model assumptions. For example, we excluded palindromic IVs (SNPs with alleles represented by the same pair of letters on the forward and reverse strands such as A/T or G/C SNPs. These SNPs can introduce ambiguity into the identity of the effect allele in the exposure and outcome GWASs.) (Hemani et al., 2018). We excluded IVs that were associated with potential confounding traits according to the GWAS catalog. Further, we employed a multivariable MR approach to adjust for potential horizontal pleiotropy acting in particular through the body mass index and smoking—the two

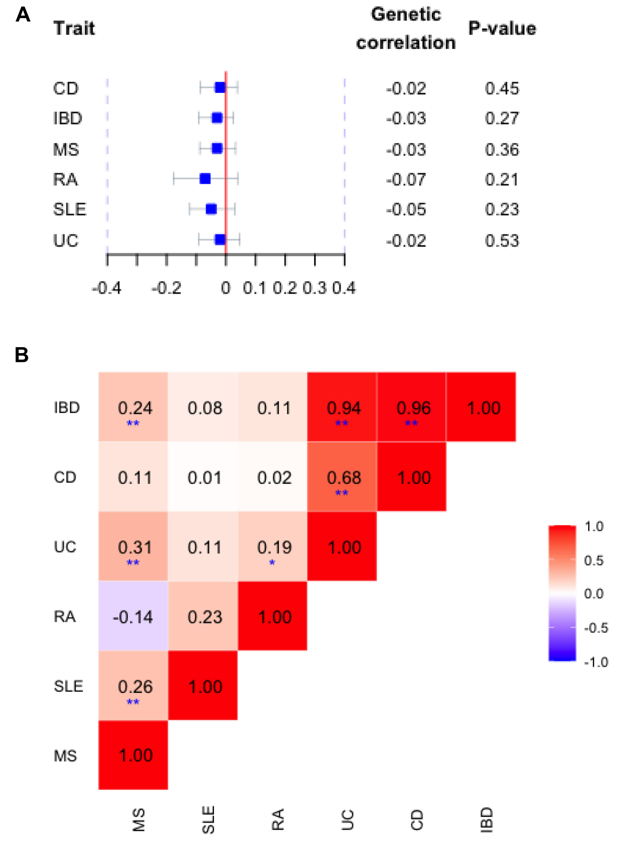


FIGURE 1 | (A) Genetic correlation between alcohol consumption (drinks per week) and risk of autoimmune inflammatory diseases. Blue squares and horizontal bars represent the point estimate and confidence intervals of genetic correlation with each disorder. **(B)** Pairwise correlation among autoimmune disorders. The color of each checker represents the magnitudes of correlation. A darker color represents a stronger correlation. **Bonferroni correct significance, *suggestive significance. IBD, inflammatory bowel disease; UC, ulcerative colitis; CD, Crohn's disease; RA, rheumatoid arthritis; SLE, systemic lupus erythematosus; MS, multiple sclerosis.

lifestyle behavioral traits tend to cluster together with alcohol consumption (Burgess and Thompson, 2015). We extracted IV-BMI effect sizes and IV-smoking effect sizes from the hitherto largest obesity (N = 700,000) (Yengo et al., 2018) and smoking (N = 1,232,091) (Liu et al., 2019) GWAS(s). Finally, we excluded one SNP at a time and performed IVW on the remaining SNPs to identify potential influence of outlying variants on the estimates.

Mendelian randomization methods evaluate an overall casual estimation; it is likely that several distinct causal mechanisms underlie the alcohol–disease relationship, in which a risk factor influences outcome with different magnitudes of causal effect. We examined such a scenario through MR-Clust (Foley et al., 2019), an approach that divides IVs into distinct clusters such that all variants in the cluster have similar causal estimates.

Finally, we complemented our main results of general drinking behavior, by incorporating genetic instruments associated with excessive or harmful drinking behavior (alcohol use disorder identification test). Given the fewer IVs associated with AUDIT (N = 13), we only performed primary analysis for this exposure (IVW and MR–Egger), as the diagnostic analyses including MVMR and MR-Clust were underpowered with the limited availability of genetic instruments.

We included four autoimmune disorders as main outcomes (CD and UC were treated as subsets of IBD) and performed analysis using different sets of instruments as well as different statistical approaches; our results were likely to suffer from false positives due to multiple comparisons. Therefore, we considered a two-sided P-threshold of 0.05 as suggestive significance. An arbitrarily corrected P-threshold of 0.01 (0.05/4) was used as statistical significance. All MR analyses were performed using R software version 4.0.2 with packages "TwoSampleMR," "MendelianRandomization," and "MRclust."

RESULTS

As shown in **Figure 1A**, using full-set GWAS summary data, we observed negligible shared genetic similarities of alcohol consumption with each disorder. Indeed, the genetic correlation estimates were all negative ranging from −0.07 to −0.02, meaning that the genetic variant associated with an increase in

TABLE 1 | The association between genetically predicted levels of alcohol consumption and risk of common autoimmune inflammatory diseases.

Methods	# SNP	OR (95%CI)	*P*-value	*P*-value for intercept	# SNP	OR (95%CI)	*P*-value	*P*-value for intercept
	Full-set				Remove palindromic SNPs			
Inflammatory bowel disease								
IVW	98	0.84 (0.54–1.29)	0.42		84	0.84 (0.52–1.34)	0.46	
Maximum likelihood	98	0.84 (0.63–1.11)	0.23		84	0.84 (0.62–1.13)	0.25	
Weighted median	98	0.92 (0.56–1.51)	0.73		84	0.92 (0.56–1.52)	0.75	
MR–Egger	98	0.96 (0.46–2.00)	0.92	0.64	84	1.01 (0.47–2.21)	0.97	0.55
Ulcerative colitis								
IVW	98	0.93 (0.59–1.49)	0.77		84	0.92 (0.56–1.52)	0.75	
Maximum likelihood	98	0.93 (0.65–1.33)	0.70		84	0.92 (0.63–1.34)	0.66	
Weighted median	98	0.99 (0.51–1.91)	0.97		84	1.00 (0.52–1.92)	1.00	
MR–Egger	98	0.97 (0.44–2.18)	0.95	0.90	84	0.99 (0.43–2.28)	0.98	0.84
Crohn's disease								
IVW	98	0.70 (0.38–1.27)	0.24		84	0.70 (0.36–1.36)	0.30	
Maximum likelihood	98	0.71 (0.48–1.03)	0.07		84	0.70 (0.47–1.03)	0.07	
Weighted median	98	0.98 (0.51–1.87)	0.95		84	0.99 (0.52–1.88)	0.98	
MR–Egger	98	0.98 (0.36–2.63)	0.97	0.40	84	1.02 (0.36–2.96)	0.96	0.37
Rheumatoid arthritis								
IVW	93	0.80 (0.54–1.19)	0.27		80	0.85 (0.56–1.29)	0.45	
Maximum likelihood	93	0.80 (0.56–1.14)	0.22		80	0.85 (0.58–1.24)	0.40	
Weighted median	93	1.38 (0.77–2.50)	0.28		80	1.43 (0.74–2.75)	0.29	
MR–Egger	93	1.45 (0.66–3.18)	0.36	0.09	80	1.58 (0.71–3.54)	0.27	0.08
Multiple sclerosis								
IVW	93	0.75 (0.49–1.12)	0.16		80	0.74 (0.47–1.16)	0.18	
Maximum likelihood	93	0.74 (0.53–1.03)	0.07		80	0.73 (0.52–1.04)	0.08	
Weighted median	93	1.13 (0.66–1.95)	0.65		80	1.13 (0.63–2.02)	0.68	
MR–Egger	93	1.17 (0.49–2.83)	0.72	0.26	80	1.27 (0.49–3.28)	0.62	0.20
Systemic lupus erythematosus								
IVW	82	1.10 (0.51–2.37)	0.80		70	1.14 (0.49–2.66)	0.76	
Maximum likelihood	82	1.11 (0.62–1.97)	0.73		70	1.14 (0.61–2.14)	0.67	
Weighted median	82	1.91 (0.71–5.12)	0.20		70	1.85 (0.65–5.27)	0.25	
MR–Egger	82	2.14 (0.29–15.69)	0.46	0.48	70	1.24 (0.13–11.89)	0.85	0.94

dose of alcohol tends to be associated with a decreased risk of autoimmune disorder. However, all these genetic correlations were not significant with confidence intervals including 1 and *P*-values > 0.05, contrasted by the significant pairwise genetic correlation observed among autoimmune disorders (**Figure 1B**).

Genetic correlation describes the intrinsic genome-wide average sharing of genetic effects between traits that are independent of environmental factors. We next performed MR analysis to elucidate a potential directional or causal association between alcohol and autoimmune disorders. We were able to match almost all alcohol-associated genetic instruments to our outcome data, ranging from 98 (99%) in IBD, 93 in RA and MS (94%), and 82 in SLE (83%)—a virtually complete coverage (**Supplementary Table 4**). These 99 alcohol-associated genetic variants constructed a strong IV with an overall F-statistic of 122.4.

As shown in **Table 1**, for most autoimmune disorders examined by us, genetically predicted alcohol consumption was associated with a slightly (10–25%) decreased risk of disease onset (IBD: $\text{OR}_{IVW} = 0.84$; UC: $\text{OR}_{IVW} = 0.93$; CD: $\text{OR}_{IVW} = 0.70$; RA: $\text{OR}_{IVW} = 0.80$; MS: $\text{OR}_{IVW} = 0.75$); for SLE, an OR_{IVW} of 1.10 was observed. However, all these associations were not statistically significant with confidence intervals covering 1.00 (95%CI, IBD: 0.54–1.29; UC: 0.59–1.49; CD: 0.38–1.27; RA: 0.54–1.19; MS: 0.49–1.12; SLE: 0.51–2.37) and *P*-values larger than 0.05. Such null findings were supported by the maximum likelihood method and the weighted median approach where we observed non-significant effects (although in opposite directions for RA and MS) with confidence intervals covering 1. MR–Egger regression did not reveal apparent signs of horizontal pleiotropy (*P*-values for the MR–Egger intercept, IBD: $P = 0.64$; UC: $P = 0.90$; CD: $P = 0.40$; RA: $P = 0.09$; MS: $P = 0.26$; SLE: $P = 0.48$).

TABLE 2 | Genetically predicted levels of alcohol consumption and the risk of autoimmune inflammatory diseases.

Methods	# SNP	OR (95%CI)	*P*-value	*P*-value for intercept
Inflammatory bowel disease				
IVW	71	0.78 (0.44–1.36)	0.38	
Maximum likelihood	71	0.77 (0.51–1.16)	0.21	
Weighted median	71	0.63 (0.33–1.21)	0.17	
MR–Egger	71	2.18 (0.50–9.59)	0.31	0.14
Ulcerative colitis				
IVW	71	0.89 (0.47–1.69)	0.72	
Maximum likelihood	71	0.89 (0.53–1.48)	0.64	
Weighted median	71	0.66 (0.30–1.46)	0.31	
MR–Egger	71	1.52 (0.27–8.39)	0.64	0.51
Crohn's disease				
IVW	71	0.61 (0.28–1.34)	0.22	
Maximum likelihood	71	0.60 (0.34–1.04)	0.07	
Weighted median	71	1.19 (0.48–2.93)	0.71	
MR–Egger	71	3.24 (0.41–25.51)	0.27	0.09
Rheumatoid arthritis				
IVW	68	0.51 (0.30–0.88)	**0.02**	
Maximum likelihood	68	0.50 (0.31–0.81)	**0.005**	
Weighted median	68	0.88 (0.42–1.85)	0.73	
MR–Egger	68	2.08 (0.42–10.30)	0.37	0.07
Multiple sclerosis				
IVW	67	0.85 (0.50–1.42)	0.54	
Maximum likelihood	67	0.85 (0.55–1.30)	0.44	
Weighted median	67	1.12 (0.60–2.13)	0.71	
MR–Egger	67	3.01 (0.62–14.71)	0.18	0.10
Systemic lupus erythematosus				
IVW	60	1.24 (0.49–3.15)	0.66	
Maximum likelihood	60	1.25 (0.61–2.54)	0.54	
Weighted median	60	0.63 (0.20–2.02)	0.44	
MR–Egger	60	4.23 (0.31–57.18)	0.28	0.32

A sensitivity analysis excluding SNPs associated with potential confounding traits.

TABLE 3 | Genetically predicted levels of alcohol consumption and risk of common autoimmune diseases.

Methods	# SNP	OR (95%CI)	*P*-value
Inflammatory bowel disease			
Body mass index	48	0.94 (0.42–2.11)	0.89
Smoking status	97	0.80 (0.51–1.30)	0.38
Ulcerative colitis			
Body mass index	48	1.20 (0.52–2.75)	0.67
Smoking status	97	0.94 (0.56–1.59)	0.83
Crohn's disease			
Body mass index	48	0.56 (0.17–1.82)	0.34
Smoking status	97	0.64 (0.33–1.25)	0.19
Rheumatoid arthritis			
Body mass index	48	0.66 (0.36–1.21)	0.18
Smoking status	92	0.79 (0.50–1.24)	0.31
Multiple sclerosis			
Body mass index	48	0.49 (0.26–0.91)	**0.02**
Smoking status	92	0.79 (0.49–1.27)	0.33
Systemic lupus erythematosus			
Body mass index	42	0.86 (0.31–2.38)	0.77
Smoking status	93	0.83 (0.34–2.03)	0.69

Multivariable analysis adjusting for the effect of body mass index and smoking status.

Palindromic SNPs introduce ambiguity for the identity of effect alleles in exposure and outcome data. Sensitivity analysis removing palindromic SNPs (**Table 1**) revealed similar null associations for all autoimmune disorders.

A search of GWAS catalog[1] reveals considerable potential for pleiotropic effects, as some IVs were identified to be associated with important potential confounders with genome-wide significance (**Supplementary Table 1**). We next performed a sensitivity analysis excluding those SNPs. As shown in **Table 2**, consistent with our primary analysis, we did not observe any significantly altered risk of autoimmune disorders with genetic predisposition to alcohol consumption. A significantly reduced risk of RA was identified [IVW, OR (95%CI) = 0.51 (0.30–0.88)], yet such an association did not pass multiple corrections and did not remain directionally consistent in other methods [MR–Egger, OR (95%CI) = 2.08 (0.42–10.30)]. In both sensitivity analyses, no

[1] https://www.ebi.ac.uk/gwas/

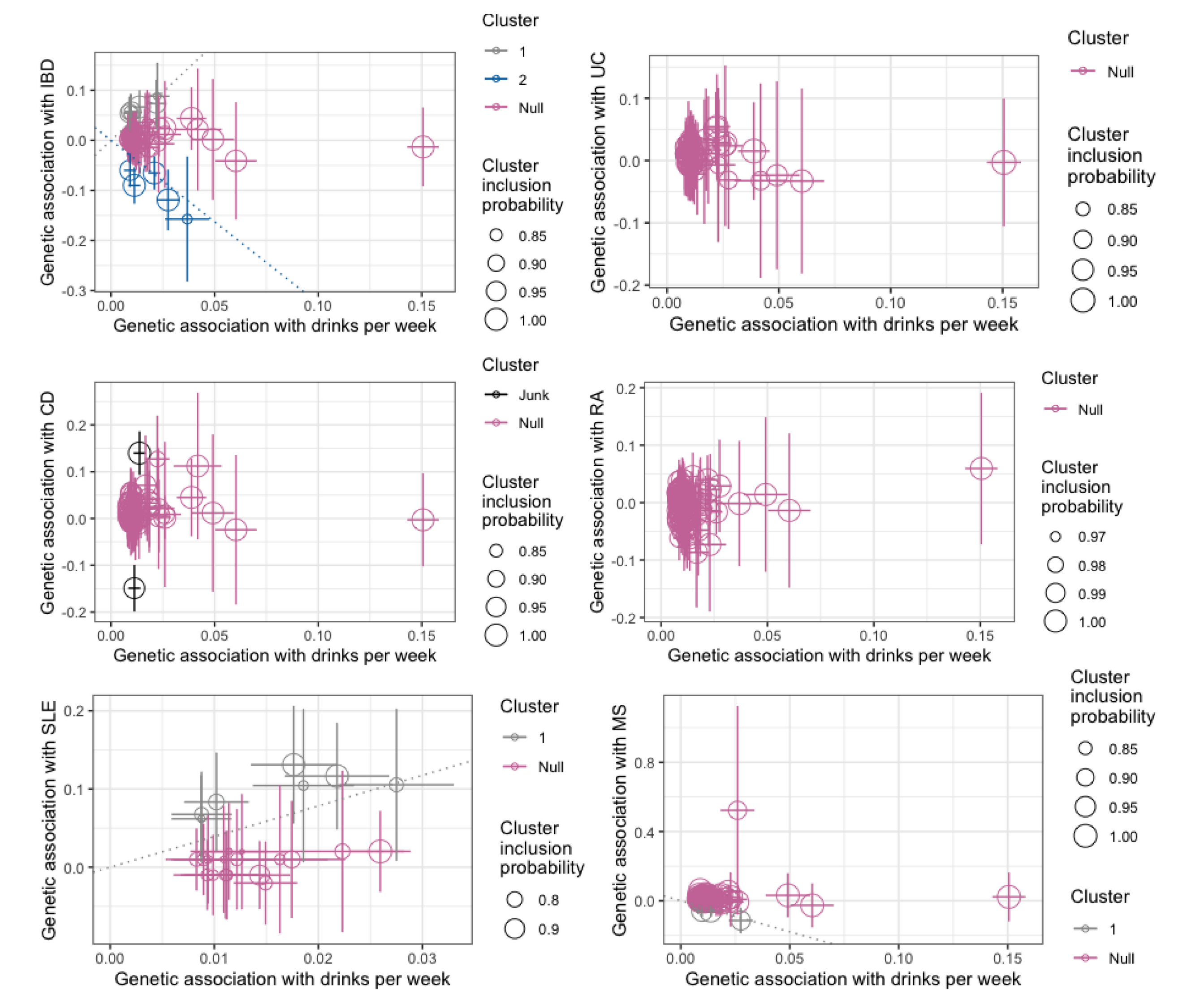


FIGURE 2 | Genetic associations with alcohol consumption (drinks per week) and risk of autoimmune inflammatory diseases (log odds) per additional alcohol consumption increasing alleles. Each genetic variant is represented by a point. Error bars are 95% confidence intervals for the genetic associations. Colors represent the clusters. Variants are only assigned to the cluster if the conditional probability is >0.8 and cluster only displayed if at least four variants are assigned to the cluster. IBD, inflammatory bowel disease; UC, ulcerative colitis; CD, Crohn's disease; RA, rheumatoid arthritis; SLE, systemic lupus erythematosus; MS, multiple sclerosis.

apparent horizontal pleiotropy was observed as reflected by the intercepts of MR–Egger regression (**Tables 1**, **2**).

Inflammatory bowel disease, RA, and MS are Th1-related autoimmune disorders, meta-analyzing across these three traits yielded to a reduced effect with marginal significance [OR_{meta}(95%CI) = 0.79 (0.63–1.01), P = 0.06 using all IVs; OR_{meta}(95%CI) = 0.70 (0.51–0.95), P = 0.02 using IVs excluding confounders]. Meta-analyzing all four traits did not reveal any significant effect [OR_{meta}(95%CI) = 0.82 (0.65–1.03), P = 0.08 using all IVs; OR_{meta}(95%CI) = 0.74 (0.54–1.02), P = 0.06 using curated IVs without pleiotropic effects].

Obesity and smoking are two important environmental risk factors clustering together with alcohol intake. We therefore employed a multivariable MR approach to adjust for potential horizontal pleiotropy acting in particular through BMI and smoking. As shown in **Table 3** and consistent with our sensitivity analysis, we did not observe apparent significant effects of alcohol consumption with risk of autoimmune disease after adjusting for BMI and smoking, except a suggestive reduced effect with MS which did not withstand multiple corrections (OR = 0.49 and P = 0.02). Leave-one-out analysis did not identify any outlying variants (**Supplementary Table 5**).

Alcohol consumption-associated variants may influence the risk of autoimmune diseases via distinct biological mechanisms. We therefore examined a scenario where variants can be divided into different clusters. According to MR-Clust, each IV is only assigned to a cluster if the conditional probability of belonging to that cluster is high (larger than 0.8) and clusters are only displayed if at least four IVs are assigned to it (Foley et al., 2019). As shown in **Figure 2**, for IBD, we observed two distinct clusters suggesting one strong positive causal effect and one strong negative causal effect; for SLE, we observed a single cluster suggesting a strong positive causal effect; and for MS, we observed a single cluster suggesting a strong negative causal effect. However, when we performed MR-Clust analysis excluding confounding IVs (corresponding to IVs used in **Table 2**), all previously observed clusters disappeared, largely consistent with an overall null finding (data not shown).

Finally, we complemented our main results by incorporating IVs associated with excessive or harmful drinking behavior (AUDIT, *N* = 13). As shown in **Table 4** and consistent with our main findings, excessive drinking did not appear to reduce the risk of autoimmune disorders. On the contrary, we observed an increased non-significant risk of IBD (OR = 1.21; 1.25 for UC and 1.14 for CD) and RA (OR = 1.16) with harmful drinking. We stress caution when interpreting these results given the very few genetic instruments associated with AUDIT.

DISCUSSION

We conducted a large-scale comprehensive genetic analysis to systemically interrogate the role of alcohol consumption in several common autoimmune inflammatory disorders. Overall, alcohol consumption and autoimmune disorder share a reverse yet non-significant genetic basis. Despite a few suggestive significant findings from MR in support of alcohol intake and a reduced risk of RA and MS, these results did not withstand multiple corrections. Meta-analyzing all traits did not reveal significant effects, and meta-analyzing three Th1-related disorders (IBD, RA, and MS) yielded to a reduced effect with significance (*P* = 0.02) not withstanding multiple corrections. Therefore, we consider an overall null association as our main conclusion.

To the best of our knowledge, the current MR study is the largest in sample size of its kind, leveraging information on 99 genetic instruments and involving data from more than one million individuals of European ancestry (941,280 individuals for exposure and 157,723 individuals for outcome). Two MR studies have been conducted for alcohol use and autoimmune disorder; none had the opportunity to achieve our power. For example, Bae and Bae and Lee (2019a,b) examined the causal relationship of alcohol intake with risk of RA and SLE, using approximately 20 alcohol-associated genome-wide significant SNPs as IVs. For outcomes, two meta-GWAS(s) were included, one with 5,539 autoantibody-positive RA patients (and 20,169 controls) and the other with 1,311 lupus patients (and 1,783 controls). No evidence of a causal relationship was identified for either RA [OR (95%CI) = 1.24 (0.82–1.89), *P* = 0.31] or lupus [OR (95%CI) = 0.46 (0.07–2.94), *P* = 0.42]. It is very likely that the few IVs did not fully capture the effect of alcohol. Our current study, with a largely augmented sample size and by incorporating additional alcohol consumption associated loci, greatly improved the strength of genetic instruments (F-statistic = 122.4) as well as both the accuracy and precision of MR estimates, as compared with previous findings.

We found an overall protective effect of alcohol intake on the three Th1-mediated autoimmune disorders (IBD, RA, and MS) as a whole; however, when breaking down into individual disorders, we did not find convincing evidence in support of a beneficial role of alcohol consumption. Our conclusion, although consistent with previous small-scale MR studies, is not

TABLE 4 | The association between genetically predicted levels of harmful alcohol consumption (alcohol use disorder identification test score) and risk of common autoimmune inflammatory diseases.

Methods	**# SNP**	**OR (95%CI)**	***P*-value**	***P*-value for intercept**	**# SNP**	**OR (95%CI)**	***P*-value**	***P*-value for intercept**
	Full-set				**Remove SNPs associated with confounders**			
Inflammatory bowel disease								
IVW	13	0.86 (0.63–1.16)	0.33		9	1.21 (0.87–1.70)	0.26	
MR–Egger	13	0.83 (0.52–1.31)	0.44	0.83	9	0.62 (0.15–2.51)	0.52	0.36
Ulcerative colitis								
IVW	13	0.95 (0.73–1.24)	0.70		9	1.25 (0.86–1.81)	0.24	
MR–Egger	13	0.90 (0.60–1.36)	0.64	0.76	9	0.40 (0.08–2.15)	0.32	0.21
Crohn's disease								
IVW	13	0.77 (0.50–1.21)	0.26		9	1.14 (0.72–1.78)	0.58	
MR–Egger	13	0.80 (0.41–1.56)	0.52	0.89	9	0.90 (0.12–6.51)	0.92	0.82
Rheumatoid arthritis								
IVW	13	**1.20 (1.00–1.44)**	**0.05**		9	1.16 (0.82–1.64)	0.42	
MR–Egger	13	1.23 (0.82–1.83)	0.34	0.91	9	1.15 (0.22–5.99)	0.87	1.00
Multiple sclerosis								
IVW	12	0.91 (0.70–1.18)	0.46		8	0.95 (0.69–1.32)	0.76	
MR–Egger	12	0.85 (0.51–1.39)	0.53	0.76	8	0.78 (0.20–3.04)	0.74	0.78
Systemic lupus erythematosus								
IVW	10	1.04 (0.64–1.69)	0.86		8	0.95 (0.59–1.54)	0.85	
MR–Egger	10	1.60 (0.40–6.37)	0.53	0.54	8	0.50 (0.10–2.53)	0.43	0.44

supported by observational studies. For example, Jin et al. (2014) summarized results from eight prospective studies containing 195,029 participants and 1,878 RA cases and found that low to moderate alcohol consumption yielded a preventive effect on the disease development [RR (95%CI) = 0.86 (0.78–0.94)]. Moreover, Wang et al. (2008) conducted a meta-analysis including six case–control studies and one cohort study and found a significantly decreased risk of lupus with moderate alcohol drinking [OR (95%CI) = 0.72 (0.55–0.95)]. Further, Zhu et al. (2015) aggregated data from nine case–control studies and one cohort study and identified an OR for the association between alcohol consumption and MS to be 0.91 (95%CI = 0.39–2.41). Reasons underlying such discrepancies can be multifactorial. Results from observational studies are likely to be plagued by measurement error or biases. For example, assessment of alcohol consumption is usually done by questionnaires, where frequency and amount of consumption are collected—precisely determining the amount of consumed alcohol is difficult. Indeed, alcohol intake can be expressed as a single measurement with "low," "medium," and "high" categories; such categorical measurement may however be of limited resolution.

Our study has several advantages in addition to its large sample size. We restricted participants to individuals of European ancestry which largely controlled for bias arising from population stratification as compared to using mixed ethnicity populations. We interrogated four common autoimmune disorders which greatly expanded pervious findings. We conducted several important sensitivity analyses to verify MR model assumptions. We selected the most significant independent SNPs identified by the largest alcohol GWAS, so all were robustly and strongly associated with exposure of interest, guaranteeing "relevance" assumption. We excluded SNPs associated with potential confounders on the exposure–outcome relationship as confirmed by GWAS catalog, to satisfy "exclusion restriction" assumption.

Nevertheless, insufficient power remains a common limitation of MR studies, because genetic variants usually explain a modest proportion of phenotypic variance. This is also a concern for alcohol consumption, a complex human behavior largely influenced by non-genetic factors. Our non-significant findings are perhaps not surprising, considering that the 99 currently reported alcohol-associated SNPs only explain ∼1% of phenotypic variance. Although improvement in the proportion of variability explained by IVs was modest, our overall statistical power was considerably raised using data from substantially augmented GWASs of four autoimmune disorders. We had 80% power at an alpha level of 0.05 to identify a ∼25–30% relative decreased risk of IBD, RA, MS, and lupus (i.e., an OR of 0.70–0.75) per SD increase in alcohol consumption. We note that most of our estimated ORs are in the expected direction, and the suggested associations for the three Th-1-mediated autoimmune diseases are in line with what have been observed previously in studies based on self-reported alcohol consumption.

Alcohol consumption plays a complicated role in human health as its effect on diseases depends on dose. In most autoimmune diseases, moderate weekly intake shows the lowest disease incidence. Such a U-shaped or J-shaped relationship cannot be identified by MR design with only summary-level data which is set out for a linear relationship. It has been proposed that a high dose of alcohol can directly suppress a wide range of immune responses (Romeo et al., 2007). We try to address this question by incorporating IVs associated with harmful drinking behavior; yet, excessive drinking does not appear to reduce the risk of autoimmune disorders. We stress caution when interpreting these results given the very few instruments available for AUDIT. Another major hypothesis for the null association is the heterogeneity of phenotypes. For example, RA can be divided into different subsets based on seropositivity. This means that even though the association is null with overall disease, signals may appear when we subtype the outcome. It is also likely that alcohol consumption, albeit with no convincing evidence to demonstrate a causal link with disease risk, may complicate symptoms or aggravate disease prognosis.

To conclude, our updated analysis, with its greatly augmented sample size and substantially improved statistical power, does not convincingly support a beneficial role of alcohol consumption in autoimmune disorder. Our findings should be interpreted with caution. Future studies may be performed to update our findings when additional alcohol-associated IVs are revealed by GWAS analysis; as well as to explore a non-linear relationship (capitalizing on individual-level data) or to understand the impact on disease prognosis.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

XJ, ZZ, and AM analyzed and interpreted the data regarding genetic correlation and mendelian randomization. LA, IK, and TO contributed significantly in writing and modifying the manuscript. All authors read and approved the final manuscript.

FUNDING

XJ was supported by a starting grant from the Swedish Research Council.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.687745/full#supplementary-material

REFERENCES

Bae, S. C., and Lee, Y. H. (2019a). Alcohol intake and risk of systemic lupus erythematosus: a Mendelian randomization study. *Lupus* 28, 174–180. doi: 10.1177/0961203318817832

Bae, S.-C., and Lee, Y. H. (2019b). Alcohol intake and risk of rheumatoid arthritis: a Mendelian randomization study. *Z. Rheumatol.* 78, 791–796. doi: 10.1007/s00393-018-0537-z

Barr, T., Helms, C., Grant, K., and Messaoudi, I. (2016). Opposing effects of alcohol on the immune system. *Prog. Neuropsychopharmacol. Biol. Psychiatry* 65, 242–251. doi: 10.1016/j.pnpbp.2015.09.001

Bentham, J., Morris, D. L., Graham, D. S. C., Pinder, C. L., Tombleson, P., Behrens, T. W., et al. (2015). Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nat. Genet.* 47, 1457–1464. doi: 10.1038/ng.3434

Bowden, J., Davey Smith, G., and Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* 44, 512–525. doi: 10.1093/ije/dyv080

Bowden, J., Davey Smith, G., Haycock, P. C., and Burgess, S. (2016). Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genet. Epidemiol.* 40, 304–314. doi: 10.1002/gepi.21965

Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., et al. (2015). An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* 47, 1236–1241. doi: 10.1038/ng.3406

Burgess, S., Butterworth, A., and Thompson, S. G. (2013). Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet. Epidemiol.* 37, 658–665. doi: 10.1002/gepi.21758

Burgess, S., Scott, R. A., Timpson, N. J., Davey Smith, G., Thompson, S. G., and Epic- InterAct Consortium (2015). Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. *Eur. J. Epidemiol.* 30, 543–552. doi: 10.1007/s10654-015-0011-z

Burgess, S., and Thompson, S. G. (2015). Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *Am. J. Epidemiol.* 181, 251–260. doi: 10.1093/aje/kwu283

Foley, C. N., Kirk, P. D. W., and Burgess, S. (2019). MR-Clust: clustering of genetic variants in Mendelian randomization with similar causal estimates. *bioRxiv* [Preprint]. doi: 10.1101/2019.12.18.881326

Hemani, G., Zheng, J., Elsworth, B., Wade, K. H., Haberland, V., Baird, D., et al. (2018). The MR-Base platform supports systematic causal inference across the human phenome. *eLife* 7:e34408. doi: 10.7554/eLife.34408

International Multiple Sclerosis Genetics Consortium (2019). Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. *Science* 365:eaav7188. doi: 10.1126/science.aav7188

Jin, Z., Xiang, C., Cai, Q., Wei, X., and He, J. (2014). Alcohol consumption as a preventive factor for developing rheumatoid arthritis: a dose-response meta-analysis of prospective studies. *Ann. Rheum. Dis.* 73, 1962–1967. doi: 10.1136/annrheumdis-2013-203323

Kranzler, H. R., Zhou, H., Kember, R. L., Vickers Smith, R., Justice, A. C., Damrauer, S., et al. (2019). Genome-wide association study of alcohol consumption and use disorder in 274,424 individuals from multiple populations. *Nat. Commun.* 10:1499. doi: 10.1038/s41467-019-09480-8

Linneberg, A., and Gonzalez-Quintela, A. (2016). The unsolved relationship of alcohol and asthma. *Int. Arch. Allergy Immunol.* 171, 155–157. doi: 10.1159/000454809

Liu, J. Z., van Sommeren, S., Huang, H., Ng, S. C., Alberts, R., Takahashi, A., et al. (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* 47, 979–986. doi: 10.1038/ng.3359

Liu, M., Jiang, Y., Wedow, R., Li, Y., Brazel, D. M., Chen, F., et al. (2019). Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat. Genet.* 51, 237–244. doi: 10.1038/s41588-018-0307-5

Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., et al. (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* 506, 376–381. doi: 10.1038/nature12873

Romeo, J., Wärnberg, J., Nova, E., Díaz, L. E., Gómez-Martinez, S., and Marcos, A. (2007). Moderate alcohol consumption and the immune system: a review. *Br. J. Nutr.* 98(Suppl. 1), S111–S115. doi: 10.1017/S0007114507838049

Smith, G. D., and Ebrahim, S. (2003). "Mendelian randomization": can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* 32, 1–22. doi: 10.1093/ije/dyg070

Wang, J., Pan, H.-F., Ye, D.-Q., Su, H., and Li, X.-P. (2008). Moderate alcohol drinking might be protective for systemic lupus erythematosus: a systematic review and meta-analysis. *Clin. Rheumatol.* 27, 1557–1563. doi: 10.1007/s10067-008-1004-z

Wang, Y.-J., Li, R., Yan, J.-W., Wan, Y.-N., Tao, J.-H., Chen, B., et al. (2015). The epidemiology of alcohol consumption and multiple sclerosis: a review. *Neurol. Sci.* 36, 189–196. doi: 10.1007/s10072-014-2007-y

Yengo, L., Sidorenko, J., Kemper, K. E., Zheng, Z., Wood, A. R., Weedon, M. N., et al. (2018). Meta-analysis of genome-wide association studies for height and body mass index in ∼700000 individuals of European ancestry. *Hum. Mol. Genet.* 27, 3641–3649. doi: 10.1093/hmg/ddy271

Zhu, T., Ye, X., Zhang, T., Lin, Z., Shi, W., Wei, X., et al. (2015). Association between alcohol consumption and multiple sclerosis: a meta-analysis of observational studies. *Neurol. Sci.* 36, 1543–1550. doi: 10.1007/s10072-015-2326-7

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

METHODS
published: 29 June 2021
doi: 10.3389/fgene.2021.634922

Identifying miRNA-mRNA Integration Set Associated With Survival Time

Yongkang Kim[1†], *Sungyoung Lee*[2,3†], *Jin-Young Jang*[4], *Seungyeoun Lee*[5] *and Taesung Park*[1,6*]

[1] Department of Statistics, Seoul National University, Seoul, South Korea, [2] Center for Precision Medicine, Seoul National University Hospital, Seoul, South Korea, [3] Department of Genomic Medicine, Seoul National University Hospital, Seoul, South Korea, [4] Department of Surgery and Cancer Research Institute, Seoul National University College of Medicine, Seoul, South Korea, [5] Department of Mathematics and Statistics, Sejong University, Seoul, South Korea, [6] Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, South Korea

In the “personalized medicine” era, one of the most difficult problems is identification of combined markers from different omics platforms. Many methods have been developed to identify candidate markers for each type of omics data, but few methods facilitate the identification of multiple markers on multi-omics platforms. microRNAs (miRNAs) is well known to affect only indirectly phenotypes by regulating mRNA expression and/or protein translation. To take into account this knowledge into practice, we suggest a miRNA-mRNA integration model for survival time analysis, called *mimi-surv*, which accounts for the biological relationship, to identify such integrated markers more efficiently. Through simulation studies, we found that the statistical power of *mimi-surv* be better than other models. Application to real datasets from Seoul National University Hospital and The Cancer Genome Atlas demonstrated that *mimi-surv* successfully identified miRNA-mRNA integrations sets associated with progression-free survival of pancreatic ductal adenocarcinoma (PDAC) patients. Only *mimi-surv* found miR-96, a previously unidentified PDAC-related miRNA in these two real datasets. Furthermore, *mimi-surv* was shown to identify more PDAC related miRNAs than other methods because it used the known structure for miRNA-mRNA regularization. An implementation of *mimi-surv* is available at http://statgen.snu.ac.kr/software/mimi-surv.

Keywords: statistical method, miRNA-mRNA integration, personalized medicine, pancreatic ductal adenocarcinoma, The Cancer Genome Atlas

OPEN ACCESS

Edited by:
Shaolong Cao,
University of Texas MD Anderson Cancer Center, United States

Reviewed by:
Wei Zhu,
Zhejiang University, China
Christine Lary,
Maine Medical Center Research Institute, United States

****Correspondence:***
Taesung Park
tspark@stats.snu.ac.kr

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *29 November 2020*
Accepted: *06 April 2021*
Published: *29 June 2021*

Citation:
Kim Y, Lee S, Jang J-Y, Lee S and Park T (2021) Identifying miRNA-mRNA Integration Set Associated With Survival Time. Front. Genet. 12:634922. doi: 10.3389/fgene.2021.634922

INTRODUCTION

MicroRNAs (miRNAs) are small, non-coding RNAs that function to regulate target messenger RNAs (mRNAs), based on sequence complementarity. It is well known that miRNAs affect nearly all developmental and pathological processes in animals, particularly in cell development, and many cancer types are affected by miRNA regulation by downregulating their target mRNAs (Ha and Kim, 2014).

Using a well-known regulation mechanism, many studies have focused on finding the target mRNAs. The biological context of regulation mechanism between miRNA and target mRNA can be easily explained by showing significant negative correlation between them and investigating their relationship with the phenotypes (Enerly et al., 2011; Xu et al., 2019). For instance, hierarchical clustering on miRNA expression profiles found that the expression levels of the tumor suppressor gene, *TP53* are associated with specific clusters (Enerly et al., 2011). When the number of target genes is small, this approach is effective. However, it is more difficult to identify novel combinations of miRNA and its target mRNAs that are concurrently associated to the phenotype.

To perform an integrated analysis of miRNA and its target mRNAs, two-step analysis has been commonly used in many studies. The first step chooses miRNAs associated with specific phenotypes. The second step further investigates expression levels of known target mRNAs that are negatively correlated with each miRNA (Enerly et al., 2011; Yonemori et al., 2017). However, this approach only focuses on the relationship between phenotypes and miRNAs without providing information about how miRNAs and their inhibited mRNAs affect observed phenotype together.

On the other hand, a hierarchical structured component analysis of miRNA-mRNA integration (*HisCoM-mimi*) has been recently proposed to investigate how miRNAs indirectly affect the phenotype with biological relationships between the miRNAs and their target mRNAs [5; 6]. *HisCoM-mimi* is a component-based method that models biological relationships as hierarchically structured "components," to efficiently identify miRNA-mRNA integration sets. *HisCoM-mimi* has an advantage of handling many types of phenotypes from an exponential family distribution under the framework of a generalized linear model. While its application to cancerous vs. normal tissues successfully identified more biologically plausible and intuitive interpretations than other methods (Kim et al., 2018), it cannot be applicable to the survival analysis which is one of prominent interest among the cancer studies.

In this study, we propose a hierarchical structured component analysis of miRNA-mRNA integration to survival phenotype, called *mimi-surv* using a Cox Proportional Hazard (Cox-PH) model (Cox, 1972; Kim, 2018; Kim et al., 2018). Like *HisCoM-mimi*, *mimi-surv* is also a component-based analysis, such as pathway models we developed for rare variant pathway analysis (Lee et al., 2016, 2019). In this respect, the proposed model introduces a latent variable for each miRNA and its target mRNAs as a component and fits one augmented model including all latent variables to determine the associations with the survival phenotype.

We applied the proposed approach, *mimi-surv*, to two real datasets from pancreatic ductal adenocarcinoma (PDAC) patients. It is noted that PDAC is one of the most lethal gastrointestinal malignancies. Despite improvements in perioperative outcomes, PDAC has a poor prognosis, with a 5-year survival rate of only 6%, worldwide (Greither et al., 2010). Because most patients are diagnosed in the advanced stages, and effective systemic therapies are lacking. Consequently, many researchers have focused on developing novel prognostic markers of PDAC. For example, several studies have identified cell-free miRNAs as prognostic markers of PDAC among which high expression of *miR-21* was shown to have a significant effect on overall survival time (Frampton et al., 2015). We considered two real PDAC datasets; one is a microarray-based dataset from PDAC patients from Seoul National University Hospital (SNUH), and the other is high-throughput sequencing data, obtained from The Cancer Genome Atlas (TCGA). From those datasets, we tried to find prognostic factors for survival after surgery of PDAC by survival analysis on integrated miRNA-mRNA sets, using *mimi-surv*.

In spite of that some prognostic miRNAs have been identified, their precise roles in the progression of PDAC have not been easy to interpret due to absence of overall grasp of vast network of miRNA-mRNA interaction. In this article, we demonstrated how well our hierarchical component-based approach can embrace such a biological concept. Moreover, the proposed *mimi-surv* was compared with many other survival analysis methods throughout the simulation studies.

MATERIALS AND METHODS

The Mimi-Surv Model

Figure 1 shows the schematic plot for *mimi-surv* model. For survival data analysis, the Cox-PH model is used (Cox, 1972). miRNA-mRNA integration set contains the miRNA, mRNA affected by the miRNA, and miRNA integration latent variable. The miRNA-mRNA integration set shows that the miRNA's direct and indirect effects on the phenotype are coming from target mRNAs. Each miRNA-mRNA integration set consists of one miRNA (z_{ij}), and mRNAs (x_{ij1}, x_{ij2}, . . ., x_{ijGj}) which were regulated by the miRNA. miRNA-mRNA integration set j is summarized by the latent variable f_{ij} which is a linear combination of z_{ij} and x_{ij1}, x_{ij2}, ..., x_{ijGj}. Thus, the effect of miRNA-mRNA integration set j on the hazard rate is computed by β_j. Detailed fitting approaches for *mimi-surv* are described as follows.

Adjusting mRNA Expression by miRNA Regulation Information

The *mimi-surv* model consists of three parts. First, the miRNA-mRNA part estimates effect of miRNA on target mRNAs. Second, the miRNA integration latent part models overall effect of each miRNA. Finally, the phenotype-latent part associates all latent variables with the target phenotype. In the miRNA-mRNA part, a simple linear combination relationship is constructed between miRNA and target mRNAs, as shown in the following Equation 1:

$$\hat{X}_{ijk} = x_{ijk} - \gamma_{jk} z_{ij}, i = 1, \cdots, N, j = 1, \cdots, J, k = 1, \cdots, G_j, \tag{1}$$

where x_{ijk} is the i^{th} individual's mRNA expression of the k^{th} gene, which is inhibited by j^{th} miRNA, z_{ij} is the i^{th} individual's j^{th} miRNA expression, γ_{jk} is the inhibition coefficient for the j^{th} miRNA for the k^{th} gene, and G_j is the number of inhibited mRNAs by the j^{th} miRNA. By estimating the miRNA inhibition coefficients γ_{jk}, the k^{th} gene's mRNA expression after adjusting the inhibition effect of the j^{th} miRNA can be obtained.

Latent Structures

The proposed *mimi-surv* models an aggregated effect of both miRNA and mRNA as a latent variable f_{ij}. As defined in Equation

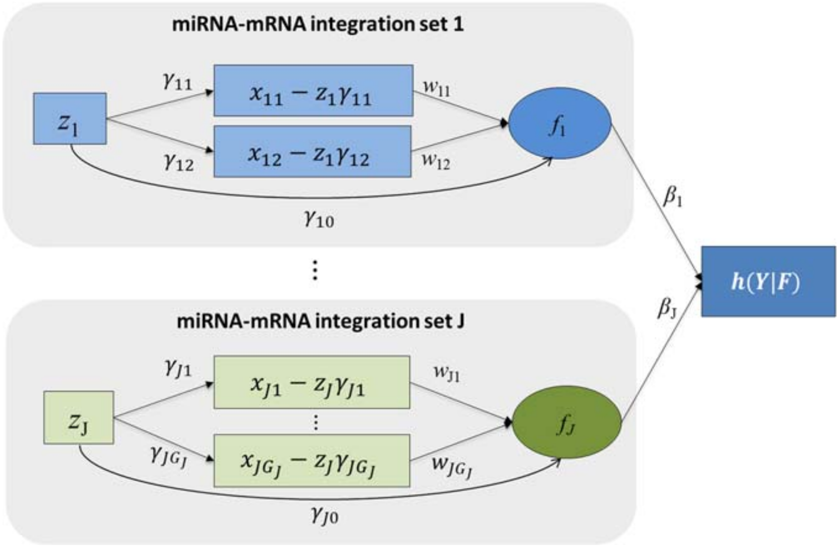


FIGURE 1 | Schematic diagram of *mimi-surv* model. Rectangles and circles indicate observed and latent variables, respectively. Arrows indicate conceptualized directions of effects between the variables. Each miRNA-mRNA integration set consists of one miRNA and its target mRNAs. Each miRNA-mRNA integration set j is summarized by the latent variable f_j which is linear combination of z_j and its adjusted mRNA expressions.

2, the latent variable f_{ij} represents the global effect of the miRNA's activity, as measured by a linear combination of both the inhibition effects (w_{jk}) of its target mRNA(s) expression and the direct effect (γ_{j0}) of the miRNA expression value.

$$f_{ij} = \gamma_{j0} z_{ij} + \sum_{k=1}^{G_j} \hat{X}_{ijk} w_{jk} \tag{2}$$

The latent variables are finally associated to the target phenotype using a Cox-PH model (Cox, 1972) as shown in Equation 3, under the assumption that the hazard rate is proportional to the risk factors over time.

$$h\left(y_i|F_i\right) = h_0\left(y_i\right) \exp\left(\sum_{j=1}^{J}\left[\gamma_{j0} z_j + \sum_{k=1}^{G_j} \hat{X}_{ijk} w_{jk}\right]\beta_j\right) = h_0\left(Y\right) \exp\left(\sum_{j=1}^{J} f_{ij}\beta_j\right), \tag{3}$$

where y_i denotes the survival time, Y denotes the vector of y_i, and $h\left(y_i \mid F\right)$ denotes the hazard function of the i^{th} sample. In addition, $h_0(Y)$ is a baseline hazard function, and β_j represents the effect of f_{ij} on the hazard rate, as a risk factor. Then, the partial likelihood function, L_p, is defined as follows:

$$L_p = \frac{\prod_{i:C_i=1} \exp\left(\sum_{j=1}^{J} f_{ij}\beta_j\right)}{\sum_{q:y_q=y_i} \exp\left(\sum_{j=1}^{J} f_{qj}\beta_j\right)},$$
$$C_i = \begin{cases} 0 & \left(i^{\text{th}} \text{ individual is censored}\right) \\ 1 & \left(i^{\text{th}} \text{ individual is deceased}\right) \end{cases} \tag{4}$$

Model Fitting

In model fitting, we estimate the parameters of *mimi-surv* by adopting the algorithm of *HisCoM-mimi* which is based on the alternating least squares (ALS) algorithm for the penalized log-likelihood function, with penalty parameters (Kim et al., 2018). In the *mimi-surv* model, the objective function to be maximized is expressed as follows:

$$\phi = \sum_{i:c_i=1}\left(\sum_{j=1}^{J} f_{ij}\beta_j - \log \sum_{q:y_q=y_i} \exp\left(\sum_{j=1}^{J} f_{qj}\beta_j\right)\right) - \frac{1}{2}\lambda_m \sum_{j=1}^{J}\sum_{k=1}^{G_j} P_{\lambda_{mm}}(w_{jk}) - \frac{1}{2}\lambda_{mm} \sum_{j=0}^{J} P_{\lambda_m}(\beta_j). \tag{5}$$

Here, the first sum consists of the partial likelihood from a Cox-PH model and the remaining term consists of two penalization parts with tuning parameters of λ_m and λ_{mm}. These two λs are so-called the tuning parameters of both the miRNA-mRNA pairs and the integrated latent components to adjust the strength of the penalty function (Cox, 1972). $P_{\lambda_{mm}}$ and P_{λ_m} denote penalty functions for w and β, respectively. Any regularization function can be used. For example, for β it can be defined as $\sum_{j=1}^{J} \beta_j^2$ for ridge, $\sum_{j=1}^{J} \left|\beta_j\right|$ for lasso, and $\left(\frac{1}{2}\sum_{j=1}^{J} \beta_j^2 + \sum_{j=1}^{J} \left|\beta_j\right|\right)$ for Elastic-Net.

We used the ALS algorithm to maximize the objective function by the two-step algorithm. The first part of the ALS algorithm is maximizing the objective function, ϕ, with the conditioning set of f_{qj}, and finding solutions for a set of β_j. The second part of algorithm is, maximizing the objective function, with a conditioning set of β_j, as calculated in the previous step, and updating the set of *f* values. Then these two steps are iterated until the solution is converged.

In the *mimi-surv* model, β_j indicates the effect size of j^{th} miRNA-mRNA integration set and w_{jk} indicates the effect size of k^{th} mRNA inhibited by j^{th} miRNA. In this study, we find the significant integrated effects of miRNA and its inhibited mRNAs, and we used *mimi-surv* to test β_j, which summarized mRNA-miRNA integration set.

We performed a simple permutation scheme to test the statistical significance of β_j and computed *p*-values and their *q*-values for the multiple testing adjustment (Ma et al., 2014). The number of permutations was set to 1,000. However, it can be increased easily to improve the accuracy of *p*-values. If one of the penalty functions is pre-specified, *mimi-surv* provides the corresponding *p*-values. However, if the choice of a penalty function is not given, *mimi-surv* can use a simple approach that picks the maximum estimate from multiple penalties, namely *maxT*. Through permutations, the null distribution of *maxT* is generated from which the *p*-value can estimated.

Comparative Models

We compared the performance of *mimi-surv* with various types of Cox-PH models, including a single miRNA Cox-PH model (single) and multiple penalized Cox-PH regression models with different penalties such as ridge, lasso, Elastic-Net (*EN*), and group lasso (*grplasso*) (Lee and Silvapulle, 1988; Tibshirani, 1996; Zou and Hastie, 2005; Meier et al., 2008)]. The objective function for multiple penalized Cox-PH model is given as follows:

$$\phi_1 = \sum_{i:c_i = 1} \left(\sum_{j=1}^{J} \delta_j z_{ij} - \log \sum_{q:y_q \leq y_i} \exp\left(\sum_{j=1}^{J} \delta_j z_{qj} \right) \right) - P_\theta\left(\delta_j\right), \tag{6}$$

where $P_\theta\left(\delta_j\right)$ denotes regularization function, which can be defined as $\theta \sum_{j=1}^{J} \delta_j^2$ for ridge,$\theta \sum_{j=1}^{J} \left|\delta_j\right|$ for lasso, and $\theta\left(\frac{1}{2}\sum_{j=1}^{J} \delta_j^2 \sum_{j=1}^{J} \left|\delta_j\right|\right)$ for *EN*. Here θ is the tuning parameter to adjust the strength of the penalty function.

For a *grplasso* Cox-PH model (Meier et al., 2008), using the group information from the miRNAs and mRNAs, the following regression model is given:

$$h(Y) = h_0(Y) \exp\left(\sum_{j=1}^{J} \delta_j z_j + \sum_{j=1}^{J} \sum_{k=1}^{G_j} \lambda_{jk} \hat{x}_{jk} \right),$$

$$\text{subject to} \left(\left|\delta_j\right| + \sum_{k=1}^{G_j} \left|\lambda_{jk}\right| \right) \geq t. \tag{7}$$

To find the optimal tuning parameter θ, we performed 10-fold cross-validation and then determined the value of θ, which minimizes the value of the objected function for the validation set.

SNUH and TCGA Datasets

The SNUH dataset consists of 95 PDAC patients in which the average of age was 65.2 years with a standard deviation 9.4 years. There were 46 male and 49 female patients. The median survival time after surgery was 795 days, which is indicated by a red vertical line in a Kaplan-Meier plot as shown in **Figure 2A**.

mRNA expression data was produced by the Human Gene 1.0 ST array (Affymetrix, Santa Clara, CA, United States). For background correction, the expression values were processed by Robust Multi-array Averaging (RMA), using the Affymetrix console, followed by quantile normalization. For the same patient, miRNA expression was obtained from the GeneChip miRNA 3.0 array (Affymetrix, Santa Clara, CA, United States). miRNA expression values were normalized by RMA, and only the human-derived miRNA targets were selected. The normalization of the background correction of the j^{th} human probe of the i^{th} sample (x_{ij}) was done using the other species' probes as background intensities as shown in Equation 8.

$$x_{ij\,(\text{norm})} = x_{ij} - \text{median}\left(x_{ij}, j \in \text{non} - \text{human miRNA}\right) \tag{8}$$

On the other hand, TCGA PDAC dataset were downloaded from the Genomic Data Commons (GDC) data portal of the U.S. National Cancer Institute[1] (Cancer Genome Atlas Research Network, Weinstein et al., 2013). To normalize mRNA-seq and miRNA-seq datasets, Fragments Per Kilobase Million (FPKM) was measured for each read count. For miRNA expression profiling, Illumina HiSeq (Illumina Inc., San Diego, CA, United States) was used. We collected 185 TCGA PDAC data sample for analysis. The read counts were log-transformed after adding a pseudo count of 0.5. In survival analysis, we excluded 25 non-PDAC samples and 47 PDAC samples whose follow-up time was less than 3 months because the cause of their deaths is not clear. After excluding these cases, we have 112 samples that consist of 48 males and 64 females. The mean age was 63.9 years with a standard deviation 11.1 years. Furthermore, the median survival time was 585 days as indicated by a red vertical line in a Kaplan-Meier plot in **Figure 2B**.

Identification of miRNA-mRNA Integration Set

For miRNA-mRNA integration analysis, we generated miRNA-mRNA integration sets which collected miRNAs and their target mRNAs satisfying two conditions as follows: (i) Reported target mRNAs by sequence-based target prediction results from TargetScan 7.1 (Agarwal et al., 2015) and (ii) significant negative correlation coefficients between miRNAs and mRNAs from SNUH dataset.

From the miRNA-mRNA pairs from TargetScan using SNUH dataset, we calculated Pearson's correlation and performed one-sided *t*-test to select the pairs with significant ($p < 0.05$) negative correlation. For those using TCGA dataset that contains many zero read counts, we first filtered out spurious pairs of miRNA-mRNA by performing one-sided *t*-test to test whether the average mRNA expression of the samples with zero miRNA read count

[1] https://portal.gdc.cancer.gov/

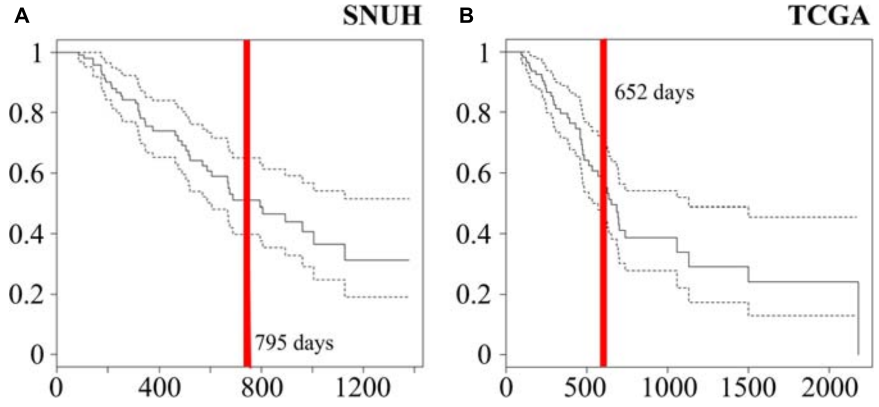


FIGURE 2 | Kaplan-Meier curves of **(A)** 95 samples of SNUH dataset and **(B)** 112 samples of TCGA dataset. Red vertical lines indicate median survival times (795 and 652 days SNUH and TCGA, respectively).

TABLE 1 | List of causal miRNAs and the numbers of target mRNAs used in simulation.

miRNA	# target mRNAs	Regulated mRNAs in SNUH data
miR-212[1,2,3]	425	PAX5, SHISA9
miR-219[1,2,3]	445	HMGA2, EGR3
miR-200b[2,3]	9	SLIT2, BNC2, CDH11
miR-32[2,3]	172	PRKAB2, SNX2
miR-362[2,3]	125	PLAT, SMAD2, CHRDL1
miR-204[3]	56	GRIN2B, HMGA2, ARNTL2, ACADL, TDRD6
miR-217[3]	449	LHX1, NR4A2, PKP1, SHOX, TRIM71, CAMK2A
miR-1297[3]	285	MCL1, RLF, RAB5IF, EDEM3
miR-496[3]	149	FLRT2, PAX6, SDHC, SERAC1, SYT5, UBXN2A
miR-670[3]	550	FRAS1, ANKRD50, LIN28B, PDE7A, SLC4A4, TP53INP1, TRIB2, CD248

[1] *miRNAs used in the simulation with two causal miRNAs.*
[2] *miRNAs used in the simulation with five causal miRNAs.*
[3] *miRNAs used in the simulation with ten causal miRNAs.*

was larger than that of the samples with non-zero miRNA read counts ($p < 0.05$). For those significant pairs, we then tested whether a correlation between target mRNAs and miRNAs was less than 0, using the samples with nonzero miRNA read counts.

Simulation Study and Real Data Analysis

To compare which method had a better power to discover the true signal miRNA-mRNA pair, we performed simulation studies to compute type I errors and power of *mimi-surv* and the compared methods, using the miRNA expression values of the SNUH PDAC dataset that consists of 64 miRNAs and 6,226 significant miRNA-mRNA pairs. Among those miRNA-mRNA pairs, we selected two, five and ten causal miRNAs to simulate phenotypes. **Table 1** lists those miRNAs and their regulated mRNAs. To generate a simulation dataset, we used the same simulation settings as we did for our previous *HisCoM-mimi* analysis (Kim et al., 2018).

We assumed a true model for generating simulated phenotype, as given in Equation 9. We considered that all causal miRNA-mRNA sets, having an effect size of β. Also, we considered regulated target mRNAs of the miRNA-mRNA sets, having the common effect size, $w_{11} = w_{1p}$, and their regulating miRNA having the effect size γ_{10}. We then considered three scenarios with different number of causal miRNAs (2, 5, and 10). For the scenario with two causal miRNAs, *miR-212* and *miR-219* were

TABLE 2 | The number of mRNAs included in the miRNA-mRNA integration set.

miRNA	# overlapped	# mRNAs (SNUH)	# mRNAs (TCGA)
miR-105	41	331	51
miR-133b	3	10	281
miR-141	28	469	37
miR-192	1	47	1
miR-200b	2	4	9
miR-200c	10	336	15
miR-206	8	50	114
miR-211	60	461	119
miR-372	7	24	207
miR-429	3	32	14
miR-488	13	43	62
miR-524	4	50	17
miR-670	2	8	131
miR-96	3	36	43

used to generate phenotypes. In the scenario with five causal miRNAs, *miR-200*, *miR-32*, *miR-362* were considered, in addition to the aforementioned two miRNAs. Lastly, five miRNAs (*miR-204*, *miR-217*, *miR-1297*, *miR-496*, *miR-670*) were additionally used in the scenario with ten causal miRNAs (see **Table 1** and section "Results"). The statistical powers were computed as the proportion of replicates whose empirical *p*-values of causal miRNAs are nonzero and significant.

$$h(Y|X,Z) = h_0(Y)\, exp\left(\beta\left(\gamma_{10}z_1 + \sum_{k=1}^{K} w_{1k}\hat{x}_k\right)\right) \tag{9}$$

In the real data analysis, to deal with the multiple testing problem, we used Benjamini-Hochberg procedure to calculate False Discovery Rate (FDR) and calculated the *q*-value. The threshold of *q*-value was set to 0.1.

RESULTS

miRNA-mRNA Pairs Extraction

We first extracted miRNA-mRNA pairs using the SNUH and TCGA datasets. For the SNUH dataset, TargetScan provided 370,075 pairs of miRNA-mRNA for 503 unique miRNAs. Our filtering strategy (see Methods) narrowed down the initial 370,075 set of pairs to 6,226 pairs that resulted in 54 unique miRNAs. For the TCGA dataset, TargetScan provided 51,014 pairs of miRNA-mRNA for 69 unique miRNAs. Unlike SNUH microarray dataset, we found that only 133 pairs of miRNA-mRNA from nine unique miRNAs were left when Pearson correlation tests were used. As noted in the Methods, the two-side filtering step resulted in 1,456 pairs with 23 unique miRNAs having at least one significant mRNA.

While two datasets showed generally concordant patterns of miRNA-mRNA selection as shown in **Table 2**, the number of mRNAs in each integration set has dataset-specific patterns. While *miR-211* integration set has the greatest number of overlapped mRNAs when combining those of SNUH and TCGA, the greatest number from each of SNUH and TCGA was *miR-141* and *miR-133b*, respectively.

Simulation Results

The simulation was conducted using the SNUH dataset with 54 miRNAs and their 6,226 miRNA-mRNA pairs, with the following parameters: two censoring fractions (δ = 0.15 and 0.3), three miRNA effect sizes (γ = 0.2, 0.3, and 0.4), three mRNA effect sizes (w = 0.5, 0.6, and 0.7). Effect of miRNA-mRNA integration set β was fixed to 1 for simplicity. The significance level α was set to 0.05. First, we estimated the type I error of each method by setting all parameters to 0 with the censoring fraction as δ. As shown in **Figure 3**, type I errors were controlled at α = 0.05 in all models, except *grplasso* (Meier et al., 2008) model which showed slightly inflated type I errors. In addition, *mimi-surv* models generally showed slightly smaller standard deviations of type I errors than the compared methods (±0.009∼0.01 for *mimi-surv*,±0.013∼0.014 for the other models). Note that the type I errors of both *mimi-surv* and the compared methods were not affected by the zero proportion of miRNA expression (zero proportion 10, 30, and 50%). In addition, we also checked an effect of penalty selection in the simulation. Since the selection of optimal penalty is challenging in Cox-PH regression (Benner et al., 2010; Ojeda et al., 2016), we applied a simple strategy that combines the three penalties by selecting the maximum of the estimates from three different penalties (lasso, ridge, and *EN*), namely *maxT*. Simulation results showed that *mimi-surv* with the proposed *maxT* approach successfully controlled type I errors

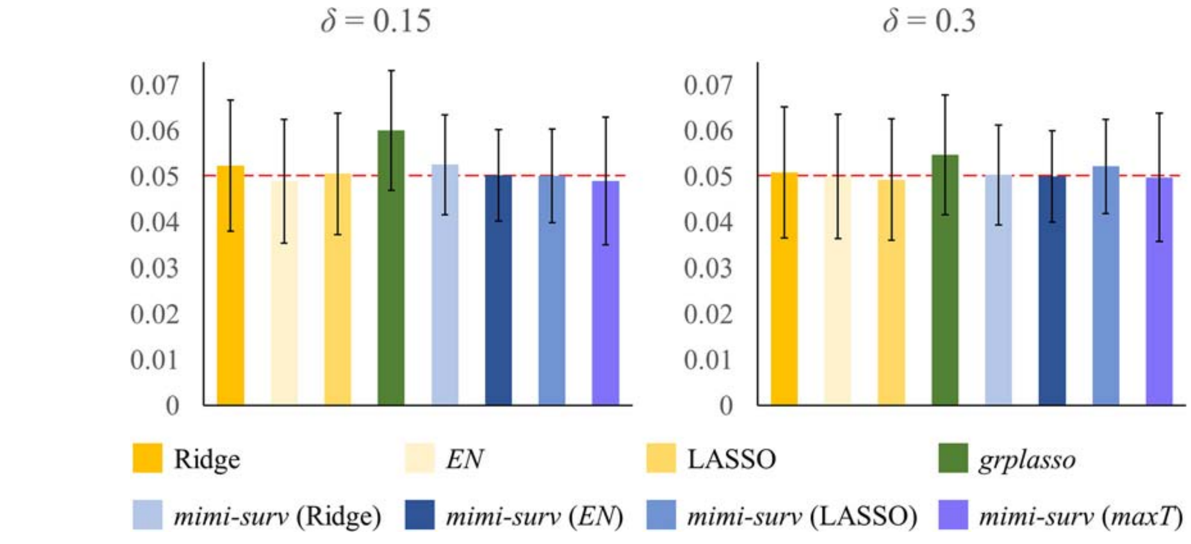


FIGURE 3 | Result of type I error evaluation. Bars indicate estimated type I error rate with given parameters (censoring fraction δ). Note that the type I errors were evaluated by fixing all parameters to 0.

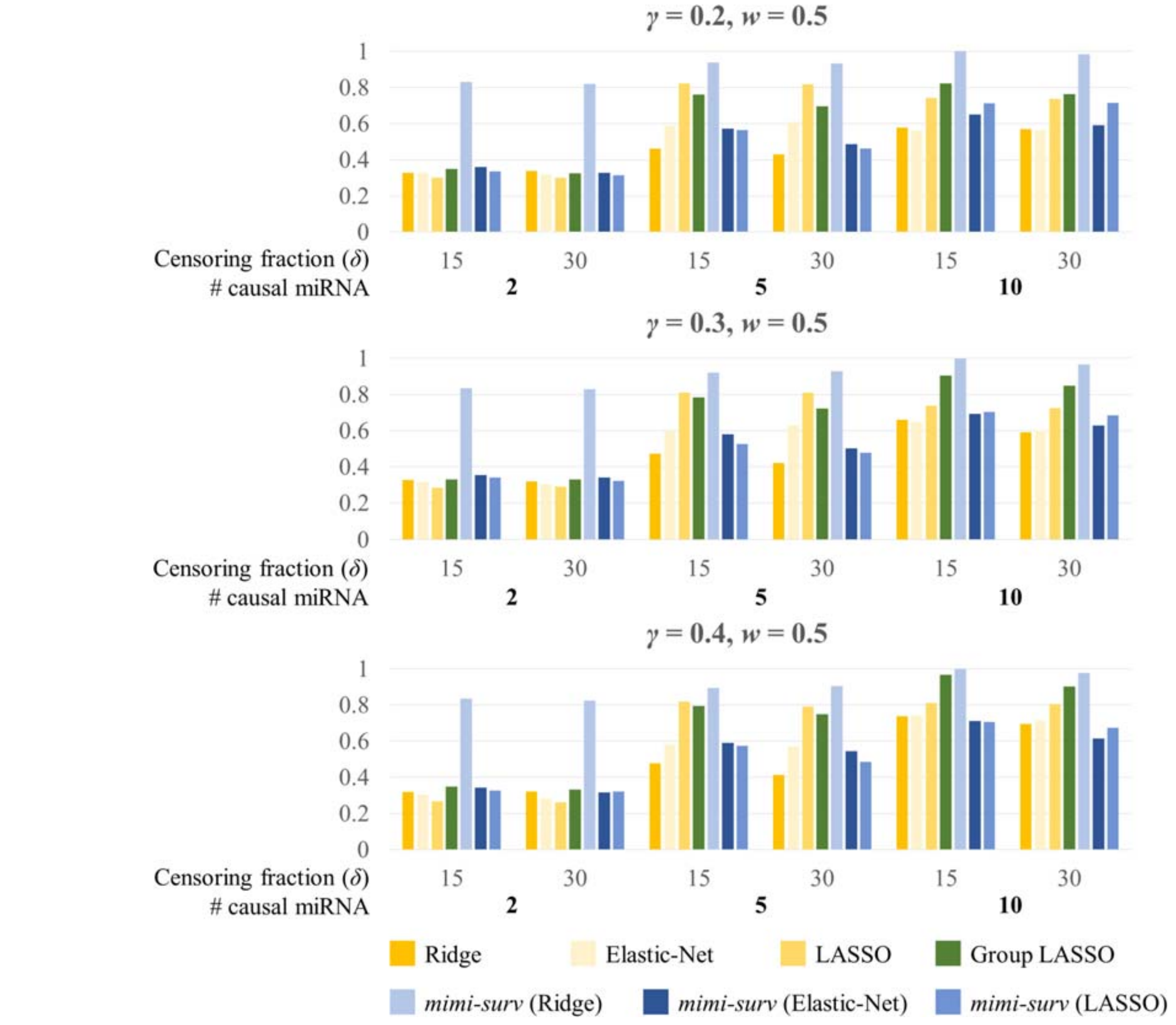


FIGURE 4 | Statistical powers of *mimi-surv* and the compared methods with different miRNA effect sizes (γ = 0.2, 0.3, and 0.4). The phenotypes were generated from two, five and ten causal miRNA-mRNA integration set and censoring fraction of 0.15 and 0.3.

with significance level of 0.05 (0.049 ± 0.014 for *mimi-surv*), as shown in **Figure 3**.

Second, we assessed the statistical powers of seven methods (*mimi-surv* with three different penalties, *grplasso*, *lasso*, *ridge*, and *EN*). Here, we generated 200 replicates of simulated phenotypes to assess the power. When variable selection methods (lasso, *EN*, *grplasso*, *mimi-surv* with lasso, and *EN* penalties) produced zero coefficients, their effects were regarded as non-significant. **Figure 4** depicts statistical powers of the compared methods with different miRNA effect sizes (0.2, 0.3, and 0.4) and two censoring fractions (0.15 and 0.3). Note that other non-causal miRNAs or mRNAs were also included to the analysis, but they actually did not contribute to the phenotypes at all. In this case, *mimi-surv* with ridge penalty and *grplasso* showed the first and second largest powers, regardless of the miRNA effect sizes. Lasso, *EN*, *mimi-surv* with *EN* and lasso penalties had smaller power than the other methods. While the powers generally increased with the miRNA effect size, their ranks vary widely (**Figure 4**). Higher censoring rate yielded generally lower power. Note that those tendencies were maintained even if γ, *w*, or the number of connected mRNAs were changed.

Figure 5 shows the barplots comparing the power of each method for a fixed miRNA effect size (γ = 0.2) and various mRNA effect sizes with censoring fractions of 0.15 and 0.3. Similarly, *mimi-surv* with ridge penalty showed the largest power. Unlike the results from **Figure 4**, *mimi-surv* with *EN* and lasso showed comparable power to *grplasso* when the number of causal miRNA increases. The same tendency was observed for various values of γ and *w*. In addition, the power differences between the results from various values of γ and *w* were small.

SNUH Dataset Analysis Result

In order to identify miRNA-mRNA integration sets, 54 miRNA-mRNA integration sets were selected to which *mimi-surv* along

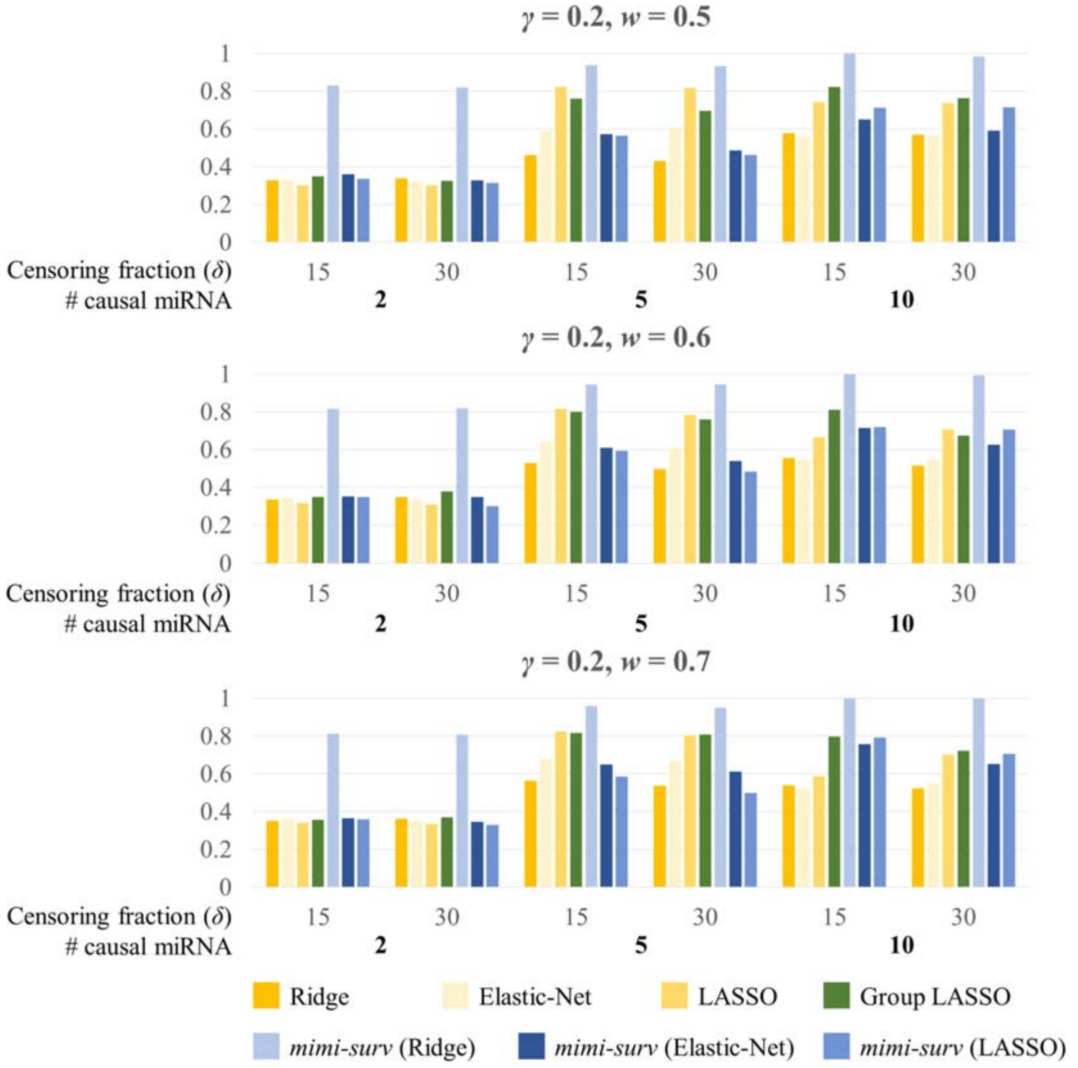


FIGURE 5 | Statistical powers of *mimi-surv* and the compared methods with different mRNA effect sizes (*w* = 0.5, 0.6, and 0.7). The phenotypes were generated from two, five and ten causal miRNA-mRNA integration set and censoring fraction of 0.15 and 0.3.

with other methods was applied to identify significant miRNA-mRNA integration sets. In this analysis, we focused on comparing the lists of significant miRNAs obtained from single, ridge, lasso, *EN*, *grplasso*, and *mimi-surv* (Lee and Silvapulle, 1988; Tibshirani, 1996; Zou and Hastie, 2005; Meier et al., 2008).

Figure 6 shows a Venn diagram displaying the number of miRNAs identified by each method, in which the number without brackets shows the number of miRNAs reported in other studies, and those within brackets show the total number of miRNAs found significant by each method. Note that the largest number of miRNAs was detected by single marker analysis. Interestingly, about half (6 out of 14) overlapped with other methods. Of these, *mimi-surv* detected a total of six miRNAs, in which four miRNAs were reported in other PDAC analyses (Ma et al., 2014; Tanaka et al., 2014; Debernardi et al., 2015; Li et al., 2015; Cheng et al., 2017). In general, the penalized Cox-PH methods identified relatively fewer miRNAs than other methods, but ridge penalty had the largest detection rate. Note that all methods commonly detected *miR-204*, which is known for the differential expression relationship between PDAC stage I and stage II-IV samples (Debernardi et al., 2015). In addition, *miR-204* has been used to distinguish solid pseudo-papillary tumors from pancreatic malignancies (Li et al., 2015).

TCGA Dataset Analysis Result and Comparison

For the analysis of TCGA data, 23 miRNA-mRNA integrations pairs were constructed. **Table 2** shows information for the miRNAs detected in the TCGA dataset analysis. For the TCGA data analysis, all the compared methods including single marker analysis and penalized regression methods failed to identify any significant miRNAs. However, *mimi-surv* detected five significant miRNAs with their significant genes, using various types of

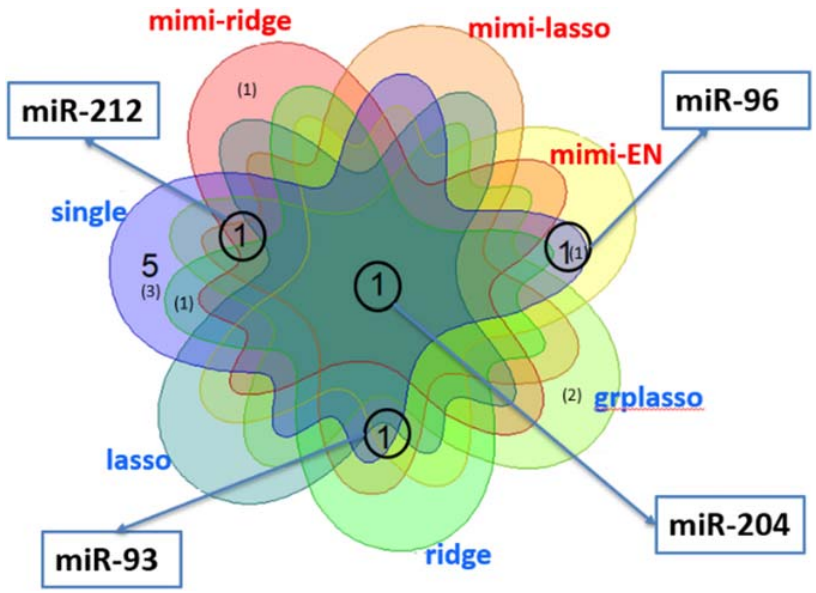


FIGURE 6 | Venn diagram for the number of miRNAs detected by each method in analysis of PDAC data from SNUH. The numbers without brackets show the numbers of miRNAs found in other PDAC analyses, while those within brackets show the number of miRNAs not previously identified.

TABLE 3 | Results of statistically significant miRNA and its significant mRNAs from both datasets using *mimi-surv*.

	miRNA	# mRNAs	# significant mRNAs (names)	β_{mimi}	p_{mimi}	q_{mimi}	*Penalty*
SNUH	*miR-204*	5	*N/A*	−0.018	0.015	0.690	Ridge
			1 (GRIN2B)	−0.179	0.004	0.221	Lasso
			1 (GRIN2B)	−0.142	0.031	0.490	*EN*
			N/A	−0.179	0.021	0.382	*maxT*
	miR-93	901	9	−0.406	0.012	0.319	Lasso
			7	−0.544	0.003	0.178	*EN*
			N/A	−0.544	0.005	0.259	*maxT*
	miR-212	2	1 (PAX5)	0.015	0.045	0.690	Ridge
			1 (PAX5)	0.008	0.033	0.601	Lasso
	miR-96	**34**	**2 (GPM6B, EPHA3)**	**0.209**	**0.017**	**0.462**	***EN***
			N/A	**0.209**	**0.020**	**0.382**	***maxT***
	miR-497	189	2 (LRRC14, PHF13)	−0.252	0.036	0.490	*EN*
			N/A	−0.252	0.046	0.620	*maxT*
	miR-339	46	*N/A*	0.024	0.045	0.690	Ridge
TCGA	*miR-133b*	281	2 (ELFN1, KCNJ12)	0.679	0.010	0.218	*EN*
			N/A	0.679	0.002	0.044	*maxT*
	miR-200c	15	2 (BASP1, LPAR1)	0.131	0.038	0.154	Lasso
			N/A	0.131	0.029	0.167	*maxT*
	miR-506	109	2 (OXSR1, RAB43)	0.023	0.040	0.249	Ridge
	miR-206	115	*N/A*	0.018	0.018	0.142	*maxT*
	miR-96	**43**	**2 (FRMD4A, SH3BP5)**	**0.419**	**0.021**	**0.244**	***EN***
			N/A	**0.419**	**0.004**	**0.046**	***maxT***

The replicated miRNA (miR-96) has embolden, and the significant mRNAs after the multiple testing adjustment (miR-96 and miR-133b) has underlined.

penalties. Among those results, we successfully replicated one miRNA *miR-96*, which was identified in the analysis of SNUH dataset. *miR-96* is a well-known marker as a suppressor of the KRAS signaling pathway (Tanaka et al., 2014). Among our detected miRNAs, *miR-200c*, *miR-506*, and *miR-96* were previously reported in other PDAC studies (Mees et al., 2010; Bryant et al., 2012; Tanaka et al., 2014; Cheng et al., 2016; Pan et al., 2018; Zhuo et al., 2018).

Table 3 lists the significant miRNAs and their significant target mRNAs detected by *mimi-surv* from both datasets. Interestingly, using the proposed *maxT* approach, *mimi-surv* successfully identified two significant miRNAs (*miR-96* and *miR-133b*) after the multiple testing adjustment (FDR q-value < 0.05), and one of those miRNAs (*miR-96*) was the replicated miRNA. In addition, our approach successfully showed the advantage of penalization approach. For instance, *miR-93* has more than 901 target mRNAs, therefore the significance level after multiple testing adjustment can be dramatically small. However, only 7 mRNAs were found significant by *EN,* and only 9 mRNAs were found significant by lasso. As a result, by using *mimi-surv*, we could reduce the number of candidate miRNA-mRNA sets.

DISCUSSION

In this study, we proposed *mimi-surv* which is a novel approach to identifying significant miRNA-mRNA sets associated with survival time, reflecting the nature of biological process between miRNA and mRNA. The objective of our analysis is to propose an integrative method for using an additional information of mRNA to the analysis of miRNA. Thus, we investigated how much the integrative analysis of miRNAs and mRNAs performs better than the other integrative methods using both miRNAs and mRNAs and the model using only miRNAs.

Through simulation studies, we compared the performance of mimi-surv, with various methods such as a single Cox-PH model, penalized Cox-PH methods with ridge, lasso, *EN* penalties and *grplasso*, including selection of optimal penalties. From the simulation results, it was shown that *mimi-surv* with ridge penalty outperformed other methods, in terms of the statistical power. The analysis of two real datasets of PDAC patients from SNUH and TCGA on which *mimi-surv* showed superior performance in identifying miRNA-mRNA integration sets for survival time. Moreover, *mimi-surv* successfully replicated one miRNA (*miR-96*) from TCGA dataset with statistical significance (q-value < 0.01), despite difference of the generation platform (Affymetrix chip vs. Illumina sequencing).

Our study remains with some limitations. First, although our simulation study based on the real SNUH dataset and simulated phenotypes showed that performance of *mimi-surv* with ridge penalty had better power than other penalties, *mimi-surv* with *maxT* approach or *EN* penalty detected more miRNAs in real PDAC data analysis. It is well known that selection of optimal penalty is challenging for Cox-PH model (Benner et al., 2010; Ojeda et al., 2016). For real data application, we recommend trying all applicable penalties to the dataset and select the penalty with less excessive shrinkage and lower dataset dependency. Although some additional simulation studies are required to evaluate performance, the *maxT* approach can be alternatively used. Finally, our permutation strategy requires an intensive computational burden to compute p-values. Thus, in future studies, we will derive a statistical distribution of the beta coefficient in *mimi-surv*, to avoid permutation procedures. Nonetheless, our *mimi-surv* remains promising for associating survival time with the expression of miRNAs and small non-coding RNAs whose misexpression is now widely accepted.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: The datasets used in this study are provided upon the approval of individual data provider. Requests to access these datasets should be directed to J-YJ, jangjy4@gmail.com.

AUTHOR CONTRIBUTIONS

YK and TP: conceptualization and methodology. SuL: software. SeL, YK, and SuL: validation. SuL and YK: formal analysis and visualization. J-YJ: resources and data curation. YK: investigation and writing—original draft preparation. SuL and TP: writing—review and editing. TP: supervision, project administration, and funding acquisition. All authors contributed to the article and approved the submitted version.

FUNDING

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, South Korea (grant number: HI16C2037010016) and Bio-Synergy Research Project of the Ministry of Science, ICT and Future Planning through the National Research Foundation (grant number: 2013M3A9C4078158). Publication of this article was sponsored by the Bio-Synergy Research Project (grant number: 2013M3A9C 4078158).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.634922/full#supplementary-material

REFERENCES

Agarwal, V., Bell, G. W., Nam, J. W., and Bartel, D. P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *Elife* 4, e05005.

Benner, A., Zucknick, M., Hielscher, T., Ittrich, C., and Mansmann, U. (2010). High-dimensional Cox models: the choice of penalty as part of the model building process. *Biom J* 52, 50–69. doi: 10.1002/bimj.200900064

Bryant, J. L., Britson, J., Balko, J. M., Willian, M., Timmons, R., Frolov, A., et al. (2012). A microRNA gene expression signature predicts response to erlotinib in epithelial cancer cell lines and targets EMT. *Br J Cancer* 106, 148–156. doi: 10.1038/bjc.2011.465

Cancer Genome Atlas Research Network, Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., et al. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* 45, 1113–1120. doi: 10.1038/ng.2764

Cheng, Y., Yang, H., Sun, Y., Zhang, H., Yu, S., Lu, Z., et al. (2017). RUNX1 promote invasiveness in pancreatic ductal adenocarcinoma through regulating miR-93. *Oncotarget* 8, 99567–99579. doi: 10.18632/oncotarget.20433

Cheng, R. F., Wang, J., Zhang, J. Y., Sun, L., Zhao, Y. R., Qiu, Z. Q., et al. (2016). MicroRNA-506 is up-regulated in the development of pancreatic ductal adenocarcinoma and is associated with attenuated disease progression. *Chin J Cancer* 35, 64.

Cox, D. R. (1972). Regression Models and Life-Tables. *J Roy Stat Soc B* 34, 187–220.

Debernardi, S., Massat, N. J., Radon, T. P., Sangaralingam, A., Banissi, A., Ennis, D. P., et al. (2015). Noninvasive urinary miRNA biomarkers for early detection of pancreatic adenocarcinoma. *Am J Cancer Res* 5, 3455–3466.

Enerly, E., Steinfeld, I., Kleivi, K., Leivonen, S. K., Aure, M. R., Russnes, H. G., et al. (2011). miRNA-mRNA integrated analysis reveals roles for miRNAs in primary breast tumors. *PloS one* 6:e16915. doi: 10.1371/journal.pone.0016915

Frampton, A. E., Krell, J., Jamieson, N. B., Gall, T. M., Giovannetti, E., Funel, N., et al. (2015). microRNAs with prognostic significance in pancreatic ductal adenocarcinoma: A meta-analysis. *Eur J Cancer* 51, 1389–1404. doi: 10.1016/j.ejca.2015.04.006

Greither, T., Grochola, L. F., Udelnow, A., Lautenschlager, C., Wurl, P., and Taubert, H. (2010). Elevated expression of microRNAs 155, 203, 210 and 222 in pancreatic tumors is associated with poorer survival. *Int J Cancer* 126, 73–80. doi: 10.1002/ijc.24687

Ha, M., and Kim, V. N. (2014). Regulation of microRNA biogenesis. *Nat. Rev. Mol. Cell Biol.* 15, 509–524.

Kim, Y., Lee, S., Choi, S., Jang, J. Y., and Park, T. (2018). Hierarchical structural component modeling of microRNA-mRNA integration analysis. *BMC Bioinformatics* 19:75.

Kim, Y. (2018). *Hierarchical Structural Component Models for Integrative Analysis of miRNA and mRNA expression data, Department of Statistics*. Seoul: Seoul National University.

Lee, S., Choi, S., Kim, Y. J., Kim, B. J., T2d-Genes Consortium, Hwang, H., et al. (2016). Pathway-based approach using hierarchical components of collapsed rare variants. *Bioinformatics* 32, i586–i594.

Lee, S., Kim, S., Kim, Y., Oh, B., Hwang, H., and Park, T. (2019). Pathway analysis of rare variants for the clustered phenotypes by using hierarchical structured components analysis. *BMC Med Genomics* 12:100.

Lee, A. H., and Silvapulle, M. J. (1988). Ridge Estimation in Logistic-Regression. *Communications in Statistics-Simulation and Computation* 17, 1231–1257.

Li, P., Hu, Y., Yi, J., Li, J., Yang, J., and Wang, J. (2015). Identification of potential biomarkers to differentially diagnose solid pseudopapillary tumors and pancreatic malignancies via a gene regulatory network. *J Transl Med* 13, 361.

Ma, C., Nong, K., Wu, B., Dong, B., Bai, Y., Zhu, H., et al. (2014). miR-212 promotes pancreatic cancer cell growth and invasion by targeting the hedgehog signaling pathway receptor patched-1. *J Exp Clin Cancer Res* 33, 54. doi: 10.1186/1756-9966-33-54

Mees, S. T., Mardin, W. A., Wendel, C., Baeumer, N., Willscher, E., Senninger, N., et al. (2010). EP300–a miRNA-regulated metastasis suppressor gene in ductal adenocarcinomas of the pancreas. *Int J Cancer* 126, 114–124. doi: 10.1002/ijc.24695

Meier, L., van de Geer, S. A., and Buhlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society Series B-Statistical Methodology* 70, 53–71. doi: 10.1111/j.1467-9868.2007.00627.x

Ojeda, F. M., Muller, C., Bornigen, D., Tregouet, D. A., Schillert, A., Heinig, M., et al. (2016). Comparison of Cox Model Methods in A Low-dimensional Setting with Few Events. *Genom Proteom Bioinf* 14, 235–243. doi: 10.1016/j.gpb.2016.03.006

Pan, Y., Lu, F., Xiong, P., Pan, M., Zhang, Z., Lin, X., et al. (2018). WIPF1 antagonizes the tumor suppressive effect of miR-141/200c and is associated with poor survival in patients with PDAC. *J Exp Clin Cancer Res* 37, 167.

Tanaka, M., Suzuki, H. I., Shibahara, J., Kunita, A., Isagawa, T., Yoshimi, A., et al. (2014). EVI1 oncogene promotes KRAS pathway through suppression of microRNA-96 in pancreatic carcinogenesis. *Oncogene* 33, 2454–2463. doi: 10.1038/onc.2013.204

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J Roy Stat Soc B Met* 58, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Xu, X., Liu, T., Wang, Y., Fu, J., Yang, Q., Wu, J., et al. (2019). miRNA-mRNA Associated With Survival in Endometrial Cancer. *Front. Genet* 10:743. doi: 10.3389/fgene.2019.00743

Yonemori, K., Kurahara, H., Maemura, K., and Natsugoe, S. (2017). MicroRNA in pancreatic cancer. *J Hum Genet* 62, 33–40.

Zhuo, M., Yuan, C., Han, T., Cui, J., Jiao, F., and Wang, L. (2018). A novel feedback loop between high MALAT-1 and low miR-200c-3p promotes cell migration and invasion in pancreatic ductal adenocarcinoma and is predictive of poor prognosis. *BMC Cancer* 18:1032.

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J Roy Stat Soc B* 67, 301–320. doi: 10.1111/j.1467-9868.2005.00503.x

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.

ORIGINAL RESEARCH
published: 02 July 2021
doi: 10.3389/fgene.2021.651546

Biomarker Categorization in Transcriptomic Meta-Analysis by Concordant Patterns With Application to Pan-Cancer Studies

Zhenyao Ye[1†], Hongjie Ke[1†], Shuo Chen[2], Raul Cruz-Cano[1], Xin He[1], Jing Zhang[1], Joanne Dorgan[2], Donald K. Milton[3] and Tianzhou Ma[1]*

[1] Department of Epidemiology and Biostatistics, School of Public Health, University of Maryland, College Park, College Park, MD, United States, [2] Department of Epidemiology and Public Health, School of Medicine, University of Maryland, Baltimore, Baltimore, MD, United States, [3] Maryland Institute for Applied Environmental Health, School of Public Health, University of Maryland, College Park, College Park, MD, United States

OPEN ACCESS

Edited by:
Chao Xu,
University of Oklahoma Health Sciences Center, United States

Reviewed by:
Lan Zhang,
University of Michigan, United States
Md Ashad Alam,
Tulane University, United States

****Correspondence:***
Tianzhou Ma
tma0929@umd.edu

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *10 January 2021*
Accepted: *28 May 2021*
Published: *02 July 2021*

Citation:
Ye Z, Ke H, Chen S, Cruz-Cano R, He X, Zhang J, Dorgan J, Milton DK and Ma T (2021) Biomarker Categorization in Transcriptomic Meta-Analysis by Concordant Patterns With Application to Pan-Cancer Studies. Front. Genet. 12:651546. doi: 10.3389/fgene.2021.651546

With the increasing availability and dropping cost of high-throughput technology in recent years, many-omics datasets have accumulated in the public domain. Combining multiple transcriptomic studies on related hypothesis via meta-analysis can improve statistical power and reproducibility over single studies. For differential expression (DE) analysis, biomarker categorization by DE pattern across studies is a natural but critical task following biomarker detection to help explain between study heterogeneity and classify biomarkers into categories with potentially related functionality. In this paper, we propose a novel meta-analysis method to categorize biomarkers by simultaneously considering the concordant pattern and the biological and statistical significance across studies. Biomarkers with the same DE pattern can be analyzed together in downstream pathway enrichment analysis. In the presence of different types of transcripts (e.g., mRNA, miRNA, and lncRNA, etc.), integrative analysis including miRNA/lncRNA target enrichment analysis and miRNA-mRNA and lncRNA-mRNA causal regulatory network analysis can be conducted jointly on all the transcripts of the same category. We applied our method to two Pan-cancer transcriptomic study examples with single or multiple types of transcripts available. Targeted downstream analysis identified categories of biomarkers with unique functionality and regulatory relationships that motivate new hypothesis in Pan-cancer analysis.

Keywords: biomarker categorization, differential expression, meta-analysis, pan-cancer, transcriptomics

INTRODUCTION

The revolutionary advancement of high-throughput technology in recent years has generated large amounts of omics data of various kinds (e.g., genetics variants, gene expression and DNA methylation, etc.), which improves our understanding of human disease and enables the development of more effective therapies in personalized medicine (Richardson et al., 2016). As more studies are conducted on a related hypothesis, meta-analysis, by combining evidence from multiple studies, has become a popular choice in genomic research to improve upon the power, accuracy, and reproducibility of individual studies (Ramasamy et al., 2008; Begum et al., 2012; Tseng et al., 2012). One of the main purposes of transcriptomics studies is to identify genes or RNAs that express differently between two or more conditions (e.g., diseased patients vs. healthy controls), also known as differential expression (DE) analysis or candidate biomarker detection. Many meta-analysis methods have been developed or applied to DE analysis, including combining p-values (Fisher, 1992) or effect sizes (Choi et al., 2003) and rank-based approaches (Hong et al., 2006). One may refer to Tseng et al. (2012) for an overview of the major meta-analysis methods in transcriptomic studies and Ma et al. (2019) for an overview of available software tools. Yet, a majority of conventional meta-analysis methods only generate a list of differentially expressed genes with strong aggregated evidence without further investigating in what studies are the genes differentially expressed.

Study or population heterogeneity always exists and has been critical to biomarker detection (Di Camillo et al., 2012). For example, The Cancer Genome Atlas (TCGA) consortium completed a Pan-Cancer Atlas of multi-platform molecular profiles spanning 33 cancer types in an effort to provide insights into the commonalities and differences across tumor lineages (Weinstein et al., 2013; Hoadley et al., 2018). When meta-analysis is performed on Pan-cancer transcriptomic studies, we expect to see both DE genes common in all tumor types as well as genes differentially expressed in some tumor types but not others. Biomarker categorization according to their DE patterns across studies is demanding in genomic studies for three reasons. First, biomarkers that share unique cross-study DE patterns are potentially involved in related functions (Berger et al., 2018). Such unique categories of genes with similar function can be used to generate new biological hypotheses. Second, biomarker categorization can make high dimensional genomic data more tractable. For example, in cancer transcriptomic studies, which frequently detect thousands of DE genes, downstream analysis methods such as pathway enrichment analysis or network analysis cannot be applied directly. By partitioning the original large set of DE genes into smaller subsets, biomarker categorization facilitates more focused downstream analysis. Third, RNA sequencing (RNA-seq) technology has led to an explosion of transcriptomic studies profiling both coding (i.e., mRNA) and noncoding RNAs (i.e., miRNA, rRNA, lncRNA, etc.) (Di Bella et al., 2020). Joint analysis of different RNA types with the same cross-study DE patterns can improve understanding of their regulatory relationships, which may lead to inferences about the underlying mechanisms of complex human diseases like cancer.

Li and Tseng (2011) first proposed an adaptively weighted Fisher (AW-Fisher) method for biomarker categorization that assigns a binary weight of 0 or 1 to each study and searches for the pattern of weights that minimizes the aggregate statistics for each gene. Though the method incorporates statistical significance by combining two-sided p-values across studies, it does not take into account the direction of regulation (e.g., up-regulated or down-regulated). Other methods incorporate biomarker categorization within the Bayesian framework and combine one-sided p-values or Bayesian posterior probabilities (Ma et al., 2017; Huo et al., 2019) but not the magnitudes of effect sizes. In practice, biological significance (i.e., large effect size) and statistical significance (i.e., small p-value) do not always occur in tandem (depending on sample size and variance) though they are equally important in interpreting study results (Sullivan and Feinn, 2012; Solla et al., 2018).

In this paper, we propose a novel meta-analysis method to detect and categorize biomarkers by simultaneously considering concordant pattern (i.e., direction of regulation), biological and statistical significance across studies. In addition, we develop a permutation test to assess the uncertainty of the proposed statistics and to control the false discovery rate (FDR). When only coding genes are included, after categorization we perform downstream pathway enrichment analysis with topological information on each category of genes for more biological insights (**Figure 1A**). In the presence of diverse RNAs, we jointly analyze all RNA species in the same category using miRNA/lncRNA target enrichment analysis and lncRNA-mRNA and miRNA-mRNA causal regulatory network analysis (**Figure 1B**). We show by simulation that our method detects both concordant and discordant biomarkers and assigns the correct weights. We apply our method to two Pan-cancer transcriptomic data examples: (1) Pan Gynecologic cancer (Pan-Gyn) data with coding genes only; (2) Pan Kidney cancer (Pan-Kidney) data that include mRNA, miRNA as well as lncRNA. The identified biomarker categories show unique functionality and informative regulatory relationships and could suggest new hypotheses about mechanisms underlying exclusive and shared features of different cancer types.

MATERIALS AND METHODS

Popular Meta-Analysis Methods

Tseng et al. (2012) reviewed the major types of meta-analysis methods for DE gene detection in microarrays and classified the methods into four main classes: combining p-values, combining effect sizes, combining ranks, and direct merging. We will discuss selected meta-analysis methods from the first two classes that are relevant to our proposed method.

Combining *P*-Values

Fisher's method (Fisher, 1992)

The conventional Fisher's method combines log transformed p-value from each study with the statistic $T_{\text{Fisher}} = -2\sum_{\text{k}=1}^{\text{K}} \log\left(p_{\text{k}}\right)$, which follows a χ^2 distribution with $2K$ degrees of freedom under the null hypothesis (i.e., genes not differentially expressed in all studies), where K is the number of studies and p_{k} is the p-value of study k, $1 \leq k \leq K$.

Stouffer's method (Stouffer, 1949)

The Stouffer's method proposes inverse normal transformation of p-value with the statistic $T_{\text{Stouffer}} \sum_{\text{k}=1}^{K} \Phi^{-1}(1-p_k)/\sqrt{K}$, which follows a standard normal distribution under the null, where $\Phi^{-1}(x)$ is the inverse cumulative distribution function of the standard normal distribution.

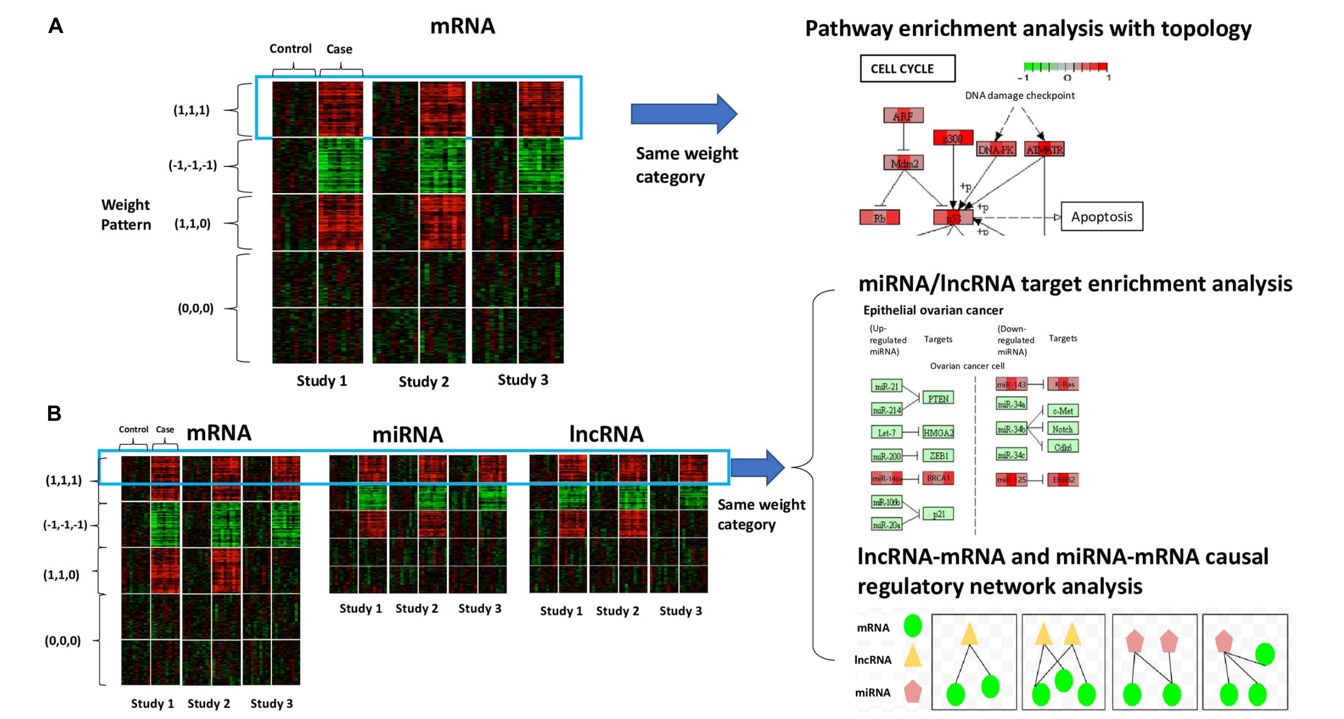


FIGURE 1 | Conceptual framework of our method. **(A)** The scenario with mRNA (or coding genes) only. The heatmap shows the gene expression of all samples from three studies. Rows refer to genes sorted by the specified weight category, columns refer to samples, and solid white lines are used to separate different conditions (control vs. case). Colors of the cells correspond to scaled expression level. The green/red indicates lower/higher expression. Pathway enrichment analysis is applied to genes belonging to the same weight category with topological information to visualize the cross-study DE patterns at the molecular level. **(B)** The scenario with diverse RNA species (e.g., mRNA, miRNA, and lncRNA). The three heatmaps show the expression of different types of transcripts of all samples from three studies, sorted by weight category. In the presence of multiple types of RNA species, we will perform integrative analysis on all the transcripts belonging to the same weight category together. Possible downstream analysis includes miRNA/lncRNA target enrichment analysis and lncRNA-mRNA and miRNA-mRNA causal regulatory network analysis.

Adaptively weighted fisher's method (AW-Fisher) (Li and Tseng, 2011)

Fisher's method does not differentiate DE in a single study or multiple studies as long as their aggregate contribution to the final statistics remains the same. To overcome this and better explain the between study heterogeneity, Li and Tseng (2011) introduced an AW-Fisher's method as a modification of the original Fisher's method. The AW-Fisher method considers $U(\overrightarrow{w}) = -2\sum_{k=1}^{K} w_k log(p_k)$ for each gene, where $\overrightarrow{w} = (w_1, \ldots, w_K)$ and each w_k is a binary weight of 0 or 1 assigned to each study k. Denote by $p\left(U(\overrightarrow{w})\right)$ the *p*-value when the weight $\overrightarrow{w}$ is given, the AW-Fisher statistic is defined as: $T_{AW} = \min_{\overrightarrow{w}} p\left(U(\overrightarrow{w})\right)$, where the optimal weight $(\widehat{w_1}, \ldots, \widehat{w_K})$ that minimizes the *p*-value indicates the subset of studies that contribute to the aggregate statistics and naturally categorizes the biomarkers. There is no closed-form distribution for AW-Fisher statistics under the null, so permutation tests and importance sampling is used to obtain the *p*-value and control the FDR.

Combining Effect Size

Fixed effect model (FEM) and random effect model (REM) (Choi et al., 2003)

Fixed effect model (FEM) combines effect sizes across all studies for each gene using a simple liner model: $T_k = \mu + \varepsilon_k,\ \ \varepsilon_k \sim N(0, s_k^2)$, where μ is the overall mean and the within-study variance s_k^2 represents the sampling error conditioned on study k. The combined point estimate of μ is a weighted average of study-specific effect sizes, where weights are equal to the inverse of s_k^2. FEM will prioritize concordant genes with the same directionality across all studies.

When strong between studies heterogeneity exists and the underlying population effect size is assumed to be unequal across studies, an REM is given hierarchically as $T_k = \theta_k + \varepsilon_k,\ \ \varepsilon_k \sim N\left(0, s_k^2\right);\ \ \theta_k = \mu + \delta_k,\ \ \delta_k \sim N(0, \tau^2)$, where between-study variance τ^2 represents the additional source of variability between studies. A homogeneity test can be performed to test whether τ^2 is zero or not, and determine the appropriateness of FEM or REM. Like FEM, REM also prioritizes concordant genes but with more flexibility across studies. Neither of FEM nor REM produces biomarker categorization results.

Remarks

P-value combination methods are powerful for detecting genes that have non-zero effects in at least one study (HS_B alternative hypothesis setting as in Chang et al. (2013) without considering the magnitudes and directionality of effects across studies. Thus, *p*-value methods cannot distinguish concordant genes (i.e., upregulated or downregulated in all studies) from discordant genes (i.e., upregulated in some studies but downregulated in others). In contrast, effect size combination methods take directionality into account but favor only concordant genes. Even so, discordant genes can still be of interest in, for example Pan-cancer analysis, to understand between tumor heterogeneity. We, therefore, propose a new meta-analysis method that incorporates both p-value and effect size combination methods, and considers concordant pattern as well as biological and statistical significance simultaneously to assist biomarker detection and categorization. Here we will introduce our method namely BCMC (**B**iomarker **C**ategorization in **M**eta-analysis by **C**oncordance).

New Meta-Analysis Method for Biomarker Detection and Categorization

Suppose there are K transcriptomic studies, each study k ($1 \leq k \leq K$) measures the gene expression of n_k samples and G genes. We use gene expression as example to introduce our method though the method is ready to analyze other types of transcripts such as miRNA and lncRNA. Our objective in meta-analysis is to detect candidate genes differentially expressed between the case (e.g., patients diagnosed with disease) and control (e.g., healthy subjects) group in multiple studies and categorize the detected genes by their DE patterns across studies. We first perform DE analysis using popular methods such as limma (Ritchie et al., 2015) for microarray or DESeq2 (Love et al., 2014) for RNA-seq in each study and obtain the summary statistics including effect size estimates (log2 fold change or LFC_{gk}) and p-values (p_{gk}) for each gene g ($1 \leq g \leq G$) in each study k. Effect sizes and p-values represent biological and statistical significance, respectively, and can be treated as DE evidence for single studies. The smaller the p-value and the larger the magnitude of effect size, the more likely a gene will be a DE gene in the study. In meta-analysis, concordance (i.e., a gene having the same sign of effect size in different studies) is regarded as additional piece of DE evidence. We define gth gene as being up-regulated in kth study when $\mathrm{LFC}_{\mathrm{gk}} > 0$ (i.e., having higher expression in case group) and being down-regulated when $\mathrm{LFC}_{\mathrm{gk}} < 0$ (i.e., having higher expression in control group).

When integrating multiple transcriptomic studies, DE genes may be altered in study-specific patterns. For example, some genes are differentially expressed in all studies while others are only differentially expressed in specific subset of studies. Meta-analysis methods also have different groups of targeted biomarkers as reflected by different statistical hypothesis settings. The null hypothesis for each gene in meta-analysis is commonly defined as: $H_0: \ \theta_{\mathrm{g1}} = \cdots = \theta_{\mathrm{gK}} = 0$, where θ_{gk} represents the true effect of gene g in study k. Depending on the types of targeted biomarkers, three alternative hypotheses have been proposed in the meta-analysis literature (Birnbaum, 1954; Tseng et al., 2012; Song and Tseng, 2014). The first setting (HS_A) aims to detect DE genes that have non-zero effect in all studies, i.e., $\theta_{\mathrm{gk}} \neq 0$ for all k. The second setting (HS_{B}) aims to detect DE genes that have non-zero effect in at least one study, i.e., $\theta_{\mathrm{gk}} \neq 0$ for some k. The third setting (HS_r) aims to detect DE genes that have non-zero effect in at least r studies, i.e., $\sum_{\mathrm{k}=1}^{K} I\left\{\theta_{\mathrm{gk}} \neq 0\right\} \geq r$. As we show next, our method generally follows HS_r setting with specifically $r = 2$ (i.e., we detect DE genes that have non-zero effect in at least two studies).

To detect DE genes and categorize them by cross-study DE patterns, we propose the following two aggregate statistics for each gene that combines DE evidence across up-regulated studies or down-regulated studies, respectively:

$$T^{+}_{\mathrm{g}(\overrightarrow{w}^{+}_{\mathrm{g}})} = \frac{\sum_{LFC_{\mathrm{gk}}>0;\ LFC_{\mathrm{gk'}}>0;\ k\neq k'} (w^{+}_{\mathrm{gk}} w^{+}_{\mathrm{gk'}} LFC_{\mathrm{gk}} LFC_{\mathrm{gk'}} \\ |log_{10} p_{\mathrm{gk}} + log_{10} p_{\mathrm{gk'}}|)}{\sum_k w^{+}_{\mathrm{gk}}}$$

$$T^{-}_{\mathrm{g}(\overrightarrow{w}^{-}_{\mathrm{g}})} = \frac{\sum_{LFC_{\mathrm{gk}}<0;\ LFC_{\mathrm{gk'}}<0;\ k\neq k'} \\ (w^{-}_{\mathrm{gk}} w^{-}_{\mathrm{gk'}} LFC_{\mathrm{gk}} LFC_{\mathrm{gk'}} |log_{10} p_{\mathrm{gk}} + log_{10} p_{\mathrm{gk'}}|)}{\sum_k w^{-}_{\mathrm{gk}}},$$

where w^{+}_{gk} and w^{-}_{gk} are binary weights of 0 or 1 assigned to the kth study for gth gene, indicating whether a study is selected for inclusion in aggregate statistics or not, +/− indicate upregulation or downregulation part, $\overrightarrow{w}^{+}_{g} = \left(w^{+}_{\mathrm{g1}}, \ldots, w^{+}_{\mathrm{gK}}\right)$ and $\overrightarrow{w}^{-}_{g} = \left(w^{-}_{\mathrm{g1}}, \ldots, w^{-}_{\mathrm{gK}}\right)$. LFC_{gk} is the $\log_2$ fold change and p_{gk} the corresponding p-value for gene g in study k obtained from single study DE analysis.

For gth gene, $T^{+}_{\mathrm{g}(\overrightarrow{w}^{+}_{\mathrm{g}})}$ aggregates the information of single study summary statistics (including both p-value and effect size) over up-regulated studies (i.e., those studies with $LFC_{\mathrm{gk}} > 0$), while $T^{-}_{\mathrm{g}(\overrightarrow{w}^{-}_{\mathrm{g}})}$ aggregates that over down-regulated studies (i.e., those studies with $LFC_{\mathrm{gk}} < 0$). The binary weights are used to indicate what studies to include to the aggregate statistics and the optimal weights that maximize the statistics will be searched for each gene. In the proposed aggregate statistics, we simultaneously account for concordant patterns (where LFC_{gk} and $LFC_{\mathrm{gk'}}$ have the same sign), biological significance (estimated as the product of LFC_{gk}) and statistical significance [estimated as the sum of $\log_{10}(p_{\mathrm{gk}})$]. This will encourage combining studies with the same directionality to find the best evidence for DE, which is consistent with the purpose of meta-analysis to identify more reproducible genes in multiple studies. Similar statistics have been proposed for concordant and discordant analysis of orthologous genes between a pair of species (Domaszewska et al., 2017). From the formula, we can see that the proposed statistic is essentially a weighted average of all study pairs with effect sizes in the same direction. A weighted average of all studies instead of study pairs is an alternative approach but it tends to exclude studies with moderate effect sizes or p-values (see a toy example in **Supplementary Table 1**).

By default, we assume $w^{+}_{\mathrm{gk}} = 0$ for studies with $LFC_{\mathrm{gk}} < 0$ and $w^{-}_{\mathrm{gk}} = 0$ for $LFC_{\mathrm{gk}} > 0$ to avoid conflict between the two statistics. When no studies are up-regulated or down-regulated for a particular gene, we suppress the corresponding $T^{+}_{\mathrm{g}(\overrightarrow{w}^{+}_{\mathrm{g}})}$ or $T^{-}_{\mathrm{g}(\overrightarrow{w}^{-}_{\mathrm{g}})}$ to zero and assign zero weights. The statistics aggregates over study pairs so we need to choose at least two studies to make it meaningful. When only one study is up-regulated or down-regulated, we also suppress the corresponding $T^{+}_{\mathrm{g}(\overrightarrow{w}^{+}_{\mathrm{g}})}$ or $T^{-}_{\mathrm{g}(\overrightarrow{w}^{-}_{\mathrm{g}})}$ to zero.

We then search for the optimal weights to identify the subset of studies that maximize each of the two aggregate statistics. Such optimal weights describe the DE patterns of each gene across studies and provide natural categorization of all genes with potential biological interpretation. The corresponding maximum statistics are defined as:

$$R^{+}_{g} = \max_{\overrightarrow{w}^{+}_{g} \in W} T^{+}_{\mathrm{g}(\overrightarrow{w}^{+}_{\mathrm{g}})};\ R^{-}_{g} = \max_{\overrightarrow{w}^{-}_{g} \in W} T^{-}_{\mathrm{g}(\overrightarrow{w}^{-}_{\mathrm{g}})},$$

where W is the pre-defined searching space of weights with aforementioned restrictions. The resulting optimal weights are denoted as $\overrightarrow{w}^{+*}_{g}$ and $\overrightarrow{w}^{-*}_{g}$. The biomarkers are then categorized according to the distribution of optimal weights among studies by merging the information of w^{+*}_{g} and w^{-*}_{g}, i.e., the final weights $\overrightarrow{w}^{*}_{g} = \overrightarrow{1} \circ \overrightarrow{w}^{+*}_{g} + \overrightarrow{-1} \circ \overrightarrow{w}^{-*}_{g}$ For example, concordantly up-regulated genes with $\overrightarrow{w}^{+*}_{g} = (0, 0, 1, 1, 1)$ and $\overrightarrow{w}^{-*}_{g} = (0,0,0,0,0)$ will be in one category $[\overrightarrow{w}^{*}_{g} = (0, 0, 1, 1, 1)]$, while concordantly down-regulated genes with $\overrightarrow{w}^{+*}_{g} = (0, 0, 0, 0, 0)$ and $\overrightarrow{w}^{-*}_{g} = (0,0,1,1,1)$ will be in the other category $[\overrightarrow{w}^{*}_{g} = (0, 0, -1, -1, -1)]$. Note that the proposed statistics can describe both up-regulated and down-regulated patterns in the same gene, thus also allowing the detection of discordant genes. In cases both patterns exist and we want to find a dominant pattern in the discordant gene, we can further define $R_{g} = \max(R^{+}_{g},\ R^{-}_{g})$ and use the corresponding $\overrightarrow{w}^{+*}_{g}$ or $\overrightarrow{w}^{-*}_{g}$ for biomarker categorization.

To assess the uncertainty of R^{+}_{g} and R^{-}_{g} and determine DE in meta-analysis, we develop a permutation-based test to calculate the p-value and FDR adjusted p-value (also known as q-value) of the statistics. We permute group labels (i.e., case or control group) in each study B times and calculate the maximum statistics in each permuted dataset. For each gene, we obtain two p-values corresponding to R^{+}_{g} and R^{-}_{g}, respectively:

$$p^{+}_{\mathrm{g}_{(\mathrm{R}^{+}_{\mathrm{g}})}} = \frac{\sum_{b=1}^{B} \sum_{\mathrm{g}'=1}^{G} I\left\{R^{+(b)}_{\mathrm{g}'} \geq R^{+}_{g}\right\} + 1}{B * G + 1};$$

$$p^{-}_{\mathrm{g}_{(\mathrm{R}^{-}_{\mathrm{g}})}} = \frac{\sum_{b=1}^{B} \sum_{\mathrm{g}'=1}^{G} I\left\{R^{-(b)}_{\mathrm{g}'} \geq R^{-}_{g}\right\} + 1}{B * G + 1},$$

where $R^{+(b)}_{\mathrm{g}'}$ and $R^{-(b)}_{\mathrm{g}'}$ are the maximum statistics for gth gene in bth $(1 \leq b \leq B)$ permutation. The value of one is added to both numerator and denominator to avoid zero p-values. After p-values are generated, we further estimate the proportion of null genes π_0 as:

$$\hat{\pi}^{+}_{0} = \frac{\sum_{g=1}^{G} I\{p^{+}_{g_{(R^{+}_{g})}} \epsilon A\}}{G * \ell(A)};\ \hat{\pi}^{-}_{0} = \frac{\sum_{g=1}^{G} I\{p^{-}_{g_{(R^{-}_{g})}} \epsilon A\}}{G * \ell(A)},$$

normally we choose $A = [0.5, 1]$ and $\ell(A) = 0.5$ to estimate the null proportion, following the guidance in the previous methods and the literature of FDR (Storey, 2002; Storey and Tibshirani, 2003; Li and Tseng, 2011). In most cases, the density of p-values beyond 0.5 is fairly flat, implying most null p-values are located in this region. In practice, depending on the problem, other common choices of A = [0.05,1] or A = [0.025,1] can also be applied. The optimal A can be empirically determined by minimizing some loss function, we do not discuss further here and refer readers to Storey (2002), Storey and Tibshirani (2003) for more details.

Then, q-values can be calculated as

$$q^{+}_{\mathrm{g}_{(\mathrm{R}^{+}_{\mathrm{g}})}} = \frac{\hat{\pi}^{+}_{0} \sum_{b=1}^{B} \sum_{\mathrm{g}'=1}^{G} I\left\{R^{+(b)}_{\mathrm{g}'} \geq R^{+}_{\mathrm{g}}\right\} + 1}{B * \sum_{\mathrm{g}'=1}^{G} I\left\{R^{+}_{\mathrm{g}'} \geq R^{+}_{\mathrm{g}}\right\} + 1},$$

$$q^{-}_{\mathrm{g}_{(\mathrm{R}^{-}_{\mathrm{g}})}} = \frac{\hat{\pi}^{-}_{0} \sum_{b=1}^{B} \sum_{\mathrm{g}'=1}^{G} I\left\{R^{-(b)}_{\mathrm{g}'} \geq R^{-}_{\mathrm{g}}\right\} + 1}{B * \sum_{\mathrm{g}'=1}^{G} I\left\{R^{-}_{\mathrm{g}'} \geq R^{-}_{\mathrm{g}}\right\} + 1}$$

Likewise, p-value and q-value of the dominant pattern statistics R_{g} (i.e., $p_{\mathrm{g}(\mathrm{R}_{\mathrm{g}})}$ and $q_{\mathrm{g}(\mathrm{R}_{\mathrm{g}})}$) can be obtained in the same way. In real data application, we determine DE in meta-analysis using the permuted p-value or q-value for the dominant pattern. Note that p-values and q-values of a zero R^{+}_{g} or R^{-}_{g} are equal to one.

Downstream Analysis on Each Identified Categories of Biomarkers

Each transcriptomic study was carefully assessed for inclusion to meta-analysis using objective criteria or systematic quality control methods (Kang et al., 2012). When only expression of mRNA data is available for the K selected transcriptomic studies, we applied our meta-analysis and identified multiple categories of mRNAs at certain BCMC p-value or q-value cutoffs, each with a unique DE pattern across the studies. DE analysis is useful to narrow down targets but focusing on single gene change across datasets is not sufficient. We still need to conduct further investigation on whether mRNAs belonging to the same category contain unifying biological theme. For each unique category of mRNAs, we then performed pathway enrichment analysis to gain more insights into their unique functions (section "Pathway Enrichment Analysis of mRNA Expression"). When expression data of mRNA, miRNA and lncRNA are all available, we applied our meta-analysis method to each type of transcripts separately and then analyzed each unique category of differentially expressed mRNA, miRNA, and lncRNA (those with the same weight or same cross-study DE pattern) together. Specifically, we performed miRNAs/lncRNAs target gene enrichment analysis (section "miRNAs/lncRNAs Target Gene Enrichment Analysis") and LncRNA-mRNA and miRNA-mRNA causal regulatory network analysis (section "LncRNA-mRNA and miRNA-mRNA Causal Regulatory Network Analysis").

Pathway Enrichment Analysis of mRNA Expression

For each category of mRNAs with unique DE pattern across the studies, we looked for biological pathways that are enriched in each category of genes more than would be expected by chance. The enriched pathways for each category can infer the unique biological functions only associated with specific study subsets and help generate new hypotheses. The *p*-value for the enrichment of a pathway was calculated using Fisher's exact test (Upton, 1992) and multiple testing was corrected by Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995). Multiple popular pathway databases were used including Gene Ontology (GO) (Ashburner et al., 2000), Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2017), Oncogenic signaling Pathways (Sanchez-Vega et al., 2018) and Reactome (Fabregat et al., 2016). Pathways in each pathway database was carefully selected for their relatedness to the problem of interest and small pathways (e.g., pathway size <10) were filtered out for the lack of power. For pathways with topological information available (e.g., pathways in KEGG), we apply the R package *"Pathview"* (Luo and Brouwer, 2013), to display the study-specific information (e.g., weights, effect sizes, etc.) on relevant pathway topology graphs.

miRNAs/lncRNAs Target Gene Enrichment Analysis

Going beyond the traditional central dogma, non-coding RNAs such as micro-RNA (or miRNA) and long non-coding RNAs (lncRNA) play important regulatory roles in mRNAs expression (Bartel, 2004; Hubé and Francastel, 2018). To understand whether miRNA/lncRNA target at mRNAs in the same category with unique cross-study DE pattern, we analyzed each unique category of mRNA, miRNA and lncRNA of the same cross-study DE pattern together and performed miRNA/lncRNAs target gene enrichment analysis on each category. Specifically, for each unique category, we first used the miRTarBase database (Chou et al., 2018) and LncRNA2Target v2.0 database (Cheng et al., 2019) to obtain common target genes of each miRNA and lncRNA in this category. We then looked for miRNA/lncRNA with target genes enriched in the gene list falling in the same category more than would be expected by chance. The *p*-value for the enrichment of miRNA/lncRNA was calculated using Fisher's exact test (Upton, 1992) and multiple testing was corrected by BH procedure (Benjamini and Hochberg, 1995).

LncRNA-mRNA and miRNA-mRNA Causal Regulatory Network Analysis

In addition to target gene enrichment analysis, we are also interested in investigating the causal regulatory relationship among the various types of transcripts in the same category using network analysis. For each unique category of mRNA and lncRNA with the same cross-study DE pattern, we followed the MSLCRN pipeline to perform module-specific lncRNA-mRNA regulatory network analysis (Zhang et al., 2019). The MSLCRN pipeline starts by using WGCNA (Langfelder and Horvath, 2008) to construct lncRNA-mRNA co-expression networks and identify modules that contain both lncRNA and mRNA. For each lncRNA-mRNA module, parallel IDA (Le et al., 2016) is then applied to learn the causal structure and estimate the causal effect of lncRNA on mRNA. IDA consists of two main steps. It first uses a parallel version of the PC algorithm (Spirtes et al., 2000; Kalisch and Bühlman, 2007; Le et al., 2016), commonly used approach for learning the causal structure of a Bayesian network, to obtain the directed acyclic graphs (DAGs) for each module. Then, the causal effect of lncRNAs on mRNAs (i.e., the lncRNA $\geq$ mRNA directed edges in the DAG) are estimated by applying do-calculus (Pearl, 2000), causal calculus that uses Bayesian conditioning to generate probabilistic formulas for the causal effect. Lastly, the module-specific causal regulatory networks are integrated to form the global lncRNA-mRNA causal regulatory network and visualized using Cytoscape (Shannon et al., 2003). In constructing the regulatory network, we use absolute values of the causal effects cutoffs to assess the regulatory strengths and confirm the regulatory relationships. More details on the use of MSLCRN to infer causal regulatory network can be found in Zhang et al. (2019). Module-specific miRNA-mRNA causal regulatory networks can be obtained in a similar way using the same tool.

SIMULATION

We conduct simulation studies to evaluate the performance of our method in biomarker detection and categorization when compared to AW-Fisher (Li and Tseng, 2011), FEM and REM methods (Choi et al., 2003). Only power is assessed for FEM and REM methods since they do not categorize biomarkers. We assume a total of $G = 2000$ genes expressed in $K = 5$ studies, each study has a total sample size of $n = 100$, evenly split into control and case groups $\left(n_{\text{case}} = n_{\text{control}} = \frac{n}{2} = 50\right)$. The details on how data are simulated are described below:

1. We generate 800 genes with 40 gene clusters (20 genes in each cluster) and another 1,200 genes that do not belong to any cluster. The cluster indexes for each gene g $\left(1 \leq g \leq 2000\right)$ is randomly sampled.
2. For genes in cluster c $(1 \leq c \leq 40)$ and study k $(1 \leq k \leq 5)$, we first generate a covariance matrix according to inverse Wishart distribution $\Sigma_{ck}^{'} \sim W^{-1}(\Psi, 60)$, where $\Psi = 0.5I_{20\times 20} + 0.5J_{20\times 20}$, I is the identity matrix and J is the matrix with all elements equal to one. Then, we standardized $\Sigma_{ck}^{'}$ into Σ_{ck} to make sure all the diagonal elements are one.
3. We sample baseline gene expression levels of the 20 genes in cluster c for sample i in study k by $\left(X_{\text{gc1ik}}^{'}, \ldots, X_{\text{gc20ik}}^{'}\right)^{T} \sim MVN(0, \Sigma_{ck})$, where $1 \leq i \leq n$ and $1 \leq k \leq K$. For those 1200 genes that are not in any cluster, we sample the baseline gene expression level independently from $N\left(0, \sigma_{\text{k}}^{2}\right)$, where $1 \leq k \leq 5$ and $\sigma_{\text{k}} \sim Unif(\sigma - 0.2, \sigma + 0.2)$ with $\sigma = 2$.
4. Denote by $\delta_{\text{gk}} \in \{0, 1, -1\}$ that gene g is non-DE, up-regulated or down-regulated in study k. We assume the first 800 genes to be DE genes divided into four mutually exclusive parts:

   (1) Concordantly up-regulated genes ($N = 225$): randomly sample $\delta_{\text{gk}} \in \{0, 1, -1\}$ such that $\sum_{\text{k}} I_{\{\delta_{\text{gk}}=1\}} \geq 2$ and $\sum_{\text{k}} I_{\{\delta_{\text{gk}}=-1\}} \leq 1$.

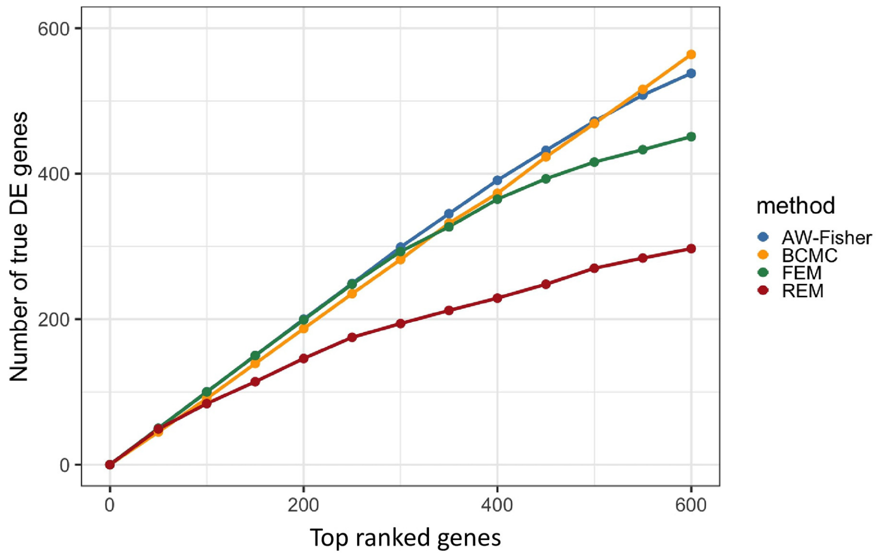


FIGURE 2 | Plot of the number of true DE genes vs. top ranked genes by *p*-value of each method.

(2) Concordantly down-regulated genes (N = 225): randomly sample $\delta_{gk} \in \{0, 1, -1\}$ such that $\sum_k I_{\{\delta_{gk}=-1\}} \geq 2$ and $\sum_k I_{\{\delta_{gk}=1\}} \leq 1$.

(3) Discordant genes with both up-regulated and down-regulated patterns (N = 150): randomly sample $\delta_{gk} \in \{0, 1, -1\}$ such that $\sum_k I_{\{\delta_{gk}=1\}} \geq 2$ and $\sum_k I_{\{\delta_{gk}=-1\}} \geq 2$.

(4) Other genes that are DE in only one study without any concordant patterns (N = 200): we randomly sample $\delta_{gk} \in \{0, 1, -1\}$ such that $\sum_k |\delta_{gk}| = 1$.

5. To simulate effect size for DE genes in each study (when $\delta_{gk} \neq 0$), we sample from a uniform distribution $\mu_{gk} \sim Unif(1,\ 3)$. The gene expression level X_{gik} are assumed to be X'_{gik} for control samples and $X_{gik} = X'_{g(i+n/2)k} + \mu_{gk} \cdot \delta_{gk}$ for case samples, where $1 \leq g \leq 2000, 1 \leq i \leq n/2$, and $1 \leq k \leq 5$.

To assess power and biomarker categorization performance, we focus on DE genes in the first three categories of genes with concordant patterns in at least two studies (N = 600). We also simulate additional scenario with smaller sample size and variance: $n = 20$ & $\sigma = 1$, results are included in the Supplement (**Supplementary Figure 1** and **Supplementary Table 2**).

Figure 2 shows the number of true DE genes detected among the top genes ranked by p-value for each method. BCMC is more powerful than AW-Fisher and FEM/REM by detecting more true DE genes among the top ranked genes. **Table 1** summarizes the number of true DE genes detected as well as with correct weight pattern in each of the three categories of DE genes identified by each method. BCMC and FEM detect more true DE genes than AW-Fisher for concordant genes. Due to the model restriction, FEM and REM fail to detect most discordant genes. AW-Fisher is equally powerful as BCMC in detecting discordant genes, however, it ignores the directionality of effects, and thus assigns the incorrect weights to genes with both up-regulated and down-regulated patterns (basically they fail to distinguish $w = -1$ from $w = 1$). Our method detects these discordant DE genes while at the same time assigns the correct weights categorizing these genes.

REAL DATA APPLICATION

Gene Expression Analysis in Pan-Gynecologic (Pan-Gyn) Studies

We applied our method to the gene expression data of TCGA Pan-Gyn studies including high-grade serous ovarian cystadenocarcinoma (OV), uterine corpus endometrial carcinoma (UCEC), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), uterine carcinosarcoma (UCS), and invasive breast carcinoma (BRCA) (Berger et al., 2018). Berger et al. (2018) identified 23 genes (e.g., BRCA1, PTEN, TP53, etc.) that were mutated at higher frequency across all Pan-Gyn cancers than non-Gyn cancers, highlighting the similarities across Pan-Gyn cohort. We focused on 19 of these genes and split samples in each study into a mutation "carrier" group and a mutation "non-carrier" group depending on whether subjects gained mutations in at least one of the genes (**Supplementary Figure 2**). Since no or very few samples were assigned to the mutation carrier group for UCS ($N_{mutation}$ = 0) and UCEC ($N_{mutation}$ = 8), we excluded those two studies and restricted our meta-analysis to only three gynecologic cancer types (i.e., number of studies K = 3) including OV (mutation carrier vs. non-carrier: 217/90), BRCA (692/408) and CESC (109/197). The purpose is to detect differentially expressed genes

TABLE 1 | Summary of number of true DE genes detected and with correct weight patterns by the four methods in each of the three categories of DE genes described in the simulation setting.

Methods	**BCMC**		**AW-Fisher**		**FEM**	**REM**
DE Gene categories	**Number of true DE genes**	**Number of true DE genes with correct weight**	**Number of true DE genes**	**Number of true DE genes with correct weight**		
Concordant up (*N* = 225)	206	116	195	106	203	151
Concordant down (*N* = 225)	210	119	195	108	201	144
Discordant (*N* = 150)	148	135	148	0	47	2
Total (*N* = 600)	564	370	538	214	451	297

between mutation carrier and non-carrier groups and categorize them according to their cross-study DE patterns. We found the overall survival differed significantly between the two groups for each cancer type (**Supplementary Figures 3–5**). This implied the differentially expressed biomarkers between these two groups can have potential prognostic values related to mutational processes and serve as optimal therapeutic intervention targets (Helleday et al., 2014; Lawrence et al., 2014).

The RNA-seq data in Transcripts Per Million (TPM) values of each cancer type were downloaded from LinkedOmics (Vasaikar et al., 2018). We first merged the three datasets by matching the gene symbols and removed genes with mean TPM < 5. A total of 9,900 mRNAs remained and were $\log_2$ transformed for analysis. We performed DE analysis by limma (Ritchie et al., 2015) and obtained the *p*-value and LFC from each of the three studies. We then performed meta-analysis using BCMC and the other methods.

All methods detected thousands of DE genes at both *q*-value cutoffs (for BCMC, *q*-value for dominant pattern was used so we focused on concordant genes only), which is common in Pan-cancer studies (**Table 2**). It becomes imperative task to partition these DE genes into smaller subsets by cross-study DE patterns before performing downstream analysis. BCMC categorized these DE biomarkers ($q < 0.05$) into eight groups according to the optimal weight assignments, each displaying a unique expression pattern across the different studies (**Figure 3** and **Supplementary Table 3**). We then merged genes with equal $|\vec{w}_{\mathrm{g}}^{*}|$ into the same group (i.e., genes with $\vec{w}_{\mathrm{g}}^{*} = (0, 1, 1)$ and those with $\vec{w}_{\mathrm{g}}^{*} = (0, -1, -1)$ are merged into the same group, allowing both up-regulated and down-regulated genes in the same pathway) and performed pathway enrichment analysis on each of the four merged groups using four pathway databases: GO (Ashburner et al., 2000), KEGG (Kanehisa et al., 2017), Oncogenic (Sanchez-Vega et al., 2018) and Reactome (Fabregat et al., 2016). The top 100 pathways enriched by each category have little overlap partly validating our speculation in motivation that the different categories of biomarkers may play different functional roles (**Figure 4**). For example, top pathways for $|\vec{w}_{\mathrm{g}}^{*}| = (1, 0, 1)$ (i.e., DE in OV and CESC but not in BRCA) are mainly involved in cell junction and adhesion related functions (**Supplementary Table 4** in **Supplemental File 1**). Top pathways for $|\vec{w}_{\mathrm{g}}^{*}| = (1, 1, 0)$ (i.e., DE in OV and BRCA but not in CESC) are mainly involved in immune and defense response. **Figure 5** shows the topology of one example KEGG pathway "Antigen processing and presentation" enriched by the genes with $|\vec{w}_{\mathrm{g}}^{*}| = (1, 1, 0)$. The highlighted DE genes showed strong DE signals (signed LFC) in OV and BRAC but not in CESC. These genes colocalized and interacted with each other as a functional unit inside the pathway.

TABLE 2 | Summary of numbers of DE genes detected by each method at different cutoffs for the Pan-Gyn study example. For BCMC, *q*-values for the dominant pattern are used.

Methods				
***q*-value**	**BCMC**	**AW-Fisher**	**FEM**	**REM**
$q < 0.05$	1,345	3,113	2,866	983
$q < 0.15$	3,931	4,743	4,342	1,641

These unique gene sets of different cross-cancer DE patterns and the associated pathways enriched help gain more insights into the homogeneous and heterogenous molecular mechanism of different Gynecologic cancer and assist the development of useful diagnostic and therapeutic strategies common or specific to cancer types. Understanding commonality and difference in drug targets can also guide the drug repurposing strategy in cancer drug development (Li et al., 2021).

Integrative Analysis of mRNA, lncRNA, and miRNA in Pan-Kidney Studies

We also used BCMC to perform integrative analysis of three different types of transcripts (mRNA, lncRNA, and miRNA) in the TCGA Pan-Kidney cohort including kidney chromophobe (KICH), kidney renal clear cell carcinoma (KIRC), and kidney renal papillary cell carcinoma (KIRP). LncRNA and miRNA have been found playing important regulatory roles on gene expression in kidney cancers (Linehan et al., 2010; Linehan, 2012; Ricketts et al., 2018). The integrative analysis of these multi-omics data provides additional insights into the biological mechanism underlying the multiple histologic subtypes of kidney cancers. We aimed to detect the differentially expressed biomarkers (mRNA, miRNA, or lncRNA) that drive the progression of kidney cancer by comparing samples from early pathologic stage (stage I and II) to late stage (stage III and stage IV) for three kidney cancer types (i.e., number of studies $K = 3$) and investigating the regulatory relationships among these biomarkers. Number of subjects in the two pathologic stages of each kidney cancer available in mRNA, miRNA and lncRNA expression data were summarized in **Supplementary Table 5**.

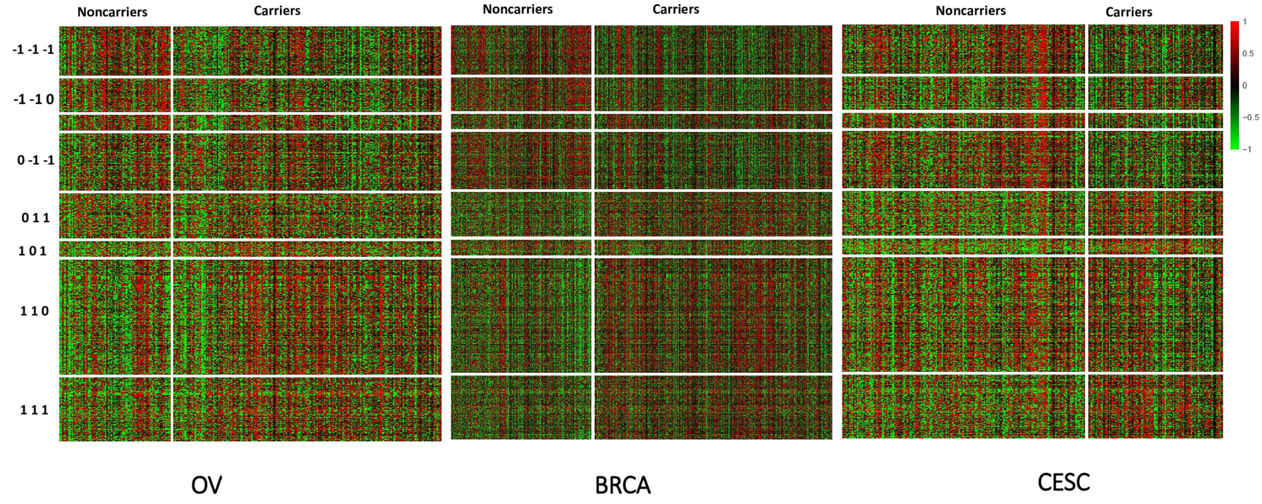


FIGURE 3 | Heatmap of standardized expression values of differentially expressed genes (BCMC $q < 0.05$) sorted by weight patterns for the Pan-Gyn cancer example.

We downloaded mRNA (in Reads Per Kilobase of transcript per Million mapped reads or RPKM) and miRNA (in Reads Per Million mapped reads or RPM) sequencing data from LinkedOmics (Vasaikar et al., 2018) and lncRNA sequencing data (in RPM) from The Atlas of Noncoding RNAs in Cancer (TANRIC) (Li et al., 2015) for all the three kidney cancer subtypes. We first merged the three subtypes by matching RNA symbols/IDs. We then separately filtered each of the three types of biomarkers by removing mRNAs with mean RPKM < 5, lncRNAs with mean RPM < 0.1, and miRNAs with mean RPM $= 0$, followed by $\log_2$ transformation. A total of 15,332 mRNAs, 2,415 lncRNAs and 719 miRNAs remained for analysis. We performed DE analysis by limma (Ritchie et al., 2015) in each study and then meta-analysis to categorize biomarkers according to cross-study DE patterns for each RNA species. For different types of RNA belonging to the same category, we further performed miRNA target gene enrichment analysis and lncRNA-mRNA causal regulatory network analysis to understand their complex interacting relationships in kidney cancer.

Both BCMC and AW-Fisher methods detected thousands of differentially expressed biomarkers (including mRNA, lncRNA, and miRNA) at both q-value cutoffs with high proportion of overlap (**Table 3**). Biomarkers detected by BCMC tend to have both significant p-values and large effect sizes in the studies indicated by optimal weights (**Supplementary Figure 6**). These biomarkers ($q < 0.05$) were partitioned into eight categories by different weight patterns (**Supplementary Table 6**). We merged biomarkers with the same $|\vec{w}_g^*|$ into the same group. We focused on the group with $|\vec{w}_g^*| = (1, 1, 1)$ to understand the common multi-omics regulatory among all histologic subtypes of kidney cancer and performed downstream analysis. In miRNA target gene enrichment analysis, we found the target gene sets of two DE miRNAs "miR-655" and "miR-326" were enriched in the DE gene list in the same group ($p < 0.05$; **Supplementary Table 7** in the **Supplementary File 1**), implying the potential regulatory relationship between different biomarker types consistent in all kidney cancer subtypes. The gene *ATAD2* targeted by miR-655 was reported as a prognostic marker for kidney disease (Chen et al., 2017). In causal network analysis, we identified two lncRNA-mRNA regulatory networks (**Supplementary Figure 8** and **Supplementary Table 8**). **Figure 6** shows the network with two hub lnRNAs, the hub lncRNA ENSG00000267449 and several mRNAs belonging to the ribosomal protein family in the same network were found consistently differentially expressed in all three subtypes, implying their potentially joint role in promoting the development of kidney cancers (Zhou et al., 2015; Dolezal et al., 2018).

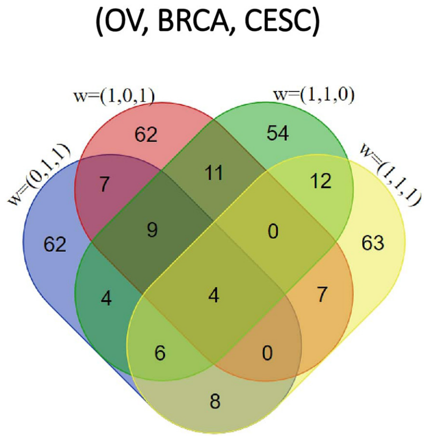


FIGURE 4 | Venn diagram of top 100 pathways enriched by each of the four categories [$|w_g^*| = (0, 1, 1), (1, 0, 1), (1, 1, 0)$, and $(1, 1, 1)$; corresponding to OV, BRCA and CESC, respectively] for the Pan-Gyn study example.

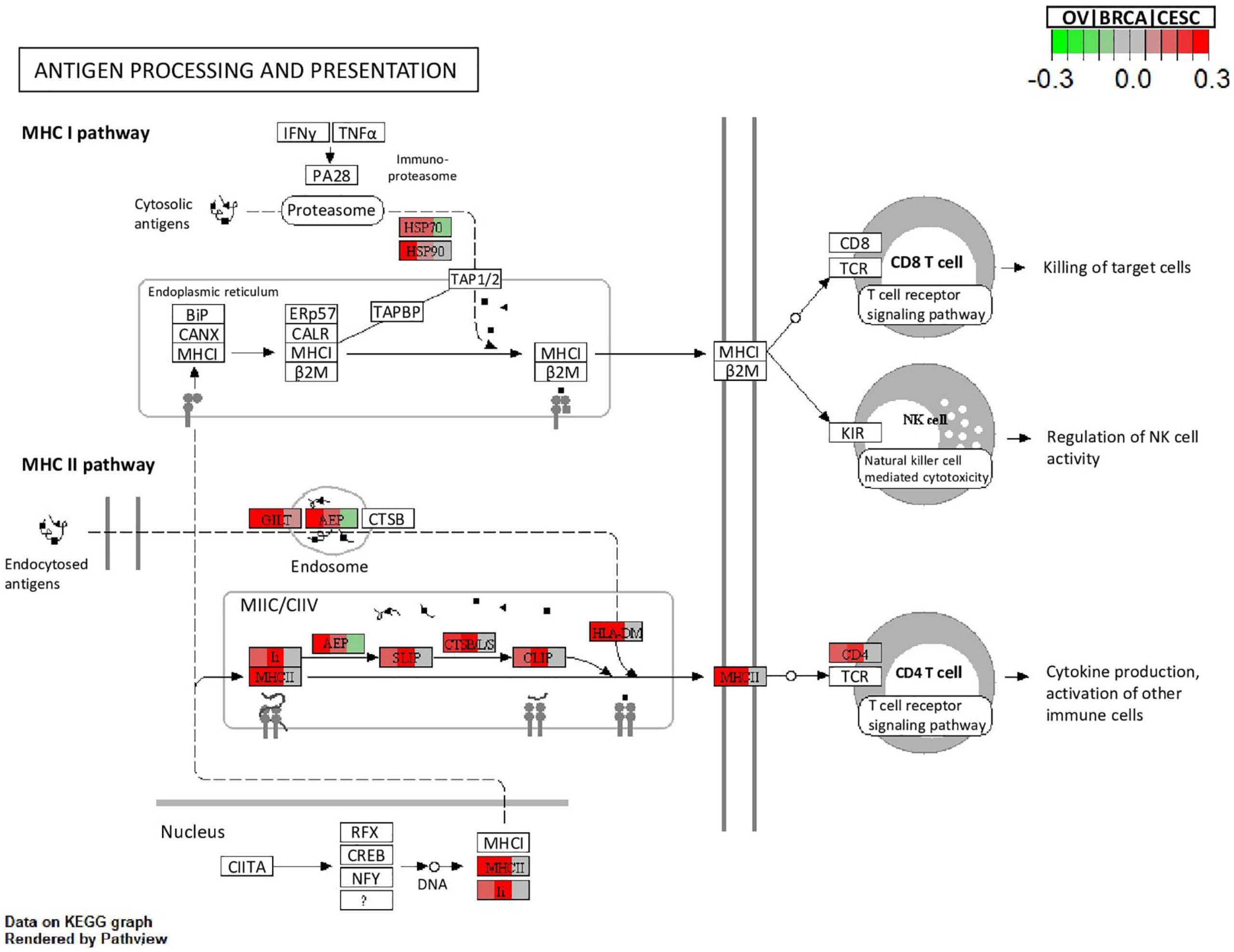


FIGURE 5 | Visualization of the topology plot of a KEGG pathway "Antigen processing and presentation" enriched by the genes with $|w_g^*| = (1, 1, 0)$ (corresponding to OV, BRCA and CESC, respectively) for the Pan-Gyn example. Each box that represents a gene is split into three parts to represent the three studies. Colors indicate the signed LFC of the mapped DE genes in the three studies.

These results demonstrate the power of our method to detect biomarkers of different types in Pan-cancer meta-analysis and to categorize them into functionally relevant biomarkers by DE patterns, which could suggest commonalities and differences in underlying mechanisms of multiple cancer types.

DISCUSSION

In this paper, we proposed a novel meta-analysis method for candidate biomarker detection in multiple transcriptomic studies that further categorizes biomarkers by concordant patterns as well as by biological and statistical significance across studies.

TABLE 3 | Summary of number of differentially expressed biomarkers among each of the three RNA species detected by each method at different cutoffs for the Pan-Kidney study example. For BCMC, q-values for the dominant pattern are used.

Type of biomarkers	mRNA		lncRNA		miRNA	
q-value	BCMC	AW-Fisher	BCMC	AW-Fisher	BCMC	AW-Fisher
$q < 0.05$	7,317	9,472	764	1,281	239	283
Intersection	6,391		622		206	
$q < 0.15$	11,810	11,440	1,468	1,464	358	358
Intersection	10,057		1,244		292	

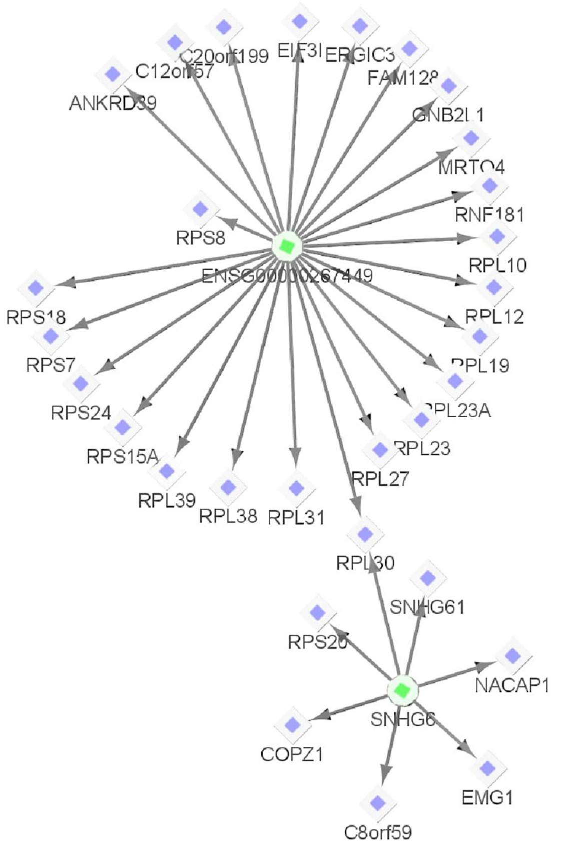


FIGURE 6 | One example lncRNA-mRNA regulatory network identified from biomarkers with $|w_g^*| = (1, 1, 1)$ (corresponding to KICH, KIRC, and KIRP, respectively) for the Pan-Kidney example. The circle shapes represent lncRNAs highlighted in green and diamond shapes represent mRNAs highlighted in purple. The arrows indicate the network relationships between lncRNAs and mRNAs.

Numerous downstream analysis tools including pathway analysis and causal network analysis are applied to each category of biomarkers with either single or multiple types of RNA species. Simulations and real data application to two Pan-cancer multi-omics studies showed the advantage of our method in classifying differentially expressed biomarkers into classes with unique biological functions and relationships that can be further investigated in future studies.

Meta-analysis is a set of statistical analytical methods and tools that combine multiple related studies to improve power and reproducibility over a single study. In recent years, we have witnessed the development of many useful meta-analysis methods applied to genomic studies for different biological purposes (Choi et al., 2003; Shen and Tseng, 2010; Li and Tseng, 2011; Huo et al., 2016, 2020; Kim et al., 2016, 2018; Zhu et al., 2017; Ma et al., 2019; Zeng et al., 2020). Genomic data is usually of high dimension and the between study heterogeneity is large due to both technological and cohort effects. In addition to improving power, post-hoc categorization of biomarkers into smaller subsets by cross-study patterns for subsequent analysis is important in genomic meta-analysis. Our meta-analysis method that aggregates over both *p*-value and effect size is a fast and intuitive solution for this purpose. Compared to other popular meta-analysis methods that include biomarker categorization, our method considers concordant pattern, and biological and statistical significance simultaneously. By calculating statistics separately for up-regulated and down-regulated parts, we can detect both concordant genes that have consistent patterns across all studies and discordant genes that are up/down regulated in some studies while down/up regulated in others. Both of these kinds of genes can be of interest in Pan-cancer analysis. For example, high expression of some genes might worsen the prognosis of all cancer types, while high expression of other genes might worsen prognosis for some cancers but be beneficial to other cancer types.

Our method also applies to the scenario when there is more than one RNA species present and proposes to jointly analyze different types of biomarkers under the same category for more biological insights. As more omics data are accumulated in the public domain, similar strategies can be applied for integrative analysis, for example with epigenomic (e.g., DNA methylation, histone modification), proteomic and metabolomic data. Unique features of each omics data type need to be addressed and will be considered as a future direction to extend our method.

Like most other two-stage meta-analysis methods, our method is based on summary measures such as *p*-values and $\log_2$ fold changes from each study. In addition, the method assigns a single optimal weight to each gene without quantifying the uncertainty in weight assignment. A more comprehensive Bayesian hierarchical model can be applied to raw data and summary measures to better capture the stochasticity and provide soft weight assignment. Our method requires the DE genes to be concordant in at least two studies to be detected, consistent with the purpose of meta-analysis in prioritizing more reproducible biomarkers. As the number of studies becomes large, the likelihood of being differentially expressed in only one study decreases. Thus, we expect the method to perform well as the number of studies increases. Since the method relies on summary measures, increasing the number of studies will not materially increase the computational burden. Additionally, use of more sophisticated parallel computing techniques will improve the speed of permutation tests. An R package called "BCMC" is available at https://github.com/kehongjie/BCMC to implement our method.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/kehongjie/BCMC.

AUTHOR CONTRIBUTIONS

ZY and HK developed the method, performed the analysis, and wrote the manuscript. TM supervised the project and took the lead in editing the manuscript. SC, RC-C, XH, JZ, JD, and DM contributed to manuscript writing and polishing. All authors provided critical feedback and helped shape the research, analysis and manuscript.

FUNDING

Research reported in this publication was supported by the National Institute on Drug Abuse (NIDA) of National Institute of Health under award number 1DP1DA048968-01 to SC and TM, and the University of Maryland Faculty-Student Research Award (FSRA) to ZY, HK, and TM.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.651546/full#supplementary-material

REFERENCES

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.

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116, 281–297.

Begum, F., Ghosh, D., Tseng, G. C., and Feingold, E. (2012). Comprehensive literature review and statistical considerations for GWAS meta-analysis. *Nucleic Acids Res.* 40, 3777–3784. doi: 10.1093/nar/gkr1255

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B (Methodol.)* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Berger, A. C., Korkut, A., Kanchi, R. S., Hegde, A. M., Lenoir, W., Liu, W., et al. (2018). A comprehensive pan-cancer molecular study of gynecologic and breast cancers. *Cancer Cell* 33, 690–705.e9.

Birnbaum, A. (1954). Combining independent tests of significance. *J. Am. Stat. Assoc.* 49, 559–574. doi: 10.2307/2281130

Chang, L.-C., Lin, H.-M., Sibille, E., and Tseng, G. C. (2013). Meta-analysis methods for combining multiple expression profiles: comparisons, statistical characterization and an application guideline. *BMC Bioinform.* 14:368. doi: 10.1186/1471-2105-14-368

Chen, D., Maruschke, M., Hakenberg, O., Zimmermann, W., Stief, C. G., and Buchner, A. (2017). TOP2A, HELLS, ATAD2, and TET3 are novel prognostic markers in renal cell carcinoma. *Urology* 102:265.e1–265.e7.

Cheng, L., Wang, P., Tian, R., Wang, S., Guo, Q., Luo, M., et al. (2019). LncRNA2Target v2. 0: a comprehensive database for target genes of lncRNAs in human and mouse. *Nucleic Acids Res.* 47, D140–D144.

Choi, J. K., Yu, U., Kim, S., and Yoo, O. J. (2003). Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics* 19, i84–i90.

Chou, C.-H., Shrestha, S., Yang, C.-D., Chang, N.-W., Lin, Y.-L., Liao, K.-W., et al. (2018). miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res.* 46, D296–D302.

Di Bella, S., La Ferlita, A., Carapezza, G., Alaimo, S., Isacchi, A., Ferro, A., et al. (2020). A benchmarking of pipelines for detecting ncRNAs from RNA-Seq data. *Brief. Bioinform.* 21, 1987–1998. doi: 10.1093/bib/bbz110

Di Camillo, B., Sanavia, T., Martini, M., Jurman, G., Sambo, F., Barla, A., et al. (2012). Effect of size and heterogeneity of samples on biomarker discovery: synthetic and real data assessment. *PLoS One* 7:e32200. doi: 10.1371/journal.pone.0032200

Dolezal, J. M., Dash, A. P., and Prochownik, E. V. (2018). Diagnostic and prognostic implications of ribosomal protein transcript expression patterns in human cancers. *BMC Cancer* 18:275. doi: 10.1186/s12885-018-4178-z

Domaszewska, T., Scheuermann, L., Hahnke, K., Mollenkopf, H., Dorhoi, A., Kaufmann, S. H., et al. (2017). Concordant and discordant gene expression patterns in mouse strains identify best-fit animal model for human tuberculosis. *Sci. Rep.* 7:12094.

Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., et al. (2016). The reactome pathway knowledgebase. *Nucleic Acids Res.* 44, D481–D487.

Fisher, R. A. (1992). "Statistical methods for research workers," in *Breakthroughs in Statistics*, eds S. Kotz and N. L. Johnson (New York, NY: Springer), 66–70.

Helleday, T., Eshtad, S., and Nik-Zainal, S. (2014). Mechanisms underlying mutational signatures in human cancers. *Nat. Rev. Genet.* 15, 585–598. doi: 10.1038/nrg3729

Hoadley, K. A., Yau, C., Hinoue, T., Wolf, D. M., Lazar, A. J., Drill, E., et al. (2018). Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* 173, 291–304.e6.

Hong, F., Breitling, R., McEntee, C. W., Wittner, B. S., Nemhauser, J. L., and Chory, J. (2006). RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics* 22, 2825–2827. doi: 10.1093/bioinformatics/btl476

Hubé, F., and Francastel, C. (2018). Coding and non-coding RNAs, the frontier has never been so blurred. *Front. Genet.* 9:140. doi: 10.3389/fgene.2018.00140

Huo, Z., Ding, Y., Liu, S., Oesterreich, S., and Tseng, G. (2016). Meta-analytic framework for sparse k-means to identify disease subtypes in multiple transcriptomic studies. *J. Am. Stat. Assoc.* 111, 27–42. doi: 10.1080/01621459.2015.1086354

Huo, Z., Song, C., and Tseng, G. (2019). Bayesian latent hierarchical model for transcriptomic meta-analysis to detect biomarkers with clustered meta-patterns of differential expression signals. *Ann. Appl. Stat.* 13:340.

Huo, Z., Tang, S., Park, Y., and Tseng, G. (2020). P-value evaluation, variability index and biomarker categorization for adaptively weighted Fisher's meta-analysis method in omics applications. *Bioinformatics* 36, 524–532. doi: 10.1093/bioinformatics/btz589

Kalisch, M., and Bühlman, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *J. Mach. Learn. Res.* 8, 613–636.

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, D353–D361.

Kang, D. D., Sibille, E., Kaminski, N., and Tseng, G. C. (2012). MetaQC: objective quality control and inclusion/exclusion criteria for genomic meta-analysis. *Nucleic Acids Res.* 40:e15. doi: 10.1093/nar/gkr1071

Kim, S., Kang, D., Huo, Z., Park, Y., and Tseng, G. C. (2018). Meta-analytic principal component analysis in integrative omics application. *Bioinformatics* 34, 1321–1328. doi: 10.1093/bioinformatics/btx765

Kim, S., Lin, C.-W., and Tseng, G. C. (2016). MetaKTSP: a meta-analytic top scoring pair method for robust cross-study validation of omics prediction analysis. *Bioinformatics* 32, 1966–1973. doi: 10.1093/bioinformatics/btw115

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. S., Stojanov, P., Mermel, C. H., Robinson, J. T., Garraway, L. A., Golub, T. R., et al. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* 505, 495–501. doi: 10.1038/nature12912

Le, T., Hoang, T., Li, J., Liu, L., Liu, H., and Hu, S. (2016). "A fast PC algorithm for high dimensional causal discovery with multi-core PCs," in *Proceedings of the IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 16, (New York, NY: IEEE), 1483–1495. doi: 10.1109/tcbb.2016.2591526

Li, J., Han, L., Roebuck, P., Diao, L., Liu, L., Yuan, Y., et al. (2015). TANRIC: an interactive open platform to explore the function of lncRNAs in cancer. *Cancer Res.* 75, 3728–3737. doi: 10.1158/0008-5472.can-15-0273

Li, J., and Tseng, G. C. (2011). An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *Ann. Appl. Stat.* 5, 994–1019.

Li, Y., Dong, Y.-P., Qian, Y.-W., Yu, L.-X., Wen, W., Cui, X.-L., et al. (2021). Identification of important genes and drug repurposing based on clinical-centered analysis across human cancers. *Acta Pharmacol. Sin.* 42, 282–289. doi: 10.1038/s41401-020-0451-1

Linehan, W. M. (2012). Genetic basis of kidney cancer: role of genomics for the development of disease-based therapeutics. *Genome Res.* 22, 2089–2100. doi: 10.1101/gr.131110.111

Linehan, W. M., Srinivasan, R., and Schmidt, L. S. (2010). The genetic basis of kidney cancer: a metabolic disease. *Nat. Rev. Urol.* 7:277. doi: 10.1038/nrurol.2010.47

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550.

Luo, W., and Brouwer, C. (2013). Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics* 29, 1830–1831. doi: 10.1093/bioinformatics/btt285

Ma, T., Huo, Z., Kuo, A., Zhu, L., Fang, Z., Zeng, X., et al. (2019). MetaOmics: analysis pipeline and browser-based software suite for transcriptomic meta-analysis. *Bioinformatics* 35, 1597–1599. doi: 10.1093/bioinformatics/bty825

Ma, T., Liang, F., and Tseng, G. (2017). Biomarker detection and categorization in ribonucleic acid sequencing meta-analysis using bayesian hierarchical models. *J. R. Stat. Soc. Ser. C Appl. Stat.* 66:847. doi: 10.1111/rssc.12199

Pearl, J. (2000). *Causality: Models, Reasoning and Inference*, Vol. 9. Cambridge, MA: Cambridge university press, 10–11.

Ramasamy, A., Mondry, A., Holmes, C. C., and Altman, D. G. (2008). Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Med.* 5:e184. doi: 10.1371/journal.pmed.0050184

Richardson, S., Tseng, G. C., and Sun, W. (2016). Statistical methods in integrative genomics. *Annu. Rev. Stat. Appl.* 3, 181–209. doi: 10.1146/annurev-statistics-041715-033506

Ricketts, C. J., De Cubas, A. A., Fan, H., Smith, C. C., Lang, M., Reznik, E., et al. (2018). The cancer genome atlas comprehensive molecular characterization of renal cell carcinoma. *Cell Rep.* 23, 313–326.e5.

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

Sanchez-Vega, F., Mina, M., Armenia, J., Chatila, W. K., Luna, A., La, K. C., et al. (2018). Oncogenic signaling pathways in the cancer genome atlas. *Cell* 173, 321–337.e10.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Shen, K., and Tseng, G. C. (2010). Meta-analysis for pathway enrichment analysis when combining multiple genomic studies. *Bioinformatics* 26, 1316–1323. doi: 10.1093/bioinformatics/btq148

Solla, F., Tran, A., Bertoncelli, D., Musoff, C., and Bertoncelli, C. M. (2018). Why a p-value is not enough. *Clin. Spine Surg.* 31, 385–388.

Song, C., and Tseng, G. C. (2014). Hypothesis setting and order statistic for robust genomic meta-analysis. *Ann. Appl. Stat.* 8:777.

Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. (2000). *Causation, Prediction, and Search*. Cambridge, MA: MIT press.

Storey, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* 64, 479–498. doi: 10.1111/1467-9868.00346

Storey, J. D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A.* 100, 9440–9445. doi: 10.1073/pnas.1530509100

Stouffer, S. (1949). A study of attitudes. *Sci. Am.* 180, 11–15.

Sullivan, G. M., and Feinn, R. (2012). Using effect size-or why the P value is not enough. *J. Graduate Med. Educ.* 4, 279–282. doi: 10.4300/jgme-d-12-00156.1

Tseng, G. C., Ghosh, D., and Feingold, E. (2012). Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Res.* 40, 3785–3799. doi: 10.1093/nar/gkr1265

Upton, G. J. (1992). Fisher's exact test. *J. R. Stat. Soc. Ser. A (Stat. Soc.)* 155, 395–402.

Vasaikar, S. V., Straub, P., Wang, J., and Zhang, B. (2018). LinkedOmics: analyzing multi-omics data within and across 32 cancer types. *Nucleic Acids Res.* 46, D956–D963.

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. doi: 10.1038/ng.2764

Zeng, X., Zong, W., Lin, C.-W., Fang, Z., Ma, T., Lewis, D. A., et al. (2020). Comparative pathway integrator: a framework of meta-analytic integration of multiple transcriptomic studies for consensual and differential pathway analysis. *Genes* 11:696. doi: 10.3390/genes11060696

Zhang, J., Le, T. D., Liu, L., and Li, J. (2019). Inferring and analyzing module-specific lncRNA–mRNA causal regulatory networks in human cancer. *Brief. Bioinform.* 20, 1403–1419. doi: 10.1093/bib/bby008

Zhou, X., Liao, W.-J., Liao, J.-M., Liao, P., and Lu, H. (2015). Ribosomal proteins: functions beyond the ribosome. *J. Mol. Cell Biol.* 7, 92–104. doi: 10.1093/jmcb/mjv014

Zhu, L., Ding, Y., Chen, C.-Y., Wang, L., Huo, Z., Kim, S., et al. (2017). MetaDCN: meta-analysis framework for differential co-expression network detection with an application in breast cancer. *Bioinformatics* 33, 1121–1129.

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: frontiersin.org/about/contact

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