MACHINE LEARNING AND NETWORK-DRIVEN INTEGRATIVE GENOMICS

EDITED BY: Mehdi Pirooznia, Richard S. Lee and Shizhong Han
PUBLISHED IN: Frontiers in Genetics and Frontiers in Plant Science

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88966-725-3
DOI 10.3389/978-2-88966-725-3

About Frontiers

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

Frontiers Journal Series

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

Dedication to Quality

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

What are Frontiers Research Topics?

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

MACHINE LEARNING AND NETWORK-DRIVEN INTEGRATIVE GENOMICS

Topic Editors:
Mehdi Pirooznia, National Heart, Lung, and Blood Institute (NHLBI), United States
Richard S. Lee, Johns Hopkins University, United States
Shizhong Han, Johns Hopkins Medicine, United States

Citation: Pirooznia, M., Lee, R. S., Han, S., eds. (2021). Machine Learning and Network-Driven Integrative Genomics. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88966-725-3

Table of Contents

frontiers in Genetics

EDITORIAL
published: 11 March 2021
doi: 10.3389/fgene.2021.660201

Editorial: Machine Learning and Network-Driven Integrative Genomics

Mehdi Pirooznia[1,2]*, *Shizhong Han*[2,3] *and Richard S. Lee*[3]

[1] Bioinformatics and Computational Biology Laboratory, National Heart, Lung, and Blood Institute National Institutes of Health, Bethesda, MD, United States, [2] Department of Psychiatry and Behavioral Sciences, Johns Hopkins School of Medicine, Baltimore, MD, United States, [3] Lieber Institute for Brain Development, Johns Hopkins Medical, Baltimore, MD, United States

Keywords: genomics, integrative analysis, systems biology, neural-network, machine learning

Editorial on the Research Topic

Machine Learning and Network-Driven Integrative Genomics

Rapid advances in high-throughput technologies have produced distinct biomedical data sets that can be analyzed using mathematical and statistical models including network science tools to decode interactions among functional molecules in living cells. Availability of data and analysis tools was critical in forming the foundation for complex networks. In the past decade, since the birth of this discipline, a robust conceptual framework known as network biology has emerged. Understanding the dimension and dynamic properties of biological data, including gene-gene and protein-protein interactions, and metabolic networks and pathways, can help elucidate the functional properties of cells, which will eventually assist further in understanding their development and disease dynamics. Machine learning (ML), on the other hand, can handle heterogeneous data in different ways such as naive Bayesian Network data integration, Tree-Based Methods such as Random Forest, and penalized linear models such as LASSO. ML-based omics analyses provide assorted integrative analysis of multiple omics data, by analyzing different omics layers together. The discipline of Network biology is rapidly emerging with most recent applications to personalized medicine. Despite great success, there remain many technical challenges, one of which is how to integrate or transform subject-specific knowledge in order to adapt to deep-learning (DL) algorithms and improve outcomes. Technical hurdles exist in data preprocessing, model selection, parametric function approximation, and model regularization and optimization. This Research Topic addresses these challenges and hurdles with a specific focus on the application of DL algorithms to disease prediction and diagnosis, which has not been adequately explored.

As summarized below, this collection of original research papers presents a significant amount of progress made in the above-mentioned scope of the Research Topic:

CL-PMI identifies pre-miRNA using neural network. In their study, Wang, Ma et al. proposed a pre-miRNA identification algorithm based on a cascaded CNN-LSTM framework, called CL-PMI. They used a convolutional neural network (CNN) and employed long short-term memory (LSTM) to automatically extract features and obtain the sequential and spatial characteristics of pre-miRNAs and capture time characteristics of pre-miRNAs to improve attention mechanisms for long-term dependence modeling. Their method overcomes the dataset imbalance problem and improves the performance of pre-miRNA identification methods.

Inferring Bayesian network using genetic node ordering. In order to study the impact of genetic variations on gene regulatory networks, Wang, Audenaert et al. proposed an alternative method for inferring high-quality Bayesian gene networks. Their method, which is easily scalable to thousands of genes, first constructs a node ordering by conducting pairwise

OPEN ACCESS

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

***Correspondence:**
Mehdi Pirooznia
mehdi.pirooznia@nih.gov

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

Received: *29 January 2021*
Accepted: *22 February 2021*
Published: *11 March 2021*

Citation:
Pirooznia M, Han S and Lee RS (2021) Editorial: Machine Learning and Network-Driven Integrative Genomics. Front. Genet. 12:660201. doi: 10.3389/fgene.2021.660201

causal inference tests between genes and then allows the user to infer a Bayesian network via a series of independent variable selection problems. In addition to higher sensitivity, this method allows for a unified false discovery rate control across genes and individual edges, and therefore provides a suitable way for tuning the sparsity level of the inferred network.

Identification of genes involved in Fetal Growth Restriction (FGR) by in-depth strategy combining methylomics and transcriptomics analyses. Chabrun et al. performed a rigorous multi-omics approach by combining methylomics and transcriptomics analyses on 36 placenta samples in a case-control study to study pathogenic mechanisms of FGR. Data-mining algorithms were used to combine the analysis of more than 1,200 significantly expressed and/or methylated genes. They used machine learning models to explore the phenotypic subgroups (premature birth, birth weight, and head circumference) associated with FGR allowing for a better description of the FGR pathophysiology as well as key genes involved.

A web server to predict Hepatocellular carcinoma (HCC) (Kaur et al.). This study employed large-scale transcriptomic profiling datasets containing a total of 2,316 HCC and 1,665 non-tumorous tissues obtained from 30 studies. They identified a panel of three genes (FCN3, CLEC1B, and PRC1) as a HCC biomarker using different feature selection techniques. The three-genes-based HCC biomarker identified HCC samples in training/validation datasets with an accuracy between 93 and 98%. Furthermore, the prognostic potential of these genes was evaluated on TCGA-LIHC and GSE14520 cohorts using univariate survival analysis. They also developed a web server HCCpred based on the above study to disseminate their tool to the scientific community.

CRISPR/Cas9 Guide RNA Activity Prediction. In order to accurately predict guide RNA (gRNA) on-target efficacy, Zhang et al. proposed CNN-SVR, a novel hybrid system that combines an improved convolutional neural network (CNN)-based method with support vector regression (SVR). The CNN-SVR system is composed of two major components, a merged CNN as the front-end for extracting gRNA features and an SVR as the back-end for regression and predicting gRNA cleavage efficiency. The authors showed that CNN-SVR can effectively learn deeper features of gRNAs and their corresponding epigenetic features, which outperforms available methods in terms of prediction accuracy, generalization, and robustness.

Developing novel computational methods for the inference of novel biological relations from multi-layered networks (Lee, Zhang et al.). Despite advances in analysis, data mining and knowledge discovery of high-dimensional multi-omics biological data remain a great challenge due to the complexity, heterogeneity, and high-dimensionality inherent in the omics data. Network has been widely used to represent relationships among entities in biological systems. In their review, the authors first discuss the properties of biological heterogeneous multi-layered network (HMLN), then surveyed four categories of state-of-the-art methods, namely matrix factorization, random walk, knowledge graph, and deep learning, and demonstrated their applications to omics data integration and analysis.

Infer the regulatory pathway from mixed observational data. In a new approach Zhong et al. presented a Mixed Directed Acyclic Graph (mDAG) algorithm and R package to infer the regulatory pathway from mixed observational data containing both continuous variables such as gene expression and categorical variables such as phenotypes or single nucleotide polymorphisms. Through extensive simulations and real data analysis, they demonstrated that the mDAG method can identify upstream causal factors and downstream effectors linked to a variable and generate hypotheses for causal direction of regulatory pathways capable of recovering a large sparse DAG with limited sample size.

A Network-based functional omics analysis server (Lee, Lee et al.). Cultivated barley is one of the most produced cereal crops worldwide and an important crop species in plant genetics, because it harbors numerous stress response alleles in its genome that can be exploited for crop engineering. In order to study the functional annotation of its genome, Lee, Lee et al. developed the BarleyNet, a co-functional network of 26,145 barley genes, along with a web server for network-based predictions of biological processes. BarleyNet has three complementary network-based algorithms for prioritizing genes to study genetic components of complex traits such as response to environmental stress: a pathway-centric search for candidate genes of pathways or complex traits; a gene-centric search to infer novel functional concepts for genes; and a context-centric search for novel genes associated with stress response to facilitate understanding of the underlying genetic components of complex traits in barley.

Predicting Autism risk genes via machine leaning approaches (Lin et al.). In order to predict Autism spectrum disorder (ASD) risk genes, the authors employed a machine learning-based approach using features from spatiotemporal gene expression patterns in the human brain, gene-level constraint metrics, and other gene variation features. They performed gene ontology enrichment analysis on these predicted risk genes that not only revealed relevant biological processes to ASD such as neuronal signaling, neurogenesis, and chromatin remodeling, but also highlighted other potential mechanisms that might underlie ASD, such as regulation of RNA alternative splicing and ubiquitination pathway related to protein degradation. They demonstrated that human brain spatiotemporal gene expression patterns and gene-level constraint metrics can help predict ASD risk genes.

Gene regulatory network inference methodologies. In their review (Van den Broeck et al.), the authors described experimental methodologies commonly used to identify regulatory interactions and generate gene regulatory networks (GRNs), which provide a blueprint of transcriptional regulations underlying development and environmental responses, including network topology, network size, and transient binding of transcription factors (TFs) to DNA. Additionally, they reviewed network inference techniques that leverage gene expression data to predict regulatory interactions that can identify new regulatory interactions and drive novel hypotheses. They also highlighted the potential of machine learning approaches to leverage gene expression data to predict phenotypic outputs.

A hybrid Approach for microbiome network inferences. Jiang et al. proposed a general framework, HARMONIES, Hybrid Approach foR MicrobiOme Network Inferences via Exploiting Sparsity, to infer a sparse microbiome network from datasets that are often high-dimensional and suffer from uneven sampling depth, over-dispersion, and zero-inflation. HARMONIES utilizes a zero-inflated negative binomial (ZINB) distribution to model the skewness and excess zeros in the microbiome data as well as incorporate a stochastic process prior to sample-wise normalization. This allows inferring a sparse and stable network by imposing non-trivial regularizations based on the Gaussian graphical model. They showed that HARMONIES can outperform other commonly used methods and discover a novel community of disease-enriched bacteria.

AUTHOR CONTRIBUTIONS

MP proposed and edited this Research Topic. SH and RL co-edited this Research Topic. All authors made a substantial, direct and intellectual contribution to this Editorial, and approved it for publication.

FUNDING

This work was in part funded by National Institutes of Health Grant Numbers: R01 MH121394 and AA024486 to SH and by the Division of Intramural Research, NHLBI, NIH, Grant Number: 1-ZIC-HL006228-04 to MP.

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: 11 October 2019
doi: 10.3389/fgene.2019.00967

CL-PMI: A Precursor MicroRNA Identification Method Based on Convolutional and Long Short-Term Memory Networks

Huiqing Wang[1], Yue Ma[1], Chunlin Dong[2], Chun Li[1], Jingjing Wang[1] and Dan Liu[1]*

[1] College of Information and Computer, Taiyuan University of Technology, Taiyuan, China, [2] Dryland Agriculture Research Center, Shanxi Academy of Agricultural Sciences, Taiyuan, China

MicroRNAs (miRNAs) are the major class of gene-regulating molecules that bind mRNAs. They function mainly as translational repressors in mammals. Therefore, how to identify miRNAs is one of the most important problems in medical treatment. Many known pre-miRNAs have a hairpin ring structure containing more structural features, and it is difficult to identify mature miRNAs because of their short length. Therefore, most research focuses on the identification of pre-miRNAs. Most computational models rely on manual feature extraction to identify pre-miRNAs and do not consider the sequential and spatial characteristics of pre-miRNAs, resulting in a loss of information. As the number of unidentified pre-miRNAs is far greater than that of known pre-miRNAs, there is a dataset imbalance problem, which leads to a degradation of the performance of pre-miRNA identification methods. In order to overcome the limitations of existing methods, we propose a pre-miRNA identification algorithm based on a cascaded CNN-LSTM framework, called CL-PMI. We used a convolutional neural network to automatically extract features and obtain pre-miRNA spatial information. We also employed long short-term memory (LSTM) to capture time characteristics of pre-miRNAs and improve attention mechanisms for long-term dependence modeling. Focal loss was used to improve the dataset imbalance. Compared with existing methods, CL-PMI achieved better performance on all datasets. The results demonstrate that this method can effectively identify pre-miRNAs by simultaneously considering their spatial and sequential information, as well as dealing with imbalance in the datasets.

Keywords: pre-miRNA identification, long short-term memory network, convolutional neural network, deep learning, imbalanced learning

OPEN ACCESS

Edited by:
Mehdi Pirooznia, National Heart, Lung, and Blood Institute (NHLBI), United States

Reviewed by:
Min Chen, Hunan Institute of Technology, China
Xiaoyong Pan, Ghent University, Belgium
André Lamúrias, University of Lisbon, Portugal

***Correspondence:**
Yue Ma
867937098@qq.com

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

Received: *01 May 2019*
Accepted: *10 September 2019*
Published: *11 October 2019*

Citation:
Wang H, Ma Y, Dong C, Li C, Wang J and Liu D (2019) CL-PMI: A Precursor MicroRNA Identification Method Based on Convolutional and Long Short-Term Memory Networks. Front. Genet. 10:967. doi: 10.3389/fgene.2019.00967

INTRODUCTION

MicroRNAs (miRNAs) are ribonucleic acid molecules of about 21–23 nucleotides that are widely found in microorganisms, viruses (Pfeffer et al., 2004), and plants (Jones-Rhoades et al. 2006). They are known to regulate thousands of human genes that account for more than one-third of the genomic coding region (Bentwich et al., 2005). miRNAs also have important roles in the pathogenesis and treatment of cancer (Wang et al., 2010; Jansson and Lund, 2012; Tüfekci, et al., 2014; Zhu et al., 2014). A study has shown that 50% of miRNAs frequently appear in tumor-associated gene regions or fragile

sites such as homozygous deletion regions, heterozygous deletion regions, amplification regions, and breakpoint regions, as well as in proximity to tumor suppressor genes and the locations of oncogenes, indicating a correlation between the localization of miRNAs on human chromosomes and tumorigenesis (Calin et al., 2004). In addition, miRNAs are potential targets for disease markers and therapeutic drugs (Schmidt, 2014), for instance, they guide the RNA-induced silencing complex to degrade or inhibit mRNA translation by pairing with bases of the target gene mRNA, thereby regulating protein expression at the post-transcriptional level (research has shown that miRNAs can also regulate gene expression at the transcriptional level). Therefore, how to identify miRNAs is a key question with implications for medical treatment. miRNAs exist in many forms; the most primitive of these is primary miRNA, which becomes precursor miRNA (pre-miRNA) after single processing. The pre-miRNA is digested by Dicer to form a mature miRNA (Agarwal et al., 2010). It is difficult to identify mature miRNAs owing to their short length; thus, most previous studies have focused on identifying pre-miRNAs.

pre-miRNA identification is a binary classification task requiring the input of a given set to be classified into two groups, producing precursors and non-precursors as the output. A large number of computational methods for identifying miRNAs have been proposed; these can be divided into experimental cloning and computer simulation prediction methods (Bartel, 2004; Jones-Rhoades et al., 2006). Experimental methods are recognized as the gold standard for miRNA identification; however, it is impossible to discover all miRNAs through experimental cloning because of the small number of discoveries and the specific development time or specific tissue expression. Computer simulation methods can be used to obtain reliable predictions and reduce the cost of research and production, and they have been proven to effectively detect miRNAs expressed in specific tissues (Bartel, 2004). Among the available computer simulation prediction methods for the identification of miRNAs, rule-based methods (Mathelier and Carbone, 2010) and machine learning methods have been widely applied. These include microPred (Batuwita and Palade, 2009b), triplet-SVM (Xue et al., 2005), and miRBoost (Tempel et al., 2015), which use different numbers of human and cross-species manual features to identify miRNAs as inputs to a support vector machine(SVM); and MiPred (Jiang et al., 2007), which selects a set of mixed features, including the minimum free energy (MFE), the local contiguous triplet structure composition, dinucleotide shuffling, and the P-values of randomization tests, to construct a random forest classifier to identify miRNAs. The context-sensitive hidden Markov model (CSHMM) method (Agarwal et al., 2010) predicts miRNAs by filtering the human dataset; whereas M0iRANN (Rahman et al., 2012), DP-miRNA (Thomas et al., 2017), and BP (Jiang et al., 2016) extracted 98 features as inputs to their neural networks. These methods use hand-crafted features as inputs to the model, including pre-miRNA structural and folding energy information such as dinucleotide and trinucleotide pair frequency, loop and sequence length, MFE, and melting temperature. Manual extraction of features often requires careful design based on the characteristics of the data; this, combined with reliance on the database, weakens the generalization ability of the model.

Many deep learning methods can automatically learn the representation of features from the data. For instance, deepMiRGene (Park et al., 2017) uses long short-term memory (LSTM) to automatically extract features and process time-dependent problems in a sequence. Do et al. (2018) introduced a convolutional neural network (CNN) to automatically extract features to identify miRNAs. Lee et al. (2016) used an automatic encoder based on a deep recurrent neural network (RNN) to determine the interaction of miRNA sequences for miRNA target prediction. All of these methods involve automatic extraction of features. However, most of the information is both spatial and sequential. The miRNA spatial structure contains miRNA functional information, as the base sequence of the miRNA affects the normal regulation of miRNA molecules. Each of these methods focuses on either time or spatial information.

In recent years, researchers have explored how to use CNN and RNN tools to construct various CNN-RNN frameworks, which can be divided into unified and cascaded combinations (Pinheiro and Collobert, 2014; Donahue et al., 2015; Vinyals et al., 2015; Zuo et al., 2015; Wang et al., 2016; You et al., 2016). In these, the cascaded framework processes the CNN and the RNN, respectively, and the RNN takes the output of the CNN as its input and returns continuous predictions at different time steps. Such cascaded frameworks can handle various tasks. For example, Pinheiro et al. replaced an RNN with LSTM to solve image subtitle tasks using CNN-RNN (Pinheiro and Collobert, 2014). The model trained the CNN to identify objects in video frames and classify them, then used the output of the CNN as input to the LSTM, creating an "instant" description for each video clip. Quang et al. proposed the DanQ CNN/LSTM combination model (Quang and Xie, 2016), which models the nature and function of introns, using a convolutional layer to capture the regulatory motif while the recursive layer captures the inter-model long-term dependencies, and demonstrated its ability to learn a regulatory grammar to improve forecasting. Compared with other models, DanQ showed great improvements with respect to many metrics. Pan et al. used CNN to learn abstract features and used bidirectional LSTM (BLSTM) to capture possible long-range dependencies between binding sequences and structural motifs recognized by CNN. In this way, they predicted sequence and structural binding preferences of RNA-protein complexes (Pan et al., 2018). These successful applications demonstrate that the ability to focus on both sequential and spatial characteristics yields better classification results.

However, these models cannot focus simultaneously on the sequential and spatial characteristics of pre-miRNAs, because they use a single neural network. They also disregard the information carried by the sequence owing to their focus on the secondary structure. Therefore, we proposed a pre-miRNA identification method based on a cascaded CNN-LSTM framework, called CL-PMI. First, CL-PMI uses the CNN to automatically learn the characteristics of the sequence and the secondary structure from the input, thereby obtaining a spatial feature representation of the pre-miRNAs. Then, a deep RNN with LSTM is used to capture pre-miRNA long-term dependence information from the effective features of CNN learning. Finally, the CL-PMI uses a fully connected layer to identify pre-miRNAs.

Our approach involves a series of nonlinear transformations on data, performed in a data-driven manner, and uses hybrid neural networks to learn the complex abstract sequential and spatial features of data with the automatic extraction of features.

MATERIALS AND METHODS

Most existing miRNA identification algorithms manually extract features, which requires strong expertise in the field and thus inevitably limits their universality. These algorithms focus on either the sequential characteristics of the miRNA or its spatial characteristics, but not both.

Using the input sequence and secondary structure to design the size, number, and sliding step size of the convolution kernel, CNN can be used to automatically extract features from the input, effectively solving the problem of manual extraction of features. In the cascaded CNN-RNN framework, the CNN describes the state of a certain space, and the RNN connects the spatial states together to form a time concept, thus enabling the model to consider spatial and sequential characteristics at the same time. Therefore, we introduced a cascaded CNN-LSTM framework to identify pre-miRNAs, called CL-PMI, which consists of a CNN layer, an LSTM layer, and a fully connected layer (FC). In this framework, we first use one-hot encoding to process the pre-miRNA sequence and its corresponding secondary structure, using encoded pre-miRNAs as the input to the CNN. The CNN automatically extracts pre-miRNAs spatial correlation features; then, LSTM takes the effective features of CNN learning as inputs and uses the three gating units to capture the long-term dependencies of the pre-miRNAs. Finally, the fully connected layer combines spatial information and sequential information for robust classification. **Figure 1** shows an overview of the CL-PMI framework.

Encoding Sequence and Structure

The information carried by the sequence and secondary structure of the pre-miRNA plays an important part in the identification process. The pre-miRNA sequence is a non-coding single-stranded RNA molecule of approximately 22 nucleotides. The secondary structure is double-stranded owing to base pairing interactions. The stem ring and hairpin structures resulting from these interactions, as shown in **Figure 2B**, are the most prominent features of pre-miRNAs. The left side of the stem is the forward chain (5'→3'), and the right is the reverse strand (3'→5'), complementary base matches between these strands result in formation of a helix. Dot bracket notation (DBN) is a widely used method for describing secondary structures. As shown in **Figure 2A**, DBN uses paired parentheses to indicate complementary pairing of bases and continuous dot numbers to indicate stem-loop structures. The pre-miRNA secondary structure is one of the inputs of CL-PMI, which is obtained by calculating the MFE of the pre-miRNA sequence with the RNAfold tool (Hofacker, 2003).

In order to capture more pre-miRNA information, we considered the sequence information and corresponding secondary structure information simultaneously. Each pre-miRNA sequence consists of four nucleotide types {A, C, G, U}, and the secondary structure has three "(", ".", ")" observable states. We used a one-hot encoding scheme to convert the nucleotides at each position of the pre-miRNA sequence into

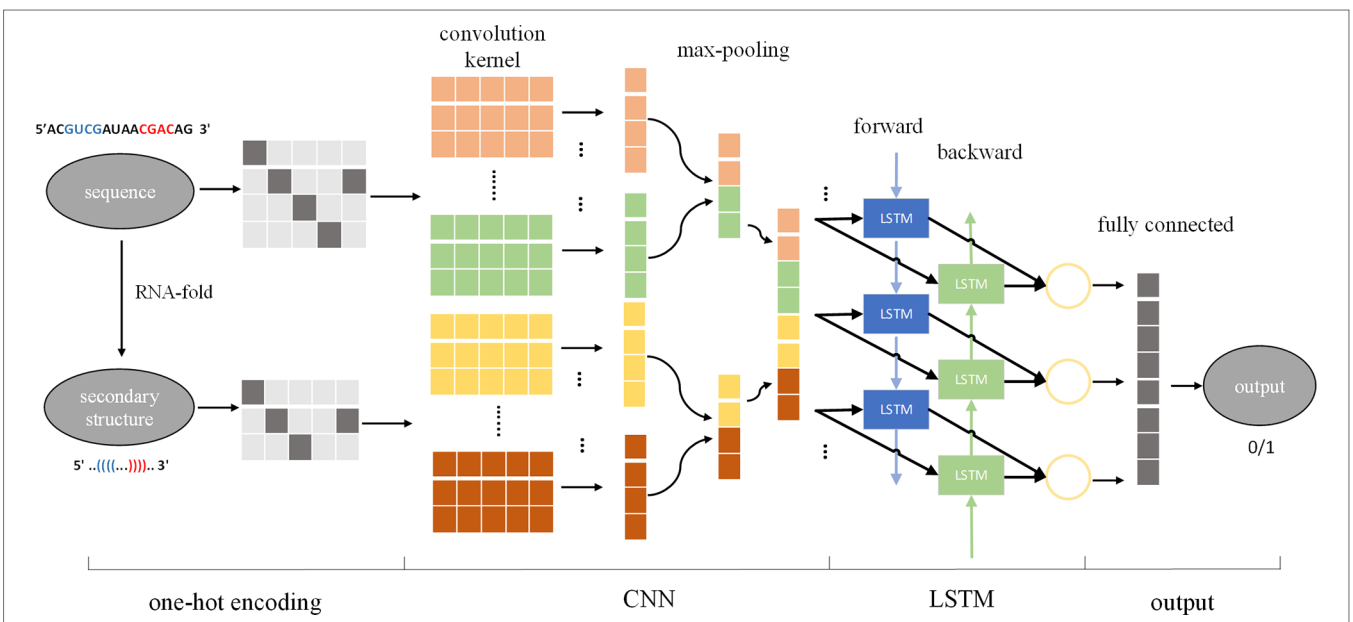

FIGURE 1 | Overview of proposed CL-PMI methodology; the sequence and the secondary structure are first encoded as four-dimensional and three-dimensional matrices, respectively. The convolutional layer acts as a scanner across the input matrix to extract features, and the max-pooling layer reduces the data dimension and retains significant features. Then two CNNs are integrated as inputs to BLSTM, and the combined information is used to learn the long-term dependence of the pre-miRNA. Finally, the fully connected layer performs a sigmoidal nonlinear transformation on the output of BLSTM to identify the pre-miRNA.

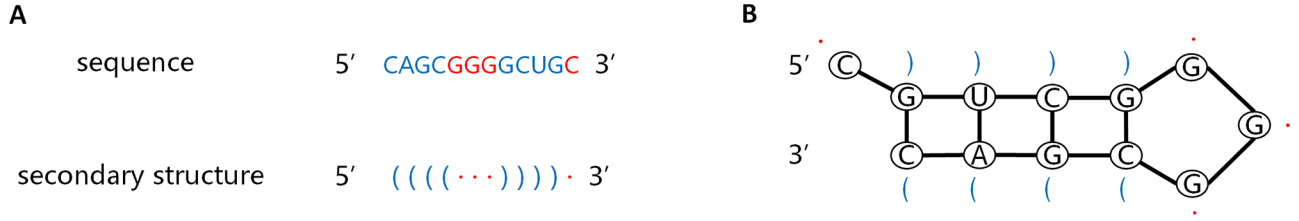

FIGURE 2 | Sequence of pre-miRNA and corresponding secondary structures predicted by RNAfold. **(A)** Sequence of pre-miRNA and secondary structures described by dot-bracket notation. **(B)** Secondary structure of the given sequence.

four-dimensional vectors, and the observable state of each position of the secondary structure was converted into a three-dimensional vector; these vectors were used as the inputs to the CNN. For example, let S_{seq} be a pre-miRNA sequence and S_{str} be the secondary structure corresponding to S_{seq}, where S_{seq} = {*A,C,G,U,U*} and S_{str} = {(,·,·,·,)}; then, S_{seq} is encoded as a four-dimensional binary tuple vector and S_{str} is encoded as a three-dimensional binary tuple vector:

$$S_{seq} : [1,0,0,0],[0,1,0,0],[0,0,1,0],[0,0,0,1],[0,0,0,1]$$

$$S_{str} : [1,0,0],[0,1,0],[0,1,0],[0,1,0],[0,0,1]$$

Convolutional Neural Network

A pre-miRNA sequence contains frequency-dependent features of two or three adjacent nucleotide and aggregated dinucleotide frequencies. The secondary structure of pre-miRNA involves different thermodynamic stability spectra of the pre-miRNA and other features, such as adjusted base pair distance, structure entropy, melting temperature, loop length, and positional entropy, which estimates the structural volatility of the secondary structure. As the pre-miRNA sequence and the secondary structure carry different information characteristics, we used CNN to train two different branches for the sequence and the secondary structure and to learn the subsequence features from the two types of input information.

Each CNN branch consisted of a convolutional layer, a rectified linear unit (ReLU), and a max-pooling layer that together extracted sequence and secondary structure features from the input. We selected the max-pooling layer to subsample the output of the convolutional layer. There are two advantages to using max-pooling. First, it reduces the offset of estimated mean caused by convolutional layer parameter errors. Second, it removes redundant information carried by the feature map, reduces parameters, and prevents overfitting. The convolutional layer extracts the features of the input data and abstracts the implicit associations in the original data through the convolution kernel matrix. In principle, convolution is a mathematical operation of point-multiplication summation of two matrices, the input data matrix and the convolution kernel (filter or feature matrix). The results obtained are expressed as specific local features extracted from the pre-miRNA. After convolution, we applied a rectifying linear unit to sparsify the output of the convolutional layer, then output the region vector obtained by the pooling layer to the LSTM layer.

For the sequence, the input matrix length was denoted by b, and the convolutional layer included N_{filter} filters, each of length k. Each sliding window range was s=1 to b-k+1. A sliding filter and point multiplication were used to obtain a feature map of size N_{filter}×(b-k+1). The convolved feature map, Z, can be represented as follows:

$$Z = f_{conv}(X) \tag{1}$$

$$Z_{s,i} = \sum_{j=1}^{N_f} \sum_{r=1}^{k} W_{i,j,r} X_{s+r-1,j} + B_i \tag{2}$$

where $Z_{s,i}$ represents the feature map generated by the sth sliding neighborhood window and the ith filter; X is the input sample, of size N_{in}×b, i∈{1,…,N_{filter}}; W is the weight, of size N_{filter}×N_{in}×k; and B is the bias value, of size N_{filter}×1. These are the trainable parameters of the convolution layer.

Next, we applied a ReLU, an activation function that keeps the convolutional layer positively matched and eliminates negative matches:

$$f_{relu}(Z) = relu(Z) = \max(0, Z) \tag{3}$$

In order to reduce the parameters and learn translational invariant features, we used max-pooling on the output of the convolution. Max-pooling preserves only the maximum output of each filter in each step to reduce the output size of the convolution layer; it was applied to the output of convolution Z of size N_{filter}×s, where s=b-k+1. In the case where the size of the pool was m, we obtained an output V

$$V = f_{maxpool}(Z) \tag{4}$$

$$V_{i,p} = \max_{j=1}^{m} Z_{i,m(s-1)+j} \tag{5}$$

where $s \in \left\{1,\ldots,\left\lfloor \frac{s}{m} \right\rfloor\right\}$, $i \in \{1,\ldots,N_{filter}\}$ and the size of V is $N_{filter} \times \left\lfloor \frac{s}{m} \right\rfloor$. Analogous definitions also hold for secondary structure.

Before entering the next layer, the sequence and the secondary structure were concatenated into a single output. The next LSTM

layer and the fully connected layer worked together on the merged sequence and the structural layer.

Long Short-Term Memory Network

We introduced LSTM (Hochreiter and Schmidhuber, 1997) to identify the combined information of the sequence and secondary structures, allowing us to use long-term dependency information to aid current predictions. An LSTM cell has an internal mechanism called a gate that regulates the flow of information. Three gate units are shown in **Figure 3**. When the LSTM cell scans each element of the input sequence, it first selectively discards the information in the cell state using the "forget" gate. The input gate records new information into the cell state and then updates the current state value. Finally, the output gates determine which values should be output. As standard LSTM often ignores the future context of the pre-miRNA when processing the sequence, a bidirectional LSTM (Graves et al., 2005) is used to solve this problem. Its main goal is to increase the information available to the RNN, including the history and future data of an input using time series data. It scans the outputs of the CNN from two directions, along and against the timing direction. The outputs for each direction are connected for subsequent classification. The calculation process of the LSTM cell at the time step t is as follows:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \tag{6}$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \tag{7}$$

$$c_t = f_t c_{t-1} + i_t tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{8}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \tag{9}$$

$$h_t = o_t \tanh(c_t) \tag{10}$$

where f_t, i_t, c_t, o_t represent the forget gate, the input gate, the cell activation vector, and the output gate, respectively; X, h, and c represent input vectors, hidden states, and memory cells, respectively; W and b are the weights and bias, that is, the model parameters to be trained; and σ is the sigmoid function:

$$\sigma(x) = 1/(1+e^{-x}) \tag{11}$$

Addressing Potential Overfitting

Overfitting is a very common problem in deep learning and may result from the model lacking control during the learning process. An overfitting model will not perform well in data identification. In the current work, the model would not be able to identify pre-miRNA correctly if overfitting were to occur and would have poor generalization ability. In order to reduce this risk, we used batch normalization, dropout, and L2 regularization to prevent or mitigate overfitting.

Batch standardization (Ioffe and Szegedy, 2015) normalizes the output of neurons in each training batch so that the output obeys a normal distribution with 0 as the mean and 1 as the standard deviation, thus avoiding the problem of internal covariate migration. In the case where the ith batch contains 100 samples, the particular neuron produces outputs $N_{i,1}$, … ,$N_{i,100}$, then standardizes it in batches:

$$\frac{N_{i,1}-\bar{N}_i}{\sigma_i},\ldots,\frac{N_{i,100}-\bar{N}_i}{\sigma_i} \tag{12}$$

where $\sigma_i^2 = \frac{1}{n-1}\sum_{j=1}^{100}(N_{i,j}-\bar{N}_i)^2$ *and* $\bar{N}_i = \frac{1}{n}\sum_{j=1}^{100}N_{i,j}$ are the sample variance and mean, respectively. Batch normalization

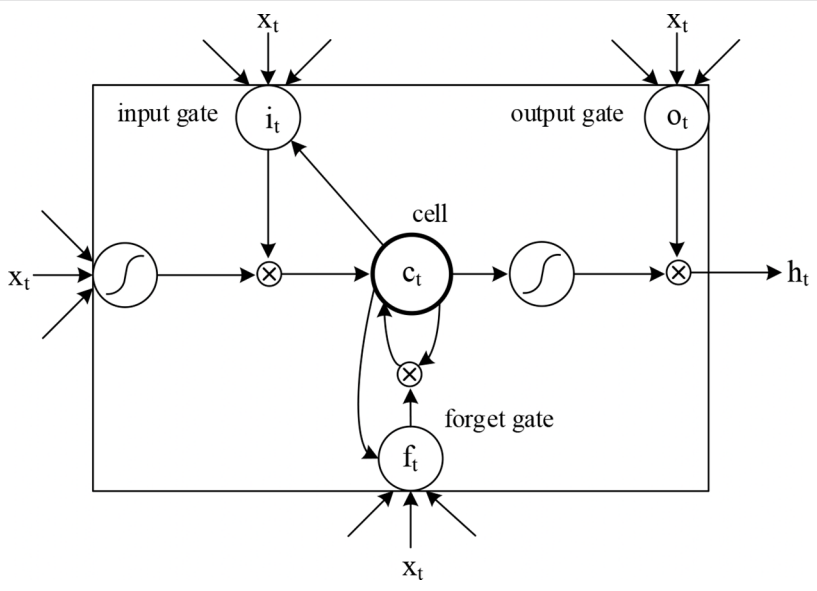

FIGURE 3 | Internal structure of the LSTM cell.

makes the feature scales consistent in all dimensions of the data to avoid excessive concentration of some dimensional data, thus alleviating gradient disappearance and overfitting. By reducing the dependence of the gradient on the parameters or their initial scales, the network can be trained with a high learning rate to accelerate network convergence. Batch normalization can be seen as a means of regularization, which can improve the generalization ability of the model and optimize the model structure. We batch-standardized the output of the max-pooling layer, the LSTM layer, the fully connected layer, and the ReLU activation in the network. During the forecast period, the batch average and variance were replaced by the total mean and variance, which were calculated when all batches had been trained.

Dropout (Srivastava et al., 2014) refers to temporarily dropping the output of the neural network unit from the network according to a fixed probability p during the training of the deep learning network. In other words, the effects of these neurons on the downstream start-up are neglected in the forward propagation, and their weights are not updated in the backpropagation. This makes the network less sensitive to changes in the weight of a neuron, increasing generalization and reducing overfitting. For the max-pooling layer, the LSTM layer, and the fully connected layer, we applied a loss rate of $p=0.5$ to the outputs. Note that dropout was only used during training. In the testing stage, dropout was not applied because a random output would affect the prediction.

Finally, we applied L2 regularization to the weight matrix of the fully connected layer, and punished the loss function by adding the square of the weight to the loss function. This reduced the complexity of the model, thereby reducing overfitting.

Training

We implemented our neural network model using Python and Keras (Chollet, 2015). The model utilized a backpropagation algorithm to calculate the loss function value between the output and the label, before calculating its gradient relative to each neuron and updating the weight according to the gradient direction. We applied focal loss (Lin et al., 2017), a dynamically scaled cross-entropy loss function, to train the samples. Formally, focal loss is defined as

$$L_{FL} = (1 - p_y)^{\gamma} L_{CE} \tag{13}$$

$$L_{CE} = -\alpha_y \log p_y \tag{14}$$

$$p_y = \begin{cases} p & if\ y = 1 \\ 1 - p & otherwise \end{cases} \tag{15}$$

where L_{CE} represents the cross-entropy of binary classification; $y \in \{+1,-1\}$ specifies the label of the sample; $p \in [0,1]$ is the model's estimated probability for the class with label $y = 1$; $\alpha \in [0,1]$ is a weighting factor corresponding to class 1, and $1-\alpha$ corresponds to class -1; γ is the focus parameter; and $(1-p_y)^{\gamma}$ is the modulation factor. In this work, we set γ to 2 and α to 0.25.

In the process of training a model, choosing a good optimizer not only accelerates the training of the model but also improves the experimental results. Kingma and Ba (2014) proposed the Adam optimizer, which combines the advantages of two algorithms, AdaGrad and RMSProp, to calculate the update step by considering the first moment estimation (the mean of the gradient) and the second moment estimation (the uncentered variance of the gradient). Furthermore, Adam is considered to be an optimizer with excellent performance and is the default choice. Therefore, we chose Adam as the optimizer while training CL-PMI, with the mini-batch size and learning rate set to 128 and 0.001 for all experiments. The details of the model parameters are shown in **Supplement A** Pseudo code can be obtained in **Supplement B**.

We performed a five-fold cross-validation on the training data to evaluate the classification performance of CL-PMI. As shown in **Figure 4**, SP, F-score, g-mean, and AUROC tend to be stable at about 20 epochs. SE and AUPR showed a slight upward trend between 20 and 300 epochs. The PPV fluctuated around 280 epochs and stabilized after 300 epochs. Loss has been in a gentle downward trend. Different indicators are stable in different epoch, in order to comprehensively consider all indicators, we stopped training after 300 epochs. Owing to the data imbalance, the prediction was biased towards the negative dataset in the early training stage, as reflected by the F-score and geometric mean (g-mean) remaining close to 1. However, the prediction was tuned and converged as learning progressed.

RESULTS AND DISCUSSION

Our method and the four comparison methods used the same datasets (Tempel et al., 2015; Park et al., 2017), *human*, *cross-species*, and *new*. Positive examples were retrieved from miRbase (Griffiths-Jones et al., 2006) (18th edition), and the negative examples were from NCBI (http://www.ncbi.nlm.nih.gov), NONCODE (Bu et al., 2011), fRNAdb (Kin et al., 2006), and snoRNA-LBME-db (Lestrade and Weber, 2006). Negative examples mainly included exonic regions from protein-coding genes and noncoding RNAs that were not miRNAs, such as tRNA, siRNA, snRNA, and snoRNA. To improve data quality and prevent overfitting, mis-annotated elements were discarded in these examples and redundant sequences were removed (Tempel et al., 2015). In addition, we obtained 690 positive samples from miRBase22, and obtained 8,246 negative examples from Xue (Xue et al., 2005) and Zou (Wei et al., 2014) as new22 dataset. For the human and cross-species datasets, we used 10% of the data as the test set and the remaining 90% to implement five-fold cross-validation for training and selecting the model. The new and new22 dataset were only used for testing. We predicted for new and new22 dataset using the model trained with cross-species dataset.

As shown in **Table 1**, the experiment involved a total of 3,230 positive examples and 23,934 negative examples. The human dataset contained 863 positive examples and 7,422 negative examples. The cross-species dataset consisted of 1,677 positive examples and 8,266 negative examples obtained

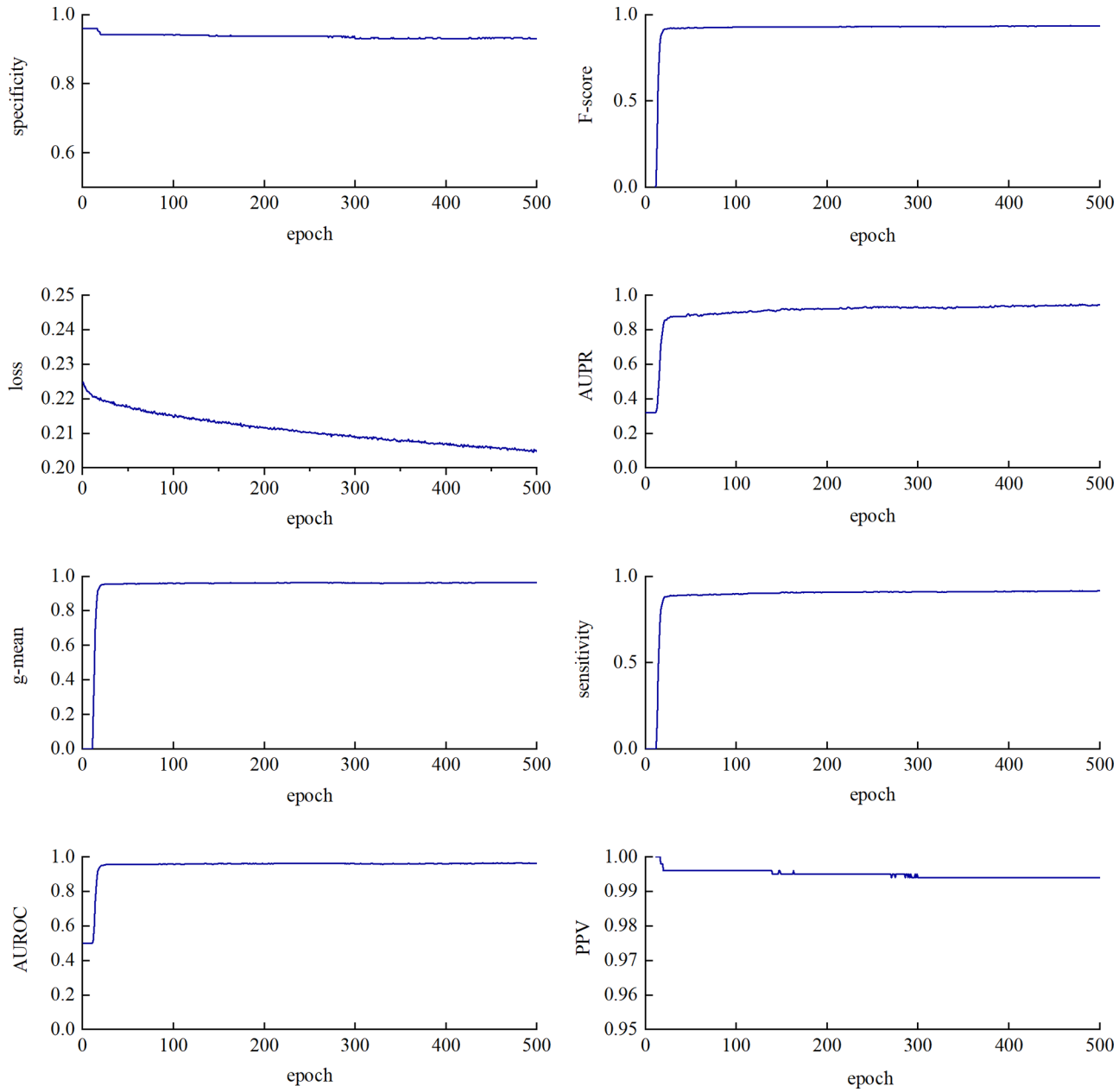

FIGURE 4 | Training loss and seven evaluation metrics using test dataset with a varying epoch. The changes in metrics and loss functions tended to be stable after about 300 epochs, so we used a model that trained 300 epochs.

TABLE 1 | The number of sequences in the datasets used in this study.

	Cross-species	Human	New	New22
Positive examples	1,677	863	690	690
Negative examples	8,266	7,422	8,246	8,246

from different species (e.g., mice, humans, and flies). The new dataset was obtained from miRBase versions 19 and 20, and consisted of 690 positive and 8,246 negative newly found examples.

Experimental Setup

For the human and cross-species datasets, we performed a five-fold cross-validation. We randomly selected 80% of the data to form the training set; the remaining 20% were used as the test set. The numbers of hidden nodes in the LSTM and FC layers were determined to be 20 and 256 by five-fold cross-validation. The mini-batch size and training epochs were set to 128 and 300, respectively.

For comparison, sensitivity (SE), specificity (SP), F-score, g-mean, positive predictive value (PPV), area under the precision-recall curve (AUPR), and area under the receiver operating

characteristic (AUROC) were used to evaluate model performance. These metrics were calculated as follows:

$$SE = TP / (TP + FN) \tag{16}$$

$$SP = TN / (TN + FP) \tag{17}$$

$$PPV = TP / (TP + FP) \tag{18}$$

$$F - score = 2TP / (2TP + FP + FN) \tag{19}$$

$$g - mean = \sqrt{SE \cdot SP} \tag{20}$$

where TN, TP, FN, and FP denote the number of true negatives, true positives, false negatives, and false positives, respectively. These formulas were based on the confusion matrix, with a decision threshold of 0.5.

Validation and Test Performance Evaluation

Next, CL-PMI was applied to three datasets for pre-miRNA identification. In order to evaluate the performance of CL-PMI, we compared it with four existing pre-miRNA identification methods.

One of these methods, miRBoost (Tempel et al., 2015), is an ensemble method that extracts the appropriate features from 187 existing features and performs classification after training the data using the enhancement techniques of the SVM component. Another, microPred (Batuwita and Palade, 2009b), selects the most discriminative feature setting to train the SVM classifier using a filtering method, handles the class imbalance problem in the dataset, and uses cross-validation to evaluate classification performance. The SVM used in these two methods is a binary classification model, which can be defined as the linear classifier with the largest interval in the feature space. The learning strategy is to maximize the interval and finally transform into a solution to a convex quadratic programming problem. These two comparison methods are traditional machine learning methods.

Park et al. (2017) combined secondary structure with the pre-miRNA sequence to form a 16-dimensional matrix, then sent the results to the RNN to improve long-term dependency modeling. The greatest advantage of this approach is that it does not require hand-crafted features. Do et al. (2018) proposed a novel joint two-dimensional multi-channel method to identify pre-miRNAs, using the secondary structure encoded by the pairing matrix format as the input to the two-dimensional convolution network to achieve automatic feature extraction. These features were fed into fully connected layers for classification. These two comparison methods are deep learning methods.

Therefore, we used miRBoost, microPred, deepMiRGene, and DCNN as comparative experiments in this paper to evaluate the performance for pre-miRNA identification under the same datasets. The experimental results for the three datasets are described and discussed below.

TABLE 2 | Results for the human dataset.

Methods\metrics	SE	SP	PPV	F-score	g-mean	AUROC	AUPR
miRBoost (CV)	0.803	0.988	0.887	0.843	0.891	–	–
microPred (CV)	0.763	**0.989**	0.888	0.820	0.869	0.974	0.890
deepMiRGene (CV)	0.799	0.988	0.885	0.839	0.888	0.984	0.915
DCNN fixed-sized (CV)	0.878	0.978	0.827	0.849	0.926	0.984	0.915
DCNN variable-sized (CV)	0.835	0.985	0.868	0.851	0.907	**0.985**	**0.922**
Proposed (CV)	**0.989**	0.935	**0.992**	**0.991**	**0.962**	0.962	0.854
miRBoost (test)	0.884	0.969	0.768	0.822	0.925	–	–
microPred (test)	0.779	0.988	0.882	0.827	0.877	0.980	0.892
deepMiRGene (test)	0.822	**0.992**	0.919	0.868	0.903	0.981	0.918
DCNN fixed-sized (test)	0.930	0.984	0.870	0.899	**0.957**	0.983	**0.946**
DCNN variable-sized (test)	0.884	0.991	0.916	0.899	0.936	**0.986**	0.934
Proposed (test)	**0.968**	0.895	**0.988**	**0.978**	0.931	0.972	0.807

Bold numbers are the highest scores in this catergory.

TABLE 3 | Results for the cross-species dataset.

Methods\metrics	SE	SP	PPV	F-score	g-mean	AUROC	AUPR
miRBoost (CV)	0.861	0.977	0.884	0.872	0.917	–	–
microPred (CV)	0.825	0.975	0.875	0.848	0.897	0.970	0.873
deepMiRGene (CV)	0.886	**0.982**	0.911	0.898	0.933	**0.985**	0.927
DCNN fixed-sized (CV)	0.903	0.978	0.894	0.898	0.940	**0.985**	0.936
DCNN variable-sized (CV)	0.881	0.981	0.906	0.893	0.930	0.983	**0.936**
Proposed (CV)	**0.995**	0.950	**0.990**	**0.992**	**0.972**	0.972	0.933
miRBoost (test)	0.856	0.844	0.526	0.651	0.850	–	–
microPred (test)	0.814	0.985	0.919	0.863	0.896	0.963	0.906
deepMiRGene (test)	0.900	0.983	0.913	0.906	0.940	0.984	**0.955**
DCNN fixed-sized (test)	0.904	0.982	0.910	0.907	0.942	0.983	0.951
DCNN variable-sized (test)	0.880	**0.988**	0.936	0.907	0.933	**0.985**	0.950
Proposed (test)	**0.977**	0.910	**0.982**	**0.979**	**0.943**	0.958	0.877

Bold numbers are the highest scores in this catergory.

Tables 2 and **3** show the cross-validation and test performance for the human and cross-species data-sets, respectively. Cross-validation performance is shown in the top half of each table and test performance in the lower half.

Our approach performed best overall in the case of the cross-species dataset. In the cross-validation, our method achieved the best values for SE, g-mean, F-score, and PPV; these were 10.19%, 3.4%, 10.47%, and 8.67% higher than the best results obtained by a comparison method, respectively, and our method ranks the second on AUPR. The reduction in SP occurred because training a classifier system with an unbalanced dataset (where the positive class is a minority) typically produces a suboptimal model with higher SP and SE (Batuwita and Palade, 2009a). In the test results, compared with DCNN, CL-PMI showed a 8.08% increase in SE. The performance with respect to the other metrics was the same in the cross-validation. This similarity indicated that overfitting had been effectively addressed. Some methods, such as miRBoost, showed fair performance in cross-validation, but poorer performance for each indicator with the test data. However, CL-PMI demonstrated the same level of performance in both cross-validation and testing, indicating that our approach has a more powerful generalization ability than the others. In order to more intuitively show the differences between the five methods for each indicator, we drew a radar chart, in which each indicator corresponded to a coordinate axis, and the relative position and angle of the axis were usually uninformed. **Figure 5** shows a comparison of the predicted performances for the cross-species dataset. We performed five-fold cross-validations and averaged the results. As illustrated by the radar chart, our method performed best on four of the seven indicators, and values for the remaining three were also close to optimal. This demonstrates that our proposed method is competitive in identifying pre-miRNAs.

For the new dataset, in the basic indicators, CL-PMI showed a 5.32% increase in SE compared with miRBoost. In the comprehensive performance indicators, although our method was slightly worse than microPred in AUROC and AUPR.CL-PMI showed increases of 27.34% and 20.07% in PPV and F-score compared with DCNN, respectively. For a better view, we also plotted the AUC curves and AUPR curves of our method on human, cross-species, and new, respectively in **Figures 6** and **7**. We performed tests on the new dataset using the model trained on the cross-species dataset. The results are shown in **Table 4**. Although CL-PMI was trained on the mixed-species dataset, it showed potential for identifying new pre-miRNAs.

For human datasets, according to the five-fold cross-validation results, CL-PMI outperformed other comparison methods with respect to SE, F-score, PPV, and g-mean. Notably, the F-score of our method was increased by 16.45% compared with that of DCNN; the other metrics also improved by 12.64%, 11.71%, and 3.89% compared with the best methods. For the test set, our method achieved the best performance in terms of SE, F-score, and PPV; although it did not give the highest scores on other metrics, the performance of CL-PMI was close to that of the best method.

We compared the performance of traditional machine learning and deep learning. In the human dataset, in addition to the highest SP achieved with microPred, the deep learning methods showed better performance for all metrics. For the cross-validation and testing of the cross-species dataset, the deep learning methods outperformed other methods. With the new dataset, the deep learning methods showed the best performance for all the evaluated metrics. All these results demonstrate that deep learning methods are superior to machine learning methods for identifying pre-miRNAs.

Our method achieved optimal results for SE, PPV, and F-score. Although deepMiRGene used LSTM to capture the long-term dependence of pre-miRNA, it did not focus on pre-miRNA spatial interaction, whereas DCNN used the CNN to focus on pre-miRNA spatial dependence but ignored complex space-time dependencies. CL-PMI considers both the sequential and spatial information of pre-miRNA, enabling the model to simultaneously express the characteristics of

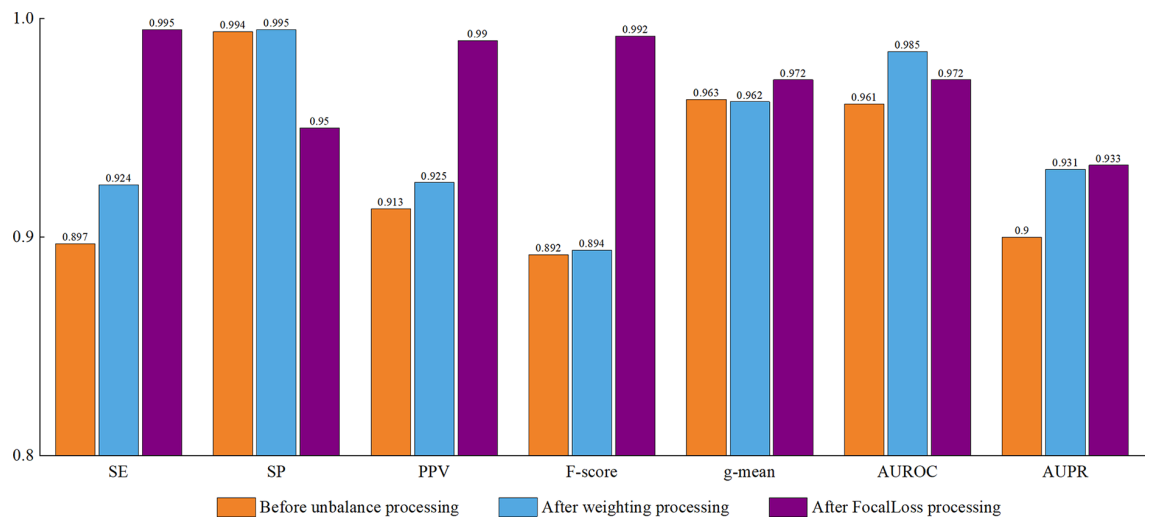

FIGURE 5 | Comparison of prediction performance of our method and other comparison methods on cross-species datasets.

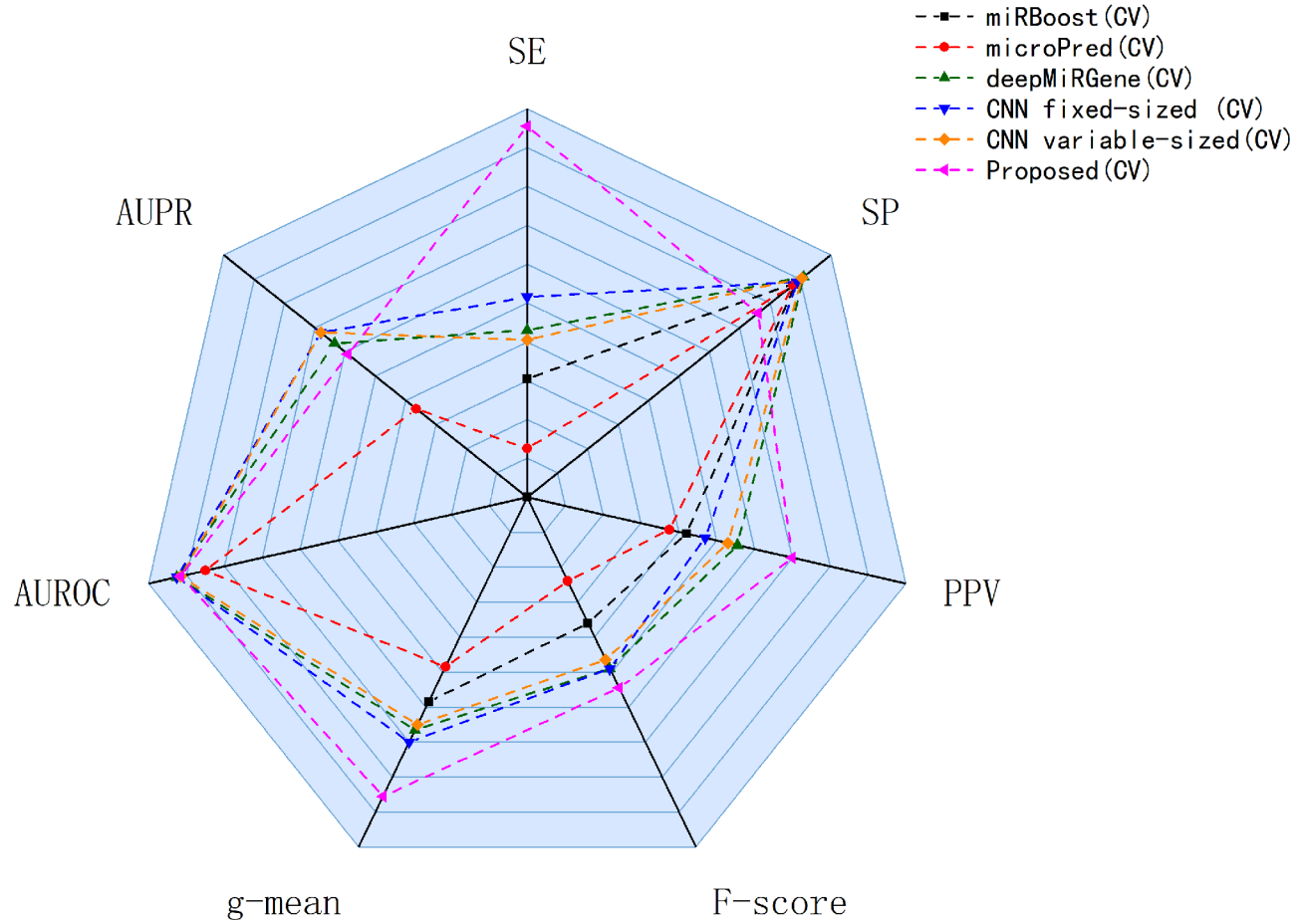

FIGURE 6 | The AUROC curves of our method on the dataset human,cross-species, and new.

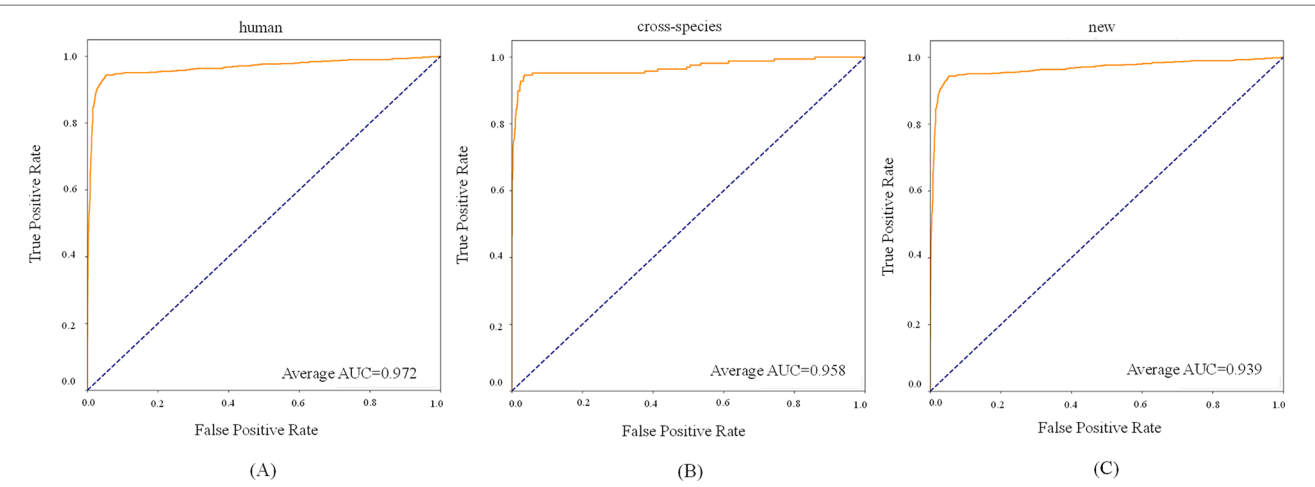

FIGURE 7 | The AUROC curves of our method on the dataset human,cross-species, and new.

pre-miRNA in the spatial and time dimensions and thus achieve better classification results. The above results and analysis confirm that CL-PMI is competitive among deep learning methods. In addition, we tested our model on the new22 with an accuracy of 0.907. Related experimental details of new22 are shown in **Supplement C.**

TABLE 4 | Results for the new dataset.

Methods\metrics	SE	SP	PPV	F-score	g-mean	AUROC	AUPR
miRBoost	0.921	0.936	0.609	0.733	0.928	–	–
microPred	0.728	0.970	0.672	0.699	0.840	0.940	0.756
deepMiRGene	0.917	0.964	0.682	0.782	0.941	**0.981**	0.808
DCNN fixed-sized	0.917	0.967	0.696	0.792	**0.942**	0.979	**0.864**
DCNN variable-sized	0.859	**0.981**	0.779	0.817	0.918	0.979	0.818
Proposed	**0.970**	0.907	**0.992**	**0.981**	0.938	0.939	0.746

Bold numbers are the highest scores in this category.

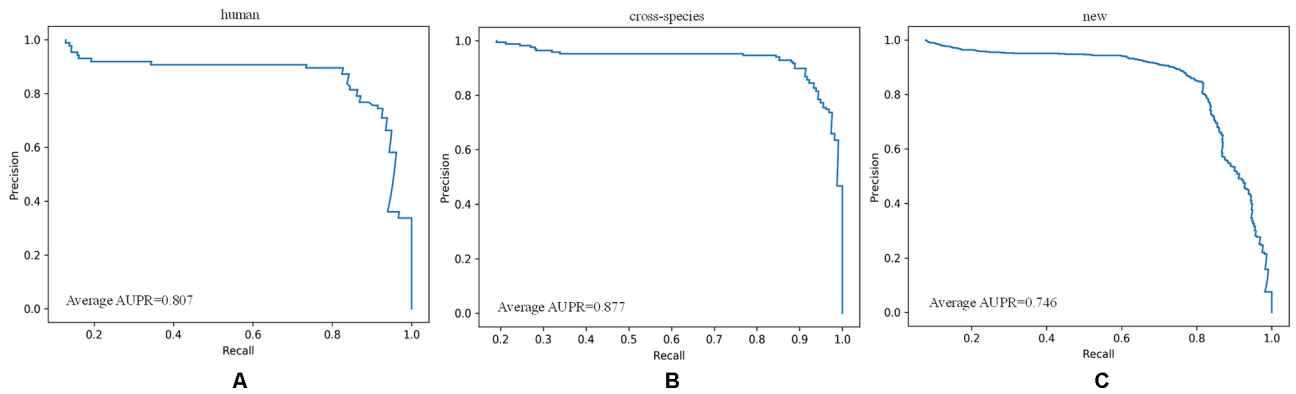

FIGURE 8 | Performance assessment of cross-species dataset before and after unbalanced processing. The average score of the five-fold results is reported.

Impact of Class Imbalance on the Model

There was a certain degree of class imbalance in our experiments. That is, the number of positive examples (3,230 pre-miRNAs) was much smaller than that of negative examples (23,934 non-pre-miRNAs). The ratio of positive and negative examples was 1:7.4. Researchers have conducted extensive explorations of class imbalances. Minority classes are largely ignored and predicting them is more difficult, leading to degraded classifier performance (Weiss, 2004). To improve the model's performance, it was necessary to solve the class imbalance problem.

In response to the dataset imbalance problem, we initially tried to use class weights, that is, we directly considered the asymmetry of the cost error during the classifier training, which embedded the output probability of each class in the cost error information. This probability was then used to define a classification rule with a 0.5 threshold. Specifically, the aim was to identify those small classes (positive pre-miRNAs) that could be used to add weight to the positive examples of the model and reduce the weight of the negative examples. This method produces a new data distribution, which allows the classifier to focus on positive examples. In this experiment, we set the positive example weight to 0.9 and the negative example weight to 0.1.

Later, we used the focal loss function, which is an elegant and effective proposal to solve the problem of class imbalance. In this function, γ is the focus parameter, which smoothly adjusts the reduced ratio of the weight of the easy sample, and $(1\text{-}p_y)^\gamma$ is the modulation factor, which reduces the loss contribution of the easy sample and broadens the range in which the sample receives low loss. When $\gamma = 0$, focal loss is equivalent to cross-entropy loss. When γ increases, the influence of the modulation factor increases accordingly. By adding the modulating factor, focal loss reduces the weight of the easy sample, making the model more focused on the hard sample during training.

As shown in **Figure 8**, all of the performance metrics were higher after the class imbalance processing, except for g-mean and SP, although these showed small drops of only 0.001 and 0.042. The reduction of SP occurred because training a classifier system with an unbalanced dataset (where the positive class is a minority) typically produces a suboptimal model with higher SP and lower SE. By applying the class imbalance learning method, it is usually possible to increase the SE by sacrificing the SP score to some extent (Batuwita and Palade, 2009a). These results showed that our model was not biased towards the negative dataset. The focal loss method was superior to the class weight approach for dealing with imbalances. To be specific, the model achieved a 7.25% higher SE using focal loss compared with class weight. Similarly, for F-score, PPV, and g-mean, using focal loss resulted in 1.68%, 1.62%, and 0.94% higher scores compared with class weight. Therefore, we propose that focal loss is an effective means to deal with the class imbalance problem in pre-miRNA datasets.

CONCLUSIONS

In this paper, we proposed a new pre-miRNA identification method, called CL-MPI. In contrast to existing methods, CL-MPI captures sequence information while also considering secondary structure in data preprocessing. CL-MPI can take into account pre-miRNA sequential and spatial information while automatically extracting pre-miRNA sequence features. We used RNAfold to predict the secondary structure of each pre-miRNA sequence, then used the secondary structure and sequence as inputs to a

CNN to automatically extract features, and finally used LSTM to mimic RNA sequences and further understand the role of the sequence. According to the experimental results, our method achieved better overall performance across cross-species datasets, even in the absence of known manual features, especially for F-score and g-mean. This demonstrates that automatic extraction of features and considering sequential and spatial characteristics at the same time are important for the identification of pre-miRNAs. Owing to the small number of known pre-miRNAs, we needed to deal with extremely unbalanced datasets, with significantly more negative than positive examples. According to Saito and Rehmsmeier (2015), PPV is more useful than other metrics for a binary classifier on an unbalanced dataset, as it varies with positive and negative ratios. As described in the section *Results and Discussion*, we used 3,230 positive examples and 23,934 negative examples to train the CL-MPI model and obtained better performance for PPV compared with alternative methods. The higher SE and PPV achieved for the unbalanced dataset metrics prove that our method can better predict pre-miRNAs.

miRNAs are directly involved in tumor formation (Søkilde et al., 2014), which makes it possible to treat tumors using target carcinogenic miRNAs to restore the function of tumor suppressor miRNAs. With advances in clinical research, miRNAs continue to provide new ideas and treatments for tumor molecular diagnosis and treatment. Biomedical researchers are reluctant to use the outputs of "black box" methods as they are difficult to interpret, which affects the credibility of the results. Our next goal will be to explore more efficient visualization methods. In the future, we will also extend our method to other miRNA-related tasks, such as miRNA target prediction (Iqbal et al., 2016) and miRNA gene expression. Owing to the limited number of known miRNA sequences, the processing of unbalanced datasets also remains a challenge for future work.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

HW and YM performed the majority of the analysis and primarily wrote the manuscript. CD performed some analysis and provided biological expertise. CD performed some analysis of data and helped conceive the project. JW and DL completed the drawing of the charts in the results analysis and the layout of the manuscripts. All authors edited and approved the manuscript.

FUNDING

This study was supported by research grants from the National key research and development plan of China (2018YFD0100204), the National Natural Science Foundation of China (61672374), and the Scientific and Technological Project of Shanxi Province (No.201603D22103-2).

ACKNOWLEDGMENTS

We sincerely thank the other members of our team for useful discussions and help.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00967/full#supplementary-material

REFERENCES

Agarwal, S., Vaz, C., Bhattacharya, A., and Srinivasan, A. (2010). Prediction of novel precursor miRNAs using a context-sensitive hidden Markov model (CSHMM). *BMC Bioinf.* 11 (1), S29. doi: 10.1186/1471-2105-11-S1-S29

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116 (2), 281–297. doi: 10.1016/S0092-8674(04)00045-5

Batuwita, R., and Palade, V. (2009a). A new performance measure for class imbalance learning. application to bioinformatics problems. *2009 International Conference on Machine Learning and Applications*, 545–550. doi: 10.1109/ICMLA.2009.126.

Batuwita, R., and Palade, V. (2009b). microPred: effective classification of pre-miRNAs for human miRNA gene prediction. *Bioinformatics* 25 (8), 989–995. doi: 10.1093/bioinformatics/btp107

Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O., et al. (2005). Identification of hundreds of conserved and nonconserved human microRNAs. *Nat. Genet.* 37 (7), 766. doi: 10.1038/ng1590

Bu, D., Yu, K., Sun, S., Xie, C., Skogerbø, G., Miao, R., et al. (2011). NONCODE v3.0: integrative annotation of long noncoding RNAs. *Nucleic Acids Res.* 40 (D1), D210–D215. doi: 10.1093/nar/gkr1175

Calin, G. A., Sevignani, C., Dumitru, C. D., Hyslop, T., Noch, E., Yendamuri, S., etal. (2004). Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. *Proc. Natl. Acad. Sci.* 101 (9), 2999–3004. doi: 10.1073/pnas.0307323101

Chollet, F. (2015). Keras: Theano-based deep learning library. Code: https://Github.Com/Fchollet. Documentation: Http://Keras.Io.

Do, B. T., Golkov, V., Gürel, G. E., and Cremers, D. (2018). Precursor microRNA identification using deep convolutional neural networks. *BioRxiv*, 414656. doi: 10.1101/414656

Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., et al. (2015). Long-term recurrent convolutional networks for visual recognition and description. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2625–2634. doi: 10.1109/CVPR.2015.7298878.

Graves, A., and Fernández, S.J. B. T.-A. N. N. F. M. & T. A. (2005). Bidirectional LSTM networks for Improved Phoneme Classification and Recognition. *Schmidhuber International Conference, Warsaw, Poland, September*.

Griffiths-Jones, S., Grocock, R. J., Van Dongen, S., Bateman, A., and Enright, A. J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.* 34 (suppl_1), D140–D144. doi: 10.1093/nar/gkj112

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9 (8), 1735–1780. doi: 10.1162/neco.1997.9.8.1735.

Hofacker, I. L. (2003). Vienna RNA secondary structure server. *Nucleic Acids Res.* 31 (13), 3429–3431. doi: 10.1093/nar/gkg599

Ioffe, S., and Szegedy, C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. *ArXiv Preprint ArXiv*: 1502.03167.

Iqbal, M. S., Hafeez, M. N., Wattoo, J. I., Ali, A., Sharif, M. N., Rashid, B., et al. (2016). Prediction of host-derived miRNAs with the potential to target PVY in potato plants. *Front. Genet.* 7, 159. doi: 10.3389/fgene.2016.00159

Jansson, M. D., and Lund, A. H. (2012). MicroRNA and cancer. *Mol. Oncol.* 6 (6), 590–610. doi: 10.1016/j.molonc.2012.09.006.

Jiang, L., Zhang, J., Xuan, P., and Zou, Q. (2016). BP neural network could help improve pre-miRNA identification in various species. *BioMed Res. Int.* 2016. doi: 10.1155/2016/9565689

Jiang, P., Wu, H., Wang, W., Ma, W., Sun, X., and Lu, Z. (2007). MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Res.* 35 (suppl_2), W339–W344. doi: 10.1093/nar/gkm368

Jones-Rhoades, M. W., Bartel, D. P., and Bartel, B. (2006). MicroRNAs and their regulatory roles in plants. *Annu. Rev. Plant Biol.* 57, 19–53. doi: 10.1146/annurev.arplant.57.032905.105218

Kin, T., Yamada, K., Terai, G., Okida, H., Yoshinari, Y., Ono, Y., et al. (2006). fRNAdb: a platform for mining/annotating functional RNA candidates from non-coding RNA sequences. *Nucleic Acids Res.* 35 (suppl_1), D145–D148. doi: 10.1093/nar/gkl837

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *ArXiv Preprint ArXiv:*1412.6980.

Lee, B., Baek, J., Park, S., and Yoon, S. (2016). deepTarget: end-to-end learning framework for microRNA target prediction using deep recurrent neural networks. *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 434–442. ACM. doi: 10.1145/2975167.2975212

Lestrade, L., and Weber, M. J. (2006). snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res.* 34 (suppl_1), D158–D162. doi: 10.1093/nar/gkj002

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22 (13), 1658. doi: 10.1093/bioinformatics/btl158

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection. *Proceedings of the IEEE International Conference on Computer Vision*, 2980–2988. doi: 10.1109/ICCV.2017.324.

Mathelier, A., and Carbone, A. (2010). MIReNA: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data. *Bioinformatics* 26 (18), 2226–2234. doi: 10.1093/bioinformatics/btq329.

Pan, X., Rijnbeek, P., Yan, J., and Shen, H.-B. (2018). Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. *BMC Genomics* 19 (1), 511. doi: 10.1186/s12864-018-4889-1

Park, S., Min, S., Choi, H.-S., and Yoon, S. (2017). Deep recurrent neural network-based identification of precursor micrornas. *Adv. Neural. Inf. Process. Syst.* 2891–2900.

Pfeffer, S., Zavolan, M., Grässer, F. A., Chien, M., Russo, J. J., Ju, J., et al. (2004). Identification of virus-encoded microRNAs. *Science* 304 (5671), 734–736. doi: 10.1126/science.1096781

Pinheiro, P. H. O., and Collobert, R. (2014). *Recurrent convolutional neural networks for scene labeling*.

Quang, D., and Xie, X. (2016). DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.* 44 (11), e107–e107. doi: 10.1093/nar/gkw226

Rahman, M. E., Islam, R., Islam, S., Mondal, S. I., and Amin, M. R. (2012). MiRANN: A reliable approach for improved classification of precursor microRNA using Artificial Neural Network model. *Genomics* 99 (4), 189–194. doi: 10.1016/j.ygeno.2012.02.001

Saito, T., and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS One* 10 (3), e0118432. doi: 10.1371/journal.pone.0118432

Schmidt, M. F. (2014). Drug target miRNAs: chances and challenges. *Trends Biotechnol.* 32 (11), 578–585. doi: 10.1016/j.tibtech.2014.09.002

Søkilde, R., Vincent, M., Møller, A. K., Hansen, A., Høiby, P. E., Blondal, T., et al. (2014). Efficient identification of miRNAs for classification of tumor origin. *J. Mol. Diagn.* 16 (1), 106–115. doi: 10.1016/j.jmoldx.2013.10.001

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15 (1), 1929–1958.

Tempel, S., Zerath, B., Zehraoui, F., and Tahi, F. (2015). miRBoost: boosting support vector machines for microRNA precursor classification. *RNA* 21 (5), 775–785. doi: 10.1261/rna.043612.113

Thomas, J., Thomas, S., and Sael, L. (2017). DP-miRNA: an improved prediction of precursor microRNA using deep learning model. *2017 IEEE International Conference on Big Data and Smart Computing BigComp.*, 96–99. IEEE. doi: 10.1109/BIGCOMP.2017.7881722

Tüfekci, K. U., Öner, M. G., Meuwissen, R. L. J., and Genç, Ş. (2014). The role of microRNAs in human diseases. In *miRNomics*: *MicroRNA Biology and Computational Analysis* (Springer), 33–55.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: a neural image caption generator. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3156–3164. doi: 10.1109/CVPR.2015.7298935

Wang, D., Qiu, C., Zhang, H., Wang, J., Cui, Q., and Yin, Y. (2010). Human microRNA oncogenes and tumor suppressors show significantly different biological patterns: from functions to targets. *PloS One* 5 (9), e13067. doi: 10.1371/journal.pone.0013067

Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., and Xu, W. (2016). Cnn-rnn: a unified framework for multi-label image classification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2285–2294. doi: 10.1109/CVPR.2016.251

Wei, L., Liao, M., Gao, Y., Ji, R., He, Z., and Zou, Q. (2014). Improved and promising identification of human microRNAs by incorporating a high-quality negative set. *IEEE/ACM Trans. Comput. Biol. Bioinf.* 11(1), 192–201. doi: 10.1109/TCBB.2013.146

Weiss, G. M. (2004). Mining with rarity: a unifying framework. *SIGKDD Explor* 6 (1), 7–19. doi: 10.1145/1007730.1007734

Xue, C., Li, F., He, T., Liu, G.-P., Li, Y., and Zhang, X. (2005). Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinf.* 6 (1), 310. doi: 10.1186/1471-2105-6-310

You, Q., Jin, H., Wang, Z., Fang, C., and Luo, J. (2016). Image captioning with semantic attention. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4651–4659. doi: 10.1109/CVPR.2016.503

Zhu, J., Zheng, Z., Wang, J., Sun, J., Wang, P., Cheng, X., et al. (2014). Different miRNA expression profiles between human breast cancer tumors and serum. *Front. Genet.* 5, 149. doi: 10.3389/fgene.2014.00149

Zuo, Z., Shuai, B., Wang, G., Liu, X., Wang, X., Wang, B., et al. (2015). Convolutional recurrent neural networks: learning spatial dependencies for image representation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 18–26. doi: 10.1109/CVPRW.2015.7301268.

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 December 2019
doi: 10.3389/fgene.2019.01196

High-Dimensional Bayesian Network Inference From Systems Genetics Data Using Genetic Node Ordering

Lingfei Wang[1,2,3], *Pieter Audenaert*[4,5] *and Tom Michoel*[1,6]*

[1] Division of Genetics and Genomics, The Roslin Institute, The University of Edinburgh, Easter Bush Campus, Midlothian, United Kingdom, [2] Broad Institute of Harvard and MIT, Cambridge, MA, United States, [3] Department of Molecular Biology, Massachusetts General Hospital, Boston, MA, United States, [4] IDLab, Ghent University—imec, Ghent, Belgium, [5] Bioinformatics Institute Ghent, Ghent University, Ghent, Belgium, [6] Computational Biology Unit, Department of Informatics, University of Bergen, Bergen, Norway

Studying the impact of genetic variation on gene regulatory networks is essential to understand the biological mechanisms by which genetic variation causes variation in phenotypes. Bayesian networks provide an elegant statistical approach for multi-trait genetic mapping and modelling causal trait relationships. However, inferring Bayesian gene networks from high-dimensional genetics and genomics data is challenging, because the number of possible networks scales super-exponentially with the number of nodes, and the computational cost of conventional Bayesian network inference methods quickly becomes prohibitive. We propose an alternative method to infer high-quality Bayesian gene networks that easily scales to thousands of genes. Our method first reconstructs a node ordering by conducting pairwise causal inference tests between genes, which then allows to infer a Bayesian network *via* a series of independent variable selection problems, one for each gene. We demonstrate using simulated and real systems genetics data that this results in a Bayesian network with equal, and sometimes better, likelihood than the conventional methods, while having a significantly higher overlap with groundtruth networks and being orders of magnitude faster. Moreover our method allows for a unified false discovery rate control across genes and individual edges, and thus a rigorous and easily interpretable way for tuning the sparsity level of the inferred network. Bayesian network inference using pairwise node ordering is a highly efficient approach for reconstructing gene regulatory networks when prior information for the inclusion of edges exists or can be inferred from the available data.

Keywords: systems genetics, network inference, Bayesian network, expression quantitative trait loci analysis, gene expression

OPEN ACCESS

Edited by:
Shizhong Han, Johns Hopkins Medicine, United States

Reviewed by:
Bochao Jia, Eli Lilly, United States
Sungwon Jung, Gachon University, South Korea

***Correspondence:**
Tom Michoel
tom.michoel@uib.no

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

Received: *03 July 2019*
Accepted: *29 October 2019*
Published: *20 December 2019*

Citation:
Wang L, Audenaert P and Michoel T (2019) High-Dimensional Bayesian Network Inference From Systems Genetics Data Using Genetic Node Ordering. Front. Genet. 10:1196. doi: 10.3389/fgene.2019.01196

INTRODUCTION

Complex traits and diseases are driven by large numbers of genetic variants, mainly located in non-coding, regulatory DNA regions, affecting the status of gene regulatory networks (Rockman, 2008; Schadt, 2009; Civelek and Lusis, 2014; Albert and Kruglyak, 2015; Boyle et al., 2017). While important progress has been made in the experimental mapping of protein–protein and protein–DNA interactions (Walhout, 2006; Gerstein et al., 2012; Luck et al., 2017), the context-specific and

dynamic nature of these interactions means that comprehensive, experimentally validated, cell-type or tissue-specific gene networks are not readily available for human or animal model systems. Furthermore, knowledge of physical protein-DNA interactions does not always allow to predict functional effects on target gene expression (Cusanovich et al., 2014). Hence, statistical and computational methods are essential to reconstruct context-specific, causal, trait-associated networks by integrating genotype and gene, protein, and/or metabolite expression data from a large number of individuals segregating for the traits of interest (Rockman, 2008; Schadt, 2009; Civelek and Lusis, 2014).

Gene network inference is a deeply studied problem in computational biology (Friedman, 2004; Albert, 2007; Bansal et al., 2007; Penfold and Wild, 2011; Emmert-Streib et al., 2012; Marbach et al., 2012; Äijö and Bonneau, 2016; Kiani et al., 2016). Among the many successful methods that have been devised, Bayesian networks are a powerful approach for modelling causal relationships and incorporating prior knowledge (Friedman et al., 2000; Friedman, 2004; Werhli and Husmeier, 2007; Mukherjee and Speed, 2008; Koller and Friedman, 2009; Pearl, 2009). In the context of complex trait genetics, the availability of genotype data leads to an especially significant prior on the direction of causality between correlated traits, which is based on the principle that genetic variation causes variation in gene expression or disease traits, but not *vice versa* (Schadt et al., 2005). Hence, Bayesian networks have become particularly popular for modelling conditional independence and causal dependence relationships among heritable traits, including molecular abundance traits (Zhu et al., 2004; Zhu et al., 2008; Neto et al., 2010; Hageman et al., 2011; Scutari et al., 2014). Using expression quantitative trait loci (eQTL) and gene expression data as input, Bayesian networks have been used for instance to identify key driver genes of type 1 diabetes (Schadt et al., 2008), Alzheimer's disease (Zhang et al., 2013; Beckmann et al., 2018), temporal lobe epilepsy (Johnson et al., 2015), and cardiovascular disease (Talukdar et al., 2016). However, Bayesian network inference is computationally demanding and limited to relatively small-scale systems. In this paper, we address the question whether Bayesian network inference from eQTL and gene expression data is feasible on a truely transcriptome-wide scale without sacrificing performance in terms of model fit and overlap with known interactions.

A Bayesian gene network consists of a directed graph without cycles, which connects regulatory genes to their targets, and which encodes conditional independence between genes. The structure of a Bayesian network is usually inferred from the data using score-based or constraint-based approaches (Koller and Friedman, 2009). Score-based approaches maximize the likelihood of the model, or sample from the posterior distribution using Markov chain Monte Carlo (MCMC), using edge additions, deletions or inversions to search the space of network structures. Score-based methods have been shown to perform well using simulated genetics and genomics data (Zhu et al., 2007; Tasaki et al., 2015). Constraint-based approaches first learn the undirected skeleton of the network using repeated conditional independence tests, and then assign edge directions by resolving directional constraints (v-structures and acyclicity) on the skeleton. They have been used for instance in the joint genetic mapping of multiple complex traits (Scutari et al., 2014). However, the computational cost of both approaches is high. Because the number of possible graphs scales super-exponentially with the number of nodes, Bayesian gene network inference with conventional methods is feasible for systems of at most a few hundred genes or traits, and usually requires a hard limit on the number of regulators a gene can have as well as a preliminary dimension reduction step, such as filtering or clustering genes based on their expression profiles (Zhu et al., 2008; Zhang et al., 2013; Talukdar et al., 2016; Beckmann et al., 2018).

Modern sequencing technologies however generate transcript abundance data for ten-thousands of coding and non-coding genes, and large sample sizes mean that ever more of those are detected as variable across individuals (Lappalainen et al., 2013; Franzén et al., 2016; GTEx Consortium, 2017). Moreover, to explain why genetic associations are spread across most of the genome, a recently proposed "omnigenic" model of complex traits posits that gene regulatory networks are sufficiently interconnected such that all genes expressed in a disease or trait-relevant cell or tissue type affect the functions of core trait-related genes (Boyle et al., 2017). The limitations of current Bayesian gene network inference methods mean that this model can be neither tested nor accomodated. Existing Bayesian network inference methods on categorical variables, e.g., Banjo (Smith et al., 2006), lack the resolution and directionality for transcriptomic datasets. Hence, there is a clear and unmet need to infer Bayesian networks from very high-dimensional systems genetics data.

Here, we propose a novel method to infer high-quality causal gene networks that scales easily to ten-thousands of genes. Our method is based on the fact that if an ordering of nodes is given, such that the parents of any node must be a subset of the predecessors of that node in the given ordering, then Bayesian network inference reduces to a series of independent variable or feature selection problems, one for each node (Koller and Friedman, 2009; Shojaie and Michailidis, 2010). While reconstructing a node ordering is challenging in most application domains, *pairwise* comparisons between nodes can sometimes be obtained. If prior information is available for the likely inclusion of every edge, our method ranks edges according to the strength of their prior evidence (e.g., p-value) and incrementally assembles them in a directed acyclic graph (DAG) which defines a node ordering, by skipping edges that would introduce a cycle. Prior pairwise knowledge in systems biology includes the existence of TF binding motifs (Bussemaker et al., 2007), or known protein-DNA and protein-protein interactions (Ernst et al., 2008; Greenfield et al., 2013), and those have been used together with score-based MCMC methods in Bayesian network inference previously (Werhli and Husmeier, 2007; Mukherjee and Speed, 2008).

In systems genetics, where genotype and gene expression data are available for the same samples, instead of using external prior interaction data, pairwise causal inference methods can be used to estimate the likelihood of a causal interaction between every pair of genes (Schadt et al., 2005; Chen et al., 2007; Millstein

et al., 2009; Li et al., 2010; Neto et al., 2013; Millstein et al., 2016; Wang and Michoel, 2017a). To accomodate the fact that the same gene expression data is used to derive the node ordering and subsequent Bayesian network inference, we propose a novel generative model for genotype and gene expression data, given the structure of a gene regulatory graph, whose log-likelihood decomposes as a sum of the standard log-likelihood for observing the expression data and a term involving the pairwise causal inference results. Our method can then be interpreted as a greedy optimization of the posterior log-likelihood of this generative model.

METHODS

An Algorithm for the Inference of Gene Regulatory Networks From Systems Genetics Data

To allow the inference of gene regulatory networks from high-dimensional systems genetics data, we developed a method that exploits recent algorithmic developments for highly efficient mapping of eQTL and pairwise causal interactions. A general overview of the method is given here, with concrete procedures for every step detailed in subsequent sections below.

A. EQTL Mapping

When genome-wide genotype and gene expression data are sampled from the same unrelated individuals, fast matrix-multiplication based methods allow for the efficient identification of statistically significant eQTL associations (Shabalin, 2012; Qi et al., 2014; Ongen et al., 2015; Delaneau et al., 2017). Our method takes as input a list of genes, and for every gene its most strongly associated eQTL (**Figure 1A**). Typically only *cis*-acting eQTLs (i.e., genetic variants located near the gene of interest) are considered for this step, but this is not a formal requirement. Multiple genes can have the same associated eQTL, and genes without significant eQTL can be included as well, although these will only be allowed to have incoming edges in the resultant Bayesian networks.

B. Pairwise Causal Ordering

Given a set of genes and their respective eQTLs, pairwise causal interactions between all genes are inferred using the eQTLs as instrumental variables (**Figure 1B**). While there is a great amount of literature on this subject (cf. *Introduction*), only two stand-alone software packages are readily available: CIT (Millstein et al., 2016) and Findr (Wang and Michoel, 2017a). In our experience, only Findr is sufficiently efficient to test for causality between millions of gene pairs.

C. Genetic Node Ordering

In *Bayesian Network Model for Systems Genetics Data*, we introduce a generative probabilistic model for jointly observing eQTL genotypes and gene expression levels given the structure of a gene regulatory network. In this model, the posterior log-likelihood of the network given the data decomposes as a sum of two terms, one measuring the fit of the undirected network to the correlation structure of the gene expression data, and the other measuring the fit of the edge directions to the pairwise causal interactions inferred using the eQTLs as instrumental variables. The latter is optimized by a maximum-weight DAG, which induces a topological node ordering, which we term "genetic node ordering" in reference to the use of individual-level genotype data to orient pairs of gene expression traits (**Figure 1C**).

D. Bayesian Network Inference

The genetic node ordering fixes the directions of the Bayesian network edges. Variable selection methods are then used to determine the optimal sparse representation of the inverse covariance matrix of the gene expression data by a subgraph of the maximum-weight DAG (**Figure 1D**). In this paper, we consider two approaches: *(i)* a truncation of the pairwise interaction scores retaining only the most confident (highest weight) edges in the maximum-weight DAG, and *(ii)* a multi-variate, L1-penalized lasso regression (Tibshirani, 1996; Wang and Michoel, 2017b) to select upstream regulators for every gene. Given a sparse DAG, maximum-likelihood linear regression is used to determine the input functions and whether an edge is activating or repressing.

Bayesian Network Model With Prior Edge Information

A Bayesian network with n nodes (random variables) is defined by a DAG G such that the joint distribution of the variables decomposes as

$$p\left(x_1,\ldots,x_n \mid G\right)=\prod_{j=1}^{n} p\left(x_j \mid \left\{x_i : i \in \mathrm{Pa}_j\right\}\right), \tag{1}$$

where Pa_j denotes the set of parent nodes of node j in the graph G. We only consider linear Gaussian networks (Koller and Friedman, 2009), where the conditional distributions are given by normal distributions whose means depend linearly on the parent values (see **Supplementary Information**).

The likelihood of observing a data matrix $\mathbf{X}\in\mathbb{R}^{n\times m}$ with expression levels of n genes in m independent samples given a DAG G is computed as

$$p\left(\mathbf{X} \mid G\right)=\prod_{k=1}^{m}\prod_{j=1}^{n} p\left(x_{jk} \mid \left\{x_{ik} : i \in \mathrm{Pa}_j\right\}\right). \tag{2}$$

Using Bayes' theorem we can then write the likelihood of observing G given the data $\mathbf{X}$, upto a normalization constant, as

$$P\left(G \mid \mathbf{X}\right) \propto p\left(\mathbf{X} \mid G\right)P\left(G\right),$$

where $P(G)$ is the prior probability of observing G. Note that we use a lower-case 'P' to denote probability density functions and upper-case 'P' to denote discrete probability distributions.

Our method is applicable if pairwise prior information is available, i.e., for prior distributions satisfying

$$\log P(G) \propto \sum_j \sum_{i \in \mathrm{Pa}_j} f_{ij},$$

with f_{ij} a set of non-negative weights that are monotonously increasing in our prior belief that there exists a directed edge from node i to node j (e.g. $f_{ij} \propto -\log p_{ij}$, where p_{ij} is a p-value). Note that setting $f_{ij} = 0$ excludes the edge (i,j) from being present in G.

Bayesian Network Model for Systems Genetics Data

When genotype and gene expression data are available for the same samples, instrumental variable methods can be used to infer the likelihood of a causal interaction between every pair of genes (Schadt et al., 2005; Chen et al., 2007; Millstein et al., 2009; Li et al., 2010; Neto et al., 2013; Millstein et al., 2016; Wang and Michoel, 2017a). Previously, such pairwise probabilities have been used as priors in conventional score-based Bayesian network inference (Zhu et al., 2004; Zhu et al., 2007), but this is unsatisfactory, because a prior, by definition, should not be inferred from the same expression data that is used to learn the model. Other methods have addressed this by augmenting the gene network model with genotypic variables (Neto et al., 2010; Hageman et al., 2011), but this increases the size and complexity of the model even further. Here we introduce a model to use pairwise causal inference that does not suffer from these limitations.

Let G and $\mathbf{X}$ again be a DAG and a matrix of gene expression data for n genes, respectively, and let $\mathbf{E} \in \mathbb{R}^{n \times m}$ be a matrix of genotype data for the same samples. For simplicity we assume that each gene has one associated genotypic variable (e.g., its most significant *cis*-eQTL), but this can be extended easily to having more than one eQTL per gene or to some genes having no eQTLs. Using the rules of conditional probability, the joint probability (density) of observing $\mathbf{X}$ and $\mathbf{E}$ given G can be written, upto a normalization constant, as

$$p(\mathbf{X},\mathbf{E} \mid G) \propto P(\mathbf{E} \mid \mathbf{X},G)\, p(\mathbf{X} \mid G). \tag{3}$$

The distribution $p(\mathbf{X}|G)$ is obtained from the standard Bayesian network equations (eq. (2)), and we define the conditional probability of observing $\mathbf{E}$ given $\mathbf{X}$ and G as

$$P(\mathbf{E} \mid \mathbf{X},G) \propto \prod_j \prod_{i \in \mathrm{Pa}_j} P(L_i \to G_i \to G_j \mid E_i, X_i, X_j), \tag{4}$$

where E_i, $X_i \in \mathbb{R}^m$ are the ith rows of $\mathbf{E}$ and $\mathbf{X}$, respectively. $P(L_i \to G_i \to G_j | E_i, X_i, X_j)$ is the probability of a causal interaction from gene G_i to G_j inferred using G_i's eQTL L_i as a causal anchor, and can be computed with pairwise causal inference methods (Millstein et al., 2016; Wang and Michoel, 2017a). In other words, conditional on a gene-to-gene DAG G and a gene expression data matrix, our model assumes that it is more likely to observe genotype data that would lead to causal inferences consistent with G than data that would lead to inconsistent inferences. Other variations on this model can be considered as well, for instance one can include a penalty for interactions that are not present in the graph, as long as the final model can be expressed in the form

$$P(\mathbf{E} \mid \mathbf{X},G) \propto \prod_j \prod_{i \in \mathrm{Pa}_j} e^{g_{ij}}, \tag{5}$$

with g_{ij} monotonously increasing in the likelihood of a causal inference $L_i \to G_i \to G_j$.

Combining eqs. (3) and (5) with Bayes' theorem and a uniform prior $P(G) = \text{const}$, leads to an expression of the posterior log-likelihood that is formally identical to the model with prior edge information,

$$\log P(G \mid \mathbf{X},\mathbf{E}) = \log p(\mathbf{X} \mid G) + \sum_j \sum_{i \in \mathrm{Pa}_j} g_{ij} + \text{const.} \tag{6}$$

As before, if $g_{ij} = 0$, the edge (i,j) is excluded from being part of G; this would happen for instance if gene i has no associated genotypic variables and consequently zero probability of being causal for any other genes given the available data. Naturally, informative pairwise graph priors of the form $P(G) = \Sigma_j \Sigma_{i \in \mathrm{Pa}j} f_{ij}$, can still be added to the model, when such information is available.

Bayesian Network Parameter Inference

Given a DAG G, the maximum-likelihood parameters of the conditional distributions [eq. (1)], in the case of linear Gaussian networks, are obtained by linear regression of a gene on its parents' expression profiles (see **Supplementary Information**). For a specific DAG, we will use the term "Bayesian network" to refer to both the DAG itself as well as the probability distribution induced by the DAG with its maximum-likelihood parameters.

Reconstruction of the Node Ordering

Without further sparsity constraints in eq. (6), and again assuming for simplicity that each gene has exactly one eQTL, the log-likelihood is maximized by a DAG with $n(n-1)/2$ edges. Such a DAG G defines a node ordering $\prec$ where $i \prec j \Leftrightarrow i \in \mathrm{Pa}_j$. Standard results in Bayesian network theory show that for a linear Gaussian network, the likelihood function (2) is invariant under arbitrary changes of the node ordering (see (Koller and Friedman, 2009) and **Supplementary Information**). Hence to maximize eq. (6) we need to find the node ordering or DAG which maximizes the term $\Sigma_j \Sigma_{i \in \mathrm{Pa}j} g_{ij}$. Finding the maximum-weight DAG is an NP-hard problem with no known polynomial approximation algorithms with a strong guaranteed error bound (Korte and Hausmann, 1978; Hassin and Rubinstein, 1994). We therefore employed a greedy algorithm, where given n genes and the log-likelihood g_{ij} of regulation between every pair of them, we first rank the regulations according to their likelihood. The regulations are then added to an empty network one at a time starting from the most probable one, but avoiding those that would create a cycle, until a maximum-weight DAG

with $n(n-1)/2$ edges is obtained. Other edges are assigned probability 0 to indicate exclusion. The heuristic maximum-weight DAG reconstruction was implemented in Findr (Wang and Michoel, 2017a) as the command netr_one_greedy, with the *vertex-guided* algorithm for cycle detection (Haeupler et al., 2012).

Causal Inference of Pairwise Gene Regulations

We used Findr 1.0.6 (pij_gassist function) (Wang and Michoel, 2017a) to perform causal inference of gene regulatory interactions based on gene expression and genotype variation data. For every gene, its strongest *cis*-eQTL was used as a causal anchor to infer the probability of regulation between that gene and every other gene. Findr outputs posterior probabilities P_{ij} (i.e., one minus local FDR), which served directly as weights in model (6), i.e., we set $g_{ij} = \log P_{ij}$. To verify the contribution from the inferred pairwise regulations, we also generated random pairwise probability matrices which were treated in the same way as the informative ones in the downstream analyses.

Findr and Random Bayesian Networks From Node Orderings

The node ordering reconstruction removes less probable, cyclic edges, and results in a (heuristic) maximum-weight DAG G with edge weights $P_{ij} = e^{g_{ij}}$. We term these weighted DAGs as *findr* or *random Bayesian networks*, depending on the pairwise information used. A significance threshold can be applied on the continuous networks, to convert them to binary Bayesian networks at any desired sparsity level and thereby perform variable selection for the parents of every gene.

Lasso-Findr and Lasso-Random Bayesian Networks Using Penalized Regression on Ordered Nodes

As a second approach to perform variable selection in the maximum-weight DAGs, we performed hypothesis testing for every gene on whether each of its predecessors (in the findr or random Bayesian network) is a regulator, using L1-penalized lasso regression (Tibshirani, 1996) with the lassopv package (Wang and Michoel, 2017b) (see **Supplementary Information**). We calculated for every regulator the p-value of the critical regularization strength when the regulator first becomes active in the lasso path. This again forms a continuous Bayesian network in which smaller p-values indicate stronger significance. These Bayesian networks were termed the *lasso-findr* and *lasso-random Bayesian networks*.

Score-Based Bnlearn-Hc and Constraint-Based Bnlearn-Fi Bayesian Networks From Package Bnlearn

For comparison with score-based Bayesian network inference methods, we applied the hc function of the R package bnlearn (Scutari, 2010), using the Akaike information criterion (AIC) penalty to enforce sparsity. This algorithm starts from a random Bayesian network and iteratively performs greedy revisions on the network to reach a local optimum of the penalized likelihood function. Since the log-likelihood is equivalent to minus the average (over nodes) log unexplained variance (see **Supplementary Information**), which diverges when the number of regulators exceeds the number of samples, we enforced the number of regulators for every gene to be smaller than 80% of the number of samples. For each AIC penalty, one hundred random restarts were carried out and only the network with highest likelihood score was selected for downstream analyses. These Bayesian networks were termed the *bnlearn-hc* Bayesian networks.

For comparison with constraint-based Bayesian network inference methods [e.g., (Kalish and Buhlmann, 2007)], we applied the fast.iamb function of the R package bnlearn (Scutari, 2010), using nominal type I error rate. These Bayesian networks were termed the *bnlearn-fi* Bayesian networks.

To account for the role and information of cis-eQTLs on gene expression, we also included the strongest cis-eQTL of every gene in the bnlearn-based network reconstructions, for an approach similar to (Neto et al., 2010; Hageman et al., 2011; Tasaki et al., 2015). Cis-eQTLs were only allowed to have outgoing edges, using the blacklist function in bnlearn. We then removed cis-eQTL nodes from the reconstructed networks, resulting in Bayesian gene networks termed *bnlearn-hc-g* and *bnlearn-fi-g* respectively.

Evaluation of False Discovery Control in Network Inference

Scoring metrics are comparable within each hypothesis test, but not neccesarily so between different hypothesis tests. Unlike p-values, the use of arbitray scores in network inference may lead to inconsistent false positive rates of candidate regulators among different target genes, which prevents consistent network-wide false discovery control (FDC) (Wang and Michoel, 2017b). However, the network-wide FDC consistency can be evaluated with the linear relation between the numbers of false positive regulators and candidate regulators for each gene. Violation of the linearity disproves the score for FDC in network inference. Due to the (in-degree) sparsity of biological networks, we discarded the top 5% of predictions to remove true positives, after which the FDC consistency was empirically evaluated with the linear relation between the numbers of false positive and candidate regulators. See (Wang and Michoel, 2017b) for method details.

Precision-Recall Curves and Points

We compared reconstructed Bayesian networks with gold standards using precision-recall (PR) curves and points, for continuous and binary networks respectively. For Geuvadis datasets, we only included regulator and target genes that are present in both the transcriptomic dataset and the gold standard.

Assessment of Predictive Power for Bayesian Networks

To assess the predictive power of different Bayesian network inference methods, we used five-fold cross-validation to compute the training and testing errors from each method, in terms of the root mean squared error (rmse) and mean log squared error (mlse) across all genes in all testing data (**Supplementary Information**, **Algorithm S1**). For continuous Bayesian networks from non-bnlearn methods, we applied different significance thresholds to obtain multiple binary Bayesian networks that form a curve of prediction errors.

Data and Software

We used the following datasets to infer and evaluate Bayesian gene networks:

- The DREAM 5 Systems Genetics challenge A (DREAM) provided a unique testbed for network inference methods that utilize genetic variations in a population (https://www.synapse.org/\#!Synapse:syn2820440/wiki/). The DREAM challenge included 15 simulated datasets of expression levels of 1000 genes and their best eQTL variations. To match the high-dimensional property of real datasets where the number of genes exceeds the number of individuals, we analyzed datasets 1, 3, and 5 with 100 individuals each. Around 25% of the genes within each dataset had a cis-eQTL, defined in DREAM as directly affecting the expression level of the corresponding gene. Since the identity of cis-eQTLs is not revealed, we used kruX (Qi et al., 2014) to identify them, allowing for one false discovery per dataset. The DREAM challenge further provides the groundtruth network for each dataset, varying from around 1,000 to 5,000 interactions.
- The Geuvadis consortium is a population study providing RNA sequencing and genotype data of lymphoblastoid cell lines in 465 individuals. We obtained gene expression levels and genotype information, as well as the eQTL mapping from the original study (Lappalainen et al., 2013). We limited our analysis to 360 European individuals, and after quality control, a total of 3172 genes with significant cis-eQTLs remained. To validate the inferred gene regulatory networks from the Geuvadis dataset, we obtained three groundtruth networks: (Rockman, 2008) differential expression data from siRNA silencing experiments of transcription-associated factors (TFs) in a lymphoblastoid cell line (GM12878) (Cusanovich et al., 2014); (Schadt, 2009) DNA-binding information of TFs in the same cell line (Cusanovich et al., 2014); (Civelek and Lusis, 2014) the filtered proximal TF-target network from (Gerstein et al., 2012). The Geuvadis dataset overlapped with 6,790 target genes, and 6 siRNA-targeted TFs and 20 DNA-binding TFs in groundtruth 1 and 2, respectively, and with 7,000 target genes and 14 TFs in groundtruth 3. Processed Geuvadis data and groundtruth networks are available at https://github.com/lingfeiwang/findr-data-geuvadis

We preprocessed all expression data by converting them to a standard normal distribution separately for each gene, as explained in (Wang and Michoel, 2017a).

Software to reproduce the results from this study is available at the following URLs:

- Findr: https://github.com/lingfeiwang/findr.
- lassopv: https://github.com/lingfeiwang/lassopv.

RESULTS

Genetic Node Ordering Permits High-Dimensional Bayesian Network Inference

We developed a method for Bayesian network inference from high-dimensional systems genetics data which reconstructs a maximum-weight DAG from the confidence scores of pairwise causal inferences between gene expression traits using eQTLs as causal anchors, and which uses the node ordering induced by this DAG (termed "genetic node ordering" in reference to the use of genotype data to orient network edges) to decompose the Bayesian network inference task into a series of independent variable selection problems (*Methods*, *An Algorithm for the Inference of Gene Regulatory Networks From Systems Genetics Data*, **Figure 1**). Using an efficient implementation for the causal inference step (Wang and Michoel, 2017a), this approach allows to reconstruct Bayesian networks with thousands to ten-thousands of nodes. Our method is based on score-based Bayesian network inference methods for systems with pre-defined node orderings (Koller and Friedman, 2009; Shojaie and Michailidis, 2010), but differs in that the ordering is inferred from the same expression data, augmented with matched genotype data from the same samples, that is used for the subsequent Bayesian network log-likelihood maximization, using a single generative model (*Methods*, *Bayesian Network Model for Systems Genetics Data*), rather than relying on external prior information to determine the node ordering. Its computational efficiency is due to restricting the graph structure search space to Bayesian gene networks compatible with this inferred node ordering. This differs substantially from conventional score-based and constraint-based methods, including those that use genotype and gene expression data (Neto et al., 2010; Hageman et al., 2011; Tasaki et al., 2015), where the search space can only be reduced by limiting the possible number of parents for each gene to an artificially small number (Koller and Friedman, 2009). For clarity, a comparison of the main characteristics of the Bayesian network inference approaches considered in this paper is included in **Supplementary Table S1**.

Lasso-Findr Bayesian Networks Correctly Control False Discoveries

We inferred findr and lasso-findr Bayesian networks for the DREAM datasets, using Findr and lassopv respectively (*Methods*). The Findr method predicts targets for each regulator using a local FDR score (Storey and Tibshirani, 2003) which allows consistent, network-wide FDC (Chen et al., 2007; Wang and Michoel, 2017a). However, the enforcement of a gene ordering/Bayesian network partly broke the FDC, as the linearity between the numbers of false positive (i.e., significant here) and candidate regulators broke down at large candidate regulator counts (**Figure 2A**,

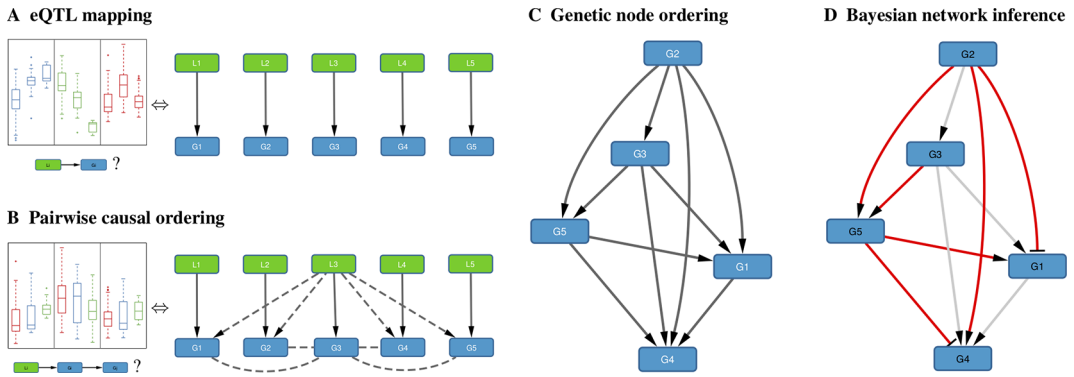

FIGURE 1 | Schematic overview of the method. **(A)** For each gene G_i, the *cis*-eQTL L_i whose genotype explains most of the variation in G_i expression is calculated; shown on the left are typical eQTL associations for three genes (colored blue, green, and red) where each box shows the distribution of expression values for samples having a particular genotype for that gene's eQTL. **(B)** Pairwise causal inference is carried out which considers in turn each gene G_i and its eQTL L_i to calculate the likelihood of this gene being causal for all others; shown on the left is a typical example where an eQTL L_i is associated with expression of G_i (red) and with expression of a correlated gene G_j (blue), but not with expression of G_j adjusted for G_i (green), resulting in a high likelihood score for the causal ordering $G_i \rightarrow G_j$. **(C)** A maximum-weight DAG having the genes as its nodes is derived from the pairwise causal interactions, which induces a "genetic" node ordering. **(D)** Variable selection is used to determine a sparse Bayesian gene network, which must be a sub-graph of the maximum-weight graph (red edges, Bayesian network; gray edges, causal orderings deemed not significant or indirect by the variable selection procedure); the signs of the maximum-likelihood linear regression coefficients determine whether an edge is activating (arrows) or repressing (blunt tips).

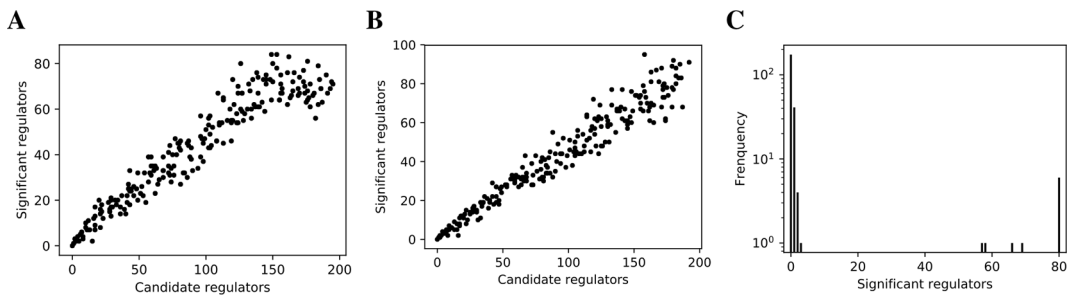

FIGURE 2 | False discovery controls of different Bayesian networks. **(A, B)** The linearity test of findr **(A)** and lasso-findr **(B)** Bayesian networks at 10,000 significant interactions on DREAM dataset 1. **(C)** The histogram of significant regulator counts for each target gene in the bnlearn-hc Bayesian network with AIC penalty 8 on DREAM dataset 1.

Methods). This effect is confirmed on the larger Geuvadis dataset in *Results on the Geuvadis Dataset Reaffirm Conclusions From Simulated Data*. By performing an extra lasso regression on top of the acyclic findr network, proper FDC was restored in terms of the linear relation in the lasso-findr Bayesian network (**Figure 2B**, **Supplementary Figure S1**).

In contrast, score-based bnlearn-hc Bayesian networks (*Methods*), inferred from multiple DREAM datasets and for a spectrum of network sparsities (AIC penalty strengths from 8 to 12 in steps of 0.5), displayed a highly skewed in-degree distribution, with most genes having few regulators, but several with near 80 regulators each, i.e., the maximum allowed (**Figure 2C**, **Supplementary Figure S2**). This is in conflict with the known in-degree sparsity of gene regulation networks, which is required for its modularity, indicating that score-based Bayesian networks lack a unified FDR control, i.e., that each gene retained incoming interactions at different FDR levels. We believe this is due to the log-likelihood score function employed by bnlearn-hc. Since the log-likelihood corresponds to the average logarithm of the unexplained variance, this score intrinsically tends to focus on the explanation of variances from a few variables/genes, especially in high-dimensional settings where this can lead to arbitrarily large score values (see **Supplementary Information**). Using the total proportion of explained variance as the score may spread regulations over more target genes, but this score is not implemented in bnlearn.

Constraint-based bnlearn-fi Bayesian networks (*Methods*) did not allow for unbiased FDC either, as they do not have a fully adjustable sparsity level. We varied its "nominal type I error rate" from 0.001 to 0.2, but the number of significant interactions varied very little on DREAM dataset 1 (**Supplementary Figure S3**).

Incorporating genotypic information in score-based (bnlearn-hc-g) or constraint-based (bnlearn-fi-g) Bayesian networks did not resolve these issues, as the problems of lacking FDC and oversparsity persisted (**Supplementary Figure S4**, **Supplementary Figure S5**).

Findr and Lasso Bayesian Networks Recover Genuine Interactions More Accurately Than MCMC or Constraint-Based Networks

We compared the inferred Bayesian networks from all methods against the groundtruth network of the DREAM challenge. We drew PR curves, or points for the binary Bayesian networks from bnlearn-based methods, as shown in **Figure 3** with areas under the PR curve (AUPR) in **Supplementary Table S2**. Bnlearn based methods could only recover ∼2% of total true regulations, after which they suffered from a sharply dropping precision and behaved like random predictions. The highest precisions they achieved could not exceed those by lasso or findr based methods at the respective recalls either. In addition, bnlearn could not obtain >10% recall within 4-day time limit with any of the methods attempted. In this sense, the findr, lasso-findr, and lasso-random Bayesian networks were more accurate predictors of the underlying network structure. The inclusion of genotypic information improved the precision of bnlearn methods, but it remained suboptimal than findr and lasso-based Bayesian networks.

Findr and Lasso Bayesian Networks Obtain Superior Predictive Performances

We validated the predictive performances of all networks in the structural equation context (see **Supplementary Information**). Under five-fold cross validation, a linear regression model for each gene on its parents is trained based on the Bayesian network structure inferred from each training set, to predict expression levels of all genes in the test set (*Methods*). Predictive errors were measured in terms of rmse and mlse (the score optimized by bnlearn-hc). The findr Bayesian network explained the highest proportion of expression variation (≈2%) in the test data and identified the highest number of regulations (200 to 300), with runners up from lasso-based networks (≈1% variation, 50 regulations, **Figure 4**). The explained variance by findr and lasso networks grew to ≈10% when more samples were added (DREAM dataset 11 with 999 samples, **Supplementary Figure S6**). Training errors did not show overfitting of predictive performances in the test data (**Supplementary Figure S7**).

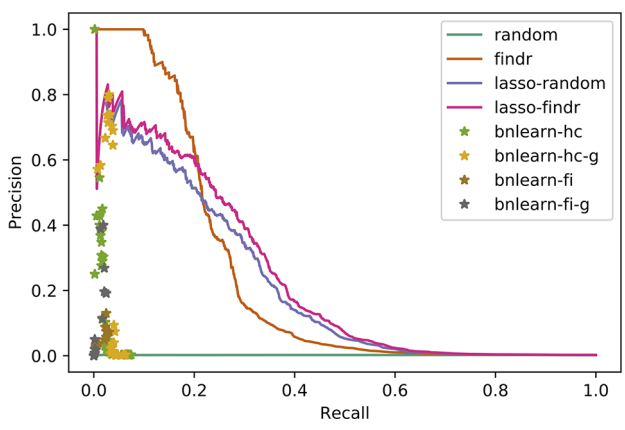

FIGURE 3 | Precision-recall curves/points of reconstructed Bayesian networks for DREAM dataset 1.

Lasso Bayesian Networks Do Not Need Accurate Prior Gene Ordering

Interestingly, the performance of lasso-based networks did not depend strongly on the prior ordering, as shown in the comparisons between lasso-findr and lasso-random in **Figure 3**, **Figure 4**, and **Supplementary Figure S7**. Further inspections revealed a high overlap of top predictions by lasso-findr and lasso-random Bayesian networks, particularly among their true positives (**Figure 5**). This suggests that lasso may be capable of prioritizing edges with correct directions, and allows us to still recover genuine interactions even if the prior gene ordering is not fully accurate.

Lasso Bayesian Networks Mistake Confounding as False Positive Interactions

We then tried to understand the differences between lasso and Findr based Bayesian networks, by comparing three types of gene relations in DREAM dataset 1, both among genes with a cis-eQTL in **Figure 6A**, and when also including genes without any cis-eQTL as only targets in **Figure 6B**. Both findr and lasso-findr showed good sensitivity for the genuine, direct interactions. However, when two otherwise independent genes are directly confounded by another gene, lasso tends to produce a false positive interaction, but not findr. As expected, to achieve optimal predictive performance, lasso regression cannot distinguish the confounding by a gene that is either unknown or ranked lower in the DAG.

Findr and Lasso Bayesian Network Inference Is Highly Efficient

The findr and lasso Bayesian networks required much less computation time compared to the bnlearn Bayesian networks, therefore allowing them to be applied on much larger datasets. To infer a Bayesian network of 230 genes from 100 samples in DREAM dataset 1, Findr required less than a second, lassopv around a minute, but bnlearn Bayesian networks took half an hour to half a day (**Table 1**). Moreover, since bnlearn only produces binary Bayesian networks, multiple recomputation is necessary to acquire the desired network sparsity.

Results on the Geuvadis Dataset Reaffirm Conclusions From Simulated Data

To test whether the results from the DREAM data also hold for real data, we inferred findr and lasso-findr Bayesian networks from the Geuvadis data using both real and random causal priors (see *Methods*); conventional bnlearn-based network inference was attempted, but none of the restarts could complete within 1000 min.

Lasso-findr Bayesian networks were previously shown to provide ideal FDR control on this dataset (Wang and Michoel, 2017b), whereas findr Bayesian networks did not obtain a satisfying FDR control (**Supplementary Figure S8**). We believe this is due to the reconstruction of the node ordering, which interferes with the FDR control in pairwise causal inference. On the other hand, and again consistent with the DREAM data, findr

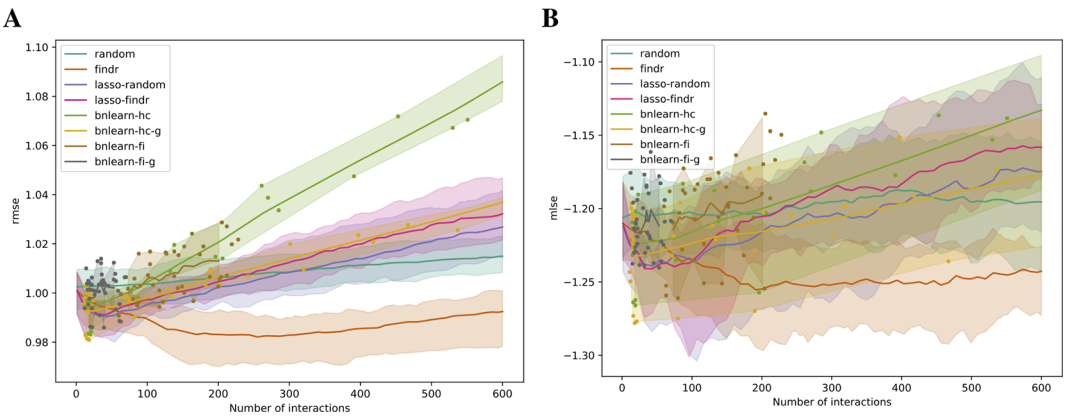

FIGURE 4 | The root mean squared error **(**rmse, **A)** and mean log squared error **(**mlse, **B)** in test data are shown as functions of the numbers of predicted interactions in five-fold cross validations using linear regression models. Shades and lines indicate minimum/maximum values and means respectively. RMSEs greater than 1 indicate over-fitting. DREAM dataset 1 with 100 samples was used.

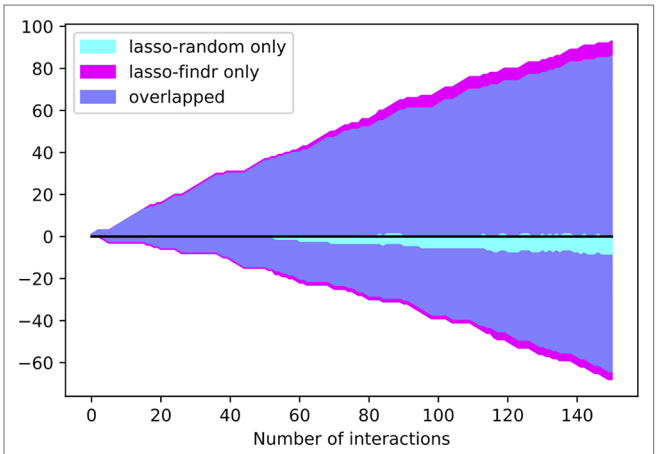

FIGURE 5 | The numbers of overlap and unique interactions (*y* axis) predicted by lasso-findr and lasso-random Bayesian networks as functions of the number of significant interactions in each network (*x* axis), on DREAM dataset 1. Positive and negative directions in *y* correspond to true and false positive interactions according to the gold standard.

Bayesian networks obtained superior results for the recovery of known transcriptional regulatory interactions inferred from ChIP-sequencing data (**Figures 7A**, **B**); neither method predicted TF targets inferred from siRNA silencing with high scores or accuracy better than random (**Figure 7C**).

Comparisons on the predictive power yielded results similar with the DREAM datasets, where predictive scores were again hardly able to distinguish network directions.

DISCUSSION

The inference of Bayesian gene regulatory networks for mapping the causal relationships between thousands of genes expressed in any given cell type or tissue is a challenging problem, due to the computational complexity of conventional hill-climbing, MCMC sampling or constraint-based methods. Here we have introduced an alternative method, which first reconstructs a topological ordering of genes, and then infers a sparse maximum-likelihood Bayesian network using variable selection of parents for every gene from its predecessors in the ordering. Our method is applicable

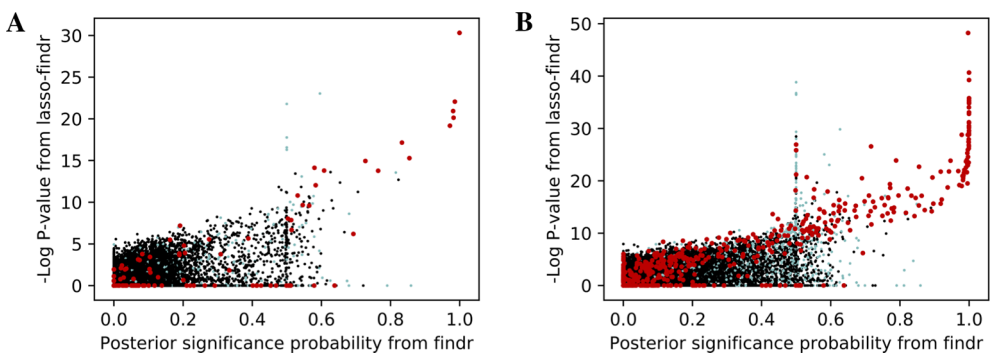

FIGURE 6 | The significance score of findr (posterior probability; x-axis) and in lasso-findr (-log P-value; y-axis) for direct true interactions (red), directly confounded gene pairs (cyan), and other, unrelated gene pairs (black) on DREAM dataset 1; in **(A)** only genes with cis-eQTLs are considered as regulator or target, whereas in **(B)** targets also include genes without cis-eQTLs. Higher scores indicate stronger significances for the gene pair tested.

TABLE 1 | Timings for different Bayesian network inference methods/programs.

Dataset	Samples	Genes	Findr	lassopv	bnlearn-hc	bnlearn-fi
DREAM	100	230	< 1 s	≈1 min	≥10 h	≥30 min
Geuvadis	360	3172	< 1 min	≈10 h	–	–

Times for bnlearn methods depend on parameter settings (e.g., nominal FDR and AIC penalty), and take longer (approx. 8 times) with genotypes included. Times for bnlearn-hc include 10 random restarts.

when pairwise prior information is available or can be inferred from auxiliary data, such as genotype data. Our evaluation of the method using simulated genotype and gene expression data from the DREAM5 competition, and real data from human lymphoblastoid cell lines from the GEUVADIS consortium, revealed several lessons that we believe to be generalizable.

A major disadvantage of conventional score-based methods, irrespective of their computational cost, was their over-fitting of the expression profiles of a very small number of target genes. In high-dimensional settings where the number of genes far exceeds the number of samples, the expression profile of any one of them can be regressed perfectly (i.e., with zero residual error) on any linearly independent subset of variables, and this causes the log-likelihood to diverge. Even when the number of parents per gene was restricted to less than the number of samples, it remained the case that at any level of network sparsity, the divergence of the log-likelihood with decreasing residual variance of even a single gene resulted in score-based networks where most genes had either the maximum number of parents, or no parents at all. Restricting the maximum number of parents to an artificially small level can circumvent this problem, but will also distort the network topology, particularly by truncating the in-degree distribution, and therefore predict a biased gene regulatory network. Optimizing the total amount of variance explained, rather than log-likelihood, might overcome this problem. This, however, is not available yet in bnlearn.

Our method reconstructs a Bayesian network as a sparse subgraph from a maximum-weight DAG determined by pairwise causal relationships inferred using instrumental variable methods. We considered two variants of the method: one where the edge weights in the maximum-weight DAG were truncated directly to form a sparse DAG, and one where an additional L1-penalized lasso regression step was used to enforce sparsity. The lasso step was introduced for two reasons. First, pairwise relations do not distinguish between direct or indirect interactions and do not account for the possibility that a true relation may only explain a small proportion of target gene variation (e.g. when the target has multiple inputs). We hypothesized that adding a multi-variate lasso regression step could address these limitations. Second, truncating pairwise relations results in non-uniform false discovery rates for the retained interactions, due to each gene starting with a different number of candidate parents in the pairwise node ordering. As we showed in this paper and our previous work (Wang and Michoel, 2017b), a model selection p-value derived from lasso regression can control the FDR uniformly for each potential regulator of each target gene, resulting in an unbiased sparse DAG.

Despite these considerations, the "naïve" procedure of truncating the original pairwise causal probabilities resulted in Bayesian networks with better overlap with groundtruth networks of known transcriptional interactions, in both simulated and real data. We believe this is due to the lack of any instrumental variables in lasso regression, which makes it hard to dissociate true causal interactions from hidden confounding. Indeed, it is known that if there are multiple strongly correlated predictors, lasso regression will randomly select one of them (Zou and Hastie, 2005), whereas in the present context it would be better to select the one that has the highest prior causal evidence. In a real biological system, findr networks and the use of instrumental variables may therefore be more robust than lasso regression, particularly in the presence of hidden confounders. We also note that the deviation from uniform FDR control for the naive truncation method was not huge and only affected genes with a very large number of candidate parents (**Figure 2**). Hence, at least in the datasets studied, adding a lasso step for better FDC did not overcome the limitations introduced by confounding interactions.

On the other hand, the lasso-random network used solely transcriptomic profiles, yet provided better performance than the conventional score-based and constrained-based networks, including those that used genotypic information. Together with its better FDC, this makes the lasso-random network an interesting method for high-dimensional Bayesian network inference with no or limited prior information.

In addition to comparing the inferred network structure against known ground-truths, we also compared the predictive performance of the various Bayesian networks. Although findr Bayesian networks again performed best, differences with lasso-based methods were modest. As is well known, using observational data alone, Bayesian networks are only defined upto Markov equivalence (Koller and Friedman, 2009; Pearl, 2009), i.e., there is usually a large class of Bayesian networks with very different topology which all explain the

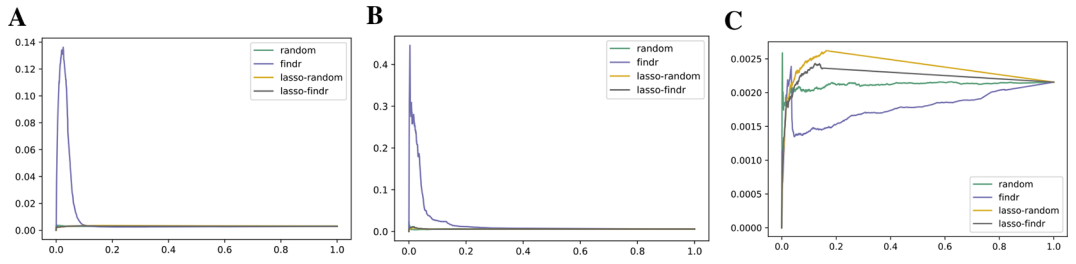

FIGURE 7 | Precision-recall curves for Bayesian networks reconstructed from the Geuvadis dataset for three groundtruth networks: DNA-binding of 20 TFs in GM12878 **(A)**, DNA-binding of 14 TFs in five ENCODE cell lines **(B)**, and siRNA silencing of six TFs in GM12878 **(C)**.

data equally well. Hence, it comes as no surprise that the prediction accuracy in edge directions has little impact on that in expression levels. This suggests that for the task of reconstructing gene networks, Bayesian network inference should be evaluated, and maybe also optimized, at the structural rather than inferential level. This also reinforces the importance of causal inference which, although challenging both statistically and computationally, demonstrated significant improvement of the global network structure even when it was restricted to pairwise causal tests.

Most of our results were derived for simulated data from the DREAM Challenges, but were qualitatively confirmed using data from human lymphoblastoid cell lines. This is because human ground-truth networks have strong limitations. They are normally reconstructed from heterogeneous, noisy, high-throughput data (e.g., ChIP-sequencing and/or knock-out experiments), and are both incomplete (many true interactions are not present) and imperfect (many detected physical interactions have no functional effect). In addition, statistical inference algorithms can hardly distinguish direct interactions from indirect ones, which operate through an unidentified third factor and should be regarded as "false positives". As such, one has to be cautious not to over-interpret results, for instance on the relative performance of findr vs. lasso-findr Bayesian networks. Much more comprehensive and accurate ground-truth networks of direct causal interactions, preferably derived from a hierachy of interventions on a much wider variety of genes and functional classes (not only transcription factors), would be required for a conclusive analysis. Emerging large-scale perturbation compendia such as the expanded Connectivity Map, which has profiled knock-downs or over-expressions of more than 5,000 genes in a variable number of cell lines using a reduced representation transcriptome (Subramanian et al., 2017), hold great promise. However, the available cell lines are predominantly cancer lines, and the relevance of the profiled interactions for systems genetics studies of human complex traits and diseases, which are usually performed on primary human cell or tissue types, remains unknown.

Lastly, we note that our study has focused on ground-truth comparisons and predictive performances, but did not evaluate how well the second part of the log-likelihood, derived from the genotype data [cf. eq. (4)], was optimized. This score is never considered in the conventional score-based algorithms, and hence a comparison would not be fair. Moreover, optimising it is known to be an NP-hard problem. We used a common greedy heuristic optimization algorithm, but for this particular problem, this heuristic has no strong guaranteed error bound. We intend to revisit this problem, and investigate whether other graph-theoretical algorithms, perhaps tailored to specific characteristics of pairwise interactions inferred from systems genetics data, are able to improve on the greedy heuristic.

To conclude, Bayesian network inference using pairwise genetic node ordering is a highly efficient approach for reconstructing gene regulatory networks from high-dimensional systems genetics data, which outperforms conventional methods by restricting the super-exponential graph structure search space to acyclic graphs compatible with the causal inference results, and which is sufficiently flexible to integrate other types of pairwise prior data when they are available.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.synapse.org//#!Synapse:syn2820440/wiki/, https://www.ebi.ac.uk/arrayexpress/files/E-GEUV-1/analysis_results/, https://github.com/lingfeiwang/findr-data-geuvadis. Findr: https://github.com/lingfeiwang/findr-R and https://github.com/lingfeiwang/findr, lassopv: https://github.com/lingfeiwang/lassopv, bnlearn: http://www.bnlearn.com/.

AUTHOR CONTRIBUTIONS

Conceptualization: LW, TM. Data curation: LW. Formal analysis: LW, PA, TM. Funding acquisition: TM. Investigation: LW, PA, TM. Methodology: LW, PA, TM. Software: LW. Supervision: TM. Writing: LW, PA, TM.

FUNDING

This work was supported by the BBSRC (grant numbers BB/J004235/1 and BB/M020053/1).

ACKNOWLEDGMENTS

This manuscript has been released as a pre-print at bioRxiv (Wang et al., 2019).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01196/full#supplementary-material

REFERENCES

Äijö, T., and Bonneau, R. (2016). Biophysically motivated regulatory network inference: progress and prospects. *Hum. Heredity* 81 (2), 62–77. doi: 10.1159/000446614

Albert, F. W., and Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.* 16, 197–212. doi: 10.1038/nrg3891

Albert, R. (2007). Network inference, analysis, and modeling in systems biology. *Plant Cell* 19 (11), 3327–3338. doi: 10.1105/tpc.107.054700

Bansal, M., Belcastro, V., Ambesi-Impiombato, A., and di Bernardo, D. (2007). How to infer gene networks from expression profiles. *Mol. Syst. Biol.* 3, 78. doi: 10.1038/msb4100120

Beckmann, N. D., Lin, W. J., Wang, M., Cohain, A. T., Wang, P., Ma, W., et al. (2018). Multiscale causal network models of Alzheimer's disease identify VGF as a key regulator of disease. *bioRxiv* p, 458430. doi: 10.1101/458430

Boyle, E. A., Li, Y. I., and Pritchard, J. K. (2017). An expanded view of complex traits: from polygenic to omnigenic. *Cell.* 169 (7), 1177–1186. doi: 10.1016/j.cell.2017.05.038

Bussemaker, H. J., Foat, B. C., and Ward, L. D. (2007). Predictive modeling of genome-wide mRNA expression: from modules to molecules. *Annu. Rev. Biophys. Biomol. Struct.* 36, 329–347. doi: 10.1146/annurev.biophys.36.040306.132725

Chen, L. S., Emmert-Streib, F., and Storey, J. D. (2007). Harnessing naturally randomized transcription to infer regulatory relationships among genes. *Genome Biol.* 8 (10), R219. doi: 10.1186/gb-2007-8-10-r219

Civelek, M., and Lusis, A. J. (2014). Systems genetics approaches to understand complex traits. *Nat. Rev. Genet.* 15 (1), 34–48. doi: 10.1038/nrg3575

Cusanovich, D. A., Pavlovic, B., Pritchard, J. K., and Gilad, Y. (2014). The functional consequences of variation in transcription factor binding. *PloS Genet.* 10 (3), e1004226. doi: 10.1371/journal.pgen.1004226

Cusanovich, D. A., Pavlovic, B., Pritchard, J. K., and Gilad, Y. (2014). The functional consequences of variation in transcription factor binding. *PloS Genet.* 10 (3), e1004226. doi: 10.1371/journal.pgen.1004226

Delaneau, O., Ongen, H., Brown, A. A., Fort, A., Panousis, N. I., and Dermitzakis, E. T. (2017). A complete tool set for molecular QTL discovery and analysis. *Nat. Commun.* 8, 15452. doi: 10.1038/ncomms15452

Emmert-Streib, F., Glazko, G., Altay, G., and De Matos Simoes, R. (2012). Statistical inference and reverse engineering of gene regulatory networks from observational expression data. *Front. Genet.* 3, 8. doi: 10.3389/fgene.2012.00008

Ernst, J., Beg, Q. K., Kay, K. A., Bala'zsi, G., Oltvai, Z. N., and Bar-Joseph, Z. (2008). A semi-supervised method for predicting transcription factor - gene interactions in Escherichia coli. *PloS Comp. Biol.* 4, e1000044. doi: 10.1371/journal.pcbi.1000044

Franzén, O., Ermel, R., Cohain, A., Akers, N., Di Narzo, A., Talukdar, H., et al. (2016). Cardiometabolic risk loci share downstream cis and trans genes across tissues and diseases. *Science* 827–830. doi: 10.1126/science.aad6970

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *J. Comput. Biol.* 7, 601–620. doi: 10.1089/106652700750050961

Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science* 308, 799–805. doi: 10.1126/science.1094068

Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K. K., Cheng, C., et al. (2012). Architecture of the human regulatory network derived from ENCODE data. *Nature* 489 (7414), 91–100. doi: 10.1038/nature11245

Greenfield, A., Hafemeister, C., and Bonneau, R. (2013). Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. *Bioinformatics.* 29 (8), 1060–1067. doi: 10.1093/bioinformatics/btt099

GTEx Consortium. (2017). Genetic effects on gene expression across human tissues. *Nature* 550 (7675), 204. doi: 10.1038/nature24277

Haeupler, B., Kavitha, T., Mathew, R., Sen, S., and Tarjan, R. E. (2012). Incremental cycle detection, topological ordering, and strong component maintenance. *ACM Trans. Algorithms* 8 (1), 3:1–3:33. doi: 10.1145/2071379.2071382

Hageman, R. S., Leduc, M. S., Korstanje, R., Paigen, B., and Churchill, G. A. (2011). A Bayesian framework for inference of the genotype-phenotype map for segregating populations. *Genetics* 187 (4), 1163–1170. doi: 10.1534/genetics.110.123273

Hassin, R., and Rubinstein, S. (1994). Approximations for the maximum acyclic subgraph problem. *Inf. Process. Lett.* 51 (3), 133–140. doi: 10.1016/0020-0190(94)00086-7

Johnson, M. R., Behmoaras, J., Bottolo, L., Krishnan, M. L., Pernhorst, K., Santoscoy, P. L. M., et al. (2015). Systems genetics identifies Sestrin 3 as a regulator of a proconvulsant gene network in human epileptic hippocampus. *Nat. Commun.* 6, 6031. doi: 10.1038/ncomms7031

Kalisch, M., and Bühlmann, P. (2007). Estimating High-Dimensional Directed Acyclic Graphs with the PC Algorithm. *J. Mach. Learn. Res.* 8, 613–636.

Kiani, N. A., Zenil, H., Olczak, J., and Tegnér, J. (2016). Evaluating network inference methods in terms of their ability to preserve the topology and complexity of genetic networks. *Semin. Cell Dev. Biol.* 51, 44–52. doi: 10.1016/j.semcdb.2016.01.012

Koller, D., and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques* (Cambridge, MA, USA: The MIT Press).

Korte, B., and Hausmann, D. (1978). "An analysis of the greedy heuristic for independence systems," in *Annals of Discrete Mathematics*, vol. 2. 65–74. doi: 10.1016/S0167-5060(08)70322-4

Lappalainen, T., Sammeth, M., Friedländer, M. R., AC't Hoen, P., Monlong, J., Rivas, M. A., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501, 506–511. doi: 10.1038/nature12531

Lappalainen, T., Sammeth, M., Friedlander, M. R., Hoen, P. A. C., Monlong, J., Rivas, M. A., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 09501 (7468), 506–511. doi: 10.1038/nature12531

Li, Y., Tesson, B. M., Churchill, G. A., and Jansen, R. C. (2010). Critical reasoning on causal inference in genome wide linkage and association studies. *Trends Genet.* 26 (12), 493–498. doi: 10.1016/j.tig.2010.09.002

Luck, K., Sheynkman, G. M., Zhang, I., and Vidal, M. (2017). Proteome-scale human interactomics. *Trends Biochem. Sci.* 42 (5), 342–354. doi: 10.1016/j.tibs.2017.02.006

Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., et al. (2012). Wisdom of crowds for robust gene network inference. *Nat. Methods* 9 (8), 796–804. doi: 10.1038/nmeth.2016

Millstein, J., Zhang, B., Zhu, J., and Schadt, E. E. (2009). Disentangling molecular relationships with a causal inference test. *BMC Genet.* 10 (1), 23. doi: 10.1186/1471-2156-10-23

Millstein, J., Chen, G. K., and Breton, C. V. (2016). cit: hypothesis testing software for mediation analysis in genomic applications. *Bioinformatics* 32, 2364–2365. doi: 10.1093/bioinformatics/btw135

Mukherjee, S., and Speed, T. P. (2008). Network inference using informative priors. *Proc. Natl. Acad. Sci.* 105 (38), 14313–14318. doi: 10.1073/pnas.0802272105

Neto, E. C., Keller, M. P., Attie, A. D., and Yandell, B. S. (2010). Causal graphical models in systems genetics: a unified framework for joint inference of causal network and genetic architecture for correlated phenotypes. *Ann. Appl. Stat.* 4 (1), 320. doi: 10.1214/09-AOAS288

Neto, E. C., Broman, A. T., Keller, M. P., Attie, A. D., Zhang, B., Zhu, J., et al. (2013). Modeling causality for pairs of phenotypes in system genetics. *Genetics* 193 (3), 1003–1013. doi: 10.1534/genetics.112.147124

Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T., and Delaneau, O. (2015). Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* 32 (10), 1479–1485. doi: 10.1093/bioinformatics/btv722

Pearl, J. (2009). *Causality* (Cambridge, UK: Cambridge University Press). doi: 10.1017/CBO9780511803161

Penfold, C. A., and Wild, D. L. (2011). How to infer gene networks from expression profiles, revisited. *Interface Focus.* 1 (6), 857–870. doi: 10.1098/rsfs.2011.0053

Qi, J., Foroughi Asl, H., Björkegren, J. L. M., and Michoel, T. (2014). kruX: Matrix-based non-parametric eQTL discovery. *BMC Bioinf.* 15, 11. doi: 10.1186/1471-2105-15-11

Qi, J., Foroughi Asl, H., Bjorkegren, J., and Michoel, T. (2014). kruX: matrix-based non-parametric eQTL discovery. *BMC Bioinf.* 15 (1), 11. doi: 10.1186/1471-2105-15-11

Rockman, M. V. (2008). Reverse engineering the genotype-phenotype map with natural genetic variation. *Nature* 456 (7223), 738–744. doi: 10.1038/nature07633

Schadt, E. E., Lamb, J., Yang, X., Zhu, J., Edwards, S., GuhaThakurta, D., et al. (2005). An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.* 37 (7), 710–717. doi: 10.1038/ng1589

Schadt, E. E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P. Y., et al. (2008). Mapping the genetic architecture of gene expression in human liver. *PloS Biol.* 6, e107. doi: 10.1371/journal.pbio.0060107

Schadt, E. E. (2009). Molecular networks as sensors and drivers of common human diseases. *Nature* 461, 218–223. doi: 10.1038/nature08454

Scutari, M., Howell, P., Balding, D. J., and Mackay, I. (2014). Multiple quantitative trait analysis using Bayesian networks. *Genetics* 198 (1), 129–137. doi: 10.1534/genetics.114.165704

Scutari, M. (2010). Learning Bayesian Networks with the bnlearn R Package. *J. Stat. Software* 35 (1), 1–22.

Shabalin, A. A. (2012). Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* 28 (10), 1353–1358. doi: 10.1093/bioinformatics/bts163

Shojaie, A., and Michailidis, G. (2010). Penalized likelihood methods for estimation of sparse high dimensional directed acyclic graphs. *Biometrika.* 97 (3), 519–538. doi: 10.1093/biomet/asq038

Smith, V. A., Yu, J., Smulders, T. V., Hartemink, A. J., and Jarvis, E. D. (2006). Computational Inference of Neural Information Flow Networks. *PloS Comput. Biol.* 2 (11), e161. doi: 10.1371/journal.pcbi.0020161

Storey, J. D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci.* Aug100 (16), 9440–9445. doi: 10.1073/pnas.1530509100

Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., et al. (2017). A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* 171 (6), 1437–1452. doi: 10.1016/j.cell.2017.10.049

Talukdar, H., Foroughi Asl, H., Jain, R., Ermel, R., Ruusalepp, A., Franzén, O., et al. (2016). Cross-tissue regulatory gene networks in coronary artery disease. *Cell Syst.* 2, 196–208. doi: 10.1016/j.cels.2016.02.002

Tasaki, S., Sauerwine, B., Hoff, B., Toyoshiba, H., Gaiteri, C., and Neto, E. C. (2015). Bayesian network reconstruction using systems genetics data: comparison of MCMC methods. *Genetics* 199 (4), 973–989. doi: 10.1534/genetics.114.172619

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B. (Methodol.)* p, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Walhout, A. J. (2006). Unraveling transcription regulatory networks by protein-DNA and protein-protein interaction mapping. *Genome Res.* 16 (12), 1445–1454. doi: 10.1101/gr.5321506

Wang, L., and Michoel, T. (2017a). Efficient and accurate causal inference with hidden confounders from genome-transcriptome variation data. *PloS Comput. Biol.* 13 (8), e1005703. doi: 10.1371/journal.pcbi.1005703

Wang, L., and Michoel, T. (2017b). doi: 10.1101/288217 Controlling false discoveries in Bayesian gene networks with lasso regression p-values. arXiv:170107011 q-bio, stat.Jan;ArXiv: 1701.07011. Available from: http://arxiv.org/abs/1701.07011.

Wang, L., Audenaert, P., and Michoel, T. (2019). High-dimensional Bayesian network inference from systems genetics data using genetic node ordering. *bioRxiv.* doi: 10.1101/501460

Werhli, A. V., and Husmeier, D. (2007). Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge. *Stat. Appl. Genet. Mol. Biol.* 6 (1), 15. doi: 10.2202/1544-6115.1282

Zhang, B., Gaiteri, C., Bodea, L. G., Wang, Z., McElwee, J., Podtelezhnikov, A. A., et al. (2013). Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell* Apr153 (3), 707–720. doi: 10.1016/j.cell.2013.03.030

Zhu, J., Lum, P. Y., Lamb, J., GuhaThakurta, D., Edwards, S. W., Thieringer, R., et al. (2004). An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenet. Genome Res.* 105, 363–374. doi: 10.1159/000078209

Zhu, J., Wiener, M. C., Zhang, C., Fridman, A., Minch, E., Lum, P. Y., et al. (2007). Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. *PloS Comput. Biol.* 3 (4), e69. doi: 10.1371/journal.pcbi.0030069

Zhu, J., Zhang, B., Smith, E. N., Drees, B., Brem, R. B., Kruglyak, L., et al. (2008). Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat. Genet.* 40, 854–861. doi: 10.1038/ng.167

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc.: Ser. B. (Stat. Methodol.)* 67 (2), 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.

frontiers in Genetics

METHODS
published: 08 January 2020
doi: 10.3389/fgene.2019.01303

A Novel Hybrid CNN-SVR for CRISPR/Cas9 Guide RNA Activity Prediction

Guishan Zhang[1], *Zhiming Dai*[2,3*] *and Xianhua Dai*[1,4*]

[1] School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, China, [2] School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China, [3] Guangdong Province Key Laboratory of Big Data Analysis and Processing, Sun Yat-sen University, Guangzhou, China, [4] Southern Marine Science and Engineering Guangdong Laboratory, Zhuhai, China

Accurate prediction of guide RNA (gRNA) on-target efficacy is critical for effective application of CRISPR/Cas9 system. Although some machine learning-based and convolutional neural network (CNN)-based methods have been proposed, prediction accuracy remains to be improved. Here, firstly we improved architectures of current CNNs for predicting gRNA on-target efficacy. Secondly, we proposed a novel hybrid system which combines our improved CNN with support vector regression (SVR). This CNN-SVR system is composed of two major components: a merged CNN as the front-end for extracting gRNA feature and an SVR as the back-end for regression and predicting gRNA cleavage efficiency. We demonstrate that CNN-SVR can effectively exploit features interactions from feed-forward directions to learn deeper features of gRNAs and their corresponding epigenetic features. Experiments on commonly used datasets show that our CNN-SVR system outperforms available state-of-the-art methods in terms of prediction accuracy, generalization, and robustness. Source codes are available at https://github.com/Peppags/CNN-SVR.

Keywords: CRISPR/Cas9, guide RNA, convolutional neural network, on-target, support vector regression

OPEN ACCESS

Edited by:
Shizhong Han,
Johns Hopkins Medicine,
United States

Reviewed by:
Xiang Li,
Harvard Medical School,
United States
Qi Liu,
Tongji University, China

***Correspondence:**
Zhiming Dai
daizhim@mail.sysu.edu.cn
Xianhua Dai
issdxh@mail.sysu.edu.cn

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

Received: *15 September 2019*
Accepted: *26 November 2019*
Published: *08 January 2020*

Citation:
Zhang G, Dai Z and Dai X (2020) A Novel Hybrid CNN-SVR for CRISPR/Cas9 Guide RNA Activity Prediction. Front. Genet. 10:1303. doi: 10.3389/fgene.2019.01303

INTRODUCTION

The CRISPR/Cas9 system, adapted from a bacterial defense mechanism, is a promising genomic editing tool that has recently revolutionized the field of biology, biotechnology, and medicine (Barrangou et al., 2007). This system consists of a nuclease activity-carrying Cas9 protein and the specificity-programming single guide RNA (gRNA), and the latter of which targets the complex to a genomic region flanked by a protospacer adjacent motif (PAM) (Jinek et al., 2012). Though the CRISPR/Cas9 system is considered to be very specific to perform the preconcerted cleavage on genome, numerous studies have indicated that Cas9 complex also binds to other unintended genomic sites, termed as off-target (Pattanayak et al., 2013; Doench et al., 2016). Thus, design of a gRNA with high on-target efficacy and low off-target effects is an important issue in CRISPR/Cas9 system. It has been shown that on-target activity is partly determined by gRNA intrinsic sequence and chromatin structure of target genomic region, but the underlying molecular mechanism is still not fully understood. Accurate prediction of gRNA on-target activity facilitates maximization of on-

target efficacy and minimization of off-target effects, further contributing to the success application of CRISPR/Cas9 system (Hsu et al., 2013; Doench et al., 2014; Xu et al., 2015; Chuai et al., 2016; Doench et al., 2016).

Previous efforts have been made to assist gRNA on-target identification and efficacy prediction based on different design rules. The alignment-based methods align the gRNAs from the given genome purely by locating the PAM [e.g. CCTop (Stemmer et al., 2015)]. Hypothesis driven-based tools empirically score the gRNA efficacy by incorporating the effect of genomic context factors [i.e. CFD (Doench et al., 2016)]. Machine learning-based methods predict the cleavage propensity of a genomic site for a given gRNA by considering different nucleotide features, such as position specific nucleotides and dinucleotides (Doench et al., 2014), GC content (Chari et al., 2015) as well as non-sequence features including thermodynamic stability of gRNA (Doench et al., 2014), amino acid cut position (Chen et al., 2017), and chromatin accessibility (Hinz et al., 2015; Horlbeck et al., 2016; Listgarten et al., 2018). For example, support vector machine (SVM)-based sgRNA Designer found that the position of the target site relative to the transcription start site and position within the protein are the most important factors for gRNA activity prediction (Doench et al., 2016). L1-regularized linear regression-based SSC reported that DNA sequence composition incorporating the preference for cytosine at the cleavage site improved the performance of gRNA on-target prediction (Xu et al., 2015). WU-CRISPR combined sequence and structural features of the gRNA to identify highly active gRNA (Wong et al., 2015). In general, no single feature but rather a combination of feature interactions governs gRNA cleavage efficacy (Wilson et al., 2018). Sophisticated models considering the interactions between the individual features achieved better performance (Aach et al., 2014; Erard et al., 2017). Nevertheless, some correlated features may result in the redundancy (Abadi et al., 2017), further rendering poor prediction outcome. Moreover, the outcomes of machine learning-based tools mainly depend on laborious manual feature engineering. They require considerable domain expertise to design the feature extractor (LeCun et al., 2015).

Deep learning allows computational models that consist of multiple processing layers to learn representations of features with multiple levels of abstraction (LeCun et al., 2015). The layers of features are learned from data by a general-purpose learning procedure instead of human engineers. Recently, several successful deep learning-based models have been provided for predicting CRISPR gRNA on-target activity. For example, Kim et al. proposed Seq-deepCpf1, which used convolutional neural networks (CNNs) to learn the nucleotide features of CRISPR gRNA, and it outperformed previous machine learning algorithms (Kim et al., 2018). Chuai et al. proposed DeepCRISPR that used deep convolutionary denosing neural network-based autoencoder to extract the CRISPR/Cas9 gRNA sequence representation and utilized the fully CNN model to predict the gRNA efficacy (Chuai et al., 2018). Extensive numerical experiments demonstrated DeepCRISPR surpassed the state-of-the-art tools across a variety of human datasets.

The above two CNN-based models showed good performance in CRISPR gRNA efficacy prediction compared with machine learning-based methods. CNNs are multi-layer architectures where the successive layers are designed to learn progressively higher-level features, until the last layer which produces the classifiers (Huang and LeCun, 2006). The last layer of CNN can be considered as a linear classifier operator on feature representation extracted by previous layers. CNN performs well in automatically learning nonlinearity features. However, CNN is not always an optimal choice for classification because the MLP layer following the feature extraction layer contains many trainable parameters. On the contrary, SVM with fixed kernel function has good utility on minimizing generalization error bound when applied to well-behaved feature vectors. Inspired by this, it is interesting to explore the hybrid CNN-SVM system in which CNN is trained to extract features and SVM computes a classifier function in the learned high dimensional feature spaces. To date, CNN-SVM models have shown impressive performance in a wide range of applications, such as object categorization (Huang and LeCun, 2006) and image recognition (Mori et al., 2005; Niu and Suen, 2012). For example, Niu et al. put forward a CNN-SVM model for handwritten digitals recognition with recognition rate of 99.81%. In their work, the proposed CNN-SVM replaced the back propagation neural network classifier with SVM in the last layer of the CNN model (Niu and Suen, 2012). Mori et al. trained a convolutional spiking neural network using different fragment images. The outputs of each layer in the model were input to the SVM model. A 100% face recognition rate was obtained for 600 images of 20 people (Mori et al., 2005). In terms of regression problem, Li et al. proposed CNN combined with support vector regression (CNN-SVR) for no-reference image quality assessment. This method achieved advanced outstanding performance compared with traditional CNN model (Li et al., 2016).

The prior success of CNN-SVM in computer vision inspired us to extend CNN-SVM application to CRISPR/Cas9 gRNA efficacy prediction. Until now, to the best of our knowledge, there is no such application. Previous studies have suggested that CRISPR gRNA efficacy prediction using linear regression achieved better performance than classification (Moreno-Mateos et al., 2015; Kim et al., 2018). Therefore, SVR, which is a common application form of SVM for regression, may be more appropriate for gRNA efficacy prediction when applied to well-behaved feature vectors. In this work, we developed a hybrid architecture incorporating CNN and SVR for CRISPR/Cas9 gRNA on-target activity prediction. The key idea of our system is to train a specialized CNN to extract robust gRNA sequence and epigenetic features, and to provide them to the SVR classifier for predicting gRNA cleavage efficacy. First, we trained the CNN model with back-propagation on the benchmark dataset, aiming at model selection and parameters tuning. Second, the initial CNN features were input into the SVR for training and evaluating. A two-step strategy was performed to select the important features from well-trained CNN intrinsic gradients features. Third, the well-trained CNN-SVR was used to test the

independent cell-line dataset. Specifically, the test data was input to the well-trained CNN model to obtain the test features. Using the test feature vector, the well-trained SVR classifier was performed to predict the gRNA cleavage efficacy. Experiments showed improved performance of the proposed CNN-SVR model for CRISPR/Cas9 gRNA on-target activity prediction compared with state-of-the-art algorithms.

MATERIALS AND METHODS

Data Resources

Benchmark Dataset

Previous studies have shown that PAM-distal region has a high tolerance for sequence mismatches (Kim et al., 2016; Kleinstiver et al., 2016). To be specific, gRNAs with two mismatches in the first two positions from the 5' end has little influence on cleavage efficiency (Doench et al., 2014; Doench et al., 2016). Inspired by these studies, Chuai et al. applied a data augmentation procedure by changing each gRNA into a new one with two mismatches in the PAM distal region (Chuai et al., 2018). Consequently, a 23-nt gRNA sequence can be expanded into 16 gRNAs with identical cleavage efficacy. The augmented dataset was generated from ~15,000 gRNAs with known on-target cleavage efficacy. By adopting this data augmentation strategy, they obtained 180512 non-redundant gRNAs. Each observation in the data contains a 23-nt gRNA sequence and its corresponding cleavage efficiency. In this work, we used this augmented dataset as the benchmark data for model selection and pre-training.

Four Cell Line Independent Test Datasets

In order to evaluate the performance of our method, we used four public experimental validated gRNA on-target cleavage efficacy independent human datasets, which were integrated and processed by Chuai et al (Chuai et al., 2018). These experimented-based datasets were originally collected from public datasets (Wang et al., 2014; Hart et al., 2015; Doench et al., 2016). They covered gRNAs targeting 1071 genes from four different cell lines, including HCT116 (4239 samples) (Hart et al., 2015), HEK293T (2333 samples) (Doench et al., 2016), HELA (8101 samples) (Hart et al., 2015), and HL60 (2076 samples) (Wang et al., 2014) with redundancy removed. The gRNA on-target activity was strictly restricted to experimental assay, where the cleavage efficiency was defined as the log-fold change in the measured knockout efficacy. Readouts of cleavage efficacies without in vivo (in vitro) experimental validation were excluded.

Each entry in the datasets contained the 23-nt gRNA sequence, four kinds of corresponding symbolic epigenetic features, as well as numerical and binary cleavage efficacy. The epigenetic features information was obtained from ENCODE (Consortium, 2004), including CTCF binding information obtained from ChIP-Seq assay, H3K4me3 information from ChIP-Seq assay, chromatin accessibility information from DNase-Seq assay, and DNA methylation information from RRBS assay. Each epigenetic feature was represented by an "A-N" symbolic sequence with length of 23. Here, the presence of the epigenetic feature at a particular base position of DNA regions was denoted by "A," and its absence was represented by "N."

Numerical cleavage efficiency of candidate gRNA was calculated using a collaborative filtering-based data normalization technique (Badaro et al., 2013). In particular, a matrix Y was formulated where each row denoted the experiments and each column represented one gRNA. y_{mn} represented the n-th gRNA on-target cleavage efficacy in the m-th experiment. Normalized numerical gRNA on-target efficiency value was defined as

$$y_{nor} = y_{mn} - (m_{row} + m_{col} + m_{all})/3 \quad (1)$$

where m_{row} denoted the mean value for each row, m_{col} represented the mean value for each column, and m_{all} denoted the mean value of Y. Next, a rank-based normalization method (Doench et al., 2016) was applied for gRNAs within each gene, and these normalized ranks were averaged across cell types, then were rescaled in [0, 1], where 1 indicated the successful on-target cleavage efficacy. The binary cleavage efficiency of gRNA was determined by using a log-fold change of 1 as the cut off, where 1 and 0 represented the high-efficiency and low-efficiency gRNAs, respectively. The processed datasets can be downloaded at https://github.com/bm2-lab/DeepCRISPR.

Sequence Encoding

We formulated one-hot encoding to encode gRNA sequence with 23 nucleotides in length. Each base in the sequence can be encoded as one of the four one-hot vectors [1,0,0,0], [0,1,0,0], [0,0,1,0] and [0,0,0,1]. Therefore, the 1-by-23 nucleotide sequence was represented by four binary channels: A-channel, C-channel, G-channel, and T-channel. Taking A-channel as an example, the presence of the nucleotide A at a particular base pair position was denoted by 1 and the absence of the nucleotide A was represented by 0. Consequently, each gRNA was expressed by a 4×23 matrix, where 23 was the length of the gRNA sequence.

Analogously, epigenetic feature information including CTCF binding, H3K4me3, chromatin accessibility, and DNA methylation were represented by a 4×23 binary matrix. Each type of epigenetic information was denoted by a 1×23 matrix using "A" and "N," with these notations meaning presence and absence of that epigenetic feature at specific position of DNA regions, respectively. To encode the epigenetic feature information, we derived a 23-length vector to encode each epigenetic feature. Thus, four epigenetic features were donated by a 4×23 binary matrix (see **Figure S1** for an example). The encoded sequence and epigenetic matrix of gRNA were then fed into CNN-based gRNA stream and epigenetic stream sub-networks for model training and testing.

CNN Model Structure

We developed a CNN model to learn deep features of gRNA sequence and its corresponding epigenetic information (**Figure S2**). The proposed CNN is composed of two branches, namely gRNA stream and epigenetic stream. These two sub-networks are structurally identical, including two one-dimensional (1D)

convolution layers, two average-pooling layers, and four fully connected layers.

Taking gRNA stream as an example, the input is a 4 (size of nucleotides vocabulary) × 23 (sequence length) binary matrix. The first layer of the sub-network is a 1D convolution layer (conv_1), which is designed for extracting the important local features between neighboring element values of gRNA sequence information using 256 convolution kernels of size 5. Rectified linear unit (ReLU) (Krizhevsky et al., 2012) is used as the activation function to the convolution outputs.

The second layer is a local average pooling layer (pool_1) with window size of 2 connected with the outputs of previous layer for down-sampling. Each of the average-pooling windows only outputs the average value of its respective convolution layer outputs.

The structures of the following convolution layer (conv_2) and average pooling layer (pool_2) are identical with the first (conv_1) and second (pool_1) layers mentioned above. After being flatten, the features are followed by four fully connected layers (fc_1, fc_2, fc_3 and fc_4) with the sizes of 256, 128, 64, and 40, respectively. We used dropout for model regularization to avoid overfitting.

The feature maps of the fourth fully connected layer from both gRNA and epigenetic branches are concatenated by the "concatenate" operator. Subsequently, the outputs of the concatenation layer are input to the last fully connected layer of the merged CNN network. The final output layer consists of one neuron corresponding to a regression score that highly correlates with gRNA activity. The loss function for our CNN is mean squared error (MSE) which was adapted in a previous study (Kim et al., 2018). We chose MSE because it is a good measure to prevent undesired outliers in the dataset.

Hybrid CNN-SVR Model

We next proposed a network combining CNN and SVR called CNN-SVR to provide a data-driven and deep learning method for CRISPR/Cas9 gRNA activity prediction. For cell line-specific prediction, CNN-SVR receives a 23-nt gRNA sequence and four "A-N" symbolic epigenetic sequences with length of 23 as inputs, and it produces a regression score of gRNA on-target cleavage efficacy. Compared with machine learning-based methods that rely heavily on hand-crafted features, CNN-SVR can get rid of the dependence on manual feature engineering. The basic flowchart of CNN-SVR consists of two major stages, namely model selection and pre-training stage as well as fine-tuning and testing stage (**Figure S3**). The dataset was randomly divided into two separate sets of training and testing, respectively. One-hot encoding converts the input sequences into binary representations for downstream processing.

In the model selection and pre-training stage, there are mainly three steps: first, the encoded benchmark dataset is fed into the proposed CNN model for pre-training by the back-propagation algorithm. Randomized five-fold cross-validation tests are conducted to determine hyperparameters of the merged CNN model. Model with the minimum average validation loss is regarded as the base model. Second, the initial CNN extracted features are input to SVR classifier for training and evaluating. SVR (i.e., cost C, gamma, and epsilon) is optimized using a grid search approach to achieve the optimal performance. Third, a two-step strategy is employed to remove the redundancy of CNN features (see details in the section *Feature Representation Optimization*). The extracted low-dimensional representative feature data and their corresponding gRNA cleavage efficacy values are fed into SVR classifier for model training.

In the fine-tuning and testing stage, there are mainly two steps: First, the well-trained CNN model is applied to extract features from new cell line data. Only the fourth fully connected layer of gRNA stream and epigenetic stream, and the top fully connected layer of the merged CNN are fine tuned. MSE loss function is minimized by back-propagation approach. Second, the extracted low-dimensional representative features are fed into the well-trained SVR classifier to complete the final gRNA activity prediction. **Figure 1** displayed the overall framework of our CNN-SVR; the procedures were described as follows:

- The gRNA sequence and epigenetic feature sequences are converted into two 4×23 binary matrices by one-hot encoding.
- The encoded gRNA and epigenetic sequences are fed into the well-trained CNN-based gRNA stream and epigenetic feature stream to fine-tune and extract features, respectively.

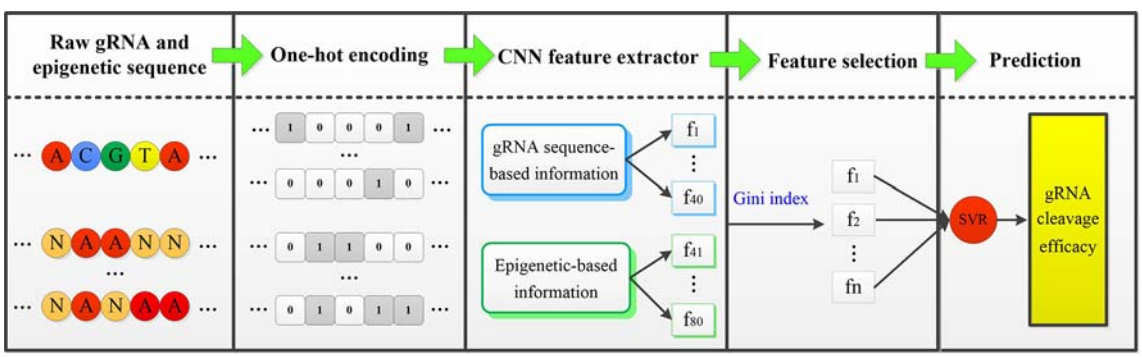

FIGURE 1 | An illustration of procedures for cell line-specific gRNA on-target activity prediction based on CNN-SVR. Here, [f_1,f_2,⋯,f_n] is the subset of [f_1,f_2, ⋯,f_80].

- SVR classifier is trained based on the optimal feature set. Ultimately, the well-trained SVR model assigns a prediction cleavage efficacy score for the candidate gRNA.

Experimental Setup

To evaluate feasibility of CNN-SVR for gRNA activity prediction, we conducted numerical experiments on public datasets. We implemented our algorithms using Keras (2.1.0) with Tensorflow (1.4.0) as the backend, running on Intel Core i7 CPU at 3.6 GHz with 16 GB RAM and NVIDIA 8 GB GTX 1080 GPU. The optimized parameters were tuned automatically under the Adam optimizer (Kingma and Ba, 2014).

Implementation of the Hybrid CNN-SVR Model

CNN Model Selection and Training

In the proposed architecture, the distribution of each network parameter was determined empirically. The main purpose of hyperparameter optimization was to choose a set of hyperparameters for a deep architecture, usually with the goal of optimizing performance of the architecture on an independent dataset. Grid search from the Scikit-learn Python library was adopted to tune the hyperparameters of the proposed architectures. Hyperparameter optimization experiments were performed sequentially as follows: the network weight initialization over the choice ("zero," "he_uniform," "uniform," "glorot_uniform," "lecun_uniform," "normal," "he_normal"), dropout regularization over the choice (0.2, 0.3, 0.4, 0.5, 0.6), batch size over the choice (64, 128, 256, 512), and number of epochs over the choice (50, 100, 200, 300).

All the constructed neural network models were trained and validated on the benchmark dataset (180512 samples). We randomly assigned the samples of the no-redundant dataset with 80% of samples for training and 20% of samples for testing with five-fold cross-validation in the training phase. Cross-validation contributed to avoiding overfitting and guaranteeing the accuracy of our model in which the datasets were divided into five equal parts randomly. In each training, one part was regarded as the testing dataset, while the remaining four parts were taken as the training dataset. Thus, we obtained 115528 training samples, 28881 validation samples, and 36103 testing samples, respectively. Separate training and validation data were applied to train the model, while the test data was used to evaluate the performance of the trained model. We chose the model that showed the minimum average validation loss as the final CNN model. After optimization, the hyperparameters were as follows: kernel_initializer: glorot_uniform; batch size: 256; epoch: 200; dropout: 0.3 (keeping 70% of the connections).

SVR Training and Testing

Next, CNN extracted features were fed into the SVR classifier. We implemented the SVR algorithm in Scikit-learn library. Grid search procedure was performed to find the optimal penalty parameter C, kernel parameter gamma, and epsilon. For training SVR with Gaussian radial basis kernel (RBF) classifier, grid search range of each parameter was as follows: cost C from the choice (1.0,1.1,⋯,1.9), kernel coefficient gamma over the choice (0.11,0.12,⋯,0.15), epsilon from the choice (0.08,0.09,⋯,0.12). We selected the parameters that maximized the maximum average area under ROC curve (AUROC) value as the final parameters of SVR classifier. The optimized parameters of the SVR were as follows: C was 1.7, gamma was 0.12, epsilon was 0.11. These parameters were then used to train the CNN-SVR model.

Feature Representation Optimization

Considering that CNN extracted features might introduce redundancy which can undermine model performance, we employed a two-step feature optimization strategy to identify important feature subsets from the initial CNN features. To be specific, we first applied random forest to the learnt feature representation from well-trained CNN model and obtained the ranked feature list based on information gain (Liaw and Wiener, 2002). We trained the random forest model with 250 decision trees using Scikit-learn. The feature importance distribution of the top 20 features based on their importance scores was illustrated in **Figure S4**. As can be seen, the seventeenth feature of CNN extracted initial features was the most predictive feature. Second, the sequential forward search (SFS) (Whitney, 2006) was performed to determine the optimal feature set. We gradually added features from random forest feature rank from higher score (lower rank) to lower score (higher rank) to reconstruct the SVR models. The feature subset with the relatively higher value of AUROC was regarded as the optimal feature set. We used the AUROC since it is a good indicator to evaluate the real performance of models. We noted that, when the feature number reached at 13, the model achieved the maximum AUROC of 0.9769. Hence, the top 13 features (i.e., "feat_17," "feat_26," "feat_9," "feat_19," "feat_30," "feat_6," "feat_12," "feat_39," "feat_36," "feat_21," "feat_22," "feat_3," "feat_25") in the random forest rank list were integrated into SVR classifier to train the prediction scheme. Here, "feat_17" means the 17th feature of CNN extracted initial features (total 80 features). Thereby, we carried out the determined hyperparameters by integrating the optimal features on the benchmark dataset under five-fold cross-validation to obtain the well-trained CNN-SVR model. The training data, validation data, and testing data were built consistent with the above mentioned data partitioning way in the *CNN Model Selection and Training* section. The well-trained CNN-SVR reached an overall Spearman correlation of 0.952, AUROC value of 0.977.

Transfer Learning for New Cell Line Specific Prediction

In this section, we proposed a fine-tune strategy by borrowing information from the benchmark data, aiming at boosting the prediction performance on small sample size cell line-specific data. To this end, four above cell-line datasets were combined together for model training and testing. We constructed the training, validation, and test data from total four datasets based on gRNA sequence composition and epigenetic feature information. The training data (13401 samples) and test data

(3748 samples) were also generated in the same way in the *CNN Model Selection and Training* section. Randomized five-fold cross-validation was implemented in the training phase.

Considering training a full CNN network with small number of cell line data may result in overfitting, which may lead to poor performance. Transfer learning (Bengio, 2012) is effective to address the challenge where the learned parameters of well-trained networks on a large dataset are shared. The main idea of transfer learning is to use a pre-trained model which is trained on large dataset and to transfer its well-trained parameters (e.g. weights) to the targeted network model. Though the dataset is different from the one that the network was trained on, the lower-level features are similar. Thus, the last fully connected layers are usually trained on the new dataset. Transfer learning has been widely applied to computer vision (Shin et al., 2016; Cheng and Malhi, 2017) and achieved a valuable efficacy in terms of accuracy. We applied transfer learning from the benchmark dataset pre-trained CNN model, and fine-tuned for small sample cell line data. Note that, the low-level features between the benchmark data and cell line-specific data are similar. Therefore, we froze the convolution layers, average pooling layers and the first three fully connected layers of both gRNA stream and epigenetic stream. After borrowing weights of the well pre-trained CNN base network, we fine-tuned the weights of the last fully connected layers of both gRNA and epigenetic sub-networks and those of the merged fully connected layer to optimize the mean validation squared error loss function. During fine tuning, we only updated 5281 free parameters. By fixing the weights parameters in the other layers, CNN-SVR could prevent overfitting and effectively learn to integrate the sequence representative and epigenetic information. For any given cell line of interest, the training process was described as follows:

- Pre-train a CNN model with the benchmark data for 200 epochs.
- Freeze the convolution layers, average pooling layers, the first three fully connected layers (for both the gRNA stream and epigenetic stream).
- Train the fourth fully connected layer of the above two streams and the top fully connected layer of the merged CNN model with training data from the cell line of interest for another 200 epochs.
- Evaluate the model on the test data.

Settings of Other Methods

For the L1-regularized linear regression (L1), we applied LassoCV from Scikit-learn Python library to find out the optimal parameters of alpha by cross-validation. Grid searching range of regularization parameter alpha was (0.01, 0.02,⋯,0.1). Other parameters were set with default values. We achieved an optimal value of 0.01. Similarly, we applied RidgeCV and ElasticNetCV with the same grid searching range of L1 to set parameter alpha for L2-regularized linear regression (L2) and L1L2-regularized linear regression (L1L2), respectively. After optimization, the best alpha values of L2 and L1L2 were 0.04 and 0.01, respectively. These parameters were then used to train the CNN-L1, CNN-L2, and CNN-L1L2 models. Other parameters of L2 and L1L2 were set with default values.

We ran the code of Seq_deepCpf1 using the same data and basic training process (downloaded from GitHub at https://github.com/MyungjaeSong/Paired-Library). Note that, the input of Seq_deepCpf1 was a 4-by-34 dimensional binary matrix. Here, we changed the input shape of Seq_deepCpf1 model into 4-by-23 to match the size of the data in this study. Besides, we used the benchmark dataset to pre-train the Seq_deepCpf1 model. To make a fair comparison, we only fine-tuned the weights parameters in the last two layers (1681 free parameters) for cell line-specific prediction. The numerical experimental condition was set consistent with DeepCRISPR. The source codes of DeepCRISPR were downloaded from https://github.com/bm2-lab/DeepCRISPR. SSC, sgRNA Designer and WU-CRISPR provided available web based applications. More details can be found in **Table S1**.

Performance Measurements

To quantitatively evaluate the performance of our CNN-SVR, Spearman correlation coefficient between predicted and measured on-target activity was calculated. We chose Spearman correlation is due to it is more robust to outliers than Pearson's correlation coefficient (Mukaka, 2012). Besides, it was adapted in previous studies (Doench et al., 2016; Chuai et al., 2018; Kim et al., 2018). Spearman correlation was calculated using SciPy library (http://scipy.org). In addition, AUROC was employed to comprehensively quantify the overall predictive model performance. The value of AUROC ranges from 0.5 to 1. A larger AUROC value represents that model achieves better and more robust performance. Note that, we used 0.5 AUROC as the baseline. Statistical test was performed using SciPy library for comparing the differences between GC content distributions of different datasets. Two-sample Kolmogorov–Smirnov test was used for testing the distance between two distributions under the null hypothesis that samples from the same continuous distribution. $P < 0.05$ was considered to indicate statistically significant difference.

RESULTS

Comparison CNN-SVR With CNN Model

To verify the feasibility of our approach, we compared our CNN-SVR with CNN model on the above four cell-line datasets. The current practice of training a model was to use cell-line specific data for prediction. Each data set was randomly split into a training set and an independent testing set with 80% and 20% classes. **Table 1** summarized the results regarding evaluation criteria including Spearman correlation and AUROC under 10-round 10-fold cross-validation tests. CNN-SVR showed substantially better performance in terms of Spearman correlation. As for AUROC, CNN-SVR was superior to CNN on datasets HEK293T, HELA, and HL60. These results showed that CNN-SVR is more predictive than CNN for gRNA on-target

TABLE 1 | Performance comparison between CNN-SVR and CNN models for gRNA activity prediction on four cell-line datasets under 10-time 10-fold cross-validation.

Model	CNN-SVR	CNN	CNN-SVR	CNN
	Spearman		AUROC	
HCT116	**0.719 ± 0.008**	0.661 ± 0.030	**0.933 ± 0.001**	0.932 ± 0.001
HEK293T	**0.807 ± 0.016**	0.725 ± 0.029	**0.983 ± 0.002**	0.972 ± 0.001
HELA	0.699 ± 0.006	**0.702 ± 0.007**	**0.933 ± 0.001**	0.916 ± 0.001
HL60	**0.589 ± 0.006**	0.576 ± 0.040	**0.934 ± 0.003**	0.914 ± 0.003

Performance is shown as mean ± standard deviation. This representation also applies to ***Table 2****. The best performance across different folds cross-validation method is highlighted in bold for clarification. These highlights also apply to* ***Tables 2*** *to* ***4*** *and* ***Tables S3*** *to* ***S5****.*

activity, further conforming the feasibility and effectiveness of the combination of CNN and SVR classifier.

Comparison of Various CNN Combined Regression Models

We then attempted to access the regression performance of CNN-SVR. To this end, we compared CNN-SVR with three CNNs plus regression approaches, including CNN plus L1 (CNN-L1), CNN plus L2 (CNN-L2), and CNN plus L1L2 (CNN-L1L2) on the above four cell lines datasets. Note that for each cell line, the training data and test data were generated in the same way as described in the section *Comparison CNN-SVR With CNN Model*. Ten-time 10-fold cross-validation tests were randomly performed and the average of the individual performance were summarized in **Table 2**. Overall, CNN-SVR performed better than CNNs with different regression methods on all datasets. These observations revealed that the regression learning performance of our SVR surpasses other regression methods on gRNA activity prediction.

Comparison With State-Of-the-Art Methods

To validate the performance of proposed CNN-SVR, we compared it with one deep learning-based method (DeepCRISPR) and three machine learning methods including sgRNA Designer, SSC, and WU-CRISPR (**Table S2**). Note that Seq-deepCpf1 only allows for receiving gRNA sequence as input. So, this approach was not compatible with other methods when considering both gRNA sequence and epigenetic information. To make a fair comparison, we trained CNN-SVR model based on the training data strictly consistent with other methods. The above four datasets were used for performance evaluation. For each cell line, the training and test data were constructed in the same way as described in the section *Comparison CNN-SVR With CNN Model*. For any given cell line of interest, the training data was built by integrating all the training data from four cell lines. The performance was evaluated on each cell line-specific testing set, respectively.

On the whole, CNN-SVR achieved the highest average Spearman correlation (**Figure 2A**). Specifically, CNN-SVR exhibited better Spearman correlation on three datasets (i.e., Total, HCT116, HELA and HL60), whereas for dataset HEK293T, it performed slightly worse than DeepCRISPR. **Figure 2B** illustrated the performance in terms of AUROC. Some interesting conclusions can be extracted from this figure. First, deep learning models were superior to machine learning methods. Second, CNN-SVR exhibited better predictive power than another deep learning model DeepCRISPR. The details of their performance can be found in **Table S3**. To sum up, these observations indicated that CNN-SVR outperforms the compared state-of-the-art methods for predicting gRNA on-target activity.

Assessment of Generalization Performance With a Leave-One-Cell-Out Procedure

Next, we investigated the generalizability ability of CNN-SVR in new cell types. For this purpose, we took turns to test the model on the above four cell-line datasets using a leave-one-cell-out approach. The training data and test data for each cell

TABLE 2 | Performance comparison of CNN-SVR and different CNNs combined regression models for gRNA activity prediction on four cell-line datasets under 10-time 10-fold cross-validation.

Model	HCT116	HEK293T	HELA	HL60
(A) Spearman correlation				
CNN-SVR	**0.719 ± 0.008**	**0.807 ± 0.016**	**0.699± 0.006**	**0.589± 0.006**
CNN-L1	0.712± 0.010	0.793 ± 0.004	0.633± 0.020	0.542± 0.033
CNN-L2	0.670± 0.025	0.731 ± 0.032	0.683± 0.009	0.517± 0.034
CNN-L1L2	0.701± 0.008	0.803 ± 0.012	0.682± 0.005	0.589± 0.018
(B) AUROC				
CNN-SVR	**0.933 ± 0.001**	**0.983 ± 0.002**	**0.933 ± 0.001**	**0.934 ± 0.003**
CNN-L1	0.931 ± 0.001	0.982 ± 0.001	0.924 ± 0.002	0.930 ± 0.003
CNN-L2	0.919± 0.002	0.975 ± 0.002	0.923± 0.002	0.895± 0.008
CNN-L2	0.918± 0.003	0.977 ± 0.001	0.915± 0.002	0.912± 0.004

The tables from top to bottom respectively record the Spearman correlation and AUROC of CNN-SVR and three CNN combined regression methods.

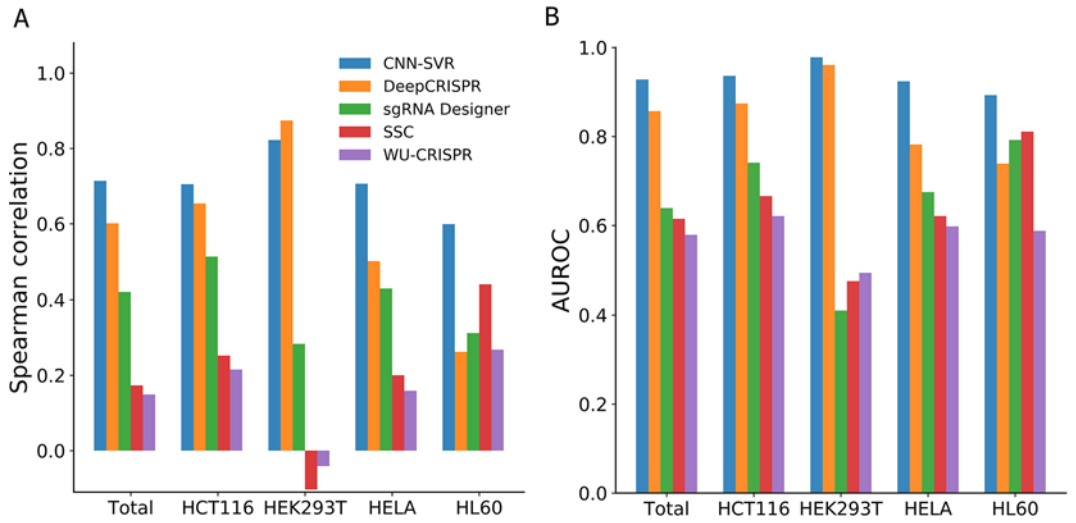

FIGURE 2 | Performance comparison of CNN-SVR and other prediction models on various testing cell line data.

line were built in advance. Note that, the partitioning method for each cell line data followed the way illustrated in the section *Comparison CNN-SVR With CNN Model* In the training phase, for a given cell line to be predicted, we just used the training data from all other three cell lines (lacking training data of given cell-line of interest). In the testing stage, we evaluated the performance on the test data of the given cell-line of interest. Taking leave-HCT116-out procedure as an example, we trained the model by combining training data of HEK293T, HELA and HL60 cell lines (without HCT116 cell line training data), and evaluated the model on HCT116 cell line testing set. For fair comparison, we tested the proposed CNN-SVR under the same condition with DeepCRISPR, sgRNA Designer, SSC, and WU-CRISPR on the four cell-line datasets.

As can be seen from **Figure 3A**, among the compared models, CNN-SVR exhibited the best predictive power, with average Spearman correlation of 0.714. Compared with DeepCRISPR, which was one of the best state-of-the-art approaches, CNN-SVR showed superior performance on all datasets except for dataset HCT116. DeepCRISPR got comparable performance with CNN-SVR on HCT116 dataset. Furthermore, CNN-SVR outperformed other methods on all datasets in terms of AUROC (**Figure 3B**). Together, these results demonstrated the excellent generalizability of CNN-SVR. More details of the performance can be found in **Table S4**.

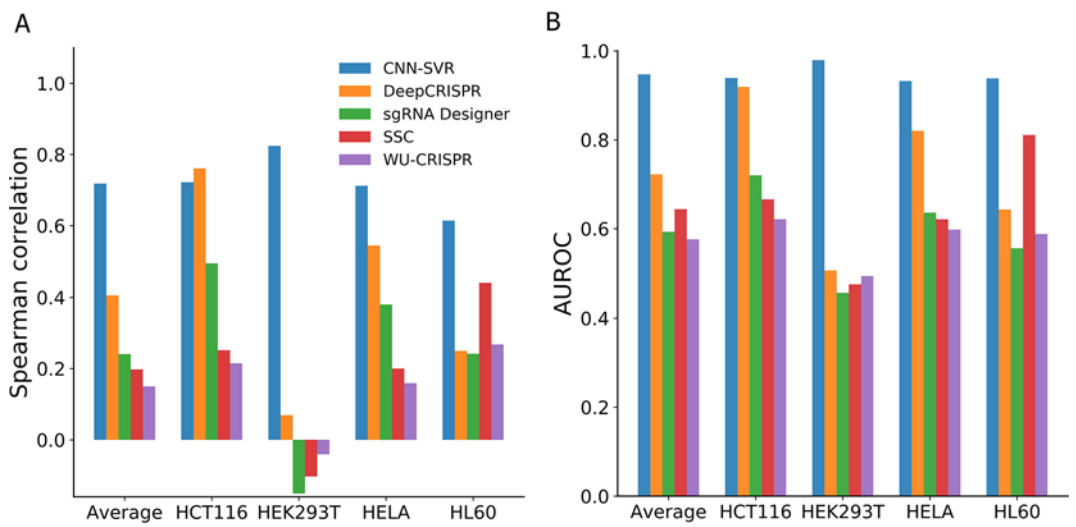

FIGURE 3 | Performance comparison of CNN-SVR and other prediction models on various testing cell line data with a leave-one-cell-out procedure.

Evaluation of Robustness of Prediction Models

In this section, we aimed to compare the robustness of the above methods. To this end, we examined the changes between Spearman correlation and AUROC values obtained by training with four cell datasets (**Figure 2**) and those produced by the leave-one-cell-out approach (**Figure 3**). For each evaluation criterion, we calculated the difference of each model by subtracting the results of training with leave-one-cell-out (**Table S4**) from the cell-line independent (**Table S3**). Taking CNN-SVR as an example, the AUROC difference of HCT116 dataset was calculated as follows:

$$\Delta\mathbf{AUROC}_{CNN-SVR} = \mathbf{0.936} - \mathbf{0.939} = -\mathbf{0.003} \quad (2)$$

where "ΔAUROC" means the difference value of AUROC. It can be seen that our CNN-SVR substantially showed smaller changes than DeepCRISPR in terms of the above mentioned two evaluation measures (**Table 3**). Interestingly, we observed that the performance of DeepCRISPR on dataset HEK293T using the whole training set was significantly better than that by leave-one-cell-out approach (with Spearman correlation difference value of 0.805, AUROC difference value of 0.455). Previous studies have shown that gRNAs with low or high GC content tended to be less active (Doench et al., 2014; Wang et al., 2014). We analyzed GC content of the four cell datasets. As expected, dataset HEK293T has the lowest GC content (vs. dataset HCT116, P=1.35E-52; vs. dataset HELA, P=1.14E-69, vs. dataset HL60, P=1.45E-07, two-sample Kolmogorov-Smirnov test, **Figure S5**).

Effect of Epigenetic Features on gRNA Cleavage Efficacy

In this section, we determined whether cell line-specific epigenetic features really boost the predictive performance. We examined the performance of deep learning-based methods on the four cell-line datasets only considering gRNA sequence composition and compared them with those considering both gRNA sequence and epigenetic information (see the section *Assessment of Generalization Performance With a Leave-One-Cell-Out Procedure*). We trained the prediction models without epigenetic information (sequence only) for each cell line with a leave-one-cell-out procedure. Note that, we trained the model just considering the gRNA stream. Other numerical experimental conditions were in accord with the section *Assessment of Generalization Performance With a Leave-One-Cell-Out Procedure*. For fair comparison, we compared our methods with two deep learning-based methods (i.e., DeepCRISPR and Seq-deepCpf1) only considering sequence composition.

Figure 4 and **Table 4** compared the prediction performance of various deep learning methods trained using different datasets. Two interesting conclusions can be drawn as below. First, CNN-SVR showed better performance compared with other models. Second, as expected, the prediction accuracies of models trained only considering sequence composition (**Figure 4A** and **Table 4A**) became lower compared with those trained with both sequence and epigenetic data (**Figure 4B** and **Table 4B**). To conclude, these observations confirm that cell line-specific epigenetic features contribute to gRNA activity and specificity. More details of their performance of Spearman correlation can be found in **Table S5**.

Visualizing Importance of Position-Specific Nucleotides

Finally, we aimed to investigate what sequence patterns of gRNA contribute to its on-target activity. Using the method in a previous study (Xie et al., 2013), we investigated the feature importance of all possible position-specific nucleotides. In brief, we constructed a specific sequence and its corresponding epigenetic features to feed the well-trained CNN model and took the outputs for visualization. More details can be found in **Supplementary Material**. **Figure 5A** depicts the importance of all four nucleotides and epigenetic features at different positions. Several interesting results can be observed: (i) Most of the top features were generated by convolving the middle region of input matrix. (ii) Thymines are found to be disfavored at the fourth position adjacent to the PAM. The same observation was obtained by Chuai et al., (2018), which is consistent with previous finding that multiple uracils in the spacer result in low gRNA expression (Doench et al., 2014). Another study also found that thymine in the seed sequence might destabilize interactions between the protein and crRNA (Kim et al., 2017). (iii) Cytosine is informative at 3-nt upstream of the PAM since the cleavage site usually resides 3 nt upstream the PAM. (iv) Our model suggests that cytosine is also preferred at position 17, which coincides with a previous finding that the cleavage is 3 nt, 4 nt or even further upstream of the PAM (Shou et al., 2018). (v) In general, the middle region contains more information of the epigenetic features. Notably, 3 nt upstream of the PAM has a consistent preference for opening-chromatin information of Dnase. This observation is in accordance with a previous study, which corroborates that consideration of target site accessibility can boost the accuracy of gRNA activity prediction (Kim et al., 2018). Besides, we presented the

TABLE 3 | The differences of Spearman correlation and AUROC between independent test and a leave-one-cell-out approach between CNN-SVR and DeepCRISPR.

Model	HCT116	HEK293T	HELA	HL60
(A) Spearman correlation				
CNN-SVR	**-0.017**	**-0.002**	**0.011**	-0.015
DeepCRISPR	-0.107	0.805	-0.043	**0.012**
(B) AUROC				
CNN-SVR	**-0.003**	**-0.001**	**-0.008**	**-0.045**
DeepCRISPR	-0.045	0.455	-0.038	0.096

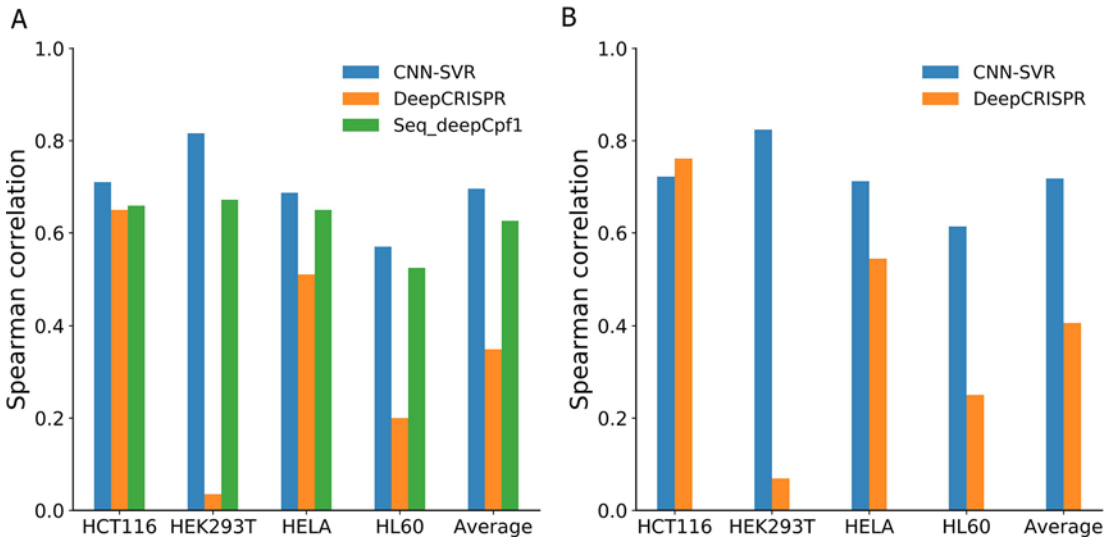

FIGURE 4 | Spearman correlation between different deep learning-based models and datasets. Models considering **(A)** gRNA sequence composition only and **(B)** both gRNA sequence and epigenetic information.

sequence logo to visualize the nucleotide differences on the benchmark dataset. Overall, the result is in line with our feature analysis (see **Figure 5B**).

We also explored the importance of dimers. Here, by adopting the method proposed above, we generated a sequence which only contains one dimer (out of 16 possible dimers) at every position k and repeated the aforementioned process for all subsequences. The scores of all the constructed subsequences for all the positions were plotted as a heatmap in **Figure S6**. We note that most of the top features were generated by convolving the region of the seed sequence of the gRNAs. This observation coincides with previous finding that a prototypical 10–12 nt PAM-proximal seed sequence largely determines target efficacy (Jinek et al., 2012; Cong et al., 2013).

DISCUSSION

Accurate prediction of gRNA cleavage efficacy is pivotal to understanding the mechanisms of CRISPR/Cas9 system. Although computational prediction of gRNA cleavage efficiency has made much progress recently, the accuracy remains to be improved. In this study, we introduced a novel and interpretable deep learning framework named CNN-SVR for CRISPR/Cas9 gRNA on-target activity prediction. Specifically, CNN works as a trainable feature extractor and SVR performs as a gRNA cleavage efficacy predictor. Compared with CNN and three CNNs combined regression-based algorithms, CNN-SVR achieved the best performance. CNN-SVR could not only automatically extract gRNA sequence and the corresponding epigenetic features using the CNN, but also improve the generalization ability of CNN and regression accuracy.

Previous studies suggested that ensemble learning (Woźniak et al., 2014) by incorporating multiple neural networks together can achieve higher accuracy than a single learner (Maqsood et al., 2004). Inspired by this, instead of using a single convolution network to train the feature vectors of gRNA like Seq-deepCpf1, we merged two sub-networks (i.e., gRNA stream and epigenetic stream) to train gRNA sequence and its corresponding epigenetic information. In addition, the architecture of the proposed sub-networks was considerably shallower than DeepCRISPR. Compared with several current state-of-the-art learning-based methods, CNN-SVR can effectively exploit deep features of gRNA sequences. Experimental results demonstrated the power of our CNN-SVR for CRISPR/Cas9 gRNA activity prediction.

TABLE 4 | AUROC of different deep learning-based methods by considering gRNA sequence only and incorporating both gRNA sequence and epigenetic features.

Model	HCT116	HEK293T	HELA	HL60	Average
(A) Sequence-only					
CNN-SVR	**0.938**	**0.976**	**0.930**	**0.928**	**0.943**
DeepCRISPR	0.887	0.474	0.788	0.584	0.683
Seq-deepCpf1	0.931	**0.976**	0.925	0.920	0.938
(B) Sequence composition and epigenetic features					
CNN-SVR	**0.939**	**0.979**	**0.932**	**0.938**	**0.947**
DeepCRISPR	0.919	0.506	0.820	0.643	0.722

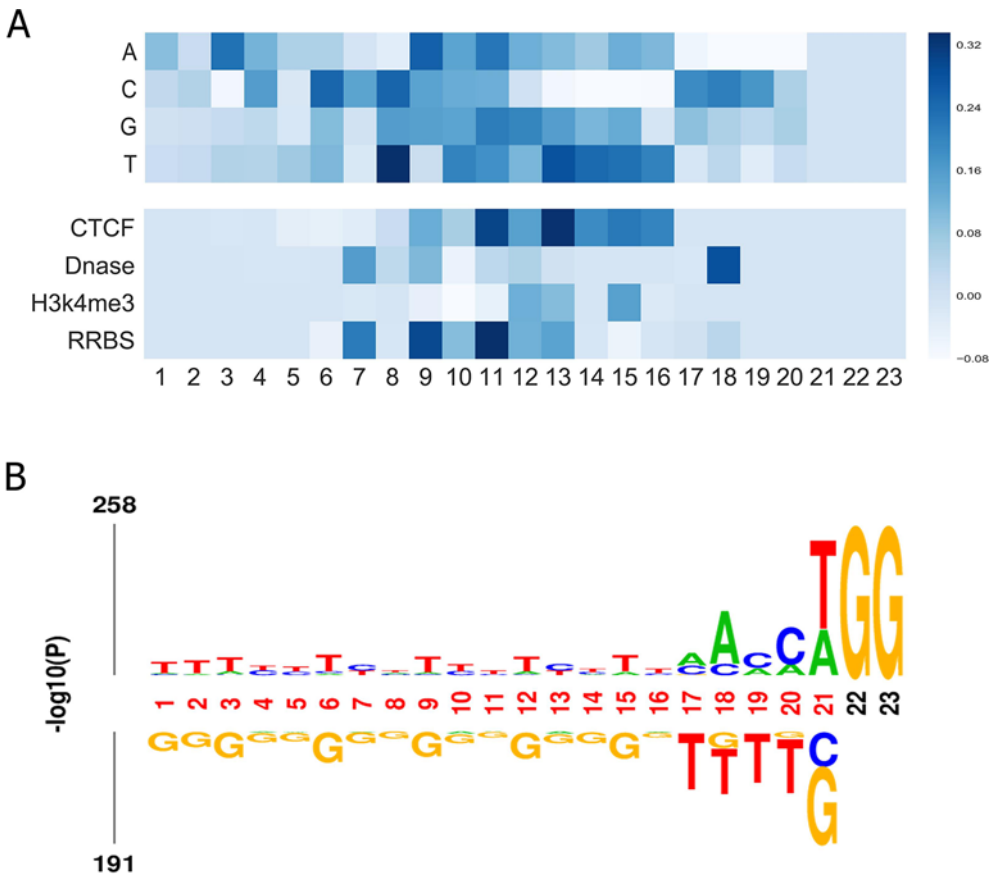

FIGURE 5 | (A) Visualization of the importance of different nucleotides and epigenetic features at different positions for our model trained on the benchmark dataset. The colors represent the contribution of the position-specific nucleotides to determining an efficient gRNA. The x-axis shows the positions of the nucleotide in the sequence. The y-axis lists all possible nucleotides. This representation also applies to. **(B)** Preference of nucleotide sequences that impact CRISPR/Cas9 gRNAs activity.

Besides, we found our CNN-SVR system has good generalizability in new cell types by using leave-one-cell-out approach on the four testing datasets. By analyzing the changes of prediction results with and without considering epigenetic information, we observed that considering of epigenetic features slightly improves the accuracy of CRISPR/Cas9 gRNA activity prediction. This result was consistent with previous studies (Chen et al., 2017; Chuai et al., 2018; Kim et al., 2018). Chuai et al. found that the prediction on HEK293T became poor (Chuai et al., 2018). They speculated that it was mainly due to the insufficient training data of the HEK293T training dataset. Note that our findings suggested that the reducing epigenetic features may be one possible explanation for the observation. Additionally, the low GC content of dataset HEK293T may be another possible explanation. We concluded that our CNN-SVR gained better generalization and robustness than DeepCRISPR.

Our model focused on gRNA sequence and four kinds of epigenetic features for CRISPR/Cas9 on-target prediction. A recent study on protein-related prediction has shown that integration of other manual extracted features, such as molecular weight and hydrophobicity into the deep learning model could improve the predictive power (Wang et al., 2016). It has been reported that GC content is associated with gRNA activity (Doench et al., 2014; Wang et al., 2014). We thus made a preliminary exploration of adding this sequence-derived feature with our CNN-SVR for gRNA activity prediction on the above four datasets. Note that, the training data and test data were constructed in the same way as described in the section *Comparison CNN-SVR With CNN Model*. Overall, addition of GC content to CNN-SVR increased the predictive ability, with Spearman correlation coefficients of 0.645, 0.656 and 0.608 on datasets HCT116, HELA and HELA, respectively. Detailed results can be found in **Table S6**. Therefore, manual design of proper gRNA features will contribute to the prediction ability. In the future, we plan to develop deep learning models incorporating indirect sequence-derived sequence features to improve the prediction performance, such as chromatin accessibility (Kim et al., 2018), RNA thermodynamics (Abadi et al., 2017), secondary structure of gRNA (Abadi et al., 2017), and GC content, which cannot be automatically obtained by deep learning models.

Visualization method was applied to our model. Note that the PAM and the core region (1-5 nt adjacent to the PAM) are very important for gRNA target efficacy. However, we observed that the most top features were generated by convolving the middle region of the input matrix. Therefore, we believe expanding the upstream and downstream of the target sequence in a proper length can enhance the generalization performance of the model. For example, Kim et al. found 34 bp (4 bp + PAM + 23bp protospacer + 3bp) was adequate as the input sequence of their models in CRISPR/Cpf1 system (Kim et al., 2018).

Several future improvements are expected. First, in the present study, taking advantage of CNN and SVR, we designed the relative concise hybrid CNN-SVR architecture. Research on the deep learning-based model for CRISPR/Cas9 system gRNA cleavage efficiency prediction is still at an early stage. Numerous complex and modern deep learning models await exploration. Second, as pre-training technique has great influence on the final predictive performance, therefore critical to know on what a model was trained before use. In general, sequencing-based models are more general applicable, but are only capable of predicting the genotype changes rather than functional result. On the contrary, phenotypic trained models are fit for recognizing target sites that cause functional changes but limited to numerical experiments with the same condition as the training set. However, the amount of available gRNA knockout data is relatively small, which provides a big challenge for training the deep learning model. Consequently, appropriate data augmentation techniques are needed to increase the training sample size. Third, reasonable encoding schemes, which provide maximum biological characteristics information as well as reducing the compute costs, will boost the CRISPR/Cas9 gRNA activity prediction accuracy. Finally, it is possible that integration of manual extracted features associated with gRNA activity can also improve predictive power of deep learning models.

CONCLUSIONS

In this study, we present CNN-SVR, an efficient and extendable method to automatically learn the sequence features for CRISPR/Cas9 gRNA activity prediction. We adopt a merged CNN architecture for gRNA and its corresponding epigenetic features extraction, and subsequently incorporate SVR classifier to predict gRNA cleavage efficiency. Compared with CNN, two state-of-the-art deep neural network based models (e.g. DeepCRISPR and Seq-deepCpf1) and three machine learning tools (i.e., sgRNA Designer, SSC, and WU-CRISPR), CNN-SVR can effectively exploit features interactions from feed-forward directions to learn deeper features of gRNAs and their corresponding epigenetic features. Experimental results on the published datasets demonstrate the superiority of our CNN-SVR for CRISPR/Cas9 gRNAs on-target activity prediction.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/Peppags/CNN-SVR.

AUTHOR CONTRIBUTIONS

All authors contributed to the project design. GZ wrote the analysis source code, analyzed the data, and drafted the full manuscript. ZD and XD critically revised the final manuscript. All authors read and approved the final manuscript.

FUNDING

This research was funded by the National Natural Science Foundation of China (NSFC) (Grant 61872396, 61872395 and U1611265), and also by Pearl River Nova Program of Guangzhou (201710010044).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01303/full#supplementary-material

REFERENCES

Aach, J., Mali, P., and Church, G. M. (2014). CasFinder: flexible algorithm for identifying specific Cas9 targets in genomes. *BioRxiv*, 005074. doi: 10.1101/005074

Abadi, S., Yan, W. X., Amar, D., and Mayrose, I. (2017). A machine learning approach for predicting CRISPR-Cas9 cleavage efficiencies and patterns underlying its mechanism of action. *PloS Comput. Biol.* 13 (10), e1005807. doi: 10.1371/journal.pcbi.1005807

Badaro, G., Hajj, H., El-Hajj, W., and Nachman, L. (2013). "A hybrid approach with collaborative filtering for recommender systems," in: 2013 9th International Wireless Communications and Mobile Computing Conference (IWCMC) (Publisher: IEEE), pp. 349–354. doi: 10.1109/IWCMC.2013.6583584

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., et al. (2007). CRISPR provides acquired resistance against viruses in Prokaryotes. *Science* 315 (5819), 1709–1712. doi: 10.1126/science.1138140

Bengio, Y. (2012). "Deep learning of representations for unsupervised and transfer learning," in: Proceedings of ICML workshop on unsupervised and transfer learning, *Conferences sand Proceedings*. pp. 17–36.

Chari, R., Mali, P., Moosburner, M., and Church, G. M. (2015). Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. *Nat. Methods* 12 (9), 823–826. doi: 10.1038/nmeth3473

Chen, L., Wang, S. P., Zhang, Y. H., Li, J. R., Xing, Z. H., Yang, J., et al. (2017). Identify key sequence features to improve CRISPR sgRNA efficacy. *IEEE Access* PP (99), 1–1. doi: 10.1109/ACCESS.2017.2775703

Cheng, P. M., and Malhi, H. S. (2017). Transfer learning with convolutional neural networks for classification of abdominal ultrasound images. *J. Digit. Imaging* 30 (2), 234. doi: 10.1007/s10278-016-9929-2

Chuai, G. H., Wang, Q. L., and Liu, Q. (2016). In Silico Meets In Vivo : towards computational CRISPR-based sgRNA design. *Trends In Biotechnol.* 35 (1), 12. doi: 10.1016/j.tibtech.2016.06.008

Chuai, G., Ma, H., Yan, J., Chen, M., Hong, N., Xue, D., et al. (2018). DeepCRISPR: optimized CRISPR guide RNA design by deep learning. *Genome Biol.* 19 (1), 80. doi: 10.1186/s13059-018-1459-4

Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., et al. (2013). Multiplex genome engineering using CRISPR/Cas systems. *Science* 339 (6121), 819–823. doi: 10.1126/science.1231143

Consortium, E. P. (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 306 (5696), 636–640. doi: 10.1126/science.1105136

Doench, J. G., Hartenian, E., Graham, D. B., Tothova, Z., Hegde, M., Smith, I., et al. (2014). Rational design of highly active sgRNAs for CRISPR-Cas9-mediated geneinactivation. *Nat. Biotechnol.* 32 (12), 1262. doi: 10.1038/nbt3026

Doench, J. G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E. W., Donovan, K. F., et al. (2016). Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* 34 (2), 184. doi: 10.1038/nbt3437

Erard, N., Knott, S. R. V., and Hannon, G. J. (2017). A CRISPR resource for individual, combinatorial, or multiplexed gene knockout. *Mol. Cell* 67 (2), 348–354 e344. doi: 10.1016/j.molcel.2017.06.030

Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K. R., MacLeod, G., et al. (2015). High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. *Cell* 163 (6), 1515–1526. doi: 10.1016/j.cell.2015.11.015

Hinz, J. M., Laughery, M. F., and Wyrick, J. J. (2015). Nucleosomes inhibit Cas9 endonuclease activity in vitro. *Biochemistry* 54 (48), 7063–7066. doi: 10.1021/acs.biochem.5b01108

Horlbeck, M. A., Gilbert, L. A., Villalta, J. E., Adamson, B., Pak, R. A., Chen, Y., et al. (2016). Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. *Elife* 5, e19760. doi: 10.7554/eLife.19760.031

Hsu, P. D., Scott, D. A., Weinstein, J. A., Ran, F. A., Konermann, S., Agarwala, V., et al. (2013). DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* 31 (9), 827. doi: 10.1038/nbt2647

Huang, F., and LeCun, Y. (2006). "Large-scale learning with svm and convolutional nets for generic object recognition," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. (Publisher: IEEE).

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., and Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* 337 (6096), 816–821. doi: 10.1126/science.1225829

Kim, D., Kim, J., Hur, J. K., Been, K. W., Yoon, S. H., and Kim, J. S. (2016). Genome-wide analysis reveals specificities of Cpf1 endonucleases in human cells. *Nat. Biotechnol.* 34 (8), 863–868. doi: 10.1038/nbt3609

Kim, H. K., Song, M., Lee, J., Menon, A. V., Jung, S., Kang, Y. M., et al. (2017). In vivo high-throughput profiling of CRISPR-Cpf1 activity. *Nat. Methods* 14 (2), 153–159. doi: 10.1038/nmeth4104

Kim, H. K., Min, S., Song, M., Jung, S., Choi, J. W., Kim, Y., et al. (2018). Deep learning improves prediction of CRISPR-Cpf1 guide RNA activity. *Nat. Biotechnol.* 36 (3), 239–241. doi: 10.1038/nbt4061

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *Comput. Sci.*

Kleinstiver, B. P., Tsai, S. Q., Prew, M. S., Nguyen, N. T., Welch, M. M., Lopez, J. M., et al. (2016). Genome-wide specificities of CRISPR-Cas Cpf1 nucleases in human cells. *Nat. Biotechnol.* 34 (8), 869–874. doi: 10.1038/nbt3620

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "ImageNet Classification with Deep Convolutional Neural Networks," in *International Conference on Neural Information Processing Systems*. (Publisher: Neural Information Processing Systems Foundation, Inc. (NIPS)).

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521 (7553), 436–444. doi: 10.1038/nature14539

Li, J., Yan, J., Deng, D., Shi, W., and Deng, S. (2016). No-reference image quality assessment based on hybrid model. *Signal Image Video Process.* 11 (6), 985–992. doi: 10.1007/s11760-016-1048-5

Liaw, A., and Wiener, M. (2002). Classification and regression by randomForest. *R News* 2 (3), 18–22.

Listgarten, J., Weinstein, M., Kleinstiver, B. P., Sousa, A. A., Joung, J. K., Crawford, J., et al. (2018). Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs. *Nat. BioMed. Eng.* 2 (1), 38–47. doi: 10.1038/s41551-017-0178-6

Maqsood, I., Khan, M. R., and Abraham, A. (2004). An ensemble of neural networks for weather forecasting. *Neural Comput. Appl.* 13 (2), 112–122. doi: 10.1007/s00521-004-0413-4

Moreno-Mateos, M. A., Vejnar, C. E., Beaudoin, J. D., Fernandez, J. P., Mis, E. K., Khokha, M. K., et al. (2015). CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nat. Methods* 12 (10), 982–988. doi: 10.1038/nmeth3543

Mori, K., Matsugu, M., and Suzuki, T. (2005). "Face Recognition Using SVM Fed with Intermediate Output of CNN for Face Detection," in *MVA*, *Conferences and Proceedings*. 410–413.

Mukaka, M. M. (2012). Statistics corner: a guide to appropriate use of correlation coefficient in medical research. *Malawi Med. J.* 24 (3), 69–71. doi: 10.2166/wh.2012.000

Niu, X.-X., and Suen, C. Y. (2012). A novel hybrid CNN–SVM classifier for recognizing handwritten digits. *Pattern Recognit.* 45 (4), 1318–1325. doi: 10.1016/j.patcog.2011.09.021

Pattanayak, V., Lin, S., Guilinger, J. P., Ma, E., Doudna, J. A., and Liu, D. R. (2013). High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nat. Biotechnol.* 31 (9), 839. doi: 10.1038/nbt2673

Shin, H. C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., et al. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans. Med. Imaging* 35 (5), 1285–1298. doi: 10.1109/TMI.2016.2528162

Shou, J., Li, J., Liu, Y., and Wu, Q. (2018). Precise and predictable CRISPR chromosomal rearrangements reveal principles of Cas9-mediated nucleotide insertion. *Mol. Cell* 71 (4), 498–509.e494. doi: 10.1016/j.molcel.2018.06.021

Stemmer, M., Thumberger, T., del Sol Keyer, M., Wittbrodt, J., and Mateo, J. L. (2015). CCTop: an intuitive, flexible and reliable CRISPR/Cas9 target prediction tool. *PloS One* 10 (4), e0124633. doi: 10.1371/journal.pone.0124633

Wang, T., Wei, J. J., Sabatini, D. M., and Lander, E. S. (2014). Genetic screens in human cells using the CRISPR-Cas9 system. *Science* 343 (6166), 80–84. doi: 10.1126/science.1246981

Wang, S., Peng, J., Ma, J., and Xu, J. (2016). Protein secondary structure prediction using deep convolutional neural fields. *Sci. Rep.* 6, 18962. doi: 10.1038/srep18962

Whitney, A. W. (2006). A direct method of nonparametric measurement selection. *IEEE Trans. Comput.* C-20 (9), 1100–1103. doi: 10.1109/T-C.1971.223410

Wilson, L. O. W., O'Brien, A. R., and Bauer, D. C. (2018). The current state and future of CRISPR-Cas9 gRNA design tools. *Front. Pharmacol.* 9, 749. doi: 10.3389/fphar.2018.00749

Woźniak, M., Graña, M., and Corchado, E. (2014). A survey of multiple classifier systems as hybrid systems. *Inf. Fusion* 16, 3–17. doi: 10.1016/j.inffus.2013.04.006

Wong, N., Liu, W., and Wang, X. (2015). WU-CRISPR: characteristics of functional guide RNAs for the CRISPR/Cas9 system. *Genome Biol.* 16, 218. doi: 10.1186/s13059-015-0784-0

Xie, B., Jankovic, B. R., Bajic, V. B., Song, L., and Gao, X. (2013). Poly(A) motif prediction using spectral latent features from human DNA sequences. *Bioinformatics* 29 (13), i316–i325. doi: 10.1093/bioinformatics/btt218

Xu, H., Xiao, T., Chen, C. H., Li, W., Meyer, C. A., Wu, Q., et al. (2015). Sequence determinants of improved CRISPR sgRNA design. *Genome Res.* 25 (8), 1147–1157. doi: 10.1101/gr.191452.115

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: 09 January 2020
doi: 10.3389/fgene.2019.01292

Data-Mining Approach on Transcriptomics and Methylomics Placental Analysis Highlights Genes in Fetal Growth Restriction

Floris Chabrun[1,2], Noémie Huetz[2,3], Xavier Dieu[1,2], Guillaume Rousseau[1,2], Guillaume Bouzillé[4,5], Juan Manuel Chao de la Barca[1,2], Vincent Procaccio[1,2], Guy Lenaers[1,2], Odile Blanchet[6], Guillaume Legendre[7], Delphine Mirebeau-Prunier[1,2], Marc Cuggia[4,5], Philippe Guardiola[8], Pascal Reynier[1,2] and Geraldine Gascoin[2,3]*

[1] *Département de Biochimie et Génétique, Centre Hospitalier Universitaire, Angers, France,* [2] *Unité Mixte de Recherche (UMR) MITOVASC, Équipe Mitolab, Centre National de la Recherche Scientifique (CNRS) 6015, Institut National de la Santé et de la Recherche Médicale (INSERM) U1083, Université d'Angers, Angers, France,* [3] *Réanimation et Médecine Néonatales, Centre Hospitalier Universitaire, Angers, France,* [4] *Laboratoire du Traitement de l'Image et du Signal, INSERM, UMR 1099, Université Rennes 1, Rennes, France,* [5] *Département d'Information médicale et dossiers médicaux, Centre Hospitalier Universitaire, Rennes, France,* [6] *Centre de Ressources Biologiques, Centre Hospitalier Universitaire, Angers, France,* [7] *Département de Gynécologie Obstétrique, Centre Hospitalier Universitaire, Angers, France,* [8] *Service de Génomique Onco-Hématologique, Centre Hospitalier Universitaire, Angers, France*

Intrauterine Growth Restriction (IUGR) affects 8% of newborns and increases morbidity and mortality for the offspring even during later stages of life. Single omics studies have evidenced epigenetic, genetic, and metabolic alterations in IUGR, but pathogenic mechanisms as a whole are not being fully understood. An in-depth strategy combining methylomics and transcriptomics analyses was performed on 36 placenta samples in a case-control study. Data-mining algorithms were used to combine the analysis of more than 1,200 genes found to be significantly expressed and/or methylated. We used an automated text-mining approach, using the bulk textual gene annotations of the discriminant genes. Machine learning models were then used to explore the phenotypic subgroups (premature birth, birth weight, and head circumference) associated with IUGR. Gene annotation clustering highlighted the alteration of cell signaling and proliferation, cytoskeleton and cellular structures, oxidative stress, protein turnover, muscle development, energy, and lipid metabolism with insulin resistance. Machine learning models showed a high capacity for predicting the sub-phenotypes associated with IUGR, allowing a better description of the IUGR pathophysiology as well as key genes involved.

Keywords: data mining, methylomics, intrauterine growth restriction, multi-omics, text-mining, transcriptomics

OPEN ACCESS

Edited by:
Mehdi Pirooznia, National Heart, Lung, and Blood Institute, United States

Reviewed by:
Amit Kumar Yadav, Translational Health Science and Technology Institute, India
Izabela Makałowska, Adam Mickiewicz University, Poland
Amanda Vlahos, Murdoch Children's Research Institute, Australia

***Correspondence:**
Floris Chabrun
floris.chabrun@chu-angers.fr

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

Received: *10 August 2019*
Accepted: *25 November 2019*
Published: *09 January 2020*

Citation:
Chabrun F, Huetz N, Dieu X, Rousseau G, Bouzillé G, Chao de la Barca JM, Procaccio V, Lenaers G, Blanchet O, Legendre G, Mirebeau-Prunier D, Cuggia M, Guardiola P, Reynier P and Gascoin G (2020) Data-Mining Approach on Transcriptomics and Methylomics Placental Analysis Highlights Genes in Fetal Growth Restriction. Front. Genet. 10:1292. doi: 10.3389/fgene.2019.01292

Abbreviations: IUGR, intrauterine growth restriction; PE, pre-eclampsia; Se, sensitivity; Sp, specificity; tf, term frequency; idf, inverse document frequency; SVM, support vector machine; RMSE, root-mean-square deviation; C-section, caesarean section.

INTRODUCTION

Intrauterine growth restriction (IUGR) is a frequent complication of pregnancy with a prevalence in up to 5% to 10% in the general population (Zhang et al., 2015). It is defined as a restriction of fetal growth during pregnancy, "a fetus that doesn't reach its growth potential" (Vayssière et al., 2015). It can lead to a birth weight and/or length below the tenth percentile for a given gestational age in newborns, thus considered as "Small for Gestational Age" (Vayssière et al., 2015). IUGR represents a major public health problem, being one of the main causes of premature birth, perinatal mortality, and neurological and respiratory morbidities (Flamant and Gascoin, 2013). It is also suspected to be a determining factor in the development of cardiovascular diseases, obesity, and type 2 diabetes in adulthood (Gascoin and Flamant, 2013).

Fetal growth is a complex process that involves fetal genetics, nutrient and oxygen availability, and maternal nutrition, as well as growth factors and hormones from maternal, fetal, and placental origin (Murki, 2014). Fetal growth is inseparable from placental growth and requires a continuous supply of nutrients that is adapted to each period of pregnancy (Sharma et al., 2016).

IUGR remains a complex problem for the clinician. Placental dysfunction and vascular underperfusion are involved in the largest proportion of cases (Kaplan, 2007; Malhotra et al., 2019). It results from utero-placental insufficiency due to abnormal uterine artery remodeling in the first trimester of pregnancy and may or may not be associated with pre-eclampsia (PE). However, while many risk factors have been identified, placental insufficiency is still unexplained in up to 60% of cases (Malhotra et al., 2019).

Epigenetics (Xiao et al., 2016) and gene expression (Buffat et al., 2007; Madeleneau et al., 2015) reprogramming play a central role in IUGR. However, the pathophysiological connections between these two fields of high-throughput analyses have only recently begun to be studied (Ding and Cui, 2017). Although many tools have been developed to analyze and integrate multi-omics data, this task remains a challenge in medicine (Gomez-Cabrero et al., 2014). Many features originating from the variance between samples and the complexity of the statistical data processing require developing data-driven approaches rather than classical hypothesis-driven approaches (van Helden, 2013). The exploration of pathophysiological conditions with such data-driven approaches must integrate many processes from clinical and biological data collection, through complex data normalization and mathematical and bioinformatics modeling, to the final interpretation and data visualization.

When dealing with a short list of genes, the exploration of their roles and underlying patterns is usually carried out through "manual" interpretation, using both annotations and personal knowledge. This "manual" interpretation may be used to categorize the genes, or to seek patterns in roles, functions, or localizations, underpinning the pathology or context studied. When dealing with thousands of significant gene features (e.g. expression levels or methylation levels), the interpretation becomes humanly untenable, due to time and memory limits. Rather than limiting our literature review to a small subset of the most significantly altered genes, we used text-mining algorithms to perform an unsupervised analysis of those genes. Those algorithms have already been used to categorize and summarize text corpora based on similarities in their content (Aggarwal and Zhai, 2012).

With the aim of having an extended vision of the pathophysiological processes at the origin of IUGR, while identifying the most predominant deregulated pathways that may be targeted for therapeutic purposes, we used machine-learning models to explore the relationship between placental transcriptomics and methylomics variations and IUGR. The highly predictive models obtained from IUGR and its sub-phenotypes were then used to highlight the genes with a high correlation with IUGR clinical severity, and thus with a high therapeutic potential.

MATERIAL AND METHODS

The global workflow is summarized in **Figure 1**.

Patients

All placentas were collected from Angers University Hospital. This study was approved by the Ethics Committee of Angers. All patients gave their informed consent for the use of their placenta. Clinical data related to the mother and the fetus, as well as neonatal data, were collected from the patients' obstetric files. The cohort was registered at the French *CNIL* (*Commission Nationale de l'Informatique et des Libertés* no. pWP03752UL, ethics committee for the collection of clinical data from patient records). The study was validated by the French *CPP* (*Comité de Protection des Personnes*) and registered to the French Ministry of Research under number DC-2011-1467. The study was conducted in accordance with the declaration of Helsinki.

Placentas were obtained from caesarean sections before onset of labor or from vaginal delivery. For the analysis, patients were classified into two groups: IUGR and control group. The IUGR group was defined by a reduction of fetal growth during gestation, with a notch observed by Echo-Doppler in at least one uterine artery and with Doppler abnormalities on umbilical Doppler and/or cerebral Doppler and/or ductus venosus, and with a birth weight below the tenth percentile according to Audipog growth curves (American College of Obstetricians and Gynecologists, 2013) and confirmed by the anatomopathological analysis of the placenta after birth. The control group was defined by women with normal pregnancy and who underwent a planned caesarean section. All obstetrical and neonatal data were collected prospectively from medical records.

Placental Samples

To avoid degradation, only placental tissues dissected within a time frame of 30 min after delivery were included. After removal

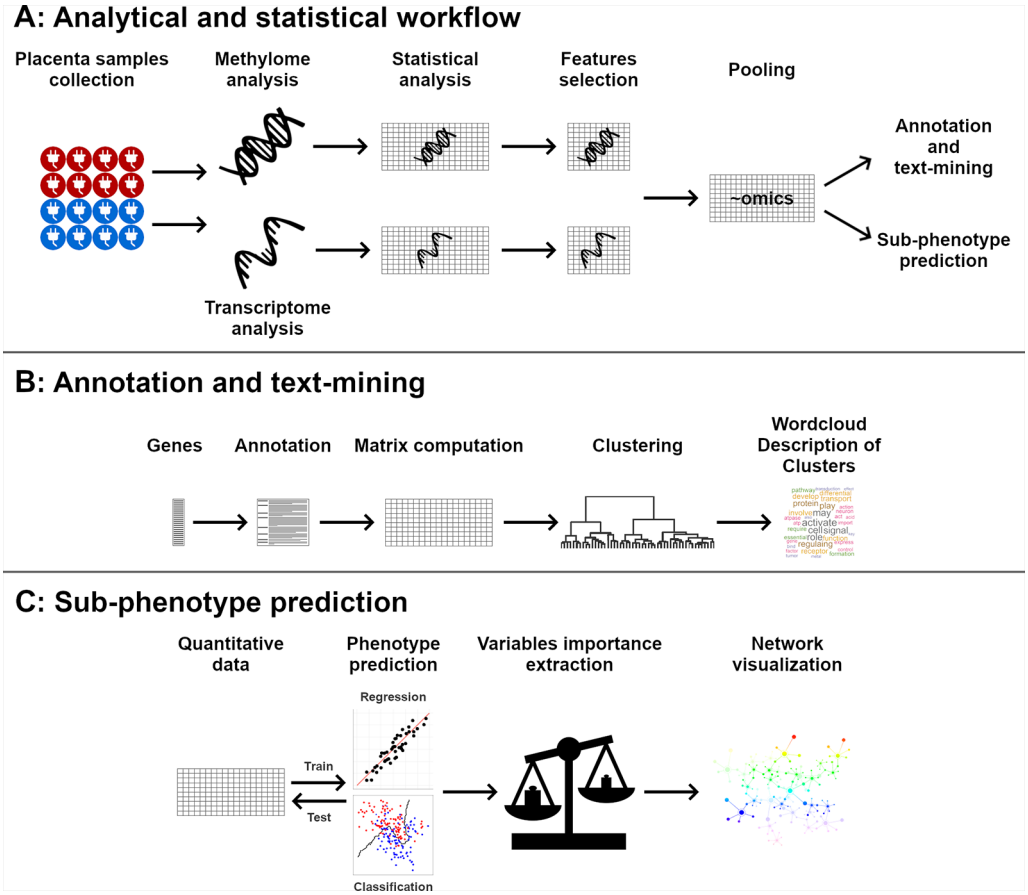

FIGURE 1 | Global workflow of the analysis. Placentas methylome and transcriptome were analyzed **(A)**. Significant genes were clustered and described using text annotations **(B)**. Quantitative data were used to predict phenotypic data, and the importance of each gene in phenotype prediction was visualized using networks **(C)**.

of the maternal decidua and amniotic membrane, sections of 1 cm^3 of placental villi were dissected from four different cotyledons between the basal and chorionic plates, as previously described (Gascoin-Lachambre et al., 2010). After vigorous washing with PBS to remove maternal blood, tissues were immediately frozen in liquid nitrogen, before storage at −80 °C, to further extract DNA and RNA. Placentas were then sent for anatomopathological analysis or stored at the biological core facility at Angers University Hospital.

DNA Preparation and Microarray Hybridization

Genomic DNA extraction was performed manually using a QIAamp DNA mini QIAcube Kit (Qiagen, Venlo, Netherlands), according to the manufacturer's protocol.

DNA was treated with bisulfite using an EZ-96 DNA Methylation Kit on a Zymo Spin I-96 column (Zymo Research, Irvine, CA, U.S.A.). Bisulfite-converted DNA was amplified, fragmented, and hybridized to Illumina Human Methylation 450k microarrays using an Illumina Hybridization Oven (Illumina, San Diego, CA, U.S.A.), according to the manufacturer's protocol. Slides were analyzed by an Illumina I-Scan (Illumina, San Diego, CA, U.S.A.).

Raw iDAT files were directly imported in R software (R Development Core Team, 2008) and processed using the R minfi package (Aryee et al., 2014). Raw data were normalized using functional normalization (Fortin et al., 2014) before constructing the beta matrix for all 36 samples and 485,512 CpG sites (methylomics dataset).

RNA Preparation and Microarray Hybridization

Total RNA was extracted after lysing samples with TRIzol reagent (Life Technologies, Carlsbad, CA, U.S.A.), using the RNeasy Micro kit (Qiagen, Venlo, Netherlands), according to the manufacturer's recommendations. Biotinylated, amplified cRNA was generated using the Illumina Total Prep RNA Amplification kit (Ambion, Life Technologies, Carlsbad, CA, U.S.A.), according to the manufacturer's recommendations. cRNA was hybridized on Illumina HumanHT-12 v4 Expression BeadChips, stained, and detected with the iScan system, according to the manufacturer's protocol (Illumina,

San Diego, CA, U.S.A.). A total of 47,323 marker probes were assessed, of which: 47,231 elements with sequences, with 46,841 with at least one genome alignment, including 34,627 elements mapped to at least one among 22,283 unique genes. GenomeStudio 2011 (version 1) and its Expression Analysis Module (version 1.9.0) were used for signal extraction and quantile normalization (Illumina, San Diego, CA, U.S.A.).

Normalized data for all 47,323 marker probes and 36 samples were imported into R software (R Development Core Team, 2008) and processed as described below (transcriptomics dataset).

Omics Data Integration

Each omics dataset was processed independently. Levene's tests were used to assess the comparability of variances between control and IUGR groups. Significant features were determined using Student's t-tests. Alpha thresholds for *p*-value significance were set to $\alpha = 0.05$. For Student's t-tests, *p*-values were adjusted into *q*-values using the Benjamini-Hochberg method in order to control the false discovery rate. The $\frac{IUGR}{control}$ fold-change was computed for all significant features. Only features with Levene's test *p*-value ≥ 0.05 and Benjamini-Hochberg adjusted Student's t-test q-value < 0.05 were considered significant.

Gene Annotation and Text-Mining

All genes showing a significant alteration in methylation or expression were annotated using abstracts available on PubMed, by automatic retrieval. Genes without available annotations were discarded. Abstracts were pre-processed by removing punctuation, short words (words of three characters or fewer) and stop words (i.e. common language non-specific words), and stemming (Willett, 2006). They were then analyzed by taking into account, in the same analytical process, unigrams, bigrams, and trigrams, commonly denoted as terms. A normalized term-frequency inverse-document-frequency (tf-idf) matrix (Aggarwal and Zhai, 2012) was then computed based on the frequency and specificity of each term in each gene summary, using the formula:

$$M_{i,j} = tf_i \times idf_i$$

With the inverse document frequency idf_i for the term *i*:

$$idf_i = \log_2\left(\frac{|D|}{|\{d|t_i \in d|\}|}\right)$$

where Mi,j is the value in the matrix for the term *i* and gene *j*,tf_i is the number of occurrences of the term *i* in the gene *j* summary divided by the total number of terms in the summary, $|D|$ is the number of genes and $|\{d|t_i \in d|\}|$ is the number of gene summaries where the term *i* appears.

Due to the large dimension of the initial tf-idf matrix, a Latent Semantic Analysis (LSA) (Evangelopoulos, 2013) was performed in order to reduce its dimension and render further analyses possible. K-means was then used to perform clustering based on gene annotations similarity. Clusters were then summarized by terms closest to the cluster centers.

Phenotype Prediction and Network Visualization

Support vector machines (SVM) are state-of-the-art machine-learning models that have already been successfully applied to several omics studies (Ben-Hur et al., 2008). They can successfully highlight non-linear correlations between genes and phenotypic traits, in order to highlight genes based on their links with several phenotypic traits (Altmann et al., 2010). Furthermore, SVM models are particularly suitable for high-dimensionality datasets, such as results of high-throughput analyses (Vanitha et al., 2015).

SVM models were trained using grid search cross-validation to predict four phenotypic traits as a function of omics data: control/IUGR group, premature birth (see below), birth weight, and head circumference at birth. These four phenotypic traits were chosen because of their known relevance in the IUGR pathophysiology. Term birth is defined by the International Classification of Diseases as between 37 (included) and 42 (excluded) weeks (Quinn et al., 2016), otherwise 39.43 ± 2.43 weeks. To simplify, pregnancy term was expressed as a variable named premature birth, computed with the formula:

$$Premature\ birth = 39 - Gestational\ Age$$

Since gestational age and the newly-created variable, premature birth, are linearly correlated, this simplifies yet does not alter the interpretation of the results of the model's predictions. Values >2 therefore indicate pre-term newborns, while values ≤-3 indicate post-term newborns.

Both head circumference at birth and birth weight were expressed as Z-scores according to the gestational age and gender, based on Olsen growth curves (Olsen et al., 2010), to standardize values between infants born at different terms. Case-control classification is important to verify the integrity of the dimension-reduced dataset. Birth weight is a criterion of severity of the IUGR. Head circumference at birth is a criterion of high severity, due to the brain sparing effect (Cohen et al., 2015). Premature birth is indirectly linked to severity of these. Indeed, in most cases during IUGR pregnancies, a delivery is induced or carried out *via* caesarean section, to prevent either maternal or fetal damage. Exploring factors correlated with the premature birth may therefore allow exploring severity symptoms not directly and only linked to IUGR.

The dimensionality of the omics dataset had to be reduced before training the SVM, to reduce noise and achieve better model predictions (Keogh and Mueen, 2010). For this reason, only features with a significant difference between IUGR and control groups were used to train SVM models ($q < 0.05$, after Benjamini-Hochberg adjustment). Several methods may be used to reduce the dimensionality of a dataset (Guyon and Elisseeff, 2003). Features selection was preferred compared to other methods like Principal Components Analysis as it allows the use of the initial variables instead of computing new, abstract dimensions, making the final interpretation easier. Student's t-tests have already been evidenced as an effective method for features selection (Haury et al., 2011). By using Student's t-tests as the features selection method, this step could be applied

seamlessly to our omics analyses results, without modifying or altering the results.

The dataset was randomly partitioned into training and test sets, with a ratio of two-thirds/one-third, using stratified sampling in order to respect the original $\frac{case}{control}$ ratio. Due to the low number of samples and the imbalance between IUGR and control samples, Synthetic Minority Over-sampling Technique (SMOTE) was used in order to synthetically increase the training set size (Chawla et al., 2002). Test sets were not modified to ensure unbiased results when measuring models' performances. Hyperparameters were fine-tuned with grid search cross-validation. Model results were assessed using accuracy for classification, and Pearson's correlation factor for regression.

The variable importance for predicting each phenotypic trait was computed for each feature by Permutation Importance (Breiman, 2001). These results were used to carry out a network visualization to assess the importance of each feature in the prediction of each phenotypic trait.

Computational Tools

R software (version 3.4.1) and Python (version 3.6) were used to carry out all data processing and analysis, as well as to output all plots (van Rossum, 1995; R Development Core Team, 2008). Heat maps were created using the gplots package (Warnes et al., 2016). Gene functional annotation analysis was performed for both gene expression and gene methylation using the DAVID 6.8 online tool (Huang et al., 2009a, 2009b). Genes were annotated with abstracts available from PubMed (10/10/2019) using easyPubMed (Fantini, 2019). Text-mining and SVM computing were processed using the python scikit-learn library (Pedregosa et al., 2011). Word clouds were created using the wordcloud R software package (Fellows, 2014). Hierarchical clustering was performed using the R software base package. Networks were constructed using Cytoscape (Shannon et al., 2003). The GIMP software was used to refine figures.

RESULTS

Cohort

Patient cohort is described in **Table 1**. It should be noted that while the control group is smaller, controls are much more homogeneous concerning clinically relevant phenotypic traits discussed below. F-tests show a significantly lower variance in this control group for gestational age at birth (in grams) (p = 4.48E-5), head circumference at birth (in centimeters) (p = 1.08E-3), and APGAR at 5 min (p = 3.48E-5).

Univariate Analyses

A total of 1651 features (1,072 DNA methylation sites, 579 transcripts) showed significantly different values between IUGR and control groups ($q < 0.05$). The full list of significant features is available in **Supplementary Table 1**.

Since a significant difference in mean gestational age had been observed between IUGR and control groups, univariate analyses were re-run after excluding IUGR samples with a gestational age lower than 37 weeks. Kendall correlation tests were then performed to compare Student's t-tests results obtained for the whole cohort and for the high gestational age

TABLE 1 | Description of the patient cohort. p-values were computed using Wilcoxon tests (quantitative values) or Fisher tests (percentages).

			Control group (n = 8)		IUGR group (n = 28)		p
Maternal data	Age (years)		35.4 ± 3.9	8	29.1 ± 5.9	28	0.006
	BMI before pregnancy	(kg/m^2)	23.7 ± 7.0	8	25.1 ± 7.9	28	N.S.
	Tobacco consumption	Before pregnancy	0 (0.0%)	8	2 (7.1%)	28	N.S.
		During pregnancy	0 (0.0%)	8	9 (32.1%)	28	N.S.
	Ethnic group	European	7 (87.5%)	8	26 (92.9%)	28	N.S.
		North African	1 (12.5%)	8	2 (7.1%)	28	N.S.
Obstetric data	Gestity		4.0 ± 2.1	8	2.5 ± 1.9	28	0.03
	Parity		2.6 ± 1.3	8	1.4 ± 0.9	28	0.005
	Weight gain (kg)		10.5 ± 10.5	8	9.1 ± 6.4	24	N.S.
	Type of delivery	Vaginal delivery	0 (0%)	8	5 (17.9%)	28	N.S.
		C-section	8 (100%)	8	23 (82.1%)	28	N.S.
	Pathology	IUGR	0 (0%)	8	16 (57.1%)	28	N/A
		IUGR + PE	0 (0%)	8	12 (42.9%)	28	N/A
Newborn data	Gestational age (week)		38.7 ± 0.7	8	34.0 ± 3.9	28	<0.001
	Gender	Boy	4 (50.0%)	8	9 (32.1%)	28	N.S.
		Girl	4 (50.0%)	8	19 (67.9%)	28	N.S.
	Birth weight	(Z-score)	−0.07 ± 0.89	8	−2.02 ± 0.75	28	<0.001
		(g)	3346 ± 444	8	1,524 ± 664	28	<0.001
	Birth size	(Z-score)	−0.47 ± 0.74	7	−1.90 ± 0.80	26	<0.001
	Birth size	(cm)	49.2 ± 1.8	7	39.2 ± 5.2	26	<0.001
	Head circumference at birth	(Z-score)	0.22 ± 0.49	7	−1.30 ± 0.86	27	<0.001
	Head circumference at birth	(cm)	34.6 ± 0.9	7	29.0 ± 3.4	27	<0.001
	APGAR at 5 min		9.88 ± 0.35	8	9.11 ± 2.08	28	N.S.
	Resuscitation at birth		0 (0%)	8	12 (42.9%)	28	0.03
	NICU		0 (0%)	8	18 (64.3%)	28	0.003

BMI, body mass index; PE, pre-eclampsia; NICU, neonatal intensive care unit; N.S., non-significant versus $\alpha = 0.05$; N/A, not applicable.

restricted subset. Gene expression and gene methylation features were significantly correlated ($p < 0.001$, $\tau = 0.45$; $p < 0.001$, $\tau = 0.40$, respectively).

Heat maps picturing all genes with significant expression (**Figure 2**) or methylation (**Figure 2**) alteration showed a global hypomethylation, as opposed to a balanced ratio between the number of overexpressed and underexpressed transcripts. While hierarchical clustering distinctly separated IUGR from control samples, IUGR samples appeared divided into two different clusters for both heatmaps, even though the exact distribution of IUGR samples is not exactly the same for epigenetic and expression alterations. In order to explain this behavior, gestational age at birth of IUGR samples according to clusters was plotted in **Figure 3**.

Gene functional annotation analysis, performed with DAVID, showed gene expression and/or methylation alterations significantly associated with several pathways ($p < 0.05$), including: NAD-binding, histone acetylation, mTOR signaling pathway, lysosome, cell-cell adhesion and cell junction, calmodulin binding, and carbohydrates metabolism. The complete results are available in **Supplementary Table 2**.

Only 25 genes were found to be altered both in methylome and transcriptome (**Table 2**). Among these 25 genes, eight show a significant linear correlation between methylation and expression.

Textual Annotation and Text-Mining

Among these 1,651 features, 1,269 unique genes could be identified, and textual annotations were successfully retrieved for 1,259 of them. A total of 196,918 abstracts were retrieved (95% confidence interval: [146;167] abstracts per gene). LSA allowed reducing the dimension from 135,220 unique terms among all abstracts to 1,000 principal components, while retaining 97% of the initial tf-idf matrix variance. Genes were classified into 24 clusters. The cluster sizes ranged from 7 (0.6%) to 241 (19.1%) genes.

These clusters were summarized by word clouds picturing the most frequent and specific terms among the gene clusters,

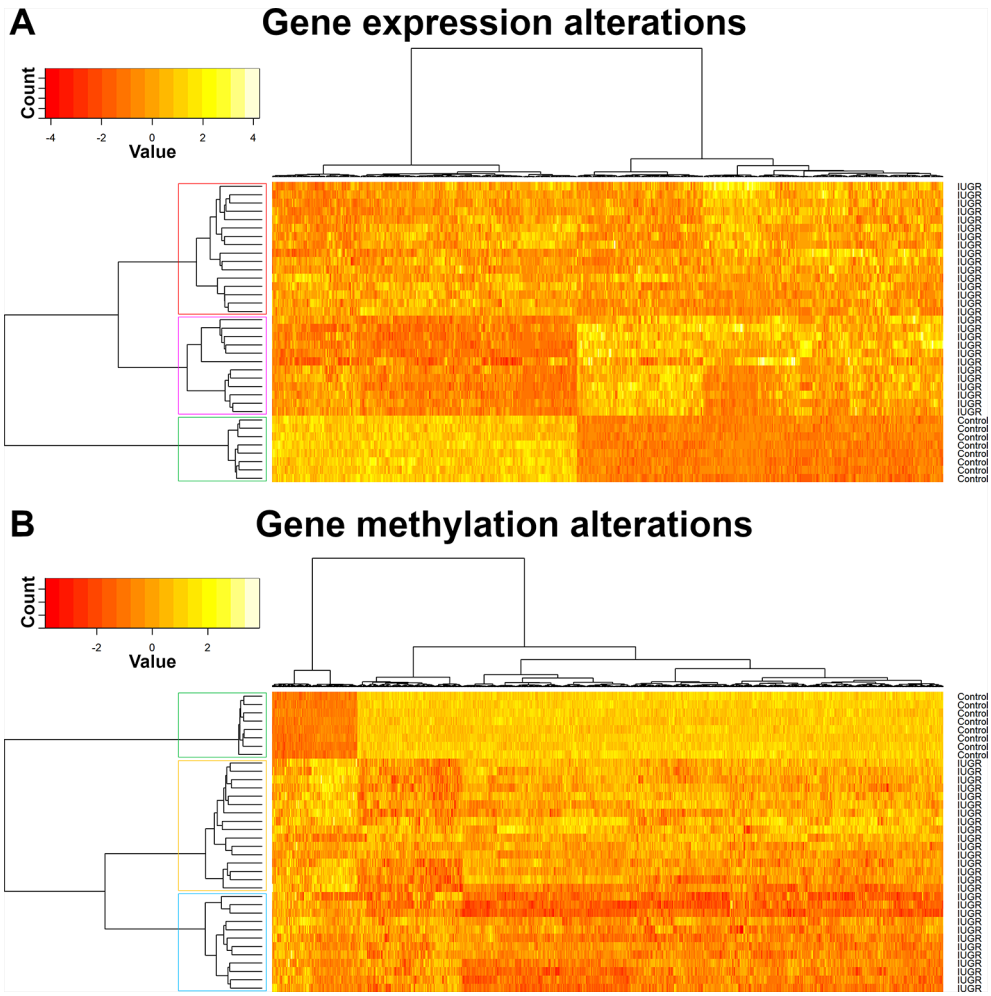

FIGURE 2 | Hierarchical clustering of samples, gene expression **(A)** and methylation **(B)**.

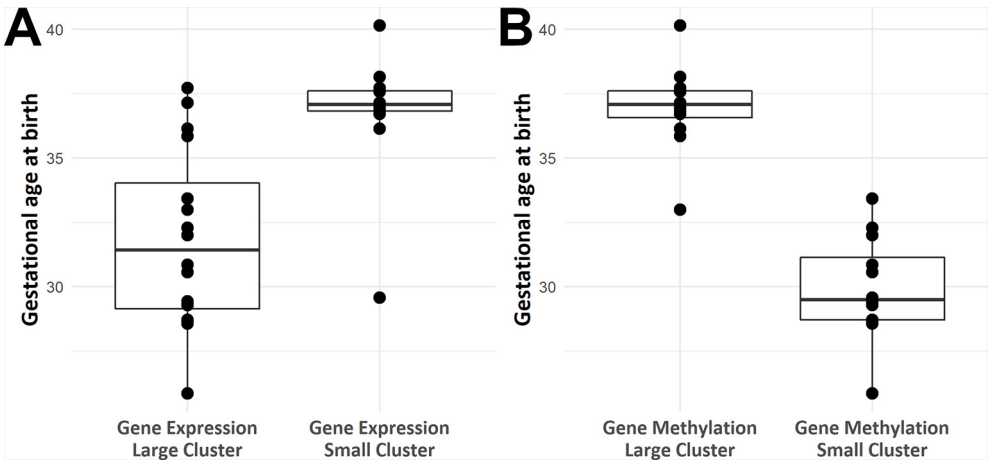

FIGURE 3 | Box plots of gestational age at birth according to IUGR samples position in hierarchical clustering based on methylomics **(A)** and transcriptomics **(B)** data.

TABLE 2 | Genes found altered in both methylome and transcriptome. Numbers in brackets refer to the number of methylation sites (methylome) and transcripts (transcriptome) found significantly altered.

Gene symbol	Gene name	Epigenetics (sites count/total)	Gene expression (transcripts count/total)	r
PAPPA2	*Pregnancy-Associated Plasma Preproprotein-A2*	Hypomethylated (2/13)	Overexpressed (2/2)	-0.76
AP2A1	*Adaptor Related Protein Complex 2 Subunit Alpha 1*	Hypomethylated (1/26)	Underexpressed (2/3)	N.S.
BCL6	*B Cell CLL/Lymphoma 6*	Hypomethylated (2/55)	Overexpressed (1/2)	-0.65
SLC2A1	*Solute Carrier Family 2 Member 1*	Hypomethylated (2/34)	Overexpressed (1/1)	-0.42
UNKL	*Unkempt Family Like Zinc Finger*	Hypomethylated (2/74)	Underexpressed (1/3)	N.S.
WSB1	*WD Repeat and SOCS Box Containing 1*	Hypomethylated (1/19)	Underexpressed (2/3)	N.S.
AFAP1	*Actin Filament Associated Protein 1*	Hypomethylated (1/103)	Overexpressed (1/3)	N.S.
ALDOA	*Aldolase, Fructose-Bisphosphate A*	Hypomethylated (1/27)	Overexpressed (1/4)	-0.43
ALKBH5	*AlkB Homolog 5, RNA Demethylase*	Hypomethylated (1/23)	Overexpressed (1/1)	N.S.
C1QTNF1	*C1q And TNF Related 1*	Hypomethylated (1/40)	Underexpressed (1/3)	0.40
CALM1	*Calmodulin 1*	Hypermethylated (1/20)	Overexpressed (1/1)	N.S.
DGKZ	*Diacylglycerol Kinase Zeta*	Hypomethylated (1/62)	Overexpressed (1/3)	N.S.
DLX5	*Distal-Less Homeobox 5*	Hypomethylated (1/47)	Overexpressed (1/1)	N.S.
FLNB	*Filamin B*	Hypomethylated (1/40)	Overexpressed (1/1)	-0.58
FOXK1	*Forkhead Box K1*	Hypomethylated (1/175)	Underexpressed (1/2)	0.36
LIMCH1	*LIM and Calponin Homology Domains 1*	Hypomethylated (1/51)	Overexpressed (1/1)	-0.51
PDP2	*Pyruvate Dehyrogenase Phosphatase Catalytic Subunit 2*	Hypomethylated (1/13)	Underexpressed (1/2)	N.S.
PDXK	*Pyridoxal Kinase*	Hypomethylated (1/37)	Underexpressed (1/1)	N.S.
PEA15	*Proliferation and Apoptosis Adaptor Protein 15*	Hypomethylated (1/12)	Overexpressed (1/1)	N.S.
PLEKHA2	*Pleckstrin Homology Domain Containing A2*	Hypermethylated (1/22)	Overexpressed (1/4)	N.S.
RALGPS1	*Ral GEF With PH Domain and SH3 Binding Motif 1*	Hypomethylated (1/20)	Underexpressed (1/1)	N.S.
RRAD	*RRAD, Ras Related Glycolysis Inhibitor and Calcium Channel Regulator*	Hypomethylated (1/13)	Overexpressed (1/2)	N.S.
SFRS8	*Splicing Factor SWAP*	Hypomethylated (1/77)	Underexpressed (1/1)	N.S.
UCKL1	*Uridine-Cytidine Kinase 1 Like 1*	Hypomethylated (1/18)	Underexpressed (1/1)	N.S.
USP5	*Ubiquitin Specific Peptidase 5*	Hypomethylated (1/23)	Underexpressed (1/1)	N.S.

Pearson's correlation coefficient r is given for genes with a significant correlation between methylation and expression. N.S., Not significant.

allowing a quick and easy grasp and visualization of the global role of the clusters (**Figure 4**).

Predicting Phenotypic Traits From Omics Data

The 1,651 features were used as input data to predict the outcome for four phenotypic traits (IUGR, premature birth, birth weight, and head circumference), in order to measure the importance of each gene in sub-phenotypic prediction. Class-control classification showed perfect predictions on the test set, with clearly distinct predicted probabilities between control and IUGR samples (**Figure 5**). This large gap of probabilities between IUGR and control samples confirmed the robustness of the model. These results were expected, as only features showing a significant difference between IUGR and control groups were selected for training the model. Furthermore, the previous unsupervised analysis (**Figure 2**) confirmed a clear distinction between IUGR and control samples.

FIGURE 4 | Word clouds summarizing the most frequent and specific terms among the 24 gene clusters **(A–X)**.

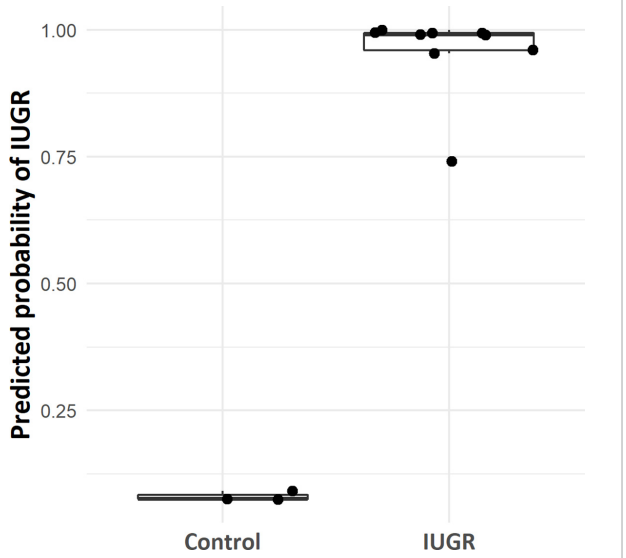

FIGURE 5 | Box plot of case-control model predicted probability according to IUGR/control group.

Premature birth, birth weight, and head circumference scores predicted on test samples were linearly correlated with actual values ($p < 0.01$) (**Figure 6**).

A network was created to represent all omics features with at least 10% importance for predicting at least one phenotypic trait (**Figure 7**). Among the nine genes with high importance (> 80%) in the prediction of at least one phenotypic trait, five (NMD3, ORC6L, MAPK8, PDCL, PLP1), in the center of the network share an importance in predicting most studied phenotypic traits.

The full list of methylomics and transcriptomics features with importance higher than 50% for phenotypic prediction is available in the **Supplementary Table 3**.

DISCUSSION

Text Annotation Clustering and Word Cloud Visualization

In most high-throughput gene studies, functional annotation analysis is a powerful tool, allowing the highlighting of pathways enriched in a particular pathophysiological context. However,

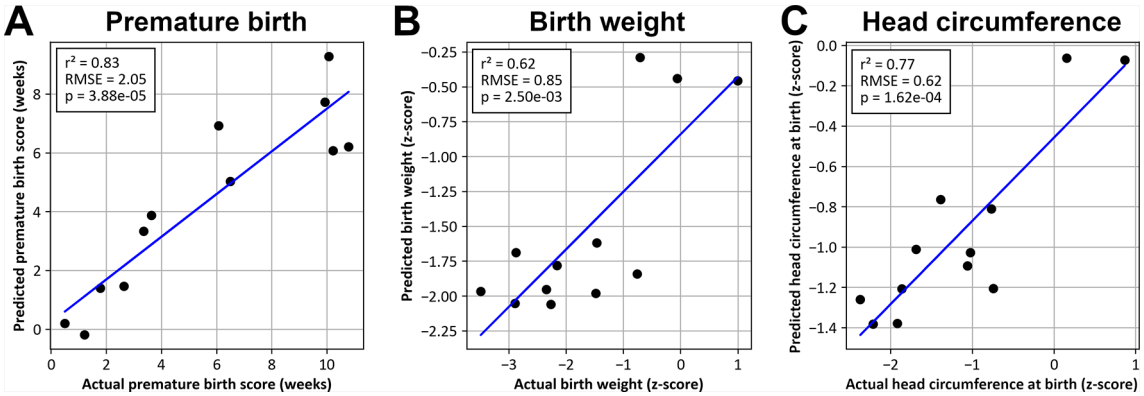

FIGURE 6 | Values predicted by SVM models as a function of actual values for premature birth **(A)**, birth weight **(B)**, and head circumference at birth **(C)**.

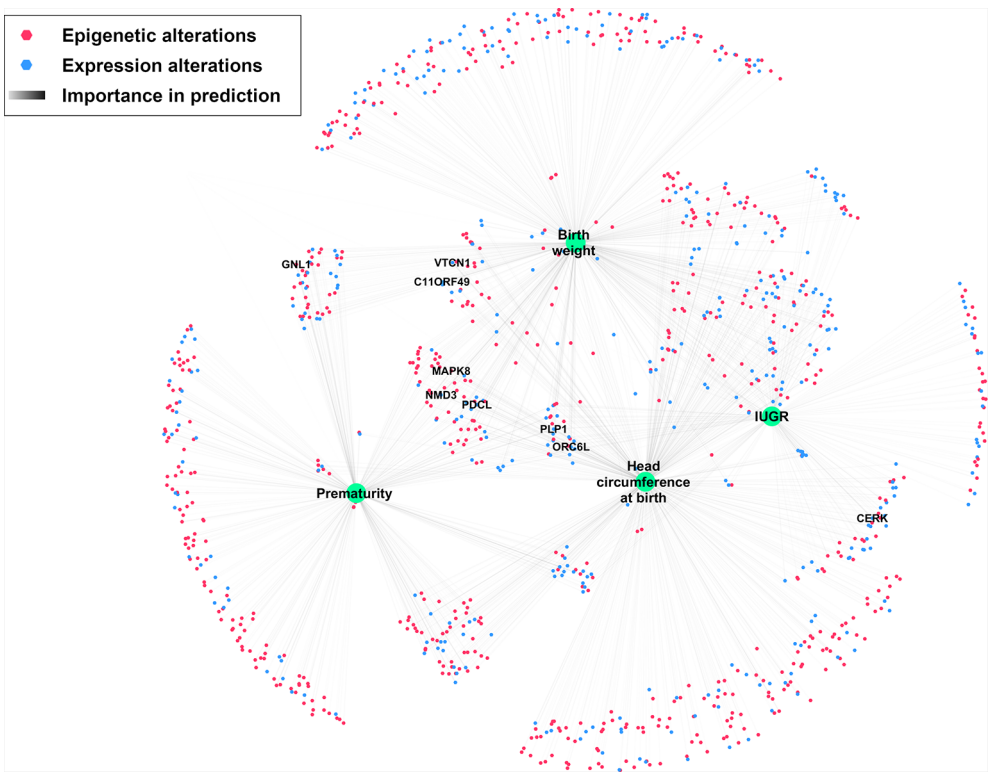

FIGURE 7 | Network depicting significantly altered features and their importance in predicting IUGR phenotype. Nodes were positioned according to an Edge-weighted Spring Embedded Layout, based on feature importance for predicting each phenotypic trait. Only genes with at least 80% importance for predicting at least one phenotypic trait are labeled.

limiting gene annotation to categorical roles or pathways leads to a significant loss of knowledge in comparison with data available in literature.

Word clouds allowed a visual description of the main biological processes and pathways involved in the IUGR pathophysiology, in order to speed up and deepen the bibliographic work on genes significantly altered in IUGR.

Cell Signaling and Proliferation

Many terms among the most frequent and specific refer to proto-oncogenes and cell proliferation and signaling and development mechanisms. This is confirmed by several genes isolated from both methylome and transcriptome (overexpression of *BCL6*, *CALM1*, *DLX5*, *PEA15*, *RRAD*, and underexpression of *FOXK1* and *UCKL1*).

DNA, RNA Regulation, Transcription, Translation

Many gene clusters (C, D, F, G, I, L, respectively 5.4%, 5.5%, 2.1%, 1.8%, 2.0% and 2.9% of genes) refer to DNA methylation and repair, regulation of transcription, and RNA splicing and translation. Epigenetic and gene expression alterations in IUGR have been evidenced here as well as in literature (Kawai et al., 2015).

Mitochondria and Oxidative Stress

Clusters H and T (2.0% and 1.8%, respectively) refer to mitochondria alterations, cell death and apoptosis, and redox reactions. Indeed, pregnancy increases ROS production and oxidative stress, causing damage to mitochondria and potentially leading to cell death, especially during pathological pregnancies like PE or IUGR (Myatt and Cui, 2004). These phenomena may have a role in the fetal programming of atherosclerosis (Leduc et al., 2010). *ALKBH5* (found hypomethylated, overexpressed) encodes a hypoxia-inducible factor playing a role in cell proliferation (Zhang et al., 2016).

Intra- and Extra-Cellular Matrix

Several clusters (E, F, W, respectively 2.3%, 2.1% and 5.0%) suggest primarily cytoskeleton and cell-cell junction alterations. Furthermore, cluster N (1.0%) refers to intra-cellular trafficking and cell mechanisms relying heavily on the cytoskeleton. Riquelme and her colleagues (Riquelme et al., 2011) have already evidenced abnormalities in the lipid raft composition of the microvillous membrane of the placental syncytiotrophoblast, linked with alterations in the expression of several cytoskeletal proteins (actin, ezrin, and cytokeratin-7) in placentas from pathological pregnancies (PE and IUGR). They suggest that these cytoskeleton alterations might be responsible for alterations in the syncytiotrophoblast microvilli, which may play a major role in the IUGR pathophysiology. Among the genes found altered in both methylome and transcriptome, *AFAP1* is a major regulator of the cytoskeleton structure (Xiao et al., 2012). *FLNB* codes for an actin-binding protein crosslinking actin filaments and playing various roles including cell proliferation and angiogenesis through mechanotransduction (Xu et al., 2017). Clusters P and X (2.0% and 6.9%, respectively) refer to extracellular matrix alterations. Such alterations have already been evidenced in IUGR (Merchant et al., 2004; Swierczewski et al., 2012).

Protein Degradation and Turnover

Cluster S (2.6%) refers to protein SUMOylation, ubiquitination, and degradation. It has been evidenced that protein ubiquitination is altered in IUGR and PE, particularly due to a modulation by oxidative stress, with an increased degradation of p53 and Mcl-1 proteins, contributing to the pathological mechanisms of the diseases (Rolfo et al., 2012). *WSB1* (underexpressed here) mediates ubiquitination and proteolytic degradation, and is also involved in cell and glucose metabolism, playing a role in hypoxia-related mechanisms (Haque et al., 2016). *USP5* (underexpressed here) codes for a deubiquitinating enzyme which has also been shown to play a role in cell cycle modulation.

Heart and Skeletal Muscle Development

Heart and skeletal muscles are referred to in cluster P (2.0%). Wang et al. (Wang et al., 2013) and Yates et al. (Yates et al., 2012) already reported that hypoxemia and hypoglycaemia undergone during IUGR decrease muscle mass in offspring. *DGKZ* (found hypomethylated, overexpressed) is known to induce muscle fiber hypertrophy and plays a role in the adaptation to energy metabolism alterations (Benziane et al., 2017). *FOXK1* induces muscle progenitor cell proliferation and inhibits their differentiation (Shi et al., 2012). *FOXK1* was found here both hypomethylated and underexpressed. This underexpression might be due to another role of *FOXK1* in repressing starvation-induced atrophy and autophagy (Bowman et al., 2014).

Energy Metabolism and Insulin Resistance

Major references are made to fat and lipid metabolism in cluster O (5.4%) and cluster Q (1.2%). These clusters support the hypothesis of an alteration of lipid and fat metabolism during IUGR, reflecting mechanisms of insulin resistance. Several genes found altered in both methylome and transcriptome support this pathway. Among these genes, *PAPPA-2* is the gene with the largest number of methylation sites significantly altered (hypomethylation), and with the largest number of transcripts significantly differently expressed (overexpression) in IUGR placentas. Its overexpression has already been reported in both maternal blood and the placenta in IUGR (Whitehead et al., 2013) and PE (Kramer et al., 2016). *PAPPA-2* encodes a protein cleaving the insulin-like growth factor 1 (IGF-1) from a ternary complex with IGF binding proteins (IGFBP-3) (Fujimoto et al., 2017). Via this regulation of the IGF-1 bioavailability, it plays a key role in both placenta development and fetal growth. Both low and high levels of IGF-1 have also been associated with insulin resistance (Friedrich et al., 2012). Interestingly, the *STC2* gene, encoding the PAPPA2 inhibitor stanniocalcin-2, was found significantly hypomethylated here, but its expression was not significantly altered between IUGR and control groups.

PEA15 encodes a phosphoprotein responsible for insulin resistance and diabetes. Higher levels of expression of *PEA15* have been reported in both patients with diabetes mellitus type 2 (Condorelli et al., 1998) and in euglycemic patients with impaired insulin sensitivity (Valentino et al., 2006). The *DGKZ* gene, already discussed above, has been proven to play a role in the protection against peripheral insulin resistance and in improving overall energy metabolism (Benziane et al., 2017). *SLC2A1*, also known as glucose transporter 1 (*GLUT1*), is the major glucose transporter in the human placenta and the rate-limiting step of glucose transport from the placenta to the fetus (Illsley, 2000). Its overexpression here might reflect mechanisms of adaptation to fetal nutrient restriction. *C1QTNF1*, also known as glucose-dependent insulinotropic polypeptide (GIP) is an adipokine, whose secretion by adipocytes is increased under hypoxia, partially under the control of HIF-1α. It stimulates proinflammatory gene expression and impairs insulin sensitivity of adipocytes (Chen et al., 2015). However, *C1QTNF1* was found underexpressed in this study.

Two more genes supporting these mechanisms of insulin resistance were found here among the most overexpressed genes:

HTRA4 (IGF binding domain containing protein, fold-change = 7.33) and *LEP* (leptin, fold-change = 4.89). This major overexpression had already been observed in both IUGR (Madeleneau et al., 2015) and PE (Brew et al., 2016).

Sub-Phenotype Prediction

Unsupervised clustering (**Figure 2**) showed a clear distinction between IUGR and controls and suggested the existence of multiple sub-phenotypes in the IUGR group (**Figure 3**).

As expected, SVM models were able to accurately predict such phenotypic traits: gestational age at birth, birth weight, and head circumference, using only a small subset of the whole data, i.e. 1,651 (0.3%) methylome and transcriptome variables. These results confirmed the high predictive value of the genes highlighted in this study in the IUGR, as well as in several variables of severity and pathophysiology of the IUGR.

In particular, nine genes with high importance in the prediction of these phenotypic traits were observed. Network visualization (**Figure 7**) showed that most of these genes are correlated with most clinically relevant traits studied here.

Among these genes, *CERK*, *GNL1*, *PLP1*, and *MAPK8* are known to be altered or play a direct role in the pathophysiology of IUGR or PE in various pathways discussed above: differentiation and proliferation regulation, response to hypoxia and oxidative stress, and neurological maturation (Vaiman et al., 2011; Reid et al., 2012; Goyal et al., 2013; Chan et al., 2019). For the other genes (*VTCN1*, *C11ORF49*, *PDCL*, *ORC6L*, *NMD3*), no obvious link with IUGR was found in literature, creating a topic for future studies regarding their exact role in the IUGR pathophysiology.

Limits

Our study was mainly limited by the imbalance between cases and controls and the relatively weak number of controls. However, as already stated, controls show a significantly lower variance for most phenotypic traits discussed in this study. Furthermore, oversampling methods were used in order to compensate this limit and prevent model overfitting, while assessing the importance of genes on unmodified test sets which were not previously used for training models.

Conclusion

Many epigenetic and gene expression alterations in IUGR placentas have been observed here, some of them confirming previous mechanisms already published, and others being new findings. Several major pathways were highlighted by annotation text-mining analysis: cell cycle and proliferation, regulation of apoptosis, epigenetic modifications, transcription, translation, oxidative stress and hypoxia, cytoskeleton and cell structure, protein degradation and turnover, autophagy, muscle development, and glucose and lipid energy metabolism. The involvement of these pathways was supported by significant differences in both methylome and transcriptome. Finally, several key genes with high correlation with phenotypic traits clinically relevant for IUGR were observed and may constitute potential targets for future study.

DATA AVAILABILITY STATEMENT

Array-based datasets for both genome methylation and expression have been deposited at the European Genome-phenome Archive (EGA), which is hosted by the EBI and the CRG, under accession number EGAS00001003467. Further information about EGA can be found on https://ega-archive.org (The European Genome-phenome Archive of human data consented for biomedical research, http://www.nature.com/ng/journal/v47/n7/full/ng.3312.html). Analysis output files are available in **Supplementary Material** (**Supplementary Tables 1–3**).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics committee of Angers. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

FC: literature search, data analysis, data interpretation, figures, writing. NH: literature search, data collection, data analysis. XD: data interpretation. GR: data interpretation. GB proofreading, expertise in data analysis methods. JB: expertise in data analysis methods. VP: proofreading. GLen: proofreading. OB: data collection. GLeg: data collection. DM-P: proofreading. MC: proofreading, expertise in data analysis methods. PG: data analysis, data interpretation, expertise in data analysis methods. PR: co-director of the study, data interpretation, expertise, writing. GG: director of the study, literature search, data collection, data interpretation, expertise, writing.

FUNDING

This study was funded by a grant from the Angers University Hospital, France.

ACKNOWLEDGMENTS

We acknowledge support from the Institut National de la Santé et de la Recherche Médicale (INSERM), the Centre National de la Recherche Scientifique (CNRS), the University of Angers, and Angers University Hospital.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01292/full#supplementary-material

REFERENCES

Aggarwal, C. C., and Zhai, C. (2012). "A Survey of Text Clustering Algorithms," in *Mining Text Data*. Eds. C. C. Aggarwal and C. Zhai (Boston, MA: Springer US), 77–128. doi: 10.1007/978-1-4614-3223-4_4

Altmann, A., Toloşi, L., Sander, O., and Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics* 26, 1340–1347. doi: 10.1093/bioinformatics/btq134

American College of Obstetricians and Gynecologists (2013). ACOG Practice bulletin no. 134: fetal growth restriction. *Obstet. Gynecol.* 121, 1122–1133. doi: 10.1097/01.AOG.0000429658.85846.f9

Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., et al. (2014). Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* 30, 1363–1369. doi: 10.1093/bioinformatics/btu049

Ben-Hur, A., Ong, C. S., Sonnenburg, S., Schölkopf, B., and Rätsch, G. (2008). Support Vector Machines and Kernels for Computational Biology. *PloS Comput. Biol.* 4, e1000173. doi: 10.1371/journal.pcbi.1000173

Benziane, B., Borg, M. L., Tom, R. Z., Riedl, I., Massart, J., Björnholm, M., et al. (2017). DGKζ deficiency protects against peripheral insulin resistance and improves energy metabolism. *J. Lipid Res.* 58, 2324–2333. doi: 10.1194/jlr.M079723

Bowman, C. J., Ayer, D. E., and Dynlacht, B. D. (2014). Foxk proteins repress the initiation of starvation-induced atrophy and autophagy programs. *Nat. Cell Biol.* 16, 1202–1214. doi: 10.1038/ncb3062

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

Brew, O., Sullivan, M. H. F., and Woodman, A. (2016). Comparison of normal and pre-eclamptic placental gene expression: a systematic review with meta-analysis. *PloS One* 11, e0161504. doi: 10.1371/journal.pone.0161504

Buffat, C., Mondon, F., Rigourd, V., Boubred, F., Bessières, B., Fayol, L., et al. (2007). A hierarchical analysis of transcriptome alterations in intrauterine growth restriction (IUGR) reveals common pathophysiological pathways in mammals. *J. Pathol.* 213, 337–346. doi: 10.1002/path.2233

Chan, K. J., Swan, K. F., Narayanappa, A., Mersereau, C., and Intapad, S. (2019). Imbalance of Sphingolipids synthesis/degradation pathway in preeclamptic mouse placenta and kidney of intrauterine growth restricted mouse fetus. *FASEB J.* 33, 593.3–593.3. doi: 10.1096/fasebj.2019.33.1_supplement.593.3

Chawla, N. V., 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

Chen, S., Okahara, F., Osaki, N., and Shimotoyodome, A. (2015). Increased GIP signaling induces adipose inflammation *via a* HIF-1α-dependent pathway and impairs insulin sensitivity in mice. *Am. J. Physiol. Endocrinol. Metab.* 308, E414–E425. doi: 10.1152/ajpendo.00418.2014

Cohen, E., Baerts, W., and Bel, F. V. (2015). Brain-sparing in intrauterine growth restriction: considerations for the neonatologist. *Neonatology* 108, 269–276. doi: 10.1159/000438451

Condorelli, G., Vigliotta, G., Iavarone, C., Caruso, M., Tocchetti, C. G., Andreozzi, F., et al. (1998). PED/PEA-15 gene controls glucose transport and is overexpressed in type 2 diabetes mellitus. *EMBO J.* 17, 3858–3866. doi: 10.1093/emboj/17.14.3858

Ding, Y., and Cui, H. (2017). Integrated analysis of genome-wide DNA methylation and gene expression data provide a regulatory network in intrauterine growth restriction. *Life Sci.* 179, 60–65. doi: 10.1016/j.lfs.2017.04.020

Evangelopoulos, N. E. (2013). Latent semantic analysis. *Wiley Interdiscip. Rev. Cogn. Sci.* 4, 683–692. doi: 10.1002/wcs.1254

Fantini, D. (2019). *easyPubMed: Search and Retrieve Scientific Publication Records from PubMed*. Available at: https://CRAN.R-project.org/package=easyPubMed [Accessed October 10, 2019].

Fellows, I. (2014). wordcloud: Word Clouds. Available at: https://CRAN.R-project.org/package=wordcloud.

Flamant, C., and Gascoin, G. (2013). Short-term outcome and small for gestational age newborn management. *J. Gynecol. Obstet. Biol. Reprod. (Paris)* 42, 985–995. doi: 10.1016/j.jgyn.2013.09.020

Fortin, J.-P., Labbe, A., Lemire, M., Zanke, B. W., Hudson, T. J., Fertig, E. J., et al. (2014). Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol.* 15 (12), 503. doi: 10.1186/s13059-014-0503-2

Friedrich, N., Thuesen, B., Jørgensen, T., Juul, A., Spielhagen, C., Wallaschofksi, H., et al. (2012). The Association Between IGF-I and Insulin Resistance. *Diabetes Care* 35, 768–773. doi: 10.2337/dc11-1833

Fujimoto, M., Hwa, V., and Dauber, A. (2017). Novel Modulators of the Growth Hormone - Insulin-Like Growth Factor Axis: Pregnancy-Associated Plasma Protein-A2 and Stanniocalcin-2. *J. Clin. Res. Pediatr. Endocrinol.* 9, 1–8. doi: 10.4274/jcrpe.2017.S001

Gascoin, G., and Flamant, C. (2013). Long-term outcome in context of intra uterine growth restriction and/or small for gestational age newborns. *J. Gynecol. Obstet. Biol. Reprod. (Paris)* 42, 911–920. doi: 10.1016/j.jgyn.2013.09.014

Gascoin-Lachambre, G., Buffat, C., Rebourcet, R., Chelbi, S. T., Rigourd, V., Mondon, F., et al. (2010). Cullins in human intra-uterine growth restriction: expressional and epigenetic alterations. *Placenta* 31, 151–157. doi: 10.1016/j.placenta.2009.11.008

Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., et al. (2014). Data integration in the era of omics: current and future challenges. *BMC Syst. Biol.* 8, I1. doi: 10.1186/1752-0509-8-S2-I1

Goyal, R., Van Wickle, J., Goyal, D., Matei, N., and Longo, L. D. (2013). Antenatal maternal long-term hypoxia: acclimatization responses with altered gene expression in ovine fetal carotid arteries. *PloS One* 8, e82200. doi: 10.1371/journal.pone.0082200

Guyon, I., and Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.* 3, 1157–1182.

Haque, M., Kendal, J. K., MacIsaac, R. M., and Demetrick, D. J. (2016). WSB1: from homeostasis to hypoxia. *J. Biomed. Sci.* 23, 61. doi: 10.1186/s12929-016-0270-3

Haury, A.-C., Gestraud, P., and Vert, J.-P. (2011). The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PloS One* 6, e28210. doi: 10.1371/journal.pone.0028210

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 37, 1–13. doi: 10.1093/nar/gkn923

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211

Illsley, N. P. (2000). Glucose transporters in the human placenta. *Placenta* 21, 14–22. doi: 10.1053/plac.1999.0448

Kaplan., C. G. (2007). Fetal and maternal vascular lesions. *Semin. Diagn. Pathol.* 24, 14–22. doi: 10.1053/j.semdp.2007.02.005

Kawai, T., Yamada, T., Abe, K., Okamura, K., Kamura, H., Akaishi, R., et al. (2015). Increased epigenetic alterations at the promoters of transcriptional regulators following inadequate maternal gestational weight gain. *Sci. Rep.* 5, 14224. doi: 10.1038/srep14224

Keogh, E., and Mueen, A. (2010). "Curse of Dimensionality," in *Encyclopedia of Machine Learning*. Eds. C. Sammut and G. I. Webb (Boston, MA: Springer US), 257–258. doi: 10.1007/978-0-387-30164-8_192

Kramer, A. W., Lamale-Smith, L. M., and Winn, V. D. (2016). Differential expression of human placental PAPP-A2 over gestation and in preeclampsia. *Placenta* 37, 19–25. doi: 10.1016/j.placenta.2015.11.004

Leduc, L., Levy, E., Bouity-Voubou, M., and Delvin, E. (2010). Fetal programming of atherosclerosis: possible role of the mitochondria. *Eur. J. Obstet. Gynecol. Reprod. Biol.* 149, 127–130. doi: 10.1016/j.ejogrb.2009.12.005

Madeleneau, D., Buffat, C., Mondon, F., Grimault, H., Rigourd, V., Tsatsaris, V., et al. (2015). Transcriptomic analysis of human placenta in intrauterine growth restriction. *Pediatr. Res.* 77, 799–807. doi: 10.1038/pr.2015.40

Malhotra, A., Allison, B. J., Castillo-Melendez, M., Jenkin, G., Polglase, G. R., and Miller, S. L. (2019). Neonatal morbidities of fetal growth restriction: pathophysiology and impact. *Front. Endocrinol.* 10, 55. doi: 10.3389/fendo.2019.00055

Merchant, S. J., Crocker, I. P., Baker, P. N., Tansinda, D., Davidge, S. T., and Guilbert, L. J. (2004). Matrix metalloproteinase release from placental explants of pregnancies complicated by intrauterine growth restriction. *J. Soc Gynecol. Investig.* 11, 97–103. doi: 10.1016/j.jsgi.2003.08.005

Murki, S. (2014). Intrauterine growth retardation - a review article. *J. Neonatal Biol.* 3, 1–13. doi: 10.4172/2167-0897.1000135

Myatt, L., and Cui, X. (2004). Oxidative stress in the placenta. *Histochem. Cell Biol.* 122, 369–382. doi: 10.1007/s00418-004-0677-x

Olsen, I. E., Groveman, S. A., Lawson, M. L., Clark, R. H., and Zemel, B. S. (2010). New intrauterine growth curves based on United States data. *PEDIATRICS* 125, e214–e224. doi: 10.1542/peds.2009-0913

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830. doi: 10.1007/978-1-4842-0958-5_8

Quinn, J.-A., Munoz, F. M., Gonik, B., Frau, L., Cutland, C., Mallett-Moore, T., et al. (2016). Preterm birth: case definition & guidelines for data collection, analysis, and presentation of immunisation safety data. *Vaccine* 34, 6047–6056. doi: 10.1016/j.vaccine.2016.03.045

R Development Core Team. (2008). "R: A language and environment for statistical computing," in *R Foundation for Statistical Computing*(Vienna, Austria). Available at: http://www.R-project.org.

Reid, M. V., Murray, K. A., Marsh, E. D., Golden, J. A., Simmons, R. A., and Grinspan, J. B. (2012). Delayed myelination in an intrauterine growth retardation model is mediated by oxidative stress upregulating bone morphogenetic protein 4. *J. Neuropathol. Exp. Neurol.* 71, 640–653. doi: 10.1097/NEN.0b013e31825cfa81

Riquelme, G., Vallejos, C., de Gregorio, N., Morales, B., Godoy, V., Berrios, M., et al. (2011). Lipid rafts and cytoskeletal proteins in placental microvilli membranes from preeclamptic and IUGR pregnancies. *J. Membr. Biol.* 241, 127. doi: 10.1007/s00232-011-9369-3

Rolfo, A., Garcia, J., Todros, T., Post, M., and Caniggia, I. (2012). The double life of MULE in preeclamptic and IUGR placentae. *Cell Death Dis.* 3, e305. doi: 10.1038/cddis.2012.44

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

Sharma, D., Shastri, S., and Sharma, P. (2016). Intrauterine Growth Restriction: Antenatal and Postnatal Aspects. *Clin. Med. Insights Pediatr.* 10, 67–83. doi: 10.4137/CMPed.S40070

Shi, X., Wallis, A. M., Gerard, R. D., Voelker, K. A., Grange, R. W., DePinho, R. A., et al. (2012). Foxk1 promotes cell proliferation and represses myogenic differentiation by regulating Foxo4 and Mef2. *J. Cell Sci.* 125, 5329–5337. doi: 10.1242/jcs.105239

Swierczewski, A., Kobos, J., Pasiński, J., Kowalska-Koprek, U., and Karowicz-Bilińska, A. (2012). Expression of metalloproteinase MMP-9 and tissue inhibitor of metalloproteinase TIMP-2 in placenta of pregnant women with intrauterine growth restriction. *Ginekol. Pol.* 83, 439–445.

Vaiman, D., Gascoin-Lachambre, G., Boubred, F., Mondon, F., Feuerstein, J.-M., Ligi, I., et al. (2011). The intensity of IUGR-induced transcriptome deregulations is inversely correlated with the onset of organ function in a rat model. *PloS One* 6, e21222. doi: 10.1371/journal.pone.0021222

Valentino, R., Lupoli, G. A., Raciti, G. A., Oriente, F., Farinaro, E., Della Valle, E., et al. (2006). The PEA15 gene is overexpressed and related to insulin resistance in healthy first-degree relatives of patients with type 2 diabetes. *Diabetologia* 49, 3058–3066. doi: 10.1007/s00125-006-0455-5

van Helden, P. (2013). Data-driven hypotheses. *EMBO Rep.* 14, 104. doi: 10.1038/embor.2012.207

van Rossum, G. (1995). Python tutorial, Technical Report CS-R9526.

Vanitha, C. D. A., Devaraj, D., and Venkatesulu, M. (2015). Gene expression data classification using support vector machine and mutual information-based gene selection. *Proc. Comput. Sci.* 47, 13–21. doi: 10.1016/j.procs.2015.03.178

Vayssière, C., Sentilhes, L., Ego, A., Bernard, C., Cambourieu, D., Flamant, C., et al. (2015). Fetal growth restriction and intra-uterine growth restriction: guidelines for clinical practice from the French College of Gynaecologists and Obstetricians. *Eur. J. Obstet. Gynecol. Reprod. Biol.* 193, 10–18. doi: 10.1016/j.ejogrb.2015.06.021

Wang, T., Liu, C., Feng, C., Wang, X., Lin, G., Zhu, Y., et al. (2013). IUGR alters muscle fiber development and proteome in fetal pigs. *Front. Biosci. Landmark Ed.* 18, 598–607. doi: 10.2741/4123

Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Liaw, W. H. A., Lumley, T., et al. (2016). *gplots: Various R Programming Tools for Plotting Data*. Available at: https://CRAN.R-project.org/package=gplots [Accessed October 15, 2018].

Whitehead, C. L., Walker, S. P., Ye, L., Mendis, S., Kaitu'u-Lino, T. J., Lappas, M., et al. (2013). Placental specific mRNA in the maternal circulation are globally dysregulated in pregnancies complicated by fetal growth restriction. *J. Clin. Endocrinol. Metab.* 98, E429–E436. doi: 10.1210/jc.2012-2468

Willett, P. (2006). The Porter stemming algorithm: then and now. *Program* 40, 219–223. doi: 10.1108/00330330610681295

Xiao, H., Han, B., Lodyga, M., Bai, X.-H., Wang, Y., and Liu, M. (2012). The actin-binding domain of actin filament-associated protein (AFAP) is involved in the regulation of cytoskeletal structure. *Cell. Mol. Life Sci. CMLS* 69, 1137–1151. doi: 10.1007/s00018-011-0812-5

Xiao, X., Zhao, Y., Jin, R., Chen, J., Wang, X., Baccarelli, A., et al. (2016). Fetal growth restriction and methylation of growth-related genes in the placenta. *Epigenomics* 8, 33–42. doi: 10.2217/epi.15.101

Xu, Q., Wu, N., Cui, L., Wu, Z., and Qiu, G. (2017). Filamin B: The next hotspot in skeletal research? *J. Genet. Genomics Yi Chuan Xue Bao* 44, 335–342. doi: 10.1016/j.jgg.2017.04.007

Yates, D. T., Macko, A. R., Nearing, M., Chen, X., Rhoads, R. P., and Limesand, S. W. (2012). Developmental programming in response to intrauterine growth restriction impairs myoblast function and skeletal muscle metabolism. *J. Pregnancy* 2012, 631038. doi: 10.1155/2012/631038

Zhang, S., Regnault, T. R. H., Barker, P. L., Botting, K. J., McMillen, I. C., McMillan, C. M., et al. (2015). Placental adaptations in growth restriction. *Nutrients* 7, 360–389. doi: 10.3390/nu7010360

Zhang, C., Zhi, W. I., Lu, H., Samanta, D., Chen, I., Gabrielson, E., et al. (2016). Hypoxia-inducible factors regulate pluripotency factor expression by ZNF217- and ALKBH5-mediated modulation of RNA methylation in breast cancer cells. *Oncotarget* 7, 64527–64542. doi: 10.18632/oncotarget.11743

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: 10 January 2020
doi: 10.3389/fgene.2019.01306

Identification of Platform-Independent Diagnostic Biomarker Panel for Hepatocellular Carcinoma Using Large-Scale Transcriptomics Data

Harpreet Kaur[1,2], *Anjali Dhall*[2], *Rajesh Kumar*[1,2] *and Gajendra P. S. Raghava*[2*]

[1] Bioinformatics Center, CSIR-Institute of Microbial Technology, Chandigarh, India, [2] Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India

OPEN ACCESS

Edited by:
Mehdi Pirooznia, National Heart, Lung, and Blood Institute (NHLBI), United States

Reviewed by:
Shi Ming, Sun Yat-sen University Cancer Center (SYSUCC), China
Yun Hak Kim, Pusan National University, South Korea

***Correspondence:**
Gajendra P. S. Raghava
raghava@iiitd.ac.in

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

Received: *04 September 2019*
Accepted: *26 November 2019*
Published: *10 January 2020*

Citation:
Kaur H, Dhall A, Kumar R and Raghava GPS (2020) Identification of Platform-Independent Diagnostic Biomarker Panel for Hepatocellular Carcinoma Using Large-Scale Transcriptomics Data. Front. Genet. 10:1306. doi: 10.3389/fgene.2019.01306

The high mortality rate of hepatocellular carcinoma (HCC) is primarily due to its late diagnosis. In the past, numerous attempts have been made to design genetic biomarkers for the identification of HCC; unfortunately, most of the studies are based on small datasets obtained from a specific platform or lack reasonable validation performance on the external datasets. In order to identify a universal expression-based diagnostic biomarker panel for HCC that can be applicable across multiple platforms, we have employed large-scale transcriptomic profiling datasets containing a total of 2,316 HCC and 1,665 non-tumorous tissue samples. These samples were obtained from 30 studies generated by mainly four types of profiling techniques (Affymetrix, Illumina, Agilent, and High-throughput sequencing), which are implemented in a wide range of platforms. Firstly, we scrutinized overlapping 26 genes that are differentially expressed in numerous datasets. Subsequently, we identified a panel of three genes (*FCN3, CLEC1B*, and *PRC1)* as HCC biomarker using different feature selection techniques. Three-genes-based HCC biomarker identified HCC samples in training/validation datasets with an accuracy between 93 and 98%, Area Under Receiver Operating Characteristic curve (AUROC) in a range of 0.97 to 1.0. A reasonable performance, i.e., AUROC 0.91–0.96 achieved on validation dataset containing peripheral blood mononuclear cells, concurred their non-invasive utility. Furthermore, the prognostic potential of these genes was evaluated on TCGA-LIHC and GSE14520 cohorts using univariate survival analysis. This analysis revealed that these genes are prognostic indicators for various types of the survivals of HCC patients (e.g., Overall Survival, Progression-Free Survival, Disease-Free Survival). These genes significantly stratified high-risk and low-risk HCC patients (p-value <0.05). In conclusion, we identified a universal platform-independent three-genes-based biomarker that can predict HCC patients with high precision and also possess significant prognostic potential. Eventually, we developed a web server HCCpred based on the above study to facilitate scientific community (http://webs.iiitd.edu.in/raghava/hccpred/).

Keywords: liver cancer, hepatocellular carcinoma, biomarker, expression, diagnosis, survival, machine learning, classification

INTRODUCTION

Cancer is a heterogeneous disease driven by genomic and epigenomic changes within the cell (Sharma et al., 2010; Dawson and Kouzarides, 2012; Nagpal et al., 2015; Flavahan et al., 2017; Kamel and Al-Amodi, 2017; Chatterjee et al., 2018; Kagohara et al., 2018; Narrandes and Xu, 2018; Nebbioso et al., 2018; Kumar et al., 2019). Gene dysregulation is considered a hallmark of cancer. Among the 22 common cancer type, hepatocellular carcinoma (HCC) ranks at sixth in terms of frequency of occurrence and fourth at cancer-related mortality (Siegel et al., 2019). The etiology of HCC can be induced by multiple factors, especially hepatitis viral infection, alcoholic cirrhosis, and consumption of aflatoxin-contaminated foods (Ho et al., 2016). Although various traditional and locoregional treatment strategies such as hepatic resection (RES), percutaneous ethanol injection (PEI), radiofrequency ablation (RFA), microwave ablation (MWA), and trans-arterial chemotherapy infusion (TACI) have improved the survival rate, patients with HCC still have a late diagnosis and poor prognosis (Tian et al., 2018).

In the past, several studies focus on the identification of biomarkers by comparing the global gene expression changes between cancer tissue and non-tumorous tissues (Shirota et al., 2001; Jia et al., 2007; Marshall et al., 2013; Gao et al., 2015; Kang et al., 2015; Liu et al., 2015; Emma et al., 2016; Komatsu et al., 2016; Cai et al., 2017; Li et al., 2017; Zhang et al., 2017; Li et al., 2018b; Liao et al., 2018; Meng et al., 2018; Wang et al., 2018; Xu et al., 2018; Zheng et al., 2018; Cai et al., 2019; Jiao et al., 2019; Xia et al., 2019; Zhang et al., 2019). Such analyses yield hundreds or thousands of gene signature that are differentially expressed in cancer tissue compared to normal tissue, thus making it difficult to identify a universal subset of genes that play a crucial role in neoplastic transformation and progression (Rhodes et al., 2004). The lack of concordance of signature genes among different studies and extensive molecular variation between the patient's samples restrains the establishment of the robust biomarkers, promising targets and their experimental validation in clinical trials (Vasudevan et al., 2018). The transcriptome signatures have yet to be translated into a clinically useful biomarker, which may be due to a lack of their satisfactory validation performance on independent patient's cohort.

In this regard, treatment of HCC remains unsatisfying as only diagnostic and prognostic biomarkers alpha-fetoprotein (AFP) has been established so far. Several other biomarkers AFP-L3, osteopontin, and glypican-3 are currently being under investigation for the early diagnosis of HCC patients (Ocker, 2018). Advancement in the genomics has created rich public repositories of microarray and high throughput datasets from numerous studies such as The Cancer Genome Atlas (TCGA) (Cancer Genome Atlas Research Network et al., 2013), Genomic Data Common (GDC), and Gene Expression Omnibus (Grossman et al., 2016), (Barrett et al., 2013), which provide the opportunity to study the various aspects of cancer. Thus, novel methods exploring the computational approach by merging multiple datasets from different platforms could provide a new way to establish a robust and universal biomarker for disease diagnosis and prognosis with increased precision and reproducibility. Recently, this approach has been used for biomarker identification of pancreatic adenocarcinoma (PDAC) (Bhasin et al., 2016; Klett et al., 2018). However, various studies employed large-scale data or meta-analysis approaches to identify protein and miRNA expression-based biomarker for HCC diagnosis (Ji et al., 2016; Ding et al., 2017; Chen et al., 2018b; Ji et al., 2018). But, to the best of our knowledge, RNA-expression data are not explored in this regard for identification of the robust biomarker for HCC diagnosis and prognosis.

In order to overcome the limitations of existing methods, we made a systematic attempt to identify genetic biomarkers for HCC diagnosis that apply to a wide range of platforms and profiling techniques. One of the objectives of this study is to identify robust gene expression signatures for discrimination of HCC samples by the integration of multiple transcriptomic datasets from various platforms. Here, we have collected and analyzed a total of 3,981 samples from published datasets, out of which 2,316 and 1,665 are of HCC and normal or non-tumorous tissue samples, respectively. From this, we identified 26 genes, which are commonly differentially expressed in uniform patterns among most of the datasets, which provides a universally activated transcriptional signatures of HCC cancer type. Further, we have established a robust "three-genes-based HCC biomarker" implementing different machine learning techniques to distinguish HCC and non-tumorous samples with high precision. Additionally, the survival analysis of HCC patient's cohorts using these genes revealed their significant prognostic potential in the stratification of high-risk and low-risk patient's groups. To the best of our knowledge, this is the first study regarding HCC cancer type for the identification of universal platform-independent diagnostic biomarkers by integrating data from multiple platforms implementing machine learning approaches.

MATERIALS AND METHODS

Dataset Collection

Collection of Gene Expression Datasets of HCC

In this study, we extract raw expression data of 30 datasets, where 29 transcriptome datasets were obtained from GEO and one is from TCGA; each dataset contains at least 10 samples. The following is the list of datasets obtained from GEO: GSE102079 (Chiyonobu et al., 2018), GSE22405, GSE98383 (Diaz et al., 2018), GSE84402 (Wang et al., 2017), GSE64041 (Makowska et al., 2016), GSE69715 (Sekhar et al., 2018), GSE51401, GSE62232 (Schulze et al., 2015), GSE45267 (Chen et al., 2018a), GSE32879 (Oishi et al., 2012), GSE19665 (Deng et al., 2010), GSE107170 (Diaz et al., 2018), GSE76427 (Grinchuk et al.,

Abbreviations: AUROC, Area under the Receiver Operating Characteristic curve; ETREES, Extra Trees Classifier; SVC-RBF, Support Vector Machine with RBF kernel; TCGA, The Cancer Genome Atlas; KNN, K Neighbors Classifier; HCC, Hepatocellular Carcinoma; MCC, Matthew's correlation coefficient; LR, Logistic Regression; NB, Naive Bayes; RF, Random Forest; PBMCs, Peripheral Blood Mononuclear Cells

2018), GSE39791 (Kim et al., 2014), GSE57957 (Mah et al., 2014), GSE87630 (Woo et al., 2017), GSE46408, GSE57555 (Murakami et al., 2015), GSE54236 (Villa et al., 2016; Zubiete-Franco et al., 2019), GSE65484 (Dong et al., 2015), GSE31370 (Seok et al., 2012), GSE84598, GSE89377, GSE29721 (Stefanska et al., 2011), GSE14323 (Mas et al., 2009), GSE25097 (Lamb et al., 2011; Tung et al., 2011; Wong et al., 2016), GSE14520 (Roessler et al., 2010; Zhao et al., 2015), GSE36376 (Lim et al., 2013), GSE36076). All GEO datasets were obtained using GEOquery package of Bioconductor in R-3.5.3. The TCGA RNA-seq dataset of TCGA-LIHC was downloaded using gdc-client from the GDC data portal. All datasets were curated manually to remove all non-human samples and ensured that only human tissue samples remain in the dataset. Besides, Probe ID mapped to gene symbols extracted from respective platform file and incorporated in the dataset matrix for each dataset. It has been observed that two datasets, i.e., GSE102079 and GSE64041, have three types of samples (HCC, adjacent non-tumor, and normal healthy). Thus, we derived two datasets from GSE102079, called GSE102079_D1 (contains HCC and adjacent non-tumor samples) and GSE102079_D2 (contains HCC and healthy normal samples). Similarly, we derived GSE64041_D1 and GSE64041_D2 datasets from GSE64041. Finally, we derived 32 datasets from original 30 datasets as we derived four datasets corresponding to GSE102079 and GSE64041. Notably, we used one non-invasive dataset (GSE36076), which contains 20 blood samples of peripheral blood mononuclear cells (PBMCs) to evaluate our models.

Pre-Processing of Datasets

Each retrieved raw dataset (**Supplementary Data**) was subjected to a detailed curation process. We have pre-processed dataset matrix individually from each profiling technique for different platforms in a standardized manner. In case of Affymetrix datasets, raw data files were pre-processed with background correction; RMA values were calculated using the Oligo package (Carvalho and Irizarry, 2010). In case of Illumina datasets, raw files were processed using Limma and Lumi packages (Du et al., 2008; Ritchie et al., 2015) and finally log2 values calculated using in-house R scripts. Similarly, raw Agilent-1-color and Agilent-2-color files were pre-processed using Limma package individually, then A-values were generated, which were further transformed to log2 values. Eventually, the average of multiple probes computed that correspond to a single gene for each dataset individually employing in-house R scripts. TCGA-LIHC dataset contains FPKM values, which were further converted to log2 values. Entrez transcript IDs were mapped to the gene symbols using GENCODE v22.

Datasets for the Identification of Differentially Expressed Genes

We divide our datasets into two parts: i) datasets for features extraction and ii) datasets for the development of the prediction models. Twenty-seven out of 32 datasets were selected for identification of differentially expressed genes (DEGs); each dataset contains more than 10 samples (**Figure 1A**). These 27 datasets were derived from 25 original GEO datasets. Out of them, 20 datasets contain HCC v/s adjacent non-tumor samples and 7 datasets contain HCC v/s healthy samples. These datasets encompass a total of 1,199 HCC and 949 normal or adjacent non-tumor samples.

Training and Validation Datasets

In this study, the GSE25097 dataset was used as a training dataset to develop prediction models; it contains 268 HCC and 243 non-tumor samples (**Figure 1B**). The performance of these models was evaluated on the following three datasets: GSE14520, GSE36376, and TCGA-LIHC, and called them as external validation datasets. As shown in **Figure 1B**, each dataset has a minimum of 400 samples. The distribution of all cohorts used in the current study based on sample size is shown in **Figure 1C**. To validate the performance of models on the non-invasive specimen, we also evaluated the performance on the GSE36076 dataset. This dataset contains 20 blood samples of PBMCs; it contains 10 HCC and 10 healthy individuals. In order to reduce the cross-platform artifacts, we performed quantile normalization using the PreprocessCore library of Bioconductor (Grossman, et al., 2016) package, for each dataset as well as for each profiling technique. This approach is well-adapted in the literature (Huang and Qin, 2018; Klett et al., 2018; Pedersen et al., 2018). These datasets contain a total of 1,117 HCC and 716 adjacent non-tumor samples.

Identification of Differentially Expressed Genes

Each gene in 32 datasets was analyzed for differential expression using Student's *t*-test (Welch *t*-test and Wilcoxon *t*-test). It is implemented using in-house R scripts after the assignment of samples to the respective class, i.e., cancer or normal. These tests have been applied previously in different studies for the identification of DEGs (WELCH, 1947; Akaiwa et al., 1999; Carvalho and Irizarry, 2010; Aino et al., 2014; Schulze et al., 2015; Best et al., 2016; Bhasin et al., 2016; Bhalla et al., 2017; Cai et al., 2017; Bhalla et al., 2019; Cai et al., 2019; Kaur et al., 2019). Wilcoxon T-test is used for paired samples and Welch T-test is used for unpaired samples. Only those sets of genes chosen to define DEGs that are statistically differentially expressed between two classes of samples with Bonferroni adjusted p-value less than 0.01. In order to identify a set of differential expression signatures or "core DEGs of hepatocellular carcinoma," DEGs in all 27 datasets were compared. Finally, only those overlapping genes were considered as "core DEGs of hepatocellular carcinoma," which have significant differential expression in at least 80% of cohorts. A similar type of approach was previously implemented in various studies (Bhasin et al., 2016; Klett et al., 2018; Li et al., 2018a).

Identification of Robust Biomarkers for HCC Diagnosis

Ranking and Selection of Features

To reduce the number of genes from the selected set of signature, i.e., "the core genes of hepatocellular carcinoma," genes were

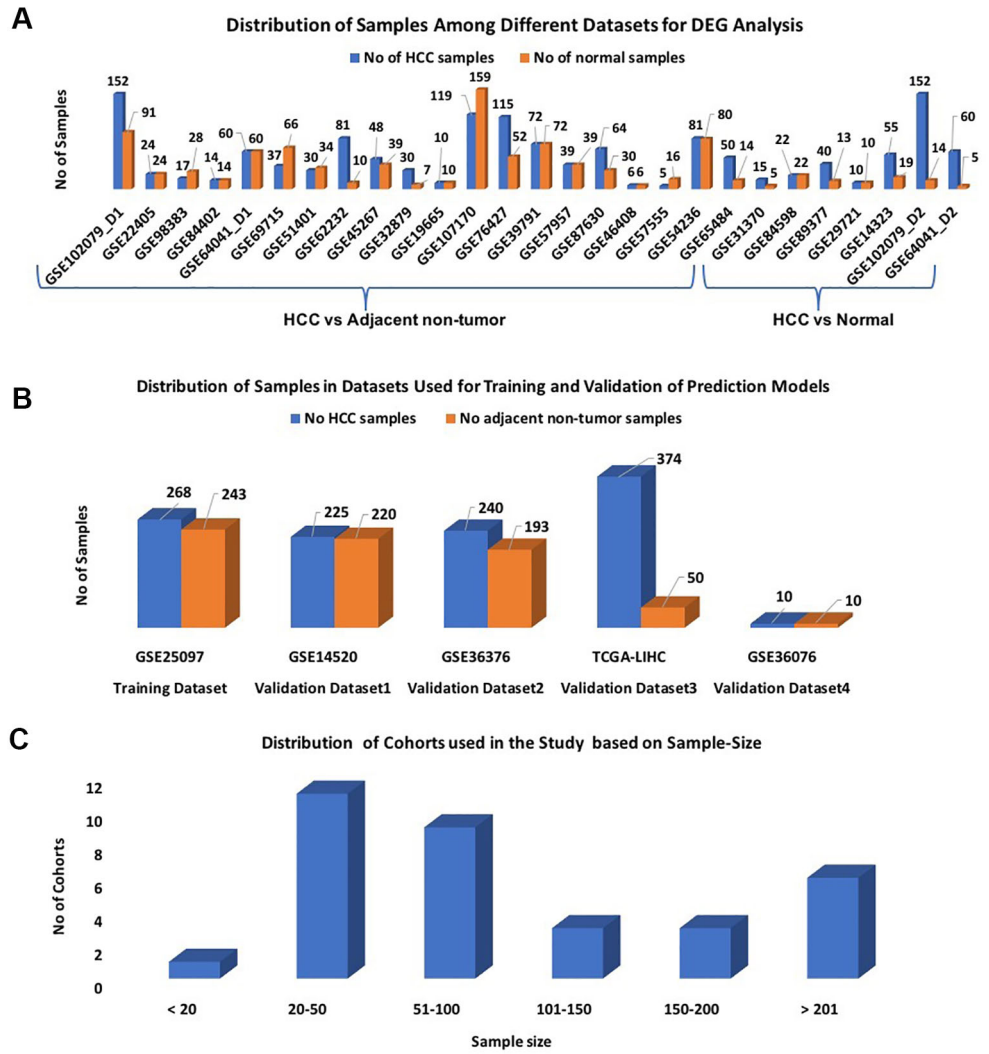

FIGURE 1 | Distribution of samples among datasets used in the study: **(A)** Datasets used for DEG analysis; **(B)** Datasets used for Development of Prediction models; **(C)** Sample-wise distribution of the datasets.

ranked on training dataset (GSE25097) using a simple threshold-based approach (Bhalla et al., 2017; Bhalla et al., 2019; Kaur et al., 2019). In the threshold-based approach, genes with a score above the threshold are assigned to cancer if it is found to be upregulated in cancer and otherwise normal; whereas sample is assigned to normal if the gene is downregulated in cancerous condition. We compute the performance of each gene based on a given threshold and identify the top 10 features having the highest performance. We further identified the top 5 features, which give the best performance when evaluated on the training dataset using a 10-fold cross-validation technique. Features were further reduced from five to four and then four to three using a wrapper-based approach. In this technique, one-by-one each feature is removed, and the prediction model is developed using the remaining features. Finally, a combination of features that give the best performance is selected. This technique is also known as the feature-reduction technique.

Development of Prediction Models

Here, we have developed the prediction models to distinguish HCC and non-tumorous samples using selected features. These models were implemented using Python package Scikit-learn (Pedregosa et al., 2011). A wide range of machine learning techniques have been used for developing these prediction models that include ExtraTrees (ETREES), Naive Bayes, K-nearest neighbor (KNN), Random Forest, Logistic Regression (LR), and SVC-RBF (radial basis function). The optimization of the parameters for the various classifiers was done by using a grid search with AUROC curve as scoring performance measure for selecting the best parameter.

Performance Evaluation of the Prediction Models

In the current study, both internal and external validation techniques were employed to evaluate the performance of models. First, the training dataset is used to develop prediction models and standard 10-fold cross-validation is used for performing internal validation, which is commonly employed in the literature (Burton et al., 2012; Bastani et al., 2013; Kourou et al., 2015; Bhalla et al., 2017; Jiang et al., 2018; Bhalla et al., 2019; Kaur et al., 2019). It is important to evaluate the realistic performance of the model on the external validation dataset, which should not be used for training and testing during model development. Therefore, we evaluated the performance of our models on four independent gene-expression cohorts that include GSE14520, GSE36376, GSE36076, and TCGA-LIHC obtained from GEO and The Cancer Genome Atlas (TCGA) (see **Figure 1B**), which were not used for training. In order to measure the performance of models, we used both threshold-dependent and threshold-independent parameters. In the case of threshold-dependent parameters, we measure sensitivity, specificity, accuracy, and Matthew's correlation coefficient (MCC) using the following equations.

$$Sensitivity\ (Sen) = \frac{TP}{TP + FN} \times 100 \quad (1)$$

$$Specificity\ (Spec) = \frac{TN}{TN + FP} \times 100 \quad (2)$$

$$Accuracy\ (Acc) = \frac{TP + TN}{TP + FP + TN + FN} \times 100 \quad (3)$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (4)$$

where FP, FN, TP, and TN are false positive, false negative, true positive, and true negative predictions, respectively.

In case of threshold-independent measures, we used a standard parameter Area under the Receiver Operating Characteristic (AUROC) curve. The AUROC curve is generated by plotting sensitivity or true positive rate against the false positive rate (1-specificity) at various thresholds. Finally, the area under the curve is calculated to compute a single parameter called AUROC.

Prognostic Potential of Identified HCC Diagnostic Biomarkers

The prognostic potential of the *"three-genes HCC biomarker"* was analyzed using gene-expression data of TCGA-LIHC and GSE14520 cohorts. The TCGA and GSE14520 datasets contain 374 and 219 tumor samples, respectively. Their clinical information was extracted from GEO, GDC, and the literature (Roessler et al., 2010; Liu et al., 2018a). The clinical characteristics of patients are given in **Table S1** (**Supplementary Information File 1**). Univariate survival analyses and risk assessments were performed by survival package in R (Therneau and Grambsch, 2000; Therneau, 2013). The distribution of the survival risk groups is done by using a log-rank test, eventually represented in the form of Kaplan-Meier plots. A p-value < 0.05 was considered the cut-off to describe the statistical significance in all survival analyses. Here, we analyzed four types of survivals, i.e., OS (Overall Survival), DSS (Disease-Specific Survival), DFS (Disease-Free Survival), and PFS (Progression-Free Survival) for TCGA-LIHC cohort, and two types of survivals, i.e., OS and RFS (Recurrence-Free Survival) (also called as DFS) for GSE14520 cohort. Besides, genes from the signature, univariate survival analysis is also performed on clinical characteristics of patients like age, gender, and tumor stage individually. Additionally, multivariate survival analysis was performed to assess the combined effect of clinical characteristics with the signature genes.

Functional Annotation of Signature Genomic Markers

In order to discern the biological relevance of the signature genes, enrichment analysis is performed using Enrichr (Kuleshov et al., 2016). Enrichr executes Fisher exact test to identify enrichment score. It provides Z-score and adjusted p-value, which is derived by applying correction on a Fisher exact test. We have considered only those Gene Ontology (GO) terms that are significantly enriched with adjusted p-value less than 0.05.

RESULTS

Overview

The pipeline of our analysis is illustrated in **Figure 2**. The detail of each step is described below.

Transcriptomic Cores for Hepatocellular Carcinoma

Identification of the Transcriptomic Cores

The individual statistical differential expression analyses of 27 gene-expression datasets resulted in the identification of hundreds of DEGs (**Supplementary Figure 1**). The 9,954 genes are present among each of the 27 datasets (**Supplementary Information File 1**, **Table S2**). Further, the comparative analysis among all 27 datasets scrutinized 26 overlapping genes that are differentially expressed in 80% or more datasets, i.e., 22 datasets. We called these genes as "core genes for hepatocellular carcinoma." Among these 26 genes, 12 are downregulated and 14 are upregulated in HCC in comparison to normal samples. The regulatory patterns of the core genes were consistent among most of the datasets (**Table 1**). Additionally, the expression pattern of these genes in training and three external validation datasets is shown in **Figure S2** (**Supplementary Information File 2**).

Gene Enrichment Analysis of the Transcriptomic Cores

Gene enrichment analysis of these "core genes of HCC" revealed their biological significance. The proteins encoded by the downregulated genes mainly enriched in complement activation and lectin pathways related processes. These genes negatively regulate cellular extravasation. They are also enriched in GO

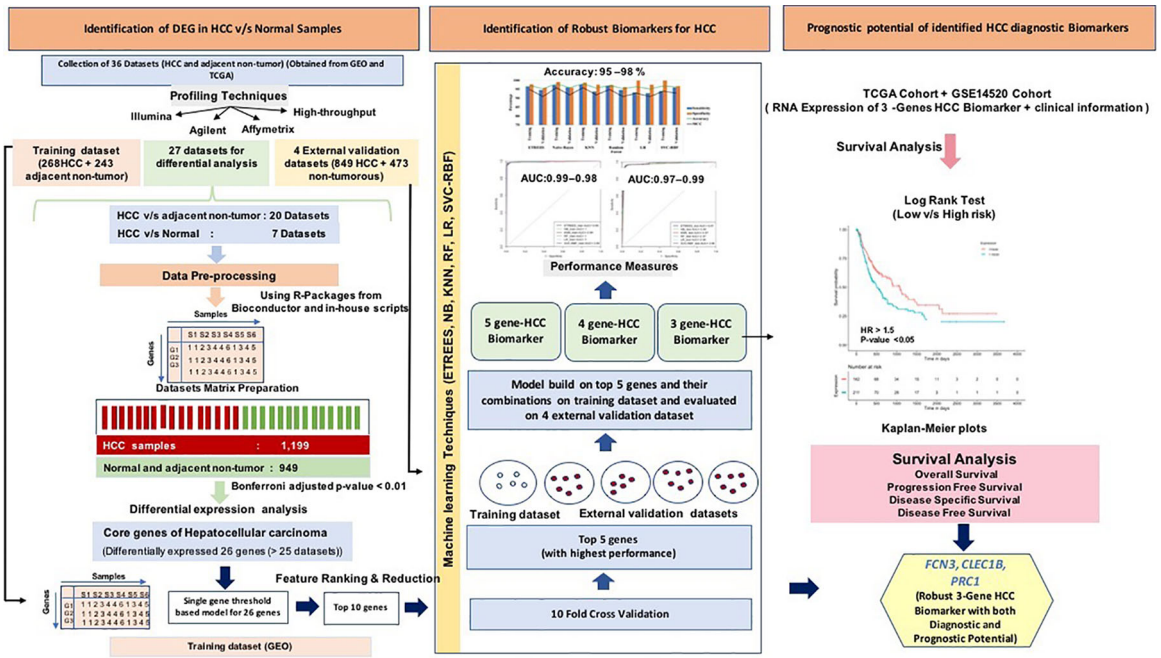

FIGURE 2 | Overview for the analysis implemented in the study.

TABLE1 | List of overlapping 26 genes that are differentially expressed (Core DEGs for HCC) between HCC and adjacent normal or adjacent non-tumor samples with Bonferroni p-values < 0.01.

Gene	#Up	#Down	#Sig	#Non-sig	Up (%)	Down (%)	Sig (%)	Regulation
FCN3	1	26	24	3	3.70	96.30	88.89	Down
CLEC4M	2	25	24	3	7.41	92.59	88.89	Down
FCN2	2	25	24	3	7.41	92.59	88.89	Down
MARCO	3	24	22	5	11.11	88.89	81.48	Down
CRHBP	2	25	22	5	7.41	92.59	81.48	Down
CFP	2	25	22	5	7.41	92.59	81.48	Down
STEAP3	2	25	25	2	7.41	92.59	92.59	Down
HGFAC	4	23	22	5	14.81	85.19	81.48	Down
CLEC1B	2	25	23	4	7.41	92.59	85.19	Down
CXCL12	3	24	24	3	11.11	88.89	88.89	Down
MT1E	3	24	24	3	11.11	88.89	88.89	Down
NSUN5	25	2	24	3	92.59	7.41	88.89	Down
MCM7	24	3	24	3	88.89	11.11	88.89	Up
MCM3	24	3	24	3	88.89	11.11	88.89	Up
ITGA6	24	3	24	3	88.89	11.11	88.89	Up
SSR2	24	3	23	4	88.89	11.11	85.19	Up
STMN1	23	4	24	3	85.19	14.81	88.89	Up
PRC1	24	3	23	4	88.89	11.11	85.19	Up
POLD1	24	3	23	4	88.89	11.11	85.19	Up
PBK	24	3	24	3	88.89	11.11	88.89	Up
IGSF3	22	5	23	4	81.48	18.52	85.19	Up
DTL	24	3	22	5	88.89	11.11	81.48	Up
ZWINT	24	3	22	5	88.89	11.11	81.48	Up
SPATS2	24	3	24	3	88.89	11.11	88.89	Up
GPSM2	23	4	23	4	85.19	14.81	85.19	Up
COL15A1	24	3	22	5	88.89	11.11	81.48	Up

Up, Upregulated in cancer or HCC; Down, Downregulated in cancer or HCC; #Up: No. of datasets in which gene is overexpressed; #Down: No. of datasets in which gene is underexpressed; #Sig: No. of datasets in which gene is significantly differentially expressed; #Non-Sig: No. of datasets in which gene is not significantly differentially expressed; Up (%): Percentage of datasets in which gene is overexpressed; Down (%): Percentage of datasets in which gene is underexpressed; Sig (%): Percentage of datasets in which gene is significantly differentially expressed.

molecular functions like serine-type endopeptidase, oxidoreductase, RNA methyltransferase activity, etc. (**Supplementary Information File 2**, **Figure S3**). Whereas, upregulated core genes are enriched in cell cycle GO biological processes like mitotic spindle organization and mitotic sister chromatid segregation, DNA synthesis and DNA replication, post-replication repair and cellular response to DNA damage stimulus, etc. They are also enriched in GO molecular functions such as exodeoxyribonuclease activity, GDP-dissociation inhibitor activity, DNA polymerase activity and insulin-like growth factor binding, etc. (**Supplementary Information File 2**, **Figure S3**).

Identification of HCC Biomarkers and Development of Prediction Models

Single-Gene Based Prediction Models

All 26 DEGs were ranked on the training dataset using threshold-based approach; ranking is based on their discriminatory power to distinguish HCC from non-tumorous samples (Bhalla et al., 2017; Kaur et al., 2019). The performance of the top 10 genes having maximum discriminatory power is shown in **Table 2**; see **Supplementary Information File 1**, **Table S3** for detail. These top 10 genes showed highest performance with an accuracy > 85%, MCC > 0.75, and AUROC > 0.85. We also evaluate the performance of these top 10 genes using 10-fold cross-validation to understand their robustness as shown in **Table S4** (**Supplementary Information File 1**). We further selected 5 genes out of 10 genes, which exhibit the maximum performance. These genes are *FCN3, CLEC1B, CLEC4M, PRC1*, and *PBK*; models based on these genes have accuracy more than 90% with AUROC > 0.95. In addition, the performance is also evaluated on the external validation datasets. The performance of the method was same on the training dataset but decreases on the external validation for few genes/features (see **Table S5**, **Supplementary Information File 1**).

Multiple-Genes Based Prediction Models

We identified the top five genes based on single gene-based prediction models, as described above. Further, we developed machine learning techniques-based classification models using these top five genes. We called these models as multiple-genes based prediction models as they take multiple genes as input. These models were evaluated on the training as well as validation datasets using internal and external cross-validation. The performance of these models on training as well as on three validation datasets is shown in **Table 3**. As shown in **Table 3**, we got AUROC approximately 0.98 on training as well as on the validation datasets. We further reduced one gene from selected set of five genes using feature reduction technique as described in *Materials and Methods* and obtained a set of four genes (*FCN3, CLEC1B, PRC1, PBK*). Subsequently, machine learning prediction models developed based on them classified HCC and non-tumor samples with accuracy more than 95% with AUROC in the range of 0.97–0.99 on both training and three independent validation datasets as shown in **Table S6** (**Supplementary Information File 1**). Results from this analysis show that we got nearly same performance using four genes-based biomarkers as we got in case of five genes-based biomarkers. Thus, reduction of one feature (five to four) does not affect the performance of our multiple-gene based prediction method. We further reduced features using feature reduction technique and got a set of three genes that contains *FCN3, CLEC1B,* and *PRC1*. Prediction models based on three genes-biomarker got accuracy 95–98% with AUROC in the range of 0.96–0.99 on training as well as independent validation datasets as shown in **Table 4**. The expression pattern of these three genes among samples of training dataset and three external validation datasets is depicted in **Figure 3**. We also tried two gene biomarkers, but there is substantial reduction in the performance on validation datasets. Thus, our final model is developed using a biomarker panel of three genes that include *FCN3, CLEC1B,* and *PRC1*. We considered three-genes based biomarker as the final model because the number of genes is limited. Hence, it is easy to implement in real life as well as economical.

Validation of Models on Blood Samples

In this study, models have been developed on tissue samples, which is complex and difficult to implement for routine testing. The question arises whether this model can also be used to discriminate the samples achieved from non-invasive techniques. Thus, we assessed the performance of our final model on PBMCs/blood samples of GSE36076. These signature genes

TABLE 2 | Top 10 genes based on the simple threshold-based approach.

Gene symbol	Thresh	Sens (%)	Spec (%)	Acc (%)	MCC	AUROC	Mean in HCC	Mean in normal	Mean diff
FCN2	9.78	97.76	99.59	98.63	0.97	0.98	5.76	10.89	–5.13
CLEC4M	7.59	97.01	98.77	97.85	0.96	0.98	4.32	9.37	–5.06
FCN3	10.76	95.15	99.18	97.06	0.94	0.97	7.87	12.32	–4.45
CLEC1B	9.46	95.52	97.94	96.67	0.93	0.97	5.96	11.38	–5.42
CFP	8.14	96.64	94.24	95.50	0.91	0.96	6.15	8.63	–2.48
CRHBP	8.69	92.54	96.71	94.52	0.89	0.95	6.35	10.30	–3.95
PRC1	7.76	91.42	97.12	94.13	0.88	0.94	10.03	6.35	3.68
PBK	6.03	91.04	93.42	92.17	0.84	0.93	8.65	4.41	4.24
DTL	6.71	85.82	94.65	90.02	0.80	0.91	8.72	5.20	3.52
IGSF3	6.93	81.34	91.77	86.30	0.73	0.88	8.10	6.08	2.01

Sens, Sensitivity; Spec, Specificity; Acc, Accuracy; MCC, Mathews Correlation Coefficient; AUROC, Area under Receiver operator curve; Thresh, Threshold; Mean diff, Mean in HCC–Mean in normal.

TABLE 3 | Performance of five genes (*FCN3*, *CLEC4M*, *CLEC1B*, *PRC1*, *PBK*) based models on training and validation datasets implementing various machine learning techniques.

Classifier	Sens (%)	Spec (%)	Acc (%)	MCC	AUROC with 95% CI	Sens (%)	Spec (%)	Acc (%)	MCC	AUROC with 95% CI
	Training Dataset					Validation Dataset1				
ETREES	97.39	98.35	97.85	0.96	0.99 (0.99-1)	97.78	94.09	95.96	0.92	0.98 (0.97-0.99)
NB	97.76	99.18	98.43	0.97	0.99 (0.99-1)	97.33	95.45	96.40	0.93	0.98 (0.97-0.99)
KNN	97.39	98.77	98.04	0.96	0.99 (0.99-1)	96.89	96.82	96.85	0.94	0.98 (0.97-0.99)
RF	97.01	97.94	97.46	0.95	0.99 (0.99-1)	97.33	94.55	95.96	0.92	0.98 (0.97-0.99)
LR	97.76	99.59	98.63	0.97	0.99 (0.99-1)	95.56	97.27	96.40	0.93	0.99 (0.98-0.99)
SVC	97.01	100	98.43	0.97	0.99 (0.99-1)	96.89	95.00	95.96	0.92	0.99 (0.98-0.99)
	Validation Dataset2					Validation Dataset3				
ETREES	95	97.41	96.07	0.92	0.98 (0.97-0.99)	97.86	96	97.64	0.89	0.99 (0.98-0.99)
NB	94.58	98.45	96.3	0.93	0.98 (0.96-0.99)	98.13	92	97.41	0.88	0.98 (0.98-0.99)
KNN	92.92	98.45	95.38	0.91	0.97 (0.96-0.99)	97.86	94	97.41	0.88	0.99 (0.98-0.99)
RF	96.67	93.26	95.15	0.9	0.98 (0.97-0.99)	98.4	90	97.41	0.88	0.99 (0.98-0.99)
LR	93.75	98.45	95.84	0.92	0.98 (0.97-0.99)	97.59	98	97.64	0.90	0.99 (0.98-0.99)
SVC-RBF	93.33	98.45	95.61	0.91	0.98 (0.97-0.99)	97.33	98	97.41	0.89	0.99 (0.98-0.99)

ETREES, Extra Trees Classifier; NB, Naive Bayes; KNN, K Neighbors Classifier; RF, Random Forest; LR, Logistic Regression; SVC-RBF, Support Vector Machine with RBF-kernel; Sens, Sensitivity; Spec, Specificity; Acc, Accuracy; MCC, Mathews Correlation Coefficient; AUROC, Area under Receiver operator curve.

TABLE 4 | Performance of three-genes HCC biomarker-A (*FCN3*, *CLEC1B*, *PRC1*) based models on training and validation datasets implementing various machine learning techniques.

Classifier	Sens (%)	Spec (%)	Acc (%)	MCC	AUROC with 95% CI	Sens (%)	Spec (%)	Acc (%)	MCC	AUROC with 95% CI
	Training Dataset					Validation Dataset1				
ETREES	96.64	97.94	97.26	0.95	0.99 (0.98-0.99)	94.67	95.91	95.28	0.91	0.97 (0.96-0.99)
NB	97.39	99.18	98.24	0.96	0.99 (0.99-1.0)	96.00	95.91	95.96	0.92	0.98 (0.97-0.99)
KNN	97.76	98.77	98.24	0.96	0.99 (0.99-1.0)	93.78	97.73	95.73	0.92	0.97 (0.96-0.99)
RF	97.01	97.53	97.26	0.95	0.99 (0.99-1.0)	94.67	96.36	95.51	0.91	0.97 (0.96-0.99)
LR	93.28	100	96.48	0.93	0.99 (0.99-1.0)	92.89	97.73	95.28	0.91	0.98 (0.97-0.99)
SVC-RBF	94.03	100	96.87	0.94	0.99 (0.98-0.99)	96.00	96.82	96.40	0.93	0.98 (0.97-0.99)
	Validation Dataset2					Validation Dataset3				
ETREES	93.75	96.37	94.92	0.90	0.98 (0.97-0.99)	95.72	98	95.99	0.84	0.99 (0.98-0.99)
NB	94.58	98.45	96.3	0.93	0.98 (0.97-0.99)	98.13	82	96.23	0.82	0.96 (0.95-0.98)
KNN	95.83	97.93	96.77	0.94	0.98 (0.97-0.99)	97.59	96	97.41	0.88	0.99 (0.98-0.99)
RF	95.42	94.3	94.92	0.90	0.98 (0.97-0.99)	95.45	96	95.52	0.82	0.98 (0.97-0.99)
LR	95.42	98.45	96.77	0.94	0.99 (0.98-0.99)	97.33	98	97.41	0.89	0.99 (0.98-0.99)
SVC-RBF	93.33	97.93	95.38	0.91	0.98 (0.97-0.99)	96.79	98	96.93	0.87	0.99 (0.98-0.99)

ETREES, Extra Trees Classifier; NB, Naive Bayes; KNN, K Neighbors Classifier; RF, Random Forest; LR, Logistic Regression; SVC-RBF, Support Vector Machine with RBF kernel; Sens, Sensitivity; Spec, Specificity; Acc, Accuracy; MCC, Mathews Correlation Coefficient; AUROC, Area under Receiver operator curve.

correctly predicted 90% of both HCC and healthy samples with ROC in the range of 0.91–0.96 and MCC 0.80–0.82. Complete results of prediction models are tabulated in **Table 5**. This demonstrates that our three genes-based models have the ability to discriminate HCC and healthy blood samples with reasonably high accuracy.

Protein-Based Biomarkers

In the past, proteins have been identified as diagnostic biomarkers for HCC. These protein biomarkers are AFP +GPC3 and AFP+GPC3+CK19 (*KRT19*) (Lou et al., 2017; Ocker, 2018). As we do not have their protein expression for these patients' samples, we employed only their gene expression values. Models based on the gene expression of *AFP+GPC3 +KRT19* classified HCC and normal samples of training dataset with an accuracy 67–75%. While this model attained accuracy of 69–77%, 51–87%, and 50–74% on external validation dataset1, dataset2 and dataset3, respectively, as shown in **Table S7** (**Supplementary Information File 1**). Further, the prediction models based on the gene expression of *AFP+GPC3* have improved performance on training dataset with an accuracy of 70–77%, but lower performance on all three validation datasets as given in **Table S8** (**Supplementary Information File 1**).

Survival Analysis to Determine the Prognostic Potential of "Three-Genes HCC Biomarker"

Univariate Survival Analysis for Three-Genes HCC Biomarker

To examine the prognostic potential of the "three-genes HCC biomarker," the univariate survival analysis was performed on TCGA-LIHC and GSE14520 cohorts. The samples were partitioned into low-risk and high-risk groups. Interestingly, all three genes of "three-genes HCC biomarker-A" are significantly

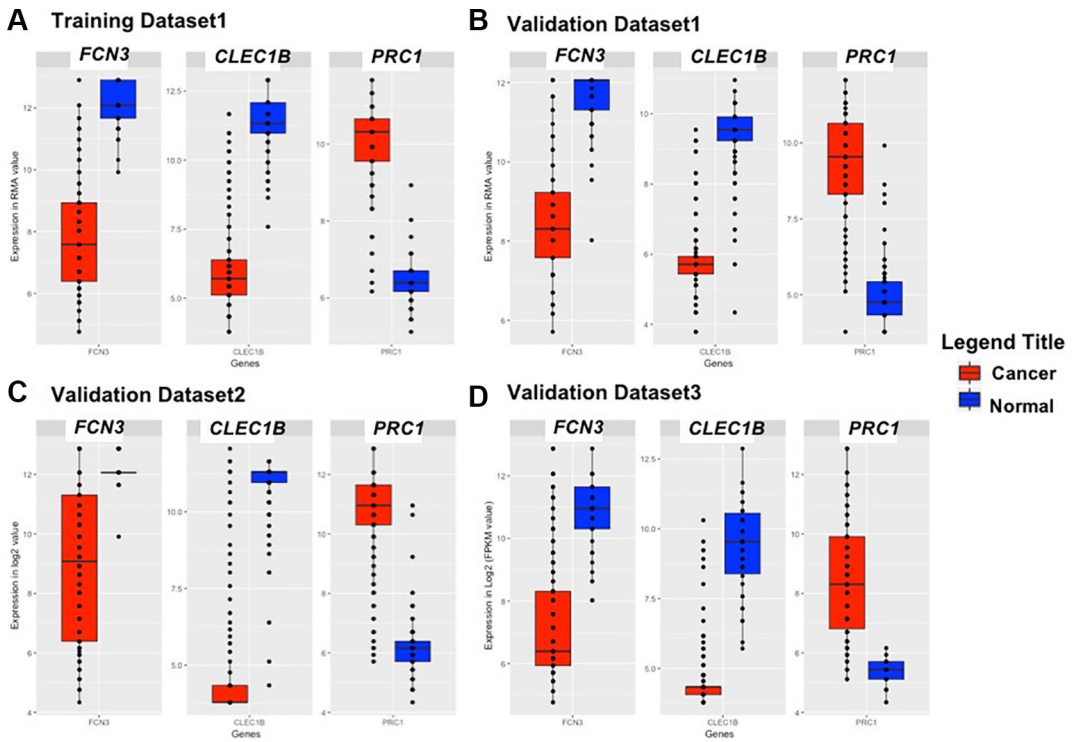

FIGURE 3 | Boxplot representing the expression pattern of three-genes panel-based HCC biomarker in the **(A)** Training Dataset, **(B)** Validation Dataset 1, **(C)** Validation Dataset 2, **(D)** Validation Dataset 3.

TABLE 5 | Performance of three-genes HCC biomarker-A (*FCN3*, *CLEC1B*, *PRC1*) based models on training and validation datasets 4 (containing blood samples, i.e., PBMCs) implementing various machine learning techniques.

Classifier	Sens (%)	Spec (%)	Acc (%)	MCC	AUROC with 95% CI	Sens (%)	Spec (%)	Acc (%)	MCC	AUROC with 95% CI
	Training Dataset					Validation Dataset4				
ETREES	94.78	99.18	96.87	0.94	0.99 (0.979-0.998)	100	80	90	0.82	0.93 (0.854-1.0)
NB	97.39	99.18	98.24	0.96	0.99 (0.989-1.0)	90	90	90	0.80	0.95 (0.81-1.0)
KNN	97.01	99.59	98.24	0.97	0.99 (0.986-1.0)	90	90	90	0.80	0.96 (0.878-1.0)
RF	95.52	99.59	97.46	0.95	0.99 (0.991-1.0)	100	80	90	0.82	0.93 (0.81-1.0)
LR	96.64	100	98.24	0.97	0.99 (0.992-1.0)	90	90	90	0.80	0.96 (0.877-1.0)
SVC	95.15	99.18	97.06	0.94	0.99 (0.988-0.999)	90	90	90	0.80	0.91 (0.744-1.0)

ETREES, Extra Trees Classifier; NB, Naive Bayes; KNN, K Neighbors Classifier; RF, Random Forest; LR, Logistic Regression; SVC-RBF, Support Vector Machine with RBF kernel; Sens, Sensitivity; Spec, Specificity; Acc, Accuracy; MCC, Mathews Correlation Coefficient; AUROC, Area under Receiver operator curve.

associated with the survival of HCC patients. For instance, higher expression (greater than mean) of *CLEC1B* and *FCN3* is significantly associated with good outcome of the patients, i.e. OS, DSS, DFS, and PFS; while the overexpression of *PRC1* is significantly associated with poor survival including DSS, DFS, or RFS and PFS of HCC patients for TCGA-LIHC dataset as shown in **Figure 4**. In the GSE14520 dataset, higher expression of *PRC1* is significantly associated with the poor outcome of patients, i.e., OS and DFS or RFS, while the higher expression of *FCN3* is significantly associated with the better outcome of HCC patients as depicted in **Figure 5**. Complete results of survival analysis with HR (Hazard Ratio), with 95% CI and p-value, are presented in **Table S9** (**Supplementary Information File 1**).

Univariate Survival Analysis for Clinical Features

The clinical characteristics of the patients like age, gender, tumor size, and stage are considered as important prognostic indicators for the survival of the patients in different malignancies including HCC (Best et al., 2016; Liu et al., 2018a; Wu et al., 2018; Yang et al., 2019). As the tumor size information is not present in one of the cohorts, therefore, we performed univariate survival analysis using only age, gender, and tumor stage of

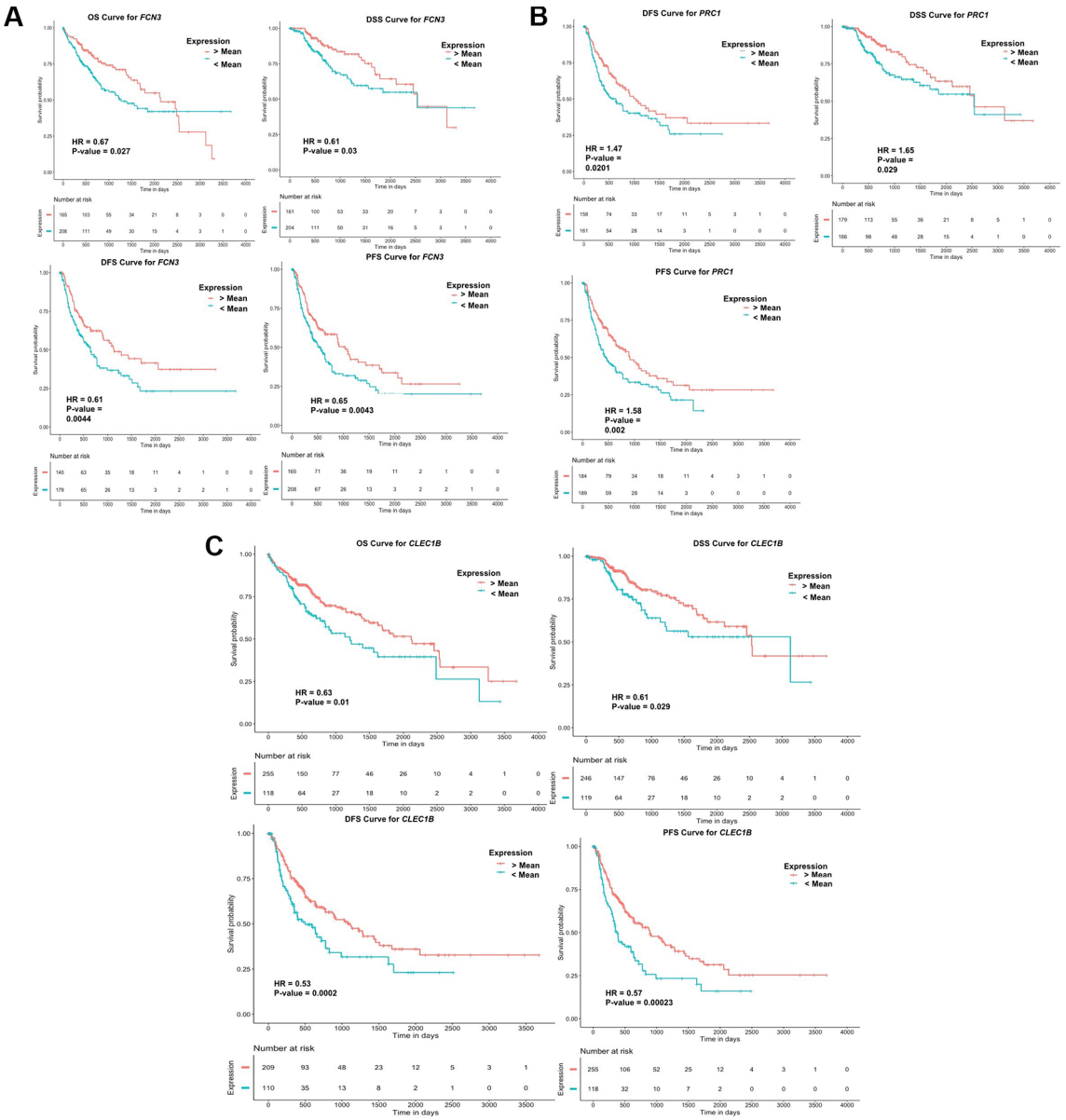

FIGURE 4 | Kaplan Meier survival curves for the risk estimation of HCC patient in TCGA cohort based on the RNA expression of **(A)** *FCN3*, **(B)** *PRC1*, and **(C)** *CLEC1B*.

the patients. This analysis shows that tumor stage is an important clinical factor with prognostic potential that significantly stratified high-risk and low-risk groups of patients in both cohorts, i.e., TCGA-LIHC and GSE14520. For instance, stage individually significantly (p-value <0.0001) stratified risk groups for OS, RFS with HR = 1.73 and HR = 1.65 of TCGA cohorts and with HR = 2.29 and HR = 1.79 of GSE14520 cohort, respectively (**Table S10**, **Supplementary Information File 1**). While the gender and age of patients do not possess high prognostic potential, as shown in **Table S10** (**Supplementary Information File 1**).

Multivariate Survival Analysis

Eventually, the multivariate analysis is performed to assess the independent impact of clinical characteristics and three genes of our signature biomarker that are determined as significant prognostic variables by univariate analysis. From this analysis, tumor stage is identified as the sole independent prognostic factor associated with the survival of HCC patients that significantly (with p-value <0.01) stratified high-risk and low-risk groups of both TCGA-LIHC and GSE1450 cohorts as presented in **Figures S4–S6** (**Supplementary Information File 2**).

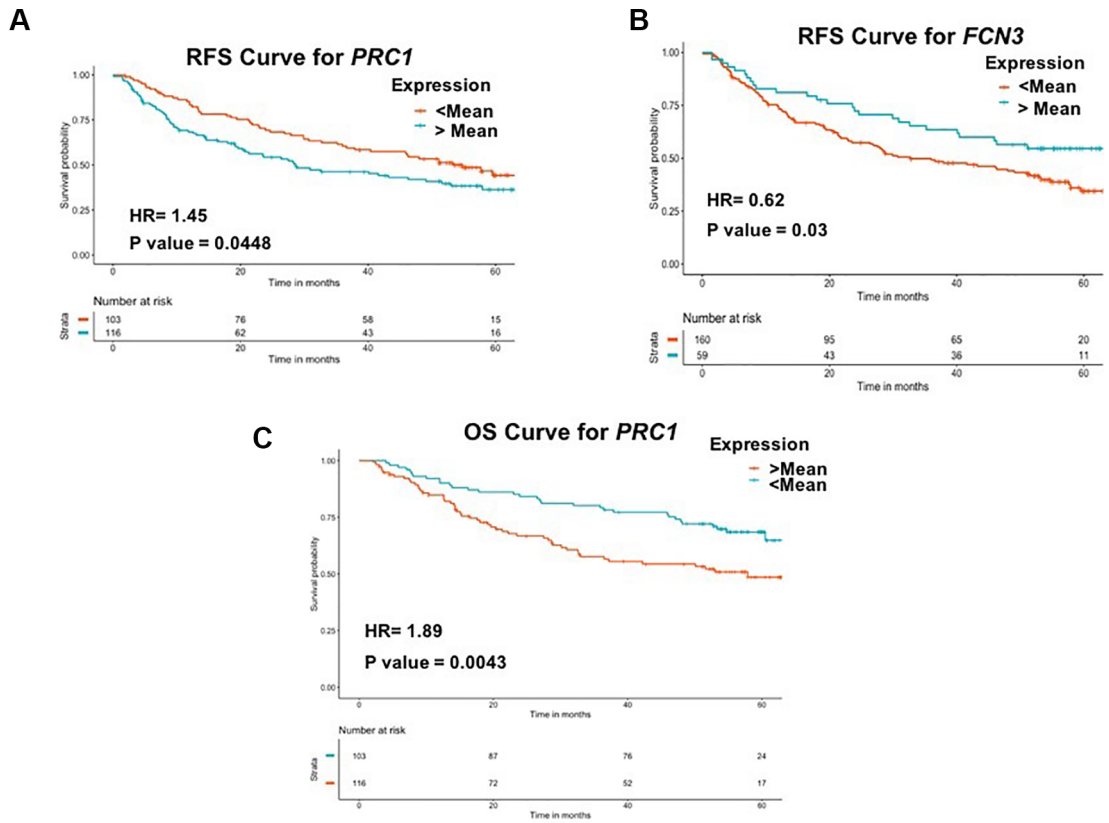

FIGURE 5 | Kaplan Meier survival curves for the risk estimation of HCC patient in GSE14520 cohort based on the RNA expression of **(A)** RFS for *PRC1*, **(B)** RFS for *FCN3*, and **(C)** OS for *PRC1*.

Web Server

To facilitates the scientific community working in the area of liver cancer research, we developed "HCCpred" (Prediction Server for Hepatocellular Carcinoma). In HCCpred, we execute mainly two modules: Prediction Module and Analysis Module based on robust five-genes, four-genes. and three-genes HCC biomarkers and 26 Core genes of HCC identified in the present study for the prediction and analysis of samples from the RNA-expression data. The prediction module permits the users to predict the disease status, i.e., cancerous or normal using RNA expression values of a subset of genes using *in silico* prediction models based on robust five-genes, four-genes, and three-genes HCC biomarkers identified in the present study. Here, the user is required to submit RMA (for Affymetrix), A-value (for Agilent), Log2 value (for Illumina), or FPKM (High throughput RNA-seq data) for a subset of genes or biomarkers. The output result displays a list for patient samples and corresponding predicted status of samples. Moreover, the user can select among the models, i.e., ETREES-based or SVC-RBF based model. Further, the Analysis Module permits the user to analyze the expression pattern of any of the top 10 ranked genes to check whether it is upregulated or downregulated in comparison to HCC samples based on the samples of the current study. This webserver is freely accessible at http://webs.iiitd.edu.in/raghava/hccpred/.

DISCUSSION

HCC is a type of tumor that is associated with the poor prognosis and a high mortality rate among the most common cancer types (Siegel et al., 2019). High recurrence rate and low rate of early detection results in poor prognosis. Accurate diagnosis of HCC may provide the opportunity for appropriate treatment, including traditional available treatment like liver transplantation resection, etc. Although the AFP and DCP proteins are well-established markers for the diagnosis of HCC, their sensitivity and specificity are not optimum (Sauzay et al., 2016). Therefore, the development of a novel robust diagnostic and prognostic biomarker for HCC is needed as it can assist in the existing clinical management of tumor. Towards this, our current report is an attempt to scrutinize a robust transcriptomic biomarker for HCC diagnosis. Briefly, in this study, we provide a novel large-scale analysis-based approach to identify a robust gene expression-based candidate diagnostic biomarker for HCC

derived from multiple transcriptomic profiles/datasets across a variety of platforms obtained from GEO and TCGA. This metadata integration approach employed to elucidate "core HCC DEGs" subset followed by a class prediction by implementing various machine learning algorithms. Eventually, validation on external independent datasets led us to the identification of multiple-genes based robust biomarkers for HCC.

Here, firstly, we have identified 26 genes named as "Core DEGs for HCC" that are uniformly differentially expressed among 80% of datasets. We have considered only these genes for downstream machine learning analysis. In an urge to identify a manageable subset with the minimum number of genes from this list that have a high discriminatory power, we further identified three genes signature-set containing *CLEC1B, FCN3*, and *PRC1*. This "three-genes based HCC biomarker" has predictive accuracy of 95–98% and AUROC 0.96–0.99 on the training and all three independent validation datasets. We further hypothesized that this biomarker gene set might be proved as quite an effective non-invasive diagnostic biomarker for HCC. Therefore, eventually, we validated their discriminatory performance on 20 PBMCs samples (GSE36076) extracted from 10 HCC and 10 healthy individuals. As anticipated, this biomarker set correctly classified 90% of the samples with AUROC in the range of 0.91–0.96. Besides, we also developed the prediction models based on the gene expression of already well-established protein biomarkers of HCC in the literature, i.e., *AFP+GPC3* and *AFP+GPC3+KRT19* (Lou et al., 2017). The prediction models based on *AFP+GPC3 +KRT19* discriminate samples of training dataset with an accuracy of 67–75% and 69–77% of validation dataset1, 55–87% of validation dataset2, and 50–74% of validation dataset3, while the models based on *AFP+GPC3* have quite lower performance on validation datasets. Further, we speculate that "three-genes HCC biomarker" can be explored as an effective novel protein based non-invasive biomarker as they have very good predictive power to distinguish HCC and non-tumor samples at gene expression level from the tissue and PBMC samples. Moreover, the product of *FCN3* gene is released in the serum and bile (Akaiwa et al., 1999; Brown et al., 2015; Pan et al., 2015; Tizzot et al., 2018); thus, this may serve as non-invasive biomarkers for diagnosis of HCC. Furthermore, recently, it has been reported that the protein product of two of the three genes from three-genes HCC biomarker, i.e., *PRC1* and *FCN3*, is also associated with HCC diagnosis and prognosis independently (Liu et al., 2018b; Shen et al., 2018). Hence, we anticipate that the three-genes signature might prove to be a good diagnostic and prognostic marker for HCC at the protein level as well. There is still a need for the validation of the protein product of these genes on a large scale of samples to confirm this hypothesis and their clinical utility.

Interestingly, the robust "three-genes HCC biomarker" contains *FCN3, PRC1,* and *CLEC1B,* has very high diagnostic ability, and also possesses prognostic potential, i.e., they are significantly associated with survival of HCC patients as determined by univariate analysis. For instance, higher expression of *CLEC1B* and *FCN3* significantly associated with the good outcome of HCC patients in TCGA-LIHC cohort; while higher expression of *PRC1* is significantly associated with the poor outcome of HCC patients in both TCGA-LIHC and GSE14520 cohorts. Besides, the role of *CLEC1B* and *PRC1* was previously also revealed in the diagnosis and prognosis of HCC (Chen et al., 2016; Chan et al., 2018; Hu et al., 2018; Kaur et al., 2019). Further, univariate analysis employing clinical factors of patients found that tumor stage of patients can act as a strong prognostic factor in the various types of survival, i.e., OS, RFS/DFS, PFS, and DSS of patients. Eventually, the multivariate survival analysis revealed the tumor stage as a sole independent prognostic factor, which was also corroborated with the previous literature (Aino et al., 2014; Wang and Li, 2019). The correct tumor stage identification is quite a tedious and challenging task in comparison to the quantification of the expression of genes.

In the past, a concern raised by Kaplan et al. is that despite the number of advantages of big studies, large sample size can also magnify the bias associated with an error resulting from sampling or study design (Kaplan et al., 2014). Thus, to reduce the overestimation of inferences from the results of large cohorts, we have included both types of cohorts, i.e., large cohort (sample size >50) and small cohort (sample size <50). We hypothesized that these results might be more reliable and applicable. Additionally, it might be practically more useful in real life, where, usually, small cohorts are available with maximum clinical parameters. Therefore, to ensure that cohort's size does not affect the results derived from the overall study, results should be validated on a small cohort as well. Towards this, we have also validated models built on the training dataset on three large cohorts of external validation dataset and one small cohort (contains 20 blood samples). Thus, these results indicate that there is no overestimation of inferences from the results of cohorts used in the study.

Taken together, we have established a robust three-gene HCC diagnostic biomarker with reasonable performance and possesses both diagnostic and prognostic potential. A meta-data integration pipeline is employed for the identification of a robust biomarker using machine learning techniques, which can work across different platforms. Further, this pipeline can also be used for the analysis of any other cancer type. Although more and more research is under the development of novel biomarkers, further work will be required to implement the clinical utilization of identified biomarker to meet real-world demand. We are anticipating that identifying novel cost-efficient biomarker using predictive technology for the detection of HCC will be promising.

CONCLUSIONS

This study identified and validated a highly accurate three-genes HCC biomarker for discriminating HCC and non-tumorous samples; it also possesses a significant prognostic potential that may facilitate more accurate early diagnosis and risk stratification upon validation in prospective clinical trials. Reasonable performance on the validation dataset of PBMCs samples indicates their non-invasive utility. Moreover, the protein product of *FCN3* is released in the serum and bile. Thus, this may serve as non-invasive protein diagnostic biomarkers. Large-scale non-invasive cohorts are required to confirm their non-invasive

clinical utility. Additionally, the uniform overexpression pattern of *PRC1* among numerous HCC samples suggests it as a novel potential therapeutic target for HCC.

DATA AVAILABILITY STATEMENT

We have taken the Gene-expression data from the public repositories, i.e., GEO (https://www.ncbi.nlm.nih.gov/geo/) and GDC data portal (https://portal.gdc.cancer.gov/).

AUTHOR CONTRIBUTIONS

HK collected the data and created the datasets. HK developed classification algorithms. HK and AD implemented algorithms. HK and AD performed the survival analysis. HK and AD created the back-end server and front-end user interface. HK and GR analyzed the results. HK, RK, and AD wrote the manuscript. GR conceived and coordinated the project, helped in the interpretation and analysis of data, refined the drafted manuscript, and gave complete supervision to the project. All of the authors have read and approved the final manuscript.

FUNDING

This research was funded by J. C. Bose National Fellowship (with Grant No. SRP076), Department of Science and Technology (DST), India.

ACKNOWLEDGMENTS

All the authors acknowledge funding agencies J. C. Bose National Fellowship DST. HK and RK are thankful to Council of Scientific and Industrial Research (CSIR) and AD is thankful to DST INSPIRE for providing fellowships, respectively.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01306/full#supplementary-material

SUPPLEMENTARY FIGURE S1 | Distribution of Significantly DEG (Differentially Expressed Genes) among various datasets with Bonferroni adjusted p-value < 0.01.

SUPPLEMENTARY FIGURE S2 | Heatmap representing the expression pattern of "Core genes of HCC" in different datasets.

SUPPLEMENTARY FIGURE S3 | Gene Enrichment analysis of 26 genes or "Core genes of HCC".

SUPPLEMENTARY FIGURE S4 | Multivariate analysis of clinical characteristics and three genes of HCC Biomarker on TCGA cohort for **(A)** OS, **(B)** RFS/DFS.

SUPPLEMENTARY FIGURE S5 | Multivariate analysis of clinical characteristics and three genes of HCC Biomarker on TCGA cohort for **(A)** DSS, **(B)** PFS.

SUPPLEMENTARY FIGURE S6 | Multivariate analysis of clinical characteristics and three genes of HCC Biomarker on GSE14520 cohort for **(A)** OS, **(B)** RFS/DFS.

REFERENCES

Aino, H., Sumie, S., Niizeki, T., Kuromatsu, R., Tajiri, N., Nakano, M., et al. (2014). Clinical characteristics and prognostic factors for advanced hepatocellular carcinoma with extrahepatic metastasis. *Mol. Clin. Oncol.* 2, 393–398. doi: 10.3892/mco.2014.259

Akaiwa, M., Yae, Y., Sugimoto, R., Suzuki, S. O., Iwaki, T., Izuhara, K., et al. (1999). Hakata Antigen, a New Member of the Ficolin/Opsonin p35 Family, Is a Novel Human Lectin Secreted into Bronchus/Alveolus and Bile. *J. Histochem. Cytochem.* 47, 777–785. doi: 10.1177/002215549904700607

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.* 41, D991–D995. doi: 10.1093/nar/gks1193

Bastani, M., Vos, L., Asgarian, N., Deschenes, J., Graham, K., Mackey, J., et al. (2013). A machine learned classifier that uses gene expression data to accurately predict estrogen receptor status. *PloS One* 8, e82144. doi: 10.1371/journal.pone.0082144

Best, J., Bilgi, H., Heider, D., Schotten, C., Manka, P., Bedreli, S., et al. (2016). The GALAD scoring algorithm based on AFP, AFP-L3, and DCP significantly improves detection of BCLC early stage hepatocellular carcinoma. *Z. Gastroenterol.* 54, 1296–1305. doi: 10.1055/s-0042-119529

Bhalla, S., Chaudhary, K., Kumar, R., Sehgal, M., Kaur, H., Sharma, S., et al. (2017). Gene expression-based biomarkers for discriminating early and late stage of clear cell renal cancer. *Sci. Rep.* 7, 44997. doi: 10.1038/srep44997

Bhalla, S., Kaur, H., Dhall, A., and Raghava, G. P. S. (2019). Prediction and Analysis of Skin Cancer Progression using Genomics Profiles of Patients. *Sci. Rep.* 9, 15790. doi: 10.1038/s41598-019-52134-4

Bhasin, M. K., Ndebele, K., Bucur, O., Yee, E. U., Otu, H. H., Plati, J., et al. (2016). Meta-analysis of transcriptome data identifies a novel 5-gene pancreatic adenocarcinoma classifier. *Oncotarget* 7, 23263–23281. doi: 10.18632/oncotarget8139

Brown, G. R., Hem, V., Katz, K. S., Ovetsky, M., Wallin, C., Ermolaeva, O., et al. (2015). Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res.* 43, D36–D42. doi: 10.1093/nar/gku1055

Burton, M., Thomassen, M., Tan, Q., and Kruse, T. A. (2012). Gene expression profiles for predicting metastasis in breast cancer: a cross-study comparison of classification methods. *Sci. World J.* 2012, 380495. doi: 10.1100/2012/380495

Cai, J., Li, B., Zhu, Y., Fang, X., Zhu, M., Wang, M., et al. (2017). Prognostic Biomarker Identification Through Integrating the Gene Signatures of Hepatocellular Carcinoma Properties. *EbioMed.* 19, 18–30. doi: 10.1016/j.ebiom.2017.04.014

Cai, C., Wang, W., and Tu, Z. (2019). Aberrantly DNA Methylated-Differentially Expressed Genes and Pathways in Hepatocellular Carcinoma. *J. Cancer* 10, 355–366. doi: 10.7150/jca.27832

Cancer Genome Atlas Research Network, Weinstein, J. N., Collisson, E. A., Mills, G. B., Mills Shaw, K. R., Ozenberger, B. A., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nat Genet.* 45 (10), 1113–1120. doi: 10.1038/ng.2764

Carvalho, B. S., and Irizarry, R. A. (2010). A framework for oligonucleotide microarray preprocessing. *Bioinformatics* 26, 2363–2367. doi: 10.1093/bioinformatics/btq431

Chan, H. L., Beckedorff, F., Zhang, Y., Garcia-Huidobro, J., Jiang, H., Colaprico, A., et al. (2018). Polycomb complexes associate with enhancers and promote oncogenic transcriptional programs in cancer through multiple mechanisms. *Nat. Commun.* 9, 3377. doi: 10.1038/s41467-018-05728-x

Chatterjee, A., Rodger, E. J., and Eccles, M. R. (2018). Epigenetic drivers of tumourigenesis and cancer metastasis. *Semin. Cancer Biol.* 51, 149–159. doi: 10.1016/J.SEMCANCER.2017.08.004

Chen, J., Rajasekaran, M., Xia, H., Zhang, X., Kong, S. N., Sekar, K., et al. (2016). The microtubule-associated protein PRC1 promotes early recurrence of hepatocellular carcinoma in association with the Wnt/β-catenin signalling pathway. *Gut* 65, 1522–1534. doi: 10.1136/gutjnl-2015-310625

Chen, C.-L., Tsai, Y.-S., Huang, Y.-H., Liang, Y.-J., Sun, Y.-Y., Su, C.-W., et al. (2018a). Lymphoid Enhancer Factor 1 Contributes to Hepatocellular Carcinoma Progression Through Transcriptional Regulation of Epithelial-Mesenchymal Transition Regulators and Stemness Genes. *Hepatol. Commun.* 2, 1392–1407. doi: 10.1002/hep41229

Chen, H., Zhang, Y., Li, S., Li, N., Chen, Y., Zhang, B., et al. (2018b). Direct comparison of five serum biomarkers in early diagnosis of hepatocellular carcinoma. *Cancer Manage. Res.* 10, 1947–1958. doi: 10.2147/CMAR.S167036

Chiyonobu, N., Shimada, S., Akiyama, Y., Mogushi, K., Itoh, M., Akahoshi, K., et al. (2018). Fatty Acid Binding Protein 4 (FABP4) Overexpression in Intratumoral Hepatic Stellate Cells within Hepatocellular Carcinoma with Metabolic Risk Factors. *Am. J. Pathol.* 188, 1213–1224. doi: 10.1016/j.ajpath.2018.01.012

Dawson, M. A., and Kouzarides, T. (2012). Cancer epigenetics: from mechanism to therapy. *Cell* 150, 12–27. doi: 10.1016/j.cell.2012.06.013

Deng, Y.-B., Nagae, G., Midorikawa, Y., Yagi, K., Tsutsumi, S., Yamamoto, S., et al. (2010). Identification of genes preferentially methylated in hepatitis C virus-related hepatocellular carcinoma. *Cancer Sci.* 101, 1501–1510. doi: 10.1111/j.1349-7006.2010.01549.x

Diaz, G., Engle, R. E., Tice, A., Melis, M., Montenegro, S., Rodriguez-Canales, J., et al. (2018). Molecular signature and mechanisms of hepatitis D virus–associated hepatocellular carcinoma. *Mol. Cancer Res.* 16, 1406–1419. doi: 10.1158/1541-7786.MCR-18-0012

Ding, Y., Yan, J.-L., Fang, A.-N., Zhou, W.-F., Huang, L., Ding, Y., et al. (2017). Circulating miRNAs as novel diagnostic biomarkers in hepatocellular carcinoma detection: a meta-analysis based on 24 articles. *Oncotarget* 8, 66402–66413. doi: 10.18632/oncotarget.18949

Dong, H., Zhang, L., Qian, Z., Zhu, X., Zhu, G., Chen, Y., et al. (2015). Identification of HBV-MLL4 integration and its molecular basis in chinese hepatocellular carcinoma. *PloS One* 10, e0123175. doi: 10.1371/journal.pone.0123175

Du, P., Kibbe, W. A., and Lin, S. M. (2008). lumi: a pipeline for processing Illumina microarray. *Bioinformatics* 24, 1547–1548. doi: 10.1093/bioinformatics/btn224

Emma, M. R., Iovanna, J. L., Bachvarov, D., Puleio, R., Loria, G. R., Augello, G., et al. (2016). NUPR1, a new target in liver cancer: implication in controlling cell growth, migration, invasion and sorafenib resistance. *Cell Death Dis.* 7, e2269. doi: 10.1038/cddis.2016.175

Flavahan, W. A., Gaskell, E., and Bernstein, B. E. (2017). Epigenetic plasticity and the hallmarks of cancer. *Science* 357, eaal2380. doi: 10.1126/science.aal2380

Gao, B., Ning, S., Li, J., Liu, H., Wei, W., Wu, F., et al. (2015). Integrated analysis of differentially expressed mRNAs and miRNAs between hepatocellular carcinoma and their matched adjacent normal liver tissues. *Oncol. Rep.* 34, 325–333. doi: 10.3892/or.20153968

Grinchuk, O. V., Yenamandra, S. P., Iyer, R., Singh, M., Lee, H. K., Lim, K. H., et al. (2018). Tumor-adjacent tissue co-expression profile analysis reveals pro-oncogenic ribosomal gene signature for prognosis of resectable hepatocellular carcinoma. *Mol. Oncol.* 12, 89–113. doi: 10.1002/1878-0261.12153

Grossman, R. L., Heath, A. P., Ferretti, V. V. H. E., Lowy, D. R., Kibbe, W. A., and Staudt, L. M. (2016). Toward a shared vision for cancer genomic data. *N. Engl. J. Med.* 375 (12), 1109–1112. doi: 10.1056/NEJMp1607591

Ho, D. W.-H., Lo, R. C.-L., Chan, L.-K., and Ng, I. O.-L. (2016). Molecular Pathogenesis of Hepatocellular Carcinoma. *Liver Cancer* 5, 290–302. doi: 10.1159/000449340

Hu, K., Wang, Z.-M., Li, J.-N., Zhang, S., Xiao, Z.-F., and Tao, Y.-M. (2018). CLEC1B Expression and PD-L1 expression predict clinical outcome in hepatocellular carcinoma with tumor hemorrhage. *Transl. Oncol.* 11, 552–558. doi: 10.1016/j.tranon.2018.02.010

Huang, H.-C., and Qin, L.-X. (2018). Empirical evaluation of data normalization methods for molecular classification. *PeerJ* 6, e4584. doi: 10.7717/peerj4584

Ji, J., Wang, H., Li, Y., Zheng, L., Yin, Y., Zou, Z., et al. (2016). Diagnostic evaluation of des-gamma-carboxy prothrombin versus α-Fetoprotein for hepatitis B virus-related hepatocellular carcinoma in China: a large-scale, multicentre study. *PloS One* 11, e0153227. doi: 10.1371/journal.pone.0153227

Ji, J., Chen, H., Liu, X.-P., Wang, Y.-H., Luo, C.-L., Zhang, W.-W., et al. (2018). A miRNA combination as promising biomarker for hepatocellular carcinoma diagnosis: a study based on bioinformatics analysis. *J. Cancer* 9, 3435–3446. doi: 10.7150/jca.26101

Jia, H.-L., Ye, Q.-H., Qin, L.-X., Budhu, A., Forgues, M., Chen, Y., et al. (2007). Gene expression profiling reveals potential biomarkers of human hepatocellular carcinoma. *Clin. Cancer Res.* 13, 1133–1139. doi: 10.1158/1078-0432.CCR-06-1025

Jiang, Y., Mei, W., Gu, Y., Lin, X., He, L., Zeng, H., et al. (2018). Construction of a set of novel and robust gene expression signatures predicting prostate cancer recurrence. *Mol. Oncol.* 12, 1559–1578. doi: 10.1002/1878-0261.12359

Jiao, Y., Li, Y., Jiang, P., Han, W., and Liu, Y. (2019). PGM5: a novel diagnostic and prognostic biomarker for liver cancer. *PeerJ* 7, e7070. doi: 10.7717/peerj7070

Kagohara, L. T., Stein-O'Brien, G. L., Kelley, D., Flam, E., Wick, H. C., Danilova, L. V., et al. (2018). Epigenetic regulation of gene expression in cancer: techniques, resources and analysis. *Brief. Funct. Genomics* 17, 49–63. doi: 10.1093/bfgp/elx018

Kamel, H. F. M., and Al-Amodi, H. S. A. B. (2017). Exploitation of gene expression and cancer biomarkers in paving the path to era of personalized medicine. *Genomics Proteomics Bioinf.* 15, 220–235. doi: 10.1016/j.gpb.2016.11.005

Kang, L., Liu, X., Gong, Z., Zheng, H., Wang, J., Li, Y., et al. (2015). Genome-wide identification of RNA editing in hepatocellular carcinoma. *Genomics* 105, 76–82. doi: 10.1016/j.ygeno.2014.11.005

Kaplan, R. M., Chambers, D. A., and Glasgow, R. E. (2014). Big Data and Large Sample Size: A Cautionary Note on the Potential for Bias. *Clin. Transl. Sci.* 7, 342–346. doi: 10.1111/cts.12178

Kaur, H., Bhalla, S., and Raghava, G. P. S. (2019). Classification of early and late stage liver hepatocellular carcinoma patients from their genomics and epigenomics profiles. *PloS One* 14, e0221476. doi: 10.1371/journal.pone.0221476

Kim, J. H., Sohn, B. H., Lee, H.-S., Kim, S.-B., Yoo, J. E., Park, Y.-Y., et al. (2014). Genomic predictors for recurrence patterns of hepatocellular carcinoma: model derivation and validation. *PloS Med.* 11, e1001770. doi: 10.1371/journal.pmed.1001770

Klett, H., Fuellgraf, H., Levit-Zerdoun, E., Hussung, S., Kowar, S., Küsters, S., et al. (2018). Identification and validation of a diagnostic and prognostic multi-gene biomarker panel for pancreatic ductal adenocarcinoma. *Front. Genet.* 9, 108. doi: 10.3389/fgene.2018.00108

Komatsu, H., Iguchi, T., Masuda, T., Ueda, M., Kidogami, S., Ogawa, Y., et al. (2016). HOXB7 expression is a novel biomarker for long-term prognosis after resection of hepatocellular carcinoma. *Anticancer Res.* 36, 2767–2773.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* 13, 8–17. doi: 10.1016/j.csbj.2014.11.005

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 44, W90–W97. doi: 10.1093/nar/gkw377

Kumar, R., Patiyal, S., Kumar, V., Nagpal, G., and Raghava, G. P. S. (2019). In Silico Analysis of Gene Expression Change Associated with Copy Number of Enhancers in Pancreatic Adenocarcinoma. *Int. J. Mol. Sci.* 20, 3582. doi: 10.3390/ijms20143582

Lamb, J. R., Zhang, C., Xie, T., Wang, K., Zhang, B., Hao, K., et al. (2011). Predictive genes in adjacent normal tissue are preferentially altered by sCNV during tumorigenesis in liver cancer and may rate limiting. *PloS One* 6, e20090. doi: 10.1371/journal.pone.0020090

Li, L., Lei, Q., Zhang, S., Kong, L., and Qin, B. (2017). Screening and identification of key biomarkers in hepatocellular carcinoma: Evidence from bioinformatic analysis. *Oncol. Rep.* 38, 2607–2618. doi: 10.3892/or.20175946

Li, J., Tan, W., Peng, L., Zhang, J., Huang, X., Cui, Q., et al. (2018a). Integrative analysis of gene expression profiles reveals specific signaling pathways associated with pancreatic duct adenocarcinoma. *Cancer Commun. (London England)* 38, 13. doi: 10.1186/s40880-018-0289-9

Li, N., Li, L., and Chen, Y. (2018b). The identification of core gene expression signature in hepatocellular carcinoma. *Oxid. Med. Cell. Longev.* 2018, 3478305. doi: 10.1155/2018/3478305

Liao, X., Liu, X., Yang, C., Wang, X., Yu, T., Han, C., et al. (2018). Distinct diagnostic and prognostic values of minichromosome maintenance gene expression in patients with hepatocellular carcinoma. *J. Cancer* 9, 2357–2373. doi: 10.7150/jca.25221

Lim, H.-Y., Sohn, I., Deng, S., Lee, J., Jung, S. H., Mao, M., et al. (2013). Prediction of disease-free survival in hepatocellular carcinoma by gene expression profiling. *Ann. Surg. Oncol.* 20, 3747–3753. doi: 10.1245/s10434-013-3070-y

Liu, F., Li, H., Chang, H., Wang, J., and Lu, J. (2015). Identification of hepatocellular carcinoma-associated hub genes and pathways by integrated microarray analysis. *Tumori* 101, 206–214. doi: 10.5301/tj.5000241

Liu, J., Lichtenberg, T., Hoadley, K. A., Poisson, L. M., Lazar, A. J., Cherniack, A. D., et al. (2018a). An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* 173, 400–416.e11. doi: 10.1016/j.cell.2018.02.052

Liu, X., Li, Y., Meng, L., Liu, X.-Y., Peng, A., Chen, Y., et al. (2018c). Reducing protein regulator of cytokinesis 1 as a prospective therapy for hepatocellular carcinoma. *Cell Death Dis.* 9, 534. doi: 10.1038/s41419-018-0555-4

Lou, J., Zhang, L., Lv, S., Zhang, C., and Jiang, S. (2017). Biomarkers for Hepatocellular Carcinoma. *Biomark. Cancer* 9, 1–9. doi: 10.1177/1179299X16684640

Mah, W.-C., Thurnherr, T., Chow, P. K. H., Chung, A. Y. F., Ooi, L. L. P. J., Toh, H. C., et al. (2014). Methylation profiles reveal distinct subgroup of hepatocellular carcinoma patients with poor prognosis. *PloS One* 9, e104158. doi: 10.1371/journal.pone.0104158

Makowska, Z., Boldanova, T., Adametz, D., Quagliata, L., Vogt, J. E., Dill, M. T., et al. (2016). Gene expression analysis of biopsy samples reveals critical limitations of transcriptome-based molecular classifications of hepatocellular carcinoma. *J. Pathol. Clin. Res.* 2, 80–92. doi: 10.1002/cjp2.37

Marshall, A., Lukk, M., Kutter, C., Davies, S., Alexander, G., and Odom, D. T. (2013). Global gene expression profiling reveals SPINK1 as a potential hepatocellular carcinoma marker. *PloS One* 8, e59459. doi: 10.1371/journal.pone.0059459

Mas, V. R., Maluf, D. G., Archer, K. J., Yanek, K., Kong, X., Kulik, L., et al. (2009). Genes involved in viral carcinogenesis and tumor initiation in hepatitis C virus-induced hepatocellular carcinoma. *Mol. Med.* 15, 85–94. doi: 10.2119/molmed.2008.00110

Meng, C., Shen, X., and Jiang, W. (2018). Potential biomarkers of HCC based on gene expression and DNA methylation profiles. *Oncol. Lett.* 16 (3), 3183–3192. doi: 10.3892/ol.20189020

Murakami, Y., Kubo, S., Tamori, A., Itami, S., Kawamura, E., Iwaisako, K., et al. (2015). Comprehensive analysis of transcriptome and metabolome analysis in Intrahepatic Cholangiocarcinoma and Hepatocellular Carcinoma. *Sci. Rep.* 5, 16294. doi: 10.1038/srep16294

Nagpal, G., Sharma, M., Kumar, S., Chaudhary, K., Gupta, S., Gautam, A., et al. (2015). PCMdb: Pancreatic Cancer Methylation Database. *Sci. Rep.* 4, 4197. doi: 10.1038/srep04197

Narrandes, S., and Xu, W. (2018). Gene Expression Detection Assay for Cancer Clinical Use. *J. Cancer* 9, 2249. doi: 10.7150/JCA.24744

Nebbioso, A., Tambaro, F. P., Dell'Aversana, C., and Altucci, L. (2018). Cancer epigenetics: Moving forward. *PloS Genet.* 14, e1007362. doi: 10.1371/journal.pgen.1007362

Ocker, M. (2018). Biomarkers for hepatocellular carcinoma: What's new on the horizon? *World J. Gastroenterol.* 24, 3974–3979. doi: 10.3748/wjg.v24.i353974

Oishi, N., Kumar, M. R., Roessler, S., Ji, J., Forgues, M., Budhu, A., et al. (2012). Transcriptomic profiling reveals hepatic stem-like gene signatures and interplay of miR-200c and epithelial-mesenchymal transition in intrahepatic cholangiocarcinoma. *Hepatology* 56, 1792–1803. doi: 10.1002/hep.25890

Pan, J.-W., Gao, X.-W., Jiang, H., Li, Y.-F., Xiao, F., and Zhan, R.-Y. (2015). Low serum ficolin-3 levels are associated with severity and poor outcome in traumatic brain injury. *J. Neuroinflammation* 12, 226. doi: 10.1186/s12974-015-0444-z

Pedersen, C. B., Nielsen, F. C., Rossing, M., and Olsen, L. R. (2018). Using microarray-based subtyping methods for breast cancer in the era of high-throughput RNA sequencing. *Mol. Oncol.* 12, 2136–2146. doi: 10.1002/1878-0261.12389

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *JMLR* 12, 2825–2830.

Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., et al. (2004). Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc. Natl. Acad. Sci. U.S.A.* 101, 9309–9314. doi: 10.1073/pnas.0401994101

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

Roessler, S., Jia, H.-L., Budhu, A., Forgues, M., Ye, Q.-H., Lee, J.-S., et al. (2010). A unique metastasis gene signature enables prediction of tumor relapse in early-stage hepatocellular carcinoma patients. *Cancer Res.* 70, 10202–10212. doi: 10.1158/0008-5472.CAN-10-2607

Sauzay, C., Petit, A., Bourgeois, A.-M., Barbare, J.-C., Chauffert, B., Galmiche, A., et al. (2016). Alpha-foetoprotein (AFP): A multi-purpose marker in hepatocellular carcinoma. *Clin. Chim. Acta* 463, 39–44. doi: 10.1016/j.cca.2016.10.006

Schulze, K., Imbeaud, S., Letouzé, E., Alexandrov, L. B., Calderaro, J., Rebouissou, S., et al. (2015). Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. *Nat. Genet.* 47, 505–511. doi: 10.1038/ng3252

Sekhar, V., Pollicino, T., Diaz, G., Engle, R. E., Alayli, F., Melis, M., et al. (2018). Infection with hepatitis C virus depends on TACSTD2, a regulator of claudin-1 and occludin highly downregulated in hepatocellular carcinoma. *PloS Pathog.* 14, e1006916. doi: 10.1371/journal.ppat.1006916

Seok, J. Y., Na, D. C., Woo, H. G., Roncalli, M., Kwon, S. M., Yoo, J. E., et al. (2012). A fibrous stromal component in hepatocellular carcinoma reveals a cholangiocarcinoma-like gene expression trait and epithelial-mesenchymal transition. *Hepatology* 55, 1776–1786. doi: 10.1002/hep.25570

Sharma, S., Kelly, T. K., and Jones, P. A. (2010). Epigenetics in cancer. *Carcinogenesis* 31, 27–36. doi: 10.1093/carcin/bgp220

Shen, S., Peng, H., Wang, Y., Xu, M., Lin, M., Xie, X., et al. (2018). Screening for immune-potentiating antigens from hepatocellular carcinoma patients after radiofrequency ablation by serum proteomic analysis. *BMC Cancer* 18, 117. doi: 10.1186/s12885-018-4011-8

Shirota, Y., Kaneko, S., Honda, M., Kawai, H. F., and Kobayashi, K. (2001). Identification of differentially expressed genes in hepatocellular carcinoma with cDNA microarrays. *Hepatology* 33, 832–840. doi: 10.1053/jhep.2001.23003

Siegel, R. L., Miller, K. D., and Jemal, A. (2019). Cancer statistics, 2019. *CA. Cancer J. Clin.* 69, 7–34. doi: 10.3322/caac.21551

Stefanska, B., Huang, J., Bhattacharyya, B., Suderman, M., Hallett, M., Han, Z.-G., et al. (2011). Definition of the Landscape of Promoter DNA Hypomethylation in Liver Cancer. *Cancer Res.* 71, 5891–5903. doi: 10.1158/0008-5472.CAN-10-3823

Therneau, T. (2013). A Package for Survival Analysis in S. R package version 2, 37–4, Available at: http://CRAN.R-project.org/package=survival

Therneau, T., and Grambsch, P. (2000). *Modeling Survival Data: Extending the Cox Model*. New York: Springer.

Tian, G., Yang, S., Yuan, J., Threapleton, D., Zhao, Q., Chen, F., et al. (2018). Comparative efficacy of treatment strategies for hepatocellular carcinoma: systematic review and network meta-analysis. *BMJ Open* 8, e021269. doi: 10.1136/bmjopen-2017-021269

Tizzot, M. R., Lidani, K. C. F., Andrade, F. A., Mendes, H. W., Beltrame, M. H., Reiche, E., et al. (2018). Ficolin-1 and Ficolin-3 Plasma Levels are altered in HIV and HIV/HCV coinfected patients from Southern Brazil. *Front. Immunol.* 9, 2292. doi: 10.3389/fimmu.2018.02292

Tung, E. K.-K., Mak, C. K.-M., Fatima, S., Lo, R. C.-L., Zhao, H., Zhang, C., et al. (2011). Clinicopathological and prognostic significance of serum and tissue Dickkopf-1 levels in human hepatocellular carcinoma. *Liver Int.* 31, 1494–1504. doi: 10.1111/j.1478-3231.2011.02597.x

Vasudevan, S., Flashner-Abramson, E., Remacle, F., Levine, R. D., and Kravchenko-Balasha, N. (2018). Personalized disease signatures through information-theoretic compaction of big cancer data. *Proc. Natl. Acad. Sci. U.S.A.* 115, 7694–7699. doi: 10.1073/pnas.1804214115

Villa, E., Critelli, R., Lei, B., Marzocchi, G., Cammà, C., Giannelli, G., et al. (2016). Neoangiogenesis-related genes are hallmarks of fast-growing hepatocellular carcinomas and worst survival. Results from a prospective study. *Gut* 65, 861–869. doi: 10.1136/gutjnl-2014-308483

Wang, C.-Y., and Li, S. (2019). Clinical characteristics and prognosis of 2887 patients with hepatocellular carcinoma: a single center 14 years experience

from China. *Med. (Baltimore).* 98, e14070. doi: 10.1097/MD.0000000000014070

Wang, H., Huo, X., Yang, X.-R., He, J., Cheng, L., Wang, N., et al. (2017). STAT3-mediated upregulation of lncRNA HOXD-AS1 as a ceRNA facilitates liver cancer metastasis by regulating SOX4. *Mol. Cancer* 16, 136. doi: 10.1186/s12943-017-0680-1

Wang, Z., Teng, D., Li, Y., Hu, Z., Liu, L., and Zheng, H. (2018). A six-gene-based prognostic signature for hepatocellular carcinoma overall survival prediction. *Life Sci.* 203, 83–91. doi: 10.1016/j.lfs.2018.04.025

WELCH, B. L. (1947). The generalisation of student's problems when several different population variances are involved. *Biometrika* 34, 28–35. doi: 10.1093/biomet/34.1-2.28

Wong, K.-F., Liu, A. M., Hong, W., Xu, Z., and Luk, J. M. (2016). Integrin α2β1 inhibits MST1 kinase phosphorylation and activates Yes-associated protein oncogenic signaling in hepatocellular carcinoma. *Oncotarget* 7, 77683–77695. doi: 10.18632/oncotarget.12760

Woo, H. G., Choi, J.-H., Yoon, S., Jee, B. A., Cho, E. J., Lee, J.-H., et al. (2017). Integrative analysis of genomic and epigenomic regulation of the transcriptome in liver cancer. *Nat. Commun.* 8, 839. doi: 10.1038/s41467-017-00991-w

Wu, G., Wu, J., Wang, B., Zhu, X., Shi, X., and Ding, Y. (2018). Importance of tumor size at diagnosis as a prognostic factor for hepatocellular carcinoma survival: a population-based study. *Cancer Manage. Res.* 10, 4401–4410. doi: 10.2147/CMAR.S177663

Xia, Q., Li, Z., Zheng, J., Zhang, X., Di, Y., Ding, J., et al. (2019). Identification of novel biomarkers for hepatocellular carcinoma using transcriptome analysis. *J. Cell. Physiol.* 234, 4851–4863. doi: 10.1002/jcp.27283

Xu, W., Rao, Q., An, Y., Li, M., and Zhang, Z. (2018). Identification of biomarkers for Barcelona Clinic Liver Cancer staging and overall survival of patients with hepatocellular carcinoma. *PloS One* 13, e0202763. doi: 10.1371/journal.pone.0202763

Yang, J. D., Addissie, B. D., Mara, K. C., Harmsen, W. S., Dai, J., Zhang, N., et al. (2019). Galad score for hepatocellular carcinoma detection in comparison with liver ultrasound and proposal of galadus score. *Cancer Epidemiol. Biomarkers Prev.* 28, 531–538. doi: 10.1158/1055-9965.EPI-18-0281

Zhang, C., Peng, L., Zhang, Y., Liu, Z., Li, W., Chen, S., et al. (2017). The identification of key genes and pathways in hepatocellular carcinoma by bioinformatics analysis of high-throughput data. *Med. Oncol.* 34, 101. doi: 10.1007/s12032-017-0963-9

Zhang, Y.-L., Ding, C., and Sun, L. (2019). High expression B3GAT3 is related with poor prognosis of liver cancer. *Open Med. (Warsaw Poland)* 14, 251–258. doi: 10.1515/med-2019-0020

Zhao, X., Parpart, S., Takai, A., Roessler, S., Budhu, A., Yu, Z., et al. (2015). Integrative genomics identifies YY1AP1 as an oncogenic driver in EpCAM+ AFP+ hepatocellular carcinoma. *Oncogene* 34, 5095–5104. doi: 10.1038/onc.2014.438

Zheng, Y., Liu, Y., Zhao, S., Zheng, Z., Shen, C., An, L., et al. (2018). Large-scale analysis reveals a novel risk score to predict overall survival in hepatocellular carcinoma. *Cancer Manage. Res.* 10, 6079–6096. doi: 10.2147/CMAR.S181396

Zubiete-Franco, I., García-Rodríguez, J. L., Lopitz-Otsoa, F., Serrano-Macia, M., Simon, J., Fernández-Tussy, P., et al. (2019). Sumoylation regulates LKB1 localization and its oncogenic activity in liver cancer. *EBioMedicine* 40, 406–421. doi: 10.1016/j.ebiom.2018.12.031

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

frontiers in Genetics

REVIEW
published: 28 January 2020
doi: 10.3389/fgene.2019.01381

Heterogeneous Multi-Layered Network Model for Omics Data Integration and Analysis

Bohyun Lee[1], *Shuo Zhang*[1], *Aleksandar Poleksic*[2] *and Lei Xie*[1,3,4,5]*

[1] Ph.D. Program in Computer Science, The City University of New York, New York, NY, United States, [2] Department of Computer Science, The University of Northern Iowa, Cedar Falls, IA, United States, [3] Ph.D. Program in Biochemistry and Biology, The City University of New York, New York, NY, United States, [4] Department of Computer Science, Hunter College, The City University of New York, New York, NY, United States, [5] Helen and Robert Appel Alzheimer's Disease Research Institute, Feil Family Brain & Mind Research Institute, Weill Cornell Medicine, Cornell University, Ithaca, NY, United States

Advances in next-generation sequencing and high-throughput techniques have enabled the generation of vast amounts of diverse omics data. These big data provide an unprecedented opportunity in biology, but impose great challenges in data integration, data mining, and knowledge discovery due to the complexity, heterogeneity, dynamics, uncertainty, and high-dimensionality inherited in the omics data. Network has been widely used to represent relations between entities in biological system, such as protein-protein interaction, gene regulation, and brain connectivity (i.e. network construction) as well as to infer novel relations given a reconstructed network (aka link prediction). Particularly, heterogeneous multi-layered network (HMLN) has proven successful in integrating diverse biological data for the representation of the hierarchy of biological system. The HMLN provides unparalleled opportunities but imposes new computational challenges on establishing causal genotype-phenotype associations and understanding environmental impact on organisms. In this review, we focus on the recent advances in developing novel computational methods for the inference of novel biological relations from the HMLN. We first discuss the properties of biological HMLN. Then we survey four categories of state-of-the-art methods (matrix factorization, random walk, knowledge graph, and deep learning). Thirdly, we demonstrate their applications to omics data integration and analysis. Finally, we outline strategies for future directions in the development of new HMLN models.

Keywords: data mining and knowledge discovery, machine learning, biological data analysis, biological network, link prediction, relation inference, deep learning

OPEN ACCESS

Edited by:
Shizhong Han, Johns Hopkins Medicine, United States

Reviewed by:
Olivier Taboureau, Paris Diderot University, France
Chih-Hsu Lin, Baylor College of Medicine, United States

***Correspondence:**
Lei Xie
lxie@iscb.org

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

Received: *28 September 2019*
Accepted: *18 December 2019*
Published: *28 January 2020*

Citation:
Lee B, Zhang S, Poleksic A and Xie L (2020) Heterogeneous Multi-Layered Network Model for Omics Data Integration and Analysis. Front. Genet. 10:1381. doi: 10.3389/fgene.2019.01381

INTRODUCTION

A fundamental task in biological studies is to identify relations, more specifically dynamic functional associations or physical interactions between various chemical and biological entities. Network has been widely used to represent relations between entities in biology such as gene regulation, signaling transduction, metabolism, brain connectivity, and species interaction. In the network, a node represents an entity such as chemical compound, gene, protein, etc. A link between nodes represents their relations. There are basically two types of relations (or links), intra-domain

relations and cross-domain relations. An intra-domain link denotes a relation between the same type of entities, e.g. a protein-protein interaction. A cross-domain link represents a relation between two entities that belong to different types, e.g. protein-chemical interactions. Given a network of nodes and links (observed relations), a computational challenge is how to predict missing relations.

Depending on the underlying algorithms, relation inference (or link prediction) can be formulated as a problem in a homogenous network, a multiplex network, or a heterogeneous multi-layered network (HMLN), as shown in **Figure 1**. In a homogenous network (**Figure 1A**), all nodes from different domains, as well as intra-domain and cross-domain relations, are treated equally. In contrast, multiplex and multi-layered networks separate different types of nodes and relations. A multiplex network is often used to represent homogeneous nodes that have different types of characterizations (a.k.a. views). For example, a gene can be characterized by multiple measurements of gene expression, essentiality, literature citation, phylogenetic profile, neighborhood in the interaction network, biological pathway involved, Gene Ontology annotation, protein domain profile etc. (Hwang et al., 2019). Each type of measurement can form a unique type of link between genes (**Figure 1B**). In a HMLN (**Figure 1C**), multiple types of heterogeneous nodes are involved. The nodes from each type are grouped into a single layer and treated separately. In the same vein, different types of intra-domain and cross-domain relations are marked differently in a multi-layered network. We note that more complex network representations, such as multiplex multi-layered network, may be needed in real applications. In this review, we focus on the cross-domain relation inference (or link prediction) problem for the HMLN. Readers can refer other excellent reviews of the multiplex networks (Chauvel et al., 2019).

Recently, multi-layered networks have been proposed to connect multiple inter-dependent heterogeneous domains in biology (Himmelstein and Baranzini, 2015; Chen et al., 2016; Kringelum et al., 2016; Li et al., 2017; Pinero et al., 2017) and ecology (Silk et al., 2018). A typical example of a multi-layered network is HetioNet (Himmelstein and Baranzini, 2015) (**Figure 1D**). HetioNet contains nine domains, namely compound, pharmacologic class, gene, pathway, biological process, disease, side effect, symptom, and anatomy. Another example of a multi-layered network is a multi-scale model that represents metabolic phenotypic response to vaccination (Li et al., 2017). It consists of four layers: blood transcriptomics, plasma metabolomics, plasma cytokines, and cell populations. The multi-layered network provides a natural way to represent the hierarchy of a biological system and its environmental context: from genetic markup to gene to biological pathway to cellular function to organismal phenotype to population dynamics. It allows us to uncover novel relations between biological entities (e.g. genotype-phenotype associations) on a multi-scale. Furthermore, the cross-layer relations may represent casual effects (e.g. loss-of-function mutation) rather than statistical correlations, e.g. Genome-Wide Association Studies (GWAS). Compared to a homogeneous single-layered network, a unique topological characteristic of a multi-layered network lies in its cross-layer relation or dependency structure in addition to intra-layer connectivity. For example, in HetioNet (Himmelstein and Baranzini, 2015), a compound can inhibit or activate a gene. This cross-layer dependency often plays a central role in a multi-layered network. The prediction of new cross-layer relations is often the key to new discoveries, such as a treatment of a new disease by an existing drug, i.e. drug repurposing.

Substantial efforts have been devoted to reconstructing a multi-layered network [e.g. HetioNet (Himmelstein and Baranzini, 2015)] from the experimentally observed or

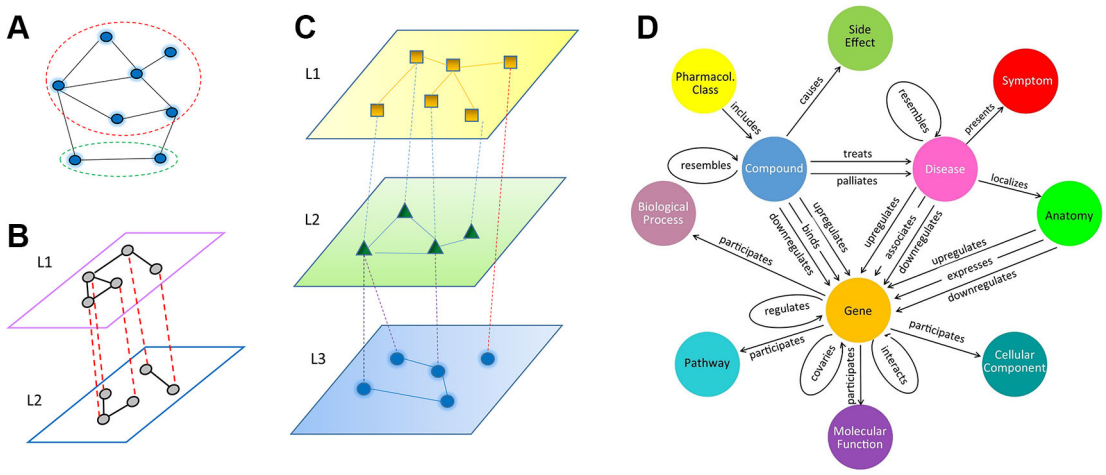

FIGURE 1 | Illustration of three types of network models, **(A)** homogeneous network, where all nodes and edges are treated equally, even though they may belong to different types (dashed red and green circles). **(B)** multiplex network, **(C)** multi-layered network, **(D)** an example of heterogeneous multi-layered network HetioNet (Himmelstein and Baranzini, 2015).

computationally inferred heterogeneous data sets. Even though the recent technology advances have enabled the generation of a vast amount of biological, physiological, and epidemiological data, the cross-layer relations observed by experiments are rarely complete, unbiased, and certain (Xie et al., 2017). Many important cross-layer relations are even completely missing. For example, there are no connections between genes and side effects in HetioNet, although such linkages are critical in understanding the molecular and genetic basis of adverse drug reactions. An unsolved computational problem is how to efficiently, accurately, and robustly infer the missing cross-layer relations in a HMLN.

In this review, we summarize the recent advances in the development of cross-layer relation inference algorithms for the HMLN, and their applications to biological discovery. The paper will be organized as follows. First, we will discuss the properties of biological HMLN. Second, we will introduce four major computational strategies for the cross-layer relation prediction, namely, matrix factorization, random walk, meta-path, and deep learning. Then, we will demonstrate the applications of these methods in biomedicine. Finally, we will discuss the unsolved issues and future directions.

CHARACTERISTICS OF BIOLOGICAL HMLN

Biological HMLN has several unique characteristics that impose great challenges for cross-layer relation inference.

Biasness

Due to limitation of experimental techniques and biases of researchers' interests, the observed data is highly skewed to certain gene families, species, diseases, etc. (Xie and Bourne, 2005) Rapid accumulation of large omics data could alleviate this problem to a certain degree. In addition, the reported positive results often greatly exceed the reported negative results, as the latter ones are seldom reported in the literature. Unless this reporting bias is taken into account, the models trained using the observed data by machine learning are unrealistic and hence unreliable when applied to unseen data.

Noisiness

Many observed cross-layer links are noisy. The source of noisiness is mainly due to the inconsistency in the experimental and clinical observations. Given the same relation, the inconsistency might result from different experimental protocols, computational pipelines, and batch effects.

Uncertainty

The relations in HMLN often come from calculated values or predictions made by heuristic algorithms. For example, many algorithms exist for computing intra-layer relations, such as chemical-chemical similarity. These methods differ in the choice of chemical representation and similarity metric employed. Similarly, no method is perfect for constructing cross-layer relations. While text mining is a popular technique, it is known to introduce a large number of false positives.

Conditionality

Biological observations could be from different cell lines, culture conditions, disease conditions, and environmental conditions. Under different circumstances, the biological relations are changed dynamically. For example, the physical strength and functional consequence of protein-ligand binding are strongly dependent on that mutation and post-modification state of protein, gene expression profile, and other factors.

Ambiguity

Many relations in HMLN are ambiguous and require proper classification. In one scenario, a relation can have opposite biological consequence. For example, the "association" relation between diseases and genes in HetioNet (**Figure 1D**) can be either "upregulate" or "downregulate". Another example is the binding of bioactivity compounds on a protein. The bioactivity of compound is often ambiguous. It could be an agonist or an antagonist.

Sparsity and Imbalance

The observed cross-layer links are highly sparse. In the real world, the number of relations of existence could be far less than the number of relations of non-existence. For example, a highly selective drug only binds to several protein isoforms among hundreds of thousands of protein isoforms in human. In addition, the observed relations are rare compared with the unobserved relations. For example, among hundreds of millions of sequenced genes, only tens of thousands of genes have the bioactivity data associated with chemical compounds. Because the negative cases often and greatly outnumber the positive ones, this imbalance imposes a great challenge in model training and evaluation.

Open World Assumption

Missing links cannot be treated as false relations, but instead as "unknown". In reality, these links could represent either a true or false relation (of different kinds, if the relation is not binary), or the lack of a relation.

ALGORITHMS FOR RELATION INFERENCE IN HMLN

Overview

The premise of relation inference or link prediction is that the missing relations can be inferred from the existing observed relations. Although such direct linkages are sparse, they can be recovered through intermediate intra-domain and cross-domain relations. For example, if a rare SNP S_x is a gain-of-function mutation of the gene G_3 and if G_3 is associated with the tall height P_1, then S_x is likely to be associated with P_1, even if the S_x-P_1 association is not statistically significant in the GWAS (**Figure 2**). However, such a simplistic inference method, based

on the existing highly sparse and highly biased observations, is prone to type I errors. In the above setting, multiple genes (e.g. G_3 and G_4) may be collectively responsible for P_1 and thus the likelihood of the inference "S_x causes P_1" has to be adjusted accordingly. To factor in the network multi-connectivity, an algorithm needs to jointly predict whether S_x is associated with other genes and whether these genes are associated with P_1, by simultaneously taking all observed cross-layer and intra-layer relations into account. In **Figure 2** example, the linkages of S_x-S_2-> G_2- G_4 and S_x- S_3-> G_2- G_4 will significantly strengthen the inferred S_x-P_1 association.

A number of algorithms have been developed to solve the relation inference problem in HMLN. All of these algorithms follow a common framework, consisting of two steps, as shown in **Figure 3**. The first step is to infer low dimension (i.e. rank) latent features for each entity and/or relation (aka node embedding and edge embedding). In the second step, the latent features from different layers are used to restore all missing cross-layer relations through a simple inner product or other more sophisticated machine learning techniques. In **Figure 3**, a chemical-gene-disease network is used to illustrate the concept. The input is a matrix representation of multi-layered network including both intra-layer relations (disease-disease similarity, gene-gene similarity, and chemical-chemical similarity) or their attributes (e.g. fingerprint representation for nodes in chemical layer, sequence representation for nodes in gene layer, and word2vec representation for nodes in disease layer), as well as a set of cross-layer relations (observed gene-disease association and chemical-gene interaction). In principle, even if we do not know any drug-disease associations, we can infer them through observed drug-gene, and gene-disease associations. The difference between the algorithms lies in the objective function for shallow or deep representations in the first step and machine learning methods for classification, regression, or ranking used in the second step.

In the next section, we review the major embedding algorithms in more details. These algorithms can be roughly classified into matrix factorization, random walk, meta-path, graph convolutional network (GCN), and their combinations.

Matrix Factorization

The cross-layer relation inference problem is conceptually related to collaborative filtering (Goldberg et al., 1992). Commonly used collaborative filtering methods can be classified into two groups: neighborhood methods (Breese et al., 1998) and latent factor methods (Koren et al., 2009). As the latent factor approach is generally more effective in capturing the implicit cross-layer relations, many variants of this methodology, such as recommended systems (Portugal et al., 2018), have been proposed to address relation inference problems in a two-layered network (Gao et al., 2019; Xuan et al., 2019). However, few methods have been developed for the multi-layered network.

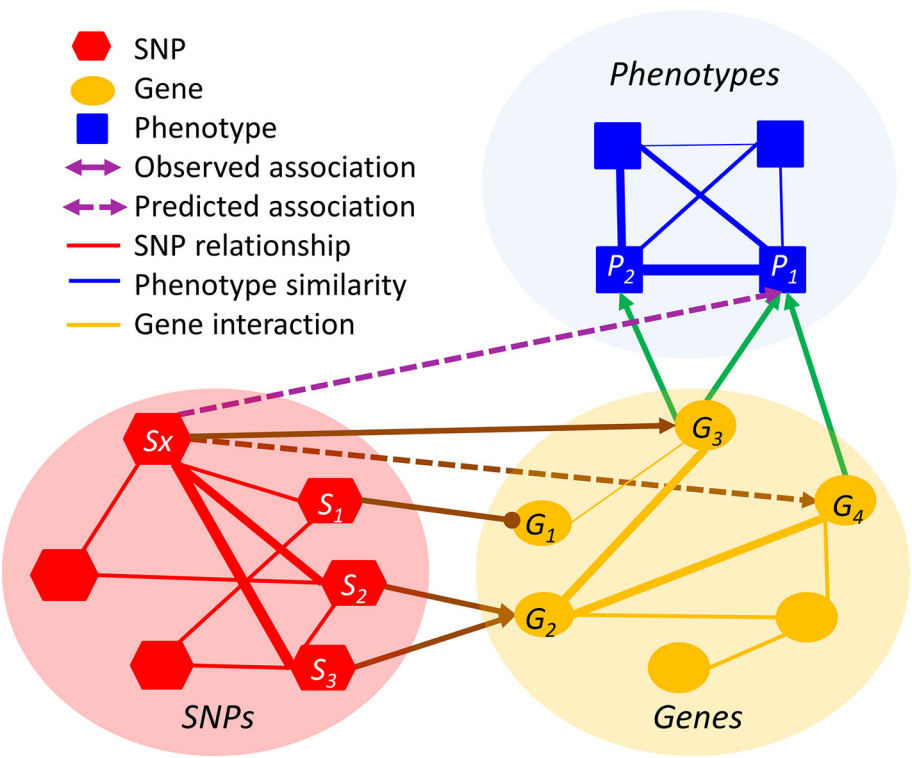

FIGURE 2 | An illustration of relation inference in the HMLN. The line thickness is proportional to the degree of relation. Arrowed and headed lines denote positive and negative relations, respectively.

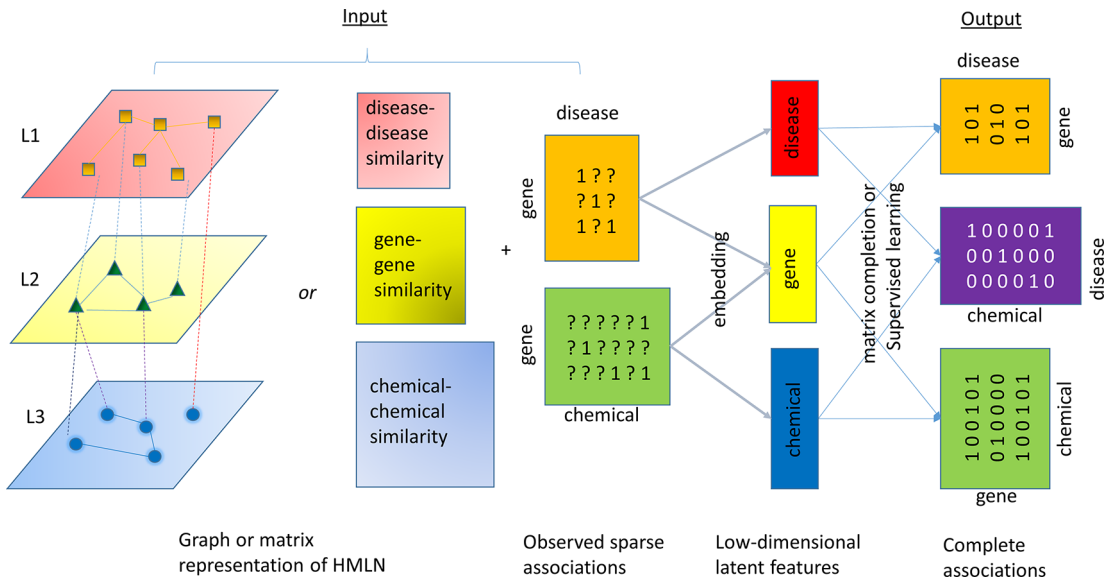

FIGURE 3 | An illustration of the common algorithmic framework for the relation inference (link prediction). HMLN is represented as a graph or collection of matrices. An inference algorithm takes the HMLN as input and generates a low-rank latent feature representation of chemicals, genes, and diseases, respectively. The inner product of latent features or supervised learning techniques will reconstruct complete gene-disease, chemical-disease, and chemical-gene association matrix.

Zitnik et al. developed a penalized matrix tri-factorization (PMTF) approach for data fusion (Zitnik and Zupan, 2015). Singh et al. proposed a collective matrix factorization (CMF) model to learn the dependencies across any two inter-dependent domains (Singh and Gordon, 2008). However, neither PMTF nor CMF takes the side information (i.e. intra-relations) into account. Moreover, both methods suffer the "cold-start" problem, which occurs when a new node arrives in the network.

Recently, Chen et al. developed the FASCINATE (Chen et al., 2016) algorithm to solve the multi-layered network inference problem, formulated as a weighted neighborhood-regularized collective one-class collaborative filtering problem. Mathematically, let $\mathbf{G}$ denotes a $g \times g$ layer-layer association matrix, where $\mathbf{G}(i, j) = 1$ if layer-j associates with layer-i, and $\mathbf{G}(i, j) = 0$ otherwise. Furthermore, let $\mathbf{A} = \{A_1, \ldots, A_g\}$ represents a set of g within-layer connectivity matrices that describe the connectivity/similarity between nodes within the same layer. Finally, denote by $\mathbf{D} = \{D_{i,j}\ i, j = 1, \ldots, g\}$ the set of cross-layer relation matrices, where $D_{i,j}$ specifies the relations between the nodes from layer i and the nodes from layer-j. (each relation is labeled 1, in case of an observed association; otherwise 0). The problem of inferring missing relations between layers is formulated as the following minimization problem:

$$\min_{F_i \geq 0 (i=1,\ldots,g)} J = \underbrace{\sum_{i,j:\, G(i,j)=1} \| W_{i,j} \odot \left(D_{i,j} - F_i F_j'\right) \|_F^2}_{\text{Matching observed cross-layer relations}} + \underbrace{\alpha \sum_{i=1}^{g} tr\left(F_i'(T_i - A_i)F_i\right)}_{\text{Node homophily}} + \underbrace{\beta \sum_{i=1}^{g} \| F_i \|_F^2}_{\text{Regularization}} \quad (1)$$

In the above loss function, $W_{i,j}$ denotes an $n_i \times n_j$ weight matrix that assigns different weights to different associations in the corresponding cross-layer relation matrix $D_{i,j}$, depending on the confidence in $D_{i,j}$. The confidence scores are extracted from the existing databases (Jensen et al., 2009; Kuhn et al., 2012). The matrix F_i gives the low-rank representation for nodes in layer *i, while* T_i is the diagonal degree matrix of A_i. Overall, the first term in Eq. 1 is used to match all the cross-layer relations calibrated by the weight matrix $W_{i,j,}$. The second term ensures that the similar nodes have similar low-rank representations. The third term is included to help prevent over-fitting. The optimization problem defined in Eq. (1) is non-convex. Block coordinate descent method is applied to find a local optima (where each F_i naturally forms a 'block'). Furthermore, the second term in Eq. (1) allows us to address the cold-start problem (namely the scenario where the query node does not have any known cross-layer links with the existing nodes in the network) based on similarity information.

There are several limitations of the existing MF-based methods for HMLN. First, the linear reconstruction of the complete matrix may not capture the complex cross-layer relations that are often non-linear. Deep neural network (DNN) has enjoyed great success in two-layered recommender system (Batmaz et al., 2019). Thus, it is interesting and tempting to extend the application of DNN to model the HMLN. Second, multiple types of links are often needed to model various biological relations between two layers. For example, there are three types of links between 'gene' and 'disease' in HetioNet: 'down-regulate', 'up-regulate', and 'associate'. And, while 'down-regulate' and 'up-regulate' are mutually exclusive, 'associate' is ambiguous (could be either 'down-regulate' or 'up-regulate'). Few

of the existing MF-based methods can handle such multi-type relations. Finally, the scalability might become an issue when the existing implementations of MF are applied to extremely large matrices. A distributed variant of MF could alleviate the problem.

Random Walk

Network propagation algorithm has been widely used in network biology (Cowen et al., 2017). Majority of applications of network propagation to biology networks are formulated in a homogeneous setting. For example, Lin et al. constructed a disease-gene-chemical network by integrating multiple data resources and then applied several homogenous network propagation algorithms for the relation inference (Lin et al., 2019). The random walk with restart (RWR) is one of the most representative network propagation algorithms. It was first developed to explore the global topology of networks, by simulating a particle that iteratively moves from a node to a randomly selected neighboring node (Lovasz, 1993). Only recently, random walk model has been extended to HMLN by allowing jumps across layers (Valdeolivas et al., 2019).

Consider an undirected graph, $G = (V, E)$ with adjacency matrix A. An imaginary particle starts a random walk at the initial node $v_0 \in V$. At a discrete time step $t \in N$, the particle is at node v_t. Then, it walks from v_t to v_{t+1}, a randomly selected neighbor of v_t, by following the transition matrix M calculated from A *via* column normalization (Lovasz, 1993). Probabilistically, $\forall x,y \in V$, $\forall t \in N$

$$P(v_{t+1} = y | v_t = x) = \begin{cases} 1/d(x) & \text{if} \quad (x,y) \in E \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where $d(x)$ is the degree of x in the graph G. The probability distribution of random walk at time $t+1$ is described by the following equation:

$$P_{t+1}^{T} = MP_t^{T} \tag{3}$$

Accounting for the restart probability r on the seed node to avoid the particle's dead-end, the random walk with restart (RWR) can be reformulated as:

$$P_{t+1}^{T} = (1-r)^{*}MP_t^{T} + r^{*}P_0^{T} \tag{4}$$

Even a multiplex graph with the collection of L undirected graphs can be formulated as a RWR problem (De Domenico et al., 2013; Kivelä et al., 2014). Each layer $\alpha = 1,\ldots,L$, can be represented by an n-by-n adjacency matrix $A^{[\alpha]} = (A^{[\alpha]}(i,j))_{i,j=1,\ldots,n}$, where $A^{[\alpha]}(i,j) = 1$, if nodes i and j are connected in layer α, and 0 otherwise (Battiston et al., 2014). The multiplex graph is defined as $G_M = (V_M, E_M)$, where:

$$V_M = \{v_i^{\alpha}, i = 1,\ldots,n, \alpha = 1,\ldots L\}, \tag{5}$$

where v_i^{α} stands for node i in layer α, and

$$E_M = \left\{ \left(v_i^{\alpha}, v_j^{\alpha}\right),\ i,j = 1,\ldots,n,\ \alpha = 1,\ldots L,\ A^{[\alpha]}(i,j) \neq 0 \right\}$$
$$\cup \left\{ \left(v_i^{\alpha}, v_i^{\beta}\right), i = 1,\ldots,n,\ \alpha \neq \beta \right\}. \tag{6}$$

The particle can walk from its current node v_i^{α} to any of its neighbors within a layer, or jump to any node v_i^{β} with $\alpha \neq \beta$ (De Domenico et al., 2013), and thereby travel from one layer to another, as shown in **Figure 1C**.

Extending classical RWR algorithm to a multiplex graph introduces a supra-adjacency matrix A of size $nL^{*}nL$, which contains different types of transitions:

$$A = \begin{pmatrix} (1-\delta)A^{[1]} & \frac{\delta}{(L-1)}I & \cdots & \frac{\delta}{(L-1)}I \\ \frac{\delta}{(L-1)}I & (1-\delta)A^{[2]} & \cdots & \frac{\delta}{(L-1)}I \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\delta}{(L-1)}I & \frac{\delta}{(L-1)}I & \cdots & (1-\delta)A^{[L]} \end{pmatrix} \tag{7}$$

In (7), I is the n-by-n identity matrix and $A^{[\alpha]}$ is the adjacency matrix of the layer α, as previously described. The diagonal elements represent potential intra-layer walks, whereas the off-diagonal elements account for possible jumps between different layers. The parameter $\delta \in [0,1]$ quantifies the probability of staying in the current layer or jumping to another layer. If $\delta = 0$, the particle will stay in the same layer after a non-restart step.

Topological features of each node or edge derived from the RW algorithm can be directly applied to link prediction. Those features are often used as the basis of the more sophisticated node embedding algorithms, such as DeepWalk (Perozzi et al., 2014), Node2Vec (Grover and Leskovec, 2016), etc. However, these algorithms focus on the homogenous network and have not been extended to HMLN yet.

One of major limitations of the network propagation algorithm is that its performance strongly depends on the topology of the input network. It is less tolerant to biasness, noisiness, and incompleteness of the network, which are the characteristics of reconstructed biological HMLN.

Meta-Path-Based Algorithms

Meta-path has been extensively studied in heterogeneous information networks (HIN) (Sun and Han, 2013). Since HMLN is a variant of HIN, the meta-path algorithm, described here, can be applied to the relation inference problem for HMLN. Given a directed graph representation: $G = (V, E)$ of HIN, an object type mapping function $\tau: V \rightarrow A$ and a link type mapping function $\varphi: E \rightarrow R$ are defined such that object $v \in V$ belongs to one particular object type $\tau(v) \in A$ and each link $e \in E$ belongs to a particular relation $\varphi(e) \in R$. A meta-path in G is a sequence of relations $R_1, \ldots, R_l$, which connect two object types A_i and A_j. In the example of **Figure 2**, the relation types include SNP-associate-Phenotype (SaP), Chemical-associate-Gene (CaG), and Gene-associate-Phenotype (GaP), SNP-similar-SNP (SsS), Phenotype-similar-Phenotype (PsP), and Gene-similar-Gene (GsG). The SNP-Phenotype association between S_x and P_1 can be defined by multiple meta-paths, e.g., SaG->GaP, SaG->GsG->GaP, and SsS->SaG->GaP, etc. By systematically designing meta-path based topological features and their measures in HLMN, supervised models can be used to learn the best weights associated with different topological features for

effective relation inference (Sun et al., 2012). In general, for a target relation $<A_i, A_j>$, any meta-path starting with type A_i and ending with type A_j (other than the target relation itself) can be used as a topological feature. All these meta-paths can be obtained by traversing on the network schema, for example, using the breadth first search. Most algorithms for HIN reconstruction enumerate a predefined set of meta-paths. Once all meta-paths are defined, the next task is to design measures on their topology. The commonly used measures include the count of the path instances and the random walk-based measures. Using topological features, either a supervised or unsupervised learning model is used for node representation. For example, the metapath2vec method (Dong et al., 2017) uses a meta-path-based random walk to form the heterogeneous neighborhood of a node, taking advantage of word representation algorithm in the Nature Language Processing to perform node embedding (Dong et al., 2017). One of the drawbacks of these algorithms is that they require manual predefinition and enumeration of meta-paths. This may be not feasible for schema-rich HMLN or the relations that involve multiple hopping paths (Cao et al., 2017), e.g. relations inferred through thousands of similar chemicals.

Graph Neural Network and Other Deep Learning Techniques

Besides the traditional algorithms, like matrix factorization, random walk, and meta-path, introduced in previous sections, the embedding of HMLN can also benefit from Deep Learning techniques, especially the Neural Networks (NNs). Though NNs are initially proposed to learn the embedding of data, such as texts, images, and videos, they have shown powerful performance when dealing with graph structured data, which exist in non-Euclidean domain. Due to the growing interests and demands in recent years, Graph Neural Networks (GNNs) have been proposed to learn the embedding of graphs (Li et al., 2015; Scarselli et al., 2008; Duvenaud et al., 2015; Kipf and Welling, 2017; Hamilton et al., 2017; Zhang et al., 2018; Ying et al., 2018; Morris et al., 2019; Xu et al., 2019; Zhang and Xie, 2019).

A GNN consists of a number of hidden layers that employ iterative, propagation procedures in order to transform different node and edge features. Each layer takes the output of the previous layer as the input. With graph structured data, GNNs adopt element (node or edge) features X and the graph structure A as input to learn the representation of each element h_i, or graph h_G, for different tasks. Each hidden layer employs the "aggregation" functions and the "update" functions (Battaglia et al., 2018). Each aggregation function ρ takes a set of node or edge features as input and reduces it to a single element which represents the aggregated information. The aggregations usually operate on the nearest neighbors or the local subgraphs of each element to capture local information gradually. Since the permutation invariance of the input holds in graph data, the ρ functions must also have the same property. These functions can take variable numbers of arguments. Commonly used ρ functions include sum (Xu et al., 2019), mean (Kipf and Welling, 2017), max-pooling (Hamilton et al., 2017) and attention mechanism (Velickovic et al., 2018; Wang et al., 2019; Fan et al., 2019). Update functions ϕ are applied across all elements to compute per-element updates after the aggregations. In the final layer, the generated embedding can be fed into the classification/prediction layer, and the whole model is trained for different (e.g. node classification, link prediction) tasks.

The design of GNNs is flexible. GNNs can be designed to fit different graph structures and different tasks. In the link prediction problem, the prediction of a feature (e.g. link or non-link) of a desired edge is based on the local structural information around that edge. For example, the method by Zhang et al. learns the link prediction heuristics from local (enclosing) subgraphs of edges rather than from the entire network (Zhang and Chen, 2018). The prediction of cross-layer relations follows a similar idea if HMLN is given as input. The model designed by Fan et al. learns the embedding of the two nodes by aggregating their neighbors (Fan et al., 2019). The embedding of two nodes is fed into a classification layer to classify the type of a given edge. Due to the topology of HMLN, GNNs can take meta-path into consideration when designing the aggregation functions ρ. In (Wang et al., 2019), the node embedding are computed by the neighbor nodes connected by meta-paths. During the training procedure, the effect of different meta-paths can be distinguished by using attention mechanism in aggregation. In (Zhang et al., 2018), the original input heterogeneous network is modified to be multi-channel network. Each channel is a homogeneous network consisting of the nodes that are connected by a similar type of meta-paths in the original network. Thus, GNNs can be used on each channel for learning the embedding, which is concatenated from all channels. As discussed in the previous section, the meta-path based GNN shares the same limitations of other meta-path based algorithms. New types of GNNs, those that explicitly take different types of relations into consideration, are needed for the link prediction problem in HMLN (Nathani et al., 2019).

Although Graph Neural Networks have been applied to heterogeneous networks and proven their ability of learning representations (Zhang et al., 2019), GNNs still exhibit limitations in several aspects. First, current GNNs proposed for the learning of heterogeneous networks do not particularly distinguish cross-layer from intra-layer relations. For example, while researchers can simply treat distinct relations as different types, the intra-layer relations in the same layer of an HMLN usually represent the similarity relation, which is semantically distinct from the cross-layer relation. The above needs to be taken into consideration when designing GNNs for HMLN. Second, the current design of GNNs relies on heuristics and empirical findings, which adds to the difficulty of learning the representations of HMLN. To enhance power of HMLN, it is crucial to properly identify the conditions that the aggregation and update functions ought to satisfy and to set those functions accordingly. Third, although GNNs can achieve promising results on different tasks for heterogeneous networks, it is hard for GNNs to have interpretability comparing to other traditional techniques. Therefore, new methods are needed to handle the

problems that need interpretability (e.g. the need to find important nodes or edges that contribute to the results).

Application of HMLN in Omics Data Integration and Analysis

Homogeneous and bi-layered network models have been widely applied in omics data integration and analysis. Recently, the HMLN has emerged as a powerful alternative. Here, we will highlight several exemplary applications of the HMLN to infer genotype-phenotype associations, and to predict chemical and other environmental perturbations.

Yao et al. integrated multi-omics data to construct a three-layered network model MetPriCNet, which consists of metabolite network, gene network, phenotype network, metabolite-phenotype network, metabolite-gene network, and gene-phenotype network (Yao et al., 2015). Afterwards, an RWR algorithm is applied to prioritize metabolites associated with diseases. The cross-validation on a benchmarking data set achieved the AUC values exceeding 0.9. An approach similar to MetPriCNet has been applied to identify and prioritize the metabolites responsible for atrial fibrillation (Yan et al., 2019), postmenopausal osteoporosis (Zhang et al., 2019), and Acute Lung Injury in Patients with Sepsis (Wang et al., 2019).

In addition to metabolite-disease association, the RWR method has been used to identify other molecular dysregulations that are associated with diseases based on the multi-layered network model. To infer disease associated m^6A RNA methylation site, Tang et al. constructed a three-layered network, that includes a m^6A site network, a gene network, a disease network, a m^6A-gene network, and a gene-disease network (Tang et al., 2019). Xu and Wang applied random walk on a three-layer heterogeneous network that uses a kinase layer as an intermediate to infer disease-phosphorylation site relation. They showed that the three-layer phosphorylation site-kinase-disease network model is superior in inferring disease-phosphorylation site relation when compared with the existing random walk models and commonly used classification methods (Xu and Wang, 2016).

HMLN provides new opportunities for inferring novel drug-target-pathway-disease-side effect associations. The identification of such missing relations could facilitate the discovery of new therapies for complex diseases.

The ANTENNA method by Wang et al. employs a one-class collaborative filtering technique based on RWR and the matrix tri-factorization to predict the drug-disease associations using a three-layered drug-gene-disease network. In a comprehensive benchmarking study, ANTENNA outperformed the more conventional OCCF methods. Using ANTENNA, Wang et al. showed that diazoxide might inhibit the growth of triple negative breast cancer (TNBC) cells efficiently (Wang et al., 2018). Lim et al. applied FASCINATE to a three-layered drug-gene-side effect network model to identify biological pathways associated with rare side effects. Their predicted side effect-causing pathways are consistent with clinical evidences (Lim et al., 2018). Fu et al. extracted meta-path based topological features from a semantic network with nine object types (compound, ChEBI type, chemical substructure, protein, GO annotation, pathway, tissue, disease, and side effect), and twelve relation types. Subsequently, they applied the extracted features to predict drug-target interactions (Fu et al., 2016).

FUTURE DIRECTIONS

Representation of Biological Hierarchy and Environment

Only a few multi-layered network models for the genotype-phenotype associations have been developed that consist of more than three layers. These models lack the power to represent the full spectrum of information flow from the genotype to the phenotype. Even in a simplified picture, a multi-layered network model needs more than three layers to connect genome to phenome *via* epigenome (DNA layer), transcriptome (RNA layer), proteome (protein layer), and metabolome (metabolite layer). The representation of DNA, RNA, protein, and metabolite in the different layer could facilitate heterogeneous omics data integration and multi-scale modeling of information flow from genotype to phenotype. Furthermore, environmental components, such as gut microbiome, play a critical role in shaping the organismal phenotypes. With the exponential growth of different omics data from the same cohorts [e.g. TCGA (Cancer Genome Atlas Research Network et al., 2013)], the multi-layered network model represents a potentially powerful tool to integrate and analyze heterogeneous data sets for novel discovery.

Incorporation of Mechanism-Based Modeling

The capability of data-driven modeling is limited by the existing data. We can enrich the missing relations in HMLN using complementary methods. For example, text mining is a commonly used tool to construct HMLN. Besides alternative machine learning approaches, mechanism-based modeling in biophysics, systems biology, and other fields can be applied to establish causal relations between entities. For example, protein-ligand docking can be applied to infer chemical-protein interactions. The mechanism argument HMLN may provide us with new opportunities for novel discovery, as demonstrated in a recent study (Lim et al., 2019). However, the potential false positives from the outside predictions should be taken into consideration when designing HMLN learning procedures.

Data Consolidation and Normalization

When reconstructing HMLN, both intra-layer and cross-layer relations can come from multiple resources. For instance, in HumanNet, gene-gene co-functional links are derived from co-citation, co-essentiality, co-expression, pathway database, protein-domain profile association, and gene neighborhood (Hwang et al., 2019). Another example is chemical-protein interaction (Gaulton et al., 2012). The binding assay could be performed using different experimental techniques, and measured by different metrics (IC50, pKi, etc.). However, mapping the entities, minimizing batch effects, and normalizing the weights of different edge types in the same network remain the challenging tasks.

Inference of Directionality and Trend of Relations

Few of relation inference algorithms can predict the directionality and trend of edges. The directionality means that one entity has effect on another but not vice versa. The trend represents distinct and often opposite functional consequence. For example, a drug can down- or up-regulate a gene. The identification of the directionality and trend of relation is pivotal to understand many biological processes such as drug action, signaling transduction and gene regulation, and determine causality between biological entities. For example, knowing that a chemical *C* interacts with a gene *G*, which is associated with a disease *D*, does not necessarily imply that the compound *C* will be effective on the disease *D*. On the other hand, if the compound *C* up-regulates the gene *G*, and the gene *G* is down-regulated in the disease *D*, than it is more likely for the compound *C* to treat or palliate the disease *D*. Recent development on signed network algorithm may provide partial solution to this problem (Kim et al., 2018).

Inference of Non-Binary and Dynamic Relations

Existing link prediction algorithms for HMLN mainly focus on binary relations. However, other types of relations, such as unary and higher-arity relations, are needed to encode more complete biological knowledge. The unary relation represents the property of an entity, for example, the expression value of a gene. When modeling a dynamic system, a relation is associated with time and location. A single binary relation is not sufficient to capture its temporal and spatial nature. In this case, the higher-arity relations might prove beneficial. An example of a ternary relation is "gene *A* with a mutation *M* down-regulates the expression of gene *B* in neuro cells". This relation includes three entities or layers (mutation, gene, and cell), and it can be expressed by three binary relations: "Mutation *M* is in gene A", "drug *A* down-regulates gene *B*", and "gene *B* is expressed in the neuro cells". However, the genes *A* and *B* might be expressed in other types of cells in addition to the neuro cells. The mutation *M* in gene *A* may not down-regulate gene *B* in other cells. As a result, the tissue-specific correspondence between mutation *M* and the neuro cell is lost.

Incorporation of Ontology

A number of biomedical ontologies have been developed to facilitate knowledge integration and discovery (Musen et al., 2012). These ontologies can serve as HMLN constraints to reduce false positives and resolve contradictory relations. There are two types of ontology constraints that can be applied to HMLN, namely deterministic constraint and functional constraint. A deterministic constraint imposes a clear dependency on relations such as "IsA" and "LocatedIn". For example, if a protein binds to zinc, it is safe to state that the protein is metal-binding, because zinc is a metal. One can precompute all relations derived from the deterministic constraint and add them to HMLN prior to learning. Functional constraints enforce mutual exclusiveness between possible values. For example, if a chemical *A* is a known inhibitor of enzyme *B*, one can exclude the relation "*A* activates *B*" from HMLN.

Sampling of Negative Relations

Many learning algorithms need a balanced number of negative examples. As mentioned in section 2, there are much less negative examples than positive examples in the biological HMLN, although, in reality, the negative cases are substantially more frequent than the positive ones. The conventional method is to randomly sample from a uniform distribution after excluding positive examples. However, this approach may not be applicable to the biological HMLN, where opposite relations exist between two entities. For example, a drug can either "down-regulate" or "up-regulate" a gene. It is not obvious how to assign the sampled relations to the opposites of "down-regulate" or "up-regulate". Cai et al. have recently developed an adversarial reinforcement learning framework to assign the negative samples (Cai and Wang, 2017). This approach can be extended to the negative sampling for different relation types in HMLN.

Visualizing HMLN

Visualization plays a key role in data mining tasks. Although many computational platforms, such as Cytoscape (Shannon et al., 2003), have been developed for the network visualization, few tools are available for efficient and intuitive visualization of HMLN, especially when the network is large (Mcgee et al., 2019). There is an urgent need to design a robust data structure for the representation and grouping of nodes and relations in HMLN in a way that they can be efficiently mapped to the graphic user interface and easily navigated by users.

AUTHOR CONTRIBUTIONS

BL, SZ, AP, and LX wrote the manuscript.

FUNDING

This work was supported by Grant Number R01LM011986 from the National Library of Medicine (NLM), Grant Number R01GM122845 from the National Institute of General Medical Sciences (NIGMS), and Grand Number R01AD057555 of National Institute on Aging on the National Institute of Health (NIH). The funding agencies do not play roles in writing this manuscript.

REFERENCES

Batmaz, Z., Yurekli, A., Bilge, A., and Kaleli, C. (2019). A review on deep learning for recommender systems: challenges and remedies. *Artif. Intell. Rev.* 52, 1–37. doi: 10.1007/s10462-018-9654-y

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., et al. (2018). Relational inductive biases, deep learning, and graph networks. arXiv preprint arXiv:1806.01261.

Battiston, F., Nicosia, V., and Latora, V. (2014). Structural measures for multiplex networks. *Phys. Rev. E* 89, 032804. doi: 10.1103/PhysRevE.89.032804

Breese, J. S., Heckerman, D., and Kadie, C. (1998). "Empirical analysis of predictive algorithms for collaborative filtering" in *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence* (Morgan Kaufmann Publishers Inc.: Burlington, MA), 43–52.

Cai, L., and Wang, W. Y. (2017). Kbgan: Adversarial learning for knowledge graph embeddings. arXiv preprint arXiv:1711.04071.

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

Cao, X., Zheng, Y., Shi, C., Li, J., and Wu, B. (2017). Meta-path-based link prediction in schema-rich heterogeneous information network. *Int. J. Data Sci. Analytics* 3, 285–296. doi: 10.1007/s41060-017-0046-1

Chauvel, C., Novoloaca, A., Veyre, P., Reynier, F., and Becker, J. (2019). Evaluation of integrative clustering methods for the analysis of multi-omics data. *Brief Bioinform*. doi: 10.1093/bib/bbz015

Chen, C., Tong, H., Xie, L., Ying, L., and He, Q. (2016). "FASCINATE: Fast Cross-Layer Dependency Inference on Multi-layered Networks," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA: ACM), 765–774.

Cowen, L., Ideker, T., Raphael, B. J., and Sharan, R. (2017). Network propagation: a universal amplifier of genetic associations. *Nat. Rev. Genet.* 18, 551–562. doi: 10.1038/nrg.2017.38

De Domenico, M., Solé-Ribalta, A., Cozzo, E., Kivelä, M., Moreno, Y., Porter, M. A., et al. (2013). Mathematical formulation of multilayer networks. *Phys. Rev. X* 3, 041022. doi: 10.1103/PhysRevX.3.041022

Dong, Y., Chawla, N. V., and Swami, A. (2017). "metapath2vec: Scalable representation learning for heterogeneous networks," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* 135–144.

Duvenaud, D. K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T., Aspuru-Guzik, A., et al. (2015). "Convolutional networks on graphs for learning molecular fingerprints," in *Advances in neural information processing systems*, 2224–2232.

Fan, W., Ma, Y., Li, Q., He, Y., Zhao, E., Tang, J., et al. (2019). "Graph Neural Networks for Social Recommendation," in *The World Wide Web Conference*, 417–426.

Fu, G., Ding, Y., Seal, A., Chen, B., Sun, Y., and Bolton, E. (2016). Predicting drug target interactions using meta-path-based semantic network analysis. *BMC Bioinf.* 17, 160. doi: 10.1186/s12859-016-1005-x

Gao, Y. L., Cui, Z., Liu, J. X., Wang, J., and Zheng, C. H. (2019). NPCMF: Nearest Profile-based Collaborative Matrix Factorization method for predicting miRNA-disease associations. *BMC Bioinf.* 20, 353. doi: 10.1186/s12859-019-2956-5

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., et al. (2012). ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 40, D1100–D1107. doi: 10.1093/nar/gkr777

Goldberg, D., Nichols, D., Oki, B. M., and Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Commun. ACM* 35, 61–70. doi: 10.1145/138859.138867

Grover, A., and Leskovec, J. (2016). "node2vec: Scalable feature learning for networks" in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 855–864.

Hamilton, W., Ying, Z., and Leskovec, J. (2017). "Inductive Representation Learning on Large Graphs," in *Adv. In Neural Inf. Process. Syst.*, 1024–1034.

Himmelstein, D. S., and Baranzini, S. E. (2015). Heterogeneous Network Edge Prediction: A Data Integration Approach to Prioritize Disease-Associated Genes. *PloS Comput. Biol.* 11, e1004259. doi: 10.1371/journal.pcbi.1004259

Hwang, S., Kim, C. Y., Yang, S., Kim, E., Hart, T., Marcotte, E. M., et al. (2019). HumanNet v2: human gene networks for disease research. *Nucleic Acids Res.* 47, D573–D580. doi: 10.1093/nar/gky1126

Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., et al. (2009). STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.* 37, D412–D416. doi: 10.1093/nar/gkn760

Kim, J., Park, H., Lee, J.-E., and Kang, U. (2018). "Side: representation learning in signed directed networks," in *Proceedings of the 2018 World Wide Web Conference, International World Wide Web Conferences Steering Committee*, 509–518.

Kipf, T. N., and Welling, M. (2017). "Semi-Supervised Classification with Graph Convolutional Networks," in *International Conference on Learning Representations*.

Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., and Porter, M. A. (2014). Multilayer networks. *J. Complex Networks* 2, 203–271. doi: 10.1093/comnet/cnu016

Koren, Y., Bell, R., and Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer* 42, 30–37. doi: 10.1109/MC.2009.263

Kringelum, J., Kjaerulff, S. K., Brunak, S., Lund, O., Oprea, T. I., and Taboureau, O. (2016). ChemProt-3.0: a global chemical biology diseases mapping. *Database*. doi: 10.1093/database/bav123

Kuhn, M., Szklarczyk, D., Franceschini, A., von Mering, C., Jensen, L. J., and Bork, P. (2012). STITCH 3: zooming in on protein-chemical interactions. *Nucleic Acids Res.* 40, D876–D880. doi: 10.1093/nar/gkr1011

Li, Y., Tarlow, D., Brockschmidt, M., and Zemel, R. (2015). Gated graph sequence neural networks. arXiv preprint arXiv:1511.05493.

Li, S., Sullivan, N. L., Rouphael, N., Yu, T., Banton, S., Maddur, M. S., et al. (2017). Metabolic Phenotypes of Response to Vaccination in Humans. *Cell* 169, 862–877 e17. doi: 10.1016/j.cell.2017.04.026

Lim, H., Poleksic, A., and Xie, L. (2018). Exploring Landscape of Drug-Target-Pathway-Side Effect Associations. *AMIA Jt Summits Transl. Sci. Proc.* 2017, 132–141.

Lim, H., He, D., Qiu, Y., Krawczuk, P., Sun, X., and Xie, L. (2019). Rational Discovery of Dual-Indication Multi-Target PDE/Kinase Inhibitor for Precision Anti-Cancer Therapy Using Structural Systems Pharmacology. *PloS Comp. Biol.* 15, e1006619. doi: 10.1371/journal.pcbi.1006619

Lin, C. H., Konecki, D. M., Liu, M., Wilson, S. J., Nassar, H., Wilkins, A. D., et al. (2019). Multimodal network diffusion predicts future disease-gene-chemical associations. *Bioinformatics* 35, 1536–1543. doi: 10.1093/bioinformatics/bty858

Lovasz, L. (1993). Random walks on graphs: a survey. *Bolyayi Soc. Math. Stud.* 2, 1–46. doi: 10.1.1.39.2847

Mcgee, F., Ghoniem, M., Melançon, G., Otjacques, B., and Pinaud, B. (2019). The state of the art in multilayer network visualization. *Computer Graphics Forum* 38, 125–149.

Morris, C., Ritzert, M., Fey, M., Hamilton, W. L., Lenssen, J. E., Rattan, G., et al. (2019). "Weisfeiler and Leman Go Neural: Higher-order Graph Neural Networks," in *Proceedings of AAAI Conference on Artificial Inteligence*.

Musen, M. A., Noy, N. F., Shah, N. H., Whetzel, P. L., Chute, C. G., Story, M. A., et al. (2012). The National Center for Biomedical Ontology. *J. Am. Med. Inform Assoc.* 19, 190–195. doi: 10.1136/amiajnl-2011-000523

Nathani, D., Chauhan, J., Sharma, C., and Kaul, M. (2019). Learning Attention-based Embeddings for Relation Prediction in Knowledge Graphs. arXiv preprint arXiv:1906.01195.

Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). "Deepwalk: online learning of social representations," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 701–710.

Pinero, J., Bravo, A., Queralt-Rosinach, N., Gutierrez-Sacristan, A., Deu-Pons, J., Centeno, E., et al. (2017). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* 45, D833–D839. doi: 10.1093/nar/gkw943

Portugal, I., Alencar, P., and Cowan, D. (2018). The use of machine learning algorithms in recommender systems: a systematic review. *Expert Syst. Appl.* 97, 205–227. doi: 10.1016/j.eswa.2017.12.020

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. (2008). The graph neural network model. *IEEE Trans. Neural Networks* 20, 61–80. doi: 10.1109/TNN.2008.2005605

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

Silk, M. J., Finn, K. R., Porter, M. A., and Pinter-Wollman, N. (2018). Can Multilayer Networks Advance Animal Behavior Research? *Trends Ecol. Evol.* 33, 376–378. doi: 10.1016/j.tree.2018.03.008

Singh, A. P., and Gordon, G. J. (2008). "Relational learning *via* collective matrix factorization" in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 650–658.

Sun, Y., and Han, J. (2013). Mining heterogeneous information networks: a structural analysis approach. *ACM Sigkdd Explor. Newslett.* 14, 20–28. doi: 10.1145/2481244.2481248

Sun, Y., Han, J., Aggarwal, C. C., and Chawla, N. V. (2012). "When will it happen?: relationship prediction in heterogeneous information networks" in *Proceedings of the fifth ACM international conference on Web search and data mining*, 663–672.

Tang, Y., Chen, K., Wu, X., Wei, Z., Zhang, S. Y., Song, B., et al. (2019). DRUM: Inference of Disease-Associated m(6)A RNA Methylation Sites From a Multi-Layer Heterogeneous Network. *Front. Genet.* 10, 266. doi: 10.3389/fgene.2019.00266

Valdeolivas, A., Tichit, L., Navarro, C., Perrin, S., Odelin, G., Levy, N., et al. (2019). Random Walk with Restart on Multiplex and Heterogeneous Biological Networks. *Bioinformatics* 35, 497–505. doi: 10.1093/bioinformatics/bty637

Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. (2018). "Graph Attention Networks," in *International Conference on Learning Representations*.

Wang, A., Lim, H., Cheng, S.-Y., and Xie, L. (2018). ANTENNA, a Multi-Rank, Multi-Layered Recommender System for Inferring Reliable Drug-Gene-Disease Associations: Repurpsing Diazoxide as an effective targeted anti-cancer therapy. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 15, 1960–1967. doi: 10.1109/TCBB.2018.2812189

Wang, X., Ji, H., Shi, C., Wang, B., Ye, Y., Cui, P., et al. (2019a). "Heterogeneous Graph Attention Network," in *The World Wide Web Conference*, 2022–2032.

Wang, P. Q., Li, J., Zhang, L. L., Lv, H. C., and Zhang, S. H. (2019b). Identification of Key Metabolites for Acute Lung Injury in Patients with Sepsis. *Iran J. Public Health* 48, 77–84.

Xie, L., and Bourne, P. E. (2005). Functional coverage of the human genome by existing structures, structural genomics targets, and homology models. *PloS Comp. Biol.* 1, e31. doi: 10.1371/journal.pcbi.0010031

Xie, L., Draizen, E., and Bourne, P. E. (2017). Harnessing Big Data for Systems Pharmacology. *Annu. Rev. Pharmacol. Toxicol.* 57, 245–262. doi: 10.1146/annurev-pharmtox-010716-104659

Xu, X., and Wang, M. (2016). Inferring Disease Associated Phosphorylation Sites *via* Random Walk on Multi-Layer Heterogeneous Network. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 13, 836–844. doi: 10.1109/TCBB.2015.2498548

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. (2019). "How Powerful are Graph Neural Networks?" in *International Conference on Learning Representations*.

Xuan, P., Cao, Y., Zhang, T., Wang, X., Pan, S., and Shen, T. (2019). Drug repositioning through integration of prior knowledge and projections of drugs and diseases. *Bioinformatics* 35, 4108–4119. doi: 10.3892/etm.2019.7443

Yan, Z. T., Huang, J. M., Luo, W. L., Liu, J. W., and Zhou, K. (2019). Combined metabolic, phenomic and genomic data to prioritize atrial fibrillation-related metabolites. *Exp. Ther. Med.* 17, 3929–3934. doi: 10.3892/etm.2019.7443

Yao, Q., Xu, Y., Yang, H., Shang, D., Zhang, C., Zhang, Y., et al. (2015). Global Prioritization of Disease Candidate Metabolites Based on a Multi-omics Composite Network. *Sci. Rep.* 5, 17201. doi: 10.1038/srep17201

Ying, Z., You, J., Morris, C., Ren, X., Hamilton, W., and Leskovec, J. (2018). "Hierarchical graph representation learning with differentiable pooling," in *Advances in Neural Information Processing Systems*, 4805–4815.

Zhang, M., and Chen, Y. (2018). "Link prediction based on graph neural networks," in *Advances in Neural Information Processing Systems*, 5165–5175.

Zhang, S., and Xie, L. (2019). Improving Attention Mechanism in Graph Neural Networks *via* Cardinality Preservation. arXiv preprint arXiv:1907.02204.

Zhang, M., Cui, Z., Neumann, M., and Chen, Y. (2018a). "An end-to-end deep learning architecture for graph classification," in *Proceedings of AAAI Conference on Artificial Inteligence*.

Zhang, Y., Xiong, Y., Kong, X., Li, S., Mi, J., and Zhu, Y. (2018b). "Deep collective classification in heterogeneous information networks," in *Proceedings of the 2018 World Wide Web Conference*, 399–408.

Zhang, C., Song, D., Huang, C., Swami, A., and Chawla, N. V. (2019a). "Heterogeneous graph neural network," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 793–803.

Zhang, C., Wang, Y., Zhang, C. L., and Wu, H. R. (2019b). Prioritization of candidate metabolites for postmenopausal osteoporosis using multi-omics composite network. *Exp. Ther. Med.* 17, 3155–3161. doi: 10.3892/etm.2019.7310

Zitnik, M., and Zupan, B. (2015). Data Fusion by Matrix Factorization. *IEEE Trans. Pattern Anal. Mach. Intell.* 37, 41–53. doi: 10.1109/TPAMI.2014.2343973

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: 07 February 2020
doi: 10.3389/fgene.2020.00008

Inferring Regulatory Networks From Mixed Observational Data Using Directed Acyclic Graphs

Wujuan Zhong[1†], Li Dong[1†], Taylor B. Poston[2], Toni Darville[2], Cassandra N. Spracklen[3], Di Wu[1,4], Karen L. Mohlke[3], Yun Li[1,3], Quefeng Li[1] and Xiaojing Zheng[1,2*]*

[1] Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States, [2] Department of Pediatrics, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States, [3] Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States, [4] Department of Oral and Craniofacial Health Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

OPEN ACCESS

Edited by:
Shizhong Han, Johns Hopkins Medicine, United States

Reviewed by:
Jian Li, Tulane University, United States
Kui Zhang, Michigan Technological University, United States

***Correspondence:**
Quefeng Li
quefeng@email.unc.edu
Xiaojing Zheng
xiaojinz@email.unc.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: *11 October 2019*
Accepted: *06 January 2020*
Published: *07 February 2020*

Citation:
Zhong W, Dong L, Poston TB, Darville T, Spracklen CN, Wu D, Mohlke KL, Li Y, Li Q and Zheng X (2020) Inferring Regulatory Networks From Mixed Observational Data Using Directed Acyclic Graphs. Front. Genet. 11:8. doi: 10.3389/fgene.2020.00008

Construction of regulatory networks using cross-sectional expression profiling of genes is desired, but challenging. The Directed Acyclic Graph (DAG) provides a general framework to infer causal effects from observational data. However, most existing DAG methods assume that all nodes follow the same type of distribution, which prohibit a joint modeling of continuous gene expression and categorical variables. We present a new mixed DAG (mDAG) algorithm to infer the regulatory pathway from mixed observational data containing both continuous variables (e.g. expression of genes) and categorical variables (e.g. categorical phenotypes or single nucleotide polymorphisms). Our method can identify upstream causal factors and downstream effectors closely linked to a variable and generate hypotheses for causal direction of regulatory pathways. We propose a new permutation method to test the conditional independence of variables of mixed types, which is the key for mDAG. We also utilize an L_1 regularization in mDAG to ensure it can recover a large sparse DAG with limited sample size. We demonstrate through extensive simulations that mDAG outperforms two well-known methods in recovering the true underlying DAG. We apply mDAG to a cross-sectional immunological study of *Chlamydia trachomatis* infection and successfully infer the regularity network of cytokines. We also apply mDAG to a large cohort study, generating sensible mechanistic hypotheses underlying plasma adiponectin level. The R package mDAG is publicly available from CRAN at https://CRAN.R-project.org/package=mDAG.

Keywords: regulatory network, directed acyclic graphs, mixed observational data, continuous and categorical variables, causal regulatory pathways

INTRODUCTION

Identification of differentially expressed genes associated with disease has become an instrumental approach, but with only limited success in mechanistic discovery, partly due to the fact that current methods based on fold-change focus only on a single gene. Co-expression network analysis (Oldham et al., 2006; Chen, 2012; Hawrylycz et al., 2012), an approach that constructs networks

of genes that tend to co-activate among a group of samples, provides a connectome of gene interaction. (Zhuang et al., 2016) proposes a more general class of undirected graphical models that can handle mixed types of variables. However, the undirected graphical model by itself cannot reveal disease causality. There is a critical need to understand regulatory pathways for discovery of therapeutic targets and disease mechanisms.

A few approaches have been proposed in recent years to estimate regulatory networks/pathways. iPoint was proposed by Atias and Sharan (2013) to infer a compact subnetwork that connects the source of the response (*anchor* genes) to the targets of the response (*terminal* genes) while optimizing local (individual path lengths) or global (likelihood) aspects of the subnetwork to solve the "anchor" reconstruction problem. The input of iPoint requires a single *anchor* gene and a list of *terminal genes*. PINE was proposed by Wilentzik and Gat-Viks (2015) to identify the particular pathways by which DNA variants perturb the signaling network. It requires prior established biological knowledge of how the stimulations affect gene expression and existence of multiple stimulation conditions. TieDie was proposed by Paull et al. (2013) to infer regulatory pathways linking genomic events (e.g. mutated genes) to transcriptional changes by a heat diffusion strategy. However, TieDie assumes that mutations necessarily lead to loss of function. All these methods assume prior knowledge of particular biological networks/pathways or functions.

Over the past few years, there has been a growing interest in utilizing directed acyclic graphs (DAG), which do not require any prior biological knowledge, to infer directional relations in a regulatory network in a large variety of disciplines such as biology, neuroscience, and psychology (Friedman et al., 2000; Huang et al., 2010; Borsboom and Cramer, 2013). The logical basis of such graphical models is the conditional independence structure of the underlying probability distributions of data. We propose to jointly model the probability distribution of mixed data composed of continuous variables (e.g., expression of proteins or genes) and discrete variables (e.g., categorical disease outcomes or single nucleotide polymorphisms) by DAG.

There are three types of methods to estimate a DAG (Nagarajan et al., 2013): constraint-based methods, score-based methods, and hybrid methods. The constraint-based methods learn a DAG by exploiting the conditional independence constraints in the observational distribution. The most prominent example of such methods is the PC algorithm (Spirtes et al., 2000). This algorithm first estimates the skeleton of the underlying DAG, and then adds orientations to the skeleton based on a set of edge orientation rules (Meek, 1995). The CPC-stable algorithm (Colombo and Maathuis, 2014) improves the PC algorithm by resolving the order-dependence issue in the determination of the skeleton. A more recent constraint-based method (Tsagris et al., 2018) proposes a symmetric conditional independence tests based on likelihood-ratio test and combines it with the existing constraint-based methods (e.g. PC algorithm) to estimate a DAG. The score-based methods (Chickering, 2002) learn a DAG by a greedy search for the optimal score of the goodness-of-fit of the estimated DAG. The hybrid methods (Nagarajan et al., 2013) learn a DAG by integrating the constraint-based and the score-based methods. An example is the Max-Min Hill-Climbing (MMHC) algorithm (Tsamardinos et al., 2006), which applies the Max-Min Parents and Children algorithm to obtain the skeleton and the Hill Climbing greedy search algorithm to orient edges in the skeleton. Another example is the causalMGM algorithm (Sedgewick et al., 2016; Sedgewick et al., 2017), which firstly estimates an undirected graph and then uses PC-stable or CPC-stable for orientation. The first step modifies the mixed graphical model method (Lee and Hastie, 2015) by using different penalty functions for different edge types. The second step uses a likelihood-ratio test to test the conditional independence in order to use the PC-stable or CPC-stable algorithm for edge orientation. Based on our experience, such an orientation method is not as efficient as score-based method, which is used in our algorithm.

However, most of these methods assume that all variables are of the same type. For example, the Gaussian graphic model assumes that the joint distribution of all variables is multivariate normal. Therefore, these methods cannot be directly applied to infer the causal relationship between continuous measurements, such as protein or gene expression, and the categorical variables, such as categorical traits or single nucleotide polymorphisms (SNPs). To this end, we propose a mixed DAG method (mDAG) that accommodates data of different types. We assume the joint distribution of all variables follow a pairwise Markov random field, which ensure that the conditional distribution of one graph node on all other nodes either follow a Gaussian distribution or a multinomial distribution. Thus, it enables joint modeling of continuous and categorical variables. We demonstrate the efficacy of our method through extensive simulations and apply it to a study of human cytokines associated with chlamydial susceptibility to infer cytokines with causal effects on a categorical disease phenotype. We also show that our method can identify gene expression levels that mediate the effect of genetic variants on traits.

MATERIALS AND METHODS

Definitions and Preliminaries

We first introduce a few key concepts in the DAG theory. A DAG of a vector of random variables $X = (X_{1,\ldots,}X_d)^T$ is a directed graph with no cycle, which is denoted by $G = (V, E)$, where V is the set of d vertices representing X, and E is the set of all directed edges. Given a path $X_{i_0} \rightarrow X_{i_1} \rightarrow \ldots \rightarrow X_{i_n}$ in a DAG, $X_{i_{l-1}}$ is called a parent of X_{i_l} and X_{i_l} is called a child of $X_{i_{l-1}}$. The d separation set S that blocks nodes i and j is a vertex set that blocks all paths that connect i and j for either the path that contains at least one arrow-emitting vertex belonging to S, or the path that contains at least one collision vertex (a vertex without emitting edges) that is outside S and no children of the collision vertex belongs to S. In a DAG, the Markov blanket of a node includes its parents, children, and the other parents of all

of its children. In an undirected graph, the Markov blanket of a node contains all nodes connecting to itself. The skeleton of a DAG is the undirected graph that results from ignoring the directionality of every edge in a DAG. In order to model the mixed data, we assume the joint distribution of all variables is faithful to a DAG, meaning that for any $i, j \in V$ and any set $S \subset V$, X_i and X_j are conditional independent given X_s if and only if node i and j are d-separated by set S (Pearl, 2009) and S is called the d-separation set of node i and j. In other words, the conditional independence can be read from the DAG. Under the faithfulness assumption, the joint distribution has the Markov property that a node is independent of all other nodes conditional on the Markov blanket. Such an assumption is widely used in Bayesian Network literature, the PC-algorithm (Spirtes et al., 2000), PC-stable and CPC-stable algorithm (Colombo and Maathuis, 2014), and MMHC algorithm (Tsamardinos et al., 2006). Meek (2013) proved that this assumption holds for a variety of Bayesian Network.

To recover the underlying DAG from the mixed data, our method consists of three main steps. First, we use a penalized nodewise maximum likelihood method (Lee and Hastie, 2015) to identify the Markov blanket of each node. Second, we use a modified PC-stable algorithm (Ha et al., 2016) to obtain the DAG's skeleton and its d-separation set. Finally, we add orientations to the skeleton using a greedy search algorithm (Tsamardinos et al., 2006). Different from the existing literature, since our data is of mixed types, we propose a new permutation test on the second step to test the conditional independence, which is the key to estimate the skeleton of the DAG for mixed data.

Identification of the Markov Blanket

We assume the distribution of $X = (X_1,\ldots,X_{p+q})^T$ follows a pairwise Markov random field with a density

$$p(x;\Theta) \propto \exp\left(\Sigma_{s=1}^{p}\Sigma_{t=1}^{p} -\frac{1}{2}\beta_{st}x_s x_t + \Sigma_{s=1}^{p}\alpha_s x_s + \Sigma_{s=1}^{p}\Sigma_{j=1}^{q}\rho_{sj}\left(x_{p+j}\right)x_s + \Sigma_{j=1}^{q}\Sigma_{r=1}^{q}\phi_{rj}\left(x_{p+j}, x_{p+r}\right)\right)$$

where we assume without loss of generality that $X_j (j = 1,\ldots p)$ are continuous variables, $X_{p+j} (j = 1,\ldots q)$ are discrete variables, and $\Theta = (\alpha_s, \beta_{st}, \rho_{sj}, \phi_{rj})$ for $s,t = 1,\ldots$ and $j,r = 1,\ldots q$ are parameters. We assume that the discrete variable X_{p+j} takes a total of L_j values. As shown in (Lee and Hastie, 2015), the conditional distribution of a pairwise Markov random field is either Gaussian or multinomial. Thus, it enables a joint modeling of mixed data. In particular, for a continuous variable X_j its density conditional on all other variables X_{-j} is given by

$$p\left(x_j \middle| x_{-j}\right) = \exp\left\{ \frac{1}{\sigma_j^2}\left[x_j x_{-j}^T\theta_j - \frac{1}{2}\left(x_{-j}^T\theta_j\right)^2 - \frac{1}{2}x_j^2\right] - \frac{1}{2}\log 2\pi\sigma_j^2 \right\}$$

where $x_{-j} = (x_1,\ldots, x_{j-1}, x_{j+1},\ldots, x_{p+q})^T$ and $\theta_j \in R^{(p+q-1)}$ and σ_j^2 are parameters from the Gaussian distribution. For a discrete variable X_j, its conditional density is given by

$$p\left(x_j = i \middle| x_{-j}\right) = \frac{\exp\left\{ w_j^{0(i)} + x_{-j}^T w_j^{(i)} \right\}}{\Sigma_{i'=1}^{L_j}\exp\left\{ w_j^{0(i')} + x_{-j}^T w_j^{(i')} \right\}}, i \in \left\{1, ., L_j\right\}$$

where $(w_j^{(0)}, \ldots\ w_j^{(L_j)})^T$ are parameters from the multinomial distribution. In order to recover the Markov blanket, we implement a nodewise penalized generalized linear model (GLM) to perform neighborhood selection for each node (Lee and Hastie, 2015). More specifically, for node j we solve a penalized maximum likelihood problem that

$$\hat{\beta}_j = \underset{\beta_j}{\arg\min}\ -\Sigma_{k=1}^{n}\log p\left(x_{kj}\middle| x_{k,-j}\right) + \lambda_j \parallel \beta_j \parallel_1$$

Where x_{kj} is the observed data for subject k at node $j, x_{k,-j} = (x_{k1}, x_{k2}, \ldots, x_{k,j+1}, \ldots, x_{kn})$ and $\Sigma_{k=1}^{n}\log p\left(x_{kj}\middle| x_{k,-j}\right)$ is the log-likelihood of all subjects. The parameter $\beta_j = \theta_j$ when X_j is Gaussian; and $\left(w_j^{(1)^T}, w_j^{(2)^T}, \cdots, w_j^{(L_j)^T}\right)^T$ when X_j is categorical. In (1), we add an L_1-penalty on the β_j to enable the neighborhood selection. If node j is continuous, we connect node i with node j if the ith element of $\hat{\beta}_j$ is nonzero. If node j is categorical, we connect node i with node j if any ith element of $\hat{w}_j^{(k)} (k = 1, \ldots, L_j)$ is nonzero.

In the next section, we will discuss how to remove false connections identified at this stage that do not belong to the skeleton of the DAG. In (1), the tuning parameter λ_j controls the level of penalization and how sparse the resulting graph will be. Its optimal value is chosen by minimizing the extended Bayesian information criteria (EBIC) (Foygel and Drton, 2010).

$$EBIC_\gamma\left(\beta_j\right) = -2\Sigma_{k=1}^{n}\log p\left(x_{kj}\middle| x_{k,-j}\right) + \parallel \beta_j \parallel_0 \log n + 2\gamma \parallel \beta_j \parallel_0 \log(p + q - 1)$$

where n is sample size, $\|\beta_j\|_0$ is number of nonzero elements of β_j and γ is a user-predefined constant.

Identification of the Skeleton

The nodewise penalized GLM results in a Mixed Graphical Model (MGM), which is graphical model on continuous and discrete variables. Next, we remove edges in a MGM that do not exist in the corresponding DAG's skeleton. In a MGM, two vertices are connected if the two variables are dependent conditional on all other variables. However, in a v-structure $X \rightarrow W \leftarrow Z$ of a DAG, co-parents X and Z are independent conditional on their parents. Therefore, X and Z are not connected in the DAG's skeleton. But since X and Z are dependent given any vertex set that contains W or its descendant, X and Z are connected in a MGM. Therefore, we need to remove false connections between co-parents of v-structures in a MGM to obtain the DAG's skeleton.

The removal of false connections between co-parents of v-structures relies on testing the conditional independence of two variables given a set of other variables. In a Gaussian graphical model, testing conditional independence is equivalent to testing a zero partial correlation coefficient (Baba et al., 2004). Therefore, such a test can be easily performed using a Fisher's z-transformation (Ha et al., 2016) on the partial correlation. However, for mixed data,

testing conditional independence will be more complicated as it is no longer equivalent to testing zero partial correlation coefficient. To this end, we propose a permutation method to test the conditional independence of mixed data. Let X_j and X_l be two variables, and X_K be the set of variables that X_j and X_l are conditioning on. We first regress X_j and X_l on X_K respectively using a GLM. When X_j is Gaussian, we calculate the residual $r_{ij} = x_{ij} - \hat{x}_{ij}$, $(i = 1, \ldots, n)$ from the ordinary linear regression, where x_{ij} is the ith observation of X_j and $\hat{x}_{ij}$ the prediction of x_{ij} from the ordinary linear regression. When X_j is discrete, we calculate the Pearson residual from a multinomial logit model

$$r_{ij} = \Sigma_{k=1}^{L_j-1} \frac{x_{ij_k} - \hat{\mu}_{ij_k}}{\sqrt{\hat{\mu}_{ij_k}\left(1 - \hat{\mu}_{ij_k}\right)}}$$

where x_{ijk} the ith observation of the kth dummy variable created for X_j and $\hat{\mu}_{ij_k}$ is its predicted value from the logit model. In a special case of binary outcome, the above form reduces to the Pearson residual from a logistic model. Then, we calculate the partial correlation

$$\hat{\rho}_{jl} = \frac{\Sigma_{i=1}^{n}\left(r_{ij} - \bar{r}_j\right)\left(r_{il} - \bar{r}_l\right)}{\sqrt{\Sigma_{i=1}^{n}\left(r_{ij} - \bar{r}_j\right)^2 \Sigma_{i=1}^{n}\left(r_{il} - \bar{r}_l\right)^2}}$$

where $\bar{r}_j = \frac{1}{n}\Sigma_{i=1}^{n} r_{ij}$ and $\bar{r}_l = \frac{1}{n}\Sigma_{i=1}^{n} r_{il}$. Next, we permute the residuals $(r_{il})_{i=1}^{n}$ to have $(r_{\pi(i)l})_{i=1}^{n}$ where $\pi(i) \in \{1,\ldots,n\}$ is the permuted label of i. The permutation is repeated for B times. For the b^{th} permutation, we calculate the partial correlation

$$\hat{\rho}_{jl}^{(b)} = \frac{\Sigma_{i=1}^{n}\left(r_{ij} - \bar{r}_j\right)\left(r_{il}^{(b)} - \bar{r}_l^{(b)}\right)}{\sqrt{\Sigma_{i=1}^{n}\left(r_{ij} - \bar{r}_j\right)^2 \Sigma_{i=1}^{n}\left(r_{il}^{(b)} - \bar{r}_l^{(b)}\right)^2}}$$

The p-value testing the conditional independence of X_j and X_l then given by $p = \frac{1}{B}\Sigma_{i=1}^{B} I(\hat{\rho}_{jl} > \hat{\rho}_{jl}^{(b)})$where $I(x)$ is the indicator function. We conclude that X_j and X_l are conditionally independent if such a p-value is greater than 0.05. Based on the above test of conditional independence, we remove the edges belonging to the MGM but not the DAG's skeleton and obtain the d-separation set.

Orientation of the Mixed DAG

In the last step, we add orientation to the skeleton of the DAG using a greedy search algorithm as proposed in (Tsamardinos et al., 2006). We aim to find the orientation such that the Bayesian Information Criterion (BIC) of the whole graph is minimized (Schwarz, 1978). For a given directed graph, the BIC score for the jth ($j = 1,2,3,\ldots,(p+q)$) node is

$$BIC^{(j)} = -2\log L^{(j)}\left(\hat{\beta}\right) + \| \hat{\beta} \|_0 \log n$$

where $L^{(j)}(\hat{\beta})$ is the log-likelihood of the GLM regressing the jth node on its parents, $\hat{\beta}$ is the estimated vector of coefficients, and $\| \hat{\beta} \|_0$ is the number of nonzero elements in $\hat{\beta}$. The overall score of a directed graph is then given by $BIC^{(overall)} = \Sigma_{j=1}^{p+q} BIC^{(j)}$. The greedy search starts from an empty graph, whose score is calculated as summation of scores of each node without any parent. Then, for a node j and any node k connected with j in the estimated skeleton, we attempt to add, delete or reverse an edge between them based on the BIC change. More specifically, if there is no directed edge between nodes j and k at the current iteration, we add a directed edge $j \rightarrow k$ if the BIC score becomes smaller after adding this directed edge. If there is a directed edge between nodes j and k, we delete or reverse it if the BIC score becomes smaller after deleting or reversing this edge. This algorithm stops when the above edge operations fail to decrease the overall BIC score and the resulting directed graph is the estimated DAG. For the pseudo code (**Supplementary Table S1**) and a small-scale illustration (**Supplementary Figure S1**) of our entire algorithm, see the **Supplementary Material**.

RESULTS

Simulation Studies

To assess our method's performance, we simulate eight scenarios with different combinations of sample size, number of nodes and edges, and percentage of categorical nodes. We vary the sample size by 100 and 1,000; the number of nodes by 100 and 500; the percentage of categorical nodes by 10% and 20%; and the number of edges by 100 and 500. For each scenario, each categorical node contains 4 levels. More details of the simulation settings are summarized in **Table S2** in the **Supplementary Material**.

For each scenario, we first use the R package spacejam to generate a DAG. We randomly select 10% or 20% of the nodes as categorical and remaining nodes as continuous. For node i with no parents, if X_i is continuous, X_i is generated from $N(0,1)$; if X_i is categorical, X_i is sampled from {1,2,3,4} with equal probabilities. For node i; with parents, if X_i is continuous, X_i is generated from $N(\Sigma_{j\in parent(i)} X_j, 1)$, where *parent* ($i$) is the parent(s) of node i; if X_i is a categorical variable, X_i is generated from *Multinomial* (1,p) where $p=(p_1, p_2, p_3, p_4)$ and $p_l = \frac{\exp(l\sum_{j\in parent(i)} X_j)}{\Sigma_{l=1}^{4}\exp(l\sum_{j\in parent(i)} X_j)}, l = 1,2,3,4.$

In simulation studies, we compared our method with the CPC-stable method (implemented the R package pcalg) and the MMHC method (implemented by the R package bnlearn). Both methods cannot distinguish categorical and continuous variables but treat all of them as continuous. For each method, we evaluated edge recovery performance in both the estimated skeleton and the estimated DAG. The edge recovery performance is assessed through sensitivity, specificity, and false discovery rate (FDR). When evaluating the estimated skeleton, we define true edges as edges appearing in the true DAG's skeleton, estimated edges as edges appearing in the estimated skeleton, true null edges as unconnected edges in the true DAG's skeleton, and estimated null edges as unconnected edges in the estimated skeleton. We further defined sensitivity, specificity, and FDR of the estimated skeleton as follows:

$$Sensitivity = \frac{\#\ of\ [(estimated\ \ edges \cap true\ \ edges)]}{\#\ of\ \ true\ \ edges},$$

$$Specificity = \frac{\#\ of\ [estimated\ \ \ null\ \ edges \cap true\ \ \ null\ \ edges]}{\#\ of\ \ true\ \ \ null\ \ edges},$$

$$FDR\ = \frac{\#\ of\ [estimated\ \ edges - true\ \ edges]}{\#\ of\ \ \ estimated\ \ edges}$$

When evaluating the estimated DAG, we defined true edges as directed edges in the true DAG, estimated edges as directed edges in the estimated DAG, undetermined edges as edges with undetermined direction in the estimated DAG, true null edges as unconnected edges in the true DAG, and estimated null edges as unconnected edges in the estimated DAG. Then, the sensitivity, specificity, and FDR of the estimated DAG is defined as follows:

$$Sensitivity = \frac{\#\ of\ [(estimated\ \ edges - undermined\ \ edges) \cap true\ \ edges]}{\#\ of\ \ true\ \ edges},$$

$$Specificity = \frac{\#\ of\ [estimated\ \ \ null\ \ edges \cap true\ \ \ null\ \ edges]}{\#\ of\ \ true\ \ \ null\ \ edges},$$

$$FDR\ \ \ (directed) = \frac{\#\ of\ [estimated\ \ edges - true\ \ edges]}{\#\ of\ \ \ estimated\ \ edges}$$

Among the three measurements, sensitivity measures how a method recovers the connected edges in the true DAG and its skeleton. In particular, for DAG, sensitivity also measures if the direction of an edge is correctly recovered. Specificity measures how a method identifies the null edges in the true DAG and its skeleton. FDR measures the rate of falsely identified edges. In **Figure 1**, we present the boxplots of sensitivity, specificity, and FDR for all simulated scenarios.

Sensitivity, specificity, and FDR should be considered simultaneously to assess the overall edge recovery performance. In **Figures 1A–D**, the true DAG is sparse, i.e., not too many edges are connected. Our method has much better specificity and FDR for recovering the DAG and its skeleton, even though its sensitivity is smaller than the two competing methods. In **Figures 1E–H**, the true DAG is dense, i.e., many edges are connected. Our method performs the best in terms of all three measurements in both recovering the DAG and its skeleton. In all cases, our method's FDR is much lower, indicating that it estimates many fewer false positive edges. These results clearly demonstrate the merit of our methods by distinguishing categorial variables from continuous variables in the mixed data, especially when the DAG is dense. For mixed data, directly applying existing methods and ignoring data type difference clearly has inferior performance.

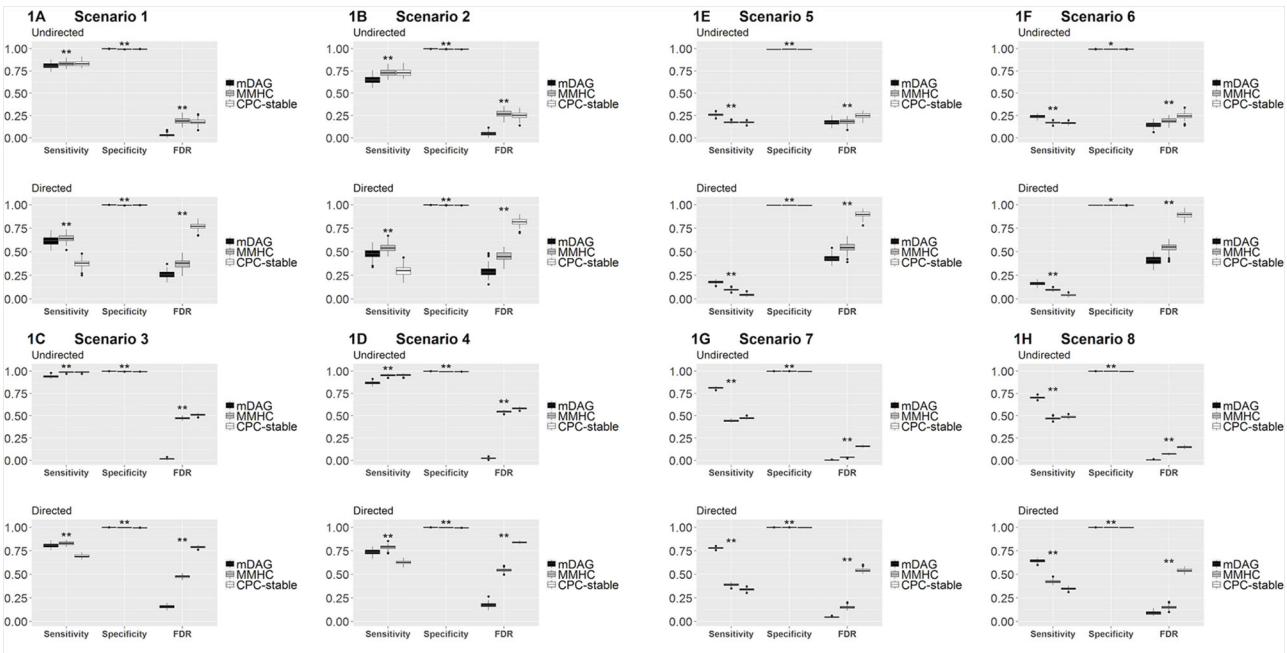

FIGURE 1 | Sensitivity, specificity, and FDR of mDAG and two alternative methods, MMHC and CPC-stable, in simulation scenarios 1–8. **(A)** Scenario 1; **(B)** Scenario 2; **(C)** Scenario 3; **(D)** Scenario 4; **(E)** Scenario 5; **(F)** Scenario 6; **(G)** Scenario 7; **(H)** Scenario 8. The X-axis indicates the measurements of performance (sensitivity, specificity, and FDR); the Y-axis indicates the corresponding values. "*" indicates the sensitivity/specificity/FDR from mDAG significantly differs from the sensitivity/specificity/FDR of CPC-stable or the sensitivity/specificity/FDR of MMHC. "**" indicates the sensitivity/specificity/FDR from mDAG significantly differs from the sensitivity/specificity/FDR of CPC-stable and the sensitivity/specificity/FDR of MMHC. Such comparisons are tested by two-sample Wilcoxon.

Real Data Application

Human *Chlamydia* Infection Dataset

Chlamydia trachomatis can ascend from the cervix to the uterus and fallopian tubes (upper genital tract) to cause long term sequelae, including chronic pelvic pain and infertility. Inflammatory cytokines and chemokines were measured in cervical secretions from 160 asymptomatic *C. trachomatis* infected women (age 15–30 years), participating in a previously described T cell Response Against Chlamydia (TRAC) cohort (Russell et al., 2015). The Institutional Review Boards for Human Subject Research at the University of Pittsburgh and the University of North Carolina approved the study and all participants provided written informed consent prior to inclusion. Ninety-six proteins were quantified using Milliplex Magnetic Bead Assay Kits (Millipore Sigma, St. Louis, MO), as previously described (Poston et al., 2019). 160 women who were infected at enrollment were assigned to two groups: women who had both cervical and endometrial infection were defined as Endo+ (cases), while those with cervical only infection were defined as Endo- (controls). To determine the regulatory networks involved in chlamydial ascension to the endometrium, we focused on 14 cytokines that were consistently detected in cervical secretions and were tentatively positively or negatively associated with endometrial infection by univariable logistic regression after adjustment for previously determined confounders, including cervical chlamydial load and gonorrhea coinfection (P<0.20) (Poston et al., 2019). We jointly modeled continuous nodes, including expression of 14 cervical cytokines and one covariate (cervical chlamydial load), with categorical nodes, including the binary disease outcome (endometrial infection: Endo+ vs. Endo-) and a binary covariate (gonorrhea coinfection) by the mDAG.

Results for our mDAG analysis are shown in **Figure 2A**, and the arrows indicate direction. We found two distinct pathways that emanate from CXCL10. The CXCL9 network is connected with ascending infection, while the CXCL11 network is distant and disconnected, which indicates a more favorable host response. The CXCL9 network includes CXCL13, IL-17A, CCL4, and TNFα as downstream regulated proteins. These cytokines are predominately associated with the induction of antibody and Th17 cells that are not protective against chlamydial genital tract infection (Andrew et al., 2013; Frazer et al., 2013; Darville et al., 2019). CXCL13, a CXCR5 ligand, is produced by multiple cell types and is a potent recruiter and activator of T follicular helper (Tfh) cells and B cells (Legler et al., 1998; Breitfeld et al., 2000). CXCL13 is a marker of germinal center activity (Havenar-Daughton et al., 2016) and may also reflect increased ectopic lymphocyte cluster development (Denton et al., 2019). Thus, increased CXCL13 levels may promote or sustain plasma cell aggregates previously observed in tissues from women with chlamydial endometritis and salpingitis (Kiviat et al., 1990). Increased CXCL13 levels that stimulate plasma cell development are consistent with detection of high serum and cervical levels of anti-chlamydial IgG and IgA in women who remain susceptible to repeated chlamydial infection (Darville et al., 2019). This is consistent with the network connectivity of CXCL13 and IL-17A, since proinflammatory CXCR5+ Th17 cells are also effective B-cell helpers capable of inducing strong antibody responses (Morita et al., 2011). Furthermore, the production of TNFα by CCL4-recruited

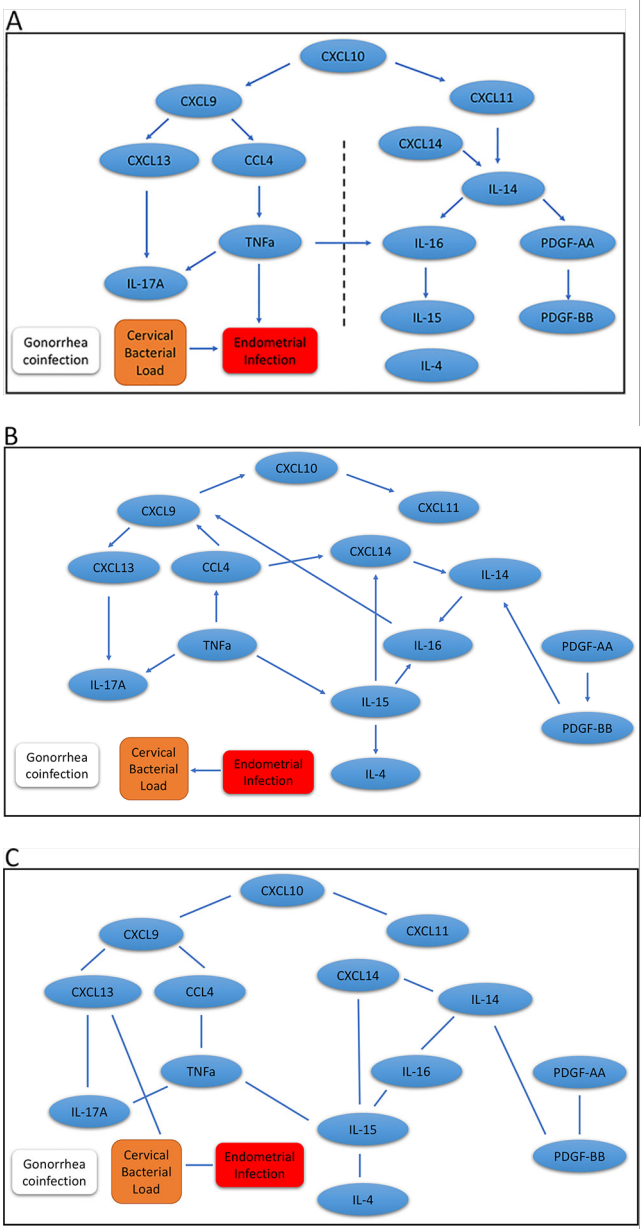

FIGURE 2 | Graphic results for causal network analysis of human *Chlamydia* infection dataset, a mixed type dataset consisting of continuous variables, including expression of 14 cervical cytokines and one covariate (cervical chlamydial load), and categorical variables, including the binary disease outcome (endometrial infection: Endo+ vs. Endo-) and a binary covariate (gonorrhea coinfection) by mDAG and two alternative methods, respectively. The arrows indicate direction of causality. **(A)** mDAG; **(B)** MMHC; **(C)** CPC-stable. The dashed line in **(A)** separates cytokines connected to ascension on the left, from cytokines disconnected from ascension on the right.

CD8 T cells may play a role in recruitment or differentiation of Th17 cells and enhance genital tract pathology (Murthy et al., 2011; Andrew et al., 2013). Besides chlamydial load, a factor we previously identified as associated with enhanced risk for upper genital tract infection, the analysis indicated TNFα production was

connected with chlamydial ascension. Previous studies have linked TNFα to infertility in *C. trachomatis*-infected women (Reddy et al., 2004; Srivastava et al., 2008).

The other major network that diverges from ascension is driven by CXCL11 and includes IL-14, CXCL14, IL-16, IL-15, PDGF-AA, and PDGF-BB. CXCL11 can induce and recruit CXCR3+ T cells shown to be protective during chlamydial infection (Perry et al., 1997), and could therefore prevent ascension. CXCL11 has strong binding affinity to its receptor, CXCR3, which is consistent with the ability of CXCL11 to increase intracellular calcium at lower doses than CXCL9 (Cole et al., 1998), and may explain the deviation of these two chemokines into separate networks. Next, the convergence of CXCL14 and CXCL11 with IL-14 could represent the ability of CXCL14 to enhance CD4 T cell activation (Chen et al., 2010). This activation would lead to the release of IL-14 and subsequently stimulate local B cell activation and proliferation (Ambrus et al., 1993). Although T cell interactions with activated antigen-presenting B cells could enhance antibody production capable of initiating Fc-mediated platelet activation and PDGF release, this cell-to-cell signaling will also trigger T cell receptor-mediated IL-16 secretion (Wu et al., 1999) and further enhance CD4 T cell recruitment (Lynch et al., 2003). IL-16 can directly stimulate mononuclear phagocyte IL-15 production (Mathy et al., 2000), which is critical for T cell survival and effector function (Borger et al., 1999; Purton et al., 2007) that would protect from chlamydial ascension. These findings are consistent with our previous analysis demonstrating that cytokines downstream of CXCL9 were associated with increased odds of endometrial infection, while cytokines downstream of CXCL11 were associated with decreased odds (Poston et al., 2019).

In addition, we applied the MMHC and CPC-stable algorithms to infer the regulatory pathways. Although the MMHC (**Figure 2B**) was able to predict the causal direction among cytokines, the directionality was completely disconnected from the disease trait, and the direction between cervical bacterial load and upper genital tract infection was reversed. Regulatory networks predicted by the CPC-stable algorithm (**Figure 2C**) completely failed to infer the direction in our cytokine dataset, which might be due to its conservative feature.

These results suggest that our proposed mDAG can infer upstream causal cytokines and downstream effector cytokines more closely linked to disease and correctly separate pathogenic and protective regulatory networks.

Metabolic Syndrome in Men Dataset

The Metabolic Syndrome in Men (METSIM) study is a population-based study with 10,197 males randomly selected from the population register of the town of Kuopio in Finland (Stancakova et al., 2009). The Ethics Committee of the University of Eastern Finland and Kuopio University Hospital approved the METSIM study, and this study was conducted in accordance with the Declaration of Helsinki. All study participants gave written informed consent. A subset of 770 participants have gene expression measurements from subcutaneous adipose tissue (Civelek et al., 2017), we analyzed genotype, gene expression, and plasma adiponectin levels using our mDAG and alternative methods. For directional inference, we focused on two GWAS loci for adiponectin (Zhong et al., 2019) and expression of genes within ± 1Mb at each locus t. Genetic variants at the first locus near the *ADIPOQ* gene may exert their effects on adiponectin levels through expression of the *ADIPOQ* gene, which is expressed in adipose tissue and encodes the adiponectin protein studied. In contrast, genetic variants identified at the second locus, where the index SNP (the SNP with the most significant *p*-value from GWAS) is an intronic SNP in *ARL15*, which might influence adiponectin levels through expression of the *FST* gene instead of *ARL15* (Civelek et al., 2017; Martin et al., 2017; Zhong et al., 2019).

We extracted genotypes of the index SNP for each locus and expression levels of genes within ± 1Mb of each index SNP. Because a gene may have *multiple probesets,* we first applied a Sobel test to each probeset to detect mediation effect of the index SNP on adiponectin levels through the probeset. We then selected the probeset with the minimum mediation *p* value. We applied our mDAG and alternative methods to estimate DAGs (**Figures 3A–C**) for the *ADIPOQ* locus and 4A-4C for the *FST-ARL15* locus]. mDAG has the feature of forcing SNPs to point to other nodes. Results of mDAG suggest that the *ADIPOQ* gene is a mediator at the first locus (**Figure 3A**), and that *FST* gene (not *ARL15*) is a mediator at the second locus (**Figure 4A**). These findings are consistent with the results in (Zhong et al., 2019). In contrast, alternative methods failed to identify the expected directional relationships (**Figures 4B**, **C**).

DISCUSSION

Jointly modeling the probability distribution of the continuous measurements of gene expression or protein abundance and the categorical nodes, such as disease traits and SNPs, identifies the regulatory paths of a disease. More importantly, it distinguishes the disease-causing pathways from the disease-reaction pathways, and identifies genes mediating the effects of GWAS loci on diseases. This leads to a better understanding of disease mechanisms, and helps generate more precise targets for new therapeutic and diagnostic interventions. The existing DAG methods cannot be applied to such a joint model, as they mostly assume all nodes are of the same type.

To this end, we proposed a mixed DAG (mDAG) algorithm to infer the regulatory paths of mixed data. Our mDAG algorithm is a hybrid method and consists of three main steps including identification of the Markov blanket, determination of the skeleton, and inference of edge orientation. There are some alternative algorithms which can be applied in each step. For example, a more general framework (Zhuang et al., 2016) can be used to estimate undirected graph and PC algorithm based approach can be applied for edge orientation. Our algorithm uses a new permutation-based method to test the conditional independence of nodes of mixed types. We compared our method with two alternative well-known methods that ignore the type difference of nodes. The simulation results show that mDAG outperforms the alternative methods in terms of the FDR, sensitivity, and specificity of the edge recovery of the

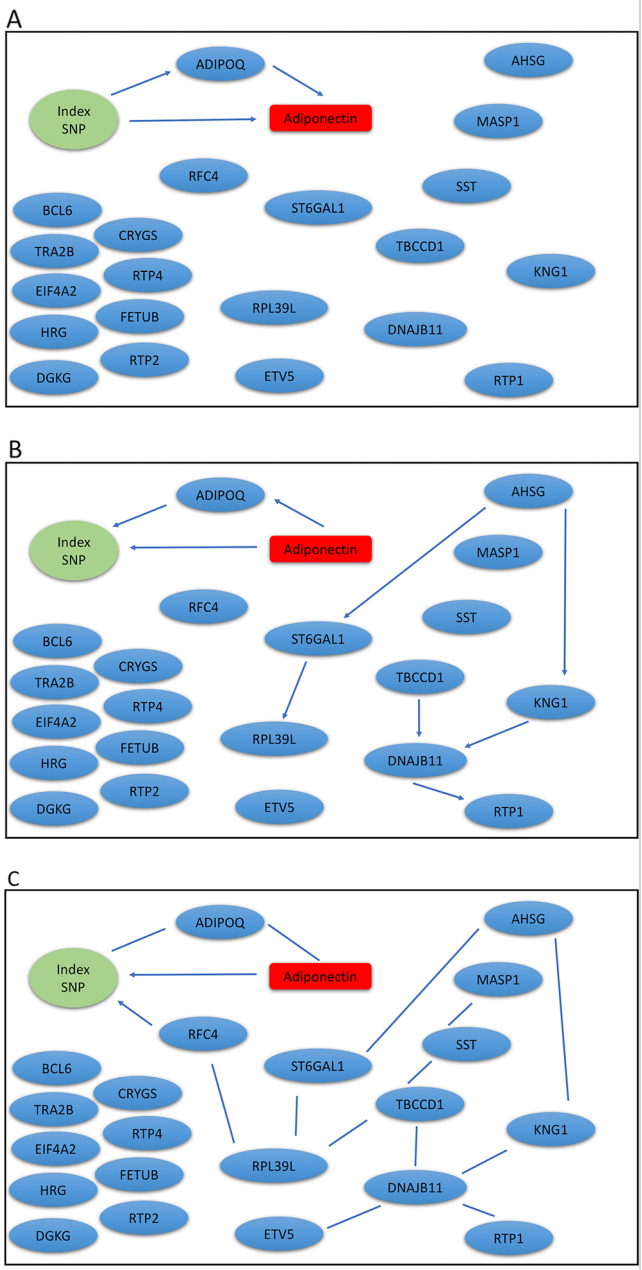

FIGURE 3 | Graphic results for causal network analysis of the Metabolic Syndrome in Men dataset, a mixed type dataset consisting of a categorical variable, genotypes of one index SNP at the *ADIPOQ* GWAS locus, and several continuous variables, including expression levels of 21 genes and plasma adiponectin levels (disease trait). The arrows indicate direction of causality. **(A)** mDAG; **(B)** MMHC; **(C)** CPC-stable.

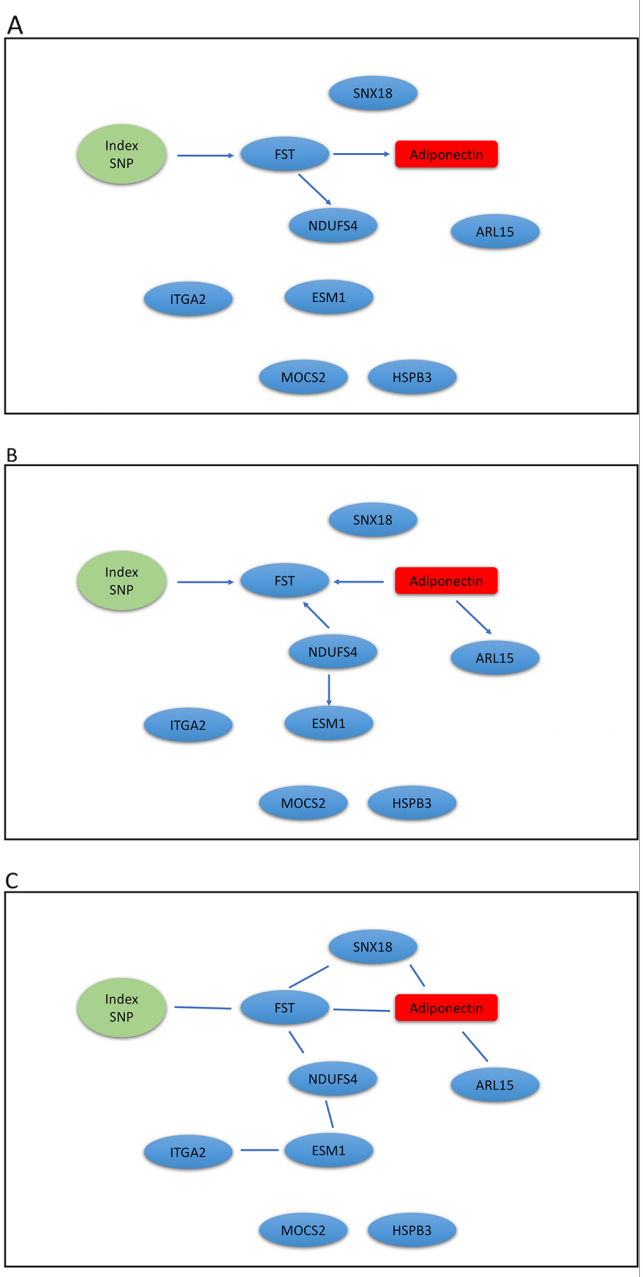

FIGURE 4 | Graphic results for causal network analysis of Metabolic Syndrome in Men dataset, a mixed type dataset consisting of a categorical variable, one index SNP at *ARL15* GWAS locus, and continuous variables, including expression of 8 genes and adiponectin levels (disease trait). The arrows indicate direction of causality. **(A)** mDAG; **(B)** MMHC; **(C)** CPC-stable.

underlying true DAG. Results from the human chlamydial infection dataset demonstrates that mDAG successfully reconstructs the pathogenic and protective regulatory networks for chlamydial ascension. The regulatory pathways inferred by our method identify upstream causal factors and generate hypotheses for causal direction of regulatory pathways, and therefore provide candidates for experimental validation. For the Metabolic Syndrome in Men dataset, mDAG also identifies the expected paths of important GWAS loci for adiponectin suggested by previous publications (Civelek et al., 2017; Martin et al., 2017), even in the presence of multiple presumably irrelevant genes in the 1D neighborhood of the loci under study in the model, indicating that mDAG can bridge the functional gap of synonymous GWAS signals and provide the mechanistic hypotheses underlying GWAS variants.

The mDAG could not only be used to infer the causality paths in mixed types of proteomic or transcriptomic data with categorical phenotypes and/or SNP data, but it could also be applied to other mixed data, such as metabolomics and DNA structural variants, including copy number variation, since it does not require prior biological knowledge. Beyond genetics, it can be applied to social, behavioral, and psychology studies.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the Gene Expression Omnibus with the accession number GSE70353.

ETHICS STATEMENT

For the TRAC study, the Institutional Review Boards for Human Subject Research at the University of Pittsburgh and the University of North Carolina approved the study and all participants provided written informed consent prior to inclusion. For the METSIM study, the Ethics Committee of the University of Eastern Finland and Kuopio University Hospital approved the METSIM study, and this study was conducted in accordance with the Declaration of Helsinki. All study participants gave written informed consent.

AUTHOR CONTRIBUTIONS

Conceptualization and supervision: QL and XZ. Data curation: XZ, TD, TP, CS, KM, and YL. Resources: XZ, TD, CS, KM, and YL. Formal analysis, visualization and writing—Original draft preparation: WZ and LD. Investigation, methodology, software and validation: WZ, LD, QL, and XZ. Writing—Review and editing: QL, XZ, TD, TP, DW, KM, and YL.

FUNDING

This work was supported by Development and Research Program awards by National Institutes of Health (www.nih.gov) to XZ (U19 AI144181, AI113170), National Institutes of Health (www.nih.gov) to TD (R01 AI119164, U19 AI084024 and AI007001), KM (DK093757), YL (R01 HL129132 and R01 GM105785), DL (R01 GM047845) and American Heart Association (www.heart.org) to CS (17POST33650016). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

ACKNOWLEDGMENTS

We thank all participants in TRAC and METSIM for agreeing to take part in the studies, and all investigators in these two studies for sharing the data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00008/full#supplementary-material

REFERENCES

Ambrus, J. L., Pippin, J., Joseph, A., Xu, C., Blumenthal, D., Tamayo, A., et al. (1993). Identification of a cDNA for a human high-molecular-weight B-cell growth factor. *Proc. Natl. Acad. Sci.* 90, 6330–6334. doi: 10.1073/pnas.90.13.6330

Andrew, D. W., Cochrane, M., Schripsema, J. H., Ramsey, K. H., Dando, S. J., O'Meara, C. P., et al. (2013). The duration of Chlamydia muridarum genital tract infection and associated chronic pathological changes are reduced in IL-17 knockout mice but protection is not increased further by immunization. *PloS One* 8, e76664. doi: 10.1371/journal.pone.0076664

Atias, N., and Sharan, R. (2013). iPoint: an integer programming based algorithm for inferring protein subnetworks. *Mol. Biosyst.* 9, 1662–1669. doi: 10.1039/c3mb25432a

Baba, K., Shibata, R., and Sibuya, M. (2004). Partial correlation and conditional correlation as measures of conditional independence. *Aust. N. Z. J. Stat.* 46, 657–664. doi: 10.1111/j.1467-842X.2004.00360.x

Borger, P., Kauffman, H. F., Postma, D. S., Esselink, M. T., and Vellenga, E. (1999). Interleukin-15 differentially enhances the expression of interferon-γ and interleukin-4 in activated human (CD4+) T lymphocytes. *Immunology* 96, 207. doi: 10.1046/j.1365-2567.1999.00679.x

Borsboom, D., and Cramer, A. O. J. (2013). Network analysis: an integrative approach to the structure of psychopathology. *Annu. Rev. Clin. Psychol.* 9, 91–121. doi: 10.1146/annurev-clinpsy-050212-185608

Breitfeld, D., Ohl, L., Kremmer, E., Ellwart, J., Sallusto, F., Lipp, M., et al. (2000). Follicular B helper T cells express CXC chemokine receptor 5, localize to B cell follicles, and support immunoglobulin production. *J. Exp. Med.* 192, 1545–1552. doi: 10.1084/jem.192.11.1545

Chen, L., Guo, L., Tian, J., He, H., Marinova, E., Zhang, P., et al. (2010). Overexpression of CXC chemokine ligand 14 exacerbates collagen-induced arthritis. *J. Immunol.* 184, 4455–4459. doi: 10.4049/jimmunol.0900525

Chen, L. S. (2012). "Using eQTLs to reconstruct gene regulatory networks," in *Quantitative Trait Loci (QTL)* (New York: Springer), 175–189. doi: 10.1007/978-1-61779-785-9_9

Chickering, D. M. (2002). Optimal structure identification with greedy search. *J. Mach. Learn. Res.* 3, 507–554.

Civelek, M., Wu, Y., Pan, C., Raulerson, C. K., Ko, A., He, A., et al. (2017). Genetic regulation of adipose gene expression and cardio-metabolic traits. *Am. J. Hum. Genet.* 100, 428–443. doi: 10.1016/j.ajhg.2017.01.027

Cole, K. E., Strick, C. A., Paradis, T. J., Ogborne, K. T., Loetscher, M., Gladue, R. P., et al. (1998). Interferon–inducible T cell alpha chemoattractant (I-TAC): a novel Non-ELR CXC Chemokine with potent activity on activated T cells through selective high affinity binding to CXCR3. *J. Exp. Med.* 187, 2009–2021. doi: 10.1084/jem.187.12.2009

Colombo, D., and Maathuis, M. H. (2014). Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.* 15, 3741–3782.

Darville, T., Albritton, H. L., Zhong, W., Dong, L., O'Connell, C. M., Poston, T. B., et al. (2019). Anti-chlamydia IgG and IgA are insufficient to prevent endometrial Chlamydia infection in women and increased anti-chlamydia IgG is associated with enhanced risk for incident infection. *Am. J. Reprod. Immunol.* 81 (5), e13103. doi: 10.1111/aji.13103

Denton, A. E., Innocentin, S., Carr, E. J., Bradford, B. M., Lafouresse, F., Mabbott, N. A., et al. (2019). Type I interferon induces CXCL13 to support ectopic germinal center formation. *J. Exp. Med.* jem–20181216. 216 (3), 621–637. doi: 10.1084/jem.20181216

Foygel, R., and Drton, M. (2010). "Extended Bayesian information criteria for Gaussian graphical models," in *Advances in neural information processing systems*, 604–612. San Diego, CA: Neural Information Processing Systems.

Frazer, L. C., Scurlock, A. M., Zurenski, M. A., Riley, M. M., Mintus, M., Pociask, D. A., et al. (2013). IL-23 Induces IL-22 and IL-17 Production in Response to C hlamydia muridarum Genital Tract Infection, but the Absence of these Cytokines does not Influence Disease Pathogenesis. *Am. J. Reprod. Immunol.* 70, 472–484. doi: 10.1111/aji.12171

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *J. Comput. Biol.* 7, 601–620. doi: 10.1089/106652700750050961

Ha, M. J., Sun, W., and Xie, J. (2016). PenPC: a two-step approach to estimate the skeletons of high-dimensional directed acyclic graphs. *Biometrics* 72, 146–155. doi: 10.1111/biom.12415

Havenar-Daughton, C., Lindqvist, M., Heit, A., Wu, J. E., Reiss, S. M., Kendric, K., et al. (2016). CXCL13 is a plasma biomarker of germinal center activity. *Proc. Natl. Acad. Sci.* 113, 2702–2707. doi: 10.1073/pnas.1520112113

Hawrylycz, M. J., Lein, E. S., Guillozet-Bongaarts, A. L., Shen, E. H., Ng, L., Miller, J. A., et al. (2012). An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* 489, 391. doi: 10.1038/nature11405

Huang, S., Li, J., Sun, L., Ye, J., Fleisher, A., Wu, T., et al. (2010). Learning brain connectivity of Alzheimer's disease by sparse inverse covariance estimation. *Neuroimage* 50, 935–949. doi: 10.1016/j.neuroimage.2009.12.120

Kiviat, N. B., Wølner-Hanssen, P., Eschenbach, D. A., Wasserheit, J. N., Paavonen, J. A., Bell, T. A., et al. (1990). Endometrial histopathology in patients with culture-proved upper genital tract infection and laparoscopically diagnosed acute salpingitis. *Am. J. Surg. Pathol.* 14, 167–175. doi: 10.1097/00000478-199002000-00008

Lee, J. D., and Hastie, T. J. (2015). Learning the structure of mixed graphical models. *J. Comput. Graph. Stat.* 24, 230–253. doi: 10.1080/10618600.2014.900500

Legler, D. F., Loetscher, M., Roos, R. S., Clark-Lewis, I., Baggiolini, M., and Moser, B. (1998). B cell–attracting chemokine 1, a human CXC chemokine expressed in lymphoid tissues, selectively attracts B lymphocytes via BLR1/CXCR5. *J. Exp. Med.* 187, 655–660. doi: 10.1084/jem.187.4.655

Lynch, E. A., Heijens, C. A. W., Horst, N. F., Center, D. M., and Cruikshank, W. W. (2003). Cutting edge: IL-16/CD4 preferentially induces Th1 cell migration: requirement of CCR5. *J. Immunol.* 171, 4965–4968. doi: 10.4049/jimmunol.171.10.4965

Martin, J. S., Xu, Z., Reiner, A. P., Mohlke, K. L., Sullivan, P., Ren, B., et al. (2017). HUGIn: Hi-C unifying genomic interrogator. *Bioinformatics* 33, 3793–3795. doi: 10.1093/bioinformatics/btx359

Mathy, N. L., Scheuer, W., Lanzendörfer, M., Honold, K., Ambrosius, D., Norley, S., et al. (2000). Interleukin-16 stimulates the expression and production of pro-inflammatory cytokines by human monocytes. *Immunology* 100, 63–69. doi: 10.1046/j.1365-2567.2000.00997.x

Meek, C. (1995). Causal inference and causal explanation with background knowledge, in: *UAI'95:Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. San Francisco, CA: Morgan Kaufmann Publishers Inc., pp. 403–410.

Meek, C. (2013). Strong completeness and faithfulness in Bayesian networks. *arXiv Prepr. arXiv1302.4973*.

Morita, R., Schmitt, N., Bentebibel, S.-E., Ranganathan, R., Bourdery, L., Zurawski, G., et al. (2011). Human blood CXCR5+ CD4+ T cells are counterparts of T follicular cells and contain specific subsets that differentially support antibody secretion. *Immunity* 34, 108–121. doi: 10.1016/j.immuni.2010.12.012

Murthy, A. K., Li, W., Chaganty, B. K. R., Kamalakaran, S., Guentzel, M. N., Seshu, J., et al. (2011). Tumor necrosis factor alpha production from CD8+ T cells mediates oviduct pathological sequelae following primary genital Chlamydia muridarum infection. *Infect. Immun.* 79, 2928–2935. doi: 10.1128/IAI.05022-11

Nagarajan, R., Scutari, M., and Lèbre, S. (2013). Bayesian networks in R. *Springer* 122, 125–127. doi: 10.1007/978-1-4614-6446-4

Oldham, M. C., Horvath, S., and Geschwind, D. H. (2006). Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proc. Natl. Acad. Sci.* 103, 17973–17978. doi: 10.1073/pnas.0605938103

Paull, E. O., Carlin, D. E., Niepel, M., Sorger, P. K., Haussler, D., and Stuart, J. M. (2013). Discovering causal pathways linking genomic events to transcriptional states using tied diffusion through interacting events (TieDIE). *Bioinformatics* 29, 2757–2764. doi: 10.1093/bioinformatics/btt471

Pearl, J. (2009). *Causality* (Cambridge, England: Cambridge University Press). doi: 10.1017/CBO9780511803161

Perry, L. L., Feilzer, K., and Caldwell, H. D. (1997). Immunity to Chlamydia trachomatis is mediated by T helper 1 cells through IFN-gamma-dependent and-independent pathways. *J. Immunol.* 158, 3344–3352.

Poston, T. B., Lee, D. E., Darville, T., Zhong, W., Dong, L., O'Connell, C. M., et al. (2019). Cervical cytokines associated with Chlamydia trachomatis susceptibility and protection. *J. Infect. Dis.* 220 (2), 330–339. doi: 10.1093/infdis/jiz087

Purton, J. F., Tan, J. T., Rubinstein, M. P., Kim, D. M., Sprent, J., and Surh, C. D. (2007). Antiviral CD4+ memory T cells are IL-15 dependent. *J. Exp. Med.* 204, 951–961. doi: 10.1084/jem.20061805

Reddy, B. S., Rastogi, S., Das, B., Salhan, S., Verma, S., and Mittal, A. (2004). Cytokine expression pattern in the genital tract of Chlamydia trachomatis positive infertile women–implication for T-cell responses. *Clin. Exp. Immunol.* 137, 552–558. doi: 10.1111/j.1365-2249.2004.02564.x

Russell, A. N., Zheng, X., O'connell, C. M., Taylor, B. D., Wiesenfeld, H. C., Hillier, S. L., et al. (2015). Analysis of factors driving incident and ascending infection and the role of serum antibody in Chlamydia trachomatis genital tract infection. *J. Infect. Dis.* 213, 523–531. doi: 10.1093/infdis/jiv438

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6, 461–464. doi: 10.1214/aos/1176344136

Sedgewick, A. J., Shi, I., Donovan, R. M., and Benos, P. V. (2016). Learning mixed graphical models with separate sparsity parameters and stability-based model selection. *BMC Bioinf.* 17, S175. doi: 10.1186/s12859-016-1039-0

Sedgewick, A. J., Ramsey, J. D., Spirtes, P., Glymour, C., and Benos, P. V. (2017). Mixed graphical models for causal analysis of multi-modal variables. *arXiv Prepr. arXiv1704.02621*. Cambridge, MA.

Spirtes, P., Glymour, C. N., and Scheines, R. (2000). *Causation, prediction, and search* (Cambridge, MA: MIT Press). doi: 10.7551/mitpress/1754.001.0001

Srivastava, P., Jha, R., Bas, S., Salhan, S., and Mittal, A. (2008). In infertile women, cells from Chlamydia trachomatis infected site release higher levels of interferon-gamma, interleukin-10 and tumor necrosis factor-alpha upon heat shock protein stimulation than fertile women. *Reprod. Biol. Endocrinol.* 6, 20. doi: 10.1186/1477-7827-6-20

Stancakova, A., Javorsky, M., Kuulasmaa, T., Haffner, S. M., Kuusisto, J., and Laakso, M. (2009). Changes in Insulin Sensitivity and Insulin Release in Relation to Glycemia and Glucose Tolerance in 6,414 Finnish Men. *Diabetes* 58, 1212–1221. doi: 10.2337/db08-1607

Tsagris, M., Borboudakis, G., Lagani, V., and Tsamardinos, I. (2018). Constraint-based causal discovery with mixed data. *Int. J. Data Sci. Anal.* 6, 19–30. doi: 10.1007/s41060-018-0097-y

Tsamardinos, I., Brown, L. E., and Aliferis, C. F. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Mach. Learn.* 65, 31–78. doi: 10.1007/s10994-006-6889-7

Wilentzik, R., and Gat-Viks, I. (2015). A statistical framework for revealing signaling pathways perturbed by DNA variants. *Nucleic Acids Res.* 43, e74–e74. doi: 10.1093/nar/gkv203

Wu, D. M. H., Zhang, Y., Parada, N. A., Kornfeld, H., Nicoll, J., Center, D. M., et al. (1999). Processing and release of IL-16 from CD4+ but not CD8+ T cells is activation dependent. *J. Immunol.* 162, 1287–1293.

Zhong, W., Spracklen, C. N., Mohlke, K. L., Zheng, X., Fine, J., and Li, Y. (2019). Multi-SNP mediation intersection-union test. *Bioinformatics* 35 (22), 4724–4729. doi: 10.1093/bioinformatics/btz285

Zhuang, R., Simon, N., and Lederer, J. (2016). Graphical models for discrete and continuous data.

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 Plant Science

TECHNOLOGY AND CODE
published: 18 February 2020
doi: 10.3389/fpls.2020.00098

BarleyNet: A Network-Based Functional Omics Analysis Server for Cultivated Barley, *Hordeum vulgare* L.

Sungho Lee[‡], *Tak Lee*[†‡], *Sunmo Yang and Insuk Lee*[*]

Department of Biotechnology, College of Life Science and Biotechnology, Yonsei University, Seoul, South Korea

OPEN ACCESS

Edited by:
Xiyin Wang,
North China University of Science and Technology, China

Reviewed by:
Le Shu,
University of California, Los Angeles, United States
Margaret Woodhouse,
Iowa State University, United States
Nils Stein,
Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Germany

***Correspondence:**
Insuk Lee
insuklee@yonsei.ac.kr

[†]**Present address:**
Tak Lee
Sainsbury Laboratory, University of Cambridge, Cambridge, United Kingdom

[‡]*These authors have contributed equally to this work*

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

Received: *12 November 2019*
Accepted: *22 January 2020*
Published: *18 February 2020*

Citation:
Lee S, Lee T, Yang S and Lee I (2020) BarleyNet: A Network-Based Functional Omics Analysis Server for Cultivated Barley, Hordeum vulgare L. Front. Plant Sci. 11:98. doi: 10.3389/fpls.2020.00098

Cultivated barley (*Hordeum vulgare* L.) is one of the most produced cereal crops worldwide after maize, bread wheat, and rice. Barley is an important crop species not only as a food source, but also in plant genetics because it harbors numerous stress response alleles in its genome that can be exploited for crop engineering. However, the functional annotation of its genome is relatively poor compared with other major crops. Moreover, bioinformatics tools for system-wide analyses of omics data from barley are not yet available. We have thus developed BarleyNet, a co-functional network of 26,145 barley genes, along with a web server for network-based predictions (http://www.inetbio.org/barleynet). We demonstrated that BarleyNet's prediction of biological processes is more accurate than that of an existing barley gene network. We implemented three complementary network-based algorithms for prioritizing genes or functional concepts to study genetic components of complex traits such as environmental stress responses: (i) a pathway-centric search for candidate genes of pathways or complex traits; (ii) a gene-centric search to infer novel functional concepts for genes; and (iii) a context-centric search for novel genes associated with stress response. We demonstrated the usefulness of these network analysis tools in the study of stress response using proteomics and transcriptomics data from barley leaves and roots upon drought or heat stresses. These results suggest that BarleyNet will facilitate our understanding of the underlying genetic components of complex traits in barley.

Keywords: barley, *Hordeum vulgare* L., gene network, network biology, crop systems genetics

INTRODUCTION

Cultivated barley (*Hordeum vulgare* L.) is one of the first cultivated grains, domesticated about 10,000 years ago in the Near East (Badr et al., 2000). It was ranked the fourth cereal crop in quantity produced after maize, bread wheat, and rice in 2017 (FAOSTAT 2017, http://fao.org/faostat/). Barley mainly serves as a source of fodder for livestock, fermentable material for alcoholic beverages, and is present in various healthy organic foods. In developing countries, it is also still a major source of carbohydrates. Furthermore, barley is a great plant model organism for studying genetic resistance to biotic or abiotic stress, since it can endure a great range of environmental stresses like drought, flood, and cold or fungal infections, either single or combined (Gürel et al., 2016). Therefore, the barley genome is a reservoir of numerous stress response alleles, which are precious

subjects for genetic engineering in other crop species. The size of the haploid Barley genome is approximately 5.3 Gbp. It is one of the largest diploid genomes sequenced to date and contains 83,105 putative genetic loci including 39,734 high-confidence ones.

Numerous studies have exploited these agronomically important traits, assisted by various new technologies such as high-throughput sequencing and mass spectrometry-based proteomics. Although they provide important clues about molecular components associated with complex plant traits, individual omics profiles are insufficient to reconstruct a holistic view of functional modules involved in these traits. Moreover, the functional interpretation of omics profile data generally requires the incorporation of other information. Therefore, a systems biology platform that integrates information derived from different data sources could effectively encapsulate the molecular network underlying complex traits. Co-functional gene networks have been applied to integrate the functional information of genes derived from heterogeneous data through a Bayesian statistics framework (Shim et al., 2017). Co-functional networks previously constructed for other major crop species have been successfully used in the genetic dissection of complex plant traits (Lee et al., 2015a; Lee et al., 2017; Lee et al., 2019). Yet, such an effective network-assisted systems genetics platform has not been developed for barley. Therefore, we developed BarleyNet, a co-functional network of barley genes and a companion web server (www.inetbio.org/barleynet/), enabling network-assisted systems genetics analysis for cultivated barley. All information on functional association between barley genes is also readily downloadable through the companion web server. Finally, the three complementary network-based algorithms implemented in the web server facilitate effective use of omics profiles for generating new functional hypotheses.

MATERIALS AND METHODS

Reference Genome

We constructed BarleyNet based on the IBSC_v2 barley genome assembly (https://plants.ensembl.org/Hordeum_vulgare/Info/Annotation/#assembly) presented by the International Barley Sequencing Consortium (Mascher et al., 2017). Among 83,105 putative genetic loci, 39,734 high-confidence loci were selected as a reference gene set for network construction. Supervised learning of co-functional gene pairs requires gold standard (GS) positive and negative gene pairs, which are generally derived from high-quality pathway annotation databases. However, both the quantity and the quality of pathway annotations for barley were not sufficient by the time we launched this project. Thus, we transferred GS-positive barley gene pairs based on sequence homology with those used for modeling *Arabidopsis* (Lee et al., 2015b) and rice (Lee et al., 2015a) gene networks. Consequently, 215,170 and 27,254 GS-positive gene pairs were transferred from rice and *Arabidopsis*, respectively. The final set of GS-positive gene pairs for training BarleyNet was a union of all transferred gene pairs, comprising 234,070 gene pairs among 7,350 barley genes (18.5% of the genome). All other possible pairwise relationships between the 7,350 barley genes were then considered GS-negatives, comprising 26,773,505 gene pairs.

Benchmarking Co-Functional Barley Gene Pairs

The likelihood of a functional association between two genes is based on the ratio between our belief after seeing the supporting data and our prior belief. Thus, we scored functional association between genes using previously developed log likelihood score (*LLS*) (Lee et al., 2004), shown as the following equation:

$$LLS = \ln\left(\frac{P(L|S)/P(\neg L|S)}{P(L)/P(\neg L)}\right)$$

where $P(L|S)$ and $P(\neg L|S)$ represent the probability of GS-positive and GS-negative gene pairs, respectively, supported by the given data, and $P(L)$ and $P(\neg L)$ represent the expected probability of GS-positive and GS-negative links, respectively.

Gene pairs are sorted by data-intrinsic scores such as the expression correlation coefficient, and then assigned into bins of 1,000 gene pairs. We computed *LLS* for each of the bins and then did a sigmoid regression between means of data-intrinsic scores and *LLS*s. Using the regression function, we calculated *LLS* for every gene pair derived from each data source.

Integrating Co-Functional Barley Gene Pairs

Functional association between barley genes can be supported by multiple data sources. We may integrate the *LLS* of their functional association by naïve Bayes integration, if there is no correlation between data sources, which is generally not true. In order to handle information correlation between supporting data sources, we previously developed the weighted sum (*WS*) method (Lee et al., 2007), shown as the following equation:

$$WS = Lo + \sum_{i=1}^{n} \frac{L_i}{D \times i}, \ for \ \ all \ \ L \geq T$$

where L_O represents the highest *LLS* of all available supporting data sources, and L_i represents the remaining *LLS*s with rank index *i*. *D* and *T* are free parameters for the weight factor and *LLS* cutoff to be considered, respectively. These free parameters were selected where the integrated network achieved the best performance based on a precision-recall curve. A total of 25 distinct data sources were finally integrated into BarleyNet (**Supplementary Table 1**).

Inferring Co-Functional Links From mRNA Co-Expression Patterns (CX)

Functionally associated genes tend to show a similar expression pattern across various biological contexts. Co-functional links between these genes were inferred from diverse sets of expression

profiles gathered from the Gene Expression Omnibus (GEO) database (Clough and Barrett, 2016), ArrayExpress (Kolesnikov et al., 2015), and Expression Atlas (Papatheodorou et al., 2018). We assessed a total of 2,385 expression profiles (1,780 by microarray and 650 by RNA-seq) and incorporated 28 datasets comprising 2,047 expression profiles into the final co-expression network. Affymetrix microarray data (Barley genome array, GPL1340) were normalized by MAS5 software. RNA-seq data were downloaded as raw data, quantified using Kallisto (Bray et al., 2016), and normalized as transcripts per million (TPM). The co-expression between two genes across expression profiles was assessed by the Pearson's correlation coefficient (*PCC*) and then benchmarked for functional associations by *LLS*. All the co-expression networks from the 28 expression datasets (**Supplementary Table 2**) were then integrated into a single co-expression network using the weighted sum method described above.

Inferring Co-Functional Links From Protein Domain Profile Association (DP)

The domain composition of a protein reflects its function. Therefore, the co-functional relationship between proteins can be inferred from the association between their domain composition profiles. We downloaded a list of barley proteins and identified domains in the InterPro database (Mitchell et al., 2018) for each protein from the Ensembl Plants database (Vullo et al., 2017). Then, mutual information scores were computed between domain profiles. We used a weighted mutual information (*WMI*) scheme, which assigns more weight on rarer domains during mutual information computation (Shim and Lee, 2016; Shim and Lee, 2020). We calculated *LLS*s for gene pairs using a regression function between *WMI* and *LLS*.

Inferring Co-Functional Links From Phylogenetic Profile Associations (PG)

During speciation, genes that operate the same biological processes tend to be inherited together. Therefore, we can infer co-functional gene pairs based on their co-inheritance pattern across a large number of species. Considering that gene inheritance across species can be represented as phylogenetic profiles, these can be used in the identification of co-inherited genes. We first aligned all the 39,734 barley protein sequences against total protein sequences from 1,626 bacterial genomes, 396 eukaryotic genomes, and 122 archaea genomes using BLASTP (Altschul et al., 1990), and then constructed phylogenetic profiles based on –log(E-value) of BLAST hit scores. Previously, we found that domain-specific phylogenetic profile analysis improved inference of co-functional links (Shin and Lee, 2015). Therefore, we calculated mutual information between two phylogenetic profiles for each of the three domains of life, resulting in three networks for profiles with bacterial, eukaryotic, and archaeal genomes. The resulting networks were scored by *LLS* and integrated into one single network for the phylogenetic profile method.

Inferring Co-Functional Links From Gene Neighborhood (GN)

Prokaryotic genes that operate in the same biological process tend to be located closely in chromosomes, often forming operons. We thus can infer functional associations between barley genes based on the proximity of their orthologs in prokaryotic genomes with two complementary measures: distance-based approach and probability-based approach (Shin et al., 2014; Szklarczyk et al., 2017). Considering 122 archaeal genomes and 1,626 bacterial genomes, the resulting two networks obtained by the different gene neighborhood measures were then scored by *LLS* and integrated into a single co-functional network for the gene neighborhood method.

In addition, we inferred co-functional links between barley genes from ortholog neighborhoods in metagenomes (Kim and Lee, 2017), which provide tremendous amounts of bacterial contigs. We used two distinct metagenomics resources, the Human Microbiome Project (HMP) database (Huttenhower et al., 2012) and the global ocean microbiome database from the TARA Oceans study (Sunagawa et al., 2015). We used DIAMOND, a fast sequence aligner (Buchfink et al., 2014), due to the enormous number of metagenomic contigs. Inferred co-functional links were scored by *LLS* and integrated with those based on neighborhood in fully sequenced prokaryotic genomes into a single network.

Inferring Co-Functional Links by Transferring Orthologous Gene Pairs From Other Species

Not only individual genes but also pathways are functionally conserved during speciation. Therefore, we may transfer functional information of orthologous gene pairs between species. This conserved co-functional relationship is called *associalog* (Kim et al., 2013). For protein homology mapping between barley and other species, we used InParanoid (Remm et al., 2001), which provides sensitive orthology mapping by taking account of co-orthologs. Associalogs were then transferred from a total of 21 co-functional networks for nine other species: AraNet v2 (Lee et al., 2015b), MaizeNet (Lee et al., 2019), RiceNet v2 (Lee et al., 2015a), HumanNet v2 (Hwang et al., 2018), MouseNet v2 (Kim et al., 2015), DanioNet (Shim et al., 2016), WormNet v3 (Cho et al., 2014), FlyNet (Shin et al., 2015), and YeastNet v3 (Kim et al., 2014).

Codes and Data Availability

Source codes for network search functions and edge information of BarleyNet are freely available from github (https://github.com/netbiolab/BarleyNet/).

RESULTS AND DISCUSSION

Construction of Barleynet *via the* Integration of Omics Data From Barley and Many Other Species

We inferred co-functional links between barley genes by analyzing various types of omics data obtained from cultivated

barley, three other plant species (*Arabidopsis thaliana*, *Zea mays*, and *Oryza sativa*), five animal species (human, *Mus musculus*, *Danio rerio*, *Caenorhabditis elegans*, and *Drosophila melanogaster*), and baker's yeast, *Saccharomyces cerevisiae*. Using our network evaluation scheme based on Bayesian statistics (see *Materials and Methods*), we selected networks with at least 2,000 inferred links more likely than those by random chance (i.e., $LLS > 0$). A total of 25 co-functional networks of barley genes inferred from distinct data sources (**Supplementary Table 1**) were integrated into a single final network mapping 1,272,200 co-functional associations between 26,145 barley genes (covering ~65.8% of 39,734 high-confidence genes) (**Figure 1A**). All edge information regarding the integrated BarleyNet and each of the component co-functional networks are freely available at the "Download" tab of the BarleyNet web server (www.inetbio.org/barleynet/download.php) and github (https://github.com/netbiolab/BarleyNet/),

Since a considerable amount of co-functional links were derived from other species rather than barley itself, we first summarized information sources supporting BarleyNet links (**Figure 1B**) using the UpSet visualization tool (Lex et al., 2014). We roughly classified network links into three groups based on the species of origin of the inferred co-functional association: "barley," "other plants (*A. thaliana*, *Z. mays*, or *O. sativa*)," and "animals or yeast (human, *M. musculus*, *D. rerio*, *C. elegans*, *D. melanogaster*, or *S. cerevisiae*)." We first found that the largest portion of BarleyNet information derived from co-functional association between orthologous genes in animals or yeast (579,005 links, 45.5% of all BarleyNet links). Given that many proteins are highly conserved between unicellular eukaryote yeast and multicellular eukaryote plant species, and much information is available from yeast interactomes, the large

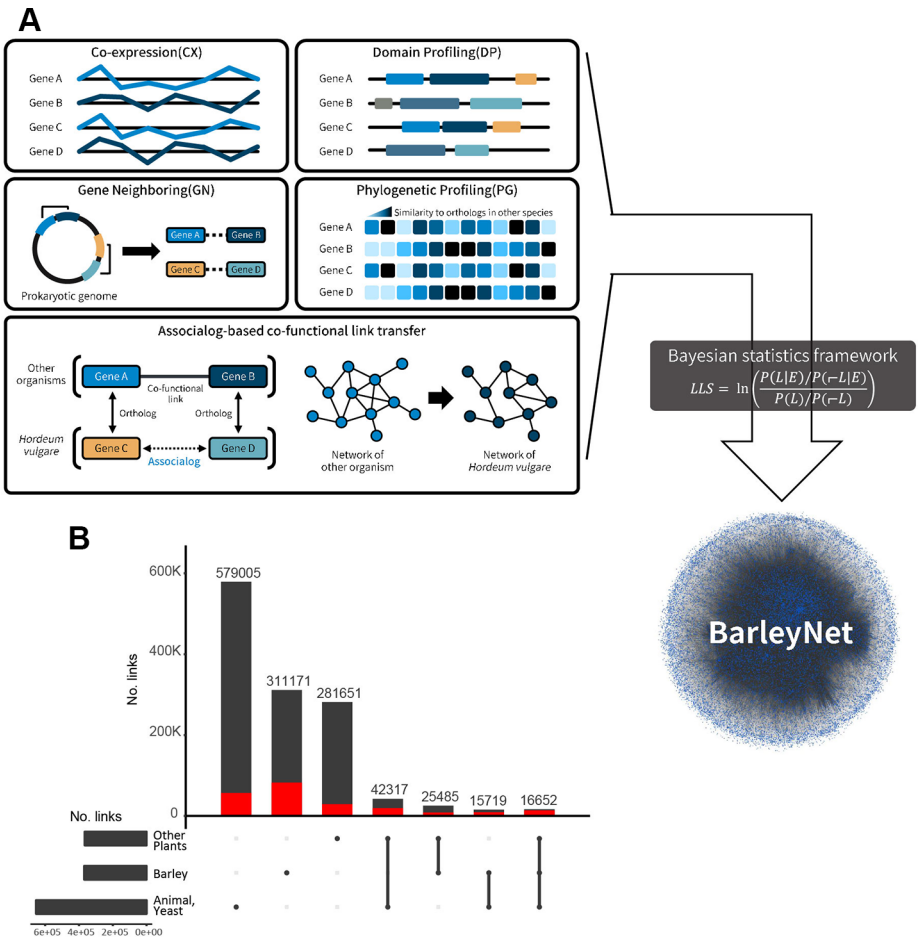

FIGURE 1 | Overview of BarleyNet. **(A)** BarleyNet was constructed by integrating functional associations between barley genes inferred from the co-expression of genes (CX), gene neighborhood (GN), association of protein domain composition profiles (DP), phylogenetic profile association (PG), and those transferred from 21 networks previously constructed for other species based on functional association between orthologous proteins (associalog). **(B)** Summary of BarleyNet edge information with UpSet visualization. Network edges were classified into three groups based on the species of origin of the inferred co-functional association: "barley," "other plants" and "animals or yeast." The bar height represents the number of BarleyNet links for each species group or their combination. The red bar represents the number of links that are 20 fold more likely than gene pairs by random (i.e., high confidence links).

observed contribution of yeast-derived information to BarleyNet was expected. In addition, we previously observed a large contribution of animal-derived information during the construction of co-functional gene networks for other plant species (Lee et al., 2010; Lee et al., 2019). Thus, we confirmed the usefulness of information derived from non-plant species in the reconstruction of a co-functional network of plant genes. Next, we observed a similar amount of co-functional links between barley genes was derived from the contribution of a group of other plant species. BarleyNet links derived from barley have a larger portion of links with high confidence (20-fold more likely than random gene pairs) than those derived from other plants (~26.5% compared with ~10%). This suggests that omics data generated from barley made critical contributions in improving the accuracy of BarleyNet. Finally, we noticed that only a small portion of BarleyNet links were supported by multiple species, although the majority of them are high confidence links (30–50% of links supported by two species groups and ~83.5% of links supported by all groups). Altogether, the contribution of different species groups to BarleyNet demonstrated the advantages of integrating omics data derived from various organisms in the construction of system-wide models with high completeness and accuracy.

Barleynet Is Highly Predictive for Biological Processes in Barley

We evaluated the overall quality and predictive power of BarleyNet. First, we assessed its accuracy against an existing barley gene network. To avoid circularity in network evaluation, we compiled a test dataset of gene pairs from the agriGO v2.0 database (Tian et al., 2017) which was not used for training the co-functional network of barley genes. The agriGO database provides gene ontology (GO) annotations for many agricultural animal and plant species, including barley. We found that gene pairs for the same GO biological process (GOBP) term comprised only 1.72% of gene pairs used for training BarleyNet, which indicates independence from the dataset used for network evaluation. The evaluation could be biased by gene pairs for GOBP terms that annotate a very large number of genes, so we ignored GOBP terms that annotated more than 1,000 genes during network evaluation. Subsequently, we compared BarleyNet and a barley network available at the STRING v11 database (Szklarczyk et al., 2019) regarding network accuracy (precision of gene pairs for the same GOBP terms) and coverage of all high confidence genes in barley (**Figure 2A**). We found that BarleyNet is substantially more accurate than the STRING database network of barley genes for any genome coverage. For example, in networks that cover 30% of the barley genome, the accuracy of BarleyNet is ~85.2% whereas that of the STRING database barley gene network is ~24.5%. Although the latter contains ~2.6 million links, it covers only 41% of all 39,734 high-confidence genes in barley, whereas the former covers ~65.8% of them. From these results, we concluded that BarleyNet is substantially more comprehensive and accurate than the STRING database network of barley genes.

Next, we evaluated the network-based gene prioritization for biological processes in barley. In an accurate and comprehensive co-functional network, the genes involved in same biological processes or pathways are highly likely to be connected by the network. If we prioritize genes for a particular pathway by network connections to the known genes of the pathway, all of the known pathway genes will be ranked generally higher than the others. Then, we may assess the network-based gene prioritization by receiver operating characteristic (ROC) analysis for the pathway genes, which can also be summarized as the area under the ROC curve (AUROC). We computed AUROC scores not only for entire ranks of predictions but also for early retrieved candidates, because only the top several hundred candidate genes are generally considered for the follow-up functional analysis in real practice. We thus computed AUROC until reaching false positive rates (FPRs) of 1% and 10%, in addition to AUROC for all predictions. We compared BarleyNet and the STRING database network of barley genes in the prediction of pathways annotated by the Plant Reactome database, ver. 59 (Gupta et al., 2016; Naithani et al., 2017), which was not used for training either BarleyNet or the STRING database network. We computed the AUROCs for 122 Plant Reactome pathways that annotate at least 10 barley genes and found that BarleyNet is significantly more predictive than the STRING database network for pathways with both early retrieved predictions and entire ranks of predictions ($P < 0.001$ by the Wilcoxon signed rank test for all comparisons, **Figure 2B**). From these results, we concluded that BarleyNet is substantially more predictive for various biological processes in barley than the existing STRING database gene network.

Since BarleyNet includes a large number of co-functional links between barley genes inferred from other species, we evaluated the contribution of network information originating from different species. For the analysis, we generated "dropout" networks that excluded the co-functional links derived from barley, plant species other than barley (*Arabidopsis*, rice, or maize), or animals and yeast (**Figure 2C**). We observed large decreases in the AUROCs for all range of FPRs by excluding links derived from barley. Notably, we observed significant decreases in the overall AUROC by excluding links inferred from other species, but not in the AUROCs for early-retrieved candidates (for FPR < 0.01 or 0.1). These results suggest that co-functional links transferred from other species by orthology contribute to the functional prediction, but not as much as those inferred from species-specific omics data sources.

We also tested robustness of BarleyNet-based functional prediction by evaluating networks with some degree of noise in network information. For the analysis, we generated 100 networks in which 20% of BarleyNet links were randomized while maintaining characteristics of network topology. Although, we observed significant decrease in AUROC with 20% of noise in network information, they were still higher than those by STRING database network (**Figure 2D**). This result suggests that BarleyNet-based functional prediction is relatively robust to some degree of noise in network information.

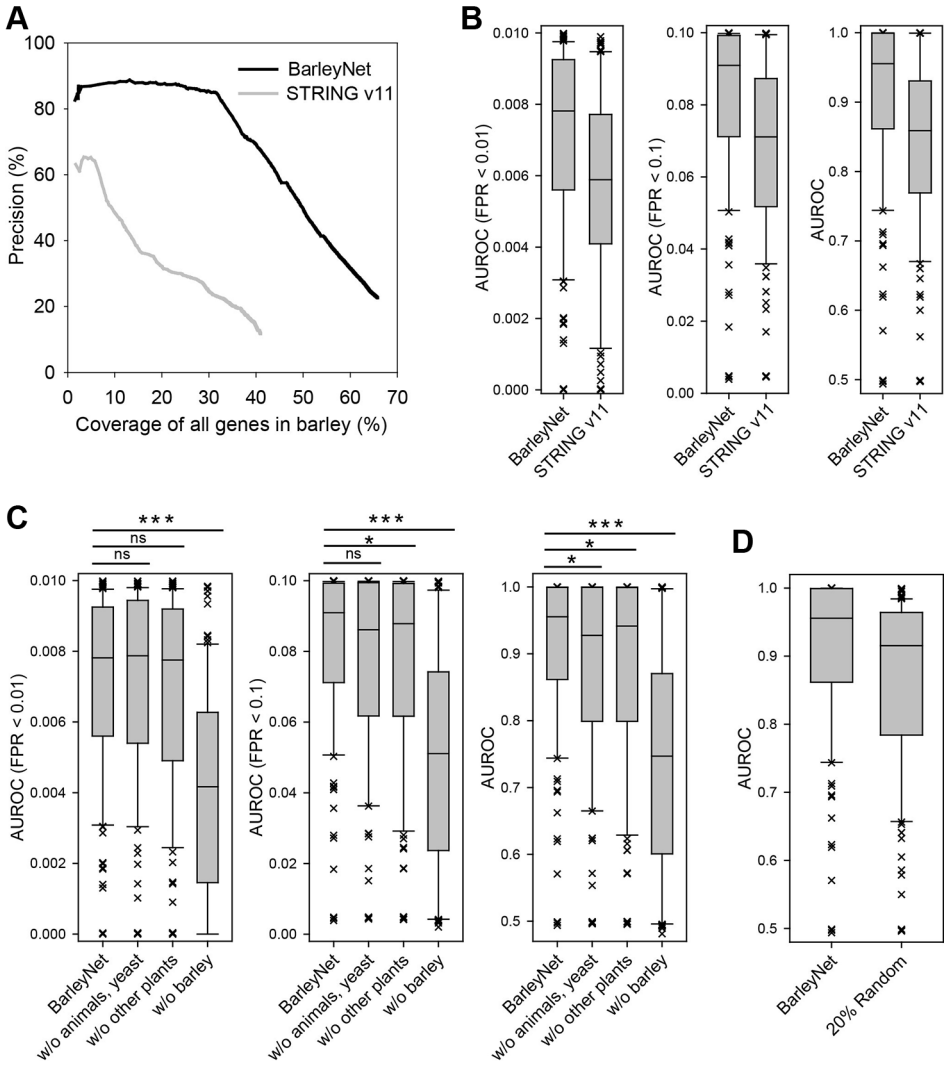

FIGURE 2 | Assessment of BarleyNet and a network of barley genes by STRING database. **(A)** The quality of the networks was evaluated based on precision for gene pairs that have the same GOBP terms by agriGO annotations and coverage of all barley genes. BarleyNet showed substantially higher precision than the network of barley genes by the STRING v11 database considering the entire range of coverage. **(B)** Comparison of area under receiver operating characteristic curve (AUROC) of 122 pathway gene sets derived from Plant Reactome database. Box-and-whisker plots represent 10%, 25%, median, 75%, and 90% of 122 AUROC scores. The same AUROC analyses were conducted until 1%, 10%, and 100% of false positive rate (FPR) were reached. BarleyNet showed a significantly higher prediction power than the STRING database barley gene network in all FPR ranges ($P < 0.001$, Wilcoxon signed rank test). **(C)** AUROC analyses were conducted as for **(B)** with BarleyNet and the following "dropout" networks by excluding links from animals and yeast (w/o animals, yeast), by excluding links from *Arabidopsis*, rice, and maize (w/o other plants), and by excluding links from barley (w/o barley). ns, not significant; *, $P < 0.05$; ***, $P < 0.001$ by Wilcoxon signed rank test. **(D)** AUROC analyses were conducted as for **(B)** with 100 networks in which 20% of BarleyNet links were randomized. Average AUROC scores for the 122 pathways gene sets across 100 networks are represented in the Box-and-whisker plot for randomized networks.

Gene Prioritization for Complex Traits Using Barleynet

The majority of omics studies on crop species aim to identify genetic components underlying economically important and complex traits such as environmental stress responses. Through the above presented benchmarking with GOBP and Plant Reactome database, BarleyNet proved to be highly predictive for pathways, but not yet for complex traits. Most human diseases are complex traits and a large portion of human disease genes were shown to be strongly associated with specific pathways (Li and Agarwal, 2009). We thus expected that genes for complex plant traits should be associated with specific pathways, and given that BarleyNet is highly predictive for pathways, it might also be predictive for complex traits. If a network is predictive for a complex trait, the genes involved in this trait might be more connected to one another than to other

genes. We thus evaluated BarleyNet in the prediction of complex traits based on the connectivity within a group of genes involved in the same traits. For this, we compiled genes for complex traits from drought-induced proteomic profiles of barley (Chmielewska et al., 2016). This study identified differentially accumulated proteins in the leaves and roots of two barley cultivars, Maresi and Cam/B1/CI (referred to as CAM), after 10 days of drought. We observed a significantly higher connectivity within a group of genes than in random gene sets of the same size in both organs of both cultivars (**Figure 3A**), which indicated that BarleyNet is significantly more predictive of drought response than random chance. The predictive power of BarleyNet for drought response was confirmed by high AUROC scores for the same groups of drought response genes (**Figure 3B**).

Considering the obtained results, we hypothesized that we might prioritize additional candidate genes for drought response through their connections to experimentally identified genes. This approach is basically a network-based search for novel candidate genes for a complex trait using previously identified genes as guides. Candidate genes were then ranked by sum of edge weight scores to the guide genes, which reflects their functional closeness. We implemented this network algorithm as a pathway-centric search method in the BarleyNet server. This server application also provides a network viewer, which visualizes a network of user-input guide genes and their closely connected neighbors. For example, **Figure 3C** shows a network of drought response genes identified from differentially accumulated proteins in CAM roots and their 50 closest neighbors. The neighbors of guide genes could be novel candidates involved in drought response in barley. Although providing a proxy for future functional studies, these candidate genes from network-based prediction should be taken with some careful consideration. The gene set analysis function of the pathway-centric search enables users to test whether these new candidates are enriched for relevant GOBP annotations. Since GOBP annotations for barley genes are still very sparse, we also employed annotations for orthologous proteins in three relatively well annotated plant species: *Arabidopsis*, rice, and maize. We found that GOBP annotations by orthology are useful in the interpretation of novel candidate genes. For example, we could not find any GOBP terms closely related to drought response among the top five enriched barley GOBP annotations. However, we found "response to heat" and "cellular response to heat," which are closely related to drought response, among the top five enriched *Arabidopsis* GOBP annotations (**Figure 3D**). Through the BarleyNet server, users can run gene set enrichment analyses for GOBP terms of all four plant species simultaneously.

A pathway-centric search provides additional information such as the list of user-input guide genes, within-group connectivity tests and AUROC analysis results for the guide gene set, as well as the list of top 100 candidate genes. By selecting a specific candidate gene, users can obtain detailed information including its connected guide genes, edge scores, data sources that support the prediction and their relative contribution, and GOBP annotations (**Figure 3E**). For example, HORVU5Hr1G072420 was a candidate drought response gene ranked 13th. The network viewer informed that six distinct data sources supported the prediction, of which yeast co-citation (SC-CC) data contributed the most (25.1% of the total prediction score). Codes for all distinct data sources are listed in **Supplementary Table 1**. Notably, the candidate genes were annotated as "response to water deprivation" in *Arabidopsis* GOBP annotation but not in barley, which demonstrates the usefulness of GOBP annotations from other plant species in the interpretation of BarleyNet predictions.

Prediction of Gene Functions Using BarleyNet

In this next step, we implemented the gene-centric search which prioritizes biological functional concepts for a gene of interest. Many proteins differentially accumulated in barley after drought stress are not yet functionally annotated. With the gene-centric search application, we can prioritize GOBP terms for genes detected in drought conditions using GOBP terms that annotate their network neighbors through information propagation. Information can be propagated to both direct and indirect neighbors in the network, and we only used the propagation to direct neighbors. We prioritized GOBP terms based on the sum of edge weight scores (log likelihood scores) to the neighbors annotated by the GOBP terms.

Figure 4A shows a screenshot of gene-centric search results for HORVU3Hr1G014120, which was differentially accumulated in CAM roots but had no GOBP annotation yet. Gene-centric search predicted "response to water" or "response to water deprivation" genes within the top five prioritized GOBP terms according to annotations for barley, *Arabidopsis*, and maize. This example clearly demonstrated that the BarleyNet gene-centric search is a useful tool in the functional interpretation of omics data in the study of complex traits of barley.

Prediction of Stress Response Genes Using Barleynet and Gene Expression Data

Finally, we provided context-centric search: a network-based prediction algorithm that uses differentially expressed genes (DEGs) along with the barley gene network to prioritize those associated with stress responses. In general, genes that respond to biotic or abiotic stresses are detected through genome-wide transcriptome profiling in which DEGs are considered to be involved in the stress response. However, some of the DEGs might play more important roles in stress response than others. Moreover, genes that do not change their transcript levels may also be involved in stress response. As discussed earlier, genes for complex plant traits such as stress response are likely to be associated with specific pathways. Therefore, we could prioritize genes involved in stress response by the changes in expression profiles of pathways they belong to. For this analysis, we pre-defined each gene and its direct neighbors in BarleyNet as subnetworks that represent pathways. We then selected subnetworks of "hub genes" that had at least 100 neighbors.

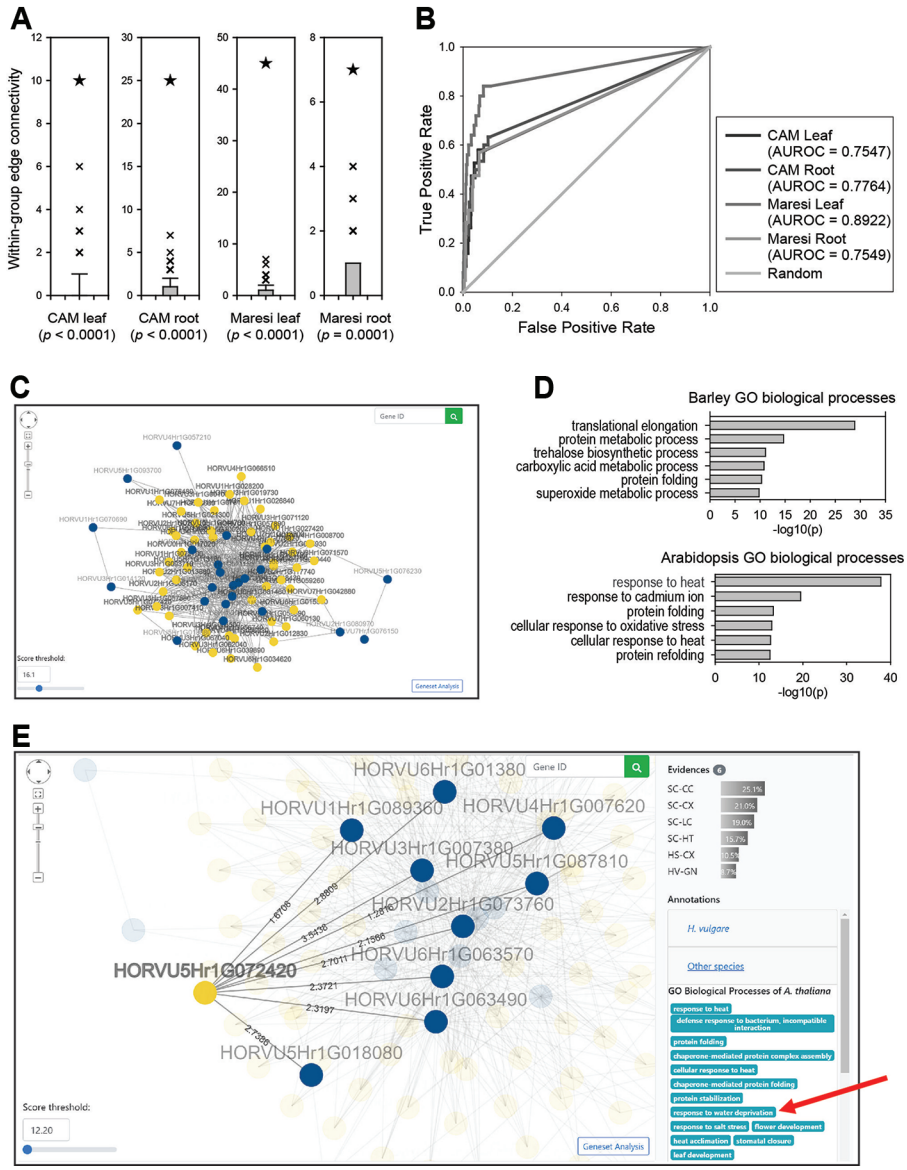

FIGURE 3 | Predictions for drought response genes using BarleyNet. **(A)** Within-group edge connectivity was computed for drought response genes identified from leaves and roots of two cultivars, Maresi and Cam/B1/CI (referred to as CAM), and 1,000 random gene sets of the same size. Asterisks indicate the within-group edge count of each trait-associated gene set in BarleyNet. Within-group edge counts for drought response genes by BarleyNet were significantly higher than those by random gene sets ($P < 0.001$ by a binomial test). **(B)** AUROC analysis for the same drought response genes. **(C)** Screenshot of network viewer, which visualizes a network of drought response genes identified from differentially accumulated proteins in CAM roots (guide genes; blue nodes) and their 50 closest neighbors (candidate genes; yellow nodes) in BarleyNet. The number of neighbors in the network can be controlled by selecting a score threshold at the bottom left area. Clicking the button at the right bottom area allows gene set enrichment analysis for the selected neighbors. **(D)** Enriched GOBP terms among the 50 closest neighbors to the drought response genes, based on barley (upper plot) and *Arabidopsis* GOBP annotations (lower plot). **(E)** Screenshot of the network viewer highlighting a selected candidate gene (yellow node), HORVU5Hr1G072420. The viewer also highlights its connected user-input guide genes (i.e., drought response genes; blue nodes) and edges with their log likelihood scores. The right-side panel shows related information such as data sources that support the prediction of HORVU5Hr1G072420 as a candidate gene (Evidences) with relative contributions (% of total prediction score), as well as GOBP annotations for the candidate gene. Notably, the selected candidate gene HORVU5Hr1G072420 was annotated for "response to water deprivation" in *Arabidopsis* GOBP annotations (marked by a red arrow).

The algorithm then computed the significance of overlap between user-submitted DEGs associated with a biological context such as stress conditions and the neighbors of each hub gene using Fisher's exact test. If the overlap between gene sets turned out to be significant, the hub gene was considered a "context-associated hub" highly likely to be involved in the biological context. The prioritized context-associated genes could be either DEGs or not.

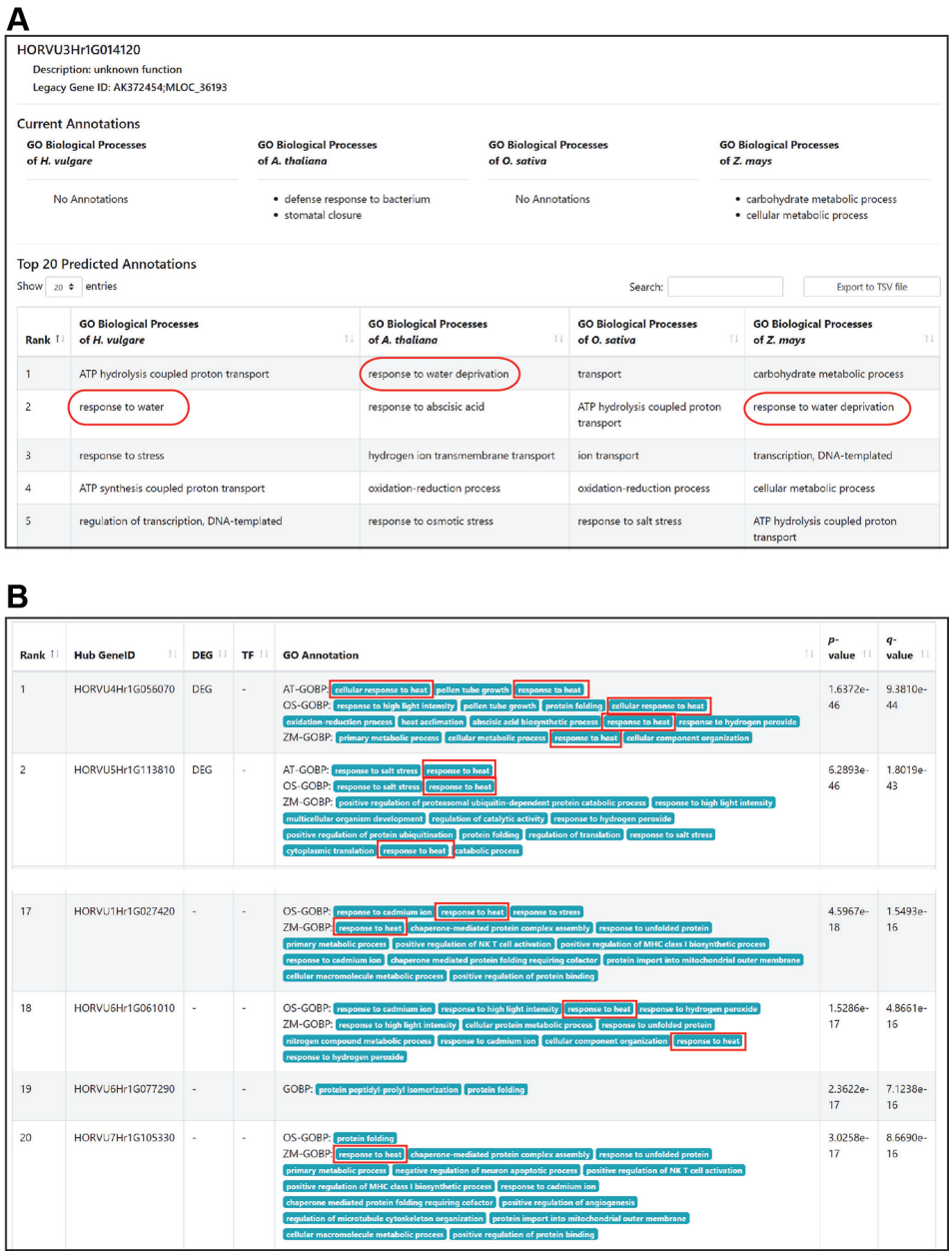

FIGURE 4 | Example results from gene-centric search and context-centric search analyses using BarleyNet. **(A)** Screenshot of BarleyNet gene-centric search results with gene HORVU3Hr1G014120, which was not annotated by barley GOBP terms. GOBP terms for drought response, "response to water" and "response to water deprivation," are marked by red circles. **(B)** Screenshots of BarleyNet context-centric search results with 625 upregulated differentially expressed genes upon heat stress in the roots of barley cultivar Rolap. The predicted genes between rank 2 and 16 were omitted. GOBP terms for heat stress response are marked by red circles.

In order to demonstrate the utility of the context-centric search application, we compiled 625 upregulated DEGs upon heat stress in barley cultivar Rolap root (Pacak et al., 2016). We manually evaluated novel candidate genes predicted by the context-centric search using the 625 upregulated DEGs (adj. p-value ≤ 0.05 and fold change ≥ 4) as input data. We found that many top ranked predictions are also DEGs that are annotated by GOBP terms for heat responses such as "response to heat" and "cellular response to heat" (**Figure 4B**). Notably, we observed candidate genes that are not DEGs but are annotated as heat response genes (see candidate genes ranked 17th, 18th, and 20th). These results clearly demonstrated that the network-based prediction along with functional genomics data facilitates the discovery of novel candidate stress response genes that could not be identified by expression profiles alone.

Because context-centric search uses network algorithm different from that of pathway-centric search, they are expected to provide different candidate genes. To investigate to what extent candidate genes vary by alternative network algorithms, we compared predictions by pathway-centric and context-centric searches for the same input genes, 30 drought response genes from differentially accumulated proteins in CAM roots. We found that 24 genes overlap between top 50 predictions from the two different network searches (48% overlap). Nevertheless, a functionally relevant GOBP term, "response to heat," was found to be enriched for both of the top 50 predictions, which indicates that both network-based methods can provide highly probable candidate genes. These results also suggest that users may use the alternative network-based methods complementarily to obtain more confident candidate genes for the follow-up functional analysis.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are freely available from https://github.com/netbiolab/BarleyNet/ and https://www.inetbio.org/barleynet.

AUTHOR CONTRIBUTIONS

SL, TL, and IL conceived the original research. SL and TL performed data analysis, constructed the network model and conducted network analysis. SY developed the web server. IL supervised the project. SL and IL wrote the manuscript with contributions from all authors. IL agrees to serve as the author responsible for contact and ensures communication.

ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) (NRF-2018M3C9A5064709, NRF-2018R1A5A2025079) to IL.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2020.00098/full#supplementary-material

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Badr, A., Muller, K., Schäfer-Pregl, R., El Rabey, H., Effgen, S., Ibrahim, H. H., et al. (2000). On the origin and domestication history of barley (Hordeum vulgare). *Mol. Biol. Evol.* 17, 499–510. doi: 10.1093/oxfordjournals.molbev.a026330

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* 34, 525. doi: 10.1038/nbt0816-888d

Buchfink, B., Xie, C., and Huson, D. H. (2014). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59. doi: 10.1038/nmeth.3176

Chmielewska, K., Rodziewicz, P., Swarcewicz, B., Sawikowska, A., Krajewski, P., Marczak, Ł., et al. (2016). Analysis of drought-induced proteomic and metabolomic changes in barley (Hordeum vulgare L.) leaves and roots unravels some aspects of biochemical mechanisms involved in drought tolerance. *Front. Plant Sci.* 7, 1108. doi: 10.3389/fpls.2016.01108

Cho, A., Shin, J., Hwang, S., Kim, C., Shim, H., Kim, H., et al. (2014). WormNet v3: a network-assisted hypothesis-generating server for Caenorhabditis elegans. *Nucleic Acids Res.* 42, W76–W82. doi: 10.1093/nar/gku367

Clough, E., and Barrett, T. (2016). "The Gene Expression Omnibus Database," in *Statistical Genomics: Methods and Protocols*. Eds. E. Mathé and S. Davis (New York, NY: Springer New York), 93–110. doi: 10.1007/978-1-4939-3578-9_5

Gürel, F., Øztürk, Z. N., Uçarlı, C., and Rosellini, D. (2016). Barley genes as tools to confer abiotic stress tolerance in crops. *Front. Plant Sci.* 7, 1137. doi: 10.3389/fpls.2016.01137

Gupta, P., Naithani, S., Tello-Ruiz, M. K., Chougule, K., D'eustachio, P., Fabregat, A., et al. (2016). Gramene database: navigating plant comparative genomics resources. *Curr. Plant Biol.* 7-8, 10–15. doi: 10.1016/j.cpb.2016.12.005

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., et al. (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234

Hwang, S., Kim, C. Y., Yang, S., Kim, E., Hart, T., Marcotte, E. M., et al. (2018). HumanNet v2: human gene networks for disease research. *Nucleic Acids Res.* 47, D573–D580. doi: 10.1093/nar/gky1126

Kim, C. Y., and Lee, I. (2017). Functional gene networks based on the gene neighborhood in metagenomes. *Anim. Cells Syst.* 21, 301–306. doi: 10.1080/19768354.2017.1382388

Kim, E., Kim, H., and Lee, I. (2013). JiffyNet: a web-based instant protein network modeler for newly sequenced species. *Nucleic Acids Res.* 41, W192–W197. doi: 10.1093/nar/gkt419

Kim, H., Shin, J., Kim, E., Kim, H., Hwang, S., Shim, J. E., et al. (2014). YeastNet v3: a public database of data-specific and integrated functional gene networks for Saccharomyces cerevisiae. *Nucleic Acids Res.* 42, D731–D736. doi: 10.1093/nar/gkt981

Kim, E., Hwang, S., Kim, H., Shim, H., Kang, B., Yang, S., et al. (2015). MouseNet v2: a database of gene networks for studying the laboratory mouse and eight other model vertebrates. *Nucleic Acids Res.* 44, D848–D854. doi: 10.1093/nar/gkv1155

Kolesnikov, N., Hastings, E., Keays, M., Melnichuk, O., Tang, Y. A., Williams, E., et al. (2015). ArrayExpress update–simplifying data submissions. *Nucleic Acids Res.* 43, D1113–D1116. doi: 10.1093/nar/gku1057

Lee, I., Date, S. V., Adai, A. T., and Marcotte, E. M. (2004). A probabilistic functional network of yeast genes. *Science* 306, 1555. doi: 10.1126/science.1099511

Lee, I., Li, Z., and Marcotte, E. M. (2007). An improved, bias-reduced probabilistic functional gene network of baker's yeast, Saccharomyces cerevisiae. *PloS One* 2, e988. doi: 10.1371/journal.pone.0000988

Lee, I., Ambaru, B., Thakkar, P., Marcotte, E. M., and Rhee, S. Y. (2010). Rational association of genes with traits using a genome-scale gene network for Arabidopsis thaliana. *Nat. Biotechnol.* 28, 149–156. doi: 10.1038/nbt.1603

Lee, T., Oh, T., Yang, S., Shin, J., Hwang, S., Kim, C. Y., et al. (2015a). RiceNet v2: an improved network prioritization server for rice genes. *Nucleic Acids Res.* 43, W122–W127. doi: 10.1093/nar/gkv253

Lee, T., Yang, S., Kim, E., Ko, Y., Hwang, S., Shin, J., et al. (2015b). AraNet v2: an improved database of co-functional gene networks for the study of Arabidopsis thaliana and 27 other nonmodel plant species. *Nucleic Acids Res.* 43, D996–D1002. doi: 10.1093/nar/gku1053

Lee, T., Hwang, S., Kim, C. Y., Shim, H., Kim, H., Ronald, P. C., et al. (2017). WheatNet: a genome-scale functional network for hexaploid bread wheat, Triticum aestivum. *Mol. Plant* 10, 1133–1136. doi: 10.1016/j.molp.2017.04.006

Lee, T., Lee, S., Yang, S., and Lee, I. (2019). MaizeNet: a co-functional network for network-assisted systems genetics in Zea mays. *Plant J.* 99, 571–582. doi: 10.1111/tpj.14341

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., and Pfister, H. (2014). UpSet: visualization of intersecting sets. *IEEE Trans. Visualization Comput. Graphics* 20, 1983–1992. doi: 10.1109/TVCG.2014.2346248

Li, Y., and Agarwal, P. (2009). A pathway-based view of human diseases and disease relationships. *PloS One* 4, e4346. doi: 10.1371/journal.pone.0004346

Mascher, M., Gundlach, H., Himmelbach, A., Beier, S., Twardziok, S. O., Wicker, T., et al. (2017). A chromosome conformation capture ordered sequence of the barley genome. *Nature* 544, 427–433. doi: 10.1038/nature22043

Mitchell, A. L., Sangrador-Vegas, A., Luciani, A., Madeira, F., Nuka, G., Salazar, G. A., et al. (2018). InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.* 47, D351–D360. doi: 10.1093/nar/gky1100

Naithani, S., Preece, J., D'eustachio, P., Gupta, P., Amarasinghe, V., Dharmawardhana, P. D., et al. (2017). Plant reactome: a resource for plant pathways and comparative analysis. *Nucleic Acids Res.* 45, D1029–D1039. doi: 10.1093/nar/gkw932

Pacak, A., Barciszewska-Pacak, M., Swida-Barteczka, A., Kruszka, K., Sega, P., Milanowska, K., et al. (2016). Heat stress affects pi-related genes expression and inorganic phosphate deposition/accumulation in barley. *Front. Plant Sci.* 7, 926. doi: 10.3389/fpls.2016.00926

Papatheodorou, I., Fonseca, N. A., Keays, M., Tang, Y. A., Barrera, E., Bazant, W., et al. (2018). Expression atlas: gene and protein expression across multiple studies and organisms. *Nucleic Acids Res.* 46, D246–D251. doi: 10.1093/nar/gkx1158

Remm, M., Storm, C. E. V., and Sonnhammer, E. L. L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* 314, 1041–1052. doi: 10.1006/jmbi.2000.5197

Shim, J. E., and Lee, I. (2016). Weighted mutual information analysis substantially improves domain-based functional network models. *Bioinformatics* 32, 2824–2830. doi: 10.1093/bioinformatics/btw320

Shim, J. E., and Lee, I. (2020). Construction of functional protein networks using domain profile associations. *Methods Mol. Biol.* 2074, 35–44. doi: 10.1007/978-1-4939-9873-9_3

Shim, H., Kim, J. H., Kim, C. Y., Hwang, S., Kim, H., Yang, S., et al. (2016). Function-driven discovery of disease genes in zebrafish using an integrated genomics big data resource. *Nucleic Acids Res.* 44, 9611–9623. doi: 10.1093/nar/gkw897

Shim, J. E., Lee, T., and Lee, I. (2017). From sequencing data to gene functions: co-functional network approaches. *Anim. Cells Syst.* 21, 77–83. doi: 10.1080/19768354.2017.1284156

Shin, J., and Lee, I. (2015). Co-inheritance analysis within the domains of life substantially improves network inference by phylogenetic profiling. *PloS One* 10, e0139006. doi: 10.1371/journal.pone.0139006

Shin, J., Lee, T., Kim, H., and Lee, I. (2014). Complementarity between distance- and probability-based methods of gene neighbourhood identification for pathway reconstruction. *Mol. Biosyst.* 10, 24–29. doi: 10.1039/C3MB70366E

Shin, J., Yang, S., Kim, E., Kim, C. Y., Shim, H., Cho, A., et al. (2015). FlyNet: a versatile network prioritization server for the Drosophila community. *Nucleic Acids Res.* 43, W91–W97. doi: 10.1093/nar/gkv453

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., et al. (2015). Structure and function of the global ocean microbiome. *Science* 348, 1261359. doi: 10.1126/science.1261359

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., et al. (2017).). 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

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. *Nucleic Acids Res.* 47, D607–D613. doi: 10.1093/nar/gky1131

Tian, T., Liu, Y., Yan, H., You, Q., Yi, X., Du, Z., et al. (2017). agriGO v2.0: a GO analysis toolkit for the agricultural community 2017 update. *Nucleic Acids Res.* 45, W122–W129. doi: 10.1093/nar/gkx382

Vullo, A., Allot, A., Zadissia, A., Yates, A., Luciani, A., Moore, B., et al. (2017). Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species. *Nucleic Acids Res.* 46, D802–D808. doi: 10.1093/nar/gkx1011

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

frontiers in Genetics

REVIEW
published: 25 May 2020
doi: 10.3389/fgene.2020.00457

Gene Regulatory Network Inference: Connecting Plant Biology and Mathematical Modeling

*Lisa Van den Broeck[1], Max Gordon[2], Dirk Inzé[3,4], Cranos Williams[2] and Rosangela Sozzani[1]**

[1] Department of Plant and Microbial Biology, North Carolina State University, Raleigh, NC, United States, [2] Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC, United States, [3] Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium, [4] VIB Center for Plant Systems Biology, Ghent, Belgium

OPEN ACCESS

Edited by:
Shizhong Han, Johns Hopkins Medicine, United States

Reviewed by:
Jiri Vohradsky, Institute of Microbiology (ASCR), Czechia
Justin William Walley, Iowa State University, United States

***Correspondence:**
Rosangela Sozzani
rsozzan@ncsu.edu

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

Received: *05 November 2019*
Accepted: *14 April 2020*
Published: *25 May 2020*

Citation:
Van den Broeck L, Gordon M, Inzé D, Williams C and Sozzani R (2020) Gene Regulatory Network Inference: Connecting Plant Biology and Mathematical Modeling. Front. Genet. 11:457. doi: 10.3389/fgene.2020.00457

Plant responses to environmental and intrinsic signals are tightly controlled by multiple transcription factors (TFs). These TFs and their regulatory connections form gene regulatory networks (GRNs), which provide a blueprint of the transcriptional regulations underlying plant development and environmental responses. This review provides examples of experimental methodologies commonly used to identify regulatory interactions and generate GRNs. Additionally, this review describes network inference techniques that leverage gene expression data to predict regulatory interactions. These computational and experimental methodologies yield complex networks that can identify new regulatory interactions, driving novel hypotheses. Biological properties that contribute to the complexity of GRNs are also described in this review. These include network topology, network size, transient binding of TFs to DNA, and competition between multiple upstream regulators. Finally, this review highlights the potential of machine learning approaches to leverage gene expression data to predict phenotypic outputs.

Keywords: gene regulatory network, network properties, network inference, machine learning, experimental methodologies

FROM GENES TO NETWORKS: A CONTINUOUS MOLECULAR SCALE FOR PLANT RESEARCH

Plant responses need to integrate environmental signals, including those from biotic and abiotic stresses. Additionally, plants integrate intrinsic signals, such as developmental or hormonal cues. Plant responses to environmental and intrinsic signals are under tight control to ensure a fast and appropriate response and at the same time prevent an indiscriminate activation of this response (Swift and Coruzzi, 2017). Accordingly, the chance of randomly activating a plant response is significantly reduced when multiple transcription factors (TFs) regulate and fine-tune this response (Swift and Coruzzi, 2017).

As such, multiple upstream TFs, connected to each other, form complex gene regulatory networks (GRNs) to redundantly control downstream responsive genes, also defined as target genes (Hernando et al., 2017). These GRNs consist of nodes that represent genes, and edges that represent the regulatory connections between genes. Overall, GRNs provide a blueprint of the molecular interactions underlying plant responses. The generation of GRNs in the context of plant responses has played a critical role in identifying new regulatory connections between genes and driving novel hypotheses. For example, the generation of a GRN at the base of the myo-inositol metabolic pathway in soybean (*Glycine max*) predicted new regulatory interactions, of which 13 interactions could be validated. The GRN was generated with transcriptome data from two mutant lines, *mips1* (*myo-inositol phosphate synthase 1*) and a triple mutant *mips1/mrp-l (multi-drug resistance protein)/mrp-n* that led to low phytic acid and a decrease in seed emergence (Redekar et al., 2017). More specifically, differentially expressed genes (DEGs) were clustered in modules based on their expression patterns. Putative regulatory interactions between the DEGs encoding TFs and the different modules were then determined based on the enrichment of known DNA-binding motifs within each module (Redekar et al., 2017). By using a systems-level approach, unknown regulatory interactions were predicted and validated, allowing for a better understanding of the myo-inositol metabolic pathway in soybean.

In another example, newly identified hub genes, i.e., highly connected genes, were hypothesized to have functional roles as stress-induced genes (Vermeirssen et al., 2014). To generate the stress-induced GRN, an *Arabidopsis* microarray compendium including 199 abiotic stress conditions was used to identify modules of co-expressed genes. Using three different network inference techniques, a set of putative upstream TFs was identified for each module resulting in a total of 200,014 regulatory interactions. Fifty percent of the predicted regulatory interactions involving seven identified hub TFs were confirmed, highlighting the capacity of GRNs to identify functional interactions (Vermeirssen et al., 2014). Furthermore, one of these seven TFs, NAC DOMAIN CONTAINING PROTEIN 32 (NAC032), was not yet shown to play a role in stress tolerance. Phenotypic analyses confirmed the involvement of NAC032 in the regulation of the osmotic stress response, demonstrating the power of GRNs to identify regulatory TFs in a biological context (Vermeirssen et al., 2014).

In addition to identifying new regulatory connections between genes with GRNs, the assessment of GRN topology can provide a system-level approach to understand network complexity and robustness, and help in identifying putative strategies for manipulating the network response. The network topology refers to the structure of the GRN and includes properties such as node connectivity, network diameter, network density, and network motifs (Hu et al., 2005). Node connectivity is the number of connections a node has to other nodes. Network diameter measures the number of connections between the most distant parts of the network. Network density is a measure of the number of connections in a network in proportion to the number of nodes. Lastly, network motifs are subgraphs that occur within a GRN with high occurrence. These aspects of network topology contribute to the understanding of network robustness and complexity.

BIOLOGICAL PROPERTIES OF GENE REGULATORY NETWORKS AND APPROACHES TO INVESTIGATE THEM

As mentioned above, complex GRNs can be identified that contribute to plant development and environmental responses. Several biological properties, including network topology, contribute to the complexity of GRNs and can be assessed when studying GRNs:

1. *Multiple upstream regulators*: Many genes are regulated by multiple upstream TFs, resulting in a complex regulatory module for every gene (Barah et al., 2016; Huang et al., 2017). Moreover, upstream TFs can act alone, form complexes, compete for binding, and act as a co-factor with or sequester other TFs (Nagel and Kay, 2012). In addition to the high number of upstream regulators, some TFs only regulate a downstream gene in combination with another TF and/or under specific conditions (Gonzalez et al., 2015). Such interactions are thus overlooked in the absence of the second TF. Furthermore, it has been shown that TFs bind to different motifs when paired with other TFs than motifs bound by single TFs, further increasing network complexity (Jolma et al., 2015). How multiple upstream TFs regulate the expression of one target gene is thus highly complex. Currently, transient luciferase assays (TEAs) can be used to quantify the effect of multiple TFs on the expression of a target gene (Vanden Bossche et al., 2013). Accordingly, by transforming protoplasts with multiple effector plasmids containing the TFs of interest and one reporter plasmid with the promoter of the target gene of interest, the combined effect of these TFs on the activity of the promoter can be evaluated. This information can be used to refine the network.
2. *Transient binding*: Transcription factors scan the DNA until they encounter the correct DNA-binding motif and bind to the DNA, which can occur transiently. A TF can execute its function through the hit-and-run principle, which means that once the TF is bound (*hit*), it establishes a transcriptional complex that regulates transcription even when the TF is no longer present (*run*) (Doidy et al., 2016; Swift and Coruzzi, 2017). Because these transient bindings occur within minutes and do not last, they are harder to detect by methods such as chromatin immunoprecipitation (ChIP), resulting in false negatives in the GRN. Performing ChIP experiments with an inducible system over multiple time points can decrease the number of false negatives (Doidy et al., 2016; Swift and Coruzzi, 2017). As such, a new class of target genes that is only transiently bound by basic LEUCINE ZIPPER 1 (bZIP1) within 1 to 5 min and not at later time points was discovered (Para et al., 2014).

3. *Size*: Depending on the molecular process, the network size can increase significantly, reaching hundreds of genes in one network. Researchers can reduce the number of genes in the network by (i) increasing the fold change or decreasing the *q*-value threshold to select a smaller subset of DEGs, (ii) focusing on a specific type of protein such as TFs, or (iii) performing an overlap with DEGs from other relevant datasets. To visualize, explore, and analyze these networks, regulatory interactions can be uploaded in Cytoscape® and analyzed with different applications such as BiNGO or NetMatch* (Su et al., 2014). Generally, these large-scale networks include hub genes with a high out-degree, i.e., the number of outgoing edges and thus the number of target genes (Lorenz et al., 2011; Barah et al., 2016). Such hub genes can be biologically important genes and thus relevant for further studies characterizing gene function.
4. *Network topology*: Within a GRN, multiple network motifs, such as feedback and feedforward loops, are found (Nohales and Kay, 2016). These network motifs can exhibit specific dynamic characteristics (**Figure 1**). Depending on the network motif, delayed, transient, or increased activation of target genes can occur (**Figure 1**; Martin et al., 2016). Thus, as a result of their dynamic behavior, network motifs contribute to GRN dynamics and complexity (**Figure 1**). As shown in **Figure 1**, multiple snapshots of the transcriptomes can be detected depending on the sampled time point (**Figure 1**). These characteristics were highlighted in Chang et al., where ChIP-seq data identifying EIN3 targets upon ethylene treatment were combined with RNA-seq analysis to construct a GRN (Chang et al., 2013). Because samples were taken at multiple time points after ethylene treatment, the dynamics of the response to ethylene could be unraveled. This study shows the power of time courses to unravel the dynamics of a GRN and view the progression of the downstream events (Chang et al., 2013).

The latter network topology also contributes to the phenotypic output of plant responses. For example, incoherent feedforward loops will generate pulses of gene expression, which in turn generate rhythmic behaviors, such as the circadian clock in *Arabidopsis* (Joanito et al., 2018). Studying phenotypic outputs is commonly achieved by eliminating or overexpressing a single gene or several genes. However, studying phenotypic outputs in the context of entire GRNs appears to be more challenging, and additional tools may be necessary to connect network characteristics and plant phenotype.

EXPERIMENTAL METHODOLOGIES TO GENERATE GENE REGULATORY NETWORKS

To reach a comprehensive understanding of plant responses, multi-level data, ranging from phenotypic analyses to gene expression analyses, are being acquired. Advances in bioinformatics and high-throughput experimental approaches, such as RNA sequencing and ChIP sequencing, allow us to study whole transcriptomes. This variety of data can be used to study genes across a molecular scale, ranging from a single gene, several genes, or interacting genes forming a GRN. A variety of experimental methodologies are used to collect data for the generation of GRNs and provide a system-level view of the plant response under study (**Figure 2**). These methodologies can (i) determine the binding of a TF to specific DNA sequences or (ii) identify target genes that are regulated by a TF of interest. Based on this information, directional edges can be drawn from the genes encoding TFs to their downstream targets.

Methodologies to identify DNA binding events of TFs are yeast one-hybrid (Y1H) assays, ChIP experiments and *in vitro* DNA binding assays (**Figure 2**). These methodologies are frequently used in studies focusing on the detailed characterization of a single gene or a small group of genes. Additionally, they can be applied in a systems-level context when performed in parallel.

- **Y1H Screens.** A large-scale Y1H screen that tested the promoters of 50 genes involved in xylem development against 467 TFs was used to construct a GRN at the base of secondary cell wall synthesis (Taylor-Teeples et al., 2015). This Y1H screen resulted in a highly interconnected GRN containing feedforward loops and led to the identification of new key TFs in the specification of the secondary cell wall (Taylor-Teeples et al., 2015). Another recently published GRN constructed from Y1H screens unraveled a GRN downstream of plant cell regeneration; subdivided this GRN in wounding, auxin, or cytokine-induced regeneration subnetworks; and identified hub TFs and novel promoter–TF interactions (Ikeuchi et al., 2018). Even though Y1H assays allow for high-throughput data generation of direct TF-DNA binding to construct GRNs, the yeast genetic background can affect the results and the identified regulatory interactions should be confirmed *in planta*.
- **ChIP.** When performing ChIP followed by high-throughput sequencing (ChIP-seq) or microarray hybridization (ChIP-chip), genome-wide TF binding loci can be determined. Although ChIP-seq is limited to one TF, the technique can be used to build GRNs when performed in parallel. A recently published study performed ChIP-seq experiments on 21 TFs related to abscisic acid (ABA) in the presence and absence of ABA, enabling the identification of dynamic TF binding; for 19 of the 21 TFs, the binding events increased after ABA treatment (Song et al., 2016). Because the authors determined the direct downstream targets of 21 TFs, they could identify highly regulated target genes that were downstream of multiple TFs, such as core ABA genes but also novel non-ABA-related genes, such as *RGL3* (*RGA-like 3*) regulated by gibberellin (GA) and *ACS2* (*ACC synthase 2*) controlling the biosynthesis of ethylene (Song et al., 2016). Expresso is available to explore and

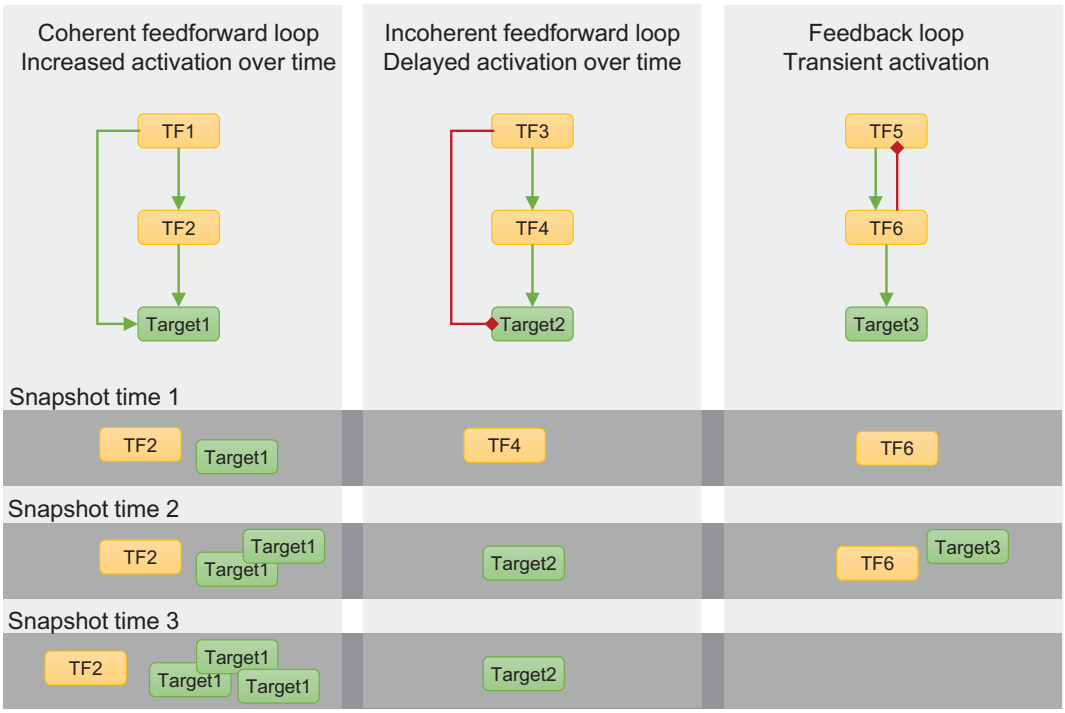

FIGURE 1 | Schematic representation of multiple snapshots of the transcriptomes in relation to the presence of network motifs, such as feedforward and feedback loops. **Left panel:** a coherent feedforward loop composed of activation interactions results in increased activation of the target gene over time as the induction of the second transcription factor (TF) only occurs after its own activation by TF1. **Middle panel:** delayed activation of target2 as a result of the delayed activation of TF4, part of an incoherent feedforward loop. **Right panel:** as a result of the feedback loop between TF5 and TF6, target3 is only transiently activated. These interactions also depend on the relationship between the two TFs, the degradation of the transcripts, and the amount of input signal. The observed transcriptomes will thus be different over multiple time points and result in different snapshots (dark gray zones). Green and red arrows represent activation and repression, respectively.

access available processed ChIP-seq data in *Arabidopsis* (Aghamirzaie et al., 2017).

– ***In vitro* DNA-Binding Experiments.** As with Y1H assays, this methodology can be used to construct GRNs; however, the large number of regulatory interactions found with these techniques are not always functional and need to be placed in a biological context. *In vitro* techniques used to determine DNA binding events of TFs include protein binding microarrays (PBM), DNA-affinity purification sequencing (DAP-seq), and Systematic Evolution of Ligands by Exponential Enrichment (SELEX). PBMs consist of dsDNA microarrays that are incubated with a tagged TF of interest. The DNA-bound TFs are detected with a fluorescent-bound antibody (Berger and Bulyk, 2009). Using PBMs, the DNA-binding motif of 2913 TFs, selected from different species, was determined in a large-scale experiment (Weirauch et al., 2014). These data are publicly available at Cis-BP[1] and forms a large resource for bioinformatics analysis and GRN inference. DAP-seq and SELEX are similar techniques; however, to our knowledge SELEX has not been used to build a GRN in plants. For SELEX, a target (e.g., TF) is incubated with a library, e.g., a synthetic library or a genome-based library of ssDNA, dsDNA, or RNA, followed by the selection and amplification of the bound complexes (Djordjevic, 2007). DAPseq makes use of a dsDNA library (inferred from genomic DNA) of which the fragments contain an adaptor sequence. A purified TF bound to beads is added to the library. Next, the bound gDNA fragments are eluted and sequenced. By mapping the sequence reads onto the genome, bound target genes can be identified (Bartlett et al., 2017). The *in vitro* DNA-binding sites of 526 *Arabidopsis* TFs are determined with DAP-seq (O'Malley et al., 2016)[2].

In addition to constructing a GRN based on the binding events of a TF, gene expression data of inducible overexpressing plant lines can be used to build GRNs (**Figure 2**). The major advantage of inducible overexpressing lines is that the desired gain or loss of function can be applied at a specific developmental stage, resulting in temporal or developmental specific GRN changes. Three inducible systems are generally used. (i) TFs translationally fused to a glucocorticoid receptor (GR) domain translocate to the nucleus in the presence of dexamethasone (DEX) (Corrado and Karali, 2009). The two other systems make use of a two-component system in which a chimeric TF induces

[1]http://cisbp.ccbr.utoronto.ca

[2]http://neomorph.salk.edu/dev/pages/shhuang/dap_web/pages/index.php

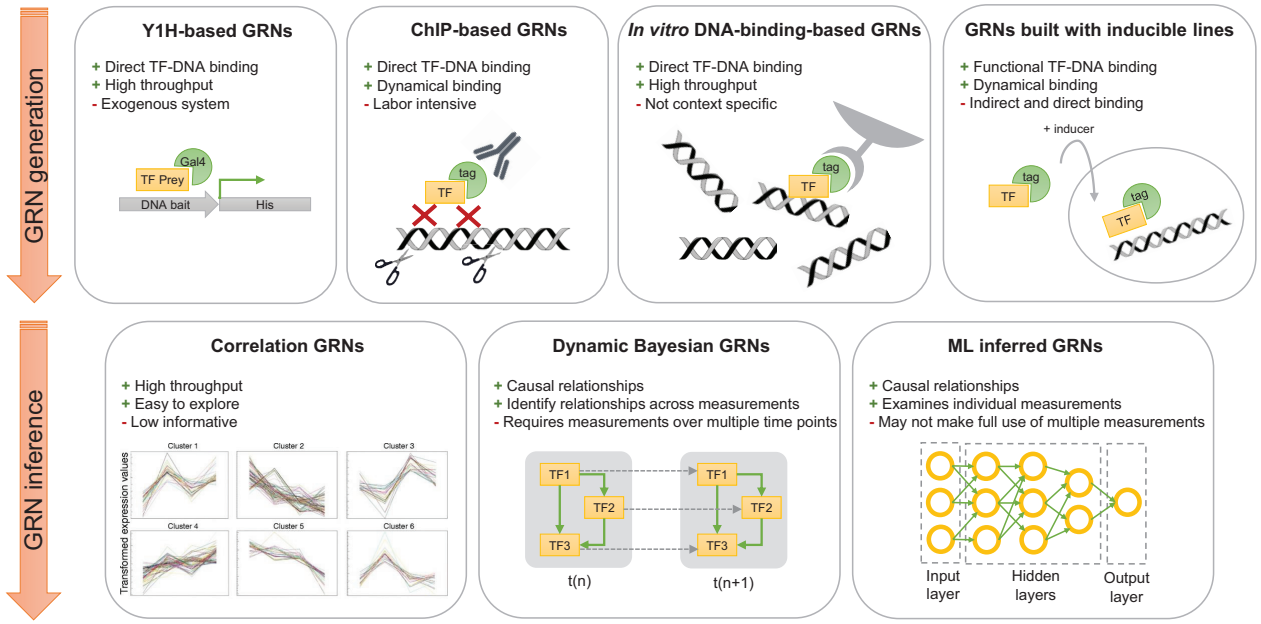

FIGURE 2 | Overview of network generation and inference methodologies described in this review. High-throughput Y1H screens, ChIP-seq assays, *in vitro* DNA-binding experiments, or expression experiments with inducible overexpressing lines can be used to generate GRNs. Three computational methodologies are described in this review to infer GRNs: Correlation networks, Dynamic Bayesian networks, and machine learning networks. Advantages and disadvantages are given for each experimental and computational methodology. GRN, gene regulatory network; ML, machine learning; Y1H, yeast one-hybrid; ChIP, chromatin immunoprecipitation.

the expression of the transgene upon a chemical inducer. (ii) First, a fusion protein, called XVE, contains a LexA DNA binding domain, the VP16 transactivation domain, and the human estrogen receptor domain and is activated when treated with estrogen (e.g., estradiol). Subsequently, the fusion protein can activate the expression of the TF of interest by binding on the LexA operator sequence upstream of the gene encoding the TF (Zuo et al., 2000). (iii) The third system, called the *alc* system, also contains two components: the first component is the AlcR TF activated in the presence of ethanol or acetaldehyde and the second component consists of the gene encoding the TF of interest downstream of the AlcA promoter. When the AlcR is active, it can bind the AlcA promoter and induces the expression of the TF of interest (Caddick et al., 1998).

These systems have been used to overexpress a gene of interest at a desired developmental stage and explore their downstream effects with, e.g., transcriptomics (Wellmer et al., 2006; Dubois et al., 2013). For example, *APETALA1* (*AP1*), a central gene in the initiation of flower development, was fused to a GR-domain and transformed into the *ap1 cal* (*cauliflower*) double mutant. By specifically activating *AP1* in the inflorescence meristems of this mutant, the temporary obstruction of flower formation in *ap1 cal* is lifted and flowers develop synchronously (Wellmer et al., 2006). In addition to inducing TFs, a system has been developed in which artificial microRNAs (amiRNAs) are specifically induced during flower development, generating new possibilities to unravel GRNs (O'Maoileidigh et al., 2015).

These GRNs contain experimentally determined transcriptional regulations but do not make a distinction between indirect or direct targets. By using cycloheximide in combination with inducible overexpressing lines, indirect and direct target genes can be distinguished. Cycloheximide will block the formation of new proteins, preventing direct targets to in turn regulate their targets and thus the detection of indirect target genes (Davies and Exworth, 1973). Based on these principles, the technique TARGET (Transient Assay Reporting Genome-wide Effects of Transcription factors) was developed (Bargmann et al., 2013). Protoplasts are transformed with a GR-TF fusion cassette that also contains a red fluorescent protein (RFP), enabling the sorting of transformed protoplast through fluorescence-activated cell sorting (FACS). With the addition of 4-thiouracil (4tU), a distinction can be made between existing and newly synthesized mRNA (Doidy et al., 2016). Using this technique, the "hit-and-run" principle was proven for bZIP1 (Para et al., 2014). However, some genes are transcriptionally induced by cycloheximide, which can render false positive. In this case, including early and later time points upon induction of overexpression can indicate whether DEGs are direct or indirect downstream targets (Van den Broeck et al., 2017). As such, the regulatory effect of 21 TFs on their downstream targets was assessed upon multiple time points after induction of overexpression. Genes differentially expressed 1, 2, and 4 h after overexpression were selected as putative direct targets and experimentally validated. The validated targets were used to

construct a GRN that is specifically activated upon osmotic stress (Van den Broeck et al., 2017).

The above-described methodologies use experimental data ranging from Y1H screens to expression data, to construct GRNs. However, these methodologies introduce uncertainties as a result of incomplete observations, background noise, and systematic errors, leading to false negatives. To this end, researchers can make use of network inference approaches to describe regulatory interactions as probabilities and built GRNs.

PROBABILISTIC NETWORK INFERENCE APPROACHES TO IDENTIFY CAUSAL RELATIONS

The inference of GRNs from large datasets is not an easy task, and different computational tools, including correlation networks, and causal inference methods such as Mutual Information and Bayesian networks, have been applied to this task (Margolin et al., 2006; Vignes et al., 2011). Co-expressed genes can be identified from microarray or RNAseq data with correlation methods, such as Pearson or Spearman correlation. This information can then be used to build correlation networks (**Figure 2**). These correlation networks are based on the principle that genes expressed in the same conditions could perform a similar biological function. Correlation networks can thus be powerful tools to predict new regulatory genes of a specific plant response. For example, a correlation network in rice was built based on 57 microarray experiments performed during different stages of anther development. This resulted in 545 clusters, with genes showing the same expression pattern across the different samples (Lin et al., 2017). By mapping DEGs identified with knock-out experiments onto the correlation network, new biologically important genes involved in anther development were identified. GRNs have been developed for a large number of species under different environmental conditions and multiple tools are available to explore correlation networks or identify sets of co-expressed genes (**Table 1**; De Bodt et al., 2010).

Correlation networks can be used to explore large datasets and identify putative central regulators/hub genes (**Figure 2**). However, these networks are unable to provide information about transcriptional relations between upstream regulators and downstream target genes. They are also limited in determining whether the interaction is direct or indirect, results in activation or repression, or involves competition between multiple upstream regulators. One technique to provide useful predictions using correlation networks despite this limitation is to integrate additional types of data. For example, combining correlation networks with metabolic data has led to the identification of key regulatory genes in metabolic pathways (Wu et al., 2016). The addition of genome-wide association studies (GWAS) can increase the power and robustness of a correlation network. A correlation network at the base of mild and severe salt stress response in roots was constructed in parallel with a GWAS of a 94-RIL (Ler/Cvi) population. Genes identified with GWAS were used to explore the clusters of the correlation network. By analyzing the neighboring genes of the identified GWAS hits, connections could be made, such as the allocation of GWAS and neighboring genes identified under mild salt stress to specific clusters (Kobayashi et al., 2016). Leveraging the advantage of combining GWAS with correlation networks, a computational framework, Camoco, was built to identify candidate SNP-associated genes, build a correlation network, and prioritize the candidates genes based on their expression correlation (Schaefer et al., 2018). This approach is especially useful for species for which the majority of the genome remains functionally uncharacterized. Other methods that integrate correlation networks with additional data are based on known DNA-binding motifs to identify the upstream regulators of a group of DEGs that cluster together (Palaniswamy et al., 2006; Lv et al., 2014; Barah et al., 2016). The TF2Network tool is such a method that allows constructing a GRN based on DNA-binding motifs by searching in a given list of genes for enriched TF-binding sites (Kulkarni et al., 2017).

While correlation networks are an adaptable and widely used computational tool, other methods are necessary to infer causal relationships from gene expression without the use of DNA-binding motifs. Using network inference methods, putative upstream regulators for DEGs can be predicted by searching for regulators that can explain observed gene expression patterns, allowing the researcher to construct a GRN (Segal et al., 2003; Phuong et al., 2004). Bayesian network (BN) inference provides one avenue to construct large, informative GRNs and infer direct causal relations between genes (**Figure 2**; Yu et al., 2004; Chen et al., 2006; Bansal et al., 2007; Vignes et al., 2011). In BNs, edges are encoded as probabilistic connections between their origin and destination nodes (Pearl, 2008). These networks are a particularly widely used tool in determining conditional dependencies among genes to predict direct interactions between an upstream gene and its downstream targets (Yu et al., 2004; Chen et al., 2006; Bansal et al., 2007; Vignes et al., 2011). In one example, a BN was used to infer conditional dependencies among *SHOOT MERISTEMLESS* (*STM*) and 56 other genes encoding TFs with publicly available datasets in *Arabidopsis*. With this network a strong dependency was identified between *STM* and *CUP-SHAPED COTYLEDON 1* (*CUC1*), which was then experimentally validated (Scofield et al., 2018). Importantly, BNs can be constructed by beginning with a set of genes of interest and iteratively adding genes that lead to a model with increased fitness. Using this approach, several *GATA* TFs were identified as possible regulators of photosynthesis in *Arabidopsis* and novel relationships were tested (Needham et al., 2009).

To lower the number of possible networks and thus sometimes extensive computation time, network inference based on Bayesian principles can make use of *a priori* knowledge about the pathway. *A priori* knowledge can be incorporated in ways such as restricting possible network structures based on known patterns of interaction or limiting the number of connections any node may have. For example, Bayesian inference with an assumption of hierarchical structure and a limited number of connections was applied to infer GRNs in *Arabidopsis* under different stress conditions. These networks identified 9 TFs as putative regulators of *DESICCATION-RESPONSIVE PROTEIN*

TABLE 1 | Summary of the available tools to explore expression datasets in different species.

Tool	Species	Specificity	References
CORNET	*Arabidopsis thaliana*	Co-expression and protein-protein interaction tool	De Bodt et al., 2010
FlowerNet	*Arabidopsis thaliana*	Includes only stamen-, pollen-, or flower-specific expression studies	Pearce et al., 2015
Genevestigator	*Arabidopsis thaliana, Hordeum vulgare, Oryza sativa, Medicago truncatula, Glycine max, Zea mays, Nicotiana tabacum, Solanum lycopersicum, Physcomitrella patens, Triticum aestivum,* and *Sorghum bicolor*	Multiple tools to analyze a set of genes, such as clustering and differential expression	Hruz et al., 2008
RapaNet	*Brassica rapa*	Includes 143 B. rapa microarrays	Kim et al., 2017
RiceAntherNet	*Oryza sativa*	Includes 57 rice anther tissue microarrays	Lin et al., 2017
RiceArrayNet/PlantArrayNet	*Oryza sativa, Arabidopsis thaliana, and Brassica rapa*	Includes diverse microarrays and links genes to pathway maps	Lee et al., 2009
PlantExpress	*Oryza sativa and Arabidopsis thaliana*	Contains two sub platforms, OryzoExpress and ArthaExpress, enabling cross-species analysis	Kudo et al., 2017
ATTED-II	*Arabidopsis thaliana, Brassica rapa, Oryza sativa, Glycine max, Populus trichocarpa, Solanum lycopersicum, Vitis vinifera, Medicago truncatula*, and *Zea mays*	Includes microarray data of crops and added RNAseq data of *Arabidopsis*	Obayashi et al., 2014, 2018
PlaNet	*Arabidopsis thaliana, Hordeum vulgare, Medicago truncatula, Populus trichocarpa, Oryza sativa, Glycine max, Triticum aestivum, Nicotiana tabacum, Brachypodium distachyon, Physcomitrella patens, and Selaginella moellendorffii*	Comparative analysis of co-expression networks across plant species and prediction of gene function	Mutwil et al., 2011
PLANEX	*Arabidopsis thaliana, Glycine max, Hordeum vulgare, Oryza sativa, Solanum lycopersicum, Triticum aestivum, Vitis vinifera, and Zea mays*	Contains microarray data from the Gene Expression Omnibus (GEO)	Yim et al., 2013

Different tools are developed to identify sets of co-expressed genes across a wide range of environmental conditions or mutant lines and explore these regulatory modules. Each tool has overlapping and distinct features.

29A (*RD29A*), a well-known stress-induced gene, in agreement with previous experimental data (Penfold et al., 2012).

Another method to infer regulatory relationships is the use of ordinary differential equation (ODE) models. These approaches are based on fitting parameterized differential equations to time-course expression data, where these equations characterize the dynamic influence of regulators on the expression patterns of target genes. These equations typically describe mechanistic interactions between regulators and targets and can vary in complexity, ranging from linear equations to more complex non-linear representations (Wu et al., 2014). Given a specific model type and time-course gene expression data, optimization routines are used to estimate the parameters of the ODE. These include least-squares methods, LASSO, Markov Chain Monte Carlo, and Genetic Algorithms (Locke et al., 2005, 2006; Krouk et al., 2010; Koryachko et al., 2019). Issues that arise when using ODEs to model GRNs include overly complex models resulting in overparameterization, sparse data resulting in unidentifiable parameters (Krouk et al., 2010), overfitted parameters resulting in models that are not generalizable (Krumsiek et al., 2010), and model structures that result in "sloppy" parameters where a wide range of parameters provide adequate fit to the data (Bujdoso and Davis, 2013). ODE models are also typically constrained to a subset of DEGs to reduce the numbers of parameters that need to be optimized. Putative upstream regulators of genes involved in the response to different light conditions in *Arabidopsis* were selected based on literature, databases such as Kyoto Encyclopedia of Genes and Genomes (KEGG), and regulator-gene predictions based on motif presence in promoter regions. Fitting ODE models to time-course expression data allowed for the removal of weak regulatory interactions and the refinement of a GRN under photosynthetic light acclimation (Yao et al., 2011). Similarly, an ODE model incorporating hidden states to represent actual protein abundances was used to infer GRNs related to nitrate response in *Arabidopsis*. In this study, SPL9 was identified as a possible regulator of nitrate signaling and experimentally validated by overexpressing *SPL9* (Krouk et al., 2010).

Importantly, each inference technique has specific advantages and limitations. For example, Bayesian inference methods are well-suited to extract useful information from noisy gene expression data and to identify linear cascades (Marbach et al., 2012). However, they cannot scale to infer large networks and

are limited in identifying feedforward loops (Marbach et al., 2012). These shortcomings can be addressed by performing a clustering step prior to inference (de Luis Balaguer et al., 2017) and extending the BN into a Dynamic Bayesian Network (DBN), respectively (Friedman et al., 1998). In DBN inference, a time-course dataset is provided to predict probabilistic dependencies between genes. As such, the value of each gene at one time point depends on the values of its regulators at the previous time point and/or at the same time point, depending on the sparsity of the time-course data that is provided. DBNs have been used to predict mechanisms that are key in regulating circadian rhythms in *Arabidopsis*. These were later confirmed in experimentally verified networks (Dondelinger et al., 2012). Moreover, DBNs have successfully been used to infer GRNs underlying molecular responses and reconstruct experimentally determined stem cell networks. Accordingly, a DBN inferred from root stem cell-specific time-course data identified *PERIANTHIA* (*PAN*) as an upstream of known stem cell regulators. Experimental evidence showed that this newly predicted stem cell regulator indeed controls columella stem-cell maintenance and QC division (de Luis Balaguer et al., 2017). Importantly, the computational pipeline used in this work, called GENIST, was made available on GitHub and through TuxNet, a simple graphical user interface for processing of RNAseq data and inferring GRNs (de Luis Balaguer et al., 2017; Spurney et al., 2019). In addition to TuxNet, other tools are available to facilitate the use of BNs and DBNs for plant biologists, such as BNArray, a tool developed in R that creates small DBNs and combines them to predict regulatory subnetworks (Chen et al., 2006). Similarly, open source Cytoscape plugins are available for network inference: (i) NetworkBMA uses Bayesian Network Averaging to infer regulatory networks (Fraley et al., 2014); (ii) Cygenexpi is based on ODEs and uses known putative regulations and time-course data to assess regulatory interactions (Modrák and Vohradskı, 2018); and (iii) ARACNE can analyze and integrate high-throughput expression steady-state data and was already successfully used in identifying previously known and new transcriptional regulations in the *Arabidopsis* root (Margolin et al., 2006; Chávez Montes et al., 2014).

BRIDGING THE GAP BETWEEN QUANTITATIVE EXPRESSION DATA AND PHENOTYPIC TRAITS WITH MACHINE LEARNING APPROACHES

Pleiotropic effects can be a major challenge in making targeted changes to biological systems. This problem can be circumvented by adjusting the specificity of the downregulation or upregulation of the gene expression. For example, the adverse effect of the constitutive overexpression of *PLASTOCHRON1* (*ZmPLA1*) in maize, such as the absence of flowering, is eliminated by targeting the ectopic expression of *PLASTOCHRON1* (*ZmPLA1*) to the transition zone of a maize leaf. This is achieved by placing *ZmPLA1* downstream of the *GA2-OXIDASE* (*ZmGA2OX*) promotor, of which the expression is limited to the transition from cell division to cell expansion and results in larger leaves (Sun et al., 2017). Predicting the need for these kinds of targeted interventions requires a detailed understanding of the complex connections between gene expression data and downstream phenotypic effects. Unraveling GRNs and understanding their dynamics provides one means to link gene expression and phenotype. However, when the link between gene expression and phenotypic output is unclear, unresolved, or highly complex machine learning (ML) approaches can provide an attractive avenue. ML approaches can yield data-driven models that offer predictions, thus providing a broadly applicable toolset to analyze biological data and predict phenotypic outputs based on gene expression data (**Figure 3**). This could help to improve the effectiveness and precision possible in modifying phenotypic traits.

Machine learning tools have been applied to biological systems at multiple scales. They have been applied to gene expression data to identify DEGs (Pirooznia et al., 2008) and transcriptional regulations between genes (**Figure 2**; Huynh-Thu et al., 2010). At the phenotypic level, ML systems have been used to analyze images for rapid phenotyping (Gonzalez-Sanchez et al., 2014; Sommer et al., 2017). Computer vision systems using ML have been used to track *Arabidopsis* growth and movement through day–night cycles, extracting patterns of movement and growth, automating extraction of phenotypic information (Bernotas et al., 2019). In another example, linear regression, support vector machines (SVMs), artificial neural networks (ANNs), random forest regression, and stochastic gradient boosting were tested for accuracy and robustness in yield prediction in almonds using orchard images, orchard-specific attributes, and weather data. After testing these ML methods, stochastic gradient boosting was found to provide the best performance in yield prediction and identifying key determinants of almond yield, such as orchard age and levels of precipitation during periods of pollinator activity (Zhang et al., 2019).

Additionally, several ML approaches such as SVMs, random forests, logistic regression, naïve Bayes classifiers, and ANNs have already been applied to genetic data for the prediction of phenotypic traits (**Figure 3**). For example, deep ANNs were used to predict yield in maize from genotype data and weather conditions. In this case, the models were able to predict yield with a root mean squared error of 12%, although this was highly sensitive to weather prediction accuracy (Khaki and Wang, 2019). ML approaches have also been used to predict genotypes. Logistic regression and naïve Bayes approaches have been used to predict the genotype of crosses between maize strains, with prediction accuracy between 82 and 85% (Seka et al., 2019). However, because of the complexity of ML approaches and lack of interpretable intermediary results, it can be difficult to understand whether the model will generalize well and operate on a wide range of input data without prohibitive amounts of testing. One approach to address this is to identify informative features that can be extracted from the data before it is used in the ML system. Extracting information about this process and using that as an input to the ML system can reduce the complexity of the relationships the ML system needs to infer.

Gene regulation is an integral mechanism for numerous biological processes. As a result, GRN topology plays a significant role in the plant response to intrinsic or environmental signals

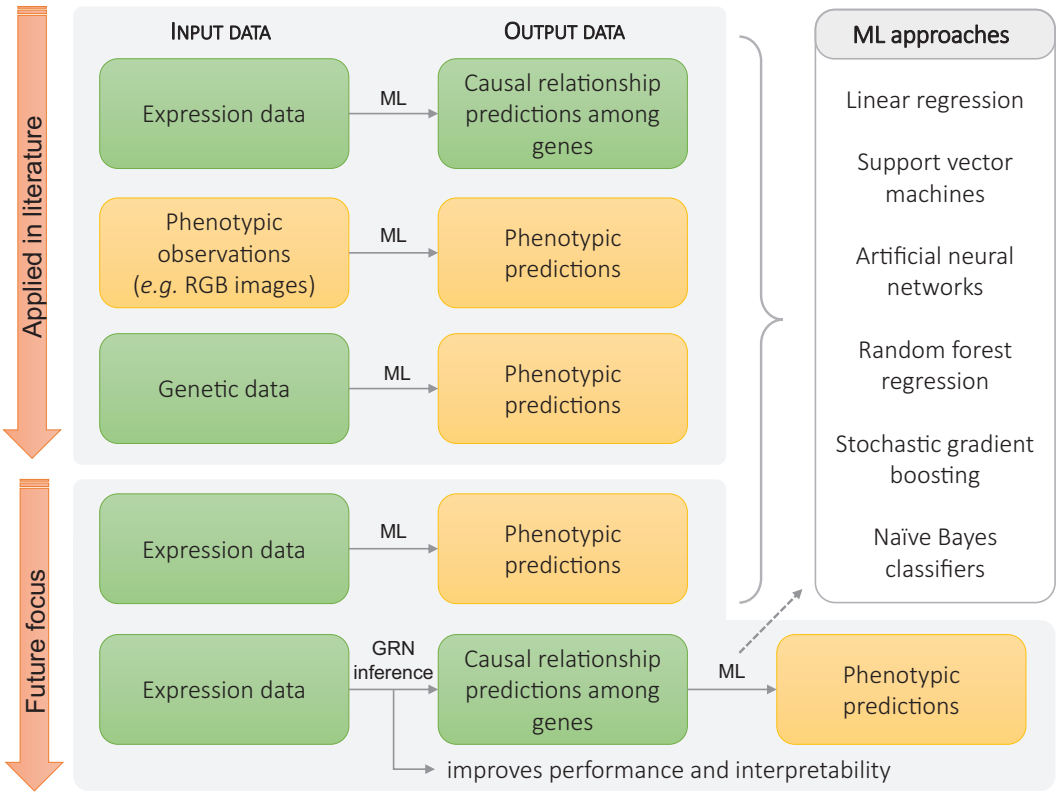

FIGURE 3 | Current and potential future applications of machine learning methods in plant biology. **Top panel**: current applications of machine learning approaches include predicting relationships from expression data, predicting phenotype from direct observational data, and predicting phenotype from genotype. **Bottom panel**: in the future, gene expression data and GRN inference methods could be used to make phenotypic predictions based on the regulatory relationships between genes.

(Stelling et al., 2002). This connection between phenotype and regulatory relationships makes constructed or inferred GRNs an attractive intermediary step between expression-level data and phenotypic predictions. Due to the key role of gene regulation in determining phenotype, features derived from the topology of GRNs, such as node connectivity, network diameter, and network density, could be used by the ML system to make predictions at a higher level of abstraction than using the raw expression data. As such, the incorporation of GRN features within the ML system can improve both phenotypic prediction performance and model interpretability (**Figure 3**). Network topological features have found use in predicting emergent behavior in systems such as protein interaction networks and metabolic networks (Hasan et al., 2006). For example, network features have been applied to identify biologically important genes in *E. coli* metabolic networks and found their predictions to agree with genome-wide knockout screens (Plaimas et al., 2008, 2010). Similarly, ML approaches that integrate network topological features have been applied to predict metabolic pathways from correlation networks in tomato plants, identifying a novel melibiose-degradation pathway (Toubiana et al., 2019).

Designing an ML system involves many tradeoffs between detail, predictive performance, availability of data, and model interpretability. While deep learning methods provide extreme detail, incorporating GRN-derived features presents an opportunity to improve predictive performance and interpretability while still making efficient use of available data.

CONCLUDING REMARKS AND FUTURE PERSPECTIVES

As shown in this review, multiple techniques, both empirical and *in silico* techniques, are available for the generation of GRNs. An environmental signal or a developmental cue can trigger transcriptional changes that are regulated by highly dynamic GRNs. Different transcriptomes are identified depending on the time upon stress or developmental signal (**Figure 1**) and as such sampling at multiple time points is crucial to fully comprehend a biological response. Moreover, as transcriptomes differ significantly between organs (root versus shoot), tissues (proliferating versus mature), and even cell types (epidermis versus stoma), the precise developmental stage at which the sampling occurs should be considered with care. Nowadays, more techniques are being developed that allow for the analysis of specific cell types using FACS, fluorescence-activated nuclei sorting (FANS), and Isolation of Nuclei TAgged in specific Cell Types (INTACT) (Bargmann and Birnbaum, 2010; Deal and Henikoff, 2011; Slane et al., 2015; Reynoso et al., 2018). Moreover,

several studies report that even within the same cell type, gene expression is heterogeneous between cells. The complexity of cellular diversity and cell-to-cell gene expression variability can be addressed with transcriptomics at scale with single-cell resolution (Denyer et al., 2019). Single-cell transcriptomics allows for the simultaneous and accurate profiling of thousands of cells, revealing detailed transcriptional pathways and developmental processes (Denyer et al., 2019). Computational techniques, such as Bayesian network inference and ML approaches, will need to be adapted to the large amounts of data generated by single-cell RNA sequencing and the cross-talk between datasets.

AUTHOR CONTRIBUTIONS

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

FUNDING

Support for this work was provided by the National Science Foundation (NSF) (CAREER MCB-1453130), the NSF/Biotechnology and Biological Sciences Research Council (BBSRC) (NSF MCB 1517058) to RS, the Ghent University "Bijzonder Onderzoeksfonds Methusalem Project" (BOF08/01M00408), the National Science Foundation Graduate Research Fellowship (Grant No. DGE-1746939), and the NIH/NCSU Molecular Biotechnology Training Program (NIH T32 GM008776).

ACKNOWLEDGMENTS

We thank Dr. Marieke Dubois for helpful suggestions to improve the manuscript.

REFERENCES

Aghamirzaie, D., Raja Velmurugan, K., Wu, S., Altarawy, D., Heath, L. S., and Grene, R. (2017). Expresso: a database and web server for exploring the interaction of transcription factors and their target genes in *Arabidopsis thaliana* using ChIP-Seq peak data. *F1000Res* 6:372. doi: 10.12688/f1000research.10041.1

Bansal, M., Belcastro, V., Ambesi-Impiombato, A., and di Bernardo, D. (2007). How to infer gene networks from expression profiles. *Mol. Syst. Biol.* 3:78. doi: 10.1038/msb4100120

Barah, P., Mahantesha, N. B. N., Jayavelu, N. D., Sowdhamini, R., Shameer, K., and Bones, A. M. (2016). Transcriptional regulatory networks in *Arabidopsis thaliana* during single and combined stresses. *Nucleic Acids Res.* 44, 3147–3164. doi: 10.1093/nar/gkv1463

Bargmann, B. O., and Birnbaum, K. D. (2010). Fluorescence activated cell sorting of plant protoplasts. *J. Vis. Exp.* 18:1673. doi: 10.3791/1673

Bargmann, B. O., Marshall-Colon, A., Efroni, I., Ruffel, S., Birnbaum, K. D., Coruzzi, G. M., et al. (2013). TARGET: a transient transformation system for genome-wide transcription factor target discovery. *Mol. Plant* 6, 978–980. doi: 10.1093/mp/sst010

Bartlett, A., O'Malley, R. C., Huang, S. C., Galli, M., Nery, J. R., Gallavotti, A., et al. (2017). Mapping genome-wide transcription-factor binding sites using DAP-seq. *Nat. Protoc.* 12, 1659–1672. doi: 10.1038/nprot.2017.055

Berger, M. F., and Bulyk, M. L. (2009). Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat. Protoc.* 4, 393–411. doi: 10.1038/nprot.2008.195

Bernotas, G., Scorza, L. C. T., Hansen, M. F., Hales, I. J., Halliday, K. J., Smith, L. N., et al. (2019). A photometric stereo-based 3D imaging system using computer vision and deep learning for tracking plant growth. *Gigascience* 8:giz056. doi: 10.1093/gigascience/giz056

Bujdoso, N., and Davis, S. (2013). Mathematical modeling of an oscillating gene circuit to unravel the circadian clock network of *Arabidopsis thaliana*. *Front. Plant Sci.* 4:3. doi: 10.3389/fpls.2013.00003

Caddick, M. X., Greenland, A. J., Jepson, I., Krause, K. P., Qu, N., Riddell, K. V., et al. (1998). An ethanol inducible gene switch for plants used to manipulate carbon metabolism. *Nat. Biotechnol.* 16, 177–180. doi: 10.1038/nbt0298-177

Chang, K. N., Zhong, S., Weirauch, M. T., Hon, G., Pelizzola, M., Li, H., et al. (2013). Temporal transcriptional response to ethylene gas drives growth hormone cross-regulation in *Arabidopsis*. *eLife* 2:e00675. doi: 10.7554/eLife.00675

Chávez Montes, R. A., Coello, G., González-Aguilera, K. L., Marsch-Martínez, N., de Folter, S., and Alvarez-Buylla, E. R. (2014). ARACNe-based inference, using curated microarray data, of *Arabidopsis thaliana* root transcriptional regulatory networks. *BMC Plant Biol.* 14:97. doi: 10.1186/1471-2229-14-97

Chen, X., Chen, M., and Ning, K. (2006). BNArray: an R package for constructing gene regulatory networks from microarray data by using Bayesian network. *Bioinformatics* 22, 2952–2954. doi: 10.1093/bioinformatics/btl491

Corrado, G., and Karali, M. (2009). Inducible gene expression systems and plant biotechnology. *Biotechnol. Adv.* 27, 733–743. doi: 10.1016/j.biotechadv.2009.05.006

Davies, M. E., and Exworth, C. P. (1973). Transient inhibition by cycloheximide of protein synthesis in cultured plant cell suspensions: a dose response paradox. *Biochem. Biophys. Res. Commun.* 50, 1075–1080. doi: 10.1016/0006-291x(73)91516-7

De Bodt, S., Carvajal, D., Hollunder, J., Van den Cruyce, J., Movahedi, S., and Inzé, D. (2010). CORNET: a user-friendly tool for data mining and integration. *Plant Physiol.* 152, 1167–1179. doi: 10.1104/pp.109.147215

de Luis Balaguer, M. A., Fisher, A. P., Clark, N. M., Fernandez-Espinosa, M. G., Möller, B. K., Weijers, D., et al. (2017). Predicting gene regulatory networks by combining spatial and temporal gene expression data in *Arabidopsis* root stem cells. *Proc. Natl. Acad. Sci. U.S.A.* 114, E7632–E7640. doi: 10.1073/pnas.1707566114

Deal, R. B., and Henikoff, S. (2011). The INTACT method for cell type-specific gene expression and chromatin profiling in *Arabidopsis thaliana*. *Nat. Protoc.* 6, 56–68. doi: 10.1038/nprot.2010.175

Denyer, T., Ma, X., Klesen, S., Scacchi, E., Nieselt, K., and Timmermans, M. C. P. (2019). Spatiotemporal developmental trajectories in the *Arabidopsis* root revealed using high-throughput single-cell RNA sequencing. *Dev. Cell* 48, 840.e5–852.e5. doi: 10.1016/j.devcel.2019.02.022

Djordjevic, M. (2007). SELEX experiments: new prospects, applications and data analysis in inferring regulatory pathways. *Biomol Eng.* 24, 179–189. doi: 10.1016/j.bioeng.2007.03.001

Doidy, J., Li, Y., Neymotin, B., Edwards, M. B., Varala, K., Gresham, D., et al. (2016). Hit-and-Run" transcription: de novo transcription initiated by a transient bZIP1 "hit" persists after the "run. *BMC Genomics* 17:92. doi: 10.1186/s12864-016-2410-2

Dondelinger, F., Husmeier, D., and Lèbre, S. (2012). Dynamic bayesian networks in molecular plant science: inferring gene regulatory networks from multiple gene expression time series. *Euphytica* 183, 361–377. doi: 10.1007/s10681-011-0538-3

Dubois, M., Skirycz, A., Claeys, H., Maleux, K., Dhondt, S., De Bodt, S., et al. (2013). ETHYLENE RESPONSE FACTOR 6 acts as a central regulator of leaf growth under water-limiting conditions in *Arabidopsis*. *Plant Physiol.* 162, 319–332. doi: 10.1104/pp.113.216341

Fraley, C., Young, W. C., Yeung, K. Y., and Raftery, A. E. (2014). { t networkBMA}: Regression-bsed network inference using Bayesian Model Averaging.

Friedman, N., Murphy, K., and Russell, S. (1998). "Learning the structure of dynamic probabilistic networks," in *Proceedings of the 1st Annual Conference on*

Uncertainty in Artificial Intelligence (UAI 1998), New York, NY: Elsevier Science Publishing Comapny, Inc, 139–147.

Gonzalez, N., Pauwels, L., Baekelandt, A., De Milde, L., Van Leene, J., Besbrugge, N., et al. (2015). A repressor protein complex regulates leaf growth in *Arabidopsis*. *Plant Cell* 27, 2273–2287. doi: 10.1105/tpc.15.00006

Gonzalez-Sanchez, A., Frausto-Solis, J., and Ojeda-Bustamante, W. (2014). Predictive ability of machine learning methods for massive crop yield prediction. *Spanish J. Agric. Res.* 12, 313–328.

Hasan, M. A., Chaoji, V., Salem, S., and Zaki, M. (2006). "Link Prediction Using Supervised Learning," in *Proceedings of the SDM 06 Workshop on Link Analysis, Counterterrorism and Security*. Available online at: http://www.cs.rpi.edu/~zaki/PaperDir/LINK06.pdf

Hernando, C. E., Romanowski, A., and Yanovsky, M. J. (2017). Transcriptional and post-transcriptional control of the plant circadian gene regulatory network. *Biochim. Biophys. Acta* 1860, 84–94. doi: 10.1016/j.bbagrm.2016.07.001

Hruz, T., Laule, O., Szabo, G., Wessendorp, F., Bleuler, S., Oertle, L., et al. (2008). Genevestigator V3: a reference expression database for the meta-analysis of transcriptomes. *Adv. Bioinformatics* 2008, 1–5. doi: 10.1155/2008/420747

Hu, Z., Mellor, J., Wu, J., Yamada, T., Holloway, D., and DeLisi, C. (2005). VisANT: data-integrating visual framework for biological networks and modules. *Nucleic Acids Res.* 33, W352–W357. doi: 10.1093/nar/gki431

Huang, X., Lei, Y., Guan, H., Hao, Y., Liu, H., Sun, G., et al. (2017). Transcriptomic analysis of the regulation of stalk development in flowering Chinese cabbage (*Brassica campestris*) by RNA sequencing. *Sci. Rep.* 7:15517. doi: 10.1038/s41598-017-15699-6

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PLoS One* 5:e12776. doi: 10.1371/journal.pone.0012776

Ikeuchi, M., Shibata, M., Rymen, B., Iwase, A., Bågman, A.-M., Watt, L., et al. (2018). A gene regulatory network for cellular reprogramming in plant regeneration. *Plant Cell Physiol.* 59, 770–782. doi: 10.1093/pcp/pcy013

Joanito, I., Chu, J. W., Wu, S. H., and Hsu, C. P. (2018). An incoherent feed-forward loop switches the *Arabidopsis* clock rapidly between two hysteretic states. *Sci. Rep.* 8:13944. doi: 10.1038/s41598-018-32030-z

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

Khaki, S., and Wang, L. (2019). Crop yield prediction using deep neural networks. *Front. Plant Sci.* 10:621. doi: 10.3389/fpls.2019.00621

Kim, J., Jun, K. M., Kim, J. S., Chae, S., Pahk, Y. M., Lee, T. H., et al. (2017). RapaNet: a web tool for the co-expression analysis of *Brassica rapa* genes. *Evol. Bioinform. Online* 13:1176934317715421. doi: 10.1177/1176934317715421

Kobayashi, Y., Sadhukhan, A., Tazib, T., Nakano, Y., Kusunoki, K., Kamara, M., et al. (2016). Joint genetic and network analyses identify loci associated with root growth under NaCl stress in *Arabidopsis thaliana*. *Plant Cell Env.* 39, 918–934. doi: 10.1111/pce.12691

Koryachko, A., Matthiadis, A., Haque, S., Muhammad, D., Ducoste, J. J., Tuck, J. M., et al. (2019). Dynamic modelling of the iron deficiency modulated transcriptome response in *Arabidopsis thaliana* roots. *Silico Plants* 1:diz005. doi: 10.1093/insilicoplants/diz005

Krouk, G., Mirowski, P., LeCun, Y., Shasha, D. E., and Coruzzi, G. M. (2010). Predictive network modeling of the high-resolution dynamic plant transcriptome in response to nitrate. *Genome Biol.* 11:R123. doi: 10.1186/gb-2010-11-12-r123

Krumsiek, J., Pölsterl, S., Wittmann, D. M., and Theis, F. J. (2010). Odefy - From discrete to continuous models. *BMC Bioinformatics* 11:233. doi: 10.1186/1471-2105-11-233

Kudo, T., Terashima, S., Takaki, Y., Tomita, K., Saito, M., Kanno, M., et al. (2017). PlantExpress: a database integrating OryzaExpress and ArthaExpress for single-species and cross-species gene expression network analyses with microarray-based transcriptome data. *Plant Cell Physiol.* 58:e1. doi: 10.1093/pcp/pcw208

Kulkarni, S. R., Vaneechoutte, D., Van de Velde, J., and Vandepoele, K. (2017). TF2Network: predicting transcription factor regulators and gene regulatory networks in *Arabidopsis* using publicly available binding site information. *Nucleic Acids Res.* 46:e31. doi: 10.1093/nar/gkx1279

Lee, T. H., Kim, Y. K., Pham, T. T., Song, S. I., Kim, J. K., Kang, K. Y., et al. (2009). RiceArrayNet: a database for correlating gene expression from transcriptome profiling, and its application to the analysis of coexpressed genes in rice. *Plant Physiol.* 151, 16–33. doi: 10.1104/pp.109.139030

Lin, H., Yu, J., Pearce, S. P., Zhang, D., and Wilson, Z. A. (2017). RiceAntherNet: a gene co-expression network for identifying anther and pollen development genes. *Plant J.* 92, 1076–1091. doi: 10.1111/tpj.13744

Locke, J. C. W., Kozma-Bognár, L., Gould, P. D., Fehér, B., Kevei, É, Nagy, F., et al. (2006). Experimental validation of a predicted feedback loop in the multi-oscillator clock of *Arabidopsis thaliana*. *Mol. Syst. Biol.* 2:59. doi: 10.1038/msb4100102

Locke, J. C. W., Millar, A. J., and Turner, M. S. (2005). Modelling genetic networks with noisy and varied experimental data: the circadian clock in *Arabidopsis thaliana*. *J. Theor. Biol.* 234, 383–393. doi: 10.1016/j.jtbi.2004.11.038

Lorenz, W. W., Alba, R., Yu, Y. S., Bordeaux, J. M., Simoes, M., and Dean, J. F. (2011). Microarray analysis and scale-free gene networks identify candidate regulators in drought-stressed roots of loblolly pine (*P. taeda* L.). *BMC Genomics* 12:264. doi: 10.1186/1471-2164-12-264

Lv, Q., Cheng, R., and Shi, T. (2014). Regulatory network rewiring for secondary metabolism in *Arabidopsis thaliana* under various conditions. *BMC Plant Biol.* 14:180. doi: 10.1186/1471-2229-14-180

Marbach, D., Costello, J. C., Kuffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., et al. (2012). Wisdom of crowds for robust gene network inference. *Nat. Methods* 9, 796–804. doi: 10.1038/nmeth.2016

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D., et al. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 1(Suppl. 1):S7. doi: 10.1186/1471-2105-7-S1-S7

Martin, O. C., Krzywicki, A., and Zagorski, M. (2016). Drivers of structural features in gene regulatory networks: from biophysical constraints to biological function. *Phys. Life Rev.* 17, 124–158. doi: 10.1016/j.plrev.2016.06.002

Modrák, M., and Vohradskı, J. (2018). Genexpi: a toolset for identifying regulons and validating gene regulatory networks using time-course expression data. *BMC Bioinformatics* 19:137. doi: 10.1186/s12859-018-2138-x

Mutwil, M., Klie, S., Tohge, T., Giorgi, F. M., Wilkins, O., Campbell, M. M., et al. (2011). PlaNet: combined sequence and expression comparisons across plant networks derived from seven species. *Plant Cell* 23, 895–910. doi: 10.1105/tpc.111.083667

Nagel, D. H., and Kay, S. A. (2012). Complexity in the wiring and regulation of plant circadian networks. *Curr. Biol.* 22, R648–R657. doi: 10.1016/j.cub.2012.07.025

Needham, C. J., Manfield, I. W., Bulpitt, A. J., Gilmartin, P. M., and Westhead, D. R. (2009). From gene expression to gene regulatory networks in *Arabidopsis thaliana*. *BMC Syst. Biol.* 3:85. doi: 10.1186/1752-0509-3-85

Nohales, M. A., and Kay, S. A. (2016). Molecular mechanisms at the core of the plant circadian oscillator. *Nat. Struct. Mol. Biol.* 23, 1061–1069. doi: 10.1038/nsmb.3327

Obayashi, T., Aoki, Y., Tadaka, S., Kagaya, Y., and Kinoshita, K. (2018). ATTED-II in 2018: a plant coexpression database based on investigation of the statistical property of the mutual rank index. *Plant Cell Physiol.* 59, 440. doi: 10.1093/pcp/pcx209

Obayashi, T., Okamura, Y., Ito, S., Tadaka, S., Aoki, Y., Shirota, M., et al. (2014). ATTED-II in 2014: evaluation of gene coexpression in agriculturally important plants. *Plant Cell Physiol.* 55:e6. doi: 10.1093/pcp/pct178

O'Malley, R. C., Huang, S. C., Song, L., Lewsey, M. G., Bartlett, A., Nery, J. R., et al. (2016). Cistrome and epicistrome features shape the regulatory DNA landscape. *Cell* 166:1598. doi: 10.1016/j.cell.2016.08.063

O'Maoileidigh, D. S., Thomson, B., Raganelli, A., Wuest, S. E., Ryan, P. T., Kwasniewska, K., et al. (2015). Gene network analysis of *Arabidopsis thaliana* flower development through dynamic gene perturbations. *Plant J.* 83, 344–358. doi: 10.1111/tpj.12878

Palaniswamy, S. K., James, S., Sun, H., Lamb, R. S., Davuluri, R. V., and Grotewold, E. (2006). AGRIS and AtRegNet: a platform to link cis-regulatory elements and transcription factors into regulatory networks. *Plant Physiol.* 140, 818–829. doi: 10.1104/pp.105.072280

Para, A., Li, Y., Marshall-Colon, A., Varala, K., Francoeur, N. J., Moran, T. M., et al. (2014). Hit-and-run transcriptional control by bZIP1 mediates rapid nutrient signaling in *Arabidopsis*. *Proc. Natl. Acad. Sci. U.S.A.* 111, 10371–10376. doi: 10.1073/pnas.1404657111

Pearce, S., Ferguson, A., King, J., and Wilson, Z. A. (2015). FlowerNet: a gene expression correlation network for anther and pollen development. *Plant Physiol.* 167, 1717–1730. doi: 10.1104/pp.114.253807

Pearl, J. (2008). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA: Morgan Kaufmann Publishers Inc.

Penfold, C. A., Buchanan-Wollaston, V., Denby, K. J., and Wild, D. L. (2012). Nonparametric Bayesian inference for perturbed and orthologous gene regulatory networks. *Bioinformatics* 28, i233–i241. doi: 10.1093/bioinformatics/bts222

Phuong, T. M., Lee, D., and Lee, K. H. (2004). Regression trees for regulatory element identification. *Bioinformatics* 20, 750–757. doi: 10.1093/bioinformatics/btg480

Pirooznia, M., Yang, J. Y., Yang, M. Q., and Deng, Y. (2008). A comparative study of different machine learning methods on microarray gene expression data. *BMC Genomics* 9:S13. doi: 10.1186/1471-2164-9-S1-S13

Plaimas, K., Eils, R., and König, R. (2010). Identifying essential genes in bacterial metabolic networks with machine learning methods. *BMC Syst. Biol.* 16:56. doi: 10.1186/1752-0509-4-56

Plaimas, K., Mallm, J.-P., Oswald, M., Svara, F., Sourjik, V., Eils, R., et al. (2008). Machine learning based analyses on metabolic networks supports high-throughput knockout screens. *BMC Syst. Biol.* 2:67. doi: 10.1186/1752-0509-2-67

Redekar, N., Pilot, G., Raboy, V., Li, S., and Saghai Maroof, M. A. (2017). Inference of transcription regulatory network in low phytic acid soybean seeds. *Front. Plant Sci.* 8:2029. doi: 10.3389/fpls.2017.02029

Reynoso, M. A., Pauluzzi, G. C., Kajala, K., Cabanlit, S., Velasco, J., Bazin, J., et al. (2018). Nuclear transcriptomes at high resolution using retooled INTACT. *Plant Physiol.* 176, 270–281. doi: 10.1104/pp.17.00688

Schaefer, R. J., Michno, J.-M., Jeffers, J., Hoekenga, O., Dilkes, B., Baxter, I., et al. (2018). Integrating coexpression networks with GWAS to Prioritize Causal Genes in Maize. *Plant Cell* 30, 2922–2942. doi: 10.1105/tpc.18.00299

Scofield, S., Murison, A., Jones, A., Fozard, J., Aida, M., Band, L. R., et al. (2018). Coordination of meristem and boundary functions by transcription factors in the SHOOT MERISTEMLESS regulatory network. *Development* 145:dev157081. doi: 10.1242/dev.157081

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

Seka, D., Bonny, B. S., Yoboué, A. N., Sié, S. R., and Adopo-Gourène, B. A. (2019). Identification of maize *(Zea mays* L.) progeny genotypes based on two probabilistic approaches: logistic regression and naïve Bayes. *Artif. Intell. Agric.* 1, 9–13. doi: 10.1016/j.aiia.2019.03.001

Slane, D., Kong, J., Schmid, M., Jurgens, G., and Bayer, M. (2015). Profiling of embryonic nuclear vs. cellular RNA in *Arabidopsis thaliana*. *Genom. Data* 4, 96–98. doi: 10.1016/j.gdata.2015.03.015

Sommer, C., Hoefler, R., Samwer, M., and Gerlich, D. W. (2017). A deep learning and novelty detection framework for rapid phenotyping in high-content screening. *Mol. Biol. Cell* 28, 3428–3436. doi: 10.1091/mbc.e17-05-0333

Song, L., Huang, S. C., Wise, A., Castanon, R., Nery, J. R., Chen, H., et al. (2016). A transcription factor hierarchy defines an environmental stress response network. *Science* 354:aag1550. doi: 10.1126/science.aag1550

Spurney, R. J., Van den Broeck, L., Clark, N. M., Fisher, A. P., de Luis Balaguer, M. A., and Sozzani, R. (2019). TuxNet: a simple interface to process RNA sequencing data and infer gene regulatory networks. *Plant J.* 101, 716–730. doi: 10.1111/tpj.14558

Stelling, J., Klamt, S., Bettenbrock, K., Schuster, S., and Gilles, E. D. (2002). Metabolic network structure determines key aspects of functionality and regulation. *Nature* 420, 190–193. doi: 10.1038/nature01166

Su, G., Morris, J. H., Demchak, B., and Bader, G. D. (2014). Biological network exploration with Cytoscape 3. *Curr. Protoc. Bioinforma.* 47 8, 1–24. doi: 10.1002/0471250953.bi0813s47

Sun, X., Cahill, J., Van Hautegem, T., Feys, K., Whipple, C., Novak, O., et al. (2017). Altered expression of maize PLASTOCHRON1 enhances biomass and seed yield by extending cell division duration. *Nat. Commun.* 8:14752. doi: 10.1038/ncomms14752

Swift, J., and Coruzzi, G. M. (2017). A matter of time - how transient transcription factor interactions create dynamic gene regulatory networks. *Biochim. Biophys. Acta* 1860, 75–83. doi: 10.1016/j.bbagrm.2016.08.007

Taylor-Teeples, M., Lin, L., De Lucas, M., Turco, G., Toal, T. W., Gaudinier, A., et al. (2015). An *Arabidopsis* gene regulatory network for secondary cell wall synthesis. *Nature* 517, 571–575. doi: 10.1038/nature14099

Toubiana, D., Puzis, R., Wen, L., Sikron, N., Kurmanbayeva, A., Soltabayeva, A., et al. (2019). Combined network analysis and machine learning allows the prediction of metabolic pathways from tomato metabolomics data. *Commun. Biol.* 2:214. doi: 10.1038/s42003-019-0440-4

Van den Broeck, L., Dubois, M., Vermeersch, M., Storme, V., Matsui, M., and Inzé, D. (2017). From network to phenotype: the dynamic wiring of an *Arabidopsis* transcriptional network induced by osmotic stress. *Mol. Syst. Biol.* 13:961. doi: 10.15252/msb.20177840

Vanden Bossche, R., Demedts, B., Vanderhaeghen, R., and Goossens, A. (2013). Transient expression assays in tobacco protoplasts. *Methods Mol Biol.* 1011, 227–239. doi: 10.1007/978-1-62703-414-2_18

Vermeirssen, V., De Clercq, I., Van Parys, T., Van Breusegem, F., and Van de Peer, Y. (2014). *Arabidopsis* ensemble reverse-engineered gene regulatory network discloses interconnected transcription factors in oxidative stress. *Plant Cell* 26, 4656–4679. doi: 10.1105/tpc.114.131417

Vignes, M., Vandel, J., Allouche, D., Ramadan-Alban, N., Cierco-Ayrolles, C., Schiex, T., et al. (2011). Gene regulatory network reconstruction using bayesian networks, the dantzig selector, the lasso and their meta-analysis. *PLoS One* 6:e29165. doi: 10.1371/journal.pone.0029165

Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., et al. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* 158, 1431–1443. doi: 10.1016/j.cell.2014.08.009

Wellmer, F., Alves-Ferreira, M., Dubois, A., Riechmann, J. L., and Meyerowitz, E. M. (2006). Genome-wide analysis of gene expression during early *Arabidopsis* flower development. *PLoS Genet.* 2:e117. doi: 10.1371/journal.pgen.0020117.eor

Wu, H., Lu, T., Xue, H., and Liang, H. (2014). Sparse additive ordinary differential equations for dynamic gene regulatory network modeling. *J. Am. Stat. Assoc.* 109, 700–716. doi: 10.1080/01621459.2013.859617

Wu, S., Alseekh, S., Cuadros-Inostroza, A., Fusari, C. M., Mutwil, M., Kooke, R., et al. (2016). Combined use of genome-wide association data and correlation networks unravels key regulators of primary metabolism in *Arabidopsis thaliana*. *PLoS Genet.* 12:e1006363. doi: 10.1371/journal.pgen.1006363

Yao, C.-W., Hsu, B.-D., and Chen, B.-S. (2011). Constructing gene regulatory networks for long term photosynthetic light acclimation in *Arabidopsis thaliana*. *BMC Bioinformatics* 12:335. doi: 10.1186/1471-2105-12-335

Yim, W. C., Yu, Y., Song, K., Jang, C. S., and Lee, B. M. (2013). PLANEX: the plant co-expression database. *BMC Plant Biol.* 13:83. doi: 10.1186/1471-2229-13-83

Yu, J., Smith, V. A., Wang, P. P., Hartemink, A. J., and Jarvis, E. D. (2004). Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics* 20, 3594–3603. doi: 10.1093/bioinformatics/bth448

Zhang, Z., Jin, Y., Chen, B., and Brown, P. (2019). California almond yield prediction at the orchard level with a machine learning approach. *Front. Plant Sci.* 10:809. doi: 10.3389/fpls.2019.00809

Zuo, J., Niu, Q. W., and Chua, N. H. (2000). Technical advance: an estrogen receptor-based transactivator XVE mediates highly inducible gene expression in transgenic plants. *Plant J.* 24, 265–273. doi: 10.1046/j.1365-313x.2000.00868.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.

frontiers in Genetics

METHODS
published: 03 June 2020
doi: 10.3389/fgene.2020.00445

HARMONIES: A Hybrid Approach for Microbiome Networks Inference via Exploiting Sparsity

Shuang Jiang[1,2], Guanghua Xiao[2], Andrew Y. Koh[3], Yingfei Chen[4], Bo Yao[2], Qiwei Li[5] and Xiaowei Zhan[2]**

[1] Department of Statistical Science, Southern Methodist University, Dallas, TX, United States, [2] Quantitative Biomedical Research Center, Department of Population and Data Sciences, University of Texas Southwestern Medical Center, Dallas, TX, United States, [3] Departments of Pediatrics, Departments of Microbiology, University of Texas Southwestern Medical Center, Dallas, TX, United States, [4] Lyda Hill Department of Bioinformatics, Bioinformatics High Performance Computing, University of Texas Southwestern Medical Center, Dallas, TX, United States, [5] Department of Mathematical Sciences, The University of Texas at Dallas, Richardson, TX, United States

The human microbiome is a collection of microorganisms. They form complex communities and collectively affect host health. Recently, the advances in next-generation sequencing technology enable the high-throughput profiling of the human microbiome. This calls for a statistical model to construct microbial networks from the microbiome sequencing count data. As microbiome count data are high-dimensional and suffer from uneven sampling depth, over-dispersion, and zero-inflation, these characteristics can bias the network estimation and require specialized analytical tools. Here we propose a general framework, HARMONIES, Hybrid Approach foR MicrobiOme Network Inferences via Exploiting Sparsity, to infer a sparse microbiome network. HARMONIES first utilizes a zero-inflated negative binomial (ZINB) distribution to model the skewness and excess zeros in the microbiome data, as well as incorporates a stochastic process prior for sample-wise normalization. This approach infers a sparse and stable network by imposing non-trivial regularizations based on the Gaussian graphical model. In comprehensive simulation studies, HARMONIES outperformed four other commonly used methods. When using published microbiome data from a colorectal cancer study, it discovered a novel community with disease-enriched bacteria. In summary, HARMONIES is a novel and useful statistical framework for microbiome network inference, and it is available at https://github.com/shuangj00/HARMONIES.

Keywords: Bayesian statistics, microbiome network, Gaussian graphical model, Dirichlet process prior, hierarchical model

OPEN ACCESS

Edited by:
Shizhong Han,
Johns Hopkins Medicine,
United States

Reviewed by:
Cuncong Zhong,
University of Kansas, United States
Guilherme Corrêa De Oliveira,
Vale Technological Institute (ITV), Brazil

****Correspondence:***
Qiwei Li
qiwei.li@utdallas.edu
Xiaowei Zhan
xiaowei.zhan@utsouthwestern.edu

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

Received: *16 December 2019*
Accepted: *14 April 2020*
Published: *03 June 2020*

Citation:
Jiang S, Xiao G, Koh AY, Chen Y, Yao B, Li Q and Zhan X (2020) HARMONIES: A Hybrid Approach for Microbiome Networks Inference via Exploiting Sparsity. Front. Genet. 11:445. doi: 10.3389/fgene.2020.00445

1. INTRODUCTION

Microbiota form complex community structures and collectively affect human health. Studying their relationship as a network can provide key insights into their biological mechanisms. The exponentially growing large datasets made available by next-generation sequencing (NGS) technology (Metzker, 2010), such as 16S rRNA gene and metagenomic profiling, motivate the development of statistic tools to quantitatively study the microbial organisms. While the number of discovered microbial taxa continues to increase, our knowledge of their interactive relationships is severely lacking. Understanding the structural organization of the human microbiome plays a vital role in revealing how the microbial taxa are collaborating or competing with each other under different physiologic conditions.

In sequencing-based microbial association studies, the enormous amount of NGS data can be summarized in a sample-by-taxon count table where each entry is a proxy to the underlying true abundance. However, there is no simple relationship between the true abundances and the observed counts. Additionally, microbiome sequencing data usually have an inflated amount of zeros, uneven sequencing depths across samples, and over-dispersion. Initial attempts at constructing microbial association networks with this type of data (Ban et al., 2015; Lo and Marculescu, 2017), first transformed the microbiome sequencing counts into their compositional formula. Specifically, a count was normalized to its proportion in the respective sample. Then, each sample was transformed by a choice of log-ratio transformations to remove the unit-sum constraint of the compositional data. While this type of normalization is simple to implement and preserves the original ordering of the counts in a sample, it fails to capture the sample to sample variation and it overlooks the excess zeros in the microbiome data. Note that these zeros can be attributed to biological or technical reasons: either certain taxa are not present among samples, or they are not sequenced due to insufficient sequencing depths. As the existing logarithmic transformation neglects the difference between these two types of zeros, it can lead to a biased estimation of the network structure. Thus, we propose a model-based normalization strategy for microbiome count data. Our normalization method simultaneously accounts for uneven sequencing depth, zero-inflation, over-dispersion, as well as the two types of zeros. Then we use the normalized abundances to estimate microbial abundance networks.

There are two major categories of statistical methods that are often used to infer microbial abundance networks. The first type is based on a taxa abundance covariance structure. For example, Faust and Raes (2016) and Weiss et al. (2016) used pairwise Pearson correlations to represent edge weights. This simple inference could be problematic since two variables (i.e., taxa) may be connected in the network due to their confounding variables (Gevers et al., 2014). The other type aims to estimate taxa abundance partial correlations, removing confounding effects. Kurtz et al. (2015) proposed a statistical model for inferring microbial ecological network, which is based on estimating the precision matrix (via exploiting sparsity) of a Gaussian multivariate model and relies on graphical lasso (Glasso) (Friedman et al., 2008). However, their data normalization step needs to be improved to account for unique characteristics observed in microbiome count data.

In this paper, we propose a general framework, HARMONIES (Hybrid Approach foR MicrobiOme Network Inferences via Exploiting Sparsity), to infer the microbiome networks. It consists of two major steps: (1) normalization of the microbiome count data by fitting a zero-inflated negative binomial (ZINB) model with the Dirichlet process prior (DPP), (2) application of Glasso to ensure sparsity and using a stability-based approach to select the tuning parameter in Glasso. The estimated network contains the information of both the degree and the direction of associations between taxa, which facilitates the biological interpretation. We demonstrated that HARMONIES could outperform other state-of-the-art tools on extensive simulated and synthetic data. Further, we used HARMONIES to uncover unique associations between disease-specific genera from microbiome profiling data generated from a colorectal cancer study. Based on these results, HARMONIES will be a valuable statistical model to understand the complex microbial associations in microbiome studies. The `R` package `HARMONIES` is freely available at https://github.com/shuangj00/HARMONIES.

2. METHODS

2.1. Microbiome Count Data Normalization

Let $\boldsymbol{Y}$ denote the n-by-p taxonomic count matrix obtained from either the 16S rRNA or the metagenomic shotgun sequencing (MSS) technology. Each entry $y_{ij}, i = 1,\ldots,n, j = 1,\ldots,p$ is a non-negative integer, indicating the total reads related to taxon j observed in sample i. It is recommended that all chosen taxa should be at the same taxonomic level (e.g., OTU for 16S rRNA or species for MSS) since that mixing different taxonomic levels in the proposed model could lead to improper biological interpretation. As the real microbiome data are characterized by zero-inflation and over-dispersion, we model y_{ij} through a zero-inflated negative binomial (ZINB) model as

$$y_{ij} \sim \pi_i \mathrm{I}(y_{ij} = 0) + (1 - \pi_i)\mathrm{NB}(\lambda_{ij}, \phi_j). \tag{1}$$

The first component in the Equation (1) models whether zeros come from a degenerate distribution with a point mass at zero. It can be interpreted as the "extra" zeros due to insufficient sequencing effort. We can assume there exists a true underlying abundance for the taxon in its sample, but we fail to observe it with the mixture probability π_i representing the proportion of "extra" zeros in sample i. The second component, $\mathrm{NB}(\lambda_{ij}, \phi_j)$, models the "true" zeros and all the nonzero observed counts. i.e., counts generated from a negative binomial (NB) distribution with the expectation of λ_{ij} and dispersion $1/\phi_j$. Here, "true" zero refers to a taxon that is truly absent in the corresponding sample. The variance of the random variable from NB distribution, under the current parameterization, equals to $\lambda_{ij}+\lambda_{ij}^2/\phi_j$. Smaller values of ϕ_j can lead to over-dispersion.

To avoid explicitly fixing the value of π_i's and ϕ_j's, we use a Bayesian hierarchical model for parameter inference. First, we rewrite the model (1) by introducing a binary indicator variable $\eta_{ij} \sim \mathrm{Bernoulli}(\pi_i)$, such that $y_{ij} = 0$ if $\eta_{ij} = 1$, and $y_{ij} \sim \mathrm{NB}(\lambda_{ij}, \phi_j)$ if $\eta_{ij} = 0$. Then, we formulate a beta-Bernoulli prior of η_{ij} by assuming $\pi_i \sim \mathrm{Beta}(a_\pi, b_\pi)$, and we let $a_\pi = b_\pi = 1$ to obtain a non-informative prior on η_{ij}. We specify independent Gamma prior $\mathrm{Ga}(a_\phi, b_\phi)$ for each dispersion parameter ϕ_j. Letting $a_\phi = b_\phi = 0.001$ results in a weakly informative gamma prior.

The mean parameter of the NB distribution, λ_{ij}, contains the key information of the true underlying abundance of the corresponding count. As λ_{ij} is affected by the varying sequencing effort across samples, we use a multiplicative characterization of the NB mean to justify the latent heterogeneity in microbiome sequencing data. Specifically, we assume $\lambda_{ij} = s_i\alpha_{ij}$. Here, s_i is the sample-specific size factor that captures the variation

in sequencing depth across samples, and α_{ij} is the normalized abundance of taxon j in sample i.

In parameter estimation, one needs to ensure identifiability between s_i and α_{ij}. For example, s_i can be the reciprocal of the total number of reads in sample i. The resulted α_{ij} is often called relative abundance, which represents the proportion of taxon j in sample i. In this setting, the relative abundances of all the taxa in one sample always sum up to 1. Similarly, other methods have been proposed with different constraints for normalizing the sequencing data (Anders and Huber, 2010; Bullard et al., 2010; Robinson and Oshlack, 2010; Paulson et al., 2013). Some normalization methods can perform better than the others in the downstream analysis (e.g., the differential abundance analysis) in certain settings. From a Bayesian perspective, fixing the values of s_i's imposes a strongly informative prior in model inference. Hence, all these methods could bias the estimations of other model parameters and degrade the performance of downstream analyses. We thus propose a regularizing prior with a stochastic constraint for estimating s_i's. Our method can simultaneously infer the size factor and other model parameters. In particular, we adopt the following mixture model for s_i,

$$\log s_i \sim \sum_{m=1}^{M} \psi_m \left[t_m \mathrm{N}(\nu_m, \sigma_s^2) + (1 - t_m)\mathrm{N}\left(-\frac{t_m \nu_m}{1 - t_m}, \sigma_s^2 \right) \right], \tag{2}$$

where ψ_m is the weight for outer mixtures of the mth component. The inner mixture of the mth component consists of two Gaussian distributions with t_m and $1-t_m$ as weights, respectively. It is straightforward to see that the inner mixture has a mean of zero and thus ensuring the stochastic constraint of $\mathrm{E}(\log s_i) = 0$. For the outer mixtures, M is an arbitrarily large positive integer. Letting $M \to \infty$ and defining the weight ψ_m by the stick-breaking procedure (i.e., $\psi_1 = V_1, \psi_m = V_m \prod_{u=1}^{m-1}(1-V_u), m = 1, 2, \ldots$) makes model (2) a special case of Dirichlet process mixture models. This class of Bayesian nonparametric infinite mixtures is widely used in quantifying the model uncertainty and allowing for flexibility in parameter estimation (Kyung et al., 2011; Taddy and Kottas, 2012). In particular, this Dirichlet process prior (DPP) has been used to account for sample heterogeneity since it is able to capture multi-modality and skewness in a distribution (Li et al., 2017; Lee and Sison-Mangus, 2018). In practice, we set M to be a large positive integer, and adopt the following hyper-prior distributions for the parameters in (2) such that $\nu_m \sim \mathrm{N}(0, \tau_\nu), t_m \sim \mathrm{Beta}(a_t, b_t)$, and $V_m \sim \mathrm{Beta}(a_m, b_m)$. We further set $\sigma_s^2 = 1$ to complete the parameter specification in the DPP prior.

In our model, the normalized abundance matrix $\boldsymbol{A} = \{\alpha_{ij}\}$ represents the true underlying abundance of the original count matrix. We further assume $\log \alpha_{ij} \sim \mathrm{N}(\mu_j, \sigma_j^2)$. This variance-stabilizing transformation on each α_{ij} not only reduces the skewness of the normalized abundance, but converts the non-negative α_{ij} to a real number. We apply the following conjugate setting to specify the priors for μ_j and $\sigma_j^2, j = 1, \ldots, p$. We let $\mu_j \sim \mathrm{N}(0, h_0\sigma_0^2)$ and $\sigma_j^2 \sim$ inverse-gamma(a_0, b_0). After integrating out μ_j and σ_j^2, the prior of the normalized abundances of taxon j follows a non-standardized Student's t-distribution, i.e.,

$$p(\boldsymbol{\alpha}_{\cdot j}) = (nh_0 + 1)^{-\frac{1}{2}} \frac{\Gamma\left(a_0 + \frac{n}{2}\right)}{\Gamma(a_0)} \frac{b_0^{a_0}}{\left\{ b_0 + \frac{1}{2}\left[\sum_{i=1}^{n} \log \alpha_{ij}^2 - \frac{\left(\sum_{i=1}^{n} \log \alpha_{ij}\right)^2}{n + \frac{1}{h_0}} \right] \right\}^{a_0 + \frac{n}{2}}}. \tag{3}$$

As for the fixed parameters a_0, b_0, h_0, and σ_0^2, we follow Li et al. (2019) and set $a_0 = 2, b_0 = 1$ to obtain a weakly informative prior for σ_j^2. We fix $\sigma_0^2 = 1$ and let $h_0 = 10$ such that the normal prior on μ_j is fairly flat. We adopt the following prior specification for the rest model parameters. First, we assume an noninformative prior for each π_i by letting $a_\pi = b_\pi = 1$. Next, we specify $a_\phi = b_\phi = 0.001$ in the Gamma prior distribution for all ϕ_j's. Then, we apply the following prior setting for the DPP: $M = n/2$, $\sigma_s = 1$, $\tau_\nu = 1$, $a_t = b_t = 1$, and $a_m = b_m = 1$.

The logarithmic scale of $\boldsymbol{A}$, denoted as $\boldsymbol{Z} = \log \boldsymbol{A}$, represents the normalized microbiome abundances on the log scale. We use Markov chain Monte Carlo (MCMC) algorithm for model parameter estimation (see details in the **Supplementary Material**), and calculate the posterior mean of $\boldsymbol{Z}$ to fit the Gaussian graphical model in the next step. Since the observed zero counts may not always represent the absence of taxa in the samples, we treat these zeros differently in the matrix $\boldsymbol{Z}$. We categorize the two types of zeros ("extra" and "true" zeros) based on the estimated η_{ij} for each observed $y_{ij} = 0$ in the data. In particular, suppose that we observe L zeros in total. We calculate the marginal posterior probability of being 1 for each $\eta_l, l = 1, \ldots, L$ as $p_l = \sum_{b=1}^{B} \mathrm{I}(\eta_l = 1)/B$, where $\mathrm{I}(\cdot)$ is the indicator function, and B is the number of MCMC iteration after burn-in. This marginal posterior probability p_l represents the proportion of MCMC iterations in which the lth 0 is essentially a missing value rather than the lowest count in the corresponding sample. Then, the observed zeros can be dichotomized by thresholding the L probabilities. The zeros with p_l greater than the threshold are considered as "true" zeros in the data, whereas the rest are imputed by the corresponding posterior mean of $\log \boldsymbol{\alpha}_{\cdot j}$. We used the method proposed by Newton et al. (2004) to determine the threshold that controls the Bayesian false discovery rate (FDR) to be smaller than c_η. Specifically, we first specify a small number c_η, which is analog to the significance level in the frequentist setting. Then we compute the threshold following Equation (4), which guarantees the imputed zeros have a Bayesian FDR to be smaller than c_η,

$$\text{Bayesian FDR} = \frac{\sum_{l=1}^{L} (1 - p_l)\, \mathrm{I}(1 - p_l < c_\eta)}{\sum_{l=1}^{L} \mathrm{I}(1 - p_l < c_\eta)}. \tag{4}$$

In practice, a choice of $c_\eta = 0.01$ guarantees that the Bayesian FDR to be at most 0.01. We set $c_\eta = 0.05$ for the simulation study and $c_\eta = 0.01$ for the real data analysis.

2.2. Graphical Model for Inferring Taxa-Taxa Association

Based on the normalized microbial abundances, we estimate their partial correlation matrix in order to construct the microbiome network under the Gaussian graphical model (GGM) framework. An undirected graph $G = (V, E)$ is used to illustrate the associations among vertices $V = \{1, \ldots, p\}$, representing the p microbial taxa. $E = \{e_{mk}\}$ is the collection of (undirected) edges, which is equivalently represented via a p-by-p adjacency matrix with $e_{mk} = 1$ or 0 according to whether vertices m and k are directly connected in G or not. GGM assumes that the joint distribution of p vertices is multivariate Gaussian $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, yielding the following relationship between the dependency structure and the network: a zero entry in the precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ indicates the corresponding vertices are conditional independent, and there is no edge between them in the graph G. Hence, a GGM can be defined in terms of the pairwise conditional independence. If $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Omega})$, then

$$\omega_{mk} = 0 \Leftrightarrow X_m \perp X_k | X_{V\setminus\{m,k\}} \Leftrightarrow \rho_{mk} = 0,$$

where $\rho_{mk} = -\omega_{mk}/\sqrt{\omega_{mm}\omega_{kk}}$ is the partial correlation between vertices m and k, representing the degree and direction of association between two vertices, conditional on the rest variables. Consequently, learning the network is equivalent to estimating the precision matrix $\boldsymbol{\Omega}$. For real microbiome data, we set the taxa (on the same taxonomic level) as vertices. Hence, a zero partial correlation in the precision matrix can be interpreted as no association between the corresponding pair of taxa, while a nonzero partial correlation can be interpreted as cooperative or competing associations between that taxa pair.

In biological applications, we often require a sparse and stable estimation of the precision matrix $\boldsymbol{\Omega}$. The sparsity can be achieved by imposing l_1-penalized log-likelihood,

$$\hat{\boldsymbol{\Omega}} = \underset{\boldsymbol{\Omega} \succ \mathbf{0}}{\operatorname{argmin}} \log\det\boldsymbol{\Omega} - \operatorname{trace}(\boldsymbol{S}\boldsymbol{\Omega}) - \lambda \|\boldsymbol{\Omega}\|_1, \qquad (5)$$

where $\boldsymbol{S}$ is the sample covariance matrix. The coordinate descent algorithm can iteratively solve p. The estimated precision matrix is sparsistent (i.e., all the parameters that are zeros would be estimated as zero with probability one) (Lam and Fan, 2009), as Glasso theoretically guarantees a consistent recovery of the sparse graph for the p vertices. When $p >> n$, the computational efficiency is often satisfactory, and thus Glasso is widely used in studying large-scale biological networks (Menéndez et al., 2010; Oh and Deasy, 2014; Zhao and Duan, 2019). We employ a stability-based approach to select the tuning parameter in the Glasso, which is named Stability Approach to Regularization Selection (StARS) (Liu et al., 2010). This method is an improved algorithm for estimating the tuning parameter λ in (5). The StARS selects the optimal sparsity parameter according to the graph reproducibility under the subsampling of the original data. In general, for each λ along the sparsity parameter path, we first obtain random subsamples from the original data. Then we estimate the graph for each subsample using the Glasso. Next, for each sparsity parameter, we calculate the overall edge selection instability from all the graphs constructed by the subsamples. Finally, the optimal sparsity parameter λ^* is chosen such that it corresponds to the smallest amount of regularization and still results in a graph instability to be lower than the pre-specified tolerance level. Liu et al. (2010) showed that StARS could provide the "sparsistent" network estimation that includes all the true associations with probability one. Further, the StARS has been widely used in biological network studies (Kurtz et al., 2015; Tipton et al., 2018; Zhao and Duan, 2019). Due to its excellent performance, here we adopt the StARS to select the tuning parameter for Glasso. In summary, we use the normalized abundances (on the log scale) as inputs, calculate the sparse estimation of the precision matrix using the Glasso, and use the StARS method to select λ in problem (5) to obtain the estimated graph that represents the microbiome network.

2.3. Simulation Scenarios

We compare the performance of the HARMONIES and several widely used methods for inferring microbiome networks. These methods include SPIEC-EASI (Kurtz et al., 2015), CClasso (Fang et al., 2015), and correlation-based network estimation used in Faust and Raes (2016) and Weiss et al. (2016). While the proposed model and SPIEC-EASI infer the network structure from sparse precision matrices, CClasso, and the correlation-based method utilize sparse correlation matrices to represent the network. We generated both simulated and synthetic datasets that mimic the real microbiome sequencing count data. We use $\boldsymbol{Y}_{n\times p}$ to denote the generated count matrix. For a comprehensive comparison, we varied the sample size and the number of taxa as $n \in \{60, 100, 200, 500\}$, and the number of taxa $p \in \{40, 60\}$.

2.3.1. Generating Simulated Data

We generated the simulated datasets from a Dirichlet-multinomial (DM) model using the following steps: (1) to generate the binary adjacency matrix; (2) to simulate the precision matrix and the corresponding covariance matrix; (3) to generate n multivariate Gaussian variables based on the covariance matrix to represent the true $n \times p$ underlying taxonomic abundances, denoted as $\boldsymbol{D}$; (4) to simulate the count table $\boldsymbol{Y}_{n\times p}$ from a DM model, with its parameters being $\exp(\boldsymbol{D})$; (5) to mimic the zero-inflation in real microbiome data by randomly setting part of entries in the count table to zeros. Note that the data generative scheme is different from the model assumption, which is given in Equation (1). The detailed generative models are described below.

We began with simulating a p-by-p adjacency matrix for the p taxa in the network. Here, the adjacency matrix was generated according to an Erdős–Rényi (ER) model. An ER model $\mathrm{ER}(p, \rho)$ generates each edge in a graph G with probability ρ independently from every other edge. Therefore, all graphs with p nodes and M edges have an equal probability of $\rho^M(1-\rho)^{\binom{p}{2}-M}$. All the edges in graph G correspond to the 1's in the resulted binary adjacency matrix. Next, we simulated the precision matrix $\boldsymbol{\Omega}$ following Peng et al. (2009). We started by setting all the diagonal elements of $\boldsymbol{\Omega}$ to be 1. Then, for the rest elements that correspond to the 1s in the adjacency matrix, we sampled their values independently from a uniform distribution

Unif($[-0.1, 0] \cup [0, 0.1]$). To ensure positive definiteness of the precision matrix, we followed Peng et al. (2009) by dividing each off-diagonal element by 1.5 times the sum of the absolute value of all the elements in its row. Finally, we averaged the rescaled precision matrix with its transpose and set the diagonal elements to 1. This process ensured the preceding matrix was positive definite and symmetric. The corresponding covariance matrix was set as $\mathbf{\Sigma} = \mathbf{\Omega}^{-1}$.

Next, we simulated n multivariate Gaussian variables from MN($\boldsymbol{\mu}, \mathbf{\Sigma}$) to represent the true underlying abundances $\boldsymbol{D}$. To obtain a count matrix that fully mimics the microbiome sequencing data, we generated counts from a DM model with parameter exp($\boldsymbol{D}$). Specifically, we first sampled the underlying fractional abundances for the ith sample from a Dirichlet distribution. The ith underlying fractional abundance was then denoted as $\boldsymbol{\psi}_i \sim$ Dirichlet(exp($\boldsymbol{D}_{i\cdot}$)). Next, the counts in the ith sample were generated from Multinomial($N_i, \boldsymbol{\psi}_i$). Finally, we randomly selected π_0% out of $n \times p$ counts and set them to zeros to mimic the zero-inflation observed in the real microbiome data. In general, the generative process had different assumptions from the proposed method. Under the appropriate choice of parameters, the simulated count data was zero-inflated, overdispersed, and the total reads varied largely between samples. In practice, we let $\rho = 0.1$ in the ER model. The mean parameter $\boldsymbol{\mu}$ of the underlying multivariate Gaussian variable was randomly sampled from a uniform distribution Unif[0, 10]. The number of total counts across samples N_i, $i = 1, \ldots, n$ was sampled from a discrete uniform distribution with range $[50,000, 100,000]$. Under each combination of n, p, and π_0, we generated 50 replicated datasets by repeating the process above.

2.3.2. Generating Synthetic Data

We generated synthetic data following the Normal-to-Anything (NorTA) approach proposed in Kurtz et al. (2015). NorTA was designed to generate multivariate random variables with an arbitrary marginal distribution from a pre-specified correlation structure (Cario and Nelson, 1997). Given the observations of p taxa from a real microbiome dataset, the NorTA generates the synthetic data with n samples as follows: (1) to calculate the p-by-p covariance matrix $\mathbf{\Sigma}_0$ from the input real dataset; (2) to generate an n-by-p matrix, denoted by $\boldsymbol{Z}_0$, from a multivariate Gaussian distribution with a mean of $\mathbf{0}_{1\times p}$ and the covariance matrix of $\mathbf{\Sigma}_0$; (3) to use standard normal cumulative distribution function to scale values in each column of $\boldsymbol{Z}_0$ within $[0, 1]$; (4) to apply the quantile function of a ZINB distribution to generate count data from those scaled values in each column of $\boldsymbol{Z}_0$. In practice, we used `R` package `SPIEC-EASI` to implement the above data generative scheme, where the real data were from those healthy control subjects in our case study presented in section 3.2. Under each combination of n and p, we generated 50 replicated datasets.

2.4. Model Performance

2.4.1. Alternative Methods in Network Learning

We considered the four commonly used network learning methods. The first two methods, SPIEC-EASI-Glasso and SPIEC-EASI-mb, use the transformed microbiome abundances which are different from the normalized abundances estimated by HARMONIES. Both infer the microbial network by estimating a sparse precision matrix. The former method (SPIEC-EASI-Glasso) measures the dependency among microbiota by their partial correlation coefficients, and the latter method (SPIEC-EASI-mb) uses the "neighborhood selection" introduced by Meinshausen and Bühlmann (2006) to construct the network. The third method, denoted as Pearson-corr, calculates Pearson's correlation coefficients between all pairs of taxa. In its estimated network, the edges correspond to large correlation coefficients. To avoid arbitrarily thresholding the correlation coefficients, the fourth method, CClasso (Fang et al., 2015), directly infers a sparse correlation matrix with l_1 regularization. However, as discussed in section 1, representing the dependency structure by the correlation matrix may lead to the detection of spurious associations.

2.4.2. Evaluation Criteria

We quantified the model performances on the simulated data by computing their receiver operating characteristic (ROC) curves and area under the ROC curve (AUC). For the HARMONIES or SPIEC-EASI, the network inference was based on the precision matrix. Hence, under each tuning parameter of Glasso, we calculated the number of edges being true positive (TP) by directly comparing the estimated precision matrix against the true one. More specifically, we considered an edge between taxon m and taxon k to be true positive if $\omega_{mk} \neq 0$, $\hat{\omega}_{mk} \neq 0$, and $\hat{\omega}_{mk}$ shared the same sign with ω_{mk}. We calculated the number of true negative (TN), false positive (FP), and false negative (FN) in a similar manner. Therefore, each tuning parameter defined a point on a ROC curve. As for the correlation-based methods, we started with ranking the absolute values in the estimated correlation matrices, denoted as $\hat{\boldsymbol{C}}$. Next, we used each value as a threshold and set all the entries in $\hat{\boldsymbol{C}}$ having their absolute values smaller than the current threshold to be zeros. Then, the number of TP, TN, FP, or FN was obtained by comparing the sparse $\hat{\boldsymbol{C}}$ against the true partial correlation matrix. Therefore, each unique absolute value in the original estimated correlation matrix defined a point on the ROC curve.

We further used the Matthew's correlation coefficient (MCC) to evaluate results from the simulated data. The MCC is defined as

$$\frac{(\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN})}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}.$$

Here, the MCC was particularly suitable for evaluating network models. As the number of conditionally independent taxa pairs was assumed to be much greater than the number of dependent pairs in a sparse network, MCC was preferable to quantify the performances under such an imbalanced situation. Note that MCC ranges from $[-1, 1]$, with a value close to 1 suggesting a better performance. Since each value of MCC was calculated using a given set of TP, TN, FP, and FN, we adopted the optimal choice of tuning parameter for the HARMONIES or SPIEC-EASI (with either Glasso or MB for network inference), given by StARS. As for the correlation-based methods, CClasso outputted a sparse correlation matrix. We used the result to calculate TP,

TN, FP, and FN directly. For Pearson-corr, we set the threshold such that the resulted number of nonzero entries in the sparse correlation matrix was the same as the number of non-zero entries in the true sparse partial correlation matrix. In fact, this choice could favor the performance of Pearson-corr for larger sample size, as shown in section 3.1.

To assess model performances on the synthetic datasets, we followed Kurtz et al. (2015) to use a metric called area under the precision-recall (AUPR) curves, in addition to AUC. Briefly speaking, the AUPR and AUC were calculated as follows: (1) to rank all possible edges according to their confidence values; (2) to generate the precision-recall curve and the ROC curve by comparing edge inclusions against the true sparse precision matrix; (3) to calculate the area under the precision-recall curve or the ROC curve. Note that the confidence values were chosen as the edge stabilities under the optimal choice of the tuning parameter selected by StARS for HARMONIES, SPIEC-EASI-Glasso, and SPIEC-EASI-mb, while for CClasso and Pearson-corr, p-values were used.

3. RESULTS

3.1. Simulation Results

Figures 1, **2** compare the AUCs and MCCs on the simulated data under various scenarios, including varying sample sizes ($n = 60, 100, 200$, or 500), total numbers of taxa ($p = 40$ or 60), extra percentages of zeros added ($\pi_0 = 10$, or 20%). In each subfigure, the HARMONIES outperformed the alternative methods in terms of both AUC and MCC, and it maintained this advantage even with the number of sample size greatly increases. Further, a smaller sample size, a larger proportion of extra zeros added ($\pi_0 = 20\%$), as well as a larger number of taxa in the network ($p = 60$), would hamper the performance of all the methods, as we expected. Two modes of SPIEC-EASI, SPIEC-EASI-Glasso, and SPIEC-EASI-mb, showed very similar performances under all the scenarios, with SPIEC-EASI-Glasso having only a marginal advantage over the other. Further, we observed that the Pearson-corr method yielded higher AUCs even than the precision matrix based methods, especially when there was a lager proportion of extra zeros or larger number of taxa in the network. This result suggested that the Pearson-corr could capture the overall rank of the signal strength in the actual network. However, under a fixed cut-off value that gave a sparse correlation network, the MCCs from the Pearson-corr were always smaller than the precision matrix based methods. Note that the cut-off value we specified for Pearson's correlation method indeed favored its performance. In general, the alternative methods considered here were able to reflect the overall rank of the signal strength by showing reasonable AUCs. However, they failed to give an accurate estimation of the network under a fixed cut-off value.

Figure 3 demonstrates that our model outperformed all others on the synthetic datasets. The performances in terms of AUC under different scenarios are summarized in **Figures 3A,B**, while those in terms of AUPR are displayed in (**Figures 3C,D**). As we can see, either increasing the sample size n or decreasing the number of features p would improve the performance of all methods and lead to greater disparity between partial and pairwise correlation-based methods. In general, our HARMONIES maintained the best in all simulation and evaluation settings except for one case, where the SPIEC-EASI-mb only showed a marginal advantage (see $n = 60$ in **Figure 3C**). Interestingly, our observation confirmed a finding mentioned by Kurtz et al. (2015), that is, the SPIEC-EASI-mb was slightly better than SPIEC-EASI-Glasso in terms of AUPR under the optimal choice of the tuning parameter. As for the two correlation-based methods, we found that Pearson-corr outperformed CClasso in most of the scenarios.

3.2. Analysis of Microbiome Data From Colorectal Cancer Patients

Colorectal cancer (CRC) is the third most common cancer diagnosed in both men and women in the United States (Arnold et al., 2017). Increasing evidence from recent studies highlights a vital role for the intestinal microbiota in malignant gastrointestinal diseases including CRC (Louis et al., 2014; Sears and Garrett, 2014; Drewes et al., 2016). In particular, studies have reported that dysbiosis of specific microbiota is directly associated with CRC (Marchesi et al., 2011; Kostic et al., 2013; Flynn et al., 2016). The current microbiome research interests have gone beyond the discovery of disease-related microbiota, with a growing number of studies investigating the interactive associations among the microbial taxa. Using the proposed model, we interrogated the microbiome profiling data of a CRC study to determine the microbiome network structures.

We analyzed the gut microbiome dataset of a CRC study published by Feng et al. (2015). We extracted from the original cohort[1] the 43 CRC patients and the 58 healthy controls. The original sequencing data at the genus level were quantified using curatedMetagenomicData (Pasolli et al., 2017). We had $p = 187$ genera for both the 43 CRC patients and the 58 healthy controls. We implemented the HARMONIES as follows. For the CRC group, we first applied the ZINB model to obtain the normalized abundance matrix $\boldsymbol{A}$, utilizing the specifications detailed in section 2.1. We then took the logarithmic transformation of the normalized abundance and imputed the missing values. Before implementing the proposed method, we filtered out the low abundant genera with zeros occurring more than half samples. Removing low abundant taxa is a common step in microbiome research (see e.g., Qin et al., 2014; Zeller et al., 2014; Kostic et al., 2015; Kurtz et al., 2015; Wadsworth et al., 2017; Yilmaz et al., 2019). The rationale being that these "zero-abundant" taxa may be less important in a network, which was also confirmed by our simulation study. This filtering process left 51 and 36 genera in the CRC and control group, respectively. The result from using a more relaxed filtering threshold is available in the supplement, where we kept the genera that had at least 10% nonzero observations across the samples.

Figures 4A,B display the estimated networks for the CRC and the control group, respectively. Each node, corresponding to a genus, was named after its phylum level. All the genera

[1]The original metagenomic shotgun sequencing data from the fecal samples are available in the European Bioinformatics Institute Database (accession number ERP008729).

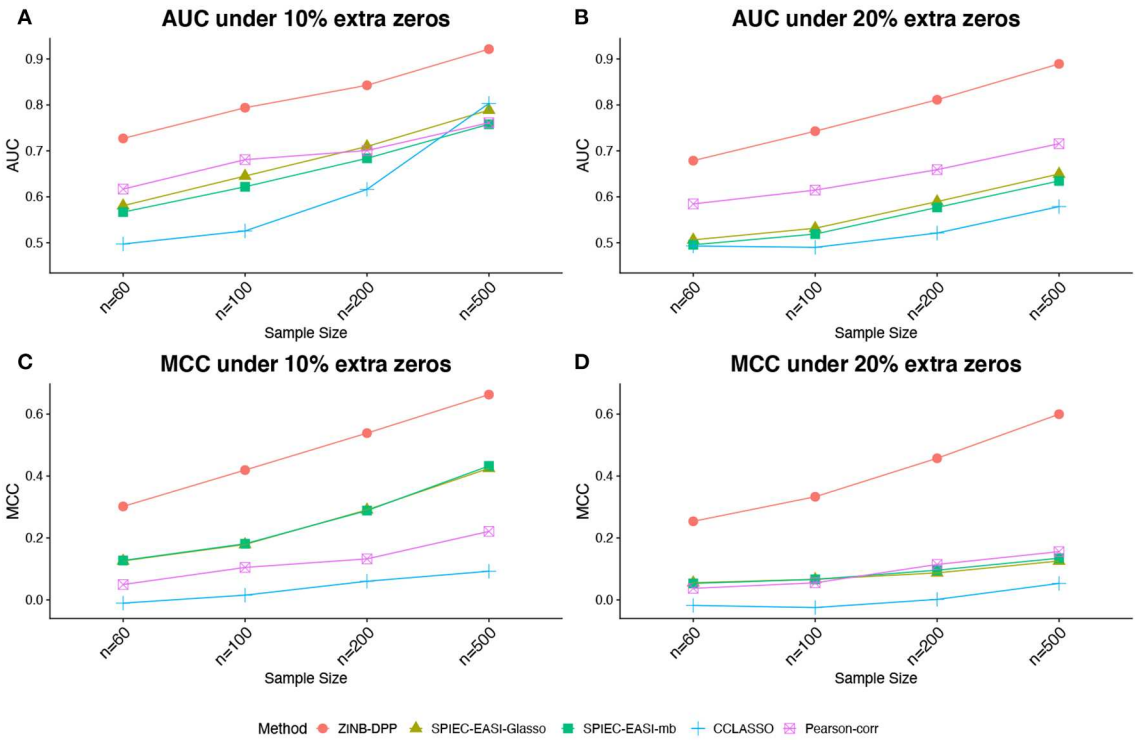

FIGURE 1 | Simulated data: **(A,B)** area under the ROC curves (AUCs) and **(C,D)** Matthew's correlation coefficient (MCCs) achieved by different methods under the number of taxa $p = 40$ and different sample sizes and zero proportions, averaged over 50 replicates.

shown in **Figure 4** belong to six phyla in total. By using their phylum name to further categorize these distinct genera, we aimed at exploring interesting patterns among them at a higher taxonomic level. **Figure S1** displays the same network using the actual genus name on each node. The node sizes are proportional to its normalized abundances in the logarithmic scale. The green or red edge indicates a positive or a negative partial correlation, respectively. And the width of an edge is proportional to the absolute value of the partial correlation coefficient. To make a clear comparison, we intentionally kept the nodes and their positions to be consistent between the two subfigures. In either of the two groups, we included a node in the current plot if there exists an edge between it with any nodes in at least one group. In general, the two groups share several edges with the same direction of partial correlations, but the majority of edges are unique within each group.

Network estimation of the CRC group demonstrated several microbial communities. For example, three genera: *Fusobacterium*, *Peptostreptococcus*, and *Parvimonas* consisted of a unique subnetwork as highlighted in **Figure 4A**. These three genera were isolated in the control group's network, as shown in **Figure 4B**. Interestingly, specific species under these three genera have been reported as enriched taxa in CRC and related to worse clinical outcome (Mima et al., 2016; Yu et al., 2017; Long et al., 2019). A previous CRC study by Kostic et al. (2013) supported the causal role of species *Fusobacterium nucleatum* by showing that *F. nucleatum* promotes tumor progression by increasing both tumor multiplicity and tumor-infiltrating myeloid cells in a preclinical CRC model. Further, a recent study (Long et al., 2019) demonstrated that *Peptostreptococcus anaerobius* accelerated colorectal tumorigenesis in a murine CRC model. This study suggested that *P. anaerobius* directly interacted with colonic epithelial cells and also promoted CRC by modifying the tumor immune microenvironment. While the causal role of the species *Parvimonas micra* has not been biologically validated, multiple clinical studies reported an elevated level of *P. micra* in CRC patients (Purcell et al., 2017; Yu et al., 2017; Dai et al., 2018). Of interest, *Parvimonas* were closely associated with animal-based diets, which have previously been shown to be significantly associated with increased risk for CRC (Chan et al., 2011). The previous studies only investigated those CRC-related taxa individually, whereas a novel finding by HARMONIES analysis suggested that all the three genera were co-aggregating in CRC patients as their pairwise associations are all positive. Interestingly, in a prior study direct positive associations between *Fusobacterium* and *Peptostreptococcus*, as well as *Peptostreptococcus* and *Parvimonas*, were identified (Hibberd et al., 2017). However, there was no direct association between *Fusobacterium* and *Parvimonas*. Similarly, another study (Drewes et al., 2017) found a direct co-occurrence pattern between two species: *F. nucleatum* and *P. micra*. Using HARMONIES, we could jointly identify the relationship among each pair of the three genera,

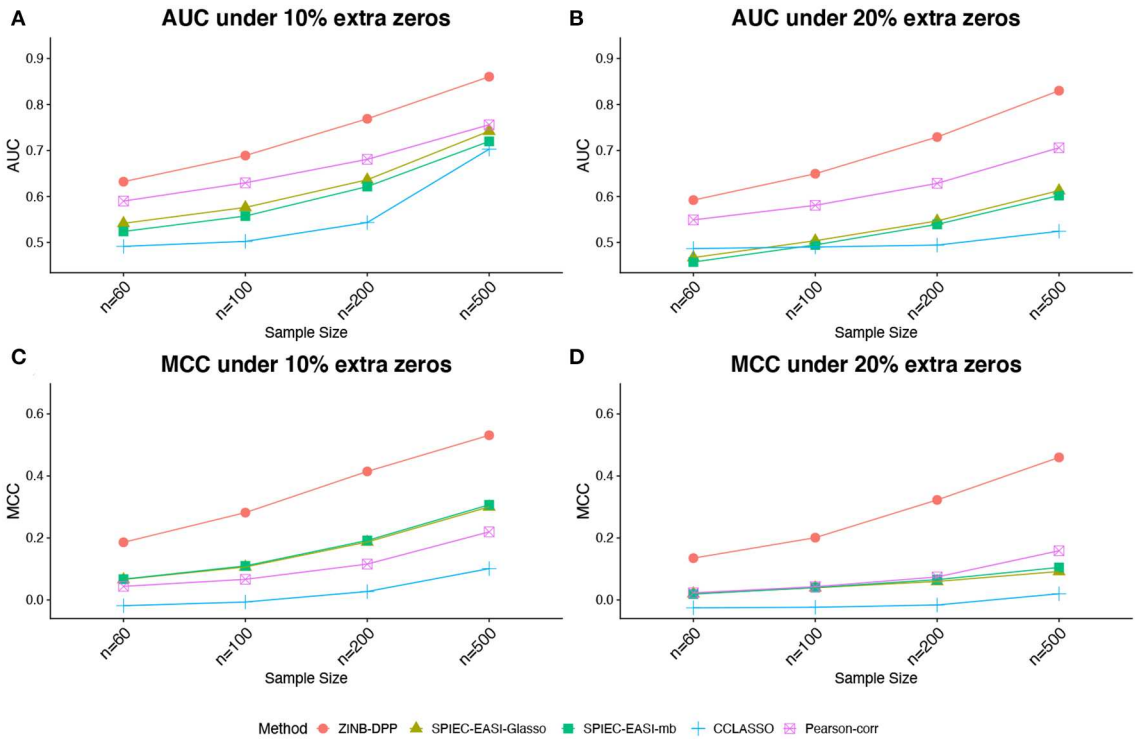

FIGURE 2 | Simulated data: **(A,B)** area under the ROC curves (AUCs) and **(C,D)** Matthew's correlation coefficient (MCCs) achieved by different methods under the number of taxa $p = 60$ and different sample sizes and zero proportions, averaged over 50 replicates.

conditional on all other genera. This novel subcommunity of three CRC-enriched genera formulated a recurring module and may function as a cooperative group in CRC patients. A closer investigation of their co-occurrence pattern could potentially elucidate both their contributions to CRC and the basic biology under their relationships. Two additional novel taxa interactions were identified by HARMONIES analysis: *Streptococcus* and *Veillonella*, and *Streptococcus* and *Haemophilus*. In fact, previous CRC studies showed enrichment of these three genera or their species in CRC patients (see e.g., Geng et al., 2014; Ugai et al., 2014; Kumar et al., 2017; Koliarakis et al., 2019), but had not detected these novel interactions. In conclusion, HARMONIES may reveal how multiple CRC-related taxa could potentially promote disease progression together.

Having shared edges between the two networks suggests that the HARMONIES is robust to the edge selection. We observed that the shared edges tended to appear for those more abundant genera. For example, we circled eight genera in **Figure 4B**, and the HARMONIES suggested multiple positive partial correlations among them. For these eight genera, we observed six shared edges between the CRC and healthy control networks. Notice that all the shared edges were consistent in the association directions, and they also corresponded to the relatively stronger association in both networks (wider in the edge width). We found these shared edges tend to connect those more abundant genera (node with larger size). Indeed, the eight genera considered here belong to phyla *Bacteroidetes* and *Firmicutes*, both were in the top three most abundant phyla for CRC patients and healthy controls reported by Gao et al. (2015) and Mori et al. (2018). Therefore, it was more likely that the highly abundant genera shared similar association patterns between the two groups, and the HARMONIES demonstrated its robustness by preserving these relatively stronger partial correlations among these genera. On the other hand, the network of the control group contained more negative partial correlations as shown in **Figure 4B**. Furthermore, the two edges linked to *Streptococcus* were different from the CRC group. Here, *Streptococcus* had a negative association with *Subdoligranulum* and a positive association with *Rothia*. There has been no evidence suggesting these two genera are CRC-related. Hence a further investigation is merited. Additionally, the CRC group has another distinct small subnetwork formed by the four genera, two from *Firmicutes*, one from *Proteobacteria*, and one from *Verrucomicrobia*. These group-specific associations were never reported. Lastly, we observed several interesting patterns between the two groups when summarizing the genera to their phylum levels. Genera in Firmicutes (labeled as "Fm" in **Figure 4**) showed more positive associations in the case group than in the control group, whereas negative associations between Firmicutes and Bacteroidetes (labeled as "Ba" in **Figure 4**) were more common in the control group. Again, these novel patterns still need further biological validations to elucidate their functions.

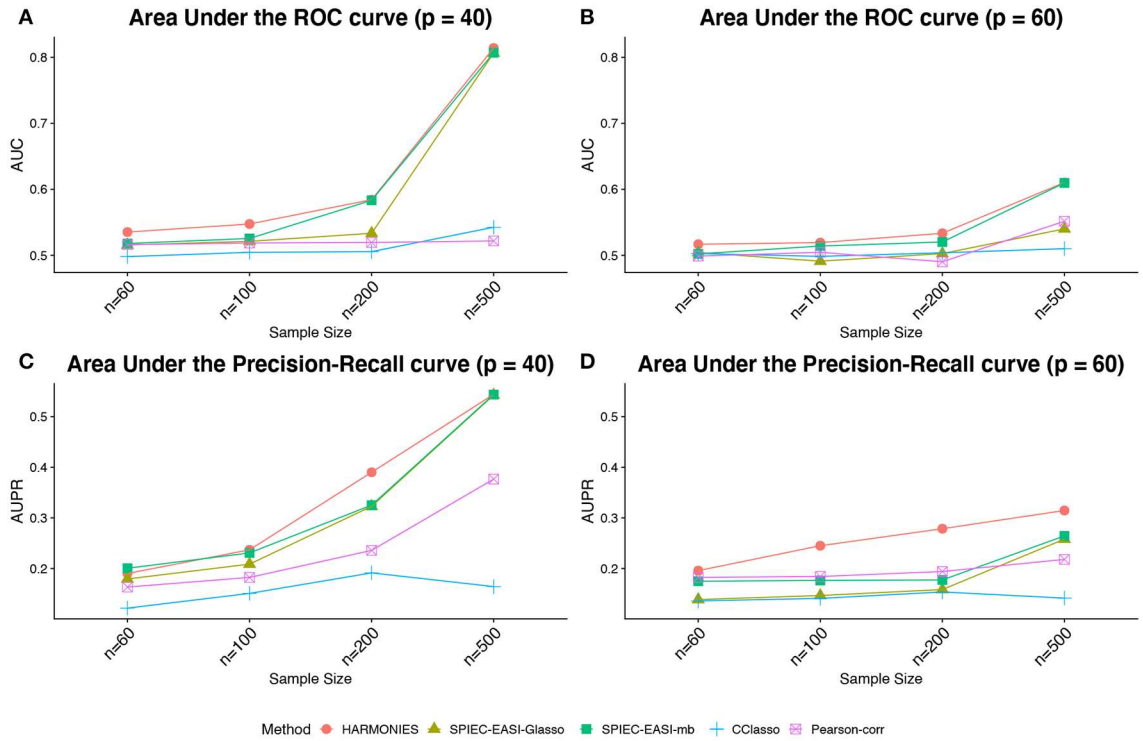

FIGURE 3 | Synthetic data: **(A,B)** area under the ROC curves (AUCs) and **(C,D)** area under the precision-recall curves (AUPRs) achieved by different methods under different sample sizes and taxa numbers, averaged over 50 replicates.

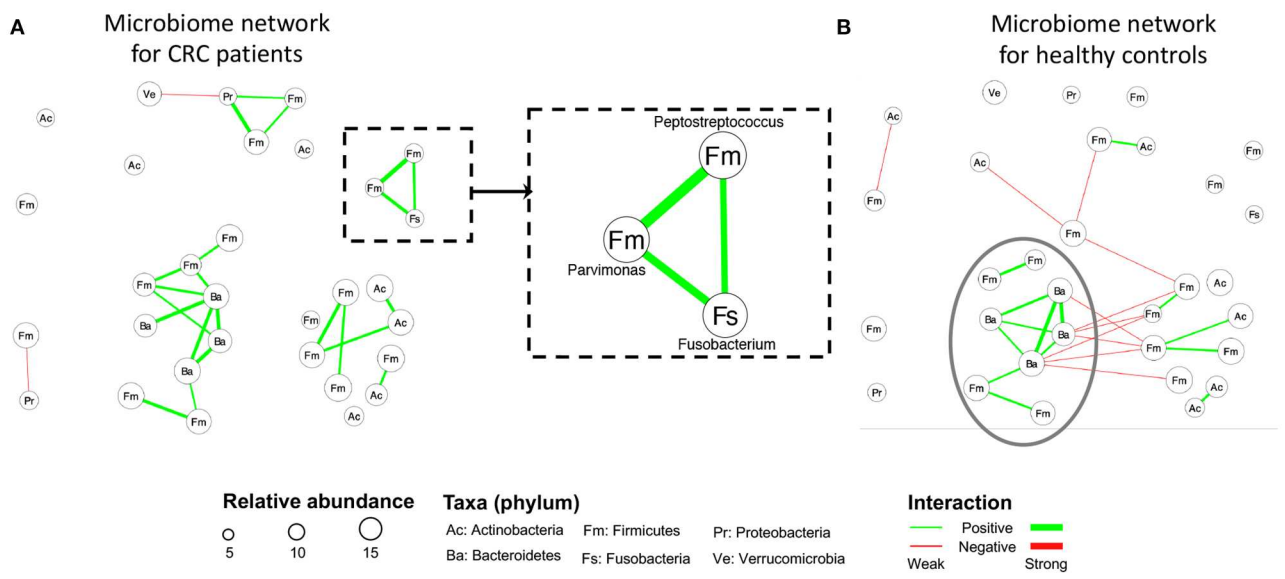

FIGURE 4 | CRC case study: The estimated networks by HARMONIES for **(A)** CRC patients and **(B)** healthy controls. Increased abundances of species under the three genera (*Fusobacterium, Peptostreptococcus, Parvimonas*) in the dashed rectangular box in **(A)** were reported to be associated with the disease. CRC patients and healthy controls shared a similar subnetwork (composed of eight genera) circled in **(B)**. Each node here represents a genus labeled by its phylum name. The version with distinct genus names is available in **Figure S1** in the supplement.

4. DISCUSSION

With the advent of next-generation sequencing technology, microbiome research now has the opportunity to explore microbial community structure and to characterize the microbial ecological association for different populations or physiology conditions (Kurtz et al., 2015). In this paper, we introduce HARMONIES as a statistical framework to infer sparse networks using microbiome sequencing data. It models the original count data by a zero-inflated negative binomial distribution to capture the large amount for zeros and over-dispersion, and it further implements Dirichlet process priors to account for sample heterogeneity. In contrast, current methods for microbiome network analyses rely on the compositional data, which could cause information loss due to ignoring the unique characteristics of the microbiome sequencing count data. Following the data normalization step, the HARMONIES explores the direct connections in the network by estimating the partial correlations. The results from the simulation study have demonstrated the advantage of the HARMONIES over alternative approaches under various conditions. When applied to an actual microbiome dataset, the HARMONIES suggests all the nodes to be taxa at the same taxonomic level, such as species, genus, family, etc. This ensures proper biologically interpretations of those detected associations. When applied to a real CRC study, the HARMONIES revealed an intriguing community among three CRC-enriched genera. Further, shared patterns between the CRC and the control networks suggest a common community pattern of disease neutral genera. Additional studies validating the biological relevance of these microbial associations, however, will need to be conducted.

Both the simulated and synthetic data showed that a larger sample size improved the performance of all the network learning methods. In practice, many disease-related microbiome studies, especially those studying rare diseases, always have small sample sizes. This limitation directly affects the estimation of the normalized matrix $\boldsymbol{A}$ from the ZINB model. Notice that for a taxon j, a small sample size could result in a large variance in the posterior distribution of $\log \boldsymbol{\alpha}_{\cdot j}$. However, many disease studies include reference groups where the measurements on the same taxonomic features are available. The additional information from the subjects in the reference group can potentially help improve the posterior inference of the normalized abundances. We generalized the proposed ZINB model to handle two groups, with the goal of borrowing information between groups in estimating the normalized abundances. These detailed model formula and implementation were included in the supplement (see **Supplementary Material**: section 2).

Our hybrid approach for microbiome network inference can be extended. One future direction is to incorporate the differential network analysis into the existing framework. It jointly considers the association strengths between each pair of taxa from different groups, and it compares the estimated individual networks to capture the significantly different connectivities. Our current method can infer the normalized abundances for two groups, and we provided the details steps in the supplement. However, an integrated differential network can be expected to better study the differential microbial community structure and link the communities to human health status.

SOFEWARE AVAILABILITY STATEMENT

We developed the R package HARMONIES that is freely available at https://github.com/shuangj00/HARMONIES. We also released a webtool that allows users to upload micorbiome datasets and run HARMONIES for microbiome network analysis. The online webtool is available at http://lce.biohpc.swmed.edu/harmonies.

DATA AVAILABILITY STATEMENT

The datasets generated for the simulation study can be found in the author's Github page—https://github.com/shuangj00/HARMONIES. The original metagenomic shotgun sequencing data analyzed in the real data study are available in the European Bioinformatics Institute Database (accession number ERP008729).

AUTHOR CONTRIBUTIONS

SJ performed the experiments. GX and AK provided resources and helpful discussions. SJ, QL, and XZ designed the experiment, performed data analysis, wrote the software, and the manuscript. SJ, YC, BY, and XZ developed the website for online implementation of HARMONIES.

FUNDING

This work was supported by the National Institutes of Health [5P30CA142543, 5R01GM126479, 5R01HG008983] and Cancer Prevention & Research Institute of Texas [CPRIT RP190107].

ACKNOWLEDGMENTS

We thank Jiwoong Kim for the helpful discussions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00445/full#supplementary-material

REFERENCES

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

Arnold, M., Sierra, M. S., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2017). Global patterns and trends in colorectal cancer incidence and mortality. *Gut* 66, 683–691. doi: 10.1136/gutjnl-2015-310912

Ban, Y., An, L., and Jiang, H. (2015). Investigating microbial co-occurrence patterns based on metagenomic compositional

data. *Bioinformatics* 31, 3322–3329. doi: 10.1093/bioinformatics/btv364

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

Cario, M. C., and Nelson, B. L. (1997). *Modeling and Generating Random Vectors With Arbitrary Marginal Distributions and Correlation Matrix*. Technical Report, Citeseer.

Chan, D. S., Lau, R., Aune, D., Vieira, R., Greenwood, D. C., Kampman, E., et al. (2011). Red and processed meat and colorectal cancer incidence: meta-analysis of prospective studies. *PLoS ONE* 6:e20456. doi: 10.1371/journal.pone.0020456

Dai, Z., Coker, O. O., Nakatsu, G., Wu, W. K., Zhao, L., Chen, Z., et al. (2018). Multi-cohort analysis of colorectal cancer metagenome identified altered bacteria across populations and universal bacterial markers. *Microbiome* 6:70. doi: 10.1186/s40168-018-0451-2

Drewes, J. L., Housseau, F., and Sears, C. L. (2016). Sporadic colorectal cancer: microbial contributors to disease prevention, development and therapy. *Br. J. Cancer* 115:273. doi: 10.1038/bjc.2016.189

Drewes, J. L., White, J. R., Dejea, C. M., Fathi, P., Iyadorai, T., Vadivelu, J., et al. (2017). High-resolution bacterial 16S rRNA gene profile meta-analysis and biofilm status reveal common colorectal cancer consortia. *NPJ Biofilms Microb*. 3, 1–12. doi: 10.1038/s41522-017-0040-3

Fang, H., Huang, C., Zhao, H., and Deng, M. (2015). CClasso: correlation inference for compositional data through lasso. *Bioinformatics* 31, 3172–3180. doi: 10.1093/bioinformatics/btv349

Faust, K., and Raes, J. (2016). CoNet app: inference of biological association networks using Cytoscape. *F1000Res*. 5:1519. doi: 10.12688/f1000research.9050.2

Feng, Q., Liang, S., Jia, H., Stadlmayr, A., Tang, L., Lan, Z., et al. (2015). Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nat. Commun*. 6:6528. doi: 10.1038/ncomms7528

Flynn, K. J., Baxter, N. T., and Schloss, P. D. (2016). Metabolic and community synergy of oral bacteria in colorectal cancer. *Msphere* 1: e00102-16. doi: 10.1128/mSphere.00102-16

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9, 432–441. doi: 10.1093/biostatistics/kxm045

Gao, Z., Guo, B., Gao, R., Zhu, Q., and Qin, H. (2015). Microbiota disbiosis is associated with colorectal cancer. *Front. Microbiol*. 6:20. doi: 10.3389/fmicb.2015.00020

Geng, J., Song, Q., Tang, X., Liang, X., Fan, H., Peng, H., et al. (2014). Co-occurrence of driver and passenger bacteria in human colorectal cancer. *Gut Pathogens* 6:26. doi: 10.1186/1757-4749-6-26

Gevers, D., Kugathasan, S., Denson, L. A., Vázquez-Baeza, Y., Van Treuren, W., Ren, B., et al. (2014). The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* 15, 382–392. doi: 10.1016/j.chom.2014.02.005

Hibberd, A. A., Lyra, A., Ouwehand, A. C., Rolny, P., Lindegren, H., Cedgåård, L., et al. (2017). Intestinal microbiota is altered in patients with colon cancer and modified by probiotic intervention. *BMJ Open Gastroenterol*. 4:e000145. doi: 10.1136/bmjgast-2017-000145

Koliarakis, I., Messaritakis, I., Nikolouzakis, T. K., Hamilos, G., Souglakos, J., and Tsiaoussis, J. (2019). Oral bacteria and intestinal dysbiosis in colorectal cancer. *Int. J. Mol. Sci*. 20:4146. doi: 10.3390/ijms20174146

Kostic, A. D., Chun, E., Robertson, L., Glickman, J. N., Gallini, C. A., Michaud, M., et al. (2013). *Fusobacterium nucleatum* potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. *Cell Host Microbe* 14, 207–215. doi: 10.1016/j.chom.2013.07.007

Kostic, A. D., Gevers, D., Siljander, H., Vatanen, T., Hyötyläinen, T., Hämäläinen, A.-M., et al. (2015). The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. *Cell Host Microbe* 17, 260–273. doi: 10.1016/j.chom.2015.01.001

Kumar, R., Herold, J. L., Schady, D., Davis, J., Kopetz, S., Martinez-Moczygemba, M., et al. (2017). *Streptococcus gallolyticus* subsp. gallolyticus promotes colorectal tumor development. *PLoS Pathogens* 13:e1006440. doi: 10.1371/journal.ppat.1006440

Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., and Bonneau, R. A. (2015). Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.* 11:e1004226. doi: 10.1371/journal.pcbi.1004226

Kyung, M., Gill, J., and Casella, G. (2011). Sampling schemes for generalized linear Dirichlet process random effects models. *Stat. Methods Appl*. 20, 259–290. doi: 10.1007/s10260-011-0168-x

Lam, C., and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Stat*. 37:4254. doi: 10.1214/09-AOS720

Lee, J., and Sison-Mangus, M. (2018). A Bayesian semiparametric regression model for joint analysis of microbiome data. *Front. Microbiol*. 9:522. doi: 10.3389/fmicb.2018.00522

Li, Q., Cassese, A., Guindani, M., and Vannucci, M. (2019). Bayesian negative binomial mixture regression models for the analysis of sequence count and methylation data. *Biometrics* 75, 183–192. doi: 10.1111/biom.12962

Li, Q., Guindani, M., Reich, B. J., Bondell, H. D., and Vannucci, M. (2017). A Bayesian mixture model for clustering and selection of feature occurrence rates under mean constraints. *Stat. Anal. Data Min*. 10, 393–409. doi: 10.1002/sam.11350

Liu, H., Roeder, K., and Wasserman, L. (2010). "Stability approach to regularization selection (StARS) for high dimensional graphical models," in *Advances in Neural Information Processing Systems* (Vancouver, CA), 1432–1440.

Lo, C., and Marculescu, R. (2017). MPLasso: Inferring microbial association networks using prior microbial knowledge. *PLoS Comput. Biol*. 13:e1005915. doi: 10.1371/journal.pcbi.1005915

Long, X., Wong, C. C., Tong, L., Chu, E. S., Szeto, C. H., Go, M. Y., et al. (2019). Peptostreptococcus anaerobius promotes colorectal carcinogenesis and modulates tumour immunity. *Nat. Microbiol*. 4, 2319–2330. doi: 10.1038/s41564-019-0541-3

Louis, P., Hold, G. L., and Flint, H. J. (2014). The gut microbiota, bacterial metabolites and colorectal cancer. *Nat. Rev. Microbiol*. 12, 661–672. doi: 10.1038/nrmicro3344

Marchesi, J. R., Dutilh, B. E., Hall, N., Peters, W. H., Roelofs, R., Boleij, A., et al. (2011). Towards the human colorectal cancer microbiome. *PLoS ONE* 6:e20447. doi: 10.1371/journal.pone.0020447

Meinshausen, N., and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Stat*. 34, 1436–1462. doi: 10.1214/009053606000000281

Menéndez, P., Kourmpetis, Y. A., ter Braak, C. J., and van Eeuwijk, F. A. (2010). Gene regulatory networks from multifactorial perturbations using Graphical Lasso: application to the DREAM4 challenge. *PLoS ONE* 5:e14147. doi: 10.1371/journal.pone.0014147

Metzker, M. L. (2010). Sequencing technologies–the next generation. *Nat. Rev. Genet*. 11:31. doi: 10.1038/nrg2626

Mima, K., Nishihara, R., Qian, Z. R., Cao, Y., Sukawa, Y., Nowak, J. A., et al. (2016). *Fusobacterium nucleatum* in colorectal carcinoma tissue and patient prognosis. *Gut* 65, 1973–1980. doi: 10.1136/gutjnl-2015-310101

Mori, G., Rampelli, S., Orena, B. S., Rengucci, C., De Maio, G., Barbieri, G., et al. (2018). Shifts of faecal microbiota during sporadic colorectal carcinogenesis. *Sci. Rep*. 8:10329. doi: 10.1038/s41598-018-28671-9

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* 5, 155–176. doi: 10.1093/biostatistics/5.2.155

Oh, J. H., and Deasy, J. O. (2014). Inference of radio-responsive gene regulatory networks using the graphical lasso algorithm. *BMC Bioinformatics* 15:S5. doi: 10.1186/1471-2105-15-S7-S5

Pasolli, E., Schiffer, L., Manghi, P., Renson, A., Obenchain, V., Truong, D. T., et al. (2017). Accessible, curated metagenomic data through experimenthub. *Nat. Methods* 14:1023. doi: 10.1038/nmeth.4468

Paulson, J. N., Stine, O. C., Bravo, H. C., and Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods* 10:1200. doi: 10.1038/nmeth.2658

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *J. Am. Stat. Assoc*. 104, 735–746. doi: 10.1198/jasa.2009.0126

Purcell, R. V., Visnovska, M., Biggs, P. J., Schmeier, S., and Frizelle, F. A. (2017). Distinct gut microbiome patterns associate with consensus molecular

subtypes of colorectal cancer. *Sci. Rep*. 7:11590. doi: 10.1038/s41598-017-11237-6

Qin, N., Yang, F., Li, A., Prifti, E., Chen, Y., Shao, L., et al. (2014). Alterations of the human gut microbiome in liver cirrhosis. *Nature* 513, 59–64. doi: 10.1038/nature13568

Robinson, M. D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*. 11:R25. doi: 10.1186/gb-2010-11-3-r25

Sears, C. L., and Garrett, W. S. (2014). Microbes, microbiota, and colon cancer. *Cell Host Microbe* 15, 317–328. doi: 10.1016/j.chom.2014.02.007

Taddy, M. A., and Kottas, A. (2012). Mixture modeling for marked Poisson processes. *Bayesian Anal*. 7, 335–362. doi: 10.1214/12-BA711

Tipton, L., Müller, C. L., Kurtz, Z. D., Huang, L., Kleerup, E., Morris, A., et al. (2018). Fungi stabilize connectivity in the lung and skin microbial ecosystems. *Microbiome* 6:12. doi: 10.1186/s40168-017-0393-0

Ugai, T., Norizuki, M., Mikawa, T., Ohji, G., and Yaegashi, M. (2014). Necrotizing fasciitis caused by haemophilus influenzae type b in a patient with rectal cancer treated with combined bevacizumab and chemotherapy: a case report. *BMC Infect. Dis*. 14:198. doi: 10.1186/1471-2334-14-198

Wadsworth, W. D., Argiento, R., Guindani, M., Galloway-Pena, J., Shelburne, S. A., and Vannucci, M. (2017). An integrative bayesian dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. *BMC Bioinformatics* 18:94. doi: 10.1186/s12859-017-1516-0

Weiss, S., Van Treuren, W., Lozupone, C., Faust, K., Friedman, J., Deng, Y., et al. (2016). Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J*. 10:1669. doi: 10.1038/ismej.2015.235

Yilmaz, B., Juillerat, P., Øyås, O., Ramon, C., Bravo, F. D., Franc, Y., et al. (2019). Microbial network disturbances in relapsing refractory Crohn's disease. *Nat. Med*. 25, 323–336. doi: 10.1038/s41591-018-0308-z

Yu, J., Feng, Q., Wong, S. H., Zhang, D., Liang, Q., Qin, Y., et al. (2017). Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut* 66, 70–78. doi: 10.1136/gutjnl-2015-309800

Zeller, G., Tap, J., Voigt, A. Y., Sunagawa, S., Kultima, J. R., Costea, P. I., et al. (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol*. 10:766. doi: 10.15252/msb.20145645

Zhao, H., and Duan, Z.-H. (2019). Cancer genetic network inference using Gaussian graphical models. *Bioinform. Biol. Insights* 13:1177932219839402. doi: 10.1177/1177932219839402

Conflict of Interest: AK is a consultant for Merck and the principal investigator on a Norvatis sponsored study.

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 10 September 2020
doi: 10.3389/fgene.2020.500064

A Machine Learning Approach to Predicting Autism Risk Genes: Validation of Known Genes and Discovery of New Candidates

Ying Lin[1], Shiva Afshar[1], Anjali M. Rajadhyaksha[2,3,4], James B. Potash[5] and Shizhong Han[5,6]**

[1] Department of Industrial Engineering, University of Houston, Houston, TX, United States, [2] Division of Pediatric Neurology, Department of Pediatrics, Weill Cornell Medicine, New York, NY, United States, [3] Feil Family Brain & Mind Research Institute, Weill Cornell Medicine, New York, NY, United States, [4] Weill Cornell Autism Research Program, Weill Cornell Medicine, New York, NY, United States, [5] Department of Psychiatry and Behavioral Sciences, Johns Hopkins School of Medicine, Baltimore, MD, United States, [6] Lieber Institute for Brain Development, Baltimore, MD, United States

OPEN ACCESS

Edited by:
Mogens Fenger, The Capital Region of Denmark, Denmark

Reviewed by:
Weihua Guan, University of Minnesota, United States
Sagiv Shifman, The Hebrew University of Jerusalem, Israel

***Correspondence:**
Anjali M. Rajadhyaksha
amr2011@med.cornell.edu
Shizhong Han
shan67@jhmi.edu

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

Received: *23 September 2019*
Accepted: *13 August 2020*
Published: *10 September 2020*

Citation:
Lin Y, Afshar S, Rajadhyaksha AM, Potash JB and Han S (2020) A Machine Learning Approach to Predicting Autism Risk Genes: Validation of Known Genes and Discovery of New Candidates. Front. Genet. 11:500064. doi: 10.3389/fgene.2020.500064

Autism spectrum disorder (ASD) is a complex neurodevelopmental condition with a strong genetic basis. The role of *de novo* mutations in ASD has been well established, but the set of genes implicated to date is still far from complete. The current study employs a machine learning-based approach to predict ASD risk genes using features from spatiotemporal gene expression patterns in human brain, gene-level constraint metrics, and other gene variation features. The genes identified through our prediction model were enriched for independent sets of ASD risk genes, and tended to be down-expressed in ASD brains, especially in frontal and parietal cortex. The highest-ranked genes not only included those with strong prior evidence for involvement in ASD (for example, *NBEA*, *HERC1*, and *TCF20*), but also indicated potentially novel candidates, such as, *MYCBP2* and *CAND1*, which are involved in protein ubiquitination. We also showed that our method outperformed state-of-the-art scoring systems for ranking curated ASD candidate genes. Gene ontology enrichment analysis of our predicted risk genes revealed biological processes clearly relevant to ASD, including neuronal signaling, neurogenesis, and chromatin remodeling, but also highlighted other potential mechanisms that might underlie ASD, such as regulation of RNA alternative splicing and ubiquitination pathway related to protein degradation. Our study demonstrates that human brain spatiotemporal gene expression patterns and gene-level constraint metrics can help predict ASD risk genes. Our gene ranking system provides a useful resource for prioritizing ASD candidate genes.

Keywords: autism, *de novo* mutation, gene expression, constraint, machine learning

INTRODUCTION

Autism spectrum disorder (ASD) is a neurodevelopmental condition characterized by impaired social interaction and communication, as well as repetitive behavior. While its etiology is complex, ASD has a strong genetic basis (Hallmayer et al., 2011; Jeste and Geschwind, 2014; Colvert et al., 2015). The role of *de novo* mutations in ASD has been firmly established through candidate

gene (Wang et al., 2016; Stessman et al., 2017), whole exome (Iossifov et al., 2012, 2014; Sanders et al., 2012; Ronemus et al., 2014), and whole genome sequencing studies (Ryan et al., 2017; Turner et al., 2017). Although the list of risk genes implicated by *de novo* mutations is growing, it is still very likely far from complete, with an estimated full set of ASD genes ranging from several hundred to more than 1,000 (Iossifov et al., 2014). In the search for additional *de novo* mutations, sequencing studies continue to be an important approach, but the current sequencing cost is still very high, especially for large samples. As an alternative strategy, advanced analytical approaches, which leverage previously implicated genes and prior knowledge, have the potential to enhance risk gene discovery in an efficient and cost-effective manner (Asif et al., 2018; Gök, 2018; Brueggeman et al., 2020).

One approach is based on the concept of guilt-by-association, i.e., assuming that genes that confer risk for ASD are likely to be functionally related, and that they thus converge on molecular networks and biological pathways implicated in disease (Gandhi et al., 2006; Xu and Li, 2006). For example, one study showed that ASD genes with *de novo* mutations converged on pathways related to chromatin remodeling and synaptic function (Krumm et al., 2014). To leverage these functional relationships, several studies have explored integrating known risk genes using a protein-protein interaction (PPI) network to identify novel genes involved in ASD (Gilman et al., 2011; Li et al., 2014; Hormozdiari et al., 2015; Liu et al., 2015). However, a PPI network is built upon general PPIs without reference to tissue or cell type specificity, and this approach may not fully capture the brain-centric functional relationships among ASD genes. Accordingly, a brain-specific network-based approach, which considered relationships within the context of the brain, was proposed to predict ASD genes (Krishnan et al., 2016; Duda et al., 2018). Studies employing this paradigm, however, did not consider the dynamic patterns of gene relationships during brain development, thereby limiting their potential for discovery, given the possibility that genes might only be functionally related within a specific developmental stage. Evidence for this comes from Willsey et al. (2013) who showed, using spatiotemporal gene expression data from human brain, that co-expression patterns of ASD risk genes varied by spatiotemporal windows, with the strongest co-expression patterns observed in the prefrontal and primary motor–somatosensory cortical regions during midfetal development, suggesting an important convergence of risk gene activity in particular places at a particular time.

In addition to having functional relationships, ASD genes affected by *de novo* mutations tend to be intolerant of variations (Samocha et al., 2014; Iossifov et al., 2015). With the availability of sequencing data from large samples, recent work has developed measures to quantify the sensitivity of genes to disruptive functional variations (Petrovski et al., 2013; Lek et al., 2016). Utilizing exome data on more than 60,000 individuals from the Exome Aggregation Consortium (ExAC), a gene-level constraint metric–the probability of being loss-of-function (LoF) intolerant (pLI)–was created, which separates genes into LoF intolerant or LoF tolerant (Lek et al., 2016). Kosmicki et al. (2017) further demonstrated that the excess of *de novo* mutations in ASD individuals was primarily driven by LoF-intolerant genes, but not LoF-tolerant genes.

We reasoned that ASD risk genes show expression patterns that are clustered in specific brain regions and developmental stages critical to disease development, and that high resolution spatiotemporal gene expression patterns in human brain can help distinguish genes that cause disease from those that do not. In addition, because ASD genes affected by *de novo* mutations are sensitive to mutational changes, we reasoned that gene-level constraint metrics can further differentiate ASD genes from normal ones. The objective of this study was to employ a machine learning-based approach to predict ASD risk genes using human brain spatiotemporal gene expression signatures, gene-level constraint metrics, and other gene variation features. We compared the performance of our method with five other state-of-the-art scoring systems for ranking ASD candidate genes, and evaluated the risk genes from our prediction model using independent sets of risk genes and differential gene expression (DGE) evidence. Gene Ontology (GO) enrichment analysis was also performed to understand the biology underlying ASD risk genes.

MATERIALS AND METHODS

Gene Set

To train the gene prediction model, we used labeled genes curated by Duda et al. (2018) as described in detail elsewhere. Briefly, the labeled genes contained 143 true positive genes and 1,145 true negative ones. The true positives came from the high confidence genes in the Simons Foundation Autism Research Initiative (SFARI) resource[1] (Category 1, Category 2, and syndromic genes) and the 65 reported genes in Sanders et al. (2015). The true negative genes were selected from the non-ASD gene list created by Krishnan et al. (2016), which were genes associated with non-mental health diseases, as annotated in OMIM. Among these genes we focused on those that had both gene expression data from the BrainSpan atlas and gene-level constraint metrics available, so that our final training gene set included 121 true positive genes and 963 true negatives.

Prediction Feature Sets

The feature sets in our prediction task included spatiotemporal gene expression patterns in human brain, network features, gene-level constraint metrics, and other gene variation features. **Supplementary Table S1** provides a summary of all features. We provide details below for each feature set.

Spatiotemporal Gene Expression

We downloaded RNA-Seq data (version 10), summarized to Gencode v20 gene-level reads per kilobase per million mapped reads (RPKM) values, from the BrainSpan website[2]. Detailed information on tissue processing, experimental and bioinformatics procedures related to the RNA-Seq data is

[1] https://www.sfari.org/resource/sfari-gene/

[2] http://www.brainspan.org/

available at the BrainSpan website. The BrainSpan dataset includes 524 gene-level expression features for each gene across 13 developmental stages in 31 brain regions from 524 brain samples spanning a variety of developmental stages and brain regions. Gene expression values were log-transformed ($\log_2$ [RPKM + 1]) and were used to predict autism genes.

To capture the functional relationships among genes, we built a weighted network for genes with both gene co-expression and PPI evidence from InWeb (Rossin et al., 2011). Specifically, the co-expression level between a gene pair was assessed by the Fischer z-transformed Pearson correlation between their spatiotemporal gene expression values. The genes with PPIs were connected and their edges were weighted by their co-expression levels. We extracted a set of network features that characterized the network topologies using *igraph* package in R. Specifically, we measured the node centralities using node degrees, clones centralities, betweenness centralities, Bonacich power centralities, eigenvector centralities, and alpha centralities (Bonacich, 1987). We captured the modules in functional relationship networks using the principle component decomposition and K-core decomposition (Batagelj and Zaversnik, 2003). The loading of the 1st principle component, hub score and coreness were obtained for each node. The importance of each node was further measured using the PageRank algorithm (Brin and Page, 1998), which counts the number and weight of links to each node. In total, 10 network features were extracted from the weighted gene network and were used for autism risk gene prediction. For genes appeared in BrainSpan but not in PPI network, we imputed their network features using the k-Nearest Neighbor algorithm.

Gene-Level Constraint Metrics and Other Gene Variation Features

We used gene-level constraint metrics developed from the exome data of more than 60,000 individuals from the ExAC to quantify the sensitivity of genes to variations (25). We considered six gene-level constraint metrics, including Z scores for synonymous (syn_z), missense (mis_z), and LoF variants (lof_z), the pLI, the probability of being intolerant of homozygous but not heterozygous LoF variants (pRec), and the probability of being tolerant of both heterozygous and homozygous LoF variants (pNull). A higher Z score or pLI indicates that the gene is more intolerant of variation (more constrained). We also included 10 general gene features, including the number of coding base pairs (bp), probabilities of mutations across the transcript for synonymous (mu_syn), missense (mu_mis), and LoF variants (mu_lof), number of rare variants (n_syn, n_mis, n_lof), and depth adjusted number of expected rare variants (exp_syn, exp_mis, exp_lof). Gene-level constraint metrics and general gene features were downloaded from the ExAC website[3]. Wilcoxon rank sum test was used to compare the group differences in above features between known ASD risk and non-risk genes.

[3] ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3.1/functional_gene_constraint/fordist_cleaned_exac_r03_march16_z_pli_rec_null_data.txt

Autism Risk Gene Prediction

We used machine learning methods to predict autism risk genes from their spatiotemporal expression signatures, network topology features, gene-level constraint metrics, and other general gene features. We applied four machine learning methods ranging from ones that are regression based [logistic regression and support vector machines (SVM) with Gaussian kernel] to others that are tree based (random forest and gradient boosted trees). The gradient boosted trees model ensembles a set of trees for prediction bias reduction and was trained in the XGBoost package (Chen et al., 2015). The optimal tuning parameters in each model were selected by a nested grid-search, and model performances were evaluated by five-fold cross validation (CV) on training data. The prediction accuracy was measured by the area under the receiver-operator curve (AUC-ROC) on the hold-out set for each fold of the CV. As the training data is unbalanced with small number of autism risk genes, we further considered the area under precision-recall curve (AUC-PRC) to measure the prediction accuracy.

Based on the average prediction accuracy over five folds, the gradient boosted trees model (BTree) was selected as the optimal algorithm. The final prediction model was built by applying the gradient boosted trees algorithm (with optimally tuned parameters) on all training genes and stored to predict over 17,000 unlabeled genes. For each labeled gene, the risk score was computed by prediction model that left the gene in the hold-out set in each CV.

Autism Risk Gene Validation Using Differential Gene Expression Evidence

Based on our gene ranking system, we classified genes into risk and non-risk genes using a threshold of risk score of 0.22 (resulting in 1,109 predicted ASD genes). We chose the risk score threshold of 0.22 because it gave the highest prediction accuracy (F_1 score = 0.59) on training data. Genes with a risk score higher than the threshold were predicted as risk genes and the remaining genes were predicted as non-risk genes. We validated the classification performance by examining whether our predicted risk genes show DGE evidence for ASD. Specifically, we obtained DGE summary statistics (beta and *p*-values) for ASD from RNA-Seq datasets for four major cortical lobes (frontal, temporal, parietal, and occipital) and their average from **Supplementary Table S1** of a previous study (Gandal et al., 2018), as well as the summary statistics for a non-psychiatric disorder inflammatory bowel disease (IBD) and two psychiatric disorders (bipolar disorder and schizophrenia) that we employed as negative controls from the same study (Gandal et al., 2018). The DGE summary statistics for IBD was derived using a linear mixed-effect model from meta-analysis of two published gene-expression microarray studies. The DGE summary statistics for ASD, bipolar disorder and schizophrenia were calculated using *limma* (Ritchie et al., 2015) with empiric Bayes moderated t-statistics from RNA-Seq analyses of post-mortem brain samples. The details for each datasets and DGE analyses were provided in the original study (Gandal et al., 2018). We used simulation-based approach to estimate the enrichment

statistics of predicted risk genes in DGE evidence. We first generated a background distribution from 100,000 random gene sets, while matching for gene size found in predicted risk genes. The enrichment fold was estimated by the ratio of the observed number of risk genes with DGE evidence ($p < 0.05$) to the average number of that from random gene sets. The *p*-value for enrichment was then the proportion of random gene sets with the same or a greater number of genes with DGE evidence, as compared to the number found for predicted risk genes. To investigate whether the enrichment of DGE evidence was specific to ASD, we also performed the same enrichment analysis for IBD, bipolar disorder, and schizophrenia.

Autism Risk Gene Validation in Independent Sequencing Studies

We further evaluated our gene ranking system utilizing genes targeted by *de novo* LoF mutations from two studies, including one that performed whole exome sequencing of 2,517 families in the Simons Simplex Collection (SSC) cohort (Iossifov et al., 2014) and another that performed whole genome sequencing of the MSSNG cohort (Ryan et al., 2017). To get independent lists of genes for validation, we excluded candidate genes from the two validation cohorts that overlapped the true positive genes in the training sample. For the SSC cohort, after excluding genes not included in BrainSpan, we compiled a list of 346 singleton LoF *de novo* mutations in probands, and 170 LoF *de novo* mutations in the unaffected siblings as negative controls. From the study of the MSSNG cohort, we created a list of 212 *de novo* LoF mutations in probands, 58 statistically significant *de novo* LoF or missense mutations, and 18 statistically significant *de novo* LoF or missense mutations that were not previously reported. For each of the five gene lists, we tested whether a larger proportion of genes were observed in the first decile of our gene ranking system than expected using a binomial test. The expected proportion (0.166) was determined using the percentage of genes with synonymous *de novo* mutations in the unaffected siblings of the SSC cohort.

Comparison With Other Ranking Systems

We compared our predictions with five autism gene prediction scores, including the ExAC score (pLI) (Lek et al., 2016), Iossifov probability score (Iossifov et al., 2015), Krishnan probability score (Krishnan et al., 2016), Zhang D score (Zhang and Shen, 2017), and Duda score (Duda et al., 2018). The former two (Iossifov et al., 2015; Lek et al., 2016) were based on measures of gene intolerance to disruptive variations, and the later three (Krishnan et al., 2016; Zhang and Shen, 2017; Duda et al., 2018) were based on machine learning methods that utilize brain-specific network features or cell-type specific gene expression signatures from mouse. Different gene scoring systems were compared in terms of ranking 173 curated candidate genes, including 130 genes with suggestive evidence from the SFARI Gene database (Category 3) and 43 recurrent *de novo* LoF genes discovered in recent studies (Wang et al., 2016; Li et al., 2017; Ryan et al., 2017; Stessman et al., 2017). We compared the overall ranking of candidate risk genes for different gene scoring systems, with a higher ranking (smaller number) indicating a greater likelihood of being ASD risk genes. We also compared the enrichment of candidate genes in the first decile of different gene scoring systems.

Gene Ontology Enrichment Analysis

We performed GO enrichment analysis to examine whether predicted risk genes were clustered into specific biological processes. Fisher's exact test was used to test the enrichment of risk genes in GO terms compared to non-risk genes. GO terms were chosen from the GO ontology of biological processes in MSigDB (v5.2) (Subramanian et al., 2005). To facilitate interpretation of the results, we included 2,758 GO terms that overlapped at least 20, but not more than 2,000 genes with our tested genes. Bonferroni correction was applied for multiple testing correction. Because GO terms were often highly overlapping in genes, we used hierarchical clustering to group significant gene sets into clusters based on similarity of their gene profiles (Chen et al., 2014). We first defined a gene overlapping matrix by counting the number of overlapping genes for each pair of gene sets. The Pearson correlation coefficient *R* was then calculated for each pair of gene sets based on their overlap profiles. The distance matrix for hierarchical clustering was then 1 - *R*. Hierarchical clustering was performed using the "ward" method implemented in the R function "hclust." The dendrogram and heatmap were plotted using the R function "heatmap.2."

RESULTS

An Overview of Study

An overview of study is provided in **Figure 1**. The basic premise is that ASD risk genes tend to show distinguishing features, including spatial-temporal gene expression patterns in human brain, gene network features, and gene-level constraint metrics. We reason that machine learning models utilizing those features can differentiate ASD genes from normal ones. To evaluate the performance of our prediction model, we examined if predicted ASD genes were enriched for DGE evidence and independent sets of ASD risk genes. We further performed GO enrichment analysis to understand the biology of predicted ASD genes.

Genome-Wide Prediction of Autism Risk Genes

We visualized gene expression patterns for 1,084 training genes across various regions and developmental stages of human brain (**Supplementary Figure S1**). There was a trend for known autism risk genes (left gene panel, red rows) to have higher expression levels than non-risk genes (left gene panel, blue rows). We further tested expression level differences between known risk and non-risk genes for each specific brain region and developmental stage (**Supplementary Figure S2**). The known autism risk genes showed significantly higher expression levels on average than non-risk genes for all tested brain regions and developmental stages ($p < 0.05$). Of note, the difference was

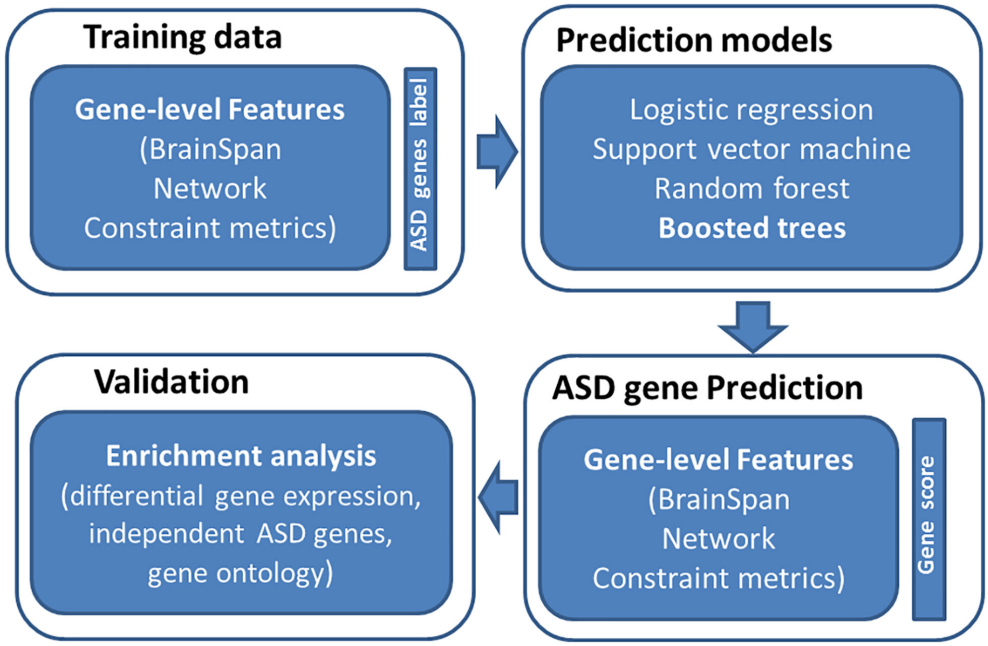

FIGURE 1 | Overview of study design. We first collected training data of labeled genes (121 true positive genes and 963 true negatives) and their associated features (spatiotemporal gene expression values from BrainSpan, network features, and gene-level constraint metrics). We then applied four machine learning algorithms to predict ASD genes, including logistic regression, support vector machine, random forest, and boosted trees. The boosted trees achieved the best prediction performance and was employed to predict ASD risk genes across the genome. We further evaluated predicted risk genes through enrichment analyses.

stronger for early to middle prenatal stages, ranging from 12 to 21 postconceptional weeks (pcw).

We compared known autism risk and non-risk genes in their sensitivity to mutational changes and other gene variation features. As shown in **Supplementary Figure S3**, compared to non-risk genes, autism risk genes were more intolerant of missense (mis_z, $p = 7 \times 10^{-16}$) and LoF mutations (lof_z, $p = 2 \times 10^{-23}$; pLI, $p = 2 \times 10^{-20}$), were less likely intolerant of homozygous, but not heterozygous LoF variants (pRec, $p = 5 \times 10^{-21}$), and had a lower probability of being tolerant of both heterozygous and homozygous LoF variants (pNull, $p = 3 \times 10^{-24}$). Autism risk genes had longer coding base pairs ($p = 4 \times 10^{-29}$), a higher probability of mutation across the transcript (mu_syn, $p = 1 \times 10^{-16}$; mu_mis, $p = 2 \times 10^{-18}$; mu_lof, $p = 4 \times 10^{-19}$), and a larger number of rare synonymous or missense variants (n_syn, $p = 4 \times 10^{-16}$; n_mis, $p = 1 \times 10^{-6}$), but less number of LoF variants (n_lof, $p = 3 \times 10^{-4}$).

We compared the prediction accuracy of four machine learning algorithms across five-fold CV. The gradient boosted trees (BTree) model achieved the best prediction accuracy for autism risk genes with AUC-ROC value of 0.86 and AUC-PRC value of 0.55 (**Figure 2**). The effects of different features on the boosted trees model were further explored by comparing the prediction accuracy under different feature sets (**Supplementary Figure S4**). We found that using the spatiotemporal gene expression features alone achieved an AUC-ROC (AUC-PRC) greater than 0.8 (0.4), and that the prediction accuracy was further improved by including either gene network features or gene-level constraint metrics, with the highest accuracy observed when all feature sets were included. We further evaluated the importance of individual features in the optimal BTree model. The feature importance was quantified as the average gain, i.e., improvement in node purity, of the feature when it was used in trees. **Supplementary Figure S5** illustrates the top 30 important features, including 28 spatiotemporal expression features and two gene-level constraint metrics (pLI and pNull). It was notable that pLI was the most predictive feature among all features used.

Autism Risk Gene Validation Using Differential Gene Expression Evidence

We predicted 1,109 risk genes using our gene ranking system under the threshold of risk score > 0.22, which generates the highest prediction accuracy measured by F_1 score on training data. We then examined whether those predicted risk genes were enriched for DGE evidence for ASD. We found that the predicted risk genes tended to be down-expressed in ASD brains, especially in frontal (fold = 1.7, $p < 1.0 \times 10^{-5}$) and parietal cortex (fold = 1.7, $p < 1.0 \times 10^{-5}$) (**Figure 3**). We did not see any significant enrichment of DGE evidence for IBD, bipolar disorder and schizophrenia, suggesting that the enriched DGE in our predicted genes was specific to ASD.

Autism Risk Gene Validation in Sequencing Studies

We further evaluated our gene ranking system using two sequencing studies (**Figure 4**). For the risk genes identified from the SSC cohort, our top decile genes were significantly enriched

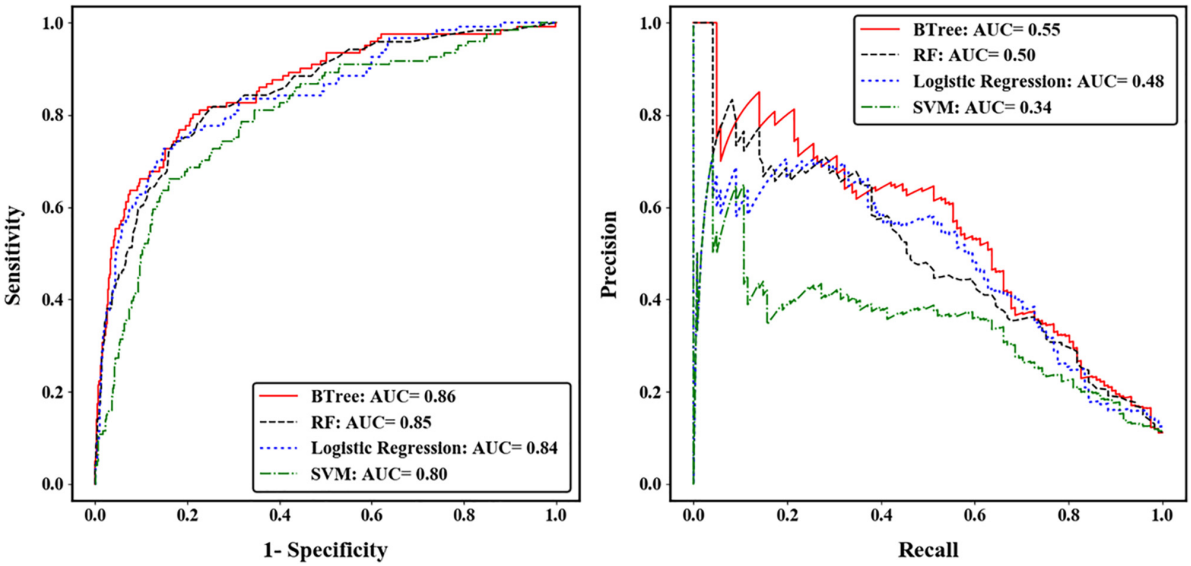

FIGURE 2 | Performance of four machine learning algorithms across five-fold cross validation. The left was measured by the area under receiver operating characteristic curve (ROC), and the right was measured by the area under precision-recall curve (PRC).

with *de novo* LoF mutations in probands. Specifically, genes in the first decile of our ranking system included 32% (88 of 273, $p = 6.8 \times 10^{-9}$) of *de novo* LOF mutations in probands. In contrast, we did not observe significant enrichment of genes with *de novo* LOF mutations in the unaffected siblings ($p = 0.65$). Similarly, for risk genes identified from the MSSNG cohort, we found significant enrichment for all three gene lists, including the *de novo* LOF mutations in probands (29%, $p = 2 \times 10^{-4}$), the 25 genes that reached genome-wide significance (72%, $p = 4.4 \times 10^{-9}$), and the 18 novel genes (67%, $p = 6.6 \times 10^{-6}$).

Comparison With Other Ranking Systems

We compared the performance of our ranking system (BTree) with five other gene scoring systems in their ability to rank curated candidate genes. When we examined the rank of an independent set of 173 autism candidate genes, our method outperformed other methods, because our method had the smallest median ranking (indicating the greatest likelihood of the set containing autism risk genes) (**Supplementary Figure S6**). We further compared the enrichment of 173 candidate genes in the first decile of each gene ranking system (**Supplementary Figure S7**). We observed the highest proportion of candidate genes in the first decile of our ranking system (52%), which was higher than the Duda score (40%), ExAC score (44%), Iossifov probability score (23%), Krishnan probability score (38%), and Zhang D score (30%). The superior performance of our method might be attributable to the human brain spatiotemporal gene expression features that were not considered in other methods.

Gene Ontology Enrichment Analysis

We conducted GO enrichment analysis to examine whether predicted 1,109 risk genes (score > 0.22) were clustered into specific biological processes. The full results of this analysis are shown in **Supplementary Table S2**. There were 179 GO terms that remained significant after Bonferroni correction ($p_{\text{corrected}} < 0.05$). Significant GO terms were grouped into five major clusters using hierarchical clustering (**Supplementary Figure S8**). These clusters included GO terms related to neuronal signaling (orange), neurogenesis (blue and black), chromatin remodeling (green), and transcriptional regulation (red). **Table 1** shows details for the top 10 enriched GO terms in enrichment fold that were particularly interesting, as they included GO terms involved in ionotropic glutamate receptor signaling, motor neuron axon guidance, and regulation of histone methylation.

DISCUSSION

A number of methods have been developed for inferring ASD risk genes. Although they employ differing computational methodologies, most methods were built upon the concept of guilt-by-association, using the assumption that risk genes are functionally related. Theoretically, ASD risk genes should exert their effects at specific developmental stages in specific brain tissues or cell types that are critical to disease development. However, most existing methods have not considered the spatial and temporal patterns of gene relationships during brain development. In addition, gene-level constraint metrics, such as loss of function intolerance, have been used to prioritize ASD candidate genes, but no studies have quantitatively examined their potential for predicting ASD genes. Employing a supervised machine learning algorithm, we have shown that a combination of human brain spatiotemporal gene expression patterns and the gene-level constraint metric features predict ASD risk genes. We further demonstrated the validity of our method through

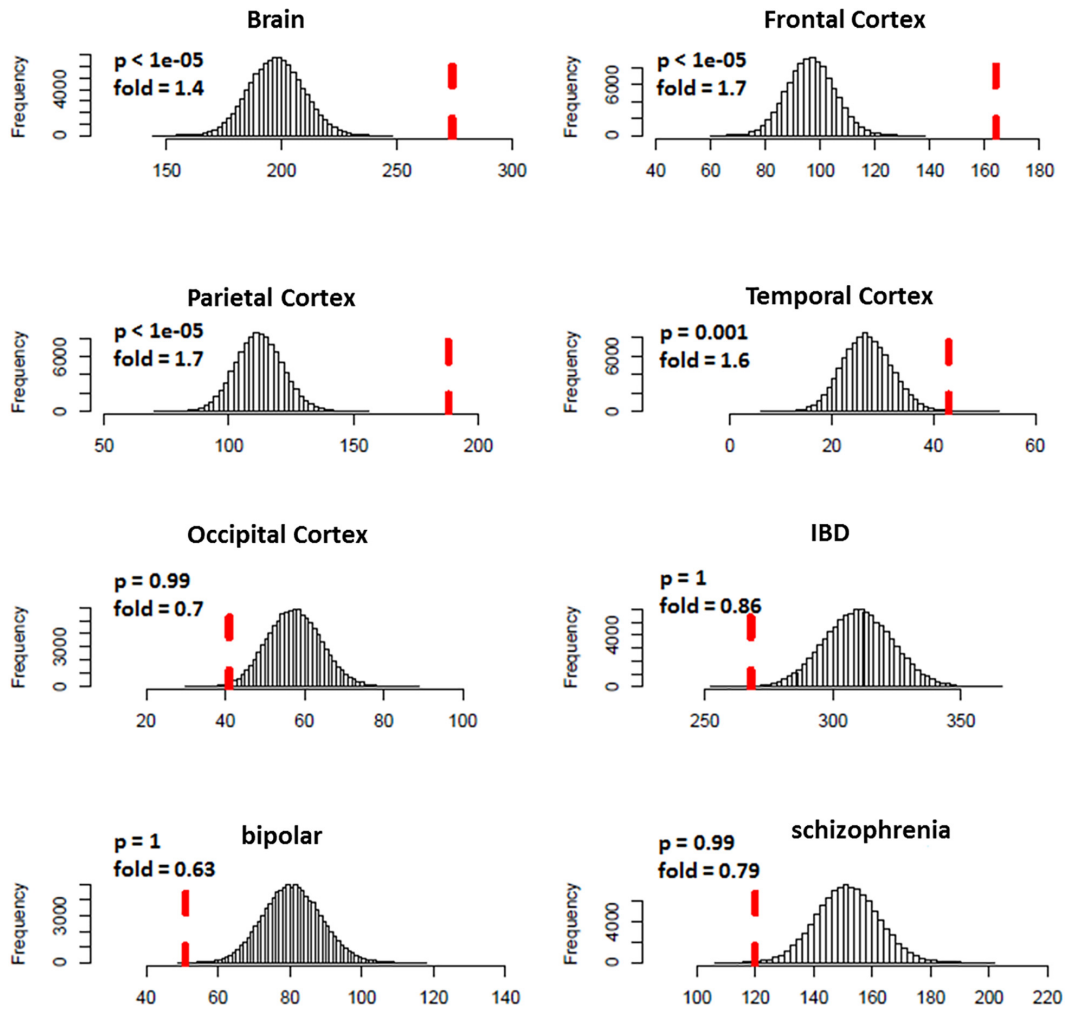

FIGURE 3 | Enrichment analysis of differential expression evidence for predicted ASD risk genes. The histogram shows the distribution for the number of genes with DGE evidence ($p < 0.05$) from random gene sets. The vertical dotted red line indicates the number of genes with DGE evidence from predicted ASD risk genes. Predicted risk genes tended to be down-expressed in brains of ASD, but not for disorders of negative control (IBD, bipolar disorder, and schizophrenia).

validations using DGE evidence and independent sets of risk genes. We have further shown the superior performance of our ranking system over several other state-of-the-art ranking systems in ranking curated candidate genes.

We explored the potential role of the top ranked genes in ASD risk. The gene *NBEA*, which encodes neurobeachin that is a brain-specific kinase-anchoring protein implicated synaptic structure and function, was assigned the highest probability for conferring ASD risk (score = 0.97). Indeed, mutations in *NBEA* have been identified in ASD (Castermans et al., 2003; Wise et al., 2015) and neurodevelopmental disorders (Mulhern et al., 2018). Another notable gene in our top list was *HERC1* (ranked third, score = 0.94), which encodes a protein that is a probable E3 ubiquitin-protein ligase. Multiple lines of evidence indicate a role for *HERC1* in ASD: (1) it was reported that *HERC1* mutations caused intellectual disability and facial dysmorphism in two Colombian siblings (Ortega-Recalde et al., 2015); (2) A nonsense variant in *HERC1* was associated with intellectual disability, megalencephaly, thick corpus callosum and cerebellar atrophy (Nguyen et al., 2016); (3) importantly, mutations in *HERC1* were reported to be associated with ASD in an exome sequencing study (Hashimoto et al., 2016). Our ranking system also successfully predicted another two ASD candidate genes *TCF20* (ranked 26th, score = 0.87) and *FBXO11* (ranked 19th, score = 0.88). Intriguingly, *TCF20* was one of the highest ranking candidate autism risk genes (category 2) according to the most recent version of the SFARI Gene resource. Mutations in *TCF20* were also implicated in Phelan-McDermid syndrome (Upadia et al., 2018), developmental disorders (Deciphering Developmental Disorders Study, 2017), and schizophrenia (Smeland et al., 2017). *FBXO11* was prioritized as a strong ASD candidate gene (Ji et al., 2016), and was recently reported to be associated with a variable neurodevelopmental disorder (Gregor et al., 2018).

Our ranking system also highlighted some potential novel candidate genes that may deserve further investigation. Four genes, *ZYG11B*, *HECTD1*, *CAND1*, and *MYCBP2*,

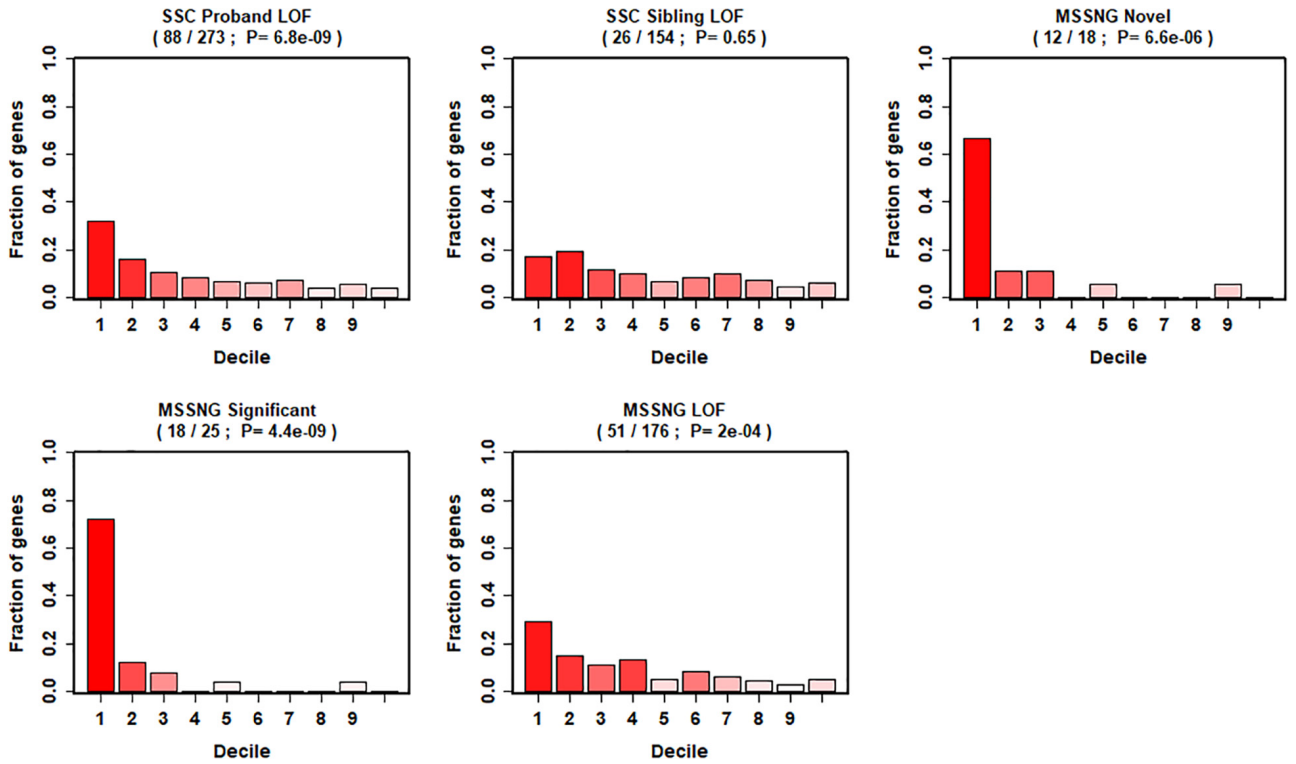

FIGURE 4 | Decile enrichment of *de novo* mutations from two independent cohorts in our gene ranking system.

TABLE 1 | Top ten enriched GO terms in predicted ASD risk genes.

GO terms	OR	95%_CI_L	95%_CI_U	*p*	p_{adj}
GO_CENTRAL_NERVOUS_SYSTEM_PROJECTION_NEURON_AXONOGENESIS	27.0	10.0	80.1	1.7E-11	4.7E-08
GO_IONOTROPIC_GLUTAMATE_RECEPTOR_SIGNALING_PATHWAY	17.5	6.9	45.3	1.9E-09	5.1E-06
GO_CENTRAL_NERVOUS_SYSTEM_NEURON_AXONOGENESIS	17.2	7.1	42.5	4.3E-10	1.2E-06
GO_DENDRITE_MORPHOGENESIS	13.0	6.2	27.0	6.5E-11	1.8E-07
GO_MOTOR_NEURON_AXON_GUIDANCE	11.2	4.4	27.6	6.7E-07	0.0018
GO_POSITIVE_REGULATION_OF_HISTONE_METHYLATION	11.1	5.0	24.4	1.4E-08	4.0E-05
GO_GLUTAMATE_RECEPTOR_SIGNALING_PATHWAY	11.1	5.4	22.4	3.1E-10	8.7E-07
GO_EXCITATORY_POSTSYNAPTIC_POTENTIAL	10.4	4.1	25.2	1.1E-06	0.003
GO_REGULATION_OF_HISTONE_H3_K4_METHYLATION	10.4	4.1	25.2	1.1E-06	0.003
GO_MODULATION_OF_EXCITATORY_POSTSYNAPTIC_POTENTIAL	9.3	3.6	23.2	7.1E-06	0.019

OR: odds ratio; 95%_CI_L: OR 95% confidence interval lower bound; 95%_CI_U: OR 95% confidence interval upper bound.

ranked second, fourth, seventh and tenth, are all involved in protein ubiquitination, which has been implicated in neuronal function and brain disorders, including ASD (Mabb and Ehlers, 2010). To our knowledge, direct genetic links between these genes with ASD have not been found. Of note, *CAND1* encodes an essential regulator of Cullin-RING ubiquitin ligases that play a critical role in ubiquitination and protein degradation (Zheng et al., 2002); *MYCBP2* encodes an E3 ubiquitin-protein ligase that plays a role in axon guidance and synapse formation in the developing nervous system. We have provided the whole list of ranked genes with their probability scores in **Supplementary Table S3**. Researchers can further explore the top-ranked genes or genes of their own interest.

Our study not only provides hundreds of new ASD candidate genes with evidence for involvement in ASD, but also shows that the predicted risk genes are biologically meaningful and are clustered around biological processes relevant to ASD. GO enrichment analysis demonstrated that the predicted risk genes were enriched in GO terms related to neuronal signaling, neurogenesis, chromatin remodeling, and histone modification, all of which are important biological processes implicated in ASD. In addition, among our top 10 ranked genes, we found that five were related to the protein ubiquitination pathway (*HERC1*, *CAND1*, *ZYG11B*, *HECTD1*, and *MYCBP2*), which is consistent with the significant enrichment of protein ubiquitination process in our GO

enrichment analysis (GO_PROTEIN_UBIQUITINATION, OR = 2.3, $p_{corrected} = 1.9 \times 10^{-6}$), supporting the merging role of ubiquitin signaling in ASD (Mabb and Ehlers, 2010; Cheon et al., 2018). Our analyses also highlighted other biological mechanisms that may underlie ASD. For example, there is evidence for roles of RNA alternative splicing (Parikshak et al., 2016) in ASD, which was represented in our top enriched GO terms (GO_RNA_SPLICING, OR = 3.5, $p_{corrected} = 8.0 \times 10^{-12}$).

Our study also sheds light on when and where ASD genes may exert their effects during brain development. Of the 28 gene expression features from the top 30 important features in the BTree model, 15 referred to brain regions in the early to mid-prenatal stage (≤24 pcw), reinforcing the important role of early prenatal development in ASD. The involved brain regions include the posteroventral (inferior) parietal cortex (IPC), primary motor cortex (area M1, area 4) (M1C), posterior (caudal) superior temporal cortex (area 22c) (STC), inferolateral temporal cortex (area TEv, area 20) (ITC), medial prefrontal cortex (MFC), cerebellum (CB), dorsolateral prefrontal cortex (DFC), and ventrolateral prefrontal cortex (VFC).

This work should be viewed in light of several limitations. First, our method was trained on genes implicated in ASD by *de novo* mutations. It was not clear how our gene ranking system was relevant to genes affected by other type of variants. Second, our gene ranking system was validated using enrichment analyses of DGE evidence in ASD brain and independent lists of candidate genes. However, a more solid validation should be a replication study for top ranked genes in independent samples through sequencing, but it is beyond the scope of current study. Third, given the strong evidence of clinical and genetic overlap between ASD and other types of neurodevelopmental disorders (Mullin et al., 2013; Srivastava and Schwartz, 2014), further work is needed to investigate whether our gene ranking system is specific to ASD.

In summary, our study has demonstrated that human brain spatiotemporal gene expression patterns and gene-level constraint metrics predict ASD risk genes. Our gene ranking system provides a useful resource for prioritizing ASD candidate genes.

DATA AVAILABILITY STATEMENT

The datasets generated from this study can be found in online **Supplementary Material**.

AUTHOR CONTRIBUTIONS

SH and AR designed the study and wrote the manuscript. YL performed the data analyses and wrote the manuscript. SA performed the data analyses. AR and JP supported the manuscript preparation and project planning. All authors contributed to the article and approved the submitted version.

FUNDING

This study was partially supported by National Institutes of Health grants R01 AA022994 and AA024486 (to SH). This manuscript has been released as a Pre-Print at bioRxiv (Ying et al., 2018).

ACKNOWLEDGMENTS

The authors would like to thank Dr. Mingyao Ying for fruitful discussions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.500064/full#supplementary-material

FIGURE S1 | Heatmap view of spatiotemporal gene expression in human brain. Each cell in the heat map corresponds to the expression level of a gene (row) in a specific brain region and development stage (column). The ASD risk and non-risk genes are denoted by red and blue rows, respectively. Brain regions are represented by the 31 colors in Color Key of Brain Regions. The ASD risk genes tend to be expressed in a higher level compared to non-risk genes across developmental stages and brain regions. The intensity of the color in each cell represents the log2-transformed expression level. Full names of each brain region: CBC, cerebellar cortex; CB, cerebellum; VFC, ventrolateral prefrontal cortex; M1C, primary motor cortex (area M1, area 4); M1C-S1C, primary motor-sensory cortex (samples); IPC, posteroventral (inferior) parietal cortex; PCx, parietal neocortex; HIP, hippocampus (hippocampal formation); DTH, dorsal thalamus; TCx, temporal neocortex; S1C, primary somatosensory cortex (area S1, areas 3,1,2); MD, mediodorsal nucleus of thalamus; A1C, primary auditory cortex (core); AMY, amygdaloid complex; STR, striatum; URL, upper (rostral) rhombic lip; OFC, orbital frontal cortex; Ocx, occipital neocortex; MGE, medial ganglionic eminence; CGE, caudal ganglionic eminence; LGE, lateral ganglionic eminence; STC, posterior (caudal) superior temporal cortex (area 22c); MFC, anterior (rostral) cingulate (medial prefrontal) cortex; V1C, primary visual cortex (striate cortex, area V1/17); ITC, inferolateral temporal cortex (area TEv, area 20); DFC, dorsolateral prefrontal cortex.

FIGURE S2 | Gene expression difference between ASD risk and non-risk genes in the spatiotemporal development of human brain. Each cell in the heat map represents the expression level difference (*t*-test) in a specific brain region (column) and development stage (row). The intensity of color represents the log-transformed *p*-value from a *t*-test. The brain regions and stages without gene expression data are marked as black.

FIGURE S3 | Boxplot of gene-level constraint metrics and other gene variation features for true positive (TP) and true negative (TN) genes.

FIGURE S4 | Boxplot of AUCs under different feature sets for BTree model. The left was measured by the area under receiver operating curve (ROC), and the right was measured by the area under precision-recall curve (PRC).

FIGURE S5 | Top 30 important features in the BTree model.

FIGURE S6 | Comparison of our gene ranking system (BTree) with five other gene ranking systems on overall rankings of 173 independent candidate genes.

FIGURE S7 | Decile enrichment of 173 independent candidate genes for each gene ranking system. The number on the top of each panel represents the number of 173 curated candidate genes appeared in the first decile of each ranking system.

FIGURE S8 | Hierarchical clustering of significant GO terms.

TABLE S1 | Feature sets included in prediction model.

TABLE S2 | GO enrichment analysis for predicted ASD risk genes.

TABLE S3 | Gene risk score predicted from Boosted tree model.

REFERENCES

Asif, M., Martiniano, H., Vicente, A. M., and Couto, F. M. (2018). Identifying disease genes using machine learning and gene functional similarities, assessed through Gene Ontology. *PLoS One* 13:e0208626. doi: 10.1371/journal.pone.0208626

Batagelj, V., and Zaversnik, M. (2003). An O(m) algorithm for cores decomposition of networks. *Adv. Data Anal. Classif.* 5, 129–145.

Bonacich, P. (1987). Power and centrality: a family of measures. *Am. J. Sociol.* 92, 1170–1182.

Brin, S., and Page, L. (1998). "The anatomy of a large-scale hypertextual web search engine," in *Proceedings of the 7th World-Wide Web Conference*, New York, NY.

Brueggeman, L., Koomar, T., and Michaelson, J. J. (2020). Forecasting risk gene discovery in autism with machine learning and genome-scale data. *Sci. Rep.* 10:4569.

Castermans, D., Wilquet, V., Parthoens, E., Huysmans, C., Steyaert, J., Swinnen, L., et al. (2003). The neurobeachin gene is disrupted by a translocation in a patient with idiopathic autism. *J. Med. Genet.* 40, 352–356. doi: 10.1136/jmg.40.5.352

Chen, T., He, T., Benesty, M., Khotilovich, V., and Tang, Y. (2015). *Xgboost: Extreme Gradient Boosting. R Package Version 0.4-2*.

Chen, Y. A., Tripathi, L. P., Dessailly, B. H., Nystrom-Persson, J., Ahmad, S., and Mizuguchi, K. (2014). Integrated pathway clusters with coherent biological themes for target prioritisation. *PLoS One* 9:e99030. doi: 10.1371/journal.pone.0099030

Cheon, S., Dean, M., and Chahrour, M. (2018). The ubiquitin proteasome pathway in neuropsychiatric disorders. *Neurobiol. Learn. Mem.* 165:106791. doi: 10.1016/j.nlm.2018.01.012

Colvert, E., Tick, B., Mcewen, F., Stewart, C., Curran, S. R., Woodhouse, E., et al. (2015). Heritability of autism spectrum disorder in a UK population-based twin sample. *JAMA Psychiatry* 72, 415–423. doi: 10.1001/jamapsychiatry.2014.3028

Deciphering Developmental Disorders Study (2017). Prevalence and architecture of de novo mutations in developmental disorders. *Nature* 542, 433–438. doi: 10.1038/nature21062

Duda, M., Zhang, H., Li, H. D., Wall, D. P., Burmeister, M., and Guan, Y. (2018). Brain-specific functional relationship networks inform autism spectrum disorder gene prediction. *Transl. Psychiatry* 8:56.

Gandal, M. J., Haney, J. R., Parikshak, N. N., Leppa, V., Ramaswami, G., Hartl, C., et al. (2018). Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. *Science* 359, 693–697. doi: 10.1126/science.aad6469

Gandhi, T. K., Zhong, J., Mathivanan, S., Karthick, L., Chandrika, K. N., Mohan, S. S., et al. (2006). Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat. Genet.* 38, 285–293. doi: 10.1038/ng1747

Gilman, S. R., Iossifov, I., Levy, D., Ronemus, M., Wigler, M., and Vitkup, D. (2011). Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. *Neuron* 70, 898–907. doi: 10.1016/j.neuron.2011.05.021

Gök, M. (2018). A novel machine learning model to predict autism spectrum disorders risk gene. *Neural Comput. Appl.* 31, 6711–6717. doi: 10.1007/s00521-018-3502-5

Gregor, A., Sadleir, L. G., Asadollahi, R., Azzarello-Burri, S., Battaglia, A., Ousager, L. B., et al. (2018). De novo variants in the F-box protein FBXO11 in 20 individuals with a variable neurodevelopmental disorder. *Am. J. Hum. Genet.* 10, 305–316. doi: 10.1016/j.ajhg.2018.07.003

Hallmayer, J., Cleveland, S., Torres, A., Phillips, J., Cohen, B., Torigoe, T., et al. (2011). Genetic heritability and shared environmental factors among twin pairs with autism. *Arch. Gen. Psychiatry* 68, 1095–1102. doi: 10.1001/archgenpsychiatry.2011.76

Hashimoto, R., Nakazawa, T., Tsurusaki, Y., Yasuda, Y., Nagayasu, K., Matsumura, K., et al. (2016). Whole-exome sequencing and neurite outgrowth analysis in autism spectrum disorder. *J. Hum. Genet.* 61, 199–206. doi: 10.1038/jhg.2015.141

Hormozdiari, F., Penn, O., Borenstein, E., and Eichler, E. E. (2015). The discovery of integrated gene networks for autism and related disorders. *Genome Res.* 25, 142–154. doi: 10.1101/gr.178855.114

Iossifov, I., Levy, D., Allen, J., Ye, K., Ronemus, M., Lee, Y. H., et al. (2015). Low load for disruptive mutations in autism genes and their biased transmission. *Proc. Natl. Acad. Sci. U.S.A.* 112, E5600–E5607.

Iossifov, I., O'roak, B. J., Sanders, S. J., Ronemus, M., Krumm, N., Levy, D., et al. (2014). The contribution of de novo coding mutations to autism spectrum disorder. *Nature* 515, 216–221.

Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., et al. (2012). De novo gene disruptions in children on the autistic spectrum. *Neuron* 74, 285–299.

Jeste, S. S., and Geschwind, D. H. (2014). Disentangling the heterogeneity of autism spectrum disorder through genetic findings. *Nat. Rev. Neurol.* 10, 74–81. doi: 10.1038/nrneurol.2013.278

Ji, X., Kember, R. L., Brown, C. D., and Bucan, M. (2016). Increased burden of deleterious variants in essential genes in autism spectrum disorder. *Proc. Natl. Acad. Sci. U.S.A.* 113, 15054–15059. doi: 10.1073/pnas.1613195113

Kosmicki, J. A., Samocha, K. E., Howrigan, D. P., Sanders, S. J., Slowikowski, K., Lek, M., et al. (2017). Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. *Nat. Genet.* 49, 504–510. doi: 10.1038/ng.3789

Krishnan, A., Zhang, R., Yao, V., Theesfeld, C. L., Wong, A. K., Tadych, A., et al. (2016). Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nat. Neurosci.* 19, 1454–1462. doi: 10.1038/nn.4353

Krumm, N., O'roak, B. J., Shendure, J., and Eichler, E. E. (2014). A de novo convergence of autism genetics and molecular neuroscience. *Trends Neurosci.* 37, 95–105. doi: 10.1016/j.tins.2013.11.005

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

Li, J., Shi, M., Ma, Z., Zhao, S., Euskirchen, G., Ziskin, J., et al. (2014). Integrated systems analysis reveals a molecular network underlying autism spectrum disorders. *Mol. Syst. Biol.* 10:774. doi: 10.15252/msb.20145487

Li, J., Wang, L., Guo, H., Shi, L., Zhang, K., Tang, M., et al. (2017). Targeted sequencing and functional analysis reveal brain-size-related genes and their networks in autism spectrum disorders. *Mol. Psychiatry* 22, 1282–1290. doi: 10.1038/mp.2017.140

Liu, L., Lei, J., and Roeder, K. (2015). Network assisted analysis to reveal the genetic basis of autism. *Ann. Appl. Stat.* 9, 1571–1600. doi: 10.1214/15-aoas844

Mabb, A. M., and Ehlers, M. D. (2010). Ubiquitination in postsynaptic function and plasticity. *Annu. Rev. Cell. Dev. Biol.* 26, 179–210. doi: 10.1146/annurev-cellbio-100109-104129

Mulhern, M. S., Stumpel, C., Stong, N., Brunner, H. G., Bier, L., Lippa, N., et al. (2018). NBEA: developmental disease gene with early generalized epilepsy phenotypes. *Ann. Neurol.* 84, 788–795.

Mullin, A. P., Gokhale, A., Moreno-De-Luca, A., Sanyal, S., Waddington, J. L., and Faundez, V. (2013). Neurodevelopmental disorders: mechanisms and boundary definitions from genomes, interactomes and proteomes. *Transl. Psychiatry* 3:e329. doi: 10.1038/tp.2013.108

Nguyen, L. S., Schneider, T., Rio, M., Moutton, S., Siquier-Pernet, K., Verny, F., et al. (2016). A nonsense variant in HERC1 is associated with intellectual disability, megalencephaly, thick corpus callosum and cerebellar atrophy. *Eur. J. Hum. Genet.* 24, 455–458. doi: 10.1038/ejhg.2015.140

Ortega-Recalde, O., Beltran, O. I., Galvez, J. M., Palma-Montero, A., Restrepo, C. M., Mateus, H. E., et al. (2015). Biallelic HERC1 mutations in a syndromic form of overgrowth and intellectual disability. *Clin. Genet.* 88, e1–e3.

Parikshak, N. N., Swarup, V., Belgard, T. G., Irimia, M., Ramaswami, G., Gandal, M. J., et al. (2016). Genome-wide changes in lncRNA, splicing, and regional

gene expression patterns in autism. *Nature* 540, 423–427. doi: 10.1038/nature20612

Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S., and Goldstein, D. B. (2013). Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.* 9:e1003709. doi: 10.1371/journal.pgen.1003709

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

Ronemus, M., Iossifov, I., Levy, D., and Wigler, M. (2014). The role of de novo mutations in the genetics of autism spectrum disorders. *Nat. Rev. Genet.* 15, 133–141. doi: 10.1038/nrg3585

Rossin, E. J., Lage, K., Raychaudhuri, S., Xavier, R. J., Tatar, D., and Benita, Y. (2011). Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet.* 7:e1001273. doi: 10.1371/journal.pgen.1001273

Ryan, K. C. Y., Merico, D., Bookman, M., Thiruvahindrapuram, B., Patel, R. V., Whitney, J., et al. (2017). Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nat. Neurosci.* 20, 602–611.

Samocha, K. E., Robinson, E. B., Sanders, S. J., Stevens, C., Sabo, A., Mcgrath, L. M., et al. (2014). A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* 46, 944–950.

Sanders, S. J., He, X., Willsey, A. J., Ercan-Sencicek, A. G., Samocha, K. E., Cicek, A. E., et al. (2015). Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. *Neuron* 87, 1215–1233.

Sanders, S. J., Murtha, M. T., Gupta, A. R., Murdoch, J. D., Raubeson, M. J., Willsey, A. J., et al. (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* 485, 237–241.

Smeland, O. B., Frei, O., Kauppi, K., Hill, W. D., Li, W., Wang, Y., et al. (2017). Identification of genetic loci jointly influencing schizophrenia risk and the cognitive traits of verbal-numerical reasoning, reaction time, and general cognitive function. *JAMA Psychiatry* 74, 1065–1075. doi: 10.1001/jamapsychiatry.2017.1986

Srivastava, A. K., and Schwartz, C. E. (2014). Intellectual disability and autism spectrum disorders: causal genes and molecular mechanisms. *Neurosci. Biobehav. Rev.* 46(Pt 2), 161–174. doi: 10.1016/j.neubiorev.2014.02.015

Stessman, H. A., Xiong, B., Coe, B. P., Wang, T., Hoekzema, K., Fenckova, M., et al. (2017). Targeted sequencing identifies 91 neurodevelopmental-disorder risk genes with autism and developmental-disability biases. *Nat. Genet.* 49, 515–526. doi: 10.1038/ng.3792

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550. doi: 10.1073/pnas.0506580102

Turner, T. N., Coe, B. P., Dickel, D. E., Hoekzema, K., Nelson, B. J., Zody, M. C., et al. (2017). Genomic Patterns of de novo mutation in simplex autism. *Cell* 171, 710.e2–722.e2.

Upadia, J., Gonzales, P. R., Atkinson, T. P., Schroeder, H. W., Robin, N. H., Rudy, N. L., et al. (2018). A previously unrecognized 22q13.2 microdeletion syndrome that encompasses TCF20 and TNFRSF13C. *Am. J. Med. Genet. A* 176, 2791–2797. doi: 10.1002/ajmg.a.40492

Wang, T., Guo, H., Xiong, B., Stessman, H. A., Wu, H., Coe, B. P., et al. (2016). De novo genic mutations among a Chinese autism spectrum disorder cohort. *Nat. Commun.* 7:13316.

Willsey, A. J., Sanders, S. J., Li, M., Dong, S., Tebbenkamp, A. T., Muhle, R. A., et al. (2013). Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* 155, 997–1007. doi: 10.1016/j.cell.2013.10.020

Wise, A., Tenezaca, L., Fernandez, R. W., Schatoff, E., Flores, J., Ueda, A., et al. (2015). Drosophila mutants of the autism candidate gene neurobeachin (rugose) exhibit neuro-developmental disorders, aberrant synaptic properties, altered locomotion, and impaired adult social behavior and activity patterns. *J. Neurogenet.* 29, 135–143. doi: 10.3109/01677063.2015.1064916

Xu, J., and Li, Y. (2006). Discovering disease-genes by topological features in human protein-protein interaction network. *Bioinformatics* 22, 2800–2805. doi: 10.1093/bioinformatics/btl467

Ying, L., Anjali, M. R., James, P., and Shizhong, H. (2018). A machine learning approach to predicting autism risk genes: validation of known genes and discovery of new candidates. *bioRxiv* [Preprint]. doi: 10.1101/463547

Zhang, C., and Shen, Y. (2017). A cell type-specific expression signature predicts haploinsufficient autism-susceptibility genes. *Hum. Mutat.* 38, 204–215. doi: 10.1002/humu.23147

Zheng, J., Yang, X., Harrell, J. M., Ryzhikov, S., Shim, E. H., and Zhang, H. (2002). CAND1 binds to unneddylated CUL1 and regulates the formation of SCF ubiquitin E3 ligase complex. *Mol. Cell.* 10, 1519–1526. doi: 10.1016/s1097-2765(02)00784-0

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