Quantitative Biology 2018, 6(2): 113–128
https://doi.org/10.1007/s40484-018-0139-4

REVIEW

Towards precise reconstruction of gene regulatory networks by data integration

Zhi-Ping Liu*

Department of Biomedical Engineering, School of Control Science and Engineering, Shandong University, Jinan 250061, China
* Correspondence: zpliu@sdu.edu.cn

Received September 10, 2017; Revised October 30, 2017; Accepted November 25, 2017

***Background*: More and more high-throughput datasets are available from multiple levels of measuring gene regulations. The reverse engineering of gene regulatory networks from these data offers a valuable research paradigm to decipher regulatory mechanisms. So far, numerous methods have been developed for reconstructing gene regulatory networks.**
***Results*: In this paper, we provide a review of bioinformatics methods for inferring gene regulatory network from omics data. To achieve the precision reconstruction of gene regulatory networks, an intuitive alternative is to integrate these available resources in a rational framework. We also provide computational perspectives in the endeavors of inferring gene regulatory networks from heterogeneous data. We highlight the importance of multi-omics data integration with prior knowledge in gene regulatory network inferences.**
***Conclusions*: We provide computational perspectives of inferring gene regulatory networks from multiple omics data and present theoretical analyses of existing challenges and possible solutions. We emphasize on prior knowledge and data integration in network inferences owing to their abilities of identifying regulatory causality.**

Keywords: gene regulatory network; computational inference; data integration; bioinformatics

Author summary: In this paper, we summarize and comment recent progresses in the important bioinformatics field of gene regulatory network inference from quantitative gene expression profiles, especially focus on the endeavors for improving the inference precision by integrating multiple resources. The paper will potentially facilitate scientists who are interested in reversely engineering gene regulatory network to quickly obtain an integrative overview and follow the start-of-the-art computational techniques.

INTRODUCTION

Due to the availability of biomedical big data, the research paradigm of biomedical science is undergoing unprecedented changes and challenges [1]. Gene regulations play central roles in transforming genotypic information to phenotypic performance [2]. Many biomolecules are involved in the biological processes of gene regulations, and their relationships are often modeled as gene regulatory networks [3]. Where the nodes are these players and the edges are their regulatory interactions. In healthy states, the dynamics underlying gene regulatory networks orchestrate a perfect concerto of various physiological processes in a cell, while they perform disorders to dysfunctions to diseases when the harmony in the network system has been broken [4]. Temporospatially rewiring regulations indicate the causality of phenotypic transitions and differences [5].

The high-throughput techniques such as ChIP-Seq provide direct recognitions of protein-DNA interactions by mapping specific binding sites [6]. The prior knowledge about gene regulations such as specific transcription factor (TF) and gene binding sites can be utilized to design and analyze the ChIP-Seq data [7]. But the production of antibodies for specific TFs and the sensitivities are still limited in these experiments. Meanwhile, there are more and more gene expression profiling datasets available for measuring the global transcriptomic status [8,9]. The epigenetic data such as DNA methylation and RNA modification provide more resources for deciphering gene regulations [10,11]. How to reconstruct gene regulatory network with these heterogeneous data have attracted dense attentions in bioinformatics research community since these transcriptomic data are available [3].

In this work, we offer a brief review with perspectives about these computational methods of reconstructing gene regulatory network from transcriptomic data. We firstly introduce the problem and the computational complexity of inferring gene regulatory network. Then, we review some available data resources and existing methods for network inference. We summarize the two types of errors in these methods individually. To achieve precision network reconstruction, these methods are formulated to remove the two types of errors in the inference, especially false positives. We highlight the data integration of prior knowledge and multiple-level omics for accurately reconstructing a comprehensive gene regulatory network, respectively.

INFERRING GENE REGULATORY NETWORKS

The problem of gene regulatory network reconstruction is a reverse engineering of inferring gene regulations from gene expressions, which are measured from the high-throughput techniques such as microarray [12] or RNA-Seq [13]. The transcriptomic concentrations measure gene expressions in parallel manners. The gene relationships are expected to be reconstructed from their expression profiles in the samples. The gene regulatory relationships will be built up from the data in the form of a network representing their causal interactions.

Usually, gene expression data is represented by matrix $\boldsymbol{G}$ as

$$\boldsymbol{G}=\begin{pmatrix} g_1 \\ \vdots \\ g_n \end{pmatrix}=\begin{pmatrix} a_{11} & \cdots & a_{i1} & \cdots & a_{p1} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{1j} & \cdots & a_{ij} & \cdots & a_{pj} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{1n} & \cdots & a_{in} & \cdots & a_{pn} \end{pmatrix},$$

where a_{ij} represents the gene expression value of the j-th gene ($1 \leqslant j \leqslant n$) in the i-th experiment ($1 \leqslant i \leqslant p$). Noted that i refers to the sample and j refers to the gene. The framework of gene regulatory network inference is shown in Figure 1. The task is to reversely induce the regulatory relationships as shown in Figure 1B from gene expression profiles (Figure 1A). For the paramount importance of gene regulations, numerous methods have been proposed to tackle gene regulatory network inference from data matrix $\boldsymbol{G}$.

Biologically, the experiments are often designed to access specific physiological conditions with limited number of samples and necessary biological replicates. For instance, a gene expression profiling experiment is designed to study the Huh7 cells after hepatitis C virus (HCV) infection. It contains three replicates at 6, 12, 18, 24, and 48 h post-infections respectively [14]. Totally, there are 18 samples of microarray with 3 controls before HCV infection. The gene chip platform contains about 25,000 genes. In this case of network inference, the aim is to reconstruct the gene regulatory network of these human cells in response to HCV infection. In these experiments, we can find that the gene expression profiles only refer to several snapshots of gene expression abundancy at several time points after viral infection. While our task is to figure out a global map of regulatory relationships in the thousands of genes. This is very similar to re-outline a 48-h movie only by 5 blurred pictures at 5 screenplays, at least for deducing the relationships among the 25,000

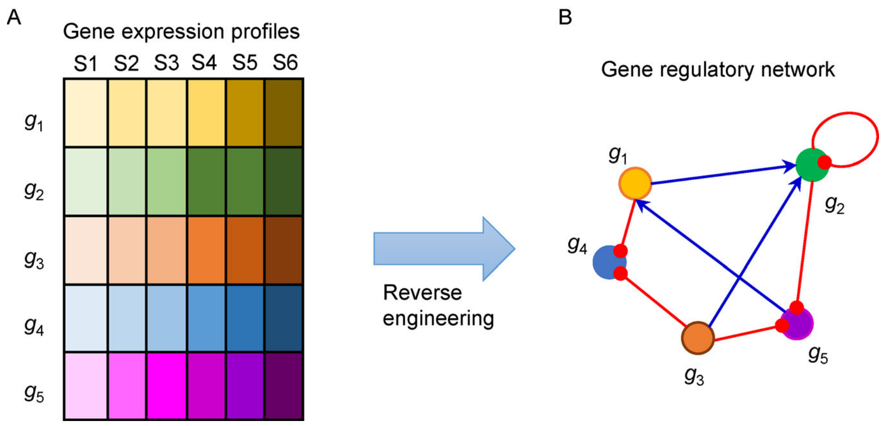

Figure 1. The framework of reconstructing gene regulatory network (B) from gene expression profiling data (A).

performers (genes). The difficulty underlies the substance of reversely engineering gene regulatory network from expression data, not to mention the noise of these high-throughput data-generating techniques in their developing periods [15].

Methodologically, gene regulatory network inference is essentially difficult because of the curse of dimensionality. It is a typical large n, small p problem, which indicates the dimension (gene number) of the inference problem is very big (large n, often ~30,000), while there are only a few samples from the high-throughput experiments (small p, often ~10), *i.e.*, $p \ll n$. The constraints generated from these samples are such few that there are many possible combinatorial solutions (the regulatory interactions of gene relationships), which makes the feasible solution space become very huge, then makes the search of optimal solution of genuine regulations become very difficult.

In computational details, supposing there are n genes in a regulatory network, we model their regulatory dynamics by the ordinary differential equations (ODE) of their expression levels as follows:

$$\begin{cases} g_1' = \dfrac{\mathrm{d}g_1}{\mathrm{d}t} = c_{11}g_1 + c_{12}g_2 + \cdots + c_{1n}g_n \\ g_2' = \dfrac{\mathrm{d}g_2}{\mathrm{d}t} = c_{21}g_1 + c_{22}g_2 + \cdots + c_{2n}g_n \\ \cdots \\ g_n' = \dfrac{\mathrm{d}g_n}{\mathrm{d}t} = c_{n1}g_1 + c_{n2}g_2 + \cdots + c_{nn}g_n \end{cases} \tag{1}$$

where g_j' simply refers to the first order derivative of the expression level of gene g_j. Under the linear assumption of gene regulatory relationships, the derivative change is caused by the combinatorial regulations of the other genes in the system reflecting by their expression levels. Thus, the gene regulatory network inference is formulated to a parameter identification problem of these equations. When the coefficients c_{ij} are determined by leveraging gene expression data, their regulatory relationships are coordinated and the regulatory network is then reconstructed.

Compare to standard linear equations in algebra, the variables in Equation (1) are different and converse. We suppose the number of experiments is p. Let $(c_{11}, c_{12}, \cdots, c_{1n})^T$ be $(x_1, x_2, \cdots, x_n)^T$, the determination of these coefficients for g_1' is formulated as linear equations with standard formats.

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_{11} \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_{21} \\ \cdots \\ a_{p1}x_1 + a_{p2}x_2 + \cdots + a_{pn}x_n = b_{p1} \end{cases} \tag{2}$$

where matrix $\boldsymbol{A} = (a_{ij})_{p\times n}$ refers to the gene expression values, a_{ij} is the gene expression of g_j in the i-th experiment. b_{i1} $(1 \leqslant i \leqslant p)$ refers to the derivation values of g_1' in the i-th experiment, which is often evaluated by approximation [16]. In the ubiquitous experiment design, the experiment number p is often small due to the limited resources. Thus, $\mathrm{Rank}(\boldsymbol{A}) \leqslant p \ll n$, and there are infinitely many solutions for these equations. It is very hard to achieve an optimal solution under few additional constraints for g_1. Note the equations in Equation (2) are only for the coefficients of g_1'. For the other genes, the similar cases exist in these equations because the left-hands of gene expressions are same and only the response derivatives on the right-hands are different. The solution space of these coefficients in Equation (1) is so huge that it is difficult to achieve the optimal solutions for individual genes, *i.e.*, the inferred regulatory coefficients between genes.

Theoretically, the parameters of the ODE system in Equation (1) are identifiable by achieving unique solutions when there are enough experimental gene expression datasets. For practical experiment constraints, the limited observational data cause the gene regulatory model partially identifiable [17]. The former model is very simple when compared with true gene regulatory processes in cells. Advanced models including time-varying parameters, dynamic gene interactions and higher order derivatives of gene expression will make the computational tasks of gene regulatory network inference much more complicated [16].

AVAILABLE RESOURCES AND METHODS

Due to the centrality of gene regulation in biological processes, numerous resources have been available for characterizing gene regulatory systems from multiple molecular levels. As described in the former section, the original gene regulatory network reconstruction is from gene expression profiling data, which are deposited in databases such as GEO [8], ArrayExpression [9] and SRA [18]. Table 1 [19–44] lists some currently available resources for inferring gene regulatory networks. For instances, the main components in gene regulations, such as gene, TF, miRNA and protein can be accessed from GenBank [21], UniProt [43], miRBase [38] and PDB [42], respectively. The other elements and effects such as ncRNA [33], RNA methylation [26] and chromatin accessibility [41] are gradually recognized for understanding gene regulations in much more details. Some prior gene regulation information are also deposited in databases such as RegNetwork [32], which collects the known and predicted transcriptional regulations (TF-gene) and posttranscriptional regulations (miRNA-mRNA) from various databases and literature. The

Table 1 Some available data resources for inferring gene regulatory networks

Category	Representative	Description	Refs.
Disease/Expression/ RNA	TCGA	The cancer genome atlas (TCGA) generates comprehensive maps of genomic changes in more than 30 types of cancer from more than 11,000 patients	[19]
	ICGC	The international cancer genome consortium (ICGC) obtains comprehensive descriptions of genomic, transcriptomic and epigenomic changes in 50 different tumor types	[20]
DNA	GenBank	GenBank is the genetic sequence database, with the multiple annotations of all DNA sequences publicly available	[21]
DNA/RNA/Protein	ENCODE	ENCODE project is a public research consortium to identify all functional elements in human genome	[22]
	modENCODE	It identifies all of the sequence-based functional elements in some model organisms	[23]
Epigenomics	ROADMAP Epigenomics	It contains the maps of histone modifications, chromatin accessibility, DNA methylation, and mRNA expression across various human cell types and tissues	[24]
	NCBI Epigenomics	A new public resource for exploring epigenomic data sets in NCBI	[25]
	RNAMDB	A database about RNA modifications containing in many of the known RNA species	[26]
	MODOMICS	MODOMICS is a comprehensive database of RNA modifications with integrated information related to RNA epigenetics	[27]
	IHEC	The international human epigenome consortium coordinates the production of reference epigenome maps of various tissues and cell types	[28]
Expression	GEO	The widely accessed database for depositing gene expression data of microarray and RNA-Seq techniques at NCBI	[8]
	ArrayExpress	It is the data archive of storing data from high-throughput functional genomics experiments, microarrays and RNA-Seq at EBI	[9]
	GTEx	The genotype-tissue expression (GTEx) program provides the information about gene expression and regulation in multiple tissues	[29]
	SRA	The database provides a repository for the studies of RNA-Seq and ChIP-Seq, as well as human microbiome project and the 1000 genomes project	[18]
Gene regulation	TRANSFAC	It provides the transcription factors of data on eukaryotes, and their consensus binding sites and target genes	[30]
	JASPAR	JASPAR contains the curated transcription factor binding profiles and sites for eukaryotes	[31]
	RegNetwork	An integrated database of transcriptional and posttranscriptional regulations	[32]
ncRNA	NONCODE	NONCODE is an integrated knowledge database of non-coding RNAs, especially for lncRNA	[33]
	RNAcentral	RNAcentral is a comprehensive database of non-coding RNA sequences	[34]
	TarBase	A comprehensive database of experimental microRNA targets	[35]
	Lncipedia	A database of human annotated lncRNA transcript sequences and structures	[36]
	lncRNAdb	lncRNAdb contains a comprehensive list of lncRNAs that are related to biological functions	[37]
	miRBase	miRBase is a searchable database of published miRNA sequences and annotation	[38]
	circBase	circBase is a comprehensive database for circular RNAs in multiple organisms	[39]
ncRNA/Interaction/ Expression	ChIPBase	A database about transcriptional regulations of long non-coding RNA and microRNA genes from ChIP-Seq	[40]
Modification	CR Cistrome	A database of chromatin regulators and histone modification linkages collected from ChIP-Seq	[41]
Protein	PDB	PDB is the resource of 3D shapes of proteins, nucleic acids, and complex assemblies	[42]
	UniProt	UniProt provides a comprehensive, high-quality and freely accessible data about protein sequences and functional annotations	[43]
Protein interaction	STRING	STRING is a database of the known and predicted protein-protein interactions	[44]

existing resources provide the materials for building computational methods of reconstructing regulatory networks.

The corresponding regulatory components and information documented in various databases also bring great challenges for integrating these available data. These elements located in different databases are often curated by different research groups and institutes. Thus, the uniform identifiers (ID) for these regulators and targets are often very important in the concrete network inferences. The ID mappings are always time consuming and some ID mapping tools, such as that of UniProt [43], provide the online services for interchanging the component IDs of different databases. In the inference of gene regulatory network, the ID should be consistent when we integrate data from diverse sources.

So far, numerous methods have been proposed for reconstructing gene regulatory networks. Table 2 lists some of the representative methods. We group them into six categories, *i.e.*, association methods, Bayesian methods, Boolean networks, differential-equation-based methods, knowledge-based methods, and machine-learning-based methods. Here, we briefly introduce the main ideas of each category individually and please see our paper [3] and the references therein for much more details of these available methods.

The first is the association-based methods. These methods are to calculate the expression relationships between genes by defining association measures. The gene pairs will be identified as regulatory interactions when their associations are significant, such as WGCNA [59]. Let gene X and gene Y with their expressions be $X=(X_1,X_2,...,X_p)$ and $Y=(Y_1,Y_2,...,Y_p)$ respectively. Based on the two sample vectors, Pearson's correlation coefficient (PCC) can be easily calculated. WGCNA defines $S_{XY}^{unsigned}=|cor(X,Y)|$ or $S_{XY}^{signed}=\frac{1}{2}+\frac{1}{2}cor(X,Y)$ as their association scores for unsigned regulation or signed regulation. Based on these pairwise correlation values, it uses a concept of topology parameter to set the threshold of linking edges between genes. Thus a gene coexpression regulatory network will be built up. Similarly, the other association measures such as mutual information can be employed in the calculation. Specifically, mutual information between X and Y in terms of entropy is $I(X,Y)=H(X)+H(Y)-H(X,Y)$, where $H(X)$, $H(Y)$ and $H(X,Y)$ are the marginal entropies of X and Y, and their joint entropy, respectively. There are many other association-based methods have been proposed for its simplicity. The association network is often the first try of investigating relationship between genes. For their popularity, we provided a comprehensive comparison study of their performances of reconstructing gene regulatory network in [60].

Due to the indirect relationships in calculating these associations, some association-based methods improve the network inference by conditional probability such as partial correlation coefficient and conditional mutual information [60]. By introducing the other gene or gene set in evaluating the association values, the false positives of indirect regulations will be eliminated for improving the inference accuracy. In this sense, an effective method named ARACNE [61] employs the information equality formula to remove the indirect mutual information between gene pairs. We proposed such a method named PCA-CMI [45] based on conditional mutual information. The strategy is proven to be effective to eliminate false positive inferences. We will introduce it in details in the next sections.

The second category is based on Bayesian network

Table 2 The main categories of these methods for inferring gene regulatory networks from high-throughput data

Category	Representative	Description	Refs.
Association methods	WGCNA, MINET, ARACNE, PCA-CMI	Calculating the association value between genes as their regulatory potential. Some of them have been improved for removing false positives	[45–48]
Bayesian methods	BNFinder, BNLEARN	Modeling the regulatory interdependence between genes by Bayesian statistics. The regulatory links will be built up from the posterior probability between structure and data	[49,50]
Boolean methods	BoolNet, PBN	The logical operations are employed to model the decision-makings of gene regulatory interactions	[51,52]
Differential-equation-based methods	D-NetWeaver, Inferelator, GeneNetWeaver	The derivatives of gene dynamics are modeled as the responsive variable of a set of regulatory factors. Then network inference is to identify the parameters of differential equations	[16,53,54]
Knowledge-based methods	Network Screening SITPR	Reconstructing gene regulations from high-throughput data by cooperating the prior knowledge of gene regulations	[55,56]
Machine-learning-based methods	TIGRESS, GENIE3	Learning the patterns of existing gene regulatory interactions by a machine-learning algorithm and predicting new regulations by the trained classifiers	[57,58]

They are alphabetically ranked by "Category" models. They are typical graphical models of representing the gene interdependence via a directed graph. Supposing a directed acyclic graph (DAG) of regulatory network be G, the probability of the consistency between graph and data be $P(G|D)$. According to Bayesian theorem $P(G|D)=\frac{P(D|G)P(G)}{P(D)}$, the posterior of measuring the consistency can be evaluated from this formula [62]. For extending DAG to general network structure, dynamic Bayesian network is employed to tackle the time series data [63]. Dynamic Bayesian network models display their powerful ability and flexibility of inferring gene regulatory networks from time-course gene expression data. These methods achieve successful applications in identifying various regulatory circuits [64].

The third category is based on Boolean networks. The logic operators "AND ($\wedge$)", "OR ($\vee$)", and "NOT ($\neg$)" are employed to address the decision-making regulatory processes of gene expressions [65]. Supposing gene X, Y, Z are in a Boolean network $G(V,F)$, where V is its node set, and F is its function set. The gene expression of Z is determined by the Boolean operation on the expressions of X and Y, such as $Z=f(X,Y)=X\wedge Y$, $f\in F$. The Boolean assumption often corresponds to the "ON" and "OFF" of gene regulations [66]. Due to the stochastics underlying gene regulations, Boolean networks have been extended to include the possibilities in the logic operations, *i.e.*, probabilistic Boolean networks [51]. The state transitions of gene regulations are shown to be feasible by modeling them as combinatorial logic computations. For the loose requirement of experimental samples in the logic computation, these methods are feasible for inferring gene regulatory networks from few samples of gene expression in single cell.

The fourth category is based on differential equations. As described in the former sections, the derivatives of gene expression are modeled as the response variable (gene) of some dependent variables (genes), *i.e.*, $\frac{\mathrm{d}x_j}{\mathrm{d}t}=\sum_{i=1}^{n}\beta x_i+\beta_0, j=1,\ldots,n$. After solving the parameters of these differential equations, the regulatory system will be built up for describing the gene regulatory interactions [16]. Equations with higher order of dependent variables and variable interactions generate more complicated models [53].

The fifth category is knowledge-based methods, which are based on the prior knowledge domain about gene regulations. For instance, if there is a regulation relationship between TF X and gene Y validated by experiments and documented in literature. The priors can be accessed freely from RegNetwork [32]. With these priors, the inference is constrained to be kept in the feasibility. For example, if we know X always activates Y, the following model constrains the knowledge of $X\rightarrow Y$ during the inferences, which benefits to avoid the possible misidentification and randomness of this regulation. With this in mind and to specify whether the regulation is existed in specific cases and conditions, the type of methods is to evaluate the probability of regulatory existence between X and Y in particular circumstances [67]. The consistency between knowledge regulation and phenotypic data are assessed for recognizing gene regulations. For instance, we proposed a network-based screening method for evaluating the prior regulatory networks in response to the gene expression profiling by graphical models [55,56]. The maximum likelihood of the consistency between them can be evaluated by a mathematical programming. For documented regulatory pathways, we identified the activated regulations in cell cycles [55] and during viral infections [56], respectively.

Last but not least, the machine-learning-based methods transform the inference of regulation between gene X and Y to a prediction of their regulatory interaction by a trained classifier [3]. The basic idea of the classifier is firstly to learn these corresponding features underlying the known regulatory relationships from gene expressions and/or sequences of X and Y. For instance, the corresponding regulatory relationship is modeled as $R=f(X,Y)$ by a machine learning method, where f is an abstract function often without its explicit form [68]. R is binary with label 1 representing the regulation between X and Y, with label 0 otherwise. Given new gene expression profiles, the trained machine-learning classifier predicts the new interacting events between the two genes. As listed in Table 2, TIGRESS presents a LARS (least angle regression) based classification with resampling the samples and variables [57]. It formulates gene regulatory network inference to feature selection. Some other machine-learning-based methods are based on sequence and/or structure features of TF and binding sites of operons [69]. The similar processes of training in the known gene regulations and predicting novel regulatory pairs are implemented to reconstruct gene regulatory networks.

TYPE I AND II ERRORS IN INFERENCE

In gene regulatory network inferences, Figure 2 shows the two types of errors, *i.e.*, type I error and type II error. The two errors are the main barriers of precisely reconstructing gene regulatory network from data. While the reasons of generating false positive and false negative inferences underlie at least three aspects. The first is the complicated regulatory relationships between transcriptional regulators and targets. The second is the developing periods of high-throughput technologies of generating the measured data. The third is about the proposed reverse engineering methods.

The gene expressions are controlled by many levels of regulators. TFs and cofactors perform their initialization of gene transcriptions. The products of mRNA are silenced by small noncoding RNAs (ncRNA), *e.g.*, miRNA [70]. The long non-coding RNA (lncRNA) is found to be crucial to coordinate gene expressions with combinatorial regulations of mRNA expression levels [71]. Recently, circular RNAs (circRNAs) are also found to be sponges to regulate gene expressions [72]. These provide clues and indications of the complexity of gene regulatory network. There might be some important regulators that still have not been revealed of performing crucial roles in controlling gene expressions.

The gene expression profiling techniques, such as microarray [12] and RNA-Seq [13], generate the high-dimensional data describing the expression abundancies. These techniques are still in their developing and maturing periods [73]. For instance, microarray splits each gene into several oligonuclotides and elaborately designs them on the probes of a microarray for measuring the expression of the corresponding gene by fluorescence [74]. The pipelines of following data preprocessing are still needed to be improved optimally. Thus, the measured datasets strongly affect the accuracy of gene regulatory network inference.

The reverse engineering methods implement various strategies of inferring the gene regulatory relationships via the measured datasets [3]. The noise of measured expressions during RNA extraction from samples, as well as the former-mentioned complexity of gene expressions restrict our precise inference [15,75]. The assumption and limitation of these computational methods also constraint the accuracy of inferring gene regulatory network from data.

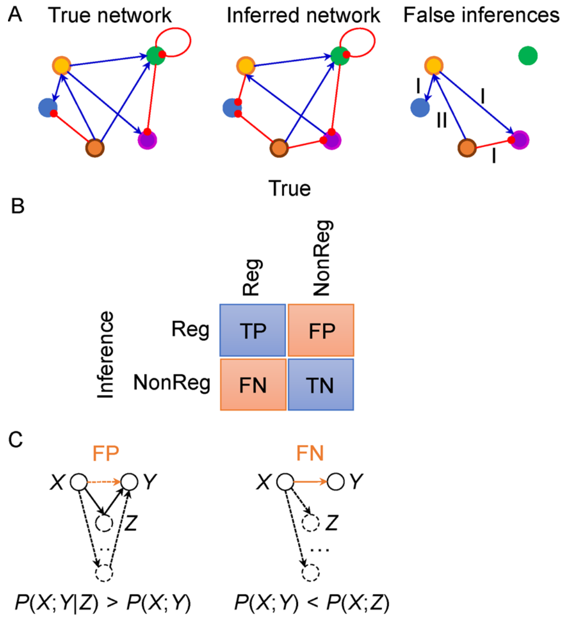

Figure 2. The false positives and false negatives in the inference network. (A) The true and the inferred gene regulatory networks, with their differences. The differences refer to the false positives and false negatives in the network inference from data. (B) The definition of two types of errors referring false positives and false negatives in the inference. Type I error: false positive (FP) refers to the inferred regulations which are not correct, Type II error: false negative (FN) refers to the inferred non-regulations which are not correct. (C) The possible reasons of false positives and false negatives. To achieve precision reconstruction of gene regulatory network, it is to remove the two types of errors for the possible reasons by introducing other gene or gene set.

Towards precise reconstruction of gene regulatory network from data is essentially to reduce the two types of errors in the inferences. Specifically, type I error is the false positive, which refers to the inferred regulations between genes which are not exist in truth. Type II error refers to the false negative, which is exist but is not inferred (as shown in Figure 2B). Moreover, the false positive inferences inherently contain numerous possibilities. The first is about the regulatory directions. As shown in Figure 2A, gene regulatory network is an oriented graph. The upstream regulators locate in the arc ends and their downstream targets locate at the arc heads. During the inferences, if the arc is upside down for some unknown reason, it will result in a wrong inference. The second is about positive and negative regulatory directions. The positive regulation is to enhance the gene expression by promoting transcriptions, while the negative regulation is refer to the inhibition of gene expression by repressing the transcriptional functions of TF, RNA polymerase, and their effective interactions with gene promoters and enhancers [76]. If the up- or down-regulations are upside down in some context-specific physiological processes, we can only obtain wrongly inferred regulations. The third is the inferences of non-existent regulations between genes, which are the major part of these false positives.

The major part of false positives are caused by many reasons, such as gene cooperation in regulations. For existing methods of reverse engineering, the inferences often reach concrete quantitative scores of gene regulation between regulator and target. The decision of the existence or non-existence of a regulatory relationship between two genes is often made by the score. As shown in Figure 2C, the regulatory information transferring from X to Y is assessed by a probability (score) $P(X;Y)$. In fact, the high score between X and Y is caused by Z, which means X and Z contain the regulatory cooperation of Y. The strong implication of a direct regulation between X and Y is caused by Z and gene Z directly regulates gene Y.

The combinatorial gene regulations of *X* and *Z* cause the indirect regulations between *X* and *Y*, which is a false positive regulation. We can use this philosophy to remove the indirect false positive regulations by introducing the other genes in the calculation of regulatory possibility. If the regulatory possibility conditioned on *Z*, $P(X;Y|Z)$ becomes bigger, this implies the regulation between *X* and *Y* is false positive. The indirect regulation between *X* and *Y* are caused by the direct regulation between *X* and *Z* and that between *Z* and *Y*. The other reasons might underlie the proposed methods, such as kernel canonical correlation analysis (CCA) of measuring the regulatory associations between genes. Kernel CCA extracts the partial correlations between genes, which easily results in high coefficient values then cause false positive inferences [60]. Due to the sparseness of gene regulation network, we can add a regularization to control these false positives.

The false negatives of type II error refer to the true regulations between genes which cannot be recognized by the proposed inference methods. There are many reasons for causing false negative inferences. The hardness of a single threshold or cut-off is such a reason of generating the false negative inferences. That is to say, we often only use a single threshold to evaluate all these candidate regulations with no regards of their temporal and spatial features underlying these regulations. For instance, the inferred coefficient between gene *X* and gene *Y* is 0.4, there is a true regulatory relationship between them during cell proliferation. When the threshold used for determining the existence of regulation between genes is 0.7, we will generate a type II error inference. The unsuitable threshold causes the missing regulatory relationship in the specific condition.

Compared with the analysis of type I error from a systematic perspective of gene regulatory cooperation, the type II errors are majorly caused by gene competition. For one target gene, the competition between several TFs will result in the dominant gene regulatory relationships between some TFs with their targets. As shown in Figure 2C, the regulatory score $P(X;Y)$ between *X* and *Y* is smaller than $P(Z;Y)$ between *Z* and *Y*. Moreover, if the value of $P(X;Y)$ is smaller than those of all the other pairs, the false negative will be generated after setting up a threshold for selecting the top-scored pairs if there really exists a regulation between *X* and *Y*. The weak regulatory signal between them causes the false negative inference.

It is very hard to remove the false negatives because the standard of threshold is often consistent for all gene pairs. The weak signal-noise-ratio of the true regulations cause the missing inferences [77]. The false negatives are caused by competitive values between gene pairs. If we can propose a dynamic threshold strategy by intelligently using different thresholds for different gene pairs according to their context-specific regulatory pathways, physiological processes and phenotypic conditions. The number of false negative interactions will be possibly decreased. That is to say, if we can optimally set up different thresholds for determining regulations in different contexts, the false negatives will be greatly reduced accordingly.

REMOVING FALSE POSITIVE REGULATIONS

To achieve accurate reconstruction of gene regulatory network from transcriptomic data, some methods have been proposed to remove the false positive predictions. So far, these available methods are often based on conditional probability and information theory. As discussed in the former sections, another gene or gene set will be introduced in the evaluation of regulations from a system biology perspective. Figure 3A demonstrates the main idea of employing the additional information from the third-party gene or gene set. For evaluating the regulatory score $P(X,Y)$ between *X* and *Y*, the other related gene or gene set will be gradually introduced in the calculation to remove possible biases and obtain a genuine regulatory score.

Suppose there is another gene or gene set *Z*, the conditional probability theory takes *Z* as conditions for accessing the genuine regulation between gene *X* and gene *Y*, *i.e.*,

$$I(X_i;Y_j|Z_k)=\sum_{X_i\in X,\ Y_j\in Y,\ Z_k\in Z}P(X_i;Y_j;Z_k)$$

$$\cdot\log\frac{P(X_i;Y_j|Z_k)}{P(X_i|Z_k)P(Y_j|Z_k)}.$$

As mentioned in the former section, when $I(X_i;Y_j|Z_k)>I(X_i;Y_j)$, the false positive regulation between *X* and *Y* is caused by the coexistence of gene *Z* or gene set *Z*. The indirect regulation between *X* and *Y* can then be removed and the inference accuracy can then be improved. Similarly, partial correlation coefficient can also be used to remove the false positives. It is defined as

$$r_{XY\cdot Z}=\frac{r_{XY}-r_{XZ}r_{YZ}}{\sqrt{(1-r_{XZ}^2)(1-r_{YZ}^2)}},$$

where $r_{\bullet\bullet}$ refers to the PCC between two genes, and $r_{XY\cdot Z}$ between *X* and *Y* is to extract the correlation between *X* and *Y* by removing the effects of *Z* [78].

Based on conditional mutual information (CMI) and path consistency algorithm (PCA), we have proposed a network inference method named PCA-CMI for removing false positive inferences [45]. As shown in Figure 3B,

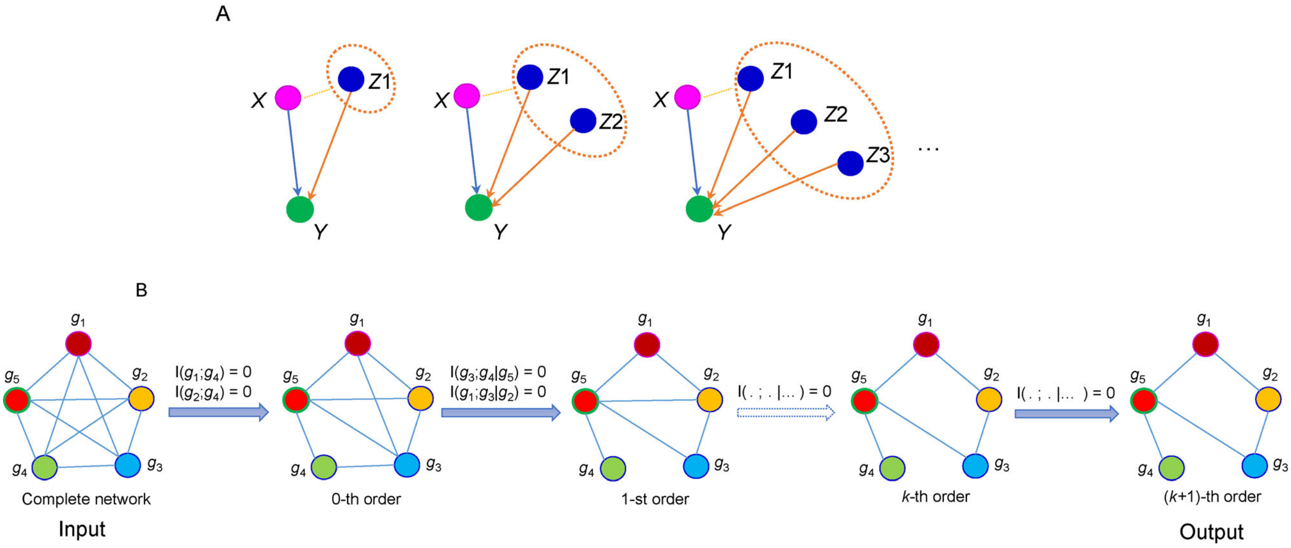

Figure 3. The employee of additional information from other genes to remove false positives. (A) Conditioning on other genes gradually to evaluate current regulation for identifying direct regulatory interaction. The conditioned gene *Z1* and gene set {*Z1*, *Z2*, …} for genes *X* and *Y*. (B) The framework of our proposed method PCA-CMI by eliminating indirect regulations from a complete network.

we firstly build a complete association network via mutual information. Then we employ the high-order CMI to eliminate the false positive regulations between genes iteratively. Under the PCA scheme, the iterative algorithm will be terminated if a consistent network has been achieved at the steps of the *k*-th order CMI and the $(k+1)$-th order CMI. Because we start our inference from a complete graph, there are no false negative regulations, *i.e.*, no type II errors at the initialization step of our algorithm. The reference or background network has $\binom{n}{2}$ possible regulations for chosen. With the conditioned gene or gene set, the false positives of indirect regulations will be removed gradually. From computational perspective, CMI can practically calculate small- or middle-size network. Fortunately, we have improved it to be a whole-genome-wide reconstruction method by parallel computing [79].

In fact, we provide a very general strategy of inferring gene regulatory networks by controlling the type I errors. For the tremendous difficulty of controlling the type II errors, we start from a complete network without any false negative although it is still possible to contain type II errors in the finally reconstructed network. It is easy to change CMI to partial correlation coefficient in the strategy. We can also calculate the conditional association when we have a prior knowledge network in some specific conditions. By introducing the related gene or gene set, the indirect regulatory relationships will be removed from the background network to achieve accurate inference of responsive gene regulations.

IMPROVING INFERENCE BY PRIOR KNOWLEDGE

Pure data-driven methods cannot always guarantee the accurate inference of gene transcriptional regulations in many aspects, such as the upside and downside from regulators to targets, the positive and negative regulatory directions. For instances, the former reviewed association network inference methods based on mutual information contain no such ability of distinguishing regulatory directions. Without the prior knowledge about regulators such as TFs, the improved conditional association-based methods still cannot obtain such information. As for our method PCA-CMI [45], it reconstructs an undirected association network without the regulatory directions indicating information transmission, *i.e.*, which ones are the regulators and which are the corresponding targets. The causality between genes cannot be modeled and revealed [60]. Easily, if we set up the prior TFs in the inference, the association-based networks will be oriented and signed. For this perspective, we need combine the prior knowledge about gene regulations with the high-throughput data to achieve more accurate network reconstruction.

Although these pure data-driven methods of inferring gene regulatory network seem to be very flexible in scope, it is impossible to infer all the negative regulations between miRNAs and targets without any incorrect identification. In our knowledge, miRNAs almost always perform their negative regulatory functions by silencing gene transcripts. If we do not include such information and infer the posttranscriptional regulations only from data, the inference will definitely contain many false positives. For instance, we employ Pearson's correlation coefficients as the quantitative association measures to access the regulatory relationships between miRNAs and their targets using the gene expression profiles of human liver tissues of HCC [80]. The values are between -1 and 1, and the negative and positive regulations are also defined accordingly. For the documented miRNA-mRNA interactions downloaded from RegNetwork [32], Figure 4A illustrates the frequency distribution of these pairwise PCC values underlying these miRNA-target pairs. We can find that almost half number (47.8%) of the correlations are with positive values, which indicate that all these pairs with positive PCC values will result in false positive regulations if we purely implement the network inference from data. As for the gene pairs with absolute values over than 0.7, we can find most of them are with positive PCCs. They are false positive inferences if we employ the PCC-based methods. The example clearly motivates the integration of such prior knowledge with pure data-driven methods. If we constrain the negative coefficients in the inferences, we will achieve precise reconstruction of gene regulations [81].

For the urgent requests of combining the prior knowledge of transcriptional and posttranscriptional regulations in network inference, we built RegNetwork for documenting the available regulatory interactions [32]. We integrated the experimental regulations from numerous databases as well as the predicted regulations from TF-binding sequence motifs. Currently, RegNetwork contains two genome-wide regulatory networks of 300,000 + edges and 20,000 + nodes for human and mouse respectively. If we start from the prior regulatory network, with the profiling data of gene expressions, we can identify the activated gene regulatory networks in response to specific biological processes and phenotypic conditions. To this end, we inferred a regulatory network during influenza A virus infection in human cells by integrating prior genome-wide networks and condition-specific gene expressions. The known regulations such as the downregulation of miRNAs are introduced as the constraints of negative coefficients, *i.e.*, $C_{XY} \leqslant 0$ between miRNA X and its target Y [81]. Under the constraints of prior regulations, the reconstructed gene regulatory network implies the synergetic regulations between transcriptional regulations and posttranscriptional regulations by the cooperation between TF and miRNA with high accuracy [81].

For accessing the transcriptional activity of prior regulatory network in specific conditions, we proposed a network screening method of identifying responsive gene regulations by measuring the consistency between network structure and expression profiles [55]. The consistency between them is measured by a maximum likelihood value of the graph consistency with the data, and then a permutation test is performed to evaluate its statistical significance. In other words, we transform the problem of reconstructing network from data to selecting a responsive network possibly with minor modifications by accessing the match between network structure and measured data. Different conditions and differential gene expression status will result in different activated regulatory network structures. The obtained gene network structure reflects the specific gene regulations in which their gene expressions have been measured by transcriptomic techniques. In different philosophy, we change a

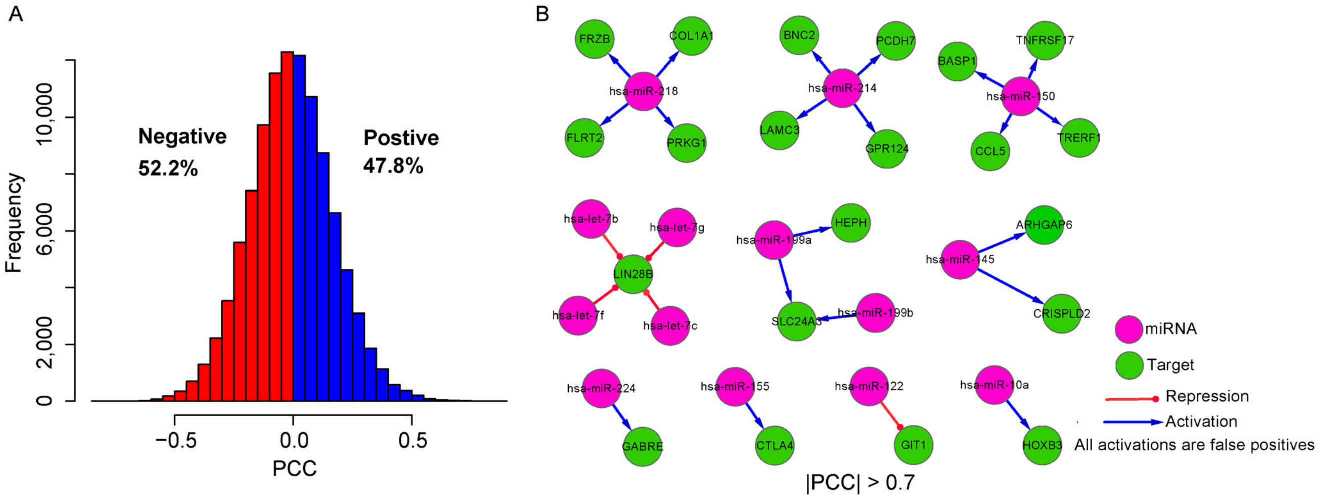

Figure 4. The PCCs between miRNAs and targets in the cases of HCC. (A) The distribution of PCCs between miRNA and target gene expressions. (B) The interactions with absolute PCCs over than 0.7. The miRNA-target interactions are downloaded from RegNetwork.

reverse engineering of inferring gene regulatory network from expression data to a forward engineering of designing network structure based on prior knowledge to achieve its match with the data [3].

Harnessing prior knowledge of gene regulations can avoid many potential pitfalls of inaccurate network inferences. If we know which one are the upstream regulators and which one are the downstream targets, we will reach a suitable inference by integrating the priors in the model. Moreover, prior knowledge indicates the truth and standard. When we set up which is regulator and which is target, the models of inference become purposeful. And in the conditional probabilities, we cannot calculate all the possible combinations in these conditioned genes. While with the guidance of prior knowledge, we can formulate the regulatory processes into a rational model of describing causal regulations [16,56]. The prior knowledge narrows down the search space of solutions and the identification of regulations become much easier. It highly benefits the precision inference of gene regulatory network from data.

IMPROVING INFERNECE BY MULTI-OMICS DATA

In current big data era, it is promising to accurately reconstruct gene regulatory network by integrating omics data. There are two lens of data integration. The first is to combine all the related data such as transcriptomics in an integrated manner. For instance, there are three datasets about the study of hepatocellular carcinoma caused by HCV infection. To achieve an accurate inference of the responsive regulations between gene *X* and gene *Y*, we need to investigate the different cases of data collection to achieve a robust inference of regulatory relationship between *X* and *Y*. The gene regulations are temporal and spatial. The data integration from various laboratories, groups, platforms and institutes will build consistent and cross-data-validated gene regulatory networks [82]. The false positives will be controlled from double-checks from multiple transcriptomic datasets.

The second aspect of data integration refers to integrate multi-level omics data for characterizing gene regulations in cells. To achieve precision reconstruction of gene regulatory network, we should integrate multi-level omics data and infer hierarchical gene regulations. Compared with the former aspect of "deep" data integration, the hierarchical regulatory network locates special emphasis on the "wide" or "broad" data integration. The determinants of gene expression coordinate in the central dogma that DNA makes RNA and RNA makes protein. The genetic information transformation from DNA to RNA to protein constructs a hierarchical information system with many events related to the transcriptional level where gene regulation takes place. The regulatory components and elements often function in specific conditions. The measured gene expressions often refer to some very specific conditions. For accurately inferring a casual regulatory network of controlling gene expression, it is needed to integrate the data of various levels to build a comprehensive regulatory map.

Figure 5 illustrates the hierarchical levels of gene regulation. For transcription, it is initialized with the binding events of RNA polymerases and TF recognition of a gene promotor region. In current high-throughput techniques, the expression-level measured data are mainly from microarray [12] and RNA-Seq [13]. Tens of thousands of gene expressions are measured from the extracted RNA concentrations. The transcriptional-level DNA modifications such as methylation and demethylation, chromatin open status of protein-DNA contacts highly affect gene expressions [83]. In the posttranscriptional level, the non-coding RNAs will highly affect RNA abundancies, such as miRNA, lncRNA and circRNA

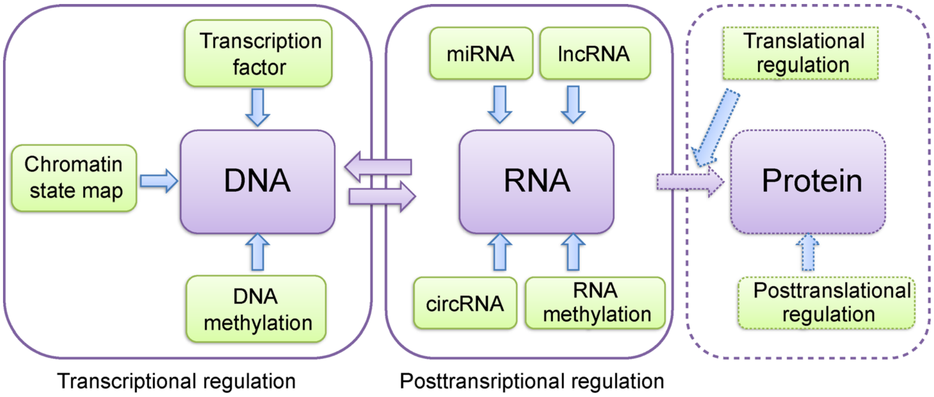

Figure 5. The hierarchical gene regulatory information flows in the central dogma. The major parts of gene regulatory network inferences refer to the transcriptional and posttranscriptional regulations, and the translational and posttranslational regulations have profound impacts for upstream information fluids and circuits, such as the activities of TF and RNA polymerase.

[84]. Moreover, the translation and posttranslational modifications will also affect the TF activity of protein abundancy [85]. In practice, gene regulations function as an integrated system. Genetics and epigenetics elements perform their critical roles in gene regulations under different temporal and spatial conditions [76]. If multi-level omics data about them are available, we should integrate these data in a hierarchical manner to build an integrative network between these regulatory components.

So far, some methods have been proposed to integrate multi-omics datasets for inferring gene regulatory networks [86]. They can roughly be categorized into three pipelines. The first is based on regression. Supposing there are M types of multi-omics data, *i.e.*, D_1, D_2, $\cdots$, D_M. The expression g_i of gene i is modeled as the response variable of these regulators (explanatory variables) such as TF f_j according to these multi-omics data. At the time point t, the regression model is formulated as

$$g_{it} = \alpha_i + \sum_{j}^{M} \beta_j D_{ij} f_{jt} + \varepsilon_{it}, \varepsilon_{it} \sim \text{Normal}(0, \sigma^2).$$

where β_j is the ordinary regression coefficient. D_{ij} is the coefficient of individual factor on the TF activity f_j. The multiple effects are from the multiple TF-influence datasets, and they are integrated into these independent variables (regulators). The problem of inferring gene regulations from multi-omics data is thus changed to identify these parameters in the linear regression equations [87]. More complicated regression techniques joint with the latter categories are expected to achieve more precision reconstruction [57,88–90].

The second is based on Bayesian theory. Similar to the first pipeline, the second category models the gene expression as a joint probability of TFs and the influence factors of TFs, *i.e.*, $g_{it} = P(f_{it}|D_{1t}, D_{2t}, ..., D_{Mt})$. According to the Bayesian theorem, the intra-relationship structure of these factors in the multi-omics can be represented as

$$P(D_{1t}, D_{2t}, ..., D_{Mt}) = \sum_{i,j,k}^{M} P(D_{it}|D_{jt}, ..., D_{kt}).$$

Joint with the combinatorial control from regulators, the combinatorial factors to these regulators reflecting in these multi-omics make the network inference model very complicated [91–93]. With prior knowledge such as the information flow in the central dogma will make sense of the former probability expansion from biological perspectives, such as TF in transcriptional regulation, miRNA in posttranscriptional regulations, and phosphorylation in posttranslational modification. In this sense, we proposed a method named SITPR to integrate multi-omics data for inferring TF-miRNA cooperative regulatory network by prior knowledge of gene regulations, TF-binding sequence motifs, protein-protein interactions, as well as mRNA and miRNA expressions [56]. We reconstructed a comprehensive transcriptional and post-transcriptional regulatory map during influenza A virus infection in human epithelial cells.

The third is based on machine learning. In this category, the methods consider all the related multi-omics data as the regulatory features and used them to train a machine learning classifier [94,95]. It often achieves good inference performance in the tests and validations [58,95]. However, the inference results is hard to be interpreted due to the "black-box" paradigm of machine learning classifiers [68]. They focus on the binary decision-making of the existence of regulations, and many inter- and intra-relationships among variables (genes, regulators, cofactors, and effectors) are often missing. In the former two pipelines, the regulatory principle and mechanism can be easily revealed due to the causal information flows underlying these modeling variables.

Currently, the high-throughput microarray and next-generation-sequencing techniques are always implemented for cell populations [96]. The cell types and their status are mixed together with few information about the specific transcription status. In fact, the gene expressions are related to these former mentioned factors, such as DNA modifications, ncRNA and RNA epigenomics, protein translation and modification [97]. Their mixed-effects of gene expressions are still not intensively considered in the network inferences. The biology of life activity is dependent on the contexts of DNA, RNA and protein and their inter- and intra-regulations in single cells [98]. To study specific cell types in different time and space by their expressions and relationships in specific organism is an alternative way to clearly understand gene regulatory mechanisms [99]. With the advance of single cell genomics and sequencing, more specific contexts of these microenvironments of gene expression will enhance to build more accurate gene regulatory networks when these data are available [100].

From computational perspectives, the multi-level omics data integration provides a promising way to tackle the difficulty of inferring of gene regulatory network. If multi-level omics data are available, we can figure out the trajectory of the dynamics of gene expressions. And the measured gene expression profiles will be modeled as the integrated results of these regulators by matching their cooperation. Generally, it is difficult to build a rational model of describing the hierarchical regulatory structures. In special cases such as with Gaussian distribution function assumptions, the mixed-effects model of these regulators is expected to generate an accurate approximation gene expressions measured from multiple experiments, *i.e.*, $Y=\sum_{i}^{m}\alpha_i X_i+\sum_{j}^{n}\beta_j Z_j+\varepsilon$, where the gene expression of Y is a response variable of the fixed effects of multi-omics information of X, and the random effects Z of the hierarchical multi-omics information of X and that of Y and their combinations. ε is an error. The information hierarchy of multiple omics can also be modeled in it by introducing causal dependences, *i.e.*, $P(X_{i_{k1}}\rightarrow X_{i_{k2}})$, $k_1, k_2 \in \{i\}$ in X [3]. In the near future, with the development of single cell techniques, the purity of gene expressions in single individual cells will bring more clean data with less noise and then achieve more accurate inference of gene regulatory networks. With more and more advanced techniques, we will definitely achieve higher resolutions of regulatory maps by data integration. The dynamics will be finally reconstructed for describing regulatory complexity with the assistance of more advanced technology such as real-time single-molecule observation [101].

CONCLUSION AND OUTLOOK

The difficulty of reversely engineering gene regulatory network from data essentially caused by the signal-noise ratio underlying the data. If our computational methods can distinguish noises from signals and the inference accuracy will be improved. For removing the two types of errors, *i.e.*, false positives and false negatives, in the inference, we can integrate the available gene expression profiles in deep and multi-level omics datasets in broad to build a comprehensive regulatory network. In this work, we provided a review of gene regulatory network inference from the perspectives of computational complexity, available resources, existing methods, and the endeavors of reducing the substantial two types of errors. We highlighted the importance of data integration to achieve accurate inferences, especially prior knowledge and multiple omics. The systems biology approaches seem to be the rational alternatives of deciphering gene regulations from data. Due to the temporal and spatial features of gene regulation, we also need to focus on specific individual conditions and single cells, and then concentrate them into the whole gene transcriptional regulatory processes. If the resolutions of experimental measures will be improved, the globally dynamic movie of gene regulations will be clearly reconstructed by integrating available high-throughput datasets.

ACKNOWLEDGEMENTS

Thanks are due to the three anonymous reviewers for their constructive comments. This work was partially supported by the National Natural Science Foundation of China (Nos. 61572287 and 61533011), the Shandong Provincial Key Research and Development Program (2018GSF118043), the Natural Science Foundation of Shandong Province, China (ZR2015FQ001), the Fundamental Research Funds of Shandong University (Nos. 2015QY001 and 2016JC007), the Scientific Research Foundation for the Returned Overseas Chinese Scholars, Ministry of Education of China.

COMPLIANCE WITH ETHICS GUIDELINES

The author Zhi-Ping Liu declares that he has no conflict of interests.

This article is a review article and does not contain any studies with human or animal subjects performed by the author.

REFERENCES

1. Marx, V. (2013) Biology: the big challenges of big data. Nature, 498, 255–260
2. Babu, M. M., Luscombe, N. M., Aravind, L., Gerstein, M. and Teichmann, S. A. (2004) Structure and evolution of transcriptional regulatory networks. Curr. Opin. Struct. Biol., 14, 283–291
3. Liu, Z. P. (2015) Reverse engineering of genome-wide gene regulatory networks from gene expression data. Curr. Genomics, 16, 3–22
4. Lee, T. I. and Young, R. A. (2013) Transcriptional regulation and its misregulation in disease. Cell, 152, 1237–1251
5. Bandyopadhyay, S., Mehta, M., Kuo, D., Sung, M. K., Chuang, R., Jaehnig, E. J., Bodenmiller, B., Licon, K., Copeland, W., Shales, M., *et al.* (2010) Rewiring of genetic networks in response to DNA damage. Science, 330, 1385–1389
6. Johnson, D. S., Mortazavi, A., Myers, R. M. and Wold, B. (2007) Genome-wide mapping of *in vivo* protein-DNA interactions. Science, 316, 1497–1502
7. Park, P. J. (2009) ChIP-seq: advantages and challenges of a maturing technology. Nat. Rev. Genet., 10, 669–680
8. Edgar, R., Domrachev, M. and Lash, A. E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res., 30, 207–210
9. Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G. G., *et al.* (2003) ArrayExpress—a public repository for microarray gene expression data at the EBI. Nucleic Acids Res., 31, 68–71
10. Jaenisch, R. and Bird, A. (2003) Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat. Genet., 33, 245–254
11. Song, C. X., Yi, C. and He, C. (2012) Mapping recently identified nucleotide variants in the genome and transcriptome. Nat. Biotechnol., 30, 1107–1116
12. Schena, M., Shalon, D., Davis, R. W. and Brown, P. O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science, 270, 467–470
13. Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet., 10, 57–63
14. Blackham, S., Baillie, A., Al-Hababi, F., Remlinger, K., You, S., Hamatake, R. and McGarvey, M. J. (2010) Gene expression

profiling indicates the roles of host oxidative stress, apoptosis, lipid metabolism, and intracellular transport genes in the replication of hepatitis C virus. J. Virol., 84, 5404–5414

15. Shi, L., Reid, L. H., Jones, W. D., Shippy, R., Warrington, J. A., Baker, S. C., Collins, P. J., de Longueville, F., Kawasaki, E. S., Lee, K. Y., *et al.*. (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat. Biotechnol., 24, 1151–1161
16. Wu, S., Liu, Z. P., Qiu, X. and Wu, H. (2014) Modeling genome-wide dynamic regulatory network in mouse lungs with influenza infection using high-dimensional ordinary differential equations. PLoS One, 9, e95276
17. Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U. and Timmer, J. (2009) Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. Bioinformatics, 25, 1923–1929
18. Leinonen, R., Sugawara, H. and Shumway, M., and the International Nucleotide Sequence Database Collaboration. (2011) The sequence read archive. Nucleic Acids Res., 39, D19–D21
19. The Cancer Genome Atlas Research Network. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature, 455, 1061–1068
20. Hudson, T. J., Anderson, W., Artez, A., Barker, A. D., Bell, C., Bernabé, R. R., Bhan, M. K., Calvo, F., Eerola, I., Gerhard, D. S., *et al.* (2010) International network of cancer genome projects. Nature, 464, 993–998
21. Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. and Sayers, E. W. (2013) GenBank. Nucleic Acids Res., 41, D36–D42
22. Maher, B. (2012) ENCODE: the human encyclopaedia. Nature, 489, 46–48
23. Muers, M. (2011) Functional genomics: the modENCODE guide to the genome. Nat. Rev. Genet., 12, 80
24. Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Beaudet, A. L., Ecker, J. R., *et al.* (2010) The NIH Roadmap Epigenomics Mapping Consortium. Nat. Biotechnol., 28, 1045–1048
25. Fingerman, I. M., McDaniel, L., Zhang, X., Ratzat, W., Hassan, T., Jiang, Z., Cohen, R. F. and Schuler, G. D. (2011) NCBI Epigenomics: a new public resource for exploring epigenomic data sets. Nucleic Acids Res., 39, D908–D912
26. Cantara, W. A., Crain, P. F., Rozenski, J., McCloskey, J. A., Harris, K. A., Zhang, X., Vendeix, F. A., Fabris, D. and Agris, P. F. (2011) The RNA Modification Database, RNAMDB: 2011 update. Nucleic Acids Res., 39, D195–D201
27. Machnicka, M. A., Milanowska, K., Osman Oglou, O., Purta, E., Kurkowska, M., Olchowik, A., Januszewski, W., Kalinowski, S., Dunin-Horkawicz, S., Rother, K. M., *et al.* (2013) MODOMICS: a database of RNA modification pathways—2013 update. Nucleic Acids Res., 41, D262–D267
28. Bujold, D., de Lima Morais, D.A., Gauthier, C., Côté, C., Caron, M., Kwan, T., Chen, K.T., Laperle, J., Markovits, A. N., Pastinen, T., *et al.* (2016) The International Human Epigenome Consortium Data Portal. Cell Syst., 3, 496–499
29. Ardlie, K. G., Deluca, D. S., Segre, A. V., Sullivan, T. J., Young, T. R., Gelfand, E. T., Trowbridge, C. A., Maller, J. B., Tukiainen, T., Lek, M., *et al.* (2015) The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science, 348, 648–660
30. Matys, V., Fricke, E., Geffers, R., Gössling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O. V., *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res., 31, 374–378
31. Bryne, J. C., Valen, E., Tang, M. H., Marstrand, T., Winther, O., da Piedade, I., Krogh, A., Lenhard, B. and Sandelin, A. (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic Acids Res., 36, D102–D106
32. Liu, Z. P., Wu, C., Miao, H. and Wu, H. (2015) RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. Database (Oxford), 2015, bav095
33. Xie, C., Yuan, J., Li, H., Li, M., Zhao, G., Bu, D., Zhu, W., Wu, W., Chen, R. and Zhao, Y. (2014) NONCODEv4: exploring the world of long non-coding RNA genes. Nucleic Acids Res., 42, D98–D103
34. The RNAcentral Consortium. (2015) RNAcentral: an international database of ncRNA sequences. Nucleic Acids Res., 43, D123–D129
35. Sethupathy, P., Corda, B. and Hatzigeorgiou, A. G. (2006) TarBase: a comprehensive database of experimentally supported animal microRNA targets. RNA, 12, 192–197
36. Volders, P. J., Helsens, K., Wang, X., Menten, B., Martens, L., Gevaert, K., Vandesompele, J. and Mestdagh, P. (2013) LNCipedia: a database for annotated human lncRNA transcript sequences and structures. Nucleic Acids Res., 41, D246–D251
37. Amaral, P. P., Clark, M. B., Gascoigne, D. K., Dinger, M. E. and Mattick, J. S. (2011) lncRNAdb: a reference database for long noncoding RNAs. Nucleic Acids Res., 39, D146–D151
38. Griffiths-Jones, S., Saini, H. K., van Dongen, S. and Enright, A. J. (2008) miRBase: tools for microRNA genomics. Nucleic Acids Res., 36, D154–D158
39. Glažar, P., Papavasileiou, P. and Rajewsky, N. (2014) circBase: a database for circular RNAs. RNA, 20, 1666–1670
40. Yang, J. H., Li, J. H., Jiang, S., Zhou, H. and Qu, L. H. (2013) ChIPBase: a database for decoding the transcriptional regulation of long non-coding RNA and microRNA genes from ChIP-Seq data. Nucleic Acids Res., 41, D177–D187
41. Wang, Q., Huang, J., Sun, H., Liu, J., Wang, J., Wang, Q., Qin, Q., Mei, S., Zhao, C., Yang, X., *et al.* (2014) CR Cistrome: a ChIP-Seq database for chromatin regulators and histone modification linkages in human and mouse. Nucleic Acids Res., 42, D450–D458
42. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. (2000) The Protein Data Bank. Nucleic Acids Res., 28, 235–242

43. The UniProt Consortium. (2008) The universal protein resource (UniProt). Nucleic Acids Res., 36, D190–D195
44. von Mering, C., Jensen, L. J., Kuhn, M., Chaffron, S., Doerks, T., Krüger, B., Snel, B. and Bork, P. (2007) STRING 7—recent developments in the integration and prediction of protein interactions. Nucleic Acids Res., 35, D358–D362
45. Zhang, X., Zhao, X. M., He, K., Lu, L., Cao, Y., Liu, J., Hao, J. K., Liu, Z. P. and Chen, L. (2012) Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. Bioinformatics, 28, 98–104
46. Zhang, B. and Horvath, S. (2005) A general framework for weighted gene co-expression network analysis. Stat. Appl. Genet. Mol. Biol., 4, Article17
47. Meyer, P. E., Lafitte, F. and Bontempi, G. (2008) *minet*: a R/Bioconductor package for inferring large transcriptional networks using mutual information. BMC Bioinformatics, 9, 461
48. Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R. and Califano, A. (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics, 7, S7
49. Wilczyński, B. and Dojer, N. (2009) BNFinder: exact and efficient method for learning Bayesian networks. Bioinformatics, 25, 286–287
50. Scutari, M. (2010) Learning Bayesian Networks with the bnlearn R Package. J. Stat. Softw., 35, 1–22
51. Shmulevich, I., Dougherty, E. R., Kim, S. and Zhang, W. (2002) Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks. Bioinformatics, 18, 261–274
52. Müssel, C., Hopfensitz, M. and Kestler, H. A. (2010) BoolNe—an R package for generation, reconstruction and analysis of Boolean networks. Bioinformatics, 26, 1378–1380
53. Schaffter, T., Marbach, D. and Floreano, D. (2011) GeneNetWeaver: *in silico* benchmark generation and performance profiling of network inference methods. Bioinformatics, 27, 2263–2270
54. Bonneau, R., Reiss, D. J., Shannon, P., Facciotti, M., Hood, L., Baliga, N. S. and Thorsson, V. (2006) The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets *de novo*. Genome Biol., 7, R36
55. Liu, Z. P., Zhang, W., Horimoto, K. and Chen, L. (2013) Gaussian graphical model for identifying significantly responsive regulatory networks from time course high-throughput data. IET Syst. Biol., 7, 143–152
56. Liu, Z. P., Wu, H., Zhu, J. and Miao, H. (2014) Systematic identification of transcriptional and post-transcriptional regulations in human respiratory epithelial cells during influenza A virus infection. BMC Bioinformatics, 15, 336
57. Haury, A. C., Mordelet, F., Vera-Licona, P. and Vert, J. P. (2012) TIGRESS: trustful inference of gene regulation using stability selection. BMC Syst. Biol., 6, 145
58. 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
59. Langfelder, P. and Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics, 9, 559
60. Liu, Z. P. (2017) Quantifying gene regulatory relationships with association measures: a comparative study. Front. Genet., 8, 96
61. Basso, K., Margolin, A. A., Stolovitzky, G., Klein, U., Dalla-Favera, R. and Califano, A. (2005) Reverse engineering of regulatory networks in human B cells. Nat. Genet., 37, 382–390
62. Friedman, N. (2004) Inferring cellular networks using probabilistic graphical models. Science, 303, 799–805
63. Zou, M. and Conzen, S. D. (2005) A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. Bioinformatics, 21, 71–79
64. Amit, I., Garber, M., Chevrier, N., Leite, A. P., Donner, Y., Eisenhaure, T., Guttman, M., Grenier, J. K., Li, W., Zuk, O., *et al.* (2009) Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. Science, 326, 257–263
65. Thomas, R. (1973) Boolean formalization of genetic control circuits. J. Theor. Biol., 42, 563–585
66. Akutsu, T., Miyano, S. and Kuhara, S. (1999) Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. Pac. Symp. Biocomput., 99, 17–28
67. Saito, S., Aburatani, S. and Horimoto, K. (2008) Network evaluation from the consistency of the graph structure with the measured data. BMC Syst. Biol., 2, 84
68. Jordan, M. I. and Mitchell, T. M. (2015) Machine learning: trends, perspectives, and prospects. Science, 349, 255–260
69. Marbach, D., Roy, S., Ay, F., Meyer, P. E., Candeias, R., Kahveci, T., Bristow, C. A. and Kellis, M. (2012) Predictive regulatory models in *Drosophila melanogaster* by integrative inference of transcriptional networks. Genome Res., 22, 1334–1349
70. Bartel, D. P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. Cell, 116, 281–297
71. Pefanis, E., Wang, J., Rothschild, G., Lim, J., Kazadi, D., Sun, J., Federation, A., Chao, J., Elliott, O., Liu, Z. P., *et al.* (2015) RNA exosome-regulated long non-coding RNA transcription controls super-enhancer activity. Cell, 161, 774–789
72. Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., Maier, L., Mackowiak, S. D., Gregersen, L. H., Munschauer, M., *et al.* (2013) Circular RNAs are a large class of animal RNAs with regulatory potency. Nature, 495, 333–338
73. Garber, M., Grabherr, M. G., Guttman, M. and Trapnell, C. (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. Nat. Methods, 8, 469–477
74. Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U. and Speed, T. P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics, 4, 249–264
75. Elowitz, M. B., Levine, A. J., Siggia, E. D. and Swain, P. S. (2002) Stochastic gene expression in a single cell. Science, 297, 1183–1186

76. Gibcus, J. H. and Dekker, J. (2012) The context of gene expression regulation. F1000 Biol. Rep., 4, 8
77. Ideker, T., Dutkowski, J. and Hood, L. (2011) Boosting signal-to-noise in complex biology: prior knowledge is power. Cell, 144, 860–863
78. de la Fuente, A., Bing, N., Hoeschele, I. and Mendes, P. (2004) Discovery of meaningful associations in genomic data using partial correlation coefficients. Bioinformatics, 20, 3565–3574
79. Zheng, G., Xu, Y., Zhang, X., Liu, Z. P., Wang, Z., Chen, L. and Zhu, X. G. (2016) CMIP: a software package capable of reconstructing genome-wide regulatory networks using gene expression data. BMC Bioinformatics, 17, 535
80. Burchard, J., Zhang, C., Liu, A. M., Poon, R. T., Lee, N. P., Wong, K. F., Sham, P. C., Lam, B. Y., Ferguson, M. D., Tokiwa, G., *et al.* (2010) microRNA-122 as a regulator of mitochondrial metabolic gene network in hepatocellular carcinoma. Mol. Syst. Biol., 6, 402
81. Liu, Z. P. (2014) Systematic identification of local structure binding motifs in protein-RNA recognition. In Proceedings of 8th International Conference on Systems Biology, pp. 74–80
82. Cheng, C., Yan, K. K., Hwang, W., Qian, J., Bhardwaj, N., Rozowsky, J., Lu, Z. J., Niu, W., Alves, P., Kato, M., *et al.* (2011) Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data. PLoS Comput. Biol., 7, e1002190
83. The ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. Nature, 489, 57–74
84. Amaral, P. P., Dinger, M. E., Mercer, T. R. and Mattick, J. S. (2008) The eukaryotic genome as an RNA machine. Science, 319, 1787–1789
85. Spitz, F. and Furlong, E. E. (2012) Transcription factors: from enhancer binding to developmental control. Nat. Rev. Genet., 13, 613–626
86. Hecker, M., Lambeck, S., Toepfer, S., van Someren, E. and Guthke, R. (2009) Gene regulatory network inference: data integration in dynamic models-a review. Biosystems, 96, 86–103
87. Jensen, S. T., Chen, G. and Stoeckert, C. J. Jr (2007) Bayesian variable selection and data integration for biological regulatory networks. Ann. Appl. Stat., 1, 612–633
88. Yeung, M. K., Tegnér, J. and Collins, J. J. (2002) Reverse engineering gene networks using singular value decomposition and robust regression. Proc. Natl. Acad. Sci. USA, 99, 6163–6168
89. Tegner, J., Yeung, M. K., Hasty, J. and Collins, J. J. (2003) Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. Proc. Natl. Acad. Sci. USA, 100, 5944–5949
90. Lam, K. Y., Westrick, Z. M., Müller, C. L., Christiaen, L. and Bonneau, R. (2016) Fused regression for multi-source gene regulatory network inference. PLoS Comput. Biol., 12, e1005157
91. 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, Article15
92. Zhu, J., Zhang, B., Smith, E. N., Drees, B., Brem, R. B., Kruglyak, L., Bumgarner, R. E. and Schadt, E. E. (2008) Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. Nat. Genet., 40, 854–861
93. Santra, T. (2014) A Bayesian framework that integrates heterogeneous data for inferring gene regulatory networks. Front. Bioeng. Biotechnol., 2, 13
94. De Smet, R. and Marchal, K. (2010) Advantages and limitations of current network inference methods. Nat. Rev. Microbiol., 8, 717–729
95. Mordelet, F. and Vert, J. P. (2008) SIRENE: supervised inference of regulatory networks. Bioinformatics, 24, i76–i82
96. Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., Cahill, D. P., Nahed, B. V., Curry, W. T., Martuza, R. L., *et al.* (2014) Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science, 344, 1396–1401
97. Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., *et al.* (2012) Landscape of transcription in human cells. Nature, 489, 101–108
98. Rosenfeld, N., Young, J. W., Alon, U., Swain, P. S. and Elowitz, M. B. (2005) Gene regulation at the single-cell level. Science, 307, 1962–1965
99. Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., Allison, K. R., Kellis, M., Collins, J. J. and Stolovitzky, G., *et al.* (2012) Wisdom of crowds for robust gene network inference. Nat. Methods, 9, 796–804
100. Moignard, V., Woodhouse, S., Haghverdi, L., Lilly, A. J., Tanaka, Y., Wilkinson, A. C., Buettner, F., Macaulay, I. C., Jawaid, W., Diamanti, E., *et al.* (2015) Decoding the regulatory network of early blood development from single-cell gene expression measurements. Nat. Biotechnol., 33, 269–276
101. Graham, J. E., Marians, K. J. and Kowalczykowski, S. C. (2017) Independent and stochastic action of DNA polymerases in the replisome. Cell, 169, 1201–1213