A Study of Transcriptional Regulatory Network Reconstruction

Chen Su

Department of Electrical and Computer Engineering
McGill University
Montreal, Quebec, Canada

Supervisor
Dr. Amin Emad

December 2021

A thesis submitted to McGill University in partial fulfillment of
the requirements of the degree of Master of Science

Acknowledgments

First and foremost, I would like to acknowledge the excellent guidance and mentorship in completing the master's degree from my supervisor, Dr. Amin Emad. We went over hundreds of slides and discussed uncountable numbers of ideas in our weekly meetings.

I have been fortunate to receive precious comments and guidance from our collaborators, Dr. Simon Rousseau and Dr. William Pastor. Those conversations trained me on critical scientific thinking and the ability to communicate with people from different backgrounds. Thank you.

I would like to thank Dr. Larry Hughes, Dr. Ghada Koleilat and Dr. Jean-François Bousquet for their invaluable support, encouragement, and mentorship through my undergraduate studies.

I would like to thank all the past and the current lab members of Dr. Emad’s COMBINE lab. Having them provide thoughtful feedback during my presentations helped me improve my work and presentation skills.

I would like to thank all the faculty members of the ECE department for their help in my course work, and I am genuinely grateful to the ECE and GPS office staff for their prompt assistance during my master’s studies at McGill University.

Last but most importantly, I would like to thank my family for always being there for me. I would like to especially thank my maternal grandparents and my mother for raising me as a person with honesty, curiosity, and always believing in my possibilities.

Abstract

Proteins are functional molecules in a cell with important roles in the structure, function, and regulation of different biological processes. Moreover, protein-coding genes contain the information necessary for producing different proteins. The process in which information from a gene is used to produce proteins is known as gene expression. Gene expression is highly regulated in a cell, and one of the primary regulators of this process are proteins known as transcription factors (TFs). Transcription factors bind to specific regions of DNA and regulate the expression of their target genes. Transcriptional regulatory networks (TRNs) are graph representations of the relationship between transcription factors and their target genes. The reconstruction of these networks based on different molecular features that capture the regulatory evidence of transcription factors and genes is a major area of research since it elucidates the regulatory mechanisms in a cell that play important roles in diseased and healthy states of a cell.

The rapid advances in high throughput sequencing technologies in the past decades that enable measuring different molecular 'omics' profiles of different samples have provided a great opportunity to unravel TRNs. Consequently, utilizing different 'omics' datasets pertaining to the same set of samples through careful data integration, known as multi-omics analysis, allows us to better and more accurately capture the regulatory relationships between TFs and their target genes. In addition, methods that enable the identification of TRNs that are related to specific biological processes or phenotypes, known as phenotype-relevant TRNs, are of great interest. This thesis focuses on the reconstruction of phenotype-relevant TRNs in two different applications and the development of methods for the reconstruction of such networks using multi-omics datasets.

After introducing the objectives of the research and thesis organizations in Chapter 1, we introduce some of the key concepts in molecular biology. Then we review the rapidly evolving approaches for reconstructing TRNs, particularly the phenotype-relevant TRNs in Chapter 2. In Chapter 3, we discuss the application of a computational pipeline that first builds a TRN associated with the response of the host to infection by SARS-CoV-2, when compared against other respiratory viruses, and then utilizes random-walk based methods to identify kinases as potential regulators of this network, as well as potential therapeutic targets. The journal manuscript of this work is currently under review as of August 2021. In Chapter 4, we introduce a novel computational method based on probabilistic graphical models that integrates ChIP-seq data, RNA-seq data, and phenotypic

data to reconstruct phenotype-relevant TRNs. Following the latter, we show the benefit of this approach compared to single-omics analysis using synthetic data. Last but not least, we use this method to study the TRNs involved in human embryonic development.

Abrégé

Les protéines sont des molécules fonctionnelles dans une cellule avec des rôles importants dans la structure, la fonction et la régulation de différents processus biologiques. De plus, les gènes codants pour les protéines contiennent les informations nécessaires à la production de différentes protéines. Le processus dans lequel l'information d'un gène est utilisée pour produire des protéines est connu sous le nom d'expression génétique. L'expression génétique est hautement régulée dans une cellule et l'un des principaux régulateurs de ce processus sont des protéines connues sous le nom de facteurs de transcription (FTs). Les facteurs de transcription se lient à des régions spécifiques de l'ADN et régulent l'expression de leurs gènes cibles. Les réseaux de régulation transcriptionnelle (RRTs) sont des représentations graphiques de la relation entre les facteurs de transcription et leurs gènes cibles. La reconstruction de ces réseaux basée sur différentes caractéristiques moléculaires, qui capturent les preuves régulatrices des facteurs de transcription et des gènes, est un domaine de recherche majeur, car elle élucide les mécanismes de régulation dans une cellule qui jouent un rôle important dans les états pathologiques et sains d'une cellule.

Les progrès rapides des technologies de séquençage à haut débit au cours des dernières décennies, qui permettent de mesurer différents profils moléculaires « omiques » de différents échantillons, ont fourni une grande occasion de démêler les RRTs. Par cons é quent, l'utilisation de différents ensembles de données « omiques » appartenant au même ensemble d'échantillons grâce à une intégration minutieuse des données, connue sous le nom d'analyse multi-omique, nous permet de mieux capturer et plus précisément les relations régulatrices entre les FTs et leurs gènes cibles. De plus, les méthodes qui permettent l'identification des RRTs qui sont liées à des processus biologiques ou à des phénotypes spécifiques, connues sous le nom de RRTs pertinents pour le phénotype, sont d'un grand intérêt. Cette thèse se concentre sur la reconstruction de RRTs pertinents pour le phénotype dans deux applications différentes, ainsi que sur le développement de méthodes pour la reconstruction de tels réseaux en utilisant des ensembles de données multi-omiques.

Après avoir présenté les objectifs de la recherche et les organisations de la thèse dans le chapitre 1, nous introduisons certains des concepts clés de la biologie moléculaire. Ensuite, nous passons en revue les approches en évolution rapide pour la reconstruction des RRTs; en particulier les RRTs pertinents pour le phénotype dans le chapitre 2. Dans le chapitre 3, nous discutons de l'application d'un pipeline de calcul qui construit d'abord un RRT associé à la réponse de l'hôte à l'infection par le SRAS-CoV-2, par rapport à d'autres virus respiratoires et utilise ensuite des méthodes basées sur la marche aléatoire pour identifier les kinases en tant que régulateurs potentiels de ce réseau, ainsi que des cibles thérapeutiques potentielles. Le manuscrit du journal de ce travail est en cours de révision en août 2021. Dans le chapitre 4, nous présentons une nouvelle méthode de calcul basée sur des modèles graphiques probabilistes qui intègrent les données ChIP-seq, les données ARN-seq et les données phénotypiques pour reconstruire les RRTs pertinents pour le phénotype. À la suite de cette dernière, nous montrons l'avantage de cette approche par rapport à l'analyse mono-omique, en utilisant des données synthétiques. Enfin, nous utilisons cette méthode pour étudier les RRTs impliqués dans le développement embryonnaire humain.

Contents

List of Figures

List of Tables

List of Abbreviations

AUROC	The Area Under the Receiver Operating Characteristic
kbp	kilobase pair, a unit of length of nucleic acids. (1 kbp = 1000 bp)
ChIP-seq	Chromatin immunoprecipitation followed by sequencing
DE analysis	Differential expression analysis
DEGs	Differentially expressed genes
DNA	Deoxyribonucleic Acid
EdgeR	An R package for empirical analysis of digital gene expression.
EN	Endoderm lineage
FoRWaRD	A computational tool to rank nodes in a heterogeneous network using random walk with restarts.
GSC	Gene Set Characterization
hESCs	Human Embryonic Stem Cells
in vitro	*In vitro* is Latin for "within the glass." It refers to work that is performed outside of a living organism
in vivo	*In vivo* is Latin for "within the living." It refers to work that is performed in a whole, living organism
KnowEnG	Knowledge Engine for Genomics analytical platform
InPheRNo	A computational method to identify "phenotype-relevant" TRNs
Limma	An R package for the analysis of gene expression data arising from microarray or RNA-seq technologies
MCMC	Monte Carlo Markov Chain
PGM	Probabilistic Graphical Model
RNA	Ribonucleic Acid
RNA-seq	A high throughput technique to examine the quantity and sequences of RNA in a sample
RNA-seq Data	Gene expression data
TFBS	Transcription Factor Binding Sites
TRN	Transcriptional Regulatory Network
TSS	Transcription Start Site

1 Introduction

1.1 Motivation

Massive studies in molecular biology have revealed that complex life phenomena result from many gene interactions guided by regulatory networks. Studying gene regulatory networks (GRNs) can reveal the operating mechanism of cells and life processes, provide new ideas to treat complex diseases, and contribute to screening drug targets and developing personalized therapeutic drugs [1]. Transcriptional regulatory networks (TRNs) are one type of GRNs in which nodes represent either transcription factors (TFs) or potential target genes, and edges represent TF-gene interactions (Figure 2.3B). There have been efforts to reconstruct TRNs from genomics data computationally, and various studies have demonstrated the effectiveness of such methods [2-7]. While this thesis focuses on TRNs, various approaches developed for the reconstruction of GRNs are also applicable to the reconstruction of TRNs; as such, we will also cover related methodologies originally developed for GRNs in chapter 2.

1.2 Thesis Organization and Contributions

This thesis records all research work I have done during my master's at McGill University under the supervision of Dr. Emad. The writing and structure of the thesis have significantly benefited from the inputs and comments obtained from Dr. Emad.

The structure of this thesis is outlined as follows:

- In Chapter 2, I present the molecular biology background required for this thesis. I start with basic molecular concepts and then describe each process in the central dogma [8] (Figure 2.2). After that, I introduce multi-omics data and the importance of integrating such data over single omics analysis. Lastly, I describe the GRN inference problem and provide a review of the rapidly growing GRN inference approaches in the past decades, with a primary focus on TRN reconstruction methods. The focus of phenotype-relevant TRN inference methods in chapter 2.4 is directly related to the research presented in the thesis.
- In Chapter 3, I present a computational pipeline for the reconstruction of SARS-CoV-2 relevant TRNs and the identification of therapeutic targets.

    - Dr. Amin Emad and our collaborator Dr. Simon Rousseau designed and conceived the study, and I implemented the computational pipeline and performed the analysis under the supervision of Dr. Emad. The computational pipeline was comprised of three existing methods: InPheRNo, FoRWaRD, KnowEnG, which were previously developed by authors in [7, 9, 10]. Furthermore, Figure 3.1 and Figure 3.2 in this chapter were adapted from the manuscript [11], which were originally created by Dr. Emad.
- In Chapter 4, I develop a new computational tool, InPheRNo-ChIP, to reconstruct phenotype-relevant TRNs using multi-omics data. Next, I show that such data integration improves the TRN reconstruction accuracy compared to using only transcriptomic data. Lastly, I carry out experiments on both synthetic and real-world biological datasets to assess model performance.
    - Dr. Emad supervised the overall conception and design of the work during our weekly one-on-one meeting. Our collaborator, Dr. William Pastor, helped to (1) select appropriate RNA-seq and ChIP-seq data from open-source databases and (2) design preprocessing modules in chapter 4.4. I generated synthetic data and developed two preprocessing pipelines for RNA-seq and ChIP-seq data. I also implemented an algorithm InPheRNo-ChIP and performed the experiments using synthetic and real-world datasets. All these works were done under the supervision of Dr. Emad. Notably, the core of the algorithm was built upon a probabilistic graphical model (PGM), and the fundamental design of that PGM was adapted from the original InPheRNo paper [7].
- In Chapter 5, I provide a summary of each chapter, the main contributions of the thesis, and propose future research directions.

1.3 Submitted Work

Some portions of this thesis are based on the following submitted manuscript [11]. The work has been performed under the supervision of Prof. Amin Emad and in collaboration with Prof. Simon Rousseau.

C. Su, S. Rousseau, and A. Emad, "Identification of COVID-19-relevant transcriptional regulatory networks and associated kinases as potential therapeutic targets," Accepted, *bioRxiv,* 2020.

We submitted the manuscript to the Scientific Reports on July 2021, with a confirmation number 7d9dce15-bcfa-4030-9854-98fc5d4b12d9. It was accepted on December 2nd, 2021.

2 Background and Literature Review

In this chapter, we start with basic concepts in molecular cell biology, such as biochemical molecules, processes in central dogma, and multi-omics data analysis. Next, we review the literature relating to the gene regulatory network reconstruction, with particular emphasis on taking phenotype information into consideration.

2.1 Molecular Biology Background

2.1.1 Biochemical Molecules

Biochemical compounds can be categorized into four major groups: nucleic acids, proteins, carbohydrates, and lipids [12].

Nucleic acids store the genetic information of an organism and can be categorized into two classes: deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). DNA is a double-stranded molecule and is made of deoxyribonucleotides and phosphate groups. A DNA molecule carries hereditary information in most organisms (one exception is RNA virus [13]). RNA is typically single-stranded and is made of ribonucleotides.

Proteins are made up of structural units (i.e., amino acids), which essentially affect the 3-dimensional structures and functions of proteins [14-16]. Proteins include enzymes, antibodies, and other amino acid-derived molecules whereas carbohydrates and lipids include sugars and fats [17, 18].

2.1.2 Genes and Genome

According to the studies, the genetic differences between humans are minuscule; two randomly selected individuals are around 99.9% genetically identical [19]. The remaining 0.1% contains crucial genetic instructions about the variations of cell types, tissues and organs, and the cause of diseases.

Genes are the basic functional units of inheritance; they encode functional products such as proteins. Except for some RNA viruses that use RNA as their genetic material [13, 20], a gene usually refers to a small stretch of the DNA sequence on a chromosome or genome. The genome is defined as a complete set of genetic instructions in a living organism. Researchers have clarified that only 2% of the human genome is coding genes (around 20k-25k genes in total [21]), whereas the last part of the genome consists of non-coding genes with other regulatory functions. Besides the nucleus that carries chromosomal DNA (22 pairs of autosomes pairs and a pair of allosomes), the cell also contains mitochondrial DNA [22].

A gene stores genetic information using different combinations of four DNA nucleotide bases: adenine (A), thymine (T), guanine (G), and cytosine (C) [23]. It contains the essential information for living cells to replicate or express particular phenotypes. During this process, gene mutations may occur in order to adapt to environmental changes.

Despite the broad similarity between eukaryotic and prokaryotic genes [24], phenotypic diversities in the natural world can be primarily attributed to the distinctive structure of the eukaryotic gene [25]. A eukaryotic gene consists of alternated regions: some of which are coding regions, and some are non-coding regions (also known as regulatory regions). The coding regions are sequences that can be expressed as functional RNA or proteins in the cell, while the non-coding regions are the sequences that are needed to express those genes, such as promoters and enhancers. As an example, we present a basic structure of a eukaryotic gene in Figure 2.1. The structure of the gene consists of protein-coding nucleotide sequences (exons), sequences that regulate gene expression before and after the coding region, and nucleotide sequences that do not code for amino acids (introns) [25]. Only approximately 1% of the 3 billion base pairs are protein-coding genes in the human genome, and only about 5% of a protein-coding gene corresponds to coding exons.

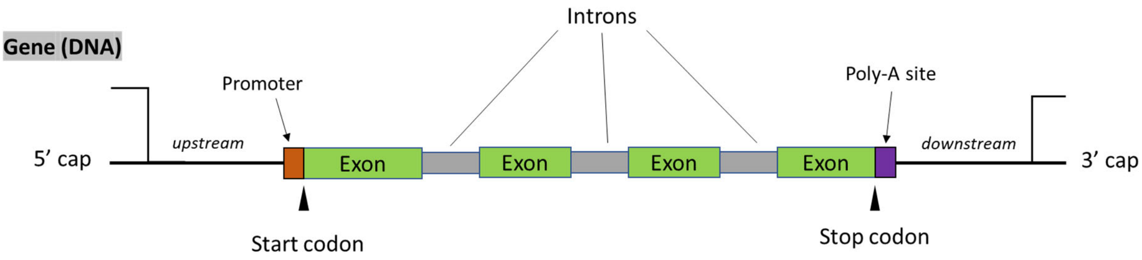


Figure 2.1 Illustration of a eukaryotic gene structure. The promoter region (orange-coloured segment) is most proximal to the start codon and contains the TATA box, the TSS, and the RNA polymerase binding site. Exons (green-coloured segments) are nucleotide sequences in DNA that code for amino acids; introns (grey-coloured segments) are nucleotide sequences in DNA that do not directly code for amino acids but help form different proteins through RNA slicing. A Poly(A) site (purple-coloured segment) is located proximal to the stop codon, which helps mRNA to move into the cytoplasm from the nucleus. This plot was inspired from Figure 1 in [24] and Figure 6 in [25].

2.1.3 DNA Replication

Semi-conservative DNA replication is a process in which each parental DNA strand acts as a template for synthesizing a new complementary DNA strand [26]. The output of DNA replication is two DNA molecules with one parental strand and one new daughter strand.

2.1.4 RNA Synthesis

Based on central dogma (Figure 2.2), there are two types of RNA synthesis. One is DNA-dependent RNA synthesis, also known as transcription. The other type is RNA-dependent RNA synthesis, also known as RNA replication.

The DNA-dependent RNA synthesis is a process in which a DNA sequence transcribes into an RNA using a DNA-dependent RNA polymerase enzyme [27]. Gene regulation, in principle, can occur at any of the stages in gene expression, but most genes are regulated primarily at the transcriptional level. Potential processing and truncation errors in RNA synthesis may lead to severe diseases [28, 29]. Thus, regulation of the RNA transcription process is essential for understanding many biological phenomena and medical problems.

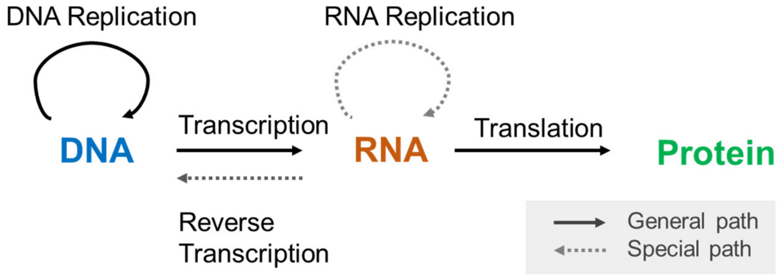


Figure 2.2 The central dogma in molecular biology. It presents the transfer of genetic information in cells among DNA, mRNA and protein [8].

The RNA-dependent RNA synthesis is catalyzed by RNA-dependent RNA polymerase and is commonly found in viruses [30, 31]. It is a unique process reserved only for RNA viruses (other than retroviruses [32]), which uses single-stranded RNA as a template to synthesize RNA in host cells.

2.1.5 Protein Synthesis

Proteins are complex molecules that carry out most biological activities (e.g., growth, reproduction, and movement); the accurate synthesis and folding of proteins are critical to the proper functioning of cells and organisms [33]. In the protein synthesis stage, cells use a ribosome machinery to translate the genetic information in messenger Ribonucleic Acid (mRNA) molecules into protein molecules [34]. This dynamic process comprises three phases: initiation, elongation and termination [35]. Proteins resulting from this process become main components of all tissues involved in biological processes such as embryonic development, metabolism, and signal transduction [36].

2.1.6 Gene Expression

Gene expression is a process describing how genetic information flows from DNA to protein through RNA transcription within a biological system (Figure 2.2). This process is composed of three stages, as mentioned earlier: replication, transcription, and translation.

The gene expression pattern is subject to temporal and spatial specificity [37-40]. It is an essential step in biological evolution: the increasingly complex genomes have a more complex and refined gene expression pattern.

Temporal-specificity refers to the fact that gene expression only occurs in a specific time frame. For example, a certain stage of infection will occur after a virus infects a host [41]. As the infection stage develops, gene expression changes in pathogens and hosts will occur: some genes get turned on, and some genes get turned off. The development process, from fertilized egg cells to multicellular organisms, is another example of temporal-specificity. Different genes are turned on or off strictly in their specific chronological orders at each stage of differentiation and development [42].

The spatial specificity refers to different expressions of the same gene in different tissues and organs at a particular stage of development in a multicellular organism. The difference between organs, tissues and cells in the same organism can be explored by differential expression analysis [43-46]. The gene expression profile of a cell, that is, the type and intensity of gene expression, determines the differentiation state and function of the cell [47-49].

2.1.7 Regulation of Gene Expression

Gene expression regulation refers to the molecular mechanism by which cells decide which genes will be expressed at a particular body site when internal and external environmental signals stimulate them [50]. It regulates how and when genes are expressed in different cell types and tissues. Since gene expression is a complex process involving many molecules and signalling pathways and determining the characteristics of a living organism, the precise regulation of gene expression is of great significance [51, 52]. Furthermore, it is now well accepted that changes in these regulatory mechanisms (introduced by mutations or other means) are associated with the development of several diseases [53]. For example, invasive ductal carcinoma is one of the common types of breast cancer. There is evidence that this disease is not caused by a single gene in a particular cell, but rather linked to changes in molecular mechanisms of gene regulation [54, 55]. Therefore, it is critical to understand and reconstruct the underlying regulatory networks under different conditions and cell types.

Transcription factors (TFs) are critical players in regulating gene transcription by binding to particular DNA sequences located in or close to their target genes, thereby controlling the expression of those genes (e.g., by promoting or repressing a gene to be transcribed into an RNA) [56]. A transcription factor can have multiple target genes, and a gene can be regulated by multiple transcription factors in combination (though in most cases, the number of TFs regulating a gene is small) [57]. These complex regulatory interactions between TFs and their targets can be conceptualized as a network, known as a transcriptional regulatory network (TRN). TRNs are directed graphs where each node is a transcription factor (TF) or a gene, and each TF-gene edge can be interpreted as regulatory relationship between a TF and its target gene (Figure 2.3B).

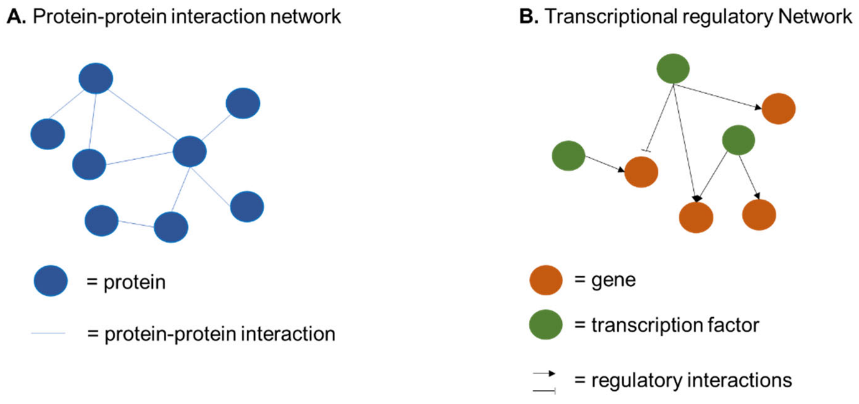

A. Protein-protein interaction network
B. Transcriptional regulatory Network
= protein
= protein-protein interaction
= gene
= transcription factor
= regulatory interactions

Figure 2.3 Examples of some common networks.a protein-protein interaction (PPI) network (B) a transcriptional regulatory network (TRN). (A) In a PPI network, nodes represent proteins, and the undirected edges show the physical protein-protein interactions between nodes [58]. (B) In a TRN, nodes represent the expression levels of genes/TFs, and the directed edges represent the causality between network nodes [59].

2.2 Multi-omics Data

Different biological and physiological processes involve different levels of biomolecules participating or having different natures of connectivity, such as objects from the epigenome, genome, proteome and metabolome [60, 61].

2.2.1 Introduction

Genomics is one of the established fields of life science. The term refers to the collective characterization and quantitative research of all genes in an organism and the comparison studies between genes [61]. Focusing on the whole genome provides the opportunity to decipher genetic information and study relations between different genes. Genes are transcribed and translated to produce proteins, which are closely related to various biochemical processes in cells that in turn affect human bodies. Therefore, there is rising attention on proteomics after genomics [62]: **proteomics** studies protein expression levels, post-translational modifications, and protein-protein interactions (PPIs) [63, 64]. A typical PPI network is shown in Figure 2.3A.

However, genomics and proteomics alone are not adequate to solve all the puzzles we have for human life. For example, the same genotype may exhibit unique characteristics due to both genetic and environmental factors, and mutations may explain the occurrence of a disease in some regions of the genome or coding errors during the translation [65]. This brings in other omics science: **Transcriptomics** studies the transcription of the whole genome and the regulation of transcription [66]; **Epigenomics** is a holistic study of the modification characteristics of genomic DNA or DNA-binding proteins [67]. Researchers can elucidate transcriptional networks by integrating gene expression and epigenetic markers/transcription factors (Figure 2.3B). **Metabolomics** is a quantitative analysis of all metabolites (e.g., amino acids, fatty acids, or carbohydrates) in biological systems (e.g., tissues, cells, or organisms) and characterizes the interactions of those with potential diseases or environments [68, 69].

2.2.2 Data Integration

Recent developments in high-throughput technologies and multi-omics databases (e.g., TCGA [70], GTRD [71], LinkedOmics [72] and Aging Atlas [73]) have led to a growing trend towards data integration. Subramanian et al. [74] provided a comprehensive review of multi-omics data integration techniques and suggested that such integration has great potential in identifying key elements in the pre-clinical research. In review papers [75, 76], researchers discussed the challenges and the necessities of developing integrative approaches to decipher regulatory mechanisms further.

To date, various studies have focused on integrating multi-omics data to model biological processes from different perspectives [77-84]. For instance, Hirai et al. [77] integrated gene expression profiles and nontargeted metabolite profiles to form gene-to-metabolite networks and depict global regulatory mechanisms under sulphur deficiency and stress response in Arabidopsis. Pirhaji et al. [81] constructed a network that integrates protein and metabolites interactions to infer disease-associated pathways and putative metabolites. They applied their model to conditionally immortalize striatal cell lines (STHdh) of Huntington's disease (HD) and identified known and novel vital metabolites. Xie et al. [82] combined DNA methylation signatures from several human tissues and corresponding gene expression profiles to identify unique signatures during cell development and tissue differentiation. As diseases are often associated with changes in gene expression, Bouchal et al. [84] performed an integrative analysis of proteomics and transcriptomics and identified eight pivotal therapeutic targets of low-grade breast cancer.

Overall, the incorporation of information from multi-omics data provides a better insight into complex biological systems and accelerates the understanding in regards to the development of specific diseases [76].

2.3 Gene Regulatory Network Reconstruction

Inferring more reliable and accurate GRNs from experimental transcriptomic data (alone or in combination with data from other sources, such as ChIP-seq and ATAC-seq data) is a key problem in the field of computational biology. Much progress has been made to tackle such a problem [85]. The available GRN reconstruction approaches are mainly designed for microarray and bulk RNA-seq data, and they can be categorized into four subsets: correlation-based methods, information-theoretic approaches, model-based methods, and regression analyses.

2.3.1 Correlation-based Methods

Correlation-based networks use different definitions of “correlation” between molecular features of genes in different conditions to determine gene dependencies. One of the earliest correlation measures used is the Pearson correlation coefficients, but other correlation measures such as

Spearman correlation coefficients or weighted correlation coefficients have also been used widely. For example, Chen et al. [86] assumed that the correlation between genes satisfies the scale-free distribution of the network and then used a weighting coefficient to re-predict the genetic correlation network by weighting the regulatory relationship between biological genes. Since one gene rarely operates in isolation from other genes [2][56], a major drawback of this method is its inability to discriminate between direct and indirect connections in the biological network. For example, if gene A is correlated with gene B and gene B is correlated with gene C, then the correlation-based method will likely draw a problematic conclusion that there is a direct dependency between A and C. Some well-known methods in addressing this problem are partial correlations [16], elastic net [87], lasso (an L1-penalized estimation of the inverse covariance matrix) [88], and information-theoretic methods [89].

2.3.2 Information-theoretic Methods

Correctly discriminating indirect connections from direct connections is essential for reducing false-positive predictions. To achieve this, sophisticated information-theoretic methods, such as Bayes conditional probability tables or mutual information, have been introduced to consider the effects of non-linearity. Information theory offers robust statistical mechanics for inferring network links from between-gene correlation measures. Some of the most reviewed and discussed methods include the Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNe) [90], Context Likelihood of Relatedness (CLR) [91], minimum redundancy maximum relevance concept (MRNET) [89] and Conservative Causal Core Network (C3NET) [92]. These algorithms take gene expression profiles as inputs, estimate pair-wise mutual information (MI) values, and filter statistically significant edges to construct the network. A detailed discussion of the functioning of these methods can be found in the review papers [89, 93-95].

The ARACNE algorithm [90] provides a conservative network estimation through the constraint of data processing inequality (DPI). If the data processing imbalance exceeds a certain threshold, ARACNE evaluates all possible edges of each triplet of genes and eliminates edges with the lowest significance values corresponding to the lowest mutual information value. One limitation of this method is that it does not allow adding new edges to the network.

Likewise, the CLR algorithm [91] takes the often sparse nature of biological networks into account and assumes most estimated MI values are insignificant. The method first computes MI values between the expression level of each TF and that of the target gene. Then, CLR compares the MI estimates of each TF-gene pair to a complete graph (i.e., a background correction of all MI values involving either the regulator or its target) and eliminates false connections. The output of this method is a network containing transcriptional regulatory interactions. However, its heavy reliance on the MI matrix prevents it from resolving a problem of causal inference from time-series data.

The MRNET algorithm [89] incorporates the maximum relevance/minimum redundancy (MRMR) feature selection method [96] to choose important edges and infer genetic interaction networks. The algorithm first considers each gene as a target gene and the remaining genes as putative regulators and then relies on a feature selection method (i.e., MRMR) to select the optimal subset of regulators. Although this method can infer the network efficiently, it suffers from high false positive rates, indicating that there may still undetected indirect regulatory interactions in the network.

The C3NET algorithm [92] is a two-step algorithm designed for non-directional network inference, as it employs MI values as test statistics among genes. It first eliminates non-significant MI links in the adjacency matrix, then selects for each gene the neighbour edge with the maximum MI value in the remaining matrix.

Since a major part of transcriptional regulation is performed by complexes of TFs and other proteins [97], models that depend only on pairwise interactions between genes are not powerful enough to capture important characteristics of regulatory relationships, such as directionality and types. In addition, employing MI scores limits the maximum possible number of inferred edges for each gene to the number of genes under consideration.

2.3.3 Model-based Methods

Model-based methods can be divided into several categories: ordinary differential equation (ODE), linear programming (LP) models, Boolean network, and probabilistic graph models (PGMs). PGMs are comprised of Bayesian network (BN) and Gaussian graphical models.

Ordinary differential equations (ODE) methods are commonly seen in applying time-series gene expression data. These methods are used to fully discover transcription generation rates, half-life periods and degradation rates. The ODE method is suitable for establishing nonlinear relationship models showing a wide range of dynamic behaviours [98]. Introducing constraints in the form of prior knowledge from known parameters or network structure is hugely beneficial for ODE-based methods [99]. An example of this method is called Inferelator [100]. Inferelator is a method imposing LASSO penalty on the kinetic parameters associated with regulatory interactions to infer the network. It is originally based on the inference of explaining independent ODEs under several experimental disturbances. The advantage of this method is that it learns the topology of networks containing TF-gene interactions and is capable of learning the rate of change in the transcription. More recently, Yang et al. proposed a complex-valued ODE model (CVODE) to identify the regulations among genes for GRN inference, where model coefficients and functions are all complex-valued [101]. The authors used Grammar-guided genetic programming (an improved variant of genetic programming) to evolve the structure of the ODE model and complex-valued firefly algorithm [102] to search for the optimal parameters. The model was tested on real-world gene expression profiles, and the result shows that CVODE has a better prediction accuracy and is more robust than the original ODE method.

Boolean network model-based methods were popularized in reconstructing gene regulatory network (GRN) by Stuart Kauffman in 2003 [103]. It aims to model the GRN as a discrete dynamic system with two states: ON and OFF: the ON/active state implies that a gene will be expressed and transcribed to generate biochemical material, while the OFF/inactive state implies that a gene will not be transcribed, and therefore no biochemical material can be produced. The changes in those two states can be described and periodically updated by Boolean functions with a few parameters [104]. The dynamic characteristics and simplicity of the model allow it to be widely applicable in modelling various genetic regulatory networks, such as the Mammalian cortical area development [105], p53-induced cell fate mechanisms [106] and CD 4+ T cell differentiation [107].

However, this type of Boolean network method is not free of limitations. The assumption of binary states is often considered an over-simplistic representation of the actual biological network [108]. To address this problem, Shmulevich et al. proposed a stochastic extension of the Boolean network, Probabilistic Boolean Networks (PBNs), which generalized Boolean network by adding a probabilistic selection of parental gene sets to the framework and integrating rule-based dependencies between genes. This method systematically studies global network dynamics and is able to work with data uncertainty and model selection. Zhao et al. adopted the PBNs framework and developed a method to determine the direct dependencies and regulation orientations of genes in time-series datasets [109].

Due to regulatory networks' complexity and uncertainty, neither differential model equations nor Boolean networks can easily handle real-world problems, such as missing data or randomness/noise [110]. To overcome such technical problems, Friedman et al. [111] proposed a new gene regulatory network model on top of Bayesian networks (BNs), and Murphy et al. [112] subsequently investigated a temporal variation of BNs (i.e., dynamic Bayesian networks) to enhance its ability to discover complex associations. The method translates probabilistic dependencies among random variables into a directed acyclic graph (DAG). In a DAG, vertices represent random variables, and directed edges represent conditional dependencies between any two variables. Also, no self-feedback loop in a DAG indicates that a variable cannot be its own ancestor or descendant. According to Markov's assumption on the conditional probabilities of each node in a BN, the joint probability can be written as follows:

$$P(X_1, X_2, \dots, X_n) = \prod_{X_i \in X} P(X_i \mid Parent(X_i)) \tag{2.1}$$

In Bayesian network structure learning, the most likely graph structure, *G*, is generally searched through the Bayesian scoring strategy [113]. However, as the number of gene nodes increases, the network structure graph predicted by the model increases exponentially. Thus, constructing a network graph structure for larger-scale networks is often time-consuming and determining the best graph that explains associations with the observed data is an NP-hard problem [114]. Several algorithms have been drawn up in order to address the issue of time complexity, such as simulated annealing [115, 116], Markov chain Monte Carlo search [117], max-min hill-climbing algorithm [118] and parallel algorithm proposed by Madsen et al. [119].

Gaussian graphical models (GGMs) are another popular method used throughout computational biology (and other fields) to characterize statistical relationships between variables. In the context of GRN inference, the underlying assumption of GGM is that the processed gene expression data follows a standard Gaussian distribution $X \sim (\mu, \sigma)$, where μ and σ represent the mean and variance of the data [120]. GGMs first fit the multivariate Gaussian distribution and then construct the accuracy matrix (i.e., the inverse of the covariance matrix) [121] to obtain a gene-gene interaction network. A differential network can be established based on the gene networks in two states to find key genes. However, such methods often fail to update beliefs and make new predictions in light of new observations. Furthermore, with the pronounced shift in technology from microarrays to RNA-seq profiles, a necessary change in how we should model gene expression patterns has occurred (where normal distribution is no longer suitable) [122]. Using RNA-seq gene expression datasets from the Cancer Genome Atlas (TCGA), Zhao et al. utilized a Gaussian graphical model to perform cancer genetic network inference and uncover hidden gene interactions across 15 specific types of human cancer [120]. In the study, the authors performed Ryan-Joiner (RJ) statistic test on each individual gene to ensure their log-transformed gene expression profiles are normally distributed.

2.3.4 Linear Regression Methods

Linear regression is a statistical term used to find the best-fitting hyper-plane through observed data points, which models the relationship between one dependent variable and two or more independent variables [123]. While this approach is widely applied to different scenarios, since the sample size is often smaller than the feature size for RNA-seq experiments, using this method may lead to an over-fitting problem.

Feature selection methods provide practical solutions in handling over-fitting problems in GRN reconstruction. Several methods have been proposed in this area, such as stepwise regression [124], least-angle regression (LARS) [125, 126], LASSO regression [127, 128], and partial least squares regression [129, 130]. For example, TIGRESS [131] is a method that introduces a stable selection mechanism and uses the least angle regression method for feature selection. Considering the complementarity of different models, the NIMEFI method [132] assembled various models based on the feature selection framework and trained them uniformly.

2.3.5 Discussion

In this chapter, we have covered common GRN reconstruction approaches for microarray or bulk RNA-seq data. While it is outside the scope of this thesis, it is worth mentioning that several studies have pointed out that some of these state-of-the-art methods, such as TIGRESS and MI, have poor performance on single-cell transcriptomic data [133, 134]. Furthermore, recent work in the field has focused on developing computational methods for dynamic GRN inference from single-cell transcriptomics alone (PIDC [135], GRISLI [133]) or from both time series and steady-state expression data (dynGENIE3 [136]).

2.4 Phenotype-relevant Regulatory Network Reconstruction

With the rapid development of high-throughput sequencing technologies, molecular phenotypes of clinical samples can now be conceived in terms of gene expression levels [137, 138]. Nowadays, more attention is being directed to reconstructing regulatory networks correlated with various quantitative traits (expression phenotypes) of the clinical samples. These phenotypic traits can be either categorical or continuous, depending on the problem setup.[139]. For example, a discrete phenotypic label could indicate the membership of a sample in different studies: whether the sample belongs to the control or case group; whether it is from a tumour or normal tissue; whether it is infected by SARS-CoV-2 or other respiratory viruses. On the other hand, a continuous phenotypic value could be a relative change in continuous manifestations of the disease.

Since the thesis focuses on reconstructing phenotype-relevant TRNs, we collect computational approaches developed to include phenotypic information in the network reconstruction and categorize them into three classes (chapter 2.4.1 - chapter 2.4.3). Next, we discuss InPhRNo, an essential building block for research projects (chapter 3 and 4) in chapter 2.4.4. Lastly, we discuss the relative strength and weaknesses of those methods in chapter 2.4.5.

2.4.1 Context-restricted Approaches

Most context-restricted approaches restrict the analysis to samples of phenotypic labels, such that only samples with identical phenotypic labels are included in the network [7]. For example, Qin et al. [140] and Streib et al. [141] reported the reconstruction of gene regulatory networks from transcriptomic data from breast cancer tissue samples or cell lines. Glass et al. [142] developed PANDA, which utilized a message-passing approach to integrate evidence from different sources and reconstruct tissue-specific regulatory networks targeting 38 tissues from GTEx database.

2.4.2 Context-specific Approaches

Context-specific approaches [7] often first discover the phenotypic variation in gene expression patterns across different biological conditions using differential expression analysis and then compute the significance of the selected differentially expressed genes (DEGs) through some statistical tests. Lastly, they relate the most significant genes with context specificity to the expression of TFs in order to reconstruct a network. For example, Raza et al. [143] inferred a cancer-specific gene regulatory network from prostate cancer microarray data. In this study, the authors first identified significant DE genes in the disease condition compared to control samples and then applied the Pearson correlation coefficient method to obtain the pair-wise correlation among the identified genes.

2.4.3 Differential Network Analysis Approaches

Differential network analysis (DiNA) approaches focus on detecting topological changes in regulatory networks under different conditions [7, 144, 145]. It examines different biological processes from context-restricted networks, each inferred from gene expression data of a specific group. Given two group-specific networks, general DiNA methods identify edges only present in one group-specific network but not the other and then form a differential network. For example, Okawa et al. [146] proposed a computational DiNA approach and applied it to some binary-fate mouse stem cell systems for lineage specifier predictions. Mall et al. [147] reported a hamming distance-based method designed to capture local topological features by evaluating topological differences between two networks (i.e., IDH-mutant versus IDH-wild-type glioma tumours) examining their statistical significance.

2.4.4 InPheRNo

However, methods discussed in chapters 2.4.1-2.4.3 may not be capable of constructing TRNs that are linked to various phenotypes due to the inefficiency of incorporating phenotypic labels/values of different biological conditions in their analysis. Addressing these deficiencies, Emad and Sinha developed InPheRNo [7], a computational tool that utilizes a probabilistic graphical model (PGM) trained to integrate the summary statistics (i.e. p values) of sample-level phenotypic labels and TF-gene associations and thereby to identify 'phenotype-relevant' TRN.

As inputs, it takes a gene expression matrix of all genes (including TFs), a list of human TFs, and phenotypes of the samples from RNA-seq data. Notably, InPheRNo uses a two-step procedure to measure (1) the combinatorial effects of multiple TFs on each gene and (2) the effect of a limited sample size.

InPheRNo first uses the ElasticNetCV function from statsmodels API library and restricts the number of non-zero coefficients in the model to be not greater than a pre-defined maximum number of candidate TFs, specified as the 'm_{max}' and a positive integer scalar. The use of the pre-defined upper limit m_{max} imposes the prior knowledge that a gene can only be regulated by a handful of TFs rather than all 1.5k TFs. The hyperparameters of the Elastic Net (i.e., alphas) are estimated by iterative fitting along a regularization path with 5-fold cross-validation and a default value p=0.5. Gene-level matrices are formed at the outputs of Elastic Net, where each matrix contains the expression of TFs for a specific gene across the samples. Then, InPheRNo performs a multivariate ordinary least squares (OLS) regression model to express the expression of each gene in terms of a set of independent features (i.e., candidate TFs). The resulting pseudo p values represent the statistical significance of the combinatorial regulatory effects of the candidate TFs on genes of interest.

The p values of associations between TFs and genes, along with the p values of associations between genes and their phenotypes, are put into step 2 of InPheRNo. In step 2, those three sets of p values are used as observations in the PGM to estimate posterior probabilities for relationships within (TF, gene, phenotype) triplets. The alternative hypothesis states that a TF regulates a gene to affect its phenotype. The last step of InPheRNo uses min-max normalization to normalize the averaged posteriors; this normalization step allows the algorithm to be widely applicable for other downstream analyses. The output of InPheRNo is a TF-gene matrix showing the confidence score for each TF-gene edge concerning its phenotype relevance.

To assess the performance of InPheRNo, the authors applied InPheRNo to infer transcriptional regulatory networks for normal/cancer tissue samples in the GTEx data portal [148] in their study. They used these tissue-relevant TRNs to compare the underlying regulatory mechanisms between cancer and normal tissues and understand how phenotype specificity manifests itself globally. By studying the predicted cancer type TF-gene edges in the inferred TRNs and comparing to the TF-gene relationships in the global TRNs reconstructed from the ENCODE’s ChIP-seq data [22], they were able to observe significant enrichments of the overlap for each cancer type and bring important insights into phenotype-relevant gene regulation.

2.4.5 Discussion

Even though various methods have been developed to incorporate phenotype information in network construction, their limitations still exist. As to the first type of approaches, context-restricted approaches can only capture regulatory mechanisms relevant to a specific context (e.g., breast cancer) but cannot embed the variation of the phenotypic values/labels (e.g., tumour grades, disease versus control) in the resulting network. Also, context-specific approaches are highly dependent on the cut-off used for choosing the phenotype relevance of genes (i.e., DEGs); therefore, they cannot utilize the phenotype information to its fullest extent. As for the last type of methods, differential network analysis methods also remedy their ignorance by neglecting the potential phenotype-relevant regulatory edges in the inferred TRN. InPheRNo [7] is an exception, which utilizes PGM to model the association within each possible (TF, gene, phenotype) triple. The use of summary statistics and the sophisticated design of PGM in the InPheRNo paper lays the foundation for future generalization and integrative analyses of other regulatory evidence such as ChIP-seq data (see chapter 4 for details).

3 Identification of TRN Associated with Response of Host Epithelial Cells to SARS-CoV-2

3.1 Introduction

As of the submission of this manuscript [11] (July 2021), the emergence of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has caused 195 million global cases of the severe pneumonia-like disease coronavirus disease 2019 (COVID-19) [149] and effective antiviral medication available for COVID-19 has remained elusive [150]. Therefore, there is a great need to unravel the underlying mechanisms involved in the host immune responses to SARS-CoV-2 infection and predict candidate therapeutic targets for the treatment of COVID-19.

To achieve this goal, we developed an automated computational pipeline that strings together algorithms previously developed in our lab to study the host response to SARS-CoV-2 infection. Through pathway enrichment analysis of identified TFs and their target genes, we demonstrated that our pipeline was able to reliably unravel the underlying mechanisms of TF-gene interactions involved in the corresponding SARS-CoV-2 infections. Moreover, we conducted an extensive literature search on kinases in our prioritized list to validate the prediction of our algorithm. The findings supported the belief that our pipeline can identify known kinase, and therefore the novel (and often less experimental studied) kinases may be used as drug target candidates with potential interest for SARS-CoV-2 treatment.

3.2 Problem Statement

Our first task is to use InPheRNo (discussed in chapter 2.4.4) to identify a transcriptional regulatory network (TRN) that differentiates the response of the host to SARS-CoV2 infection versus other respiratory viruses (denoted by SvOV). The SvOV network is a bipartite directed graph where nodes are either transcription factors (TFs) or genes, and edges are regulatory interactions between TFs and regulated genes. Considering the direct and indirect connections between identified TFs and human kinases, our second task is to identify protein kinase genes and prioritize candidate therapeutic targets for treating SARS-CoV-2 infection. We completed this task by combining the inferred SvOV TRN with other network information (i.e., HumanNet [151]) to form a heterogenous network and then apply FoRWaRD to it.

3.3 Methodology

3.3.1 Data Sources

We downloaded a list of kinase information of human non-pseudogenes from www.kinase.com (Kincat Hsap 08.02; update December 07th) [152]; a list of human TFs from AnimalTFDB 2.0 [153]; as well as a HumanNet 2.0 [151] integrated network of gene-gene interaction from the address https://github.com/KnowEnG/KN_Fetcher/blob/master/Contents.md, on July 28th, 2020. According to KnowEnG's documentation, the integrated HumanNet has 469,784 edges, 15,999 gene nodes, and a density measure of 0.00367. The log-likelihood score in the dataset ranges from 0 (the worst) to 10 (the best).

We downloaded RNA-seq profiles of human lung epithelial cells infected by SARS-CoV-2, respiratory syncytial virus (RSV), human parainfluenza virus type 3 (HPIV3), influenza A virus (IAV), and IAV that lacks the NS1 protein (IAVdNS1) from the GEO database under the accession number GSE147507. This bulk RNA-seq dataset was generated from both virus-infected and mock-infected experiments across various cell lines and animal models, and stored in *.tsv* format [154]. The downloaded matrix contains 110 columns (samples) and 21,797 rows (human genes), and each entry of the matrix is a raw read count for a gene in a given sample. Out of 110 samples, we selected samples corresponding to three cell lines: normal human bronchial epithelial cell lines (NHBE), human lung adenocarcinoma cell lines (A549), and human bronchial/sub-bronchial gland cell lines (Calu-3). These viruses-infected cells were collected at various time points post-infection with different multiplicity of infections (MOIs). Detailed sample information can be found in Table 3.1.

We downloaded the LINCS L1000 dataset from Broad Institute LINCS L1000 Phase I datasets (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE92742). This dataset contained the expression levels of 978 landmark genes induced by short hairpin RNA (shRNA) knockdown in different cell lines and provided five levels of data depending on the different stages of data preprocessing. In this project, we only used its level-5 differential gene expression signatures since this level of processing is often considered more reliable for downstream analysis than other levels [155].

We obtained a list of target genes that support the findings of our study from the GTRD database at the following address:
http://gtrd.biouml.org:8888/downloads/20.06/intervals/target_genes/Homo%20sapiens/genes%20promoter%5b-1000,+100%5d/.

In the GTRD dataset, if a gene's promoter region (defined as the interval [-1000bp, 100bp] relative to its TSS) contains at least one GTRD meta-cluster for a given TF, then this gene was identified as a target of the TF. The meta-cluster was a collection of the binding sites for individual TFs from all experimental conditions, such as tissues, cell lines and treatments.

To form an aggregated kinase-substrate interaction (KSI) network that was only focusing on experimentally determined kinase-substrate interaction pairs, we downloaded three datasets from the following studies/databases: PhosphoSitePlus database [156], PhosphoNetworks database [157] and the supplementary material of an independent study [158]. The aggregated network contained 29,594 kinase-substrate relationships corresponding to 406 unique kinases and 3942 unique substrates

3.3.2 Computational Pipeline

Computational pipelines are joining the advantages of previously developed computational tools (mostly open source) to speed up the process of identifying the hidden relationships in the large-scale experimental data, and provide easy access to non-programmers [159]. Numbers of pipelines have been developed to combat COVID-19 pandemics. For example, V-Pipe [160] performed an analysis of SARS-CoV-2 high-throughput sequencing data. UniProt portal [161] provided early access to SARS-CoV-2 annotated protein sequence, protein information from the same coronavirus family as well as visualization for data scientists. In June 2021, Le et al. [162] applied their existing computational drug repositioning pipeline to differentially expressed genes from three transcriptomics data and identified potential drugs associated with SARS-CoV-2.

In this project, we designed a computational pipeline to identify TRNs associated with the host's response to SARS-CoV-2 infection and identified associated kinases as potential therapeutic targets.

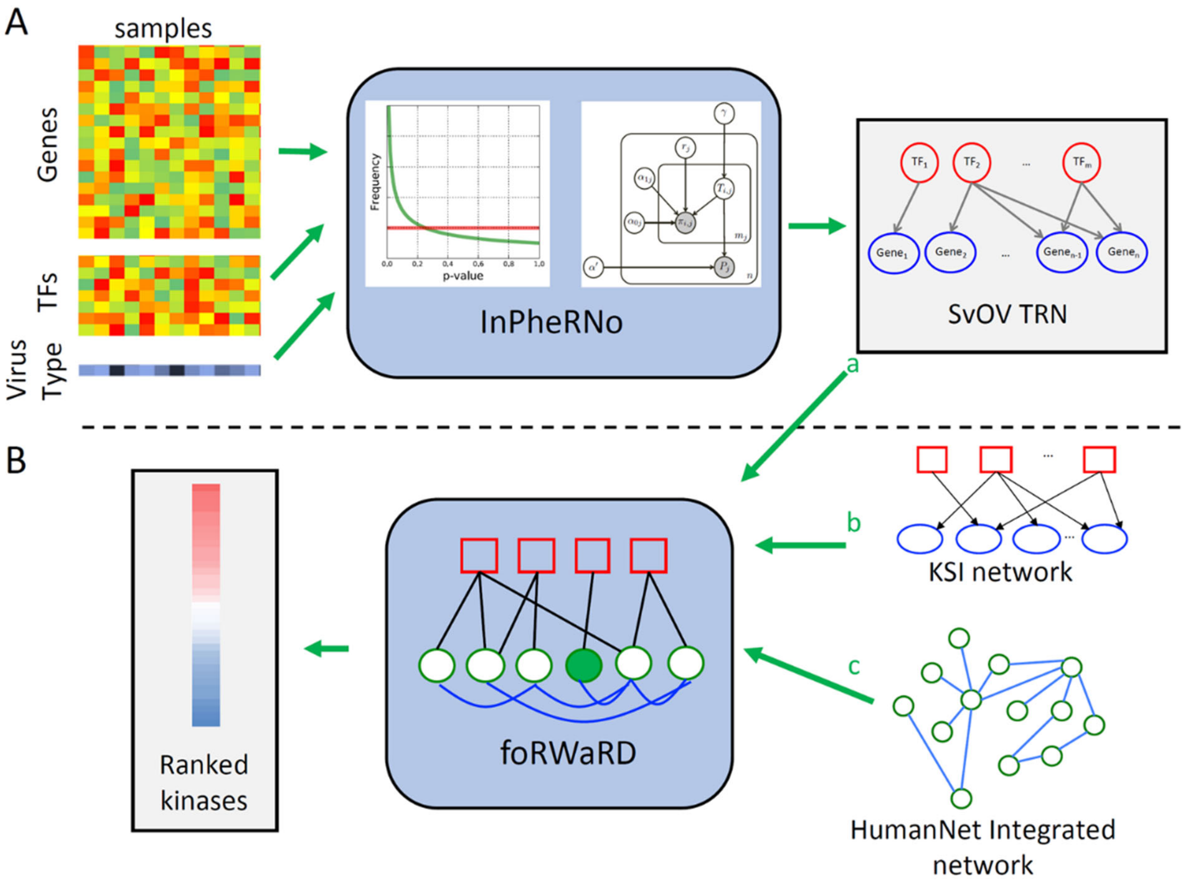


Figure 3.1 Illustration of the overall computational pipeline. (A) A differential expression (DE) analysis is performed to compare the expression levels of genes in SARS-CoV-2 infected samples versus samples infected by other viruses. The results of the DE analysis, along with the gene expression profiles of all samples, are inputted to InPheRNo [7]. InPheRNo first models the distribution of TF-gene association, and gene-phenotype associations, then assigns a confidence score to every possible TF-gene edge relevant to a phenotypic label and utilizes them to construct a SvOV TRN. (B) FoRWaRD [9] integrates several inputs, (a) a list of top TFs identified by InPheRNo, (b) an aggregated kinase-substrate interaction network, and (c) a HumanNet dataset, to construct a heterogeneous network. Then, FoRWaRD utilizes a random walk with restart (RWR) algorithm to compute a proximity score for each kinase and form a ranked list. Adapted from [11].

This pipeline combined our recently developed computational tools (i.e., InPheRNo [7], FoRWaRD [9], KnowEnG [10]) in a stepwise manner (Figure 3.1):

(1) The pipeline uses InPheRNo [7] to construct a TRN associated with gene expression changes between SARS-CoV-2 infected samples, and other respiratory viruses infected samples.

(2) The pipeline uses FoRWaRD [9] to score and identify kinases associated with the SvOV network.

(3) The pipeline uses KnowEnG [10] to identify the functional characterization of the SvOV network.

The detailed implementations and documentation of three previously developed algorithms (InPheRNo [7], FoRWaRD [9], KnowEnG [10]) are available in their publications [7, 9, 10].

InPheRNo

InPheRNo utilizes a probabilistic graphical model to integrate regulatory evidence between TFs and target genes with phenotype information to reconstruct a phenotype-relevant TRN. The inputs to InPheRNo are as follows: (1) a list of p values of TF-gene associations; (2) a list of p values of gene-phenotype associations; and (3) a list of human TFs. The output of InPheRNo is a TRN associated with the host response to SARS-CoV-2 infection. Details of the algorithm are provided in Chapter 2.4.4 and its original paper [7].

To obtain significantly differentially expressed genes between two biological conditions, we began by downloading gene expression profiles from the GEO database. After filtering several series corresponding to human sapiens in GSE147507, we obtained RNA-seq profiles containing gene expression information for genes measured across SARS-CoV-2 infected samples, and other viruses infected samples (Table 3.1).

To prioritize genes that were more likely present in one of these two biological conditions: (1) SARS-CoV-2-infected epithelial cells, (2) other respiratory viruses infected epithelial cells, we employed data corresponding to samples contaminated by SARS-CoV-2 and samples contaminated by the aforementioned other viruses (e.g., RSV, IAV, and HPIV3). Next, we executed differential expression analysis (DE analysis) using the package EdgeR [163]. During the DE analysis, we categorized samples into two groups based on the virus types of concern and adjusted for confounding factors (i.e., the duration of infected time and the cell type). As outputs, we extracted the top 500 DEGs for each sample with FDR < 1.43E-3, and we hypothesized that these DEGs tended to be associated with the relevant phenotypes (i.e., SARS-CoV-2 or other respiratory viruses). In line with the DE analysis, we applied quantile normalization with voom in

Limma package [164], followed by a z-score normalization to obtain a list of p values indicating the gene-phenotype associations.

Table 3.1 GSE147507 sample information. The column 'sample_title' provides information about which series those replicates are coming from, which cell lines, and the specific virus it infected. The second column indicates the number of replicates for each sample. The fourth column shows the duration of infected time for each sample and is used as one of the confounding variables in the DE analysis. The last column specifies the phenotypic labels in our analysis. (LUAD: Lung adenocarcinoma; HBECs: primary human bronchial epithelial cells.)

Series	Rep	Cell Type	Time	Group
Series1_NHBE_SARS-CoV-2	3	HBECs	24 hours	Sars-Cov-2
Series2_A549_SARS-CoV-2	3	LUAD	24 hours	Sars-Cov-2
Series3_A549_RSV	2	LUAD	24 hours	Other-Viruses
Series4_A549_IAV	2	LUAD	9 hours	Other-Viruses
Series5_A549_SARS-CoV-2	3	LUAD	24 hours	Sars-Cov-2
Series6_A549-ACE2_SARS-CoV-2	3	LUAD	24 hours	Sars-Cov-2
Series7_Calu3_SARS-CoV-2	3	LUAD	24 hours	Sars-Cov-2
Series8_A549_RSV	3	LUAD	24 hours	Other-Viruses
Series8_A549_HPIV3	3	LUAD	24 hours	Other-Viruses
Series9_NHBE_IAV	4	HBECs	12 hours	Other-Viruses
Series9_NHBE_IAVdNS1	4	HBECs	12 hours	Other-Viruses

Lastly, a list of p values of TF-gene associations, p values of gene-phenotype associations, along with a list of human TFs downloaded from AnimalTFDB 2.0 [153], were then fed into InPheRNo [7] (the methodology of this method is discussed in Chapter 2.4.4). We iteratively ran InPheRNo 1000 times with 500 repeats and default settings for other parameters. The output of this step was a TRN capturing high confidence TF-gene interactions involved in the host response to SARS-CoV-2 infection versus the host response to other virus infections.

FoRWaRD

In order to identify kinases associated with the top TFs in the SvOV TRN, another computational tool, FoRWaRD [9], was added to our pipeline.

FoRWaRD is a method based on the random walk with restart (RWR) algorithm to rank the kinases in a heterogeneous network based on their relevance to top TFs identified in the SvOV TRN. A heterogeneous network is a superimposing network containing two networks with different nodes and edges linked through bipartite interactions [165].

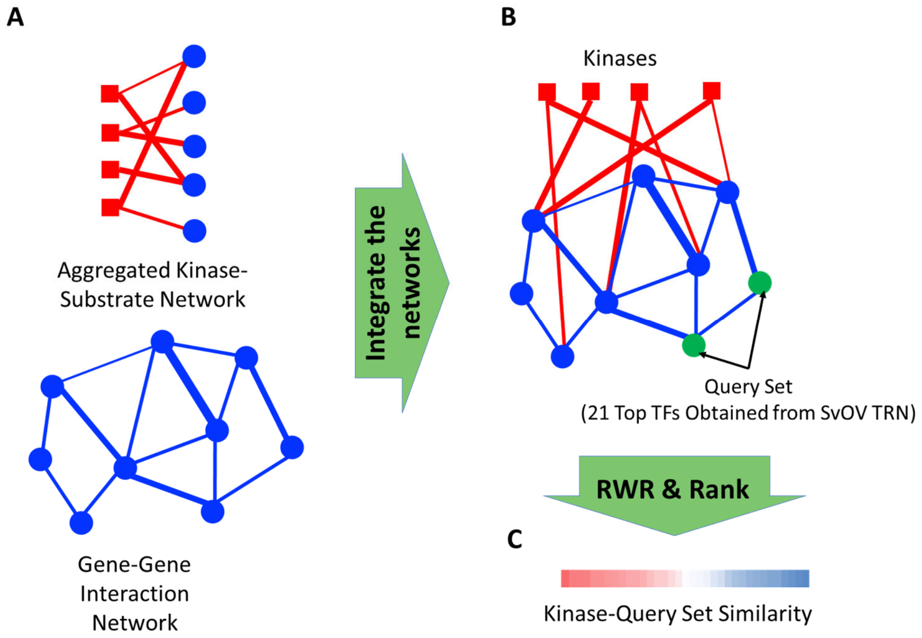

A
Aggregated Kinase-Substrate Network
Gene-Gene Interaction Network
Integrate the networks
B
Kinases
Query Set
(21 Top TFs Obtained from SvOV TRN)
RWR & Rank
C
Kinase-Query Set Similarity

Figure 3.2 Illustration of the FoRWaRD algorithm. The inputs to FoRWaRD are an aggregated KSI network and an integrated humanNet as inputs. The output of FoRWaRD is a table with the kinases and their importance measures, including 'difference score,' 'normalized difference,' and 'ratio of the former metrics.' Adapted from [11].

The inputs to FoRWaRD were the known kinase-substrate association pairs, a gene-gene interaction network, and a query set containing TFs found in the inferred TRN. The algorithm first ranked the network nodes according to their proximity to the query set and ran RWR twice on this formed heterogeneous network using either (1) the query set or (2) all network nodes as the restart set; and then produced a probability rating for each node in the network. The sets of scores were essentially standing for the relevance of the node to the restart set. As the last step, FoRWaRD normalized those scores and ranked kinases based on how much more significant their query set score was than their control score. The output of FoRWaRD was a table with the kinases and their importance measure including 'difference score,' 'normalized difference,' and 'ratio of the former metrics.'

KnowEnG

In order to identify critical pathways and biological processes associated with TFs and their targets in the inferred TRN, we incorporated the gene set characterization (GSC) computational tool of Knowledge Engine for Genomics analytical (KnowEnG) platform [10] into our pipeline.

KnowEnG is a platform that integrates multiple tools, such as GSC analysis, signature analysis, and sample clustering, for analyzing genomics data sets. The GSC pipeline of KnowEnG platform has two modes: (1) the standard mode and (2) the network-guided mode. The standard mode uses a simple Fisher's exact test to identify important pathways, while the network-guided mode uses an algorithm called DRaWR [166] to encode prior knowledge from other well-established networks (e.g., a protein-protein interaction network) in their pipeline.

The input to the GSC analysis was a TF-Gene matrix, and the output of the analysis was a *.txt* file with two columns: a list of genes of interest and a list of relevant biological processes or pathways with corresponding statistical significance of the enrichment scores.

3.4 Results

3.4.1 Identification of Top TFs in the SvOV TRN

Given the SvOV TRN, we ranked TFs based upon the number of differentially expressed target genes in the network. We hypothesized that top-ranked TFs were key regulators for distinguishing the host response to infection by SARS-CoV-2 versus other viruses.

The sorted listing of the top 21 TFs targeting at the very least 1% of the considered DEGs in the SvOV TRN is shown in Table 3.2. Based on this list, we mined the literature to scope the evidence concerning the role of those TFs in the host response to SARS-CoV-2 infection. Encouragingly, some of those top TFs have been formerly revealed to be activated throughout COVID-19, such as STAT1 [167], STAT2 [168], TP53 and IRF9 [169].

Table 3.2 Top 21 TFs implicated in the SvOV (SARS-CoV-2 vs. other viruses) TRN. We ranked TFs based on their number of DEGs identified by InPheRNo. The 1st column represents the top 21 TFs, and the 2nd column represents the percent of the considered genes for each TF.

Transcription Factors	Percent of target genes
STAT1	5.89%
STAT2	2.95%
MLX	2.74%
EGR4	1.47%
RCOR1	1.47%
SP140L	1.47%
TP53	1.26%
RCOR2	1.26%
MAX	1.26%
ZNF496	1.26%
ZNF512B	1.05%
SMAD7	1.05%
SOX12	1.05%
IRF2	1.05%
HDX	1.05%
EGR1	1.05%
SP110	1.05%
IRF9	1.05%
ZNF143	1.05%
NFIX	1.05%
ZBTB32	1.05%

Next, we sought to assess the regulatory relationships between the identified TFs and their targets in the SvOV network. While a systematic evaluation was not possible because most experimental databases were formed by conditions that did not purely target the host response to respiratory viral infections, we could still expect the identified regulatory relationships in the SvOV network to appear in some comprehensive databases such as Gene Transcription Regulation Database

(GTRD) [170]. GTRD contains published ChIP-seq data sets from various studies and reprocessed them in a uniform computational pipeline, and we hypothesized that regulation relationships identified in SvOV would highly overlap with regulatory evidence found in GTRD ChIP-seq data.

Among the 21 top TFs identified by InPheRNo (Table 3.2), four of them were present in the GTRD ChIP-seq dataset. Using these four shared TFs, we first filtered a subset of the SvOV TRN, which contained those four TFs and their target genes identified in the SvOV network. Next, we treated the filtered SvOV TRN as a sample of *n* selected individuals (without replacement) and the GTRD database as the population set. Each individual in the sample set can be characterized as 'found' or 'not-found' in the population. Since this setting fulfilled all the assumptions of the hypergeometric test, we then employed that test to measure whether our subset of the SvOV network (sample set N) was significantly enriched in known regulatory mechanisms from the GTRD (population set M). As a result, 37 out of 45 target genes were confirmed, with a highly significant enrichment p-value of 2.36E-15.

3.4.2 Identification of Significant Signalling Pathways

In order to gain mechanistic insights into gene expression programs involved in the SvOV TRN, we performed the network-guided pathway enrichment analysis for each inferred leading TFs (Table 3.2) and their target genes in the inferred network.

As depicted in Figure 3.3, we identified the biologically relevant pathways enriched in each TF and their target genes. Cytokine signalling and interferon (IFN) signalling pathways had relatively high enrichment scores among those pathways. Also, induction of type I interferons has been shown to be a promising method in mediating immune response against SARS-CoV-2 infection [171, 172].

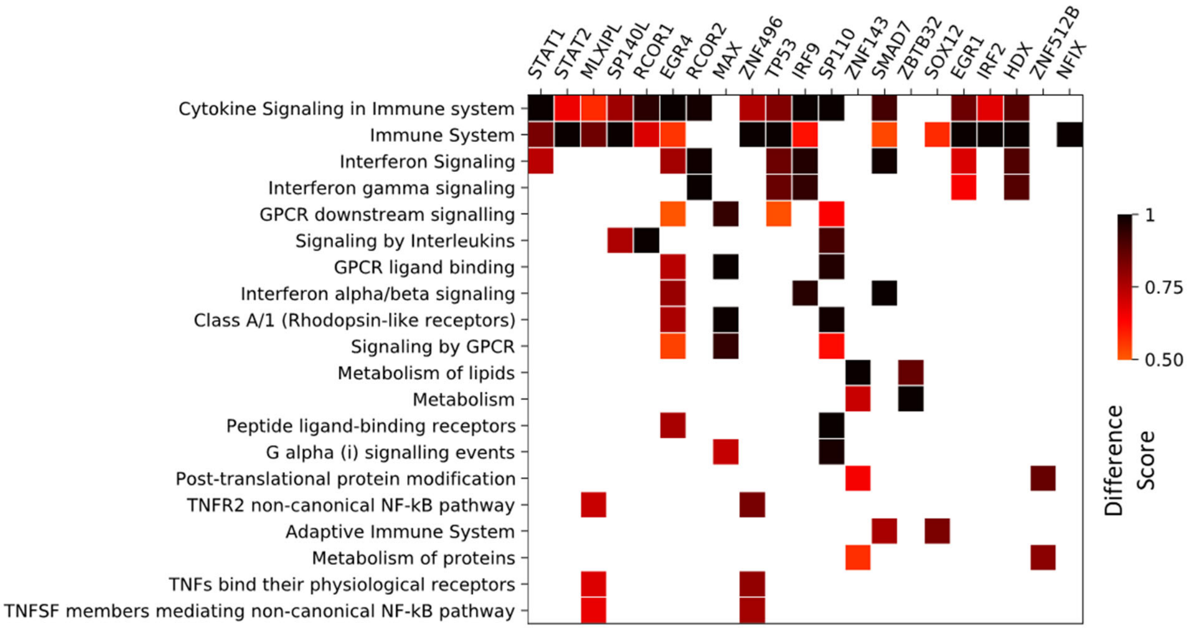


Figure 3.3 Pathway enrichment analysis using network-guided GSC pipeline of KnowEnG platform. Annotation on top of the heatmap shows top TFs (and their target genes identified in SvOV TRN); annotation on the left-hand side of the heatmap shows pathways that have been implicated for at least two TFs (along with their targets). The color intensity scale represents the 'difference score' in the analysis.

3.4.3 Identification of Kinases Associated with the SvOV TRN as Possible Therapeutic Targets

Kinases are enzymes involved in regulating most human proteins (i.e., substrates) activities through phosphorylation and are considered as one of the significant categories of drug targets for human diseases [173-175]. Turning our attention to identifying the potential therapeutic targets associated with the inferred TRN, we aimed to aggregate across the three aforementioned experimental KSI networks and retain protein kinases encoded in the human genome [152]. Our desired KSI network was a directed graph where each node was a kinase or its substrate protein, and an edge was associated with a phosphorylation reaction between a kinase and its substrate.

Based on the human genome project [152], at least 518 protein kinases were considered to be included in the human kinome, and these kinases phosphorylate a majority of human proteins (i.e. substrates). Therefore, after downloading all KSI pairs collected from three publicly available datasets mentioned earlier, we performed a filtering step to keep putative protein kinase genes in the aggregated network and remove the redundant kinase-substrate pairs. As a result, we obtained an aggregated KSI network consisting of 406 kinases and 3942 non-kinase substrates (29,594 unique KSI pairs in total).

Next, we applied FoRWaRD [9] to rank kinases based on their relevance to the query set (i.e., top TFs and target genes). As shown in Table 3.3, MYO3A, JAK3 and VRK3 are the kinases with the highest-ranking scores. We searched for other studies to support our findings to ensure that our identification was scientifically substantiated. As expected, other research groups have found that JAK-STAT pathway inhibitions may help alleviate adverse inflammatory responses [176-178].

Table 3.3 Top 15 kinases identified using foRWaRD for the top TFs in the SvOV network.

Top Kinase	FoRWaRD Score
MYO3A	11.60
JAK3	10.74
VRK3	10.34
ADCK1	9.74
JAK1	8.69
MAP2K5	8.19
BMX	7.97
HIPK4	7.48
JAK2	7.46
MAP3K13	7.34
LCK	6.63
CAMK2B	6.34
MAPK14	6.32
BTK	6.13
MAPK11	6.05

3.4.4 Evaluation of the Predicted Kinase-gene Interactions

Above, we applied FoRWaRD [9] to utilize direct and indirect interactions between kinases and the input query gene set to identify the top kinases. To complete the analysis, we aimed to examine the effects of kinase knockdown on the expression of genes identified by InPheRNo. More specifically, we would like to test whether the knockdown of a specific kinase gene can directly or indirectly interfere with the expression of its substrate (i.e., TF) and associated target genes in the SvOV TRN. To achieve this goal, we first visualized the interconnectivity of interactions between kinases, top TFs and/or substrates, as well as target genes in the SvOV network using Cytoscape [179].

In Figure 3.4, we showed *direct* interactions between kinases, TFs, and their target genes in the SvOV network. Only direct kinase-TF interactions presented in the aggregated KSI were included in this scenario. In Figure 3.5, we showed *indirect* interactions between kinases, substrates, (non-substrate) TFs and their target genes in the SvOV network. Kinases were recruited to TFs through indirect interactions mediated by non-TF substrates in this scenario.

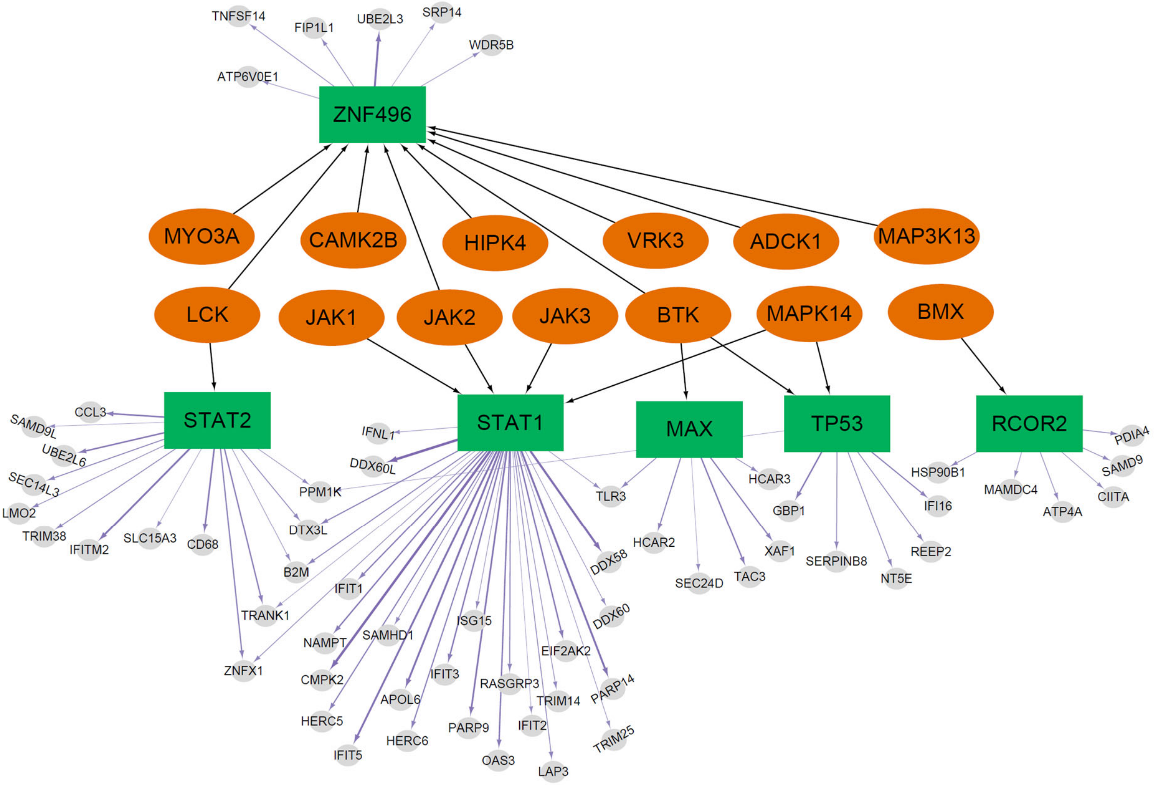


Figure 3.4 Network representation of the direct interactions between kinases, TFs, and their target genes using Cytoscape software. Kinases are depicted as orange ellipses, TFs are depicted as green rectangles, and target genes are depicted as grey ellipses.

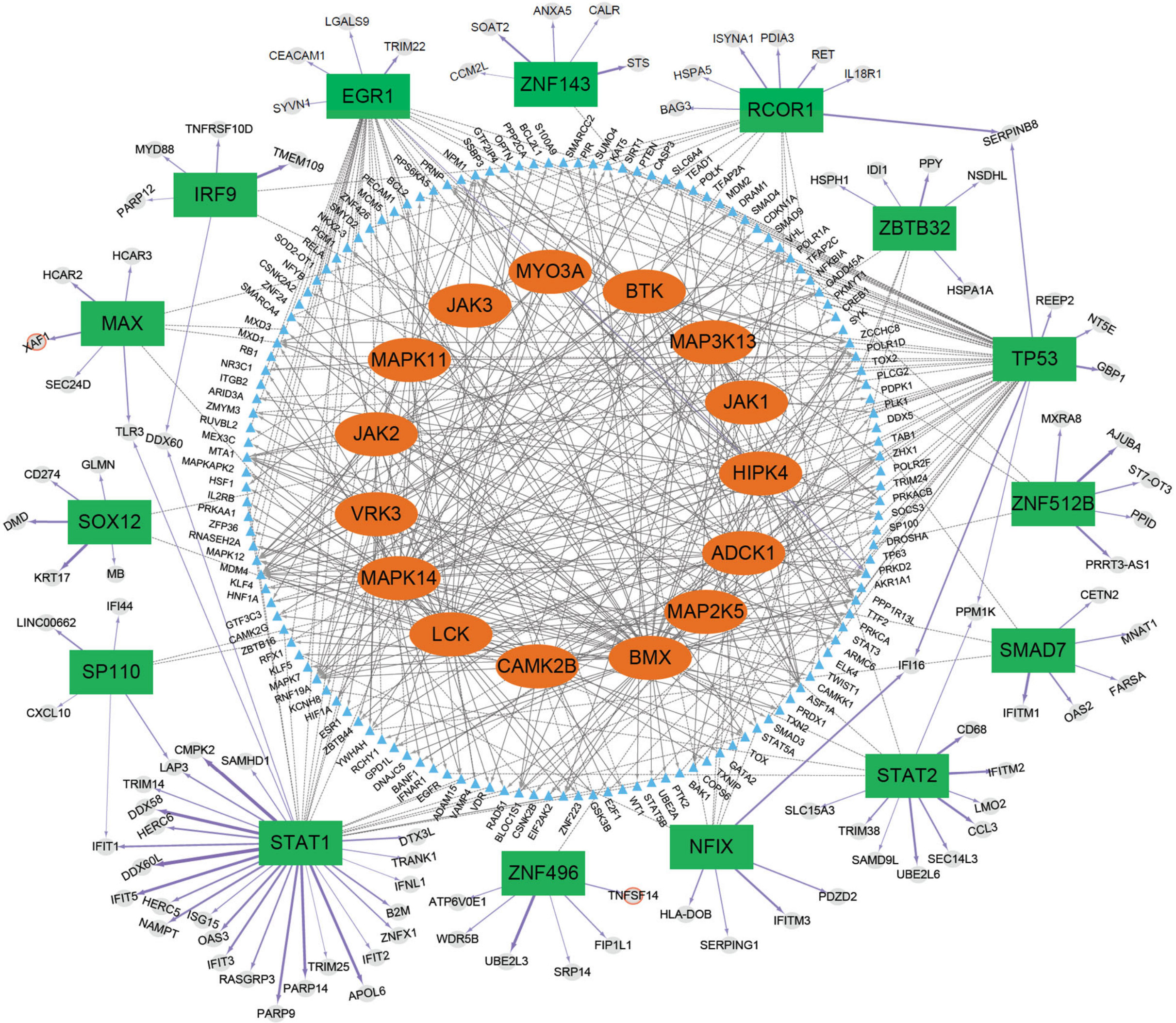


Figure 3.5 Network representation of indirect interactions between kinases, (non-TF) substrates, TFs, and their targets using Cytoscape version 3.8.2. Kinases are depicted as orange ellipses, non-TF substrates are blue triangles, TFs are green rectangles, and target genes are grey ellipses. Only the substrates with a minimum of one link to the implicated TFs in the HumanNet network are selected and coloured as light blue.

Next, we compared implicated TFs and their targets to gene expression signatures in the LINCS L1000 shRNAs-perturbation database [155]. Although shRNAs-mediated gene knockdown experiments were conducted on multiple disease cell lines, we only focused on experiments corresponding to the A549 cell line as this cell line was part of what we have included in the reconstruction of TRN.

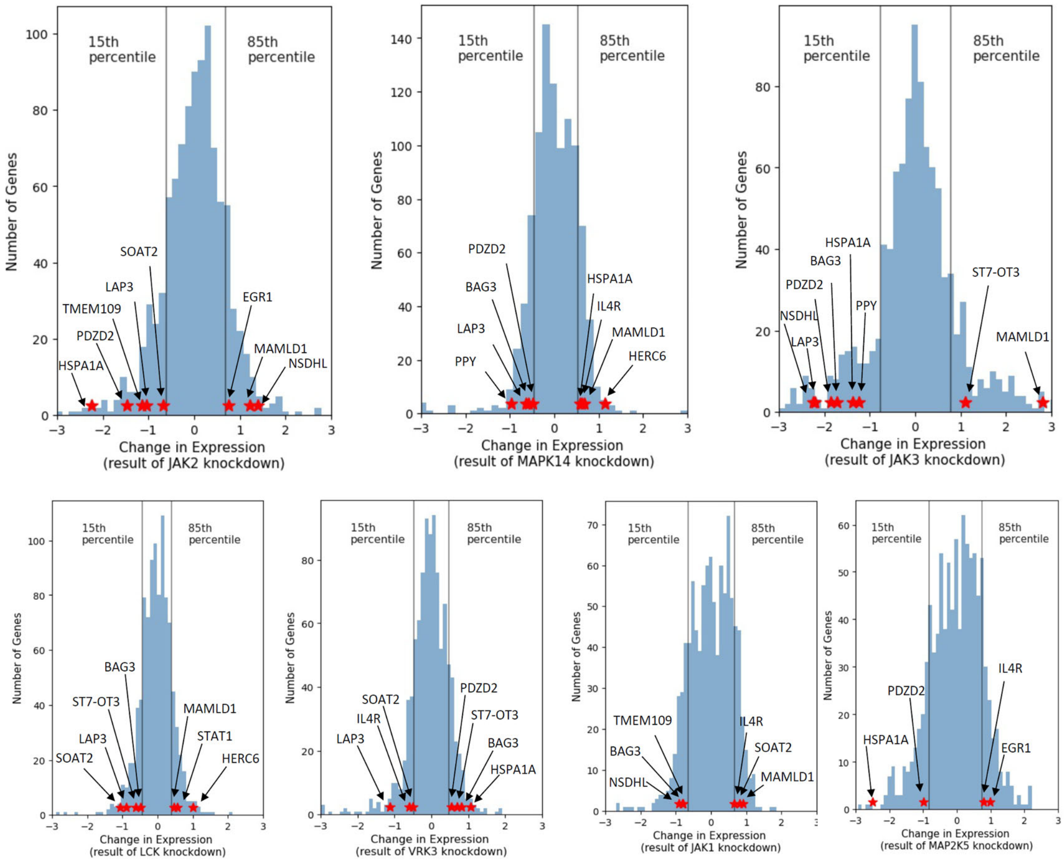


Figure 3.6 Histogram of z-score normalized gene expression changes of LINCS L1000 landmark genes in A549 cells due to knockdowns of kinases implicated in the SvOV analysis [11]. The 15th and 85th percentiles are shown as two vertical lines, and the gene expression changes in knockdown experiments are shown as red stars with arrows linking to their gene symbols.

The gene expression signatures in the GSE92742 dataset correspond to z-score normalized expression changes versus control in a series of shRNA knockdowns of a target gene of interest [155]. We only focused on using gene expression signatures for ‘L1000 landmark genes’, considering these were the only ones experimentally determined in the database (the other genes were computationally forecasted). Out of 15 implicated kinases using the ForWaRD algorithm in Table 3.3, only 7 of them have knockdown signatures. Meanwhile, out of all identified TFs and targets in the SvOV TRN, only 14 targets and 3 TFs were presented in the list of L1000 landmark genes. To better visualize our analysis, we only reported genes with their normalized expression changes (captured in the signature) amongst the top (or bottom) 15% of all landmark genes in Figure 3.6. In each histogram, the knockdown of a kinase was shown to positively (negatively) influence a landmark gene if its expression was increased (decreased).

3.5 Discussion

This study identified regulatory mechanisms specific to the SARS-CoV-2 virus compared to other viruses in contract to patient data. This decision was made based on the limited accessibility of data on different epithelial cell lines that infected several respiratory viruses that were experimented with under the same lab procedure. Combining varied data across multiple sources or laboratories can be a challenge: we will have to consider correcting lab-specific and sample-specific measurement errors with multiple references because these differences may introduce batch effects or significant technical/biological variations in the processed data.

The genetic diversity and frequent recombination nature of the coronavirus genome render the variation of this virus highly unpredictable [180, 181]. Despite our efforts to include the response of epithelial cells to infection by SARS-CoV-2 and various respiratory viruses, these findings may not translate to new SARS-CoV-2 variants such as B.1.351 (Beta), B.1.617.2 (Delta), C.37 (Lambda) initially found in the United Kingdom, South Africa and Peru [182-184]. Future studies using integrated bioinformatics methods to explore biomarkers of SARS-CoV-2 variants and the combined effects of these variants remain to be done. Nonetheless, the fact that we only included one RNA-seq dataset in our analysis may limit the predictive power of our models. Thus, we hypothesize that the accuracy of the TRN reconstruction can be improved with a cross-dataset integration (i.e., more samples could be included in the study to infer the TRN). In the future, as more datasets become available in these matters, we will incorporate multi-omics analyses in the proposed pipeline.

Despite those limitations mentioned above, the implications of the study's results are promising. In this study, we developed a computational pipeline that combines three practical algorithms (InPheRNo, KnowEnG GSC and FoRWaRD) to identify TRN, significant pathways and kinases. Among those results, a significant number of them have been proved to be related to SARS-CoV-2 infection in previous research studies. This study facilitates understanding of putative regulatory mechanisms associated with the response of host epithelial cells to SARS-CoV2; and then identifies novel therapeutic targets that can function as the key components for future development of medicines, which tend to reduce the symptoms of COVID-19. Our collaborator is currently working on the experimental verification of some of the identified therapeutic targets.

4 InPheRNo-ChIP: Inference of Phenotype-relevant Transcriptional Regulatory Networks using Multi-omics Data

4.1 Problem Statement

The focus of this study was to develop a method that incorporates multi-omics data (i.e., RNA-seq and ChIP-seq data) and phenotype information together to study human embryonic development.

4.1.1 Biological Aspect

One of biology's basic but fascinating questions is how a single fertilized egg cell eventually develops into a mature, multicellular organism containing different cell types capable of organ development and regeneration, while the cells' genomic content remains the same [185]. Once fertilization takes place, the resulting zygote starts dividing and then forms a blastocyst [186]. After implantation, the blastocyst's innermost layer forms three germ layers: ectoderm, mesoderm, and endoderm (Figure 4.1). The differentiation from the germ layers into different organs and tissues is of specific interest to studying human development and stem cell research [187, 188].

Studies have shown that abnormal early embryonic development contributes to adverse pregnancy outcomes, including recurrent implantation failure [189], recurrent pregnancy loss and congenital disabilities such as Craniosynostosis and Anophthalmia [190]. With the advent of genomic sequencing and statistical methods, regulatory mechanisms involved in embryogenic development in mice and other types of organisms have been well-studied. However, due to ethical constraints in scientific research [191], scientists cannot directly study human embryonic development *in vivo*. Fortunately, incredible progress in in-vitro models for mammalian embryos provides an excellent opportunity to study regulatory networks and other pathogenic mechanisms that lead to abnormal human development.

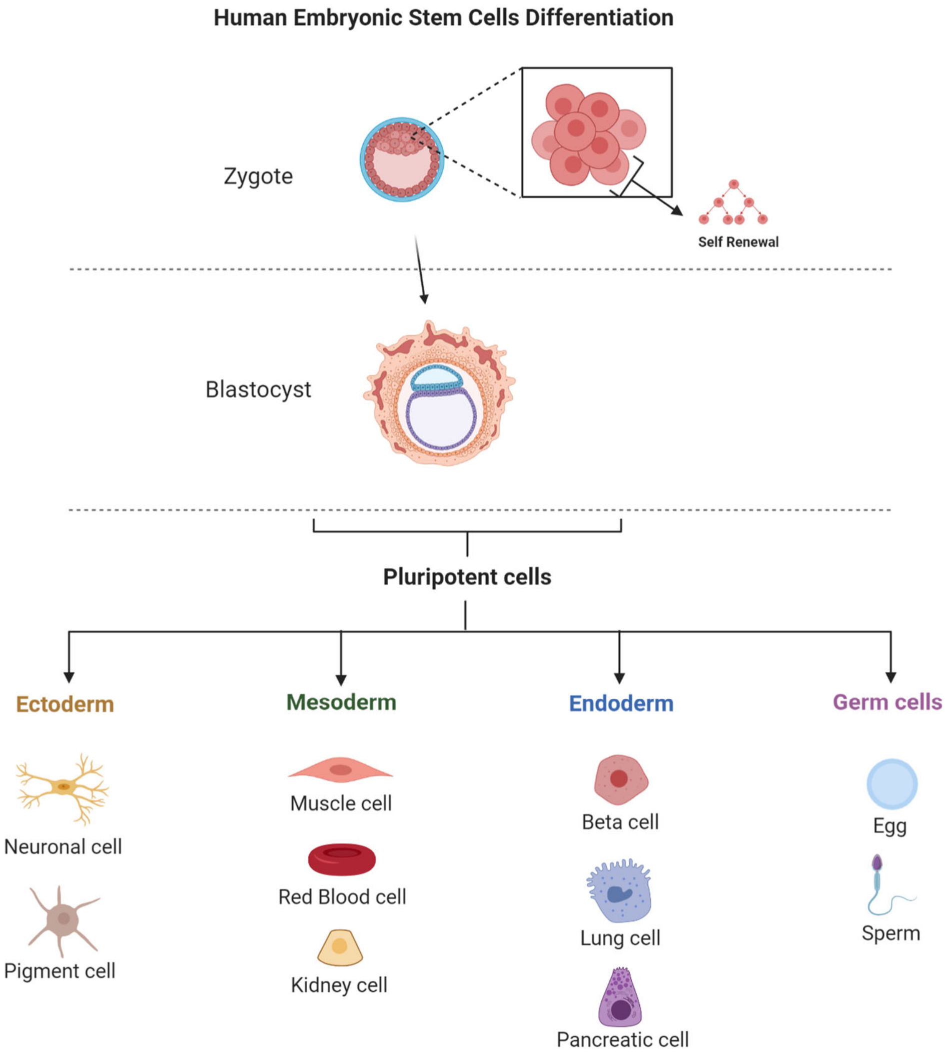


Figure 4.1 Illustration of human embryonic stem cells (hESCs) differentiation. Adapted from "Human Embryonic Stem Cell Differentiation" by BioRender.com (2020). Retrieved from https://app.biorender.com/biorender-templates.

4.1.2 Computational Aspect

Over the past two decades, decreasing cost of high throughput sequencing and new biotechnological approaches allow researchers to infer regulatory mechanisms in different biological processes. Among these methodologies, gene expression profiling is the most common approach for relating molecular-level differences in the expression level of the gene in response to different phenotypes and using this relationship to construct gene/transcriptional regulatory networks, as discussed in chapter 2. Furthermore, multi-omics data integration has been shown to better characterize complex biological processes than using a single source of data [75, 192-194].

Chromatin immunoprecipitation combined with sequencing (ChIP-seq) [195] is one such data modality that can provide evidence of regulatory relationships between TFs and their target genes, complementary to gene expression data. ChIP-seq is a technique to facilitate the study of protein-DNA interactions at the genomic level [196]. This technique is used primarily to determine how TFs and other chromatin-associated proteins, such as histones and RNA polymerase, influence phenotype-affecting mechanisms, such as morphology, biochemical or physiological properties and behaviours. Commonly, there are two different classes of protein-chromatin interactions: histone modifications with broad peaks and transcription factors ChIP-seq with narrow peaks. The most common application is to identify transcription factor binding sites (TFBS), and the resulting genomic regions are called "peaks."

Even though various methods have been developed to take advantage of multi-omics studies and construct regulatory networks [76], the main problem of these methods is their ignorance of any gene-phenotype association existing under different experimental conditions (e.g., control versus case samples) [7]. On the other hand, to the best of knowledge, network inference methods incorporating phenotype information (discussed in Chapter 2.4) cannot integrate multiple data sources.

To address both issues, we proposed InPheRNo-ChIP, an improvement variant of the original InPheRNo [7], that integrated high-quality RNA-seq, ChIP-seq, and phenotypic information of the samples to reconstruct a TRN, which was specific to the differentiation of human embryonic stem cells (hESCs) to endoderm (EN). It is acknowledged that ChIP-seq data provides genome-wide information about interactions of DNA target sites (i.e., target genes) against their corresponding TFBS [75]. Thus, we hypothesized that its integration with gene expression profiles and sample-level phenotypic data could improve the identification accuracy of phenotype-relevant regulatory mechanisms, which may also enhance our understanding of molecular dynamics during the early differentiation of hESCs.

The schematic illustration of InPheRNo-ChIP is deciphered in Figure 4.2. In order to verify our hypothesis for data integration, we first generated *in silico* gene expression data and ChIP-seq peaks with certain constraints based on real-world biological datasets, and then reported several performance assessments of InPheRNo-ChIP concerning different input data options: using (1) only gene expression data, (2) perfectly matched RNA-seq and ChIP-seq data, and (3) partially unmatched RNA-seq and ChIP-seq data (Chapter 4.3). To further illustrate the insights InPheRNo-ChIP enables, we applied InPheRNo-ChIP to *in vitro* RNA-seq and ChIP-seq data corresponding to hESCs and hESC-derived endodermal lineage (Chapter 4.4). Last but not least, we summarized the key findings and indicated some future perspectives in Chapter 4.5.

4.2 Methodology

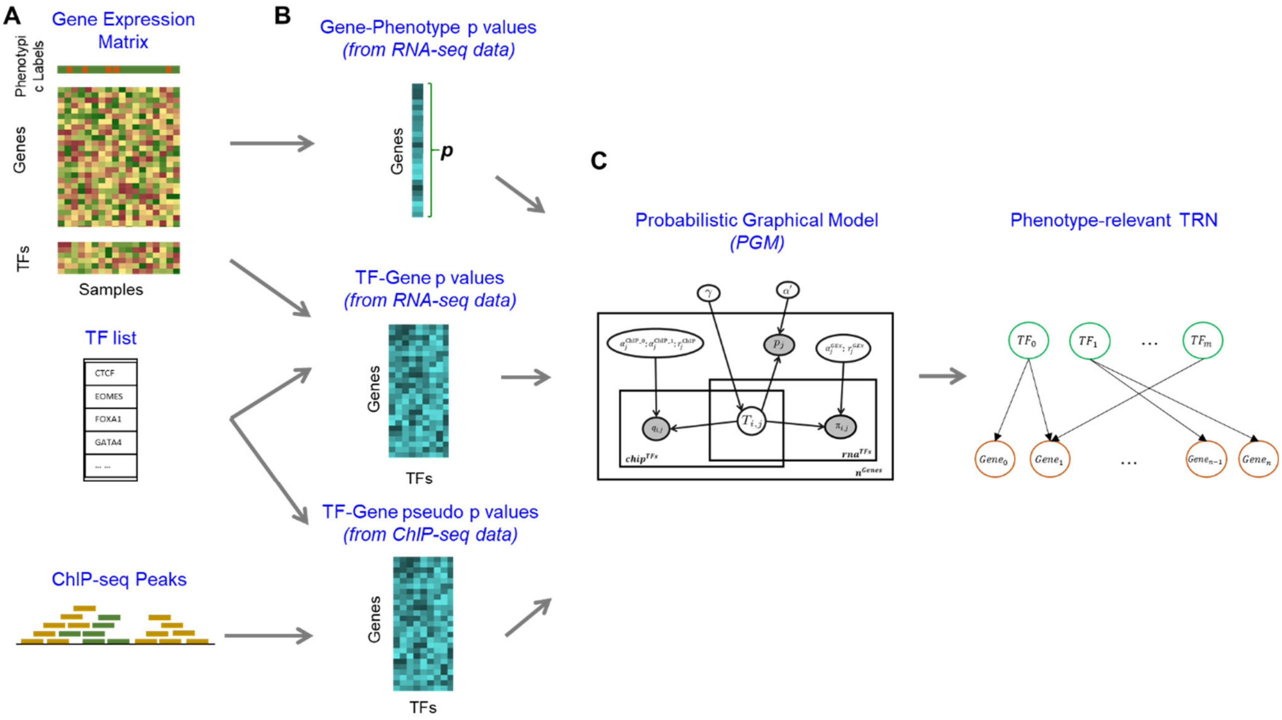


Figure 4.2 The overall depiction of the InPheRNo-ChIP framework. (A) The inputs include a list of TFs, a matrix of gene expression profiles, multiple ChIP-seq peaks and a phenotypic vector indicating which group the sample belongs to. The colour and intensity of the boxes in the heatmap representation indicate changes in gene (or TF) expression. The list of TFs is used to filter legit human TFs and separate the expression matrix into two matrices: a matrix of TF expression data and a matrix of gene expression data. The ChIP-seq peaks are obtained from peak calling algorithms such as MACS/MACS2, SISSRs and PICS [197]. (B) Prior to inferring phenotype-relevant TRN, InPheRNo-ChIP computes three sets of (pseudo/true) p values, representing gene-phenotype associations (denoted by P_j), TF-gene associations from RNA-seq (denoted by π_{ij}), and TF ChIP-seq peaks (denoted by q_{ijk}), respectively. (C) The core of the InPheRNo-ChIP is a probabilistic graphical model models the variables of interest. The output of the algorithm is a phenotype-relevant TRN composed of TFs, genes with a confidence score associated with TF-gene edges.

We present InPheRNo-ChIP, a computational tool for reconstructing phenotype-relevant TRNs from gene expression and ChIP-seq data. As shown in Figure 4.2, the main pipeline is comprised of four steps:

1) Calculation of gene-phenotype p values using gene expression data.
2) Calculation of TF-gene p values using gene expression data.
3) Calculation of TF-gene pseudo p values using TF ChIP-seq peaks.
4) Calculation of confidence scores for the identified TF-gene edges and constructing the phenotype-relevant TRN by utilizing a carefully designed probabilistic graphical model (PGM).

4.2.1 InPheRNo-ChIP step 1: Gene-phenotype Associations from RNA-seq Data

The inputs to step 1 of InPheRNo-ChIP are a gene expression count matrix where rows correspond to genes and columns correspond to samples, and a vector containing the phenotypic variation for all samples.

The output of this step is a list of p values (denoted as P_j) that summarizes the significance of the associations between the gene expression and the phenotypic labels for each gene. Under the null hypothesis that TF_i is not associated with $Gene_j$ to affect the phenotype, the distribution of P_j is uniform (based on the behaviour of a p-value under the null hypothesis).

We specify the term ‘phenotypic labels’ (or ‘phenotypic scores’) to recognize distinct phenotypic attributes (i.e., discrete versus continuous data types) that are associated with each sample. Depending on whether the type of phenotype is continuous or categorical, InPheRNo-ChIP generates p values accordingly:

1. In the case of having categorical phenotypic labels, such as SARS-CoV-2 versus other viruses or cases versus control in disease study, the p values of gene-phenotype associations can be obtained using differential expression analysis, such as EdgeR [163], DESeq2 [198], and EBSeq [199].
2. In the case of having continuous-valued phenotypic scores, where the range of variability of the phenotype is continuous (not categorical), such as IC50 drug response measurements in the pharmacogenomic study, the p values of gene-phenotype associations can be found through regression or correlation analysis.

4.2.2 InPheRNo-ChIP step 2: TF-gene Associations from RNA-seq Data

Step 2 of InPheRNo-ChIP takes properly normalized gene expression data and applies one of the following two options to estimate the p values of gene-TF associations. Although our algorithm prefers quantile normalized count-per-million (CPM) values across all samples, other normalized measurements are also acceptable as inputs in this step.

It is worth mentioning that in this step, we assume that the sample size is greater than or equal to the feature size (i.e., the number of TFs with known RNA-seq and ChIP-seq data). This is because we are currently only interested in applying this methodology to publicly available human embryonic datasets, where we have 28 RNA-seq samples and 25 TFs that are shared in both RNA-seq and ChIP-seq (see chapter 4.4 for more details). On the other hand, if the sample size is smaller than feature size, a two-step procedure similar to the one used in InPheRNo can be adapted to compute 'pseudo' p values instead of the 'true' p values above (a direction that we are going to pursue in the future for other applications).

Two options are introduced in this step to compute a p-value of association between $Gene_j$ and TF_i:

(1) We use linear correlation coefficients, such as Pearson correlation coefficients, as a first option.

(2) We use the multivariate regression model, such as Ordinary Least Squares (OLS), as a second option.

In the first option, InPheRNo-ChIP computes Pearson's correlation coefficients as a baseline mode to obtain the p values of associations between the gene and its regulator. However, such methods likely fail to consider the nature of TRNs that comprises simultaneous observations and analysis of more than one feature. Thus, InPheRNo-ChIP addresses this potential concern by introducing another option to generate the p values of TF-gene associations.

In the second option, InPheRNo-ChIP applies OLS regression model to estimate a statistical dependency between the response (i.e., the expression of a gene) and the feature variables (i.e., the expression of gene j's candidate TFs). Notably, OLS is a linear model estimation method for estimating the unknown parameters in the linear regression model by minimizing the following objective function between the response variable and the features:

$$L_{basic\,form}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2, \qquad 4.1$$

where n is the sample size, y_i is the response variable, x_i is the feature variable corresponding to the i-th sample. In this most basic form, our features are the TF expression, and the response variables are the expression profile of each gene. Our study hypothesizes that the variation of the gene-expression values of a gene across samples can be explained by a linear combination of the expression of this gene's regulators.

The output of the second step is a p-value of TF-gene association, denoted by π_{ij}. Under the Null hypothesis that TF_i is not associated with $Gene_j$, the p-value π_{ij} is uniformly distributed between 0 and 1 [200].

4.2.3 InPheRNo-ChIP Step 3: TF-gene Associations from ChIP-seq Data

Step 3 of InPheRNo-ChIP utilizes a set of p values capturing the regulatory effects of the TF on its putative target genes using ChIP-seq peaks, denoted as q_{ijk}.

Since the upstream analysis is not the objective of this study and various methods have been developed to address similar problems [201-204], we use an existing tool T-Gene [204] to identify the TF-gene relationships from ChIP-seq data. T-Gene hypothesizes that genes which locate near TF binding sites (TFBS) are more likely to be regulated than those which locate far from TFBS. The algorithm defines a uniform null model to describe the distribution of the genome distance (i.e., the number of base pairs) between the closest edge of the TFBS and transcription start site (TSS) of the transcript in the reference genome, and outputs a set of Distance p-values.

After running T-Gene on our ChIP-seq datasets, we remove non-CT links and keep only those p values of CT links smaller than one for the following two reasons. (1) T-Gene manually assigns a Distance p-value of 1 to the putative link if its length (denoted by d) exceeds the user-defined maximum distance $D = 500kbp$; (2) In their paper [204], a comparison in terms of the prediction accuracy using closest_TSS links (CT links) versus non-CT links shows that CT links can produce more reliable results than non-CT links. Because of this post-processing step, those filtered Distance p values no longer satisfy the complete characteristics of true p values. Thus, we have to treat them as 'pseudo' p values. Adapting the same terminology from the InPheRNo paper [7], under the null hypothesis that the expression of a $Gene_j$ is not regulated by a TF_i, each 'pseudo' p value can be modelled as a Beta distribution in the step 4 of InPheRNo-ChIP.

4.2.4 InPheRNo-ChIP Step 4: PGM

With the above three sets of p values capturing either TF-gene associations or gene-phenotype associations, the biological hypothesis about how those elements (i.e., TFs, genes, phenotypic variations) influence one another remains unmodelled. Thus, in the fourth step of InPheRNo-ChIP, a Bayesian network (BN), one of the well-known branches of PGMs, is used to represent conditional independencies between nodes (i.e., a set of random variables) in a directed acyclic graph (DAG). Using plate notation, Figure 4.4 illustrates a schematic view of the model for identifying which TFs regulate genes and which biological processes are associated with these genes. Each variable in the PGM has an associated set of parameters underlying the distribution, and each directed edge represents the conditional dependency between the latent variable (T_{ij}s) and the observed variables ($P_j s$, π_{ij}s, and q_{ijk}s) for a $Gene_j$.

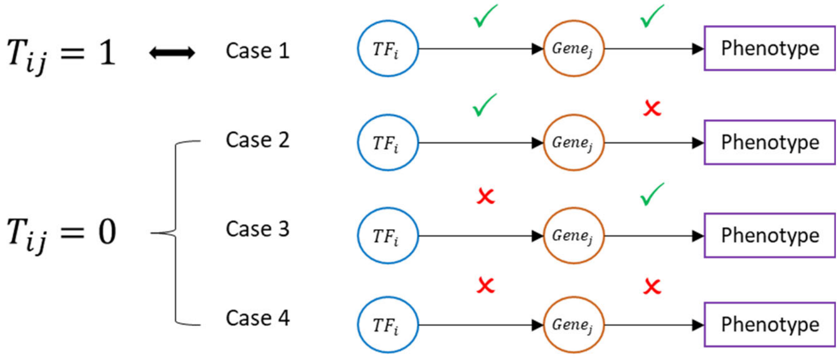


Figure 4.3 Illustration of four cases of the latent variable T_{ij}. $T_{ij} = 1$ implies TFi regulates $Gene_j$ to affect phenotype (case 1), $T_{ij} = 0$ indicates its logical complement (case 2-4).

The binary latent random variable T_{ij} represents the idea of 'a TF_i regulates the $Gene_j$ to affect its phenotype' (the alternative hypothesis). On the one hand, when the value of T_{ij} is 1, the definition of T_{ij} implies that there is a valid association between TF_i and $Gene_j$ to affect the phenotype of $Gene_j$'s (i.e., the first case in Figure 4.3). On the other hand, if the value of T_{ij} is 0, then the definition of T_{ij} implies one of the following: (a) TF_i regulates $Gene_j$ but there is no association between $Gene_j$ and its phenotype (i.e., the second case in Figure 4.3); (b) TF_i does not regulates

$Gene_j$ to affect the phenotype (i.e., the third case in the Figure 4.3); and (c) TF_i does not regulate $Gene_j$ and no valid association exists between a gene and its phenotype (i.e., the fourth case in the Figure 4.3).

We model the prior distribution of this random variable T_{ij} by using a Bernoulli distribution with parameter γ. Hence, the probability mass function (PMF) of T_{ij} is given by:

$$p(1) = P(T_{ij} = 1) = \gamma \qquad 4.2$$

$$p(0) = P(T_{ij} = 0) = 1 - \gamma, \qquad 4.3$$

and the cumulative distribution function (CDF) of T_{ij} is given by:

$$F(t_{ij}) = \begin{cases} 0, & t_{ij} < 0 \\ 1-\gamma, & 0 \le t_{ij} < 1, \\ 1, & t_{ij} \ge 1 \end{cases} \qquad 4.4$$

where γ is held fixed to the multiplicative inverse of the number of TFs in this project.

Next, as depicted in Figure 4.4, we use three observed variables to encode the following three sets of p values in the PGM: (1) P_j - representing a p-value of the association between target genes and the phenotypic labels (e.g., cell types or virus types); (2) π_{ij} - representing a p-value of the association between the expression of each gene and its regulator; (3) q_{ijk} - representing a Distance p value of TF-gene regulatory relationship obtained from the T-Gene algorithm [204].

Since the first observed variable P_j is an actual p-value and it is uniformly distributed when the null hypothesis is true. Under the null hypothesis, the phenotypic variation is not associated with the expression of $Gene_j$. In other words, if $T_{1,j} = T_{2,j} = \cdots = T_{m,j} = 0$, then the conditional distribution of P_j can be modelled by a uniform distribution. In contrast, if any T_{ij} is equal to 1, following the same approach used in InPheRNo paper [7], the conditional distribution of P_j under the alternative hypothesis (where the definition of T_{ij} implies that $Gene_j$ is associated with the phenotype) can be modelled by a Beta distribution. Since the shape of the beta distribution is often characterized by two parameters, α and β, we fix the value of β to 1 to limit the value of α to a range of [0,1]. Note that the uniform distribution on the interval [0,1] is a special case of this family (i.e., $Beta(\alpha = 1, \beta = 1)$).

Overall, the conditional distribution of this observed variable P_j given the value of T_{ij} can be modelled as:

$$P_j \sim \begin{cases} Unif(0,\ 1), & if\ T_{1,j}\ = T_{2,j}\ = \cdots = T_{m,j}\ = 0 \\ Beta\left(\alpha = \alpha_{gp},\ \beta = 1\right), & if\ any\ of\ Tij = 1 \end{cases}, \qquad 4.5$$

where we estimate the prior distribution of the unknown α by fitting a mixture of a uniform and distribution to the histogram based on the p values of gene-phenotype associations across all genes.

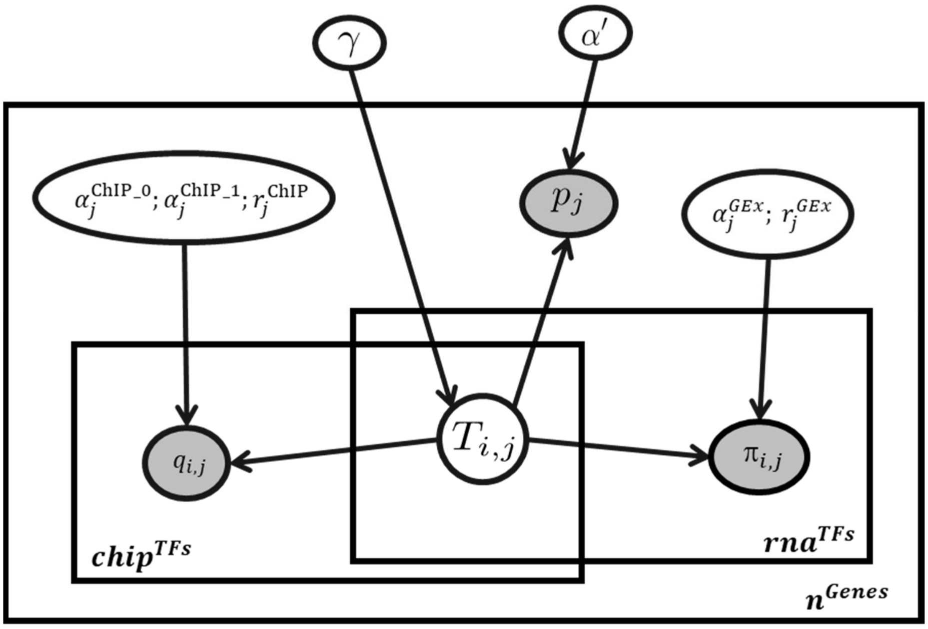

γ
α'
$\alpha_j^{\text{ChIP\_0}}; \alpha_j^{\text{ChIP\_1}}; r_j^{\text{ChIP}}$
p_j
$\alpha_j^{GEx}; r_j^{GEx}$
$q_{i,j}$
$T_{i,j}$
$\pi_{i,j}$
$\boldsymbol{chip^{TFs}}$
$\boldsymbol{rna^{TFs}}$
$\boldsymbol{n^{Genes}}$

Figure 4.4 The PGM used in InPheRNo-ChIP is represented in the plate notation. The PGM characterizes the relationship between the latent variable (T_{ij}) and a set of the observed variables (P_j, π_{ij}, and q_{ijk}). The plate notation provides an easier way to encode models with repeated structure and shared variables, whereas circular shaded nodes are observed random variables with a distribution, and plain nodes are learnable parameters and latent variables. The directed edges show how information flows from a set of parent variables of a node to itself. In terms of model construction, the most outer plate is for all the genes we have selected for InPheRNo, and n^{Genes} specifies the number of genes of interest. For each gene j (j ranges from 0 to $n^{Genes} - 1$), we construct two inner plates to separate RNA-seq and ChIP-seq information. As shown in the figure, the observed variable π_{ij} is associated only with the *RNA-seq* plate, and the observed variable q_{ijk} is associated only with the *ChIP-seq* plate. The interpretation of such dependence is that, for any pair of (TF, Gene) objects i and j, the node T_{ij} depends on (1) attribute q_{ijk}, (2) attribute π_{ij}, and (3) attribute P_j in the outer plate. The value of observed variables refits for each (TF, gene) pair based on their data sources.

The second observed variable, denoted by π_{ij}, represents the p-value of the association between TF_i and $Gene_j$ (regardless of the relationship between the expression of $Gene_j$ and its phenotype).

Under the assumption that the number of samples is equal to or greater than that of features, the obtained π_{ij} is uniformly distributed under the null hypothesis. Based on the implications of T_{ij}, we model the conditional distribution of $\pi_{i,j}$ differently. As shown in Equation 4.8, we use a Beta distribution to model the conditional distribution of a TF-Gene association when TF_i regulates $Gene_j$ to affect the phenotype ($T_{i,j} = 1$). On the other hand, when $T_{i,j} \neq 1$, it is indicating one of the following two scenarios (correspond to cases 2-4 in Figure 4.3): (1) TF_i does not regulate $Gene_j$ in a phenotype-independent manner; (2) TF_i regulates $Gene_j$ but is not relevant to a phenotype. As a result, we use a mixture of a Beta and uniform distribution to model the conditional distribution of $\pi_{i,j}$.

$$r_j^{GEx} \sim Unif(0,1) \tag{4.6}$$

$$\alpha_j^{GEx} \sim Unif(0,\ 1) \tag{4.7}$$

$$\pi_{i,j} \sim \begin{cases} Beta\left(\alpha = \alpha_j^{GEx},\ \beta = 1\right), & if\ T_{i,j}\ =\ 1 \\ r_j^{GEx}\ Beta\left(\alpha = \alpha_j^{GEx},\ \beta = 1\right) + \left(1 - r_j^{GEx}\right) Unif(0,1), & otherwise \end{cases}, \tag{4.8}$$

where we assigned α_j^{GEx} as a weak prior distribution to the shape parameter α in the Beta distribution and assigned r_j^{GEx} to the mixing proportions.

The last set of observed variable q_{ijk} represents the p values of regulatory effects of TFs and their putative target genes identified by the algorithm T-Gene [204]. Depending on the value of T_{ij}, we model the conditional distribution differently. When T_{ij} equals to 1, the conditional distribution of this random variable under the null hypothesis is modelled by a Beta distribution with parameters $\alpha = \alpha_j^{ChIP}$ and $\beta = 1$. Here, we use a Beta distribution instead of a uniform distribution to illustrate the fact that the distribution of p-value q_{ijk} is bias towards small values because we have restricted T-Gene's outputs to CT links with Distance p values smaller than 1. When T_{ij} equals to 0, the definition of T_{ij} implies that TF_i does not regulate $Gene_j$. Thus, following a similar modelling strategy described in InPheRNo, the conditional distribution of q_{ijk} can be modelled by a mixture of two Beta distributions.

$$\alpha_j^{ChIP\_1} \sim Unif(0,\ 0.5) \qquad 4.9$$

$$\alpha_j^{ChIP\_0} \sim Unif(0.5,\ 1) \qquad 4.10$$

$$r_j^{ChIP} \sim Unif(0,1) \qquad 4.11$$

$$q_{ijk} \sim \begin{cases} Beta\left(\alpha = \alpha_j^{ChIP},\ \beta = 1\right), & if\ T_{i,j}\ = 1 \\ r_j^{ChIP}\ Beta\left(\alpha = \alpha_j^{ChIP\_1},\ \beta = 1\right) + \left(1 - r_j^{ChIP}\right) Beta\left(\alpha = \alpha_j^{ChIP\_0},\ \beta = 1\right), & if\ T_{i,j}\ = 0 \end{cases}, \qquad 4.12$$

where r_j^{ChIP} stands for the proportion in the mixture.

As for prior distributions, we restrict the range of α to (0,1) and set the second shape parameter β equal to 1 to have the flexibility in modelling an extensive range of distributions tend for small values. As α increases, the degree of bias drops; when it approaches 1, the resulting distribution will be closed to a uniform distribution. On top of that restriction, we assign different upper/lower limits to the probability density functions for uniform distributions to ensure that $\alpha_j^{ChIP\_0} > \alpha_j^{ChIP\_1}$ (i.e., a more significant bias towards small values exists when TF_i is regulates $Gene_j$).

4.2.5 Inference

In terms of approximate inference, we conduct Markov chain Monte Carlo (MCMC) sampling to generate the empirical posterior probabilities of the hidden variable T_{ij} given the observations. A python module PyMC3 [205], is introduced in the process.

Table 4.1 Main parameters in the MCMC sampling The thin value is determined based on the settings in the original InPheRNo [7] and the other three parameters simply used default values of pymc3.sample() function, suggested by PyMC3 documentation [205].

Parameter	Description
N_b	Number of iterations to thin, set to 100
N_c	Number of chains, by default, set to 4, for multi-process sampling and model checking
N_i	Number of iterations by default, set to 1000
N_t	Number of tuned/burn-in samples to be discarded by default, set to 500

By default, PyMC3 assigns a Metropolis-Hastings sampler (MH) for binary latent variables such as $T_{i,j}$ and assigns a No-U-Turn Sampler (NUTS) for continuous variables such as q_{ijk}. MCMC parameters are set to values that aim to balance the trade-off between inference accuracy and computational efficiency, and PyMC3 has an auto-tuning step during the warmup/burn-in phrase to optimally tune the MCMC algorithm and speed up the convergence (Table 4.1). With this setting, the total number of iterations per chain is $(N_t + N_i)$. Since the N_t tuned samples are generated and auto removed during the sampling process, and N_b posteriors are thinned afterwards, the number of posterior samples for downstream analysis is $(N_i - N_b)$.

During the sampling, we printed out the divergence rate for each run of PGM and implemented the model in a way such that zero divergence was detected by PyMC3. Furthermore, we randomly selected some genes, plotted traces for each random variable in PGM as differences among chains can indicate problems with tuning and convergence. Also, we used summary() function of ArviZ [206] to check their Gelmen-Rubin split index R-hat values as MCMC chains can be assumed to be converged to the stationary distribution if R-hat values are less than 1.1 [207]. The convergence of chains was achieved for these selected genes, with R-hat values less than 1.1.

Given that the Markov chain can sometimes get stuck in the local minima [208], the PGM step is designed to run multiple times with different random initializations and then to average all those repeated posteriors for each T_{ij} to ameliorate this problem. InPheRNo-ChIP then forms a phenotype-relevant TRN based on the posterior probabilities of $T_{i,j}$s.

4.3 Experiments on Artificial Data

One of the main challenges in assessing unsupervised learning methods for TRN reconstruction is the absence of the underlying network (i.e., gold standard/ground truth) [209]. Since the experimental ground truth is unknown and our focus of the study is about embryonic development with emphasis on differentiation of hESCs to endoderm, there is a clear need to generate synthetic data that can statistically mimic some properties of RNA-seq and ChIP-seq data to test our model performance in a reproducible manner.

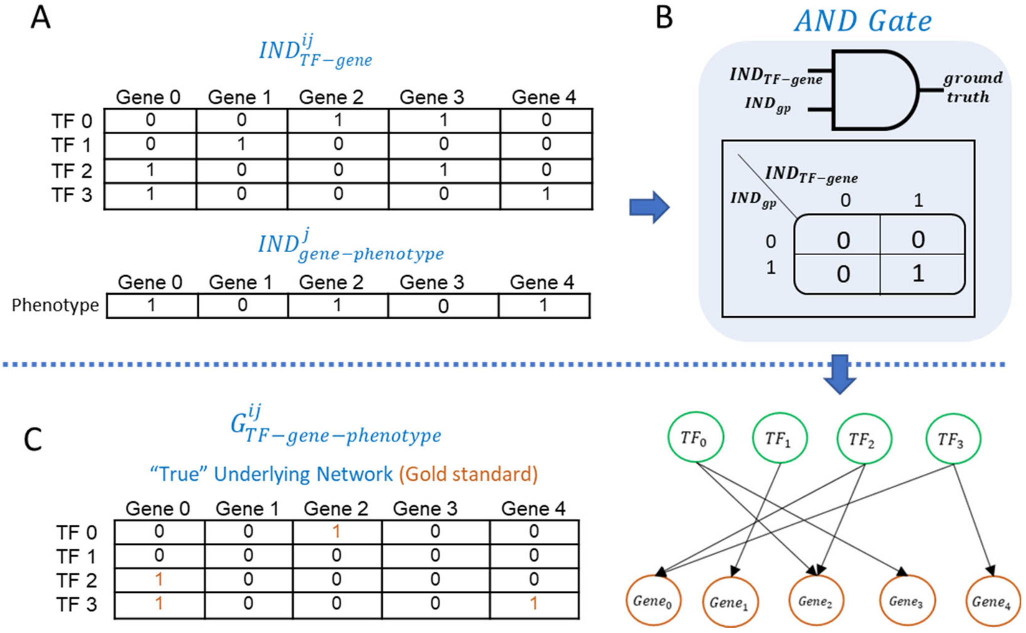
4.3.1 Assumptions

Although gene regulation is generally considered a nonlinear problem [210], linear models have simplicity as an asset and may give researchers insights into how some genes are regulated. To simplify the analysis of the data generation process, we made the following assumptions: (1) A gene must have at least one regulator (i.e., TF), and a TF must regulate at least one gene [211]; (2) The expression of the gene follows a linear relationship that embeds cellular phenotypes and noise.

In addition, based on our actual biological data from the human embryogenesis study, we fixed the sample size to 28 and the feature size (i.e., the number of TFs) to 25. We also synthesized the distribution of peak information from ChIP-seq data (see Chapter 4.4 for details).

4.3.2 Data Generation Process

We began by constructing a bipartite adjacency matrix $IND_{TF-gene}$ to indicate regulatory relationships between TFs and genes. Meanwhile, we constructed a binary vector IND_{gp} to indicate whether the expression of a gene is associated with biological conditions (e.g., case versus control).



$IND^{ij}_{TF-gene}$

	Gene 0	Gene 1	Gene 2	Gene 3	Gene 4
TF 0	0	0	1	1	0
TF 1	0	1	0	0	0
TF 2	1	0	0	1	0
TF 3	1	0	0	0	1

$IND^{j}_{gene-phenotype}$

	Gene 0	Gene 1	Gene 2	Gene 3	Gene 4
Phenotype	1	0	1	0	1

IND_{gp} \ $IND_{TF-gene}$	0	1
0	0	0
1	0	1

$G^{ij}_{TF-gene-phenotype}$

"True" Underlying Network (Gold standard)

	Gene 0	Gene 1	Gene 2	Gene 3	Gene 4
TF 0	0	0	1	0	0
TF 1	0	0	0	0	0
TF 2	1	0	0	0	0
TF 3	1	0	0	0	1

Figure 4.5 Illustration of the underlying network for synthetic gene expression data. (A) A bipartite adjacency matrix for TF-gene relationships from simulated RNA-seq data, denoted by $IND^{ij}_{TF-gene}$, whose element corresponds to the regulatory interaction between each pair of a TF i and a gene j. A binary vector for gene-phenotype relationships, denoted by IND^{j}_{gp}. A value of 1 indicates that there is a valid association between a gene and a TF (or a phenotype), and a value of 0 indicates the opposite. (B) A logic AND gate is used to determine the underlying network. The truth table represents all possible outcomes of input-output combinations for the gate. (C) A 0/1 matrix whose row indices correspond to TFs, column indices correspond to genes, and cell values correspond to regulatory interactions among TFs, genes, and phenotypes. A graph representation of the gold standard is also shown in the figure.

In a simple example (Figure 4.5), we illustrated the construction process of the ‘true’ underlying network using those two tensors. To build an index matrix $IND_{TF-gene}$, we used a Bernoulli distribution with parameter P_{tg} to determine whether a TF is regulating a gene or not. Similarly, to build an index vector IND_{gp} , we used another Bernoulli distribution with parameter P_{gp} to determine whether a gene is associated with a phenotypic label or not. Then, those two tensors were fed into a 2-input AND gate to obtain a new index matrix specifying which TF regulates which gene to affect its phenotype. The resulting index matrix $G_{TF-gene-phenotype}^{ij}$ (or G_{tgp}^{ij} for short), that came about as follows: for a specific pair of TF i and gene j, if $IND_{TF-gene}^{ij}$=IND_{gp}^{j}=1, then a value of 1 was given to G_{tgp}^{ij}. Otherwise, a value of 0 was given. The final matrix G_{tgp} was then used as the gold standard to evaluate the performance of InPheRNo-ChIP in the following chapter.

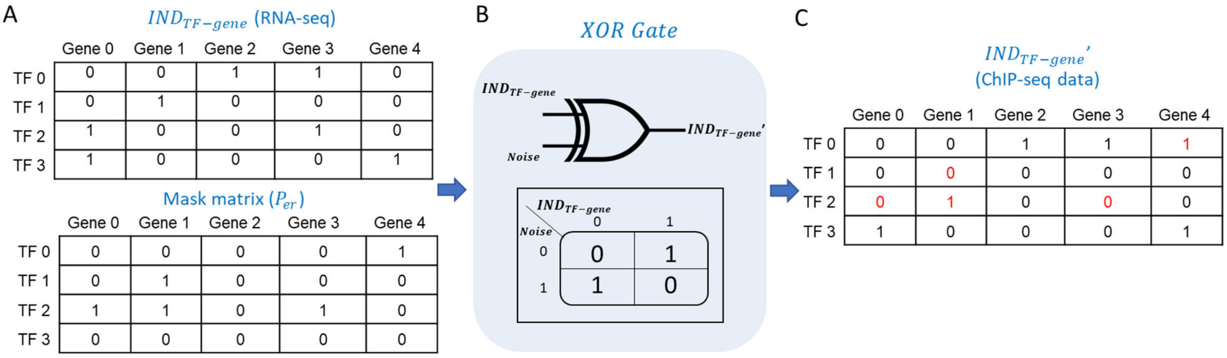
Next, we simulated gene expression data based on the ground truth G_{tpg}^{ij}. For each gene, the general form of its expression could be written as:

$$Gene_j = \begin{cases} noise, & G_{tpg}^{ij} = 0 \\ \alpha_0 TF_0 + \alpha_1 TF_1 + \cdots + \alpha_{24} TF_{24} + pheno_j + noise, & G_{tpg}^{ij} = 1 \end{cases}. \quad 4.13$$

In the case of $G_{tpg}^{ij} = 0$, a noise vector was assigned to the expression of the gene j. The noise vector was assumed to be normally distributed, whose parameter was computed based on the signal-to-noise ratio SNR_{dB} [212] defined by the user. We included this noise term because unpredictable features (i.e., biological/technical variation) are often observed in gene expression measurements [110]. In addition, adding noise to the synthetic data allowed us to access the generalization performance of the proposed method.

In the case of $G_{tpg}^{ij} = 1$, the response variable (i.e., a gene j’s expression level) was assumed to be a linear combination of features (i.e., expression of TFs), phenotypic information, and random noise. The parameter α_i in equation 4.13 was a coefficient varying from 0 to 1, which indicated the strength of the linear relationship between the expression of a specific TF i and the expression of gene j. The expression of TF was generated by a Gaussian distribution whose shape was equivalent to the sample size.

In addition to simulation of gene expression profiles, we aimed to synthesize ChIP-seq data to approximate real-world experimental data. One general strategy to generate such data is to draw samples from a real data distribution [213]. To achieve this, we plotted a real statistical distribution of the average number of peaks for all TF-gene pairs in T-Gene's processed data. Then, we randomly drew samples from this simulated population (Appendix A.4) to obtain the number of peaks for each TF-gene pair. Note that the number of peaks corresponds to the number of Distance p values obtained from the T-Gene algorithm.



	Gene 0	Gene 1	Gene 2	Gene 3	Gene 4
TF 0	0	0	1	1	0
TF 1	0	1	0	0	0
TF 2	1	0	0	1	0
TF 3	1	0	0	0	1

	Gene 0	Gene 1	Gene 2	Gene 3	Gene 4
TF 0	0	0	0	0	1
TF 1	0	1	0	0	0
TF 2	1	1	0	1	0
TF 3	0	0	0	0	0

Noise \ $IND_{TF-gene}$	0	1
0	0	1
1	1	0

	Gene 0	Gene 1	Gene 2	Gene 3	Gene 4
TF 0	0	0	1	1	1
TF 1	0	0	0	0	0
TF 2	0	1	0	0	0
TF 3	1	0	0	0	1

Figure 4.6 Illustration of the underlying network for synthetic ChIP-seq data. (A) As inputs, the XOR gate accepts an index matrix of the TF-gene association from RNA-seq, and a mask matrix specifies the level of mismatches between simulated RNA-seq and ChIP-seq data. The mask matrix was generated based on a user-defined error rate P_{er} with an acceptable range of 0% to 50% (Table 4.2). (B) A logic XOR gate with two inputs, one output and a truth table. (C) As the output, the XOR gate produces a new binary matrix $IND_{TF-gene}^{ij}$'. For a pair of TF i and gene j in the matrix: if $Noise_{ij}$ equals to 1, a logic complement value of $IND_{TF-gene}^{ij}$ will be assigned to the pair (red-colored values); otherwise, the same value of $IND_{TF-gene}^{ij}$ will be assigned to the pair (black-colored values).

To test the effects of model performance on the integration of RNA-seq and ChIP-seq data when different data types contribute complementary or slightly contradictory information to each other, we generated ChIP-seq data based on two scenarios:

(1) Ideal scenario: we hypothesized that the regulatory evidence from ChIP-seq was complementary to that from RNA-seq. Therefore, we used the same bipartite adjacency matrix from RNA-seq data (i.e., $IND_{TF-gene}$) as the underlying network to generate ChIP-seq data. The underlying network indicates whether a TF regulates a gene.

(2) Mismatch scenario: we hypothesized that the regulatory evidence from ChIP-seq data was slightly contradictory to that from RNA-seq data. Therefore, we constructed a new Boolean network (denoted as $IND_{TF-gene}'$) by applying a logical exclusive-OR gate to the noise matrix (denoted as N) and the original index matrix $IND_{TF-gene}$: $IND_{TF-gene}' = \mathrm{XOR}(IND_{TF-gene}, \mathrm{N})$. A simple example to illustrate this process is shown in Figure 4.6.

It is worth mentioning that the ground truth for both scenarios remain the same: a binary matrix G_{tgp}^{ij} where a value of 1 indicates a valid TF-gene-phenotype association and a value of 0 indicates its logical complement (Figure 4.5).

Considering the randomness and reproducibility of the data generation process, we ran the data generation process ten times with different global random seeds (using the 'random' module from the 'NumPy' library) and obtained ten sets of synthetic data for each ideal/mismatch scenario. The detailed pseudo-code for the data generation process can be found in Appendix A.1.

4.3.3 Results

As discussed in Chapter 4.2.2, there are two options available for obtaining the p values of TF-gene associations in step 2 of InPheRNo-ChIP:

- *Model A*: InPheRNo-ChIP with Pearson correlation as step 2.
- *Model B*: InPheRNo-ChIP with OLS method as step 2.

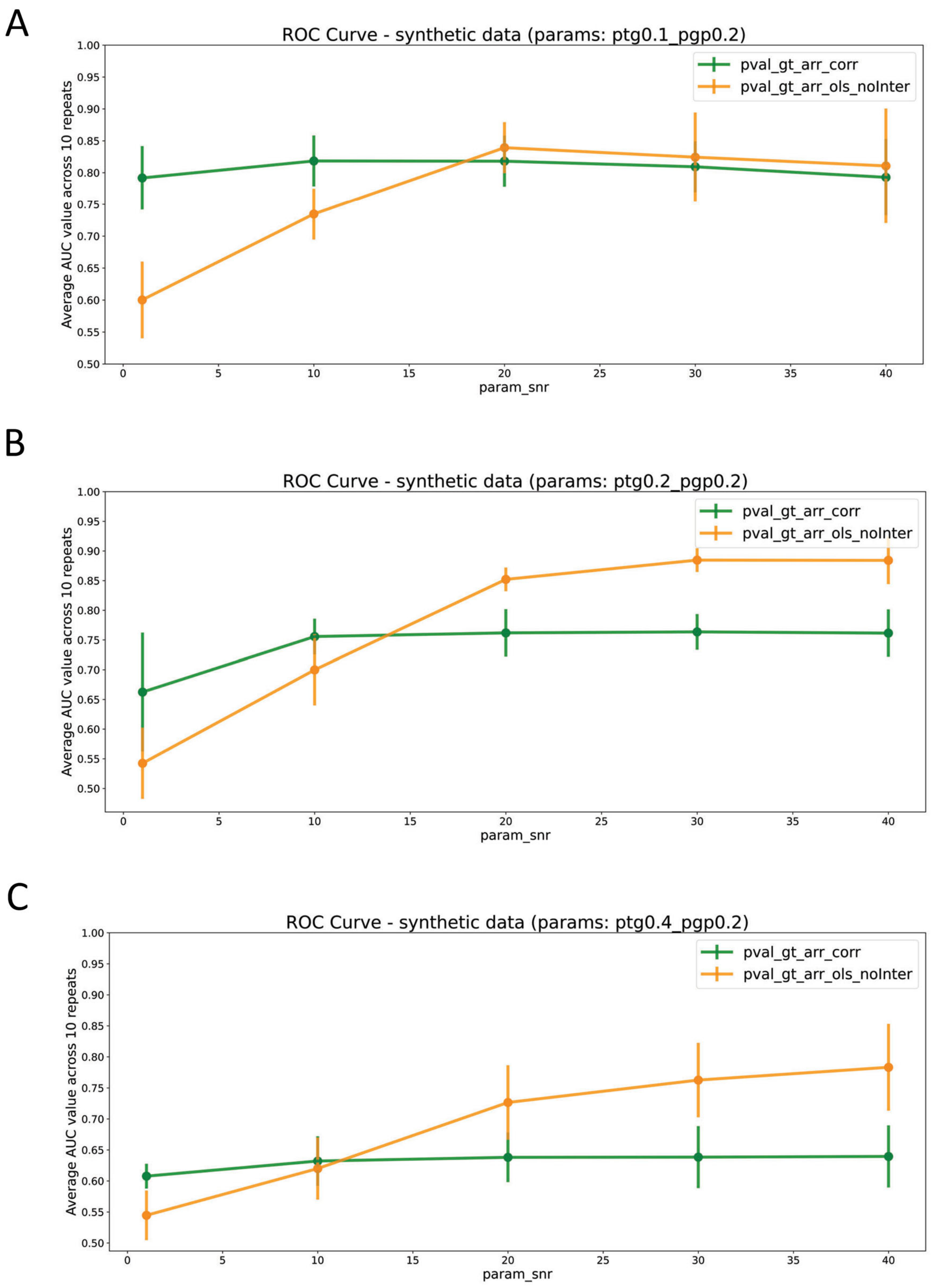
Thus, we conducted performance tests on these two options to examine which method was more appropriate for further analysis. In terms of the evaluation metric, we used the Area Under Receiver Operating Characteristic (AUROC) [214], also known as the area under the ROC curve. It measures both the specificity (i.e., $\frac{\mathrm{TN}}{\mathrm{TN+FP}}$) and sensitivity (i.e., $\frac{\mathrm{TP}}{\mathrm{TP+FN}}$) of continuous variables across all possible thresholds range from 0 to 1. The false negative (FN) is defined as the case where a TF-gene-phenotype link can be found in the ground truth network but cannot be found in the reconstructed network. The true positive (TP) is defined as those links on the ground truth network that remain on the reconstructed network, and the false positive (FP) is defined as the number of false links inferred in the reconstructed network.

We aimed to evaluate two models using only synthetic gene expression data with different values of P_{gp}, P_{tg} and SNR (Table 4.2) to have a baseline understanding of their performance. Note that there was no P_{ER} involved because we were not including ChIP-seq information.

Table 4.2 User-definable parameters in the synthetic data generation.

Parameter	**Description**	**Value**
P_{tg}	The probability of a TF is regulating a gene	[0.1, 0.2, 0.4, 0.6, 0.8]
P_{gp}	The probability of a gene is associated with phenotype	[0.1, 0.2, 0.4, 0.6, 0.8]
SNR_{dB}	Singal-to-noise ratio for gene expression data, in dB	[1, 10, 20, 30, 40, 50]
P_{er}	The error rate in the ChIP-seq underlying network	[0.05, 0.1, 0.15, 0.25, 0.5]

In Figure 4.7, we observed that both models were pretty sensitive to P_{tg}: as we increased the value of P_{tg} from 0.1 to 0.8, the averaged AUROC scores of both models dropped from 0.75 to somewhere close to 0.55. Meanwhile, we observed better performance of Model A compared to model B when SNR was relatively small (e.g., SNR=5 or SNR=10). On the other hand, model B tended to achieve better AUROC results than model A when $SNR \geq 20$ (regardless of the choice of P_{tg}). Note that we only included AUROC results for the fixed $P_{gp} = 0.2$ because both models were robust with respect to the change of P_{gp}.


A
ROC Curve - synthetic data (params: ptg0.1_pgp0.2)
pval_gt_arr_corr
pval_gt_arr_ols_noInter
Average AUC value across 10 repeats
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0
5
10
15
20
25
30
35
40
param_snr
B
ROC Curve - synthetic data (params: ptg0.2_pgp0.2)
pval_gt_arr_corr
pval_gt_arr_ols_noInter
Average AUC value across 10 repeats
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0
5
10
15
20
25
30
35
40
param_snr
C
ROC Curve - synthetic data (params: ptg0.4_pgp0.2)
pval_gt_arr_corr
pval_gt_arr_ols_noInter
Average AUC value across 10 repeats
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0
5
10
15
20
25
30
35
40
param_snr

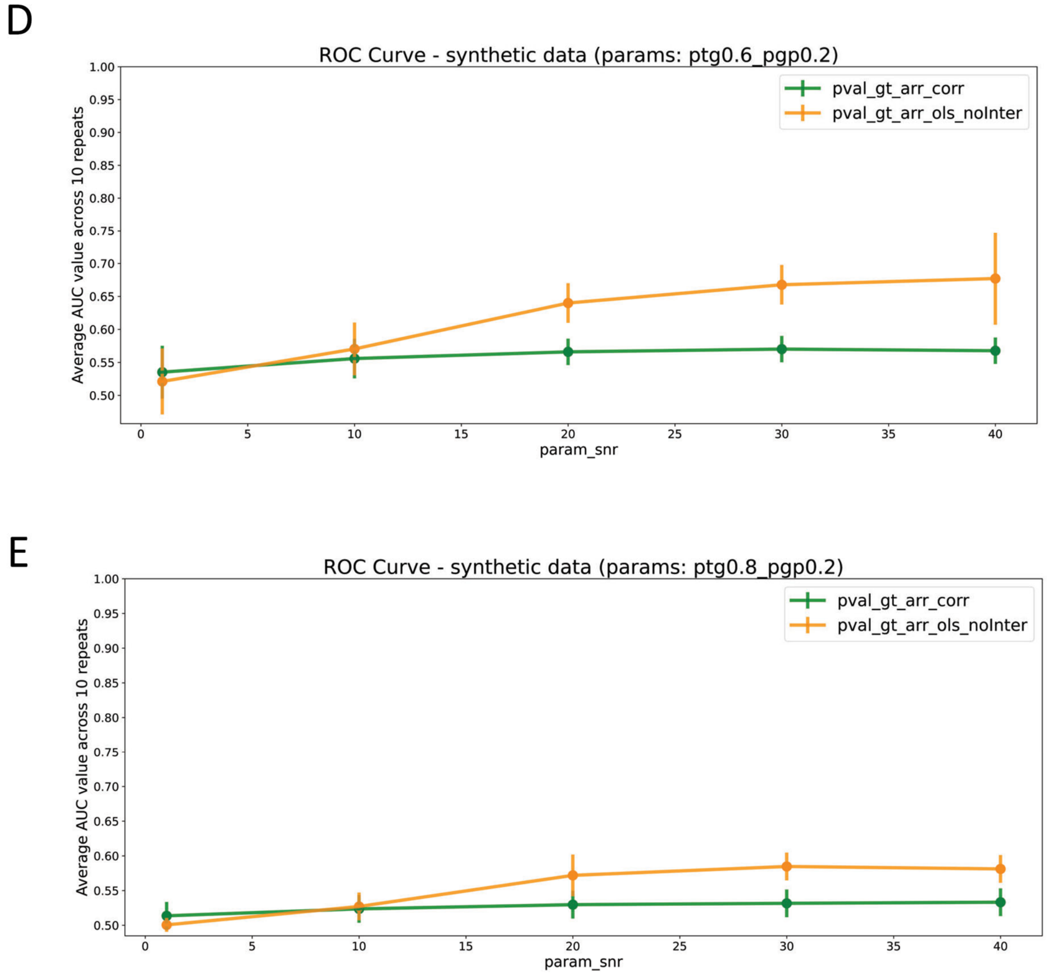


Figure 4.7 Average AUROC performance using only simulated RNA-seq data. Here, model A and model B were trained on 10 randomly generated gene expression profiles with various SNR and P_{tg}. Each subplot depicts the overall tendency for a particular value of P_{tg} in Table 5. In each figure, each dot represents the mean AUC scores across ten different TRNs generated by InPheRNo-ChIP, and the error bar overlays on each dot representing the standard deviation of the data. The Y-axis shows the AUROC score ranges from 0 to 1, and the x-axis shows the variation in the noise levels. The green-coloured line shows the performance of the model when Pearson correlation is used to compute the p-value of TF-gene associations, while the black line shows the AUC value when OLS is used.

In order to investigate model performances with respect to different error rates, we conducted two experiments: (1) fixing $P_{tg} = 0.2, P_{gp} = 0.2, SNR = 20$ and varying the error rate P_{er} from 5% to 50% (2) fixing $P_{tg} = 0.4, P_{gp} = 0.2, SNR = 20$ and varying the error rate P_{er} from 5% to 50%. We randomly generated ten datasets for each experiment and computed the average AUC score across those datasets concerning different error rates for each model.

As shown in Figure 4.8, increasing the error rate from 5% to 50% showed proportional decreases in model A performance (yellow-coloured line) and model B performance (purple-coloured line). The observed tendency indicated that a high perception of contradictory information from multi-data sources could result in noticeable changes in AUROC performance.

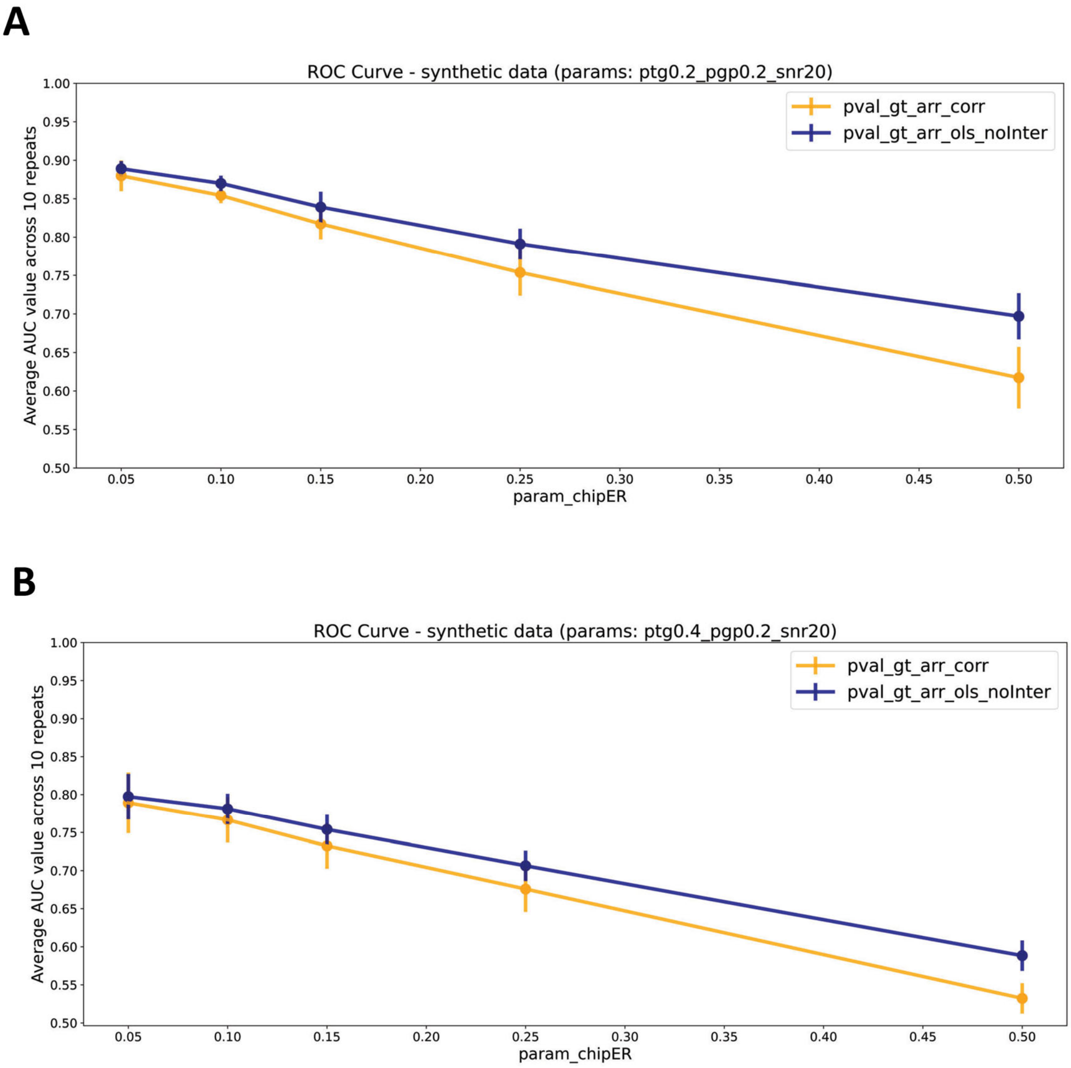


Figure 4.8 Average AUROC performance using simulated RNA-seq and ChIP-seq data. Here, model A and model B were trained on 10 randomly generated synthetic datasets with various P_{er} and P_{tg}. (A) The figure depicts the result obtained from experiment #1: fixing $P_{tg} = 0.2, P_{gp} = 0.2, SNR = 20$ and varying the error rate P_{er} between 5% and 50%. (B) The figure depicts the results obtained from experiment #2: fixing $P_{tg} = 0.4, P_{gp} = 0.2, SNR = 20$ and varying the error rate P_{er} between 5% and 50%.

Also, regardless of how much noise was generated or what value was used for P_{tg}, we observed that the performance of model B (i.e., InPheRNo-ChIP with OLS model as step 2) was generally better than that of model A (i.e., InPheRNo-ChIP with simple correlation analysis as step 2).

To test the hypothesis that incorporating ChIP-seq and RNA-seq improves the performance of InPheRNo-ChIP compared to that of utilizing only gene expression data, we combined ideal and mismatch scenarios in chapter 4.3.2 and introduced the following settings:

- ***Setting I**:* testing the model performance on gene expression data only, with simulation parameters: $P_{gp} = 0.2, P_{tg} = 0.2, SNR = 20$.
- ***Setting II and III*** (ideal/mismatch scenarios): testing the model performance on perfectly/partially matched RNA-seq and ChIP-seq data, with simulation parameters: $P_{gp} = 0.2, P_{tg} = 0.2, SNR = 20, P_{er} = 0.1\ or\ 0$.
  - *Setting II:* If $P_{er} = 0$, then the ChIP-seq data is perfectly matched to the underlying regulatory relationships in gene expression data.
  - *Setting III:* If $P_{er} = 0.1$, then the underlying network for simulated ChIP-seq data is partially matched (or mismatched) to that for gene expression data.

Table 4.3 reports AUROC scores for models in different settings. Each cell contains the mean and standard deviation for AUC measurements over ten reconstructed TRNs with different random initializations.

Table 4.3 Performance comparison of model A and model B in different settings. The AUROC is indicative of the overall performance of the inference algorithm [121]: a perfect predictor tends to give an AUC score of 1, while a random predictor will have a value close to 0.5. If the score is less than 0.5, it implies that the algorithm does not confer any predictive power (its performance is worse than random selection).

	Setting I: RNA-seq only	**Setting II: RNA-seq +** ***ideal*** **ChIP-seq ($P_{er} = 0$)**	**Setting III: RNA-seq +** ***mismatch*** **ChIP-seq ($P_{er} = 10\%$)**
Model A	0.762±0.036	0.916±0.01	0.854±0.013
Model B	0.852±0.017	0.919±0.015	0.870±0.009

As depicted in Table 4.3, regardless of choice in estimating TF-gene associations in step 2 of InPheRNo-ChIP (i.e., Pearson correlation versus OLS method), it was apparent that the integration of ChIP-seq data resulted in a vast improvement. In the first row, the accuracy of results was increased from 76.2% to 91.6% when ChIP-seq information was perfectly matched RNA-seq's story. This value dropped to 85.4% when the error rate was 10%, which was still more remarkable than the AUROC score using only gene expression data. A similar conclusion can be drawn for the 2nd row of the table, where we used OLS as the first step to estimate TF-gene associations.

In summary, we conducted several experiments on synthetic data and were able to explore the behaviour of our proposed method using different input data (1) RNA-seq data only; (2) RNA-seq with perfect ChIP-seq data that matches the underlying assumption; and (3) RNA-seq with mismatch ChIP-seq information; and different options (e.g., model A and model B). We evaluated the performance of InPheRNo-ChIP on our simulated data in different settings by computing the corresponding AUROC scores. Overall, the results indicated that Model B (utilizing the OLS method as the second step of InPheRNo-ChIP) would be more appropriate to determine the TF-gene associations. Therefore, we employed this model as the default option in our algorithm.

Although the construction process was biased and the model's effectiveness was partially dependent on the quality of the data generation process and the choice of parameters, these results indicated that the integration of ChIP-seq and RNA-seq could still improve the TRN reconstruction accuracy compared to utilizing only gene expression data. Another limitation of this experiment was that synthetic data might not mimic the complexity of regulatory mechanisms in the living organism. For example, although we used SNR to include additive noise in the gene expression, it is still unlikely to be a good approximation of the technical/biological noise resulting from a series of complex experimental processes in the real world. Despite these limitations, we believe synthetic data generation can still provide some guidance to help us evaluate the performance of our algorithm under controlled conditions and set a ground truth for different measures.

4.4 Experiments on Real Data

Having analyzed the impact of integrating different data types on the resulting performance in the previous chapter, we now take a closer look at a real-world application of InPheRNo-ChIP in human embryogenesis.

As mentioned in chapter 4.1.1, hESCs have the potential to self-differentiate into ectodermal, mesodermal, and endodermal cells of multiple tissues. It is widely used in research fields such as early embryonic development, diseases, epigenetics, and human pathophysiology and provides an unlimited source of multiple tissue cells for cell replacement therapy in regenerative medicine, which is of utmost clinical importance. As a use case, we applied the proposed InPheRNo-ChIP algorithm to multi-omics data from the GEO repository [215], and reconstructed a TRN unravelling the underlying molecular mechanisms in the embryonic specification of endoderm. The inputs to InPheRNo-ChIP were published RNA-seq and ChIP-seq data obtained from several human embryogenesis studies, with an emphasis on employing samples from hESCs and endoderm lineages. The output of the algorithm was a phenotype-relevant transcriptional regulatory network, whereas phenotypic labels (i.e., hESC or EN) indicate the group information for each sample. In this study, we hypothesized that integrating multi-omics data would potentially improve the predictive performance of our model.

4.4.1 Data Sources

We downloaded a list of 1665 human TFs from a comprehensive animal TF database, AnimalTFDB 3.0 [216], on February 4th, 2021.

We downloaded TF ChIP-seq data from the GEO repository [215] under accession GSE61475. This dataset contains 204 ChIP-seq samples, generated using an MNase-based ChIP-seq technique from early stages of endoderm, mesoderm, ectoderm and mesendoderm tissues derived from human embryonic stem (ES) cells.

Furthermore, we had searched GEO for related gene (including TF) expression profiles. Two keywords, "endoderm" and "hESCs differentiation," were used to identify potential human datasets of interest. As of April 4th, three GSE datasets GSE164361, GSE143371, and GSE160981, have been considered for this use case. The details of the RNA-seq profiles in terms of sample information are given below:

The GSE164361 dataset contains gene expression data measured across sets of neural progenitor and definitive endoderm lineage differentiation in rat and homo sapiens (118 samples in total). The GSE143371 dataset contains 55 samples corresponding to transcriptional profiles of four lineages (i.e., definitive endoderm, early mesoderm or neuroectoderm, and embryoid bodies) differentiated from mutant and control hESCs, with two or three replicates each. The GSE160981 dataset contains 24 cell line samples, including pre-treated hESCs (i.e., H9 cells) and differentiating lineages for three days with three biological replicates generated by RNA-seq technology.

Complete experimental details are available in reference publications in the summary table in Appendix A.2.

4.4.2 RNA-seq Data Processing

In order to obtain two sets of p values of TF-gene associations and gene-phenotype associations, we first identified consistently up/down-regulated genes in all selected datasets using two R packages: EdgeR (Empirical Analysis of Digital Gene Expression Data in R) [163] and Limma (Linear Models for Microarray Data) [164].

The essential steps in the preprocessing of analyzing bulk RNA-seq data are outlined in Figure 4.9. The pipeline started with downloading gene-level counts from the Gene Expression Omnibus (GEO) repository for a given GSE accession number. After downloading gene expression profiles, we kept samples (see the table in Appendix A.2) relevant to hESCs and endoderm lineages because those two lineages were our focus in this application. As a result, the total number of samples across multiple datasets was narrowed down to 28 samples: 12 samples for GSE164361, ten samples for GSE143371 and six for GSE160981.

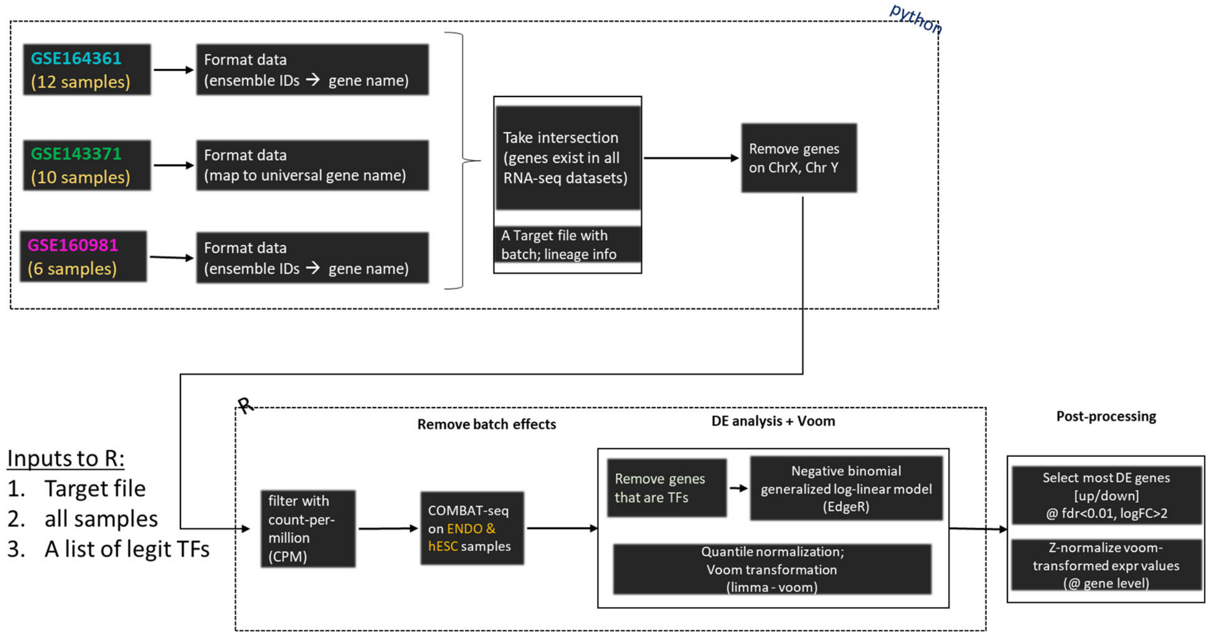

python
GSE164361
(12 samples)
Format data
(ensemble IDs → gene name)
GSE143371
(10 samples)
Format data
(map to universal gene name)
GSE160981
(6 samples)
Format data
(ensemble IDs → gene name)
Take intersection
(genes exist in all RNA-seq datasets)
A Target file with batch; lineage info
Remove genes on ChrX, Chr Y
Inputs to R:
1. Target file
2. all samples
3. A list of legit TFs
R
Remove batch effects
filter with count-per-million (CPM)
COMBAT-seq on ENDO & hESC samples
DE analysis + Voom
Remove genes that are TFs
Negative binomial generalized log-linear model (EdgeR)
Quantile normalization; Voom transformation (limma - voom)
Post-processing
Select most DE genes [up/down] @ fdr<0.01, logFC>2
Z-normalize voom-transformed expr values (@ gene level)

Figure 4.9 RNA-Seq data processing and gene expression analysis.

To leverage heterogeneous gene annotations across different data sources, we translated common gene IDs (e.g., Ensembl gene IDs [217] or Entrez gene IDs) to universal gene symbols. Furthermore, we filtered autosomal genes.

As previously mentioned, out of 200 raw samples from the GEO repository, only 28 fulfilled our planned criteria (i.e., samples correspond to hESCs or hESC-derived endoderm). To overcome the problem of the small sample size in each dataset, we decided to aggregate three datasets together and then performed a systematic analysis on the aggregated dataset. However, unwanted noise and unmodeled artifacts could emerge if we integrated datasets processed at different times and in different facilities [218]. These batch effects sometimes create bias and add variability to the results of biology experiments, which may cause a dramatic reduction in the accuracy of statistical inference later. Thus, it is critical to adjust batch effects on the aggregated RNA profiles and a package ComBat-seq [218] was introduced in this process (Appendix A.3).

After correcting batch effects, we performed DE analysis (using EdgeR [163]) and quantile transformation (using Limma-Voom [164]) on the remaining genes. For DE analysis, the goal was to find DEGs between two or more phenotypic labels (e.g., biological/treatment conditions) and compute the statistical significance of the DEGs. In general, a gene is differentially expressed if a statistically significant change is observed in expression levels between two or more conditions [198].

The inputs to EdgeR were (1) a text file containing raw read counts after aggregation and (2) a target file specifying sample information (e.g., lineages and confounding factors in experiments). EdgeR first filtered out lowly expressed genes (i.e., genes that have CPM below 1 CPM in less than two libraries), normalized the library size and fitted a negative binomial generalized log-linear model [46] (and corrected the confounding effects of day information) to the remaining counts for each gene in each sample. The output of EdgeR was a list of p values presenting gene-phenotype associations. With Benjamini & Hochberg FDR (False discovery rate) strictly smaller than 0.01 (i.e., $\mathrm{FDR} < 0.01$) and absolute fold change greater than 2 (i.e., $|\mathrm{logFC}| > 2$) as the screening criteria, we obtained a total of 2298 DEGs that were all consistently up-/down-regulated in the aggregated dataset.

Separately, we performed a quantile normalization for the genes from gene expression profiles using the package Limma-Voom and performed the z-score transformation across the samples on the filtered counts. The transformed values were then used to compute p values of the associations among genes and their regulators (π_{ij}s).

4.4.3 ChIP-seq Data Processing

Similar to RNA-seq analysis, we also limited our ChIP-seq dataset (GSE61475) to hESCs and endoderm lineages. The GSE61475 dataset contains multiple peak files identified by the MACS algorithm using aligned ChIP-Seq reads [219]. Each peak file contains the identified peak locations together with peak summits and statistical significance scores such as p values and q values. The flow of preprocessing is shown in Figure 4.10, and three existing packages were applied to accomplish the steps: Irreproducibility Discovery Rare (IDR) analysis from ENCODE [220, 221]; BEDTools [222]; and T-Gene [204].

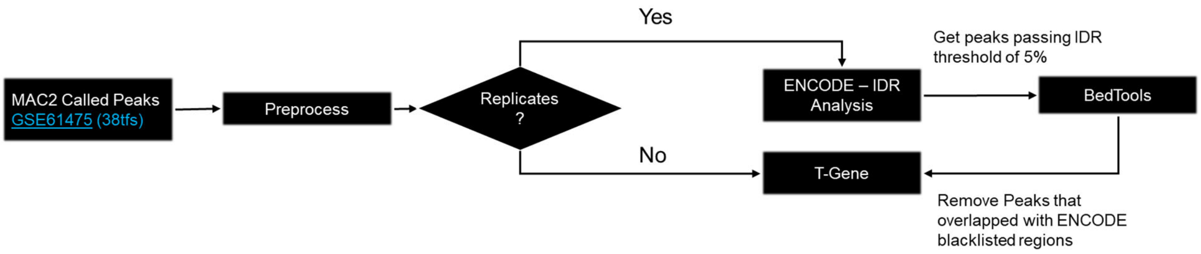

Yes
Get peaks passing IDR threshold of 5%
MAC2 Called Peaks
GSE61475 (38tfs)
Preprocess
Replicates ?
ENCODE – IDR Analysis
BedTools
No
T-Gene
Remove Peaks that overlapped with ENCODE blacklisted regions

Figure 4.10 illustration of the pipeline for ChIP-seq Pre-processing.

Among these ChIP-seq BED files, we first removed samples of (1) ES cell lines derived from shRNA-mediated knockdown of GATA4 and differentiation toward endoderm (labelled as "dEN_shGATA4" in the series matrix file downloaded from the GEO database) and (2) lineages from mesoderm or ectoderm. After that, we filtered the antibodies (TFs) with a human TF list obtained from HumanTFDB. This filtering step led to 10 unique TFs for ectoderm, 25 unique TFs for mesoderm, 24 TFs for endoderm, 22 for mesendoderm, and 29 TFs for hESCs (Table 4.4).

Next, we examined the data quality of ChIP-seq replicates and aimed to find the highly reproducible peaks for each TF using the IDR method [221]. Since IDR analysis took a pair of peaks as input, we had to categorize our samples from different lineages into 3 cases: samples with one replicate, samples with two replicates, and samples with more than two replicates (Table 4.4).

Table 4.4 MNChIP-seq data sample information in GSE61475. Shaded rows are lineages of interests. dEN designates samples of hESC-derived endoderm, HUES64 designates samples of hESCs. dEC designates samples of hESC-derived ectoderm, dMS designates samples of hESC-derived mesendoderm, and dME designates samples of hESC-derived mesoderm.

Cell Type	Number of Unique TFs	1 rep	2 reps	>2 reps
dEC	10	9	1	0
dMS	25	23	2	0
dEN	24	13	7	4
dME	22	18	3	1
HUES64	29	21	3	5

We ran IDR analysis on TF files with two replicates and ignored samples with only one replicate. Because ENCODE's IDR package [221] can only handle two replicates simultaneously, we first obtained all pairwise combinations from an individual TF's replicates, performed IDR on each pair, and then selected the paired one with the most significant number of peaks post-merging. As a final step in IDR analysis, we removed peaks that failed to pass the IDR threshold of 0.05. Here, an IDR score of 0.05 is equivalent to a score of $\mathrm{int}(-125\log2(0.05)) = 540$ in the outputted IDR table and a higher IDR score stands for a more reproducible peak [221].

Since the blacklisted regions (e.g., unstructured/anomalous reads) were reduced in the latest genome assemblies (GRCh38), "blacklist filtering" has been excluded in most analyses. However, since our dataset was produced in 2016 and used hg19 as a reference genome, it was still best practice to remove these troubling regions in the preprocessing pipeline. To achieve this, we used the intersect() function from BEDTools [222] to intersect IDRed peaks with ENCODE blacklist [223] and only kept the ones that did not appear in the blacklist. As a result, we obtained 21 endoderm-specific TFs and 28 hESC-specific TFs.

After removing IDRed peaks in blacklisted regions, we used T-Gene to compute the likelihood that a TF regulates a putative gene in a cell line of interest. The inputs to T-Gene were processed peak files. Inside each file, each row corresponded to a peak identified by MAC and passed IDR analysis; each column corresponded to the chromosome's name and start/end coordinates of a peak. We ran T-Gene with default settings and obtained a list of p values corresponding to the TFs of interest and their putative regulatory targets. In addition, we filtered the links with 'Closest_TSS=T' and only kept links with Distance p values smaller than 1, as shown in the histogram (Appendix A.5). As a result, we obtained a third set of p values of TF-gene associations from ChIP-seq peaks.

4.4.4 Results

The p values of TF-gene associations from ChIP-seq, the p values of TF-gene associations from RNA-seq, along with the p values of gene-phenotype associations, are fed into InPheRNo-ChIP to obtain a TRN specific to hESC-derived endoderm.

Instead of using an arbitrary value of 0.5 to threshold the confidence scores in the inferred network (as in chapter 3), we employed a Kneedle algorithm [224] to eliminate any edge with a confidence score lower than the knee/elbow score. This algorithm is a mathematical tool aiming to find the maximal curvature of a decreasing convex curve, and the result is highly dependent on the choice of sensitivity (S). In order to select the appropriate value of S, we converted the SvOV TRN to a ranked edge list according to the confidence score (Figure 4.11). We observed a decreasing number of qualified TF-gene edges as we increased the value of S from 1 to 100. Since no edge was detected using $S = 1$, we chose $S = 2$ as the final sensitivity value to report.

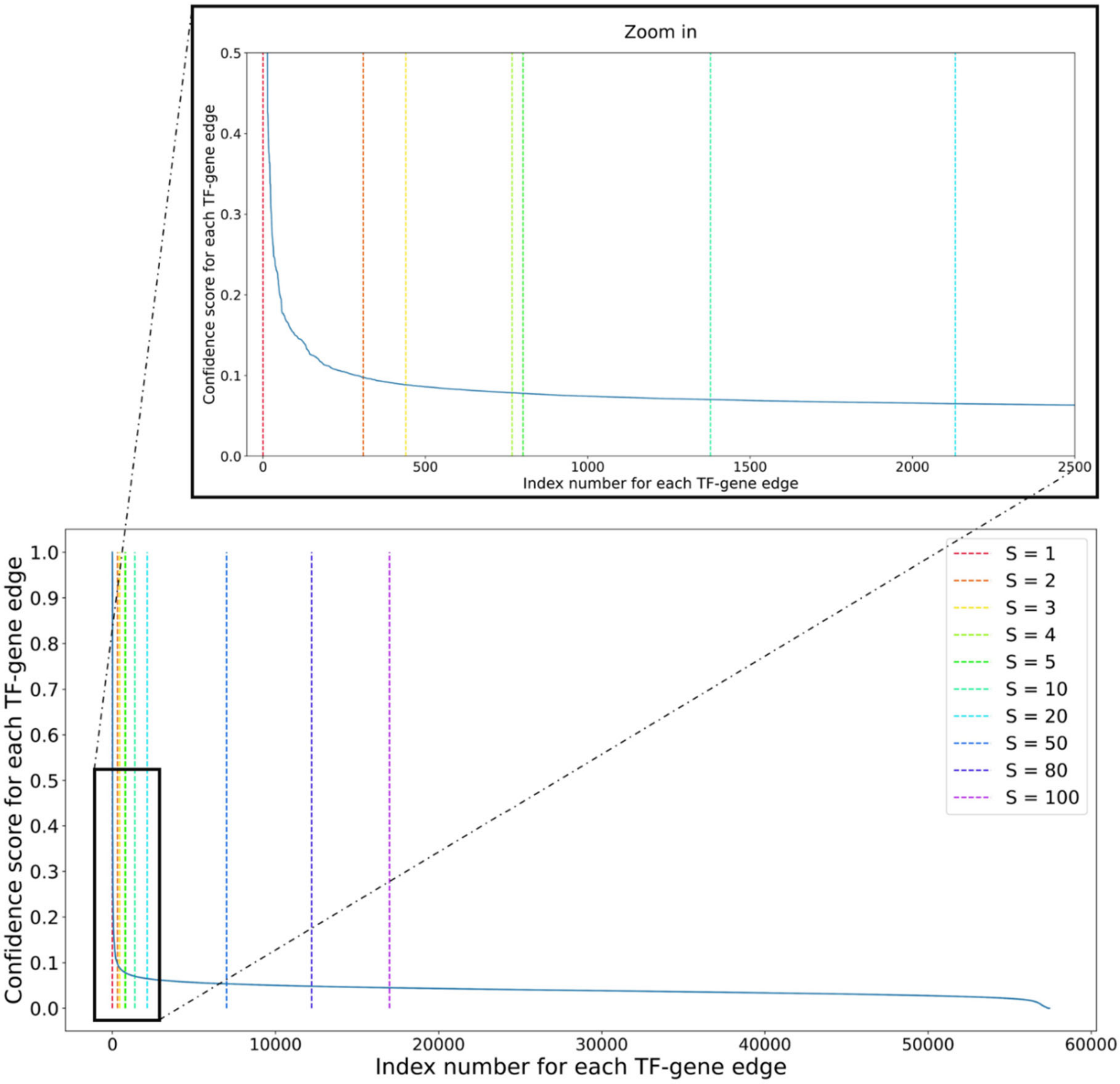


Figure 4.11 Illustration of the Kneedle algorithm with different sensitivity values. We converted the SvOV TRN to an edge list, and then ranked the list by the confidence score for each identified TF-gene pair. Each pair has its unique index number to differentiate itself from other pairs. In both plots, the x-axis depicts the index of each identified TF-gene edge in the ranking list (e.g., index #0 depicts the 1st TF-gene edge with the highest confidence score, index #1 depicts the 2nd TF-gene edge in the list), and the y-axis depicts the confidence score for each TF-gene association. The plot at the top depicts the zoom-in view of the bottom one.

As depicted in Figure 4.12, after applying the Kneedle algorithm with sensitivity $S = 2$ to InPheRNo-ChIP identified TF-gene edges (consisting of 25 TFs and 2298 DEGs), we obtained 310 TF-gene edges (consisting of 24 TFs and 259 DE genes).

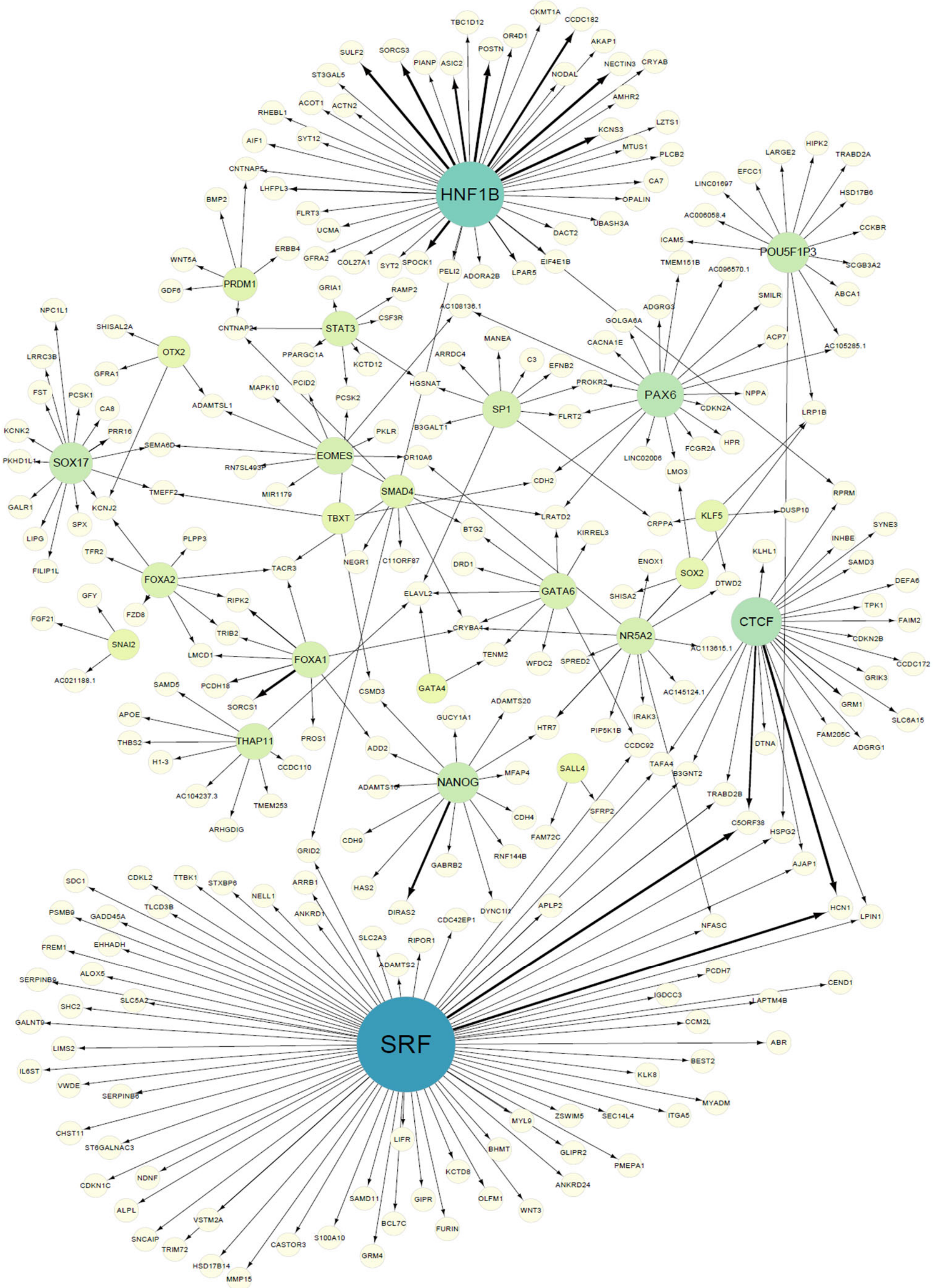


Figure 4.12 Network representation of the filtered endoderm network. In this network, nodes correspond to identified TFs/genes in the endoderm network; directed edges correspond to regulatory effects of TFs (gradient colours from cyan blue to light green) on their targets (light yellow colour). The size and colour of each TF node are associated with its out-degree: a TF with more target genes will have a darker and a larger node; and the size and colour of each Gene node are fixed. The edge thickness is associated with the confidence score assigned by the algorithm. Notably, the filtered network only includes interactions with a confidence score above 0.0974 (which is determined by the Kneedle algorithm with S=2). Overall, the filtered endoderm network comprises 310 edges, including 24 TFs and 259 target genes.

Next, we ranked the filtered TFs based on their number of target genes and listed the percent of target genes for each TF_i in Table 4.5. As proof of predictions, we mined literature to search for evidence regarding the role of top TFs (i.e., TFs with the most significant number of phenotype-relevant targets) in early human endoderm development. It has been acknowledged that the knockdown of CTCF increases the contact between endodermal enhancers and IGF2 promoters [225]. Also, there is some evidence that PAX6 is a key factor in the differentiation of pancreatic endoderm derived from hESCs [226], and SOX17 is an indispensable factor in the differentiation of extraembryonic endoderm (ExEn) cells [227, 228]. Furthermore, Li et al. [229] studied the functions of SMAD4 in mouse embryos and found that the lack of SMAD4 can lead to the failure of anterior embryonic patterning and head induction.

Table 4.5 List of top TFs identified by InPheRNo-ChIP and evidence for their roles in human embryogenesis. The evidence is denoted as "strong" if more than three pieces of literature evidence were found.

TFs	Genes	Percent of target genes	Literature Evidence
SRF	75	28.96%	Strong
HNF1B	42	16.22%	Moderate
CTCF	23	8.88%	Strong
PAX6	19	7.34%	Strong
SOX17	15	5.79%	Strong
NANOG	14	5.41%	Strong
POU5F1P3	14	5.41%	Moderate
NR5A2	10	3.86%	Moderate
EOMES	10	3.86%	Strong
SP1	10	3.86%	Moderate
THAP11	9	3.47%	Strong
GATA6	9	3.47%	Strong
FOXA1	9	3.47%	Strong
STAT3	8	3.09%	Strong
FOXA2	8	3.09%	Strong
SMAD4	6	2.32%	Strong
PRDM1	6	2.32%	Strong
SOX2	4	1.54%	Moderate
OTX2	4	1.54%	Strong
TBXT	4	1.54%	Moderate
KLF5	4	1.54%	Strong
SNAI2	3	1.16%	Weak

Next, we ranked the TF-gene relationships according to their confidence scores in the inferred network and extracted the top 20 edges with the highest confidence scores in Table 4.6. We validated the identified 310 TF-gene edges using experimentally validated TF-gene associations in the mouse embryogenesis ESCAPE database [230]. The ESCAPE database contains TF-gene regulatory evidence in mouse embryonic stem cells (mESC) based on loss-of-function or gain-of-function. We performed a hypergeometric test and found that the InPheRNo-ChIP edges are enriched in experimental TF-gene mESC edges with a significant p-value of 9.20e-09.

Table 4.6 Top 20 TF-target edges in the inferred TRN with their confidence scores. A higher score indicates a stronger TF-Gene-Phenotype association.

TFs	Genes	Scores
HNF1B	KCNS3	1
HNF1B	SORCS3	1
HNF1B	POSTN	0.9998
HNF1B	NECTIN3	0.9980
HNF1B	SULF2	0.9895
SRF	C5ORF38	0.9843
SRF	HCN1	0.9619
CTCF	HCN1	0.9486
FOXA1	SORCS1	0.9419
HNF1B	SPOCK1	0.9106
NANOG	DIRAS2	0.8749
HNF1B	CCDC182	0.8631
HNF1B	ASIC2	0.8343
CTCF	C5ORF38	0.6713
CTCF	GRM1	0.4267
HNF1B	LPAR5	0.4224
SOX17	PRR16	0.4036
HNF1B	LHFPL3	0.3874
FOXA1	RIPK2	0.3767
FOXA2	FZD8	0.3747

Next, we performed gene ontology (GO) analysis by using the standard GSC computational pipeline (with the Fisher's exact test) of KnowEnG [10] to assess functional relationships between identified TFs' targets. Then, we performed multiple testing corrections for KnowEnG results using the standard Benjamini and Hochberg's False Discovery Rate (BH-FDR) analysis. As a result, we found that the regulator PRDM1 was associated with some critical GO terms such as cell fate commitment, epithelial to mesenchymal transition and positive regulation of cartilage development.

In addition, we performed GO analysis by using the advanced network-guided GSC computational pipeline of KnowEnG [10] to assess functional relationships between identified TFs' targets. This advanced mode encodes prior knowledge of gene-gene interactions (e.g., an experimentally verified protein-protein interaction (PPI) network from the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database [231]) into its statistical analysis of the input query data set. Such knowledge-guided analysis allows more important GO terms to be exploited in grouping gene sets. We first formed a matrix of genes (rows) by TFs (columns), which embedded the regulatory effect of each TF on each gene, and then we analyzed this matrix through a network-guided (using STRING PPI network) GSC pipeline with default parameters. Heatmap of the results from GSC analysis with GO terms was drawn in rows and TFs in columns (Figure 4.13). For a more straightforward presentation of important GO terms that were associated with TFs (and their target genes), we chose a cut-off value of 0.5 for the difference score and only included the enriched GO terms that were implicated for at least two TFs (and their target genes) in the final TRN. Notably, GO terms related to the control of dendrite development and eye development were implicated in this analysis.

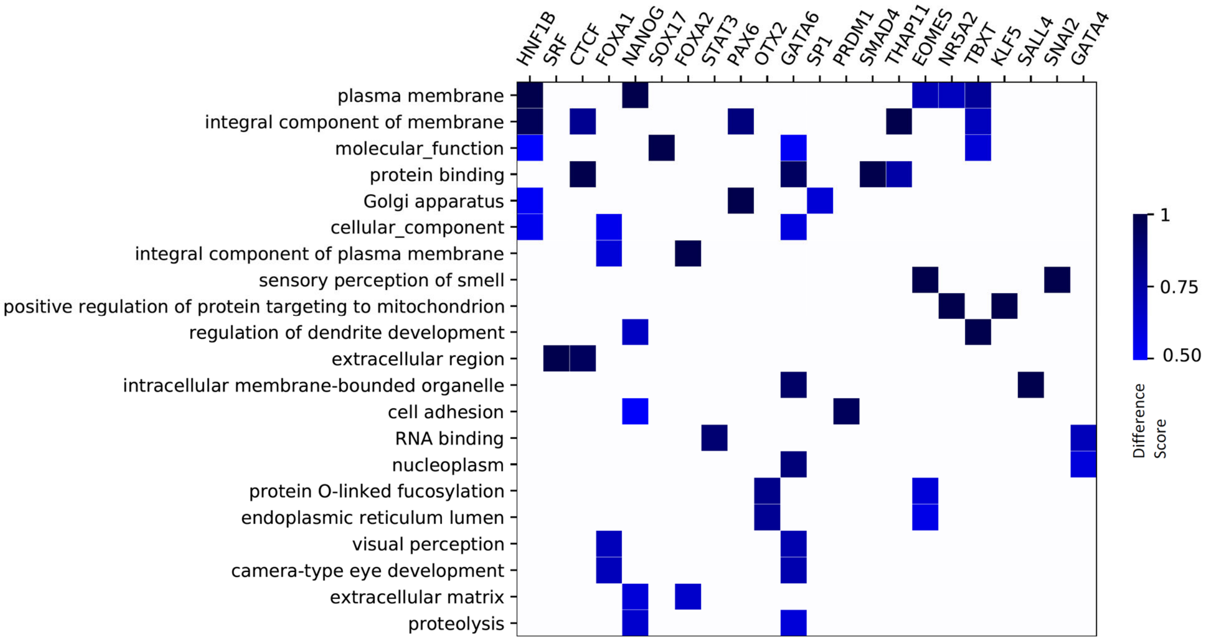


Figure 4.13 Heatmap of GO analysis using network-guided (STRING PPI) GSC pipeline of KnowEnG platform. The columns depict TFs (and their target genes) identified using InPheRNo-ChIP; the rows depict GO terms that have been implicated for at least two TFs (and their targets). The gradient colour from black (high difference score) to blue (low) depicts the ‘difference score’ above 0.5 and indicates that the GO term is associated with the input gene sets; the white colour indicates that there is no association between the GO term and the input gene sets.

Last but not least, we used the same set of TFs (and their targets) as a query gene set and performed network-guided pathway enrichment analysis (using Reactome pathways [232]) to determine the functional characterization of the identified TRN that were associated with hESC-derived endoderm. As shown in Figure 4.14, the functional role of several TFs (e.g., NANOG and STAT3) and their target genes are related to WNT signalling pathway activation, and the WNT signalling pathway has been known in inducing the differentiation potential of definitive endoderm from hESCs [233, 234]. Additionally, our results suggest that HNF1B, CTCF, and OTX2, along with their targets, may be key components of the neuronal system.

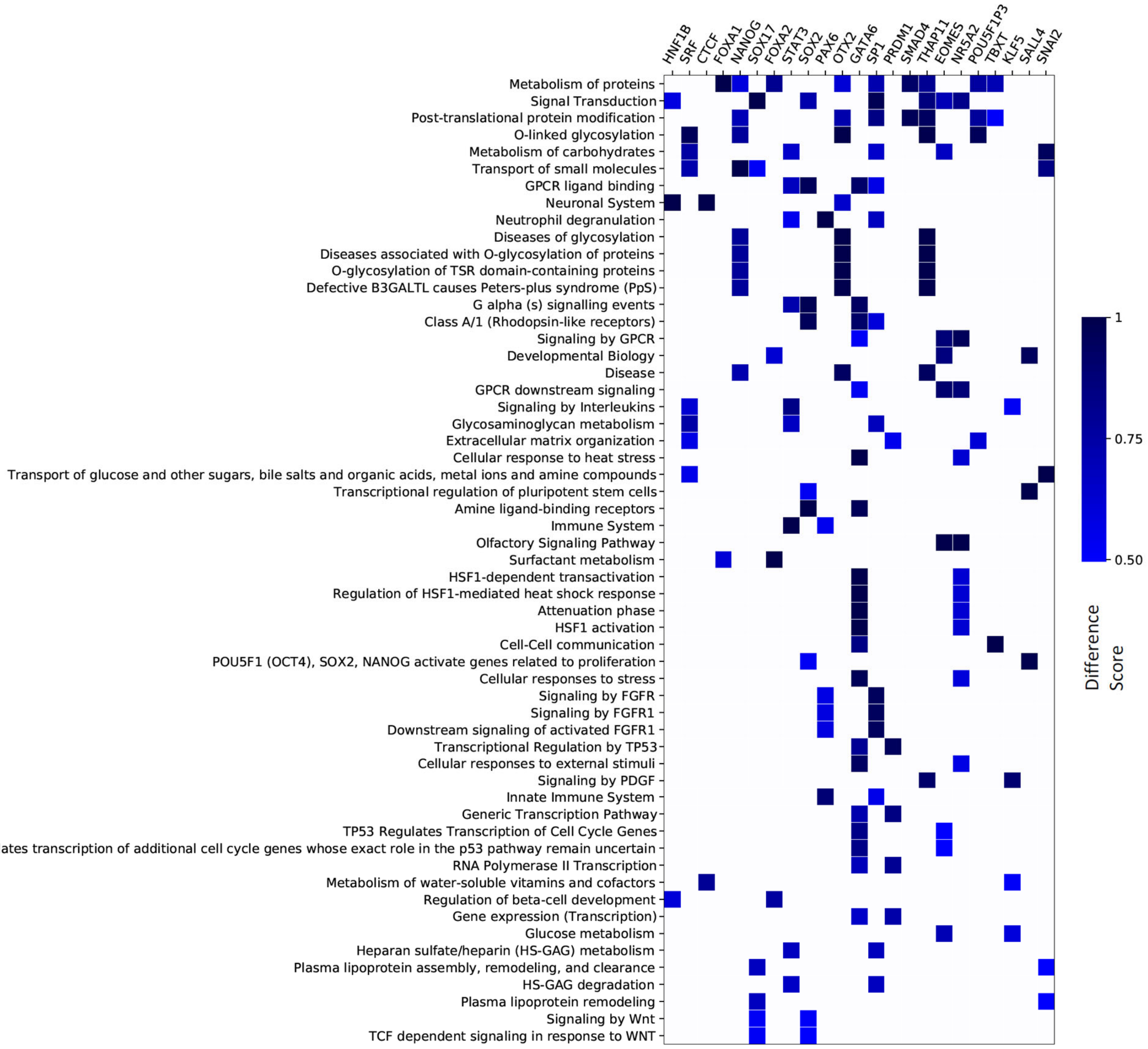


Figure 4.14 Heatmap of pathway enrichment analysis using STRING PPI-guided GSC pipeline of KnowEnG platform. The columns depict TFs (and their target genes) identified using InPheRNo-ChIP; the rows depict pathways that have been implicated for at least two TFs (and their targets). The gradient color from black (high difference score) to blue (low) depicts the 'difference score' above 0.5 and indicates that a pathway is associated with the input gene sets; the white color indicates that a pathway is not associated with the input gene sets.

4.5 Discussion

As the cost of high-throughput technologies decreases, more astounding findings are becoming available, and the possibility of modelling biological processes from different perspectives (instead of purely relying on gene expression profiles) is becoming a reality.

In this chapter, we presented InPheRNo-ChIP, a computational approach that enables the identification of phenotype-relevant TRNs. InPheRNo-ChIP generalizes InPheRNo [7] to incorporate information from both gene expression and ChIP-seq data under the assumption that the sample size is no smaller than the feature size in gene expression data. The method is applicable to any differentiation processes in human embryogenesis, such as the differentiation from hESCs to three germ layers and the differentiation of germ layers to their sub-lineages, for which gene expression data and/or ChIP-seq data are available. Additionally, it is important to note that both InPheRNo-ChIP and InPheRNo are scalable, meaning that both methods can efficiently include many genes and TFs. Performance assessments based on the generated synthetic data verified the hypothesis of how data integration can improve predictive accuracy.

Next, we applied InPheRNo-ChIP to publicly available RNA-seq and ChIP-seq data for investigating the mechanism underlying the *in vitro* differentiation of hESCs to endoderm lineages. Notably, the transcription factors HNF1B, CTCF, SRF and SOX17 identified by InPheRNo-ChIP are vital elements in the mechanisms because they actively control the expression of downstream targets to govern hESCs differentiation to endoderm. We plan to validate some of the identified TFs in wet-lab experiments.

Although we have addressed important challenges for integrating multi-omics data to reconstruct phenotype-relevant TRN, the application of InPheRNo-ChIP in the study of human embryogenesis is not free of limitations. For instance, the ChIP-seq dataset was highly biased toward a small fraction of well-studied TFs known to be relevant to either endoderm or hESCs, and therefore no novel TFs could be extracted from the inferred TRN. In order to address this issue, we plan to apply the algorithm to a more extensive database (e.g. GTRD database [71]) that contains more TFs, which can hopefully unlock InPheRNo-ChIP to its full potential. Finally, while we focus our attention on introducing more complex data into model training, future extensions could attempt a more optimized way of implementing the PGM that reduces the overall computational complexity of InPheRNo-ChIP.

5 Conclusion

5.1 Summary

In Chapter 1, we started with the motivation of the research and detailed organization. In Chapter 2, we introduced basic concepts in molecular biology, talked about the importance of reconstructing GRN/TRN, and performed a literature review about this topic. In Chapter 3, we focused on constructing a computational pipeline that can reconstruct a TRN specific to the host response to SARS-CoV-2 infection and combine the inferred TRN with other tools for the identification of potential therapeutic targets. In Chapter 4, we proposed a computational method, InPheRNo-ChIP, to infer TRNs by integrating RNA-seq, ChIP-seq and phenotype information. Instead of modelling the raw data from ChIP-seq and RNA-seq, we used summary statistics (e.g., p values) to reduce computational complexity. P values can come from different types of tests depending on the different types of information, which extends its applicability to a wide range of data sources. In addition, we conducted several experiments on synthetic data to analyze the performance of InPheRNo-ChIP under different setting scenarios. Using simulated data, an assessment of InPheRNo-ChIP verified our hypothesis that successive integration of various data types (i.e., RNA-seq, ChIP-seq and phenotypic information) could improve the TRN reconstruction accuracy. Lastly, InPheRNo-ChIP was applied to gene expression profiles and ChIP-seq samples corresponding to hESCs and endoderm lineages, where we were able to identify well-studied TFs and predicted target genes that are involved in the earliest stages of endoderm development. We believe that methods like InPheRNo-ChIP would help provide mechanistic insights into the dynamic process in stem cell lineage specification.

5.2 Future Work

From a methodology perspective, the current algorithm can be extended to incorporate other data sources such as histone ChIP-seq and ATAC-seq data [235-237]. Furthermore, the high computational costs during the inference can be circumvented by replacing MCMC sampling with variational inference [205, 238], or implementing the model on the GPU. From an application perspective, our study of human embryonic development (discussed in chapter 4.4) focused only on the shared TFs between gene expression and ChIP-seq data due to the lack of matching data sources, limiting the power of InPheRNo-ChIP in inferring relationships for unseen TFs in the human genome. One way to solve this issue is to include other unmatched TFs from ChIP-seq datasets from the GTRD database [71]. Another alternative is to use *in vivo* patient data from the Japan NBDC human database [239], under accession numbers hum0086.v3 and hum0112.v1, to construct more robust lineage-relevant TRNs.

Appendices

A.1. Pseudo Code

Algorithm: Synthetic Data Generator

Inputs: A set of parameters: P_{tg}, P_{gp}, P_{tgp}, SNR_{dB}, P_{er}, $N_{samples}$, N_{genes}, N_{TFs}, ChIP_real_distribution

Outputs: A dictionary contains RNA-seq data, ChIP-seq peaks, parameters, and ground truth

```
Initialization:
Initialize the number of total samples $N_{samples} = 2 * N_{experiments}$; set eps to 1e-100.
Initialize a dictionary expr_gene to store gene expression values; each key-value pair has shape $(1, N_{samples})$
Generate a normalized vector pheno_vec of size $N_{samples}$, where -0.1 means control sample, +0.1 indicates treated samples.
Generate a binary matrix TF_gene_ind with probability $P_{tg}$, shape: $(N_{genes}, N_{TFs})$
Generate a binary vector gene_pheno_ind with probability $P_{gp}$ with a length of $N_{genes}$

RNA-seq data generation:
Initialize a dictionary expr_TF with $N_{TFs}$ as keys to store expression values for each TF
FOR $j = 0, \ldots, N_{genes} - 1$ DO
  Expr_gene[j] = np.random.normal(0, eps, $N_{samples}$)
  IF ∃j: (j, val) ∈ gene_pheno_ind:                /* a valid gene-phenotype association*/
    expr_gene[j] += pheno_vec
  FOR TF $i = 0, \ldots, N_{TFs} - 1$ DO
    IF TF_gene_ind[i, j] == 1:                     /* a valid TF-gene binding*/
      expr_TF[i] = Draw $N_{samples}$ random samples from a normal distribution with $\mathbf{mean} = 0, \mathbf{std} = 1$
      expr_gene[j] += coefficients * expr_TF[i]
  END FOR
Compute the variance of expr_gene[j], stored in $Signal_{dB}$.
Compute $Noise_{dB} = Signal_{dB} - SNR_{dB}$, and convert to linear form $Noise_{linear}$.
noise_vec = draw $N_{samples}$ random samples from a normal distribution with $\mathbf{mean} = 0, \mathbf{std} = (Noise_{linear})^{0.5}$
Update Expr_gene[j] += noise_vec
END FOR
Output 1: synthetic gene expression data.
```

ChIP-seq peaks generation (mismatch case):

```
Initialize a dictionary tf_gene_peak to store simulated ChIP-seq peaks.
Construct an error rate mask using Bernoulli distribution of shape ($N_{genes}, N_{TFs}$), with probability $P_{er}$
ChIP_TF_gene_ind = Take bitwise XOR of (1) TF_gene_ind for gene expression data, and (2) error_rate mask.
FOR $j = 0, \dots, N_{genes} - 1$ DO
  Initialize a tmp variable to store TF-peaks information for one gene.
  FOR TF $i = 0, \dots, N_{TFs} - 1$ DO
    Draw random variable $N_{peaks}$ from ChIP_real_distribution
    IF ChIP_TF_gene_ind[i, j] == 1:                /*There is a valid TF-gene binding*/
      Generate $N_{peaks}$ random samples from a Beta distribution with two shape parameters: $P_{tgp}$ and 1
      Store in tmp[i]
  END FOR
tf_gene_peak [j] = tmp
END FOR
Output 2: synthetic ChIP-seq data.
```

A.2. Sample Information

Table A.2 Sample information for three gene expression datasets in Chapter 4. Each sample is associated with GSE number, lineage information (main factor), number of replicates, and other confounding factors. Among these samples, only lineages labelled as "EN|hESC|EIM|ESC" are used in the analysis. **NANs mean not applicable.*

Accession number	Lineage	Rep	Other Factor	Day Info
GSE143371	DE	2	CTRL	NANs
GSE143371	DE	2	EmptyVec	NANs
GSE143371	hESC	3	CTRL	NANs
GSE143371	hESC	3	EmptyVec	NANs
GSE143371	ME	3	CTRL	NANs
GSE143371	ME	3	EmptyVec	NANs
GSE143371	NE	3	CTRL	NANs
GSE143371	NE	3	EmptyVec	NANs
GSE160981	EIM	3	DMSO	NANs
GSE160981	hESC	3	DMSO	NANs
GSE160981	MIM	3	DMSO	NANs
GSE160981	NIM	3	DMSO	NANs
GSE164361	DE	2	Day1	Day1
GSE164361	DE	2	Day2	Day2
GSE164361	DE	2	Day3	Day3
GSE164361	DE	2	Day4	Day4
GSE164361	DE	2	Day5	Day5
GSE164361	ESC	2	Day0	Day0
GSE164361	NPC	2	Day1	Day1
GSE164361	NPC	2	Day2	Day2
GSE164361	NPC	2	Day3	Day3
GSE164361	NPC	2	Day4	Day4
GSE164361	NPC	2	Day5	Day5
GSE164361	NPC	2	Day6	Day6
GSE164361	NPC	2	Day7	Day7
GSE164361	NPC	2	Day8	Day8

A.3. Batch Effect Adjustments

Some batch effects are expected when we integrate pre-computed raw read counts for RNA-seq datasets generated by three different labs. Principal component analysis (PCA) [197] was performed on the unadjusted counts to show the potential batch effects in the aggregated data. PCA is an exploratory data analysis approach that provides visualization on samples as groups based on their overall pattern of gene expression values [240]. The figure below shows PCA results before and after batch correction, where samples are coloured according to their original accession number and lineage information. The PCA plot in the left upper corner suggests a distinct grouping influenced by which dataset it came.

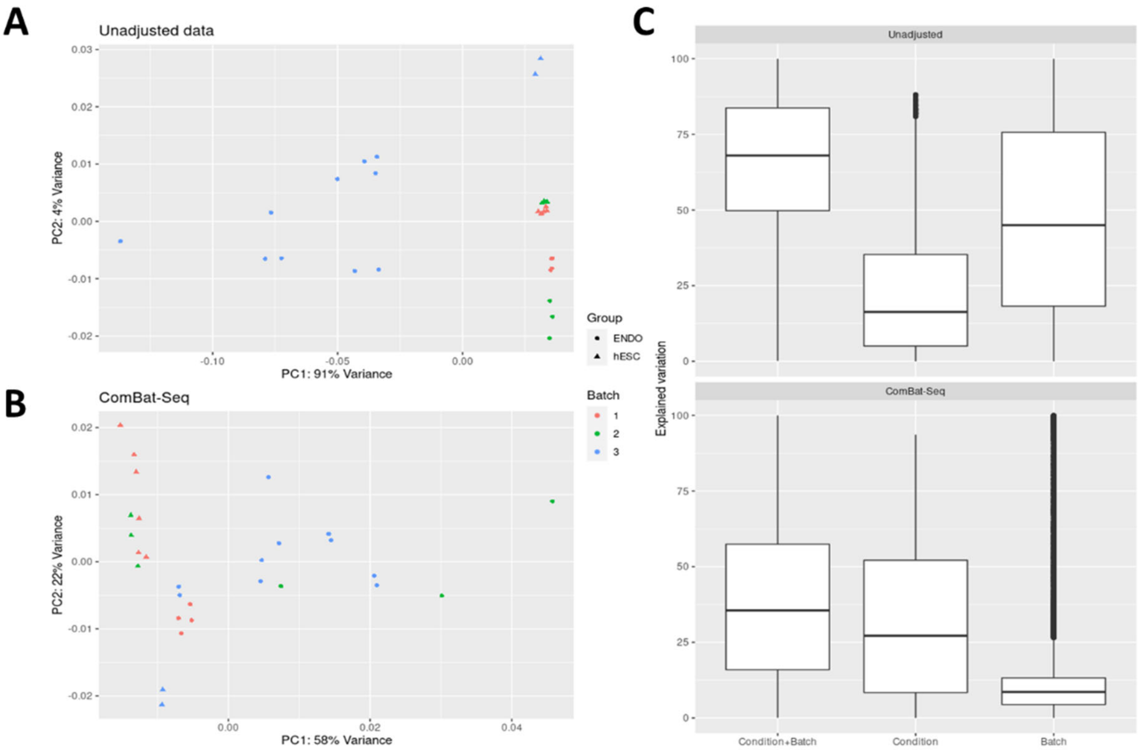


Figure A.3 PCA and Boxplot before and after applying ComBat-seq. (LHS) Note that prior to correction, the unadjusted samples group into three fairly distinct clusters. It means that the overall expression signatures of these samples reflect both the biological condition (e.g., hESCs versus Endoderm), and sources (e.g., different GEO accession numbers). After correcting for the batch effect of different libraries, we observed that hESCs samples from three GSE datasets are clustered together (as if there was no batch effect in the first place). (RHS) Compared to variation explained by batch in unadjusted data, variation explained by the batch is significantly reduced after adjustment.

A.4. Histogram of the Average Number of Peaks (ChIP-seq data)

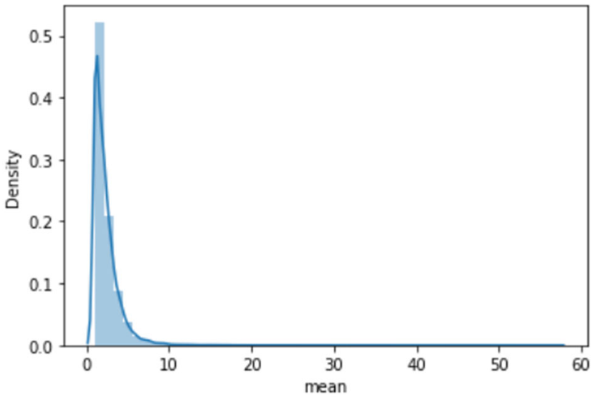
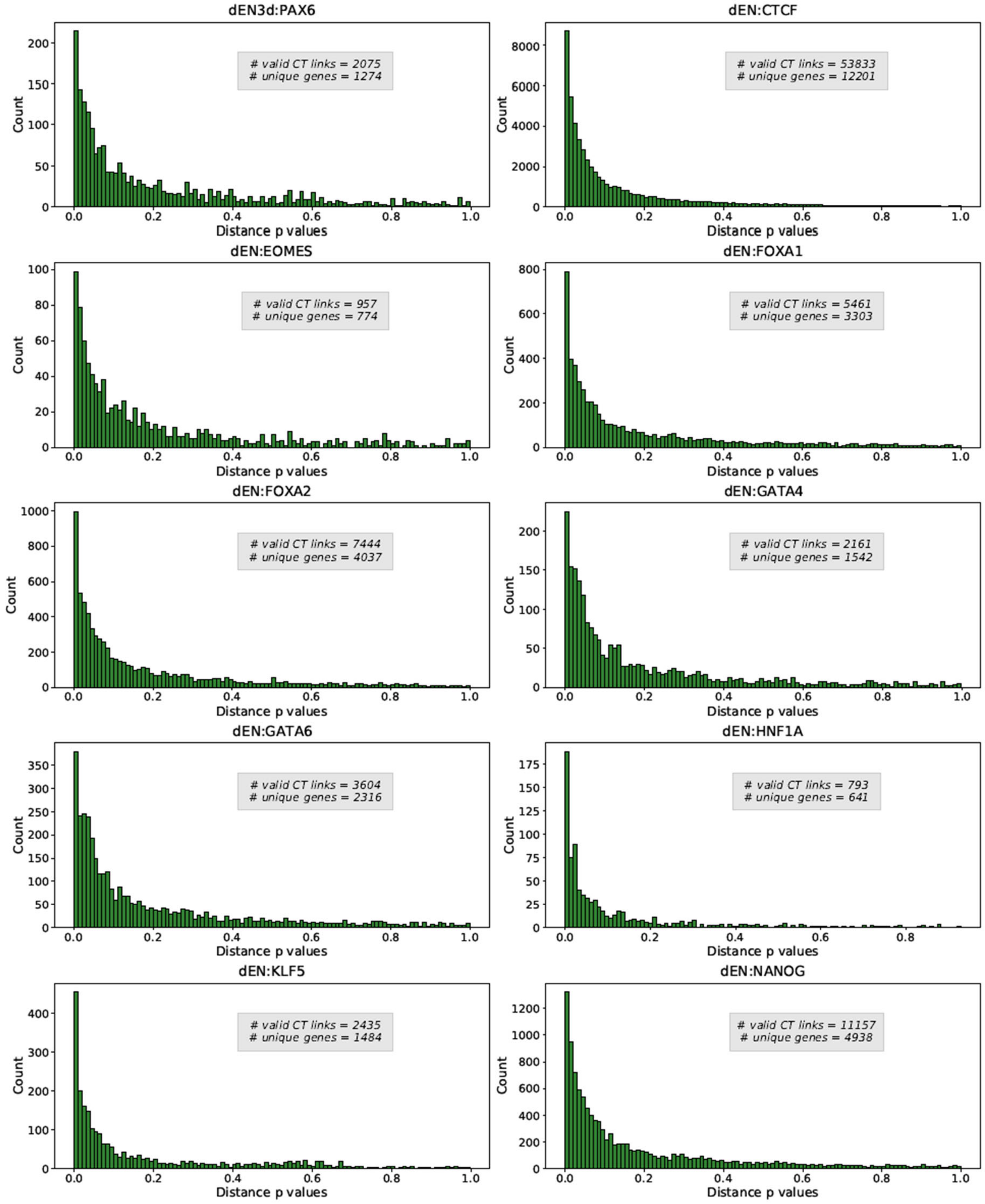
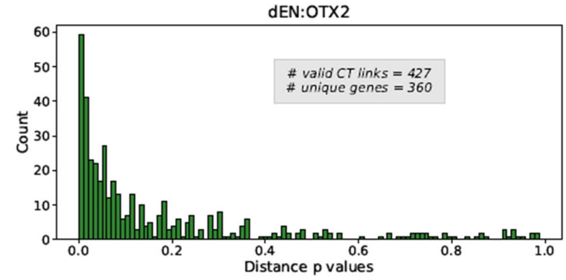
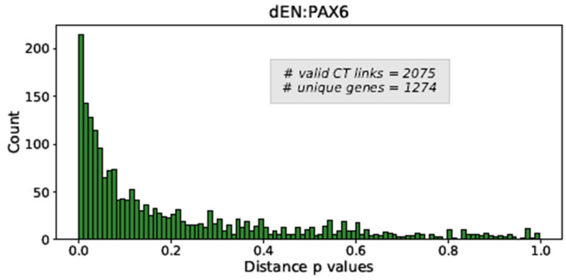


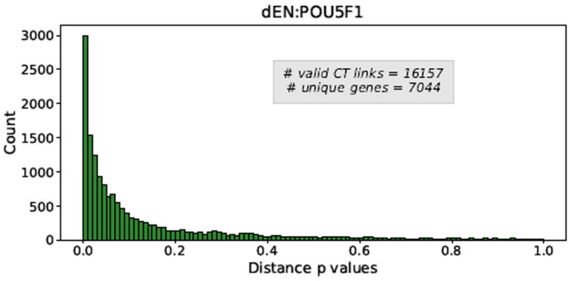
Figure A.4 Distribution of the average number of peaks per gene across all 25 TFs obtained by applying T-Gene and filtered with a condition ‘Closest_TSS=True’ to real-world ChIP-seq data (i.e., GSE61475).

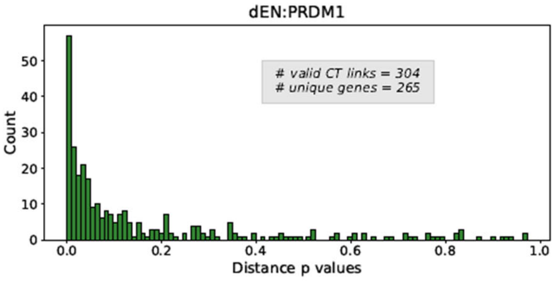
A.5. Histogram of filtered T-Gene's outputs (ChIP-seq)

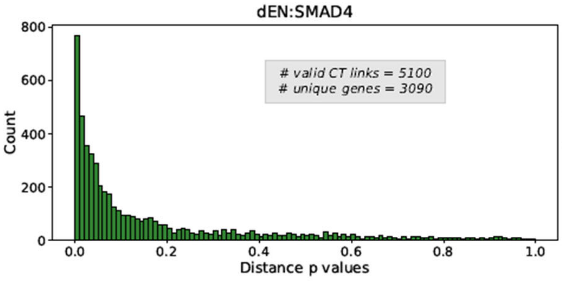


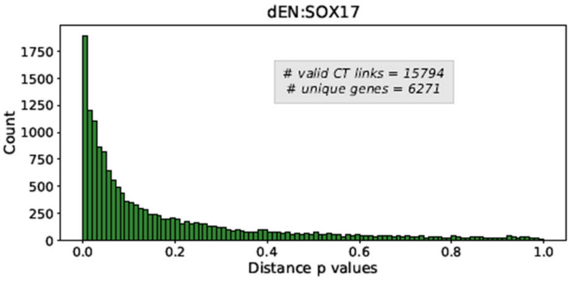

dEN:OTX2
# valid CT links = 427
# unique genes = 360
Count
0
10
20
30
40
50
60
0.0
0.2
0.4
0.6
0.8
1.0
Distance p values

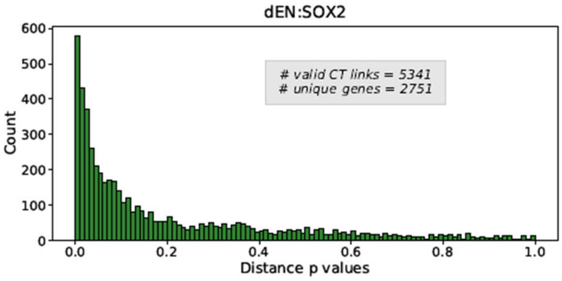

dEN:PAX6
# valid CT links = 2075
# unique genes = 1274
Count
0
50
100
150
200
0.0
0.2
0.4
0.6
0.8
1.0
Distance p values

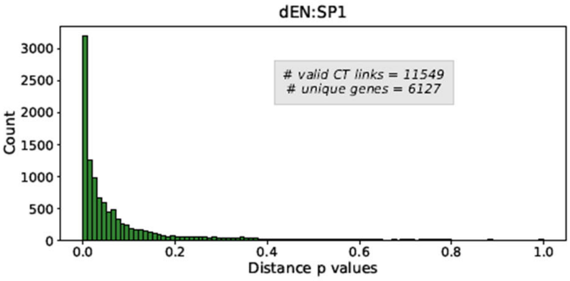

dEN:POU5F1
# valid CT links = 16157
# unique genes = 7044
Count
0
500
1000
1500
2000
2500
3000
0.0
0.2
0.4
0.6
0.8
1.0
Distance p values

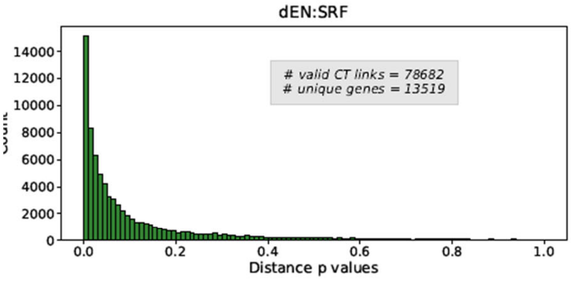

dEN:PRDM1
# valid CT links = 304
# unique genes = 265
Count
0
10
20
30
40
50
0.0
0.2
0.4
0.6
0.8
1.0
Distance p values

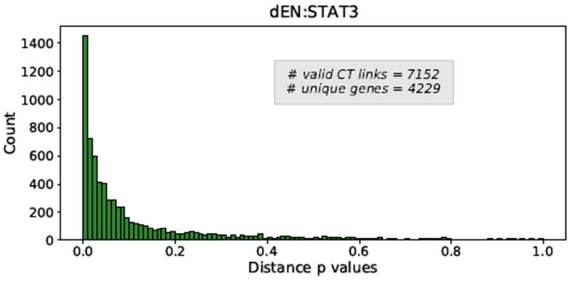

dEN:SMAD4
# valid CT links = 5100
# unique genes = 3090
Count
0
200
400
600
800
0.0
0.2
0.4
0.6
0.8
1.0
Distance p values

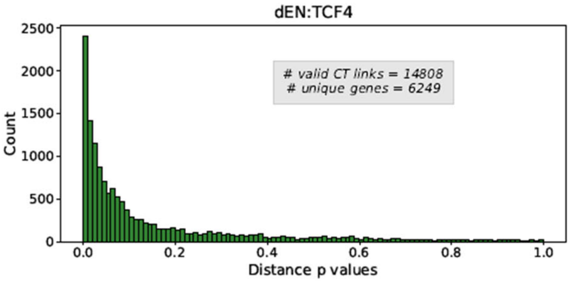

dEN:SOX17
# valid CT links = 15794
# unique genes = 6271
Count
0
250
500
750
1000
1250
1500
1750
0.0
0.2
0.4
0.6
0.8
1.0
Distance p values

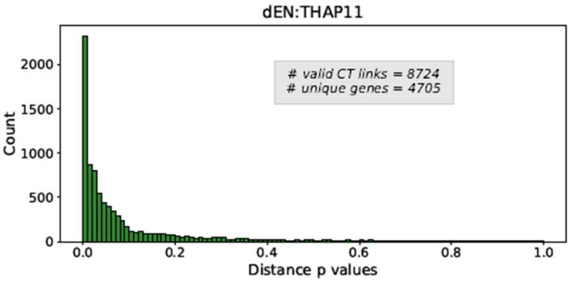

dEN:SOX2
# valid CT links = 5341
# unique genes = 2751
Count
0
100
200
300
400
500
600
0.0
0.2
0.4
0.6
0.8
1.0
Distance p values

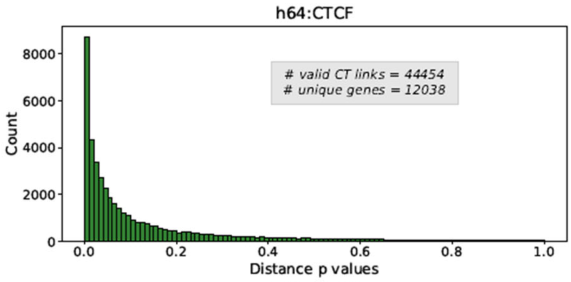

dEN:SP1
# valid CT links = 11549
# unique genes = 6127
Count
0
500
1000
1500
2000
2500
3000
0.0
0.2
0.4
0.6
0.8
1.0
Distance p values

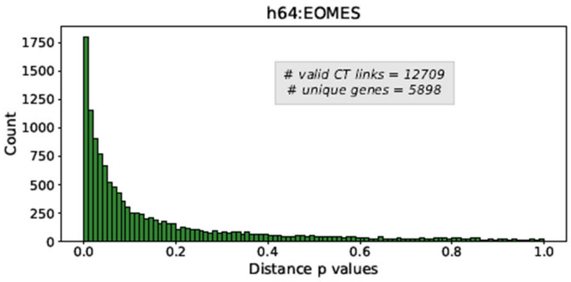

dEN:SRF
# valid CT links = 78682
# unique genes = 13519
0
2000
4000
6000
8000
10000
12000
14000
0.0
0.2
0.4
0.6
0.8
1.0
Distance p values

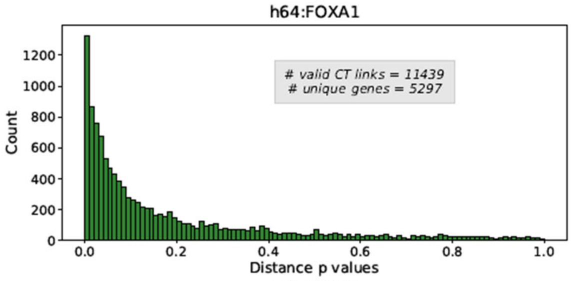

dEN:STAT3
# valid CT links = 7152
# unique genes = 4229
Count
0
200
400
600
800
1000
1200
1400
0.0
0.2
0.4
0.6
0.8
1.0
Distance p values

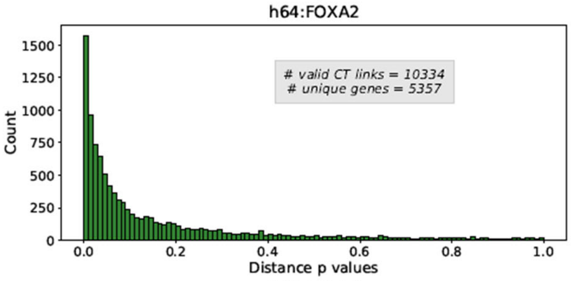

dEN:TCF4
# valid CT links = 14808
# unique genes = 6249
Count
0
500
1000
1500
2000
2500
0.0
0.2
0.4
0.6
0.8
1.0
Distance p values

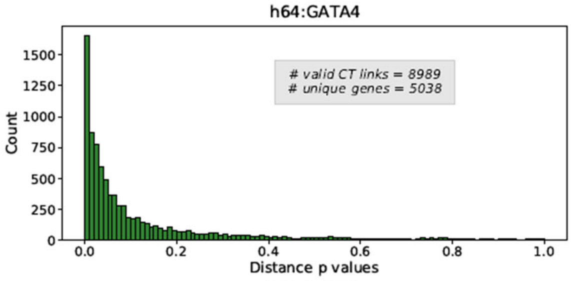

dEN:THAP11
# valid CT links = 8724
# unique genes = 4705
Count
0
500
1000
1500
2000
0.0
0.2
0.4
0.6
0.8
1.0
Distance p values

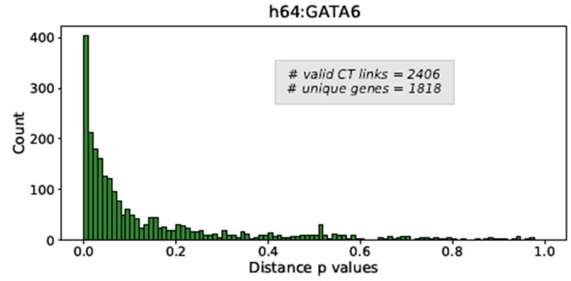

h64:CTCF
# valid CT links = 44454
# unique genes = 12038
Count
0
2000
4000
6000
8000
0.0
0.2
0.4
0.6
0.8
1.0
Distance p values

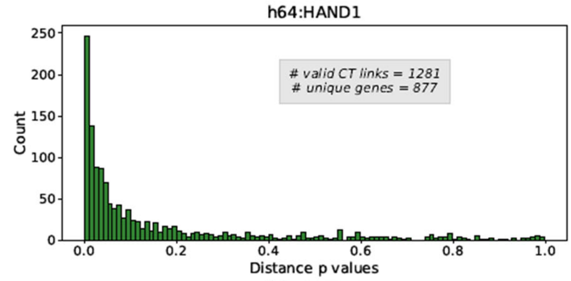

h64:EOMES
# valid CT links = 12709
# unique genes = 5898
Count
0
250
500
750
1000
1250
1500
1750
0.0
0.2
0.4
0.6
0.8
1.0
Distance p values

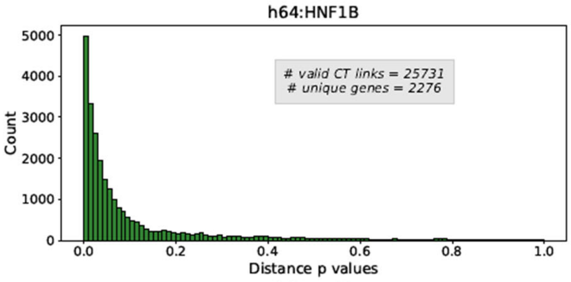

h64:FOXA1
# valid CT links = 11439
# unique genes = 5297
Count
0
200
400
600
800
1000
1200
0.0
0.2
0.4
0.6
0.8
1.0
Distance p values

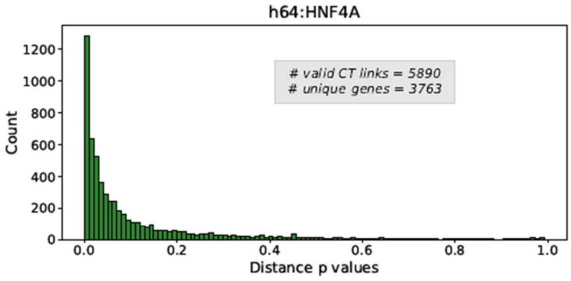

h64:FOXA2
# valid CT links = 10334
# unique genes = 5357
Count
0
250
500
750
1000
1250
1500
0.0
0.2
0.4
0.6
0.8
1.0
Distance p values

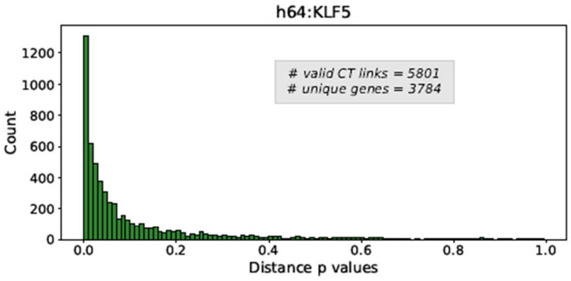

h64:GATA4
# valid CT links = 8989
# unique genes = 5038
Count
0
250
500
750
1000
1250
1500
0.0
0.2
0.4
0.6
0.8
1.0
Distance p values

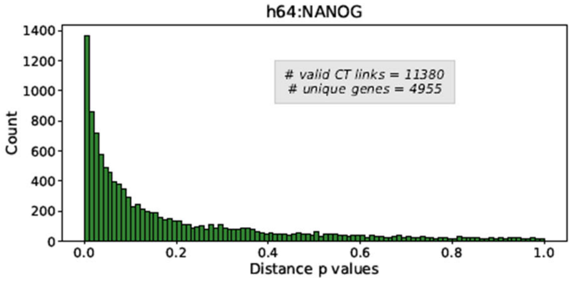

h64:GATA6
# valid CT links = 2406
# unique genes = 1818
Count
0
100
200
300
400
0.0
0.2
0.4
0.6
0.8
1.0
Distance p values

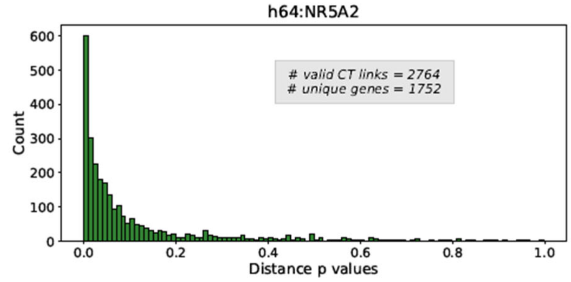

h64:HAND1
# valid CT links = 1281
# unique genes = 877
Count
0
50
100
150
200
250
0.0
0.2
0.4
0.6
0.8
1.0
Distance p values

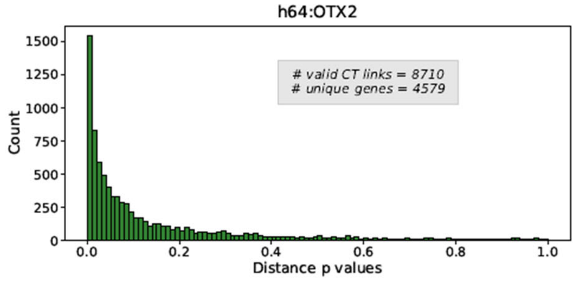

h64:HNF1B
# valid CT links = 25731
# unique genes = 2276
Count
0
1000
2000
3000
4000
5000
0.0
0.2
0.4
0.6
0.8
1.0
Distance p values

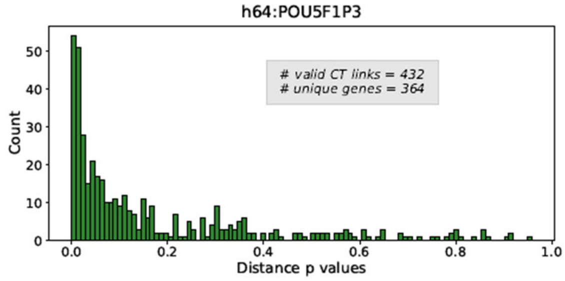

h64:HNF4A
# valid CT links = 5890
# unique genes = 3763
Count
0
200
400
600
800
1000
1200
0.0
0.2
0.4
0.6
0.8
1.0
Distance p values

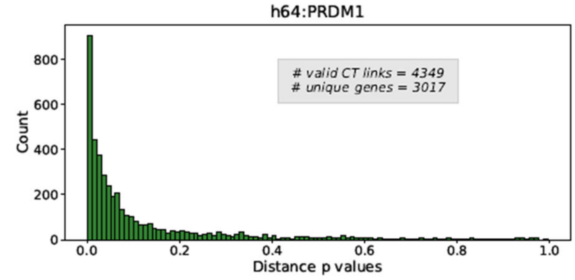

h64:KLF5
# valid CT links = 5801
# unique genes = 3784
Count
0
200
400
600
800
1000
1200
0.0
0.2
0.4
0.6
0.8
1.0
Distance p values

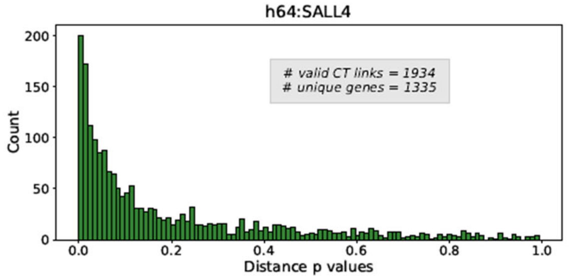

h64:NANOG
# valid CT links = 11380
# unique genes = 4955
Count
0
200
400
600
800
1000
1200
1400
0.0
0.2
0.4
0.6
0.8
1.0
Distance p values

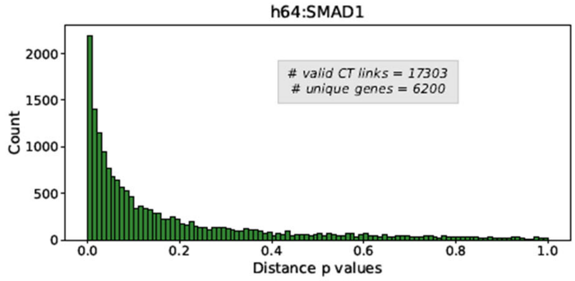

h64:NR5A2
# valid CT links = 2764
# unique genes = 1752
Count
0
100
200
300
400
500
600
0.0
0.2
0.4
0.6
0.8
1.0
Distance p values

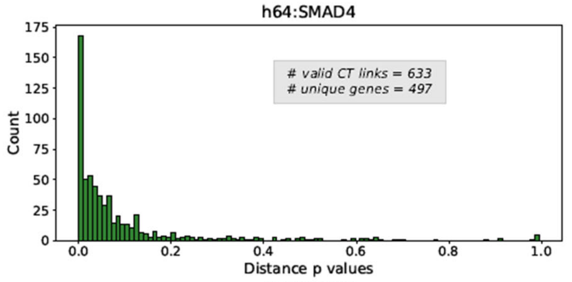

h64:OTX2
# valid CT links = 8710
# unique genes = 4579
Count
0
250
500
750
1000
1250
1500
Distance p values
0.0
0.2
0.4
0.6
0.8
1.0

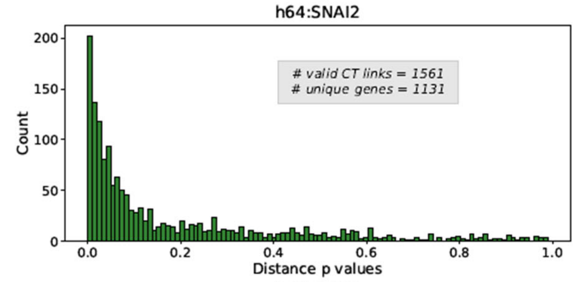

h64:POU5F1P3
# valid CT links = 432
# unique genes = 364
Count
0
10
20
30
40
50
Distance p values
0.0
0.2
0.4
0.6
0.8
1.0

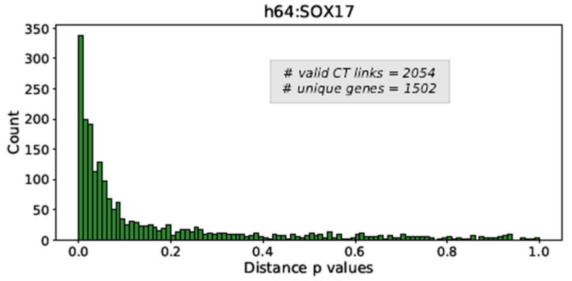

h64:PRDM1
# valid CT links = 4349
# unique genes = 3017
Count
0
200
400
600
800
Distance p values
0.0
0.2
0.4
0.6
0.8
1.0

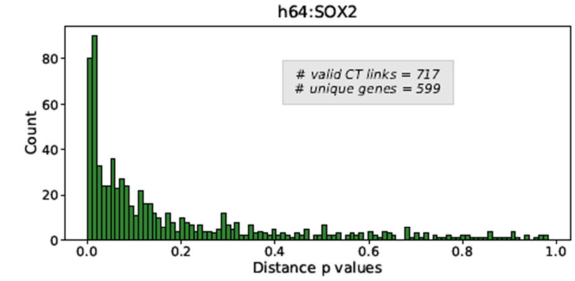

h64:SALL4
# valid CT links = 1934
# unique genes = 1335
Count
0
50
100
150
200
Distance p values
0.0
0.2
0.4
0.6
0.8
1.0

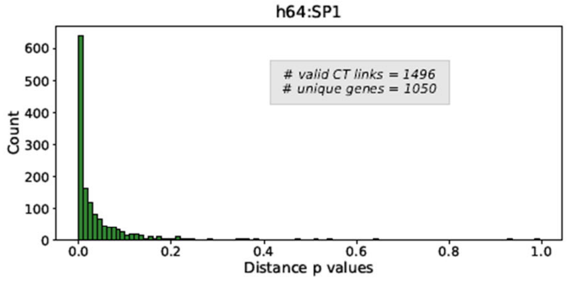

h64:SMAD1
# valid CT links = 17303
# unique genes = 6200
Count
0
500
1000
1500
2000
Distance p values
0.0
0.2
0.4
0.6
0.8
1.0

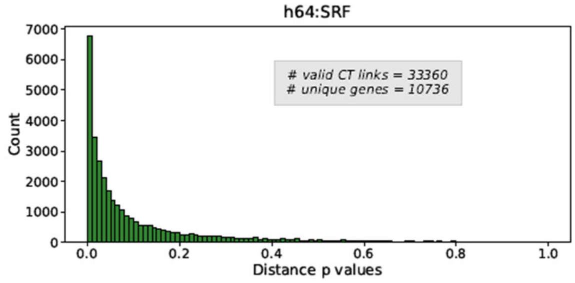

h64:SMAD4
# valid CT links = 633
# unique genes = 497
Count
0
25
50
75
100
125
150
175
Distance p values
0.0
0.2
0.4
0.6
0.8
1.0

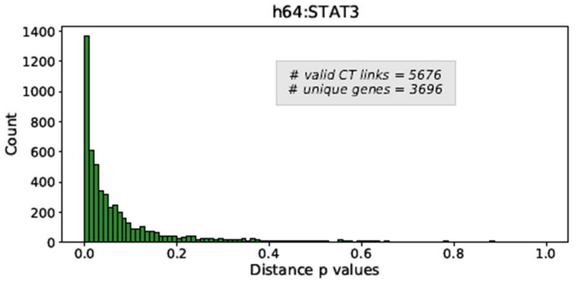

h64:SNAI2
# valid CT links = 1561
# unique genes = 1131
Count
0
50
100
150
200
Distance p values
0.0
0.2
0.4
0.6
0.8
1.0


h64:SOX17
# valid CT links = 2054
# unique genes = 1502
Count
0
50
100
150
200
250
300
350
Distance p values
0.0
0.2
0.4
0.6
0.8
1.0


h64:SOX2
# valid CT links = 717
# unique genes = 599
Count
0
20
40
60
80
Distance p values
0.0
0.2
0.4
0.6
0.8
1.0


h64:SP1
# valid CT links = 1496
# unique genes = 1050
Count
0
100
200
300
400
500
600
Distance p values
0.0
0.2
0.4
0.6
0.8
1.0


h64:SRF
# valid CT links = 33360
# unique genes = 10736
Count
0
1000
2000
3000
4000
5000
6000
7000
Distance p values
0.0
0.2
0.4
0.6
0.8
1.0


h64:STAT3
# valid CT links = 5676
# unique genes = 3696
Count
0
200
400
600
800
1000
1200
1400
Distance p values
0.0
0.2
0.4
0.6
0.8
1.0

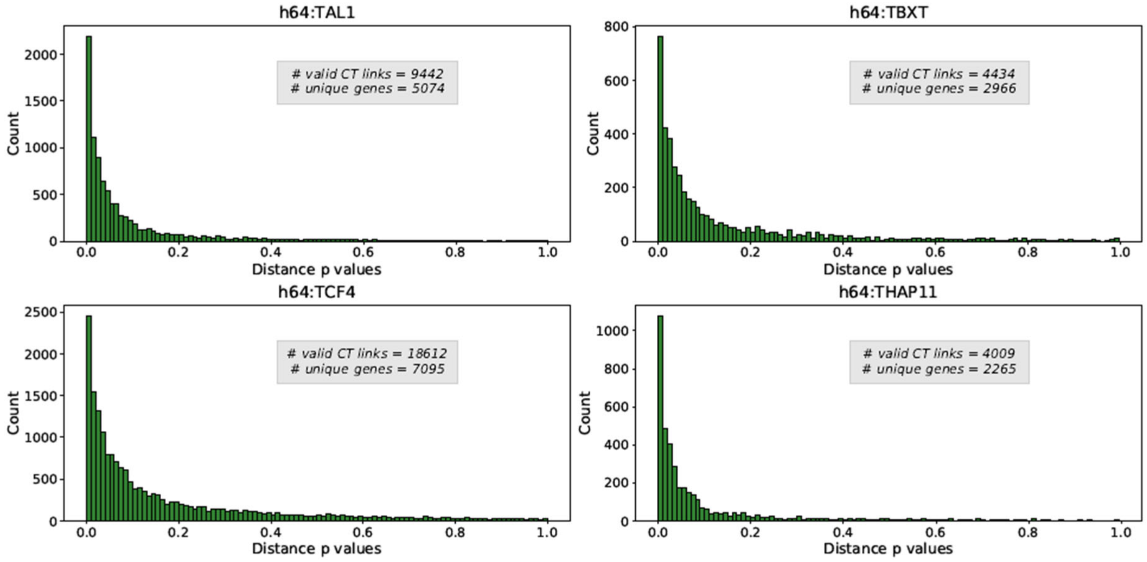


Figure A.5 Histograms of Distance p values of associations between TFs and their putative targets obtained by applying T-Gene [204] to real ChIP-seq data (i.e., GSE61475). Each figure corresponds to one TF on a specific lineage in a form of “lineage: TF name”, where we use a phenotypic label “h64” for samples from hESCs, and another label “dEN” for samples from hESC-derived endoderm lineages.

Bibliography

[1] F. M. Delgado and F. Gómez-Vela, "Computational Methods for Gene Regulatory Networks Reconstruction and Analysis: A Review," *Artificial intelligence in medicine,* vol. 95, pp. 133-145, 2019.

[2] E. Sauta, A. Demartini, F. Vitali, A. Riva, and R. Bellazzi, "A Bayesian Data Fusion Based Approach for Learning Genome-Wide Transcriptional Regulatory Networks," *BMC bioinformatics,* vol. 21, pp. 1-28, 2020.

[3] S. Aibar *et al.*, "Scenic: Single-Cell Regulatory Network Inference and Clustering," *Nature methods,* vol. 14, no. 11, pp. 1083-1086, 2017.

[4] S. A. Ament *et al.*, "Transcriptional Regulatory Networks Underlying Gene Expression Changes in Huntington's Disease," *Molecular systems biology,* vol. 14, no. 3, p. e7435, 2018.

[5] H. Han *et al.*, "Trrust V2: An Expanded Reference Database of Human and Mouse Transcriptional Regulatory Interactions," *Nucleic acids research,* vol. 46, no. D1, pp. D380-D386, 2018.

[6] C.-N. Chow *et al.*, "Plantpan3. 0: A New and Updated Resource for Reconstructing Transcriptional Regulatory Networks from Chip-Seq Experiments in Plants," *Nucleic acids research,* vol. 47, no. D1, pp. D1155-D1163, 2019.

[7] A. Emad and S. Sinha, "Inference of Phenotype-Relevant Transcriptional Regulatory Networks Elucidates Cancer Type-Specific Regulatory Mechanisms in a Pan-Cancer Study," *NPJ systems biology and applications,* vol. 7, no. 1, pp. 1-14, 2021.

[8] F. Crick, "Central Dogma of Molecular Biology," *Nature,* vol. 227, no. 5258, pp. 561-563, 1970.

[9] J. Ding *et al.*, "A Network-Informed Analysis of Sars-Cov-2 and Hemophagocytic Lymphohistiocytosis Genes’ Interactions Points to Neutrophil Extracellular Traps as Mediators of Thrombosis in Covid-19," *PLoS Computational Biology,* vol. 17, no. 3, p. e1008810, 2021.

[10] C. Blatti III *et al.*, "Knowledge-Guided Analysis of" Omics" Data Using the Knoweng Cloud Platform," *PLoS biology,* vol. 18, no. 1, p. e3000583, 2020.

[11] C. Su, S. Rousseau, and A. Emad, "Identification of Covid-19-Relevant Transcriptional Regulatory Networks and Associated Kinases as Potential Therapeutic Targets," *bioRxiv,* 2020.

[12] H. Lodish and S. L. Zipursky, "Molecular Cell Biology," *Biochem Mol Biol Educ,* vol. 29, pp. 126-133, 2001.

[13] A. E. Gorbalenya, L. Enjuanes, J. Ziebuhr, and E. J. Snijder, "Nidovirales: Evolving the Largest Rna Virus Genome," *Virus research,* vol. 117, no. 1, pp. 17-37, 2006.

[14] D. Fischer, D. Rice, J. U. Bowie, and D. Eisenberg, "Assigning Amino Acid Sequences to 3 ‐ Dimensional Protein Folds," *The FASEB journal,* vol. 10, no. 1, pp. 126-136, 1996.

[15] P. C. Ng and S. Henikoff, "Predicting the Effects of Amino Acid Substitutions on Protein Function," *Annu. Rev. Genomics Hum. Genet.,* vol. 7, pp. 61-80, 2006.

[16] P. C. Ng and S. Henikoff, "Sift: Predicting Amino Acid Changes That Affect Protein Function," *Nucleic acids research,* vol. 31, no. 13, pp. 3812-3814, 2003.

[17] G. Guidotti, "Membrane Proteins," *Annual review of biochemistry,* vol. 41, no. 1, pp. 731-752, 1972.

[18] R. A. Capaldi and D. E. Green, "Membrane Proteins and Membrane Structure," *FEBS letters,* vol. 25, no. 2, pp. 205-209, 1972.

[19] F. S. Collins and M. K. Mansoura, "The Human Genome Project: Revealing the Shared Inheritance of All Humankind," *Cancer: Interdisciplinary International Journal of the American Cancer Society,* vol. 91, no. S1, pp. 221-225, 2001.

[20] J. E. Novak and K. Kirkegaard, "Coupling between Genome Translation and Replication in an Rna Virus," *Genes & development,* vol. 8, no. 14, pp. 1726-1737, 1994.

[21] F. S. Collins, M. Morgan, and A. Patrinos, "The Human Genome Project: Lessons from Large-Scale Biology," *Science,* vol. 300, no. 5617, pp. 286-290, 2003.

[22] E. P. Consortium, "An Integrated Encyclopedia of DNA Elements in the Human Genome," *Nature,* vol. 489, no. 7414, p. 57, 2012.

[23] M. Chee *et al.*, "Accessing Genetic Information with High-Density DNA Arrays," *Science,* vol. 274, no. 5287, pp. 610-614, 1996.

[24] T. Shafee and R. Lowe, "Eukaryotic and Prokaryotic Gene Structure," *WikiJournal of Medicine,* vol. 4, no. 1, pp. 1-5, 2017.

[25] M. Lynch, "The Origins of Eukaryotic Gene Structure," *Molecular biology and evolution,* vol. 23, no. 2, pp. 450-468, 2006.

[26] S. P. Bell and A. Dutta, "DNA Replication in Eukaryotic Cells," *Annual review of biochemistry,* vol. 71, no. 1, pp. 333-374, 2002.

[27] W. S. Dynan and R. Tjian, "Control of Eukaryotic Messenger Rna Synthesis by Sequence-Specific DNA-Binding Proteins," *Nature,* vol. 316, no. 6031, pp. 774-778, 1985.

[28] H. Jakubowski and E. Goldman, "Editing of Errors in Selection of Amino Acids for Protein Synthesis," *Microbiological reviews,* vol. 56, no. 3, pp. 412-429, 1992.

[29] M. M. Scotti and M. S. Swanson, "Rna Mis-Splicing in Disease," *Nature Reviews Genetics,* vol. 17, no. 1, pp. 19-32, 2016.

[30] M. Janda and P. Ahlquist, "Rna-Dependent Replication, Transcription, and Persistence of Brome Mosaic Virus Rna Replicons in S. Cerevisiae," *Cell,* vol. 72, no. 6, pp. 961-970, 1993.

[31] C. C. Kao, P. Singh, and D. J. Ecker, "De Novo Initiation of Viral Rna-Dependent Rna Synthesis," *Virology,* vol. 287, no. 2, pp. 251-260, 2001.

[32] H. Varmus, "Retroviruses," *Science,* vol. 240, no. 4858, pp. 1427-1435, 1988.

[33] H. P. Davis and L. R. Squire, "Protein Synthesis and Memory: A Review," *Psychological bulletin,* vol. 96, no. 3, p. 518, 1984.

[34] T. Preiss and M. W. Hentze, "Starting the Protein Synthesis Machine: Eukaryotic Translation Initiation," *Bioessays,* vol. 25, no. 12, pp. 1201-1211, 2003.

[35] N. Sonenberg, J. W. Hershey, and M. Mathews, "Translational Control of Gene Expression," 2000.

[36] J. C. Lacey Jr and D. W. Mullins Jr, "Experimental Studies Related to the Origin of the Genetic Code and the Process of Protein Synthesis-a Review," 1983.

[37] C. J. Greenwald, T. Kasuga, N. L. Glass, B. D. Shaw, D. J. Ebbole, and H. H. Wilkinson, "Temporal and Spatial Regulation of Gene Expression During Asexual Development of Neurospora Crassa," *Genetics,* vol. 186, no. 4, pp. 1217-1230, 2010.

[38] A. M. Koltunow, J. Truettner, K. H. Cox, M. Wallroth, and R. B. Goldberg, "Different Temporal and Spatial Gene Expression Patterns Occur During Anther Development," *The Plant Cell,* vol. 2, no. 12, pp. 1201-1224, 1990.

[39] D. Wang, Y. Pan, X. Zhao, L. Zhu, B. Fu, and Z. Li, "Genome-Wide Temporal-Spatial Gene Expression Profiling of Drought Responsiveness in Rice," *BMC genomics,* vol. 12, no. 1, pp. 1-15, 2011.

[40] J.-X. Zhu, Y. Sasano, I. Takahashi, I. Mizoguchi, and M. Kagayama, "Temporal and Spatial Gene Expression of Major Bone Extracellular Matrix Molecules During Embryonic Mandibular Osteogenesis in Rats," *The Histochemical Journal,* vol. 33, no. 1, pp. 25-35, 2001.

[41] J. M. van den Brand *et al.*, "Comparison of Temporal and Spatial Dynamics of Seasonal H3n2, Pandemic H1n1 and Highly Pathogenic Avian Influenza H5n1 Virus Infections in Ferrets," 2012.

[42] T. A. Brevini, F. Cillo, S. Antonini, V. Tosetti, and F. Gandolfi, "Temporal and Spatial Control of Gene Expression in Early Embryos of Farm Animals," *Reproduction, Fertility and Development,* vol. 19, no. 1, pp. 35-42, 2006.

[43] A. Oshlack, M. D. Robinson, and M. D. Young, "From Rna-Seq Reads to Differential Expression Results," *Genome biology,* vol. 11, no. 12, pp. 1-10, 2010.

[44] S. Anders and W. Huber, "Differential Expression Analysis for Sequence Count Data," *Nature Precedings,* pp. 1-1, 2010.

[45] S. Anders *et al.*, "Count-Based Differential Expression Analysis of Rna Sequencing Data Using R and Bioconductor," *Nature protocols,* vol. 8, no. 9, pp. 1765-1786, 2013.

[46] D. J. McCarthy, Y. Chen, and G. K. Smyth, "Differential Expression Analysis of Multifactor Rna-Seq Experiments with Respect to Biological Variation," *Nucleic acids research,* vol. 40, no. 10, pp. 4288-4297, 2012.

[47] S. S. Shen-Orr *et al.*, "Cell Type–Specific Gene Expression Differences in Complex Tissues," *Nature methods,* vol. 7, no. 4, pp. 287-289, 2010.

[48] A. T. McKenzie *et al.*, "Brain Cell Type Specific Gene Expression and Co-Expression Network Architectures," *Scientific reports,* vol. 8, no. 1, pp. 1-19, 2018.

[49] A. Kamb and M. Ramaswami, "A Simple Method for Statistical Analysis of Intensity Differences in Microarray-Derived Gene Expression Data," *BMC biotechnology,* vol. 1, no. 1, pp. 1-8, 2001.

[50] R. Jaenisch and A. Bird, "Epigenetic Regulation of Gene Expression: How the Genome Integrates Intrinsic and Environmental Signals," *Nature genetics,* vol. 33, no. 3, pp. 245-254, 2003.

[51] V. E. Velculescu, L. Zhang, B. Vogelstein, and K. W. Kinzler, "Serial Analysis of Gene Expression," *Science,* vol. 270, no. 5235, pp. 484-487, 1995.

[52] S. Audic and J.-M. Claverie, "The Significance of Digital Gene Expression Profiles," *Genome research,* vol. 7, no. 10, pp. 986-995, 1997.

[53] O. Corradin and P. C. Scacheri, "Enhancer Variants: Evaluating Functions in Common Disease," *Genome medicine,* vol. 6, no. 10, pp. 1-14, 2014.

[54] Y. Bao *et al.*, "Transcriptome Profiling Revealed Multiple Genes and Ecm-Receptor Interaction Pathways That May Be Associated with Breast Cancer," *Cellular & molecular biology letters,* vol. 24, no. 1, pp. 1-20, 2019.

[55] J. Hannemann, A. Velds, J. B. Halfwerk, B. Kreike, J. L. Peterse, and M. J. van de Vijver, "Classification of Ductal Carcinoma in Situ by Gene Expression Profiling," *Breast Cancer Research,* vol. 8, no. 5, pp. 1-20, 2006.

[56] J. M. Vaquerizas, S. K. Kummerfeld, S. A. Teichmann, and N. M. Luscombe, "A Census of Human Transcription Factors: Function, Expression and Evolution," *Nature Reviews Genetics,* vol. 10, no. 4, pp. 252-263, 2009.

[57] J. Potash, "Unraveling Transcriptional Networks," *Nature Methods,* vol. 4, no. 3, pp. 198-198, 2007.

[58] U. Stelzl *et al.*, "A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome," *Cell,* vol. 122, no. 6, pp. 957-968, 2005.

[59] Y. Wang, T. Joshi, X.-S. Zhang, D. Xu, and L. Chen, "Inferring Gene Regulatory Networks from Multiple Microarray Datasets," *Bioinformatics,* vol. 22, no. 19, pp. 2413-2420, 2006.

[60] Y. Hasin, M. Seldin, and A. Lusis, "Multi-Omics Approaches to Disease," *Genome biology,* vol. 18, no. 1, pp. 1-15, 2017.

[61] R. P. Horgan and L. C. Kenny, "'Omic'technologies: Genomics, Transcriptomics, Proteomics and Metabolomics," *The Obstetrician & Gynaecologist,* vol. 13, no. 3, pp. 189-195, 2011.

[62] M. Tyers and M. Mann, "From Genomics to Proteomics," *Nature,* vol. 422, no. 6928, pp. 193-197, 2003.

[63] S. Franklin and T. M. Vondriska, "Genomes, Proteomes, and the Central Dogma," *Circulation: Cardiovascular Genetics,* vol. 4, no. 5, pp. 576-576, 2011.

[64] J. Micallef *et al.*, "Applying Mass Spectrometry Based Proteomic Technology to Advance the Understanding of Multiple Myeloma," *Journal of hematology & oncology,* vol. 3, no. 1, pp. 1-11, 2010.

[65] O. Al-Harazi, S. Al Insaif, M. A. Al-Ajlan, N. Kaya, N. Dzimiri, and D. Colak, "Integrated Genomic and Network-Based Analyses of Complex Diseases and Human Disease Network," *Journal of Genetics and Genomics,* vol. 43, no. 6, pp. 349-367, 2016.

[66] A. Conesa *et al.*, "A Survey of Best Practices for Rna-Seq Data Analysis," *Genome biology,* vol. 17, no. 1, pp. 1-19, 2016.

[67] P. A. Jones and S. B. Baylin, "The Epigenomics of Cancer," *Cell,* vol. 128, no. 4, pp. 683-692, 2007.

[68] J. L. Spratlin, N. J. Serkova, and S. G. Eckhardt, "Clinical Applications of Metabolomics in Oncology: A Review," *Clinical cancer research,* vol. 15, no. 2, pp. 431-440, 2009.

[69] J. G. Bundy, M. P. Davey, and M. R. Viant, "Environmental Metabolomics: A Critical Review and Future Perspectives," *Metabolomics,* vol. 5, no. 1, pp. 3-21, 2009.

[70] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "The Cancer Genome Atlas (Tcga): An Immeasurable Source of Knowledge," *Contemporary oncology,* vol. 19, no. 1A, p. A68, 2015.

[71] I. Yevshin, R. Sharipov, S. Kolmykov, Y. Kondrakhin, and F. Kolpakov, "Gtrd: A Database on Gene Transcription Regulation—2019 Update," *Nucleic acids research,* vol. 47, no. D1, pp. D100-D105, 2019.

[72] S. V. Vasaikar, P. Straub, J. Wang, and B. Zhang, "Linkedomics: Analyzing Multi-Omics Data within and across 32 Cancer Types," *Nucleic acids research,* vol. 46, no. D1, pp. D956-D963, 2018.

[73] A. A. Consortium, "Aging Atlas: A Multi-Omics Database for Aging Biology," *Nucleic Acids Research,* vol. 49, no. D1, pp. D825-D830, 2021.

[74] I. Subramanian, S. Verma, S. Kumar, A. Jere, and K. Anamika, "Multi-Omics Data Integration, Interpretation, and Its Application," *Bioinformatics and biology insights,* vol. 14, p. 1177932219899051, 2020.

[75] S. Jiang and A. Mortazavi, "Integrating Chip-Seq with Other Functional Genomics Data," *Briefings in functional genomics,* vol. 17, no. 2, pp. 104-115, 2018.

[76] N. Wani and K. Raza, "Integrative Approaches to Reconstruct Regulatory Networks from Multi-Omics Data: A Review of State-of-the-Art Methods," *Computational biology and chemistry,* vol. 83, p. 107120, 2019.

[77] M. Y. Hirai *et al.*, "Integration of Transcriptomics and Metabolomics for Understanding of Global Responses to Nutritional Stresses in Arabidopsis Thaliana," *Proceedings of the National Academy of Sciences,* vol. 101, no. 27, pp. 10205-10210, 2004.

[78] N. Nagaraj *et al.*, "Deep Proteome and Transcriptome Mapping of a Human Cancer Cell Line," *Molecular systems biology,* vol. 7, no. 1, p. 548, 2011.

[79] M. Michaut *et al.*, "Integration of Genomic, Transcriptomic and Proteomic Data Identifies Two Biologically Distinct Subtypes of Invasive Lobular Breast Cancer," *Scientific reports,* vol. 6, no. 1, pp. 1-13, 2016.

[80] J. Parsons and C. Francavilla, "'Omics Approaches to Explore the Breast Cancer Landscape," *Frontiers in cell and developmental biology,* vol. 7, p. 395, 2020.

[81] L. Pirhaji *et al.*, "Revealing Disease-Associated Pathways by Network Integration of Untargeted Metabolomics," *Nature methods,* vol. 13, no. 9, pp. 770-776, 2016.

[82] L. Xie, B. Weichel, J. E. Ohm, and K. Zhang, "An Integrative Analysis of DNA Methylation and Rna-Seq Data for Human Heart, Kidney and Liver," *BMC systems biology,* vol. 5, no. 3, pp. 1-11, 2011.

[83] A. Akcakanat *et al.*, "Genomic, Transcriptomic, and Proteomic Profiling of Metastatic Breast Cancer," *Clinical Cancer Research,* vol. 27, no. 11, pp. 3243-3252, 2021.

[84] P. Bouchal *et al.*, "Combined Proteomics and Transcriptomics Identifies Carboxypeptidase B1 and Nuclear Factor Kb (Nf-Kb) Associated Proteins as Putative Biomarkers of Metastasis in Low Grade Breast Cancer," *Molecular & Cellular Proteomics,* vol. 14, no. 7, pp. 1814-1830, 2015.

[85] L. Zappia and F. J. Theis, "Over 1000 Tools Reveal Trends in the Single-Cell Rna-Seq Analysis Landscape," *Genome biology,* vol. 22, no. 1, pp. 1-18, 2021.

[86] C. Chen *et al.*, "Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods," *PloS one,* vol. 6, no. 2, p. e17238, 2011.

[87] H. Zou and T. Hastie, "Regularization and Variable Selection Via the Elastic Net," *Journal of the royal statistical society: series B (statistical methodology),* vol. 67, no. 2, pp. 301-320, 2005.

[88] R. Tibshirani, "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society: Series B (Methodological),* vol. 58, no. 1, pp. 267-288, 1996.

[89] P. E. Meyer, K. Kontos, F. Lafitte, and G. Bontempi, "Information-Theoretic Inference of Large Transcriptional Regulatory Networks," *EURASIP journal on bioinformatics and systems biology,* vol. 2007, pp. 1-9, 2007.

[90] A. A. Margolin *et al.*, "Aracne: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context," in *BMC bioinformatics*, 2006, vol. 7, no. 1: Springer, pp. 1-15.

[91] J. J. Faith *et al.*, "Large-Scale Mapping and Validation of Escherichia Coli Transcriptional Regulation from a Compendium of Expression Profiles," *PLoS biology,* vol. 5, no. 1, p. e8, 2007.

[92] G. Altay and F. Emmert-Streib, "Inferring the Conservative Causal Core of Gene Regulatory Networks," *BMC systems biology,* vol. 4, no. 1, pp. 1-13, 2010.

[93] R. Küffner, T. Petri, P. Tavakkolkhah, L. Windhager, and R. Zimmer, "Inferring Gene Regulatory Networks by Anova," *Bioinformatics,* vol. 28, no. 10, pp. 1376-1382, 2012.

[94] M. Banf and S. Y. Rhee, "Computational Inference of Gene Regulatory Networks: Approaches, Limitations and Opportunities," *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms,* vol. 1860, no. 1, pp. 41-52, 2017.

[95] A. Madar, A. Greenfield, E. Vanden-Eijnden, and R. Bonneau, "Dream3: Network Inference Using Dynamic Context Likelihood of Relatedness and the Inferelator," *PloS one,* vol. 5, no. 3, p. e9803, 2010.

[96] C. Ding and H. Peng, "Minimum Redundancy Feature Selection from Microarray Gene Expression Data," *Journal of bioinformatics and computational biology,* vol. 3, no. 02, pp. 185-205, 2005.

[97] J. Ish-Horowicz and J. Reid, "Mutual Information Estimation for Transcriptional Regulatory Network Inference," *bioRxiv,* p. 132647, 2017.

[98] J. Cao, X. Qi, and H. Zhao, "Modeling Gene Regulation Networks Using Ordinary Differential Equations," in *Next Generation Microarray Bioinformatics*: Springer, 2012, pp. 185-197.

[99] D. Mercatelli, L. Scalambra, L. Triboli, F. Ray, and F. M. Giorgi, "Gene Regulatory Network Inference Resources: A Practical Overview," *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms,* vol. 1863, no. 6, p. 194430, 2020.

[100] R. Bonneau *et al.*, "The Inferelator: An Algorithm for Learning Parsimonious Regulatory Networks from Systems-Biology Data Sets De Novo," *Genome biology,* vol. 7, no. 5, pp. 1-16, 2006.

[101] B. Yang *et al.*, "Reverse Engineering Gene Regulatory Network Based on Complex-Valued Ordinary Differential Equation Model," *BMC bioinformatics,* vol. 22, no. 3, pp. 1-19, 2021.

[102] C. Song, "A Complex-Valued Firefly Algorithm," in *International Conference on Intelligent Computing*, 2019: Springer, pp. 700-707.

[103] S. Kauffman, C. Peterson, B. Samuelsson, and C. Troein, "Random Boolean Network Models and the Yeast Transcriptional Network," *Proceedings of the National Academy of Sciences,* vol. 100, no. 25, pp. 14796-14799, 2003.

[104] R. Barbuti, R. Gori, P. Milazzo, and L. Nasti, "A Survey of Gene Regulatory Networks Modelling Methods: From Differential Equations, to Boolean and Qualitative Bioinspired Models," *Journal of Membrane Computing,* pp. 1-20, 2020.

[105] C. E. Giacomantonio and G. J. Goodhill, "A Boolean Model of the Gene Regulatory Network Underlying Mammalian Cortical Area Development," *PLoS computational biology,* vol. 6, no. 9, p. e1000936, 2010.

[106] S. Gupta, D. A. Silveira, and J. C. M. Mombach, "Towards DNA-Damage Induced Autophagy: A Boolean Model of P53-Induced Cell Fate Mechanisms," *DNA repair,* vol. 96, p. 102971, 2020.

[107] M. E. Martinez-Sanchez, L. Huerta, E. R. Alvarez-Buylla, and C. Villarreal Luján, "Role of Cytokine Combinations on Cd4+ T Cell Differentiation, Partial Polarization, and Plasticity: Continuous Network Modeling Approach," *Frontiers in physiology,* vol. 9, p. 877, 2018.

[108] I. Shmulevich, I. Gluhovsky, R. F. Hashimoto, E. R. Dougherty, and W. Zhang, "Steady‐State Analysis of Genetic Regulatory Networks Modelled by Probabilistic Boolean Networks," *Comparative and functional genomics,* vol. 4, no. 6, pp. 601-608, 2003.

[109] W. Zhao, E. Serpedin, and E. R. Dougherty, "Inferring Gene Regulatory Networks from Time Series Data Using the Minimum Description Length Principle," *Bioinformatics,* vol. 22, no. 17, pp. 2129-2135, 2006.

[110] W. J. Blake, M. Kærn, C. R. Cantor, and J. J. Collins, "Noise in Eukaryotic Gene Expression," *Nature,* vol. 422, no. 6932, pp. 633-637, 2003.

[111] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian Networks to Analyze Expression Data," *Journal of computational biology,* vol. 7, no. 3-4, pp. 601-620, 2000.

[112] K. P. Murphy, "Dynamic Bayesian Networks," *Probabilistic Graphical Models, M. Jordan,* vol. 7, p. 431, 2002.

[113] L. Bouchaala, A. Masmoudi, F. Gargouri, and A. Rebai, "Improving Algorithms for Structure Learning in Bayesian Networks Using a New Implicit Score," *Expert Systems with Applications,* vol. 37, no. 7, pp. 5470-5475, 2010.

[114] M. Bartlett and J. Cussens, "Integer Linear Programming for the Bayesian Network Structure Learning Problem," *Artificial Intelligence,* vol. 244, pp. 258-271, 2017.

[115] S. Lee and S. B. Kim, "Parallel Simulated Annealing with a Greedy Algorithm for Bayesian Network Structure Learning," *IEEE Transactions on Knowledge and Data Engineering,* vol. 32, no. 6, pp. 1157-1166, 2019.

[116] M. A. Carrillo, F. J. C. Ortiz, R. Morales-Menéndez, and L. E. Garza-Castañón, "Learning Bayesian Network Structures from Small Datasets Using Simulated Annealing and Bayesian Score," in *Artificial Intelligence and Applications*, 2005, pp. 375-380.

[117] X. Zhou, X. Wang, and E. R. Dougherty, "Construction of Genomic Networks Using Mutual-Information Clustering and Reversible-Jump Markov-Chain-Monte-Carlo Predictor Design," *Signal Processing,* vol. 83, no. 4, pp. 745-761, 2003.

[118] I. Tsamardinos, L. E. Brown, and C. F. Aliferis, "The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm," *Machine learning,* vol. 65, no. 1, pp. 31-78, 2006.

[119] A. L. Madsen, F. Jensen, A. Salmerón, H. Langseth, and T. D. Nielsen, "A Parallel Algorithm for Bayesian Network Structure Learning from Large Data Sets," *Knowledge-Based Systems,* vol. 117, pp. 46-55, 2017.

[120] H. Zhao and Z.-H. Duan, "Cancer Genetic Network Inference Using Gaussian Graphical Models," *Bioinformatics and biology insights,* vol. 13, p. 1177932219839402, 2019.

[121] G. Sanguinetti and V. A. Huynh-Thu, *Gene Regulatory Networks*. Springer, 2019.

[122] J. C. Mar, "The Rise of the Distributions: Why Non-Normality Is Important for Understanding the Transcriptome and Beyond," *Biophysical reviews,* vol. 11, no. 1, pp. 89-94, 2019.

[123] G. A. Seber and A. J. Lee, *Linear Regression Analysis*. John Wiley & Sons, 2012.

[124] M. L. Thompson, "Selection of Variables in Multiple Regression: Part I. A Review and Evaluation," *International Statistical Review/Revue Internationale de Statistique,* pp. 1-19, 1978.

[125] J. Sun *et al.*, "Tslrf: Two-Stage Algorithm Based on Least Angle Regression and Random Forest in Genome-Wide Association Studies," *Scientific reports,* vol. 9, no. 1, pp. 1-10, 2019.

[126] J. C. Engelmann and R. Spang, "A Least Angle Regression Model for the Prediction of Canonical and Non-Canonical Mirna-Mrna Interactions," *PloS one,* vol. 7, no. 7, p. e40634, 2012.

[127] V. Fonti and E. Belitser, "Feature Selection Using Lasso," *VU Amsterdam Research Paper in Business Analytics,* vol. 30, pp. 1-25, 2017.

[128] R. Muthukrishnan and R. Rohini, "Lasso: A Feature Selection Technique in Predictive Modeling for Machine Learning," in *2016 IEEE international conference on advances in computer applications (ICACA)*, 2016: IEEE, pp. 18-20.

[129] T. Mehmood, K. H. Liland, L. Snipen, and S. Sæbø, "A Review of Variable Selection Methods in Partial Least Squares Regression," *Chemometrics and intelligent laboratory systems,* vol. 118, pp. 62-69, 2012.

[130] S. Datta, "Exploring Relationships in Gene Expressions: A Partial Least Squares Approach," *Gene Expression The Journal of Liver Research,* vol. 9, no. 6, pp. 249-255, 2001.

[131] A.-C. Haury, F. Mordelet, P. Vera-Licona, and J.-P. Vert, "Tigress: Trustful Inference of Gene Regulation Using Stability Selection," *BMC systems biology,* vol. 6, no. 1, pp. 1-17, 2012.

[132] J. Ruyssinck, V. A. Huynh-Thu, P. Geurts, T. Dhaene, P. Demeester, and Y. Saeys, "Nimefi: Gene Regulatory Network Inference Using Multiple Ensemble Feature Importance Algorithms," *PLoS One,* vol. 9, no. 3, p. e92709, 2014.

[133] P.-C. Aubin-Frankowski and J.-P. Vert, "Gene Regulation Inference from Single-Cell Rna-Seq Data with Linear Differential Equations and Velocity Inference," *Bioinformatics,* vol. 36, no. 18, pp. 4774-4780, 2020.

[134] H. Matsumoto *et al.*, "Scode: An Efficient Regulatory Network Inference Algorithm from Single-Cell Rna-Seq During Differentiation," *Bioinformatics,* vol. 33, no. 15, pp. 2314-2321, 2017.

[135] T. E. Chan, M. P. Stumpf, and A. C. Babtie, "Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures," *Cell systems,* vol. 5, no. 3, pp. 251-267. e3, 2017.

[136] P. Geurts, "Dyngenie3: Dynamical Genie3 for the Inference of Gene Networks from Time Series Expression Data," *Scientific reports,* vol. 8, no. 1, pp. 1-12, 2018.

[137] V. G. Cheung and R. S. Spielman, "The Genetics of Variation in Gene Expression," *Nature genetics,* vol. 32, no. 4, pp. 522-525, 2002.

[138] M. Morley *et al.*, "Genetic Analysis of Genome-Wide Variation in Human Gene Expression," *Nature,* vol. 430, no. 7001, pp. 743-747, 2004.

[139] A. Subramanian *et al.*, "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles," *Proceedings of the National Academy of Sciences,* vol. 102, no. 43, pp. 15545-15550, 2005.

[140] S. Qin, F. Ma, and L. Chen, "Gene Regulatory Networks by Transcription Factors and Micrornas in Breast Cancer," *Bioinformatics,* vol. 31, no. 1, pp. 76-83, 2015.

[141] F. Emmert-Streib, R. de Matos Simoes, P. Mullan, B. Haibe-Kains, and M. Dehmer, "The Gene Regulatory Network for Breast Cancer: Integrated Regulatory Landscape of Cancer Hallmarks," *Frontiers in genetics,* vol. 5, p. 15, 2014.

[142] K. Glass, C. Huttenhower, J. Quackenbush, and G.-C. Yuan, "Passing Messages between Biological Networks to Refine Predicted Interactions," *PloS one,* vol. 8, no. 5, p. e64832, 2013.

[143] K. Raza and R. Jaiswal, "Reconstruction and Analysis of Cancer-Specific Gene Regulatory Networks from Gene Expression Profiles," *arXiv preprint arXiv:1305.5750,* 2013.

[144] H. A. Chowdhury, D. K. Bhattacharyya, and J. K. Kalita, "(Differential) Co-Expression Analysis of Gene Expression: A Survey of Best Practices," *IEEE/ACM transactions on computational biology and bioinformatics,* vol. 17, no. 4, pp. 1154-1173, 2019.

[145] J. Yang *et al.*, "Dcgl V2. 0: An R Package for Unveiling Differential Regulation from Differential Co-Expression," *PloS one,* vol. 8, no. 11, p. e79729, 2013.

[146] S. Okawa, V. E. Angarica, I. Lemischka, K. Moore, and A. Del Sol, "A Differential Network Analysis Approach for Lineage Specifier Prediction in Stem Cell Subpopulations," *NPJ systems biology and applications,* vol. 1, no. 1, pp. 1-8, 2015.

[147] R. Mall, L. Cerulo, H. Bensmail, A. Iavarone, and M. Ceccarelli, "Detection of Statistically Significant Network Changes in Complex Biological Networks," *BMC systems biology,* vol. 11, no. 1, pp. 1-17, 2017.

[148] G. Consortium, "The Genotype-Tissue Expression (Gtex) Pilot Analysis: Multitissue Gene Regulation in Humans," *Science,* vol. 348, no. 6235, pp. 648-660, 2015.

[149] L. Goodfellow, "Globaldata Epidemiologist Report: Global Covid Cases near 195 Million," in *Pharmaceutical Technology*, ed: Advanstar Communications Inc., 2021.

[150] W. S. T. Consortium, "Repurposed Antiviral Drugs for Covid-19—Interim Who Solidarity Trial Results," *New England journal of medicine,* vol. 384, no. 6, pp. 497-511, 2021.

[151] S. Hwang *et al.*, "Humannet V2: Human Gene Networks for Disease Research," *Nucleic acids research,* vol. 47, no. D1, pp. D573-D580, 2019.

[152] G. Manning, D. B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam, "The Protein Kinase Complement of the Human Genome," *Science,* vol. 298, no. 5600, pp. 1912-1934, 2002.

[153] H.-M. Zhang *et al.*, "Animaltfdb 2.0: A Resource for Expression, Prediction and Functional Study of Animal Transcription Factors," *Nucleic acids research,* vol. 43, no. D1, pp. D76-D81, 2015.

[154] D. Blanco-Melo *et al.*, "Sars-Cov-2 Launches a Unique Transcriptional Signature from in Vitro, Ex Vivo, and in Vivo Systems," *BioRxiv,* 2020.

[155] A. Subramanian *et al.*, "A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles," *Cell,* vol. 171, no. 6, pp. 1437-1452. e17, 2017.

[156] P. V. Hornbeck *et al.*, "Phosphositeplus: A Comprehensive Resource for Investigating the Structure and Function of Experimentally Determined Post-Translational Modifications in Man and Mouse," *Nucleic acids research,* vol. 40, no. D1, pp. D261-D270, 2012.

[157] J. Hu, H.-S. Rho, R. H. Newman, J. Zhang, H. Zhu, and J. Qian, "Phosphonetworks: A Database for Human Phosphorylation Networks," *Bioinformatics,* vol. 30, no. 1, pp. 141-142, 2014.

[158] F. Cheng, P. Jia, Q. Wang, and Z. Zhao, "Quantitative Network Mapping of the Human Kinome Interactome Reveals New Clues for Rational Kinase Inhibitor Discovery and Individualized Cancer Therapy," *Oncotarget,* vol. 5, no. 11, p. 3697, 2014.

[159] F. Hufsky *et al.*, "Computational Strategies to Combat Covid-19: Useful Tools to Accelerate Sars-Cov-2 and Coronavirus Research," *Briefings in bioinformatics,* vol. 22, no. 2, pp. 642-663, 2021.

[160] S. Posada-Céspedes, D. Seifert, I. Topolsky, K. P. Jablonski, K. J. Metzner, and N. Beerenwinkel, "V-Pipe: A Computational Pipeline for Assessing Viral Genetic Diversity from High-Throughput Data," *Bioinformatics,* 2021.

[161] U. Consortium, "Uniprot: A Worldwide Hub of Protein Knowledge," *Nucleic acids research,* vol. 47, no. D1, pp. D506-D515, 2019.

[162] B. L. Le *et al.*, "Transcriptomics-Based Drug Repositioning Pipeline Identifies Therapeutic Candidates for Covid-19," *Scientific reports,* vol. 11, no. 1, pp. 1-14, 2021.

[163] Y. Chen, D. McCarthy, M. Robinson, and G. K. Smyth, "Edger: Differential Expression Analysis of Digital Gene Expression Data User's Guide," *Bioconductor User's Guide,* 2014.

[164] G. K. Smyth, "Limma: Linear Models for Microarray Data," in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*: Springer, 2005, pp. 397-420.

[165] Y. Li and J. C. Patra, "Genome-Wide Inferring Gene–Phenotype Relationship by Walking on the Heterogeneous Network," *Bioinformatics,* vol. 26, no. 9, pp. 1219-1224, 2010.

[166] C. Blatti and S. Sinha, "Characterizing Gene Sets Using Discriminative Random Walks with Restart on Heterogeneous Biological Networks," *Bioinformatics,* vol. 32, no. 14, pp. 2167-2175, 2016.

[167] K. G. Lokugamage *et al.*, "Type I Interferon Susceptibility Distinguishes Sars-Cov-2 from Sars-Cov," *Journal of virology,* vol. 94, no. 23, pp. e01410-20, 2020.

[168] R. Boudewijns *et al.*, "Stat2 Signaling Restricts Viral Dissemination but Drives Severe Pneumonia in Sars-Cov-2 Infected Hamsters," *Nature communications,* vol. 11, no. 1, pp. 1-10, 2020.

[169] R. Vishnubalaji, H. Shaath, and N. M. Alajez, "Protein Coding and Long Noncoding Rna (Lncrna) Transcriptional Landscape in Sars-Cov-2 Infected Bronchial Epithelial Cells Highlight a Role for Interferon and Inflammatory Response," *Genes,* vol. 11, no. 7, p. 760, 2020.

[170] I. Yevshin, R. Sharipov, T. Valeev, A. Kel, and F. Kolpakov, "Gtrd: A Database of Transcription Factor Binding Sites Identified by Chip-Seq Experiments," *Nucleic acids research,* p. gkw951, 2016.

[171] E. Mantlo, N. Bukreyeva, J. Maruyama, S. Paessler, and C. Huang, "Antiviral Activities of Type I Interferons to Sars-Cov-2 Infection," *Antiviral research,* vol. 179, p. 104811, 2020.

[172] X. Lei *et al.*, "Activation and Evasion of Type I Interferon Responses by Sars-Cov-2," *Nature communications,* vol. 11, no. 1, pp. 1-12, 2020.

[173] J. Schindler, J. Monahan, and W. Smith, "P38 Pathway Kinases as Anti-Inflammatory Drug Targets," *Journal of dental research,* vol. 86, no. 9, pp. 800-811, 2007.

[174] P. Cohen, "Protein Kinases—the Major Drug Targets of the Twenty-First Century?," *Nature reviews Drug discovery,* vol. 1, no. 4, pp. 309-315, 2002.

[175] S. Lapenna and A. Giordano, "Cell Cycle Kinases as Therapeutic Targets for Cancer," *Nature reviews Drug discovery,* vol. 8, no. 7, pp. 547-566, 2009.

[176] M. Bouhaddou *et al.*, "The Global Phosphorylation Landscape of Sars-Cov-2 Infection," *Cell,* vol. 182, no. 3, pp. 685-712. e19, 2020.

[177] J. M. Grimes and K. V. Grimes, "P38 Mapk Inhibition: A Promising Therapeutic Approach for Covid-19," *Journal of Molecular and Cellular Cardiology,* vol. 144, pp. 63-65, 2020.

[178] F. Seif, M. Pornour, and D. Mansouri, "Combination of Jakinibs with Methotrexate or Anti-Cytokine Biologics in Patients with Severe Covid-19," *International Archives of Allergy and Immunology,* vol. 181, no. 8, pp. 648-649, 2020.

[179] M. E. Smoot, K. Ono, J. Ruscheinski, P.-L. Wang, and T. Ideker, "Cytoscape 2.8: New Features for Data Integration and Network Visualization," *Bioinformatics,* vol. 27, no. 3, pp. 431-432, 2011.

[180] C. Wang *et al.*, "The Establishment of Reference Sequence for Sars ‐ Cov ‐ 2 and Variation Analysis," *Journal of medical virology,* vol. 92, no. 6, pp. 667-674, 2020.

[181] T. Koyama, D. Platt, and L. Parida, "Variant Analysis of Sars-Cov-2 Genomes," *Bulletin of the World Health Organization,* vol. 98, no. 7, p. 495, 2020.

[182] H. Tegally *et al.*, "Detection of a Sars-Cov-2 Variant of Concern in South Africa," *Nature,* vol. 592, no. 7854, pp. 438-443, 2021.

[183] J. W. Tang, P. A. Tambyah, and D. S. Hui, "Emergence of a New Sars-Cov-2 Variant in the Uk," *Journal of Infection,* vol. 82, no. 4, pp. e27-e28, 2021.

[184] E. B. Hodcroft *et al.*, "Emergence and Spread of a Sars-Cov-2 Variant through Europe in the Summer of 2020," *MedRxiv,* 2020.

[185] B. Gautam, K. Goswami, N. S. Mishra, G. Wadhwa, and S. Singh, "The Role of Bioinformatics in Epigenetics," in *Current Trends in Bioinformatics: An Insight*: Springer, 2018, pp. 39-53.

[186] A.-K. Hadjantonakis, E. D. Siggia, and M. Simunovic, "In Vitro Modeling of Early Mammalian Embryogenesis," *Current opinion in biomedical engineering,* vol. 13, pp. 134-143, 2020.

[187] J. S. Odorico, D. S. Kaufman, and J. A. Thomson, "Multilineage Differentiation from Human Embryonic Stem Cell Lines," *Stem cells,* vol. 19, no. 3, pp. 193-204, 2001.

[188] J. Itskovitz-Eldor *et al.*, "Differentiation of Human Embryonic Stem Cells into Embryoid Bodies Comprising the Three Embryonic Germ Layers," *Molecular medicine,* vol. 6, no. 2, pp. 88-95, 2000.

[189] T. Timeva, A. Shterev, and S. Kyurkchiev, "Recurrent Implantation Failure: The Role of the Endometrium," *Journal of reproduction & infertility,* vol. 15, no. 4, p. 173, 2014.

[190] R. Nievelstein, J. Van der Werff, F. Verbeek, J. Valk, and C. Vermeij ‐ Keers, "Normal and Abnormal Embryonic Development of the Anorectum in Human Embryos," *Teratology,* vol. 57, no. 2, pp. 70-78, 1998.

[191] J. Rossant and P. P. Tam, "New Insights into Early Human Development: Lessons for Stem Cell Derivation and Differentiation," *Cell stem cell,* vol. 20, no. 1, pp. 18-28, 2017.

[192] S. Pepke, B. Wold, and A. Mortazavi, "Computation for Chip-Seq and Rna-Seq Studies," *Nature methods,* vol. 6, no. 11, pp. S22-S32, 2009.

[193] J. Qin, Y. Hu, F. Xu, H. K. Yalamanchili, and J. Wang, "Inferring Gene Regulatory Networks by Integrating Chip-Seq/Chip and Transcriptome Data Via Lasso-Type Regularization Methods," *Methods,* vol. 67, no. 3, pp. 294-303, 2014.

[194] C. Angelini and V. Costa, "Understanding Gene Regulatory Mechanisms by Integrating Chip-Seq and Rna-Seq Data: Statistical Solutions to Biological Problems," *Frontiers in cell and developmental biology,* vol. 2, p. 51, 2014.

[195] E. T. Liu, S. Pott, and M. Huss, "Q&A: Chip-Seq Technologies and the Study of Gene Regulation," *BMC biology,* vol. 8, no. 1, pp. 1-6, 2010.

[196] T. S. Furey, "Chip–Seq and Beyond: New and Improved Methodologies to Detect and Characterize Protein–DNA Interactions," *Nature Reviews Genetics,* vol. 13, no. 12, pp. 840-852, 2012.

[197] R. N. Sharipov, I. S. Yevshin, Y. V. Kondrakhin, A. S. Ryabova, S. K. Kolmykov, and F. A. Kolpakov, "Peak Caller Comparison through Quality Control of Chip-Seq Datasets," in *Bioinformatics of Genome Regulation and Structure/Systems Biology (BGRS/SB-2020)*, 2020, pp. 105-106.

[198] M. Love, S. Anders, and W. Huber, "Differential Analysis of Count Data–the Deseq2 Package," *Genome Biol,* vol. 15, no. 550, pp. 10-1186, 2014.

[199] N. Leng *et al.*, "Ebseq: An Empirical Bayes Hierarchical Model for Inference in Rna-Seq Experiments," *Bioinformatics,* vol. 29, no. 8, pp. 1035-1043, 2013.

[200] P. H. Westfall and S. S. Young, *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*. John Wiley & Sons, 1993.

[201] N. Wani and K. Raza, "Raw Sequence to Target Gene Prediction: An Integrated Inference Pipeline for Chip-Seq and Rna-Seq Datasets," in *Applications of Artificial Intelligence Techniques in Engineering*: Springer, 2019, pp. 557-568.

[202] G. Yu, L.-G. Wang, and Q.-Y. He, "Chipseeker: An R/Bioconductor Package for Chip Peak Annotation, Comparison and Visualization," *Bioinformatics,* vol. 31, no. 14, pp. 2382-2383, 2015.

[203] X. Chen *et al.*, "Chip-Bit: Bayesian Inference of Target Genes Using a Novel Joint Probabilistic Model of Chip-Seq Profiles," *Nucleic acids research,* vol. 44, no. 7, pp. e65-e65, 2016.

[204] T. O'Connor, C. E. Grant, M. Bodén, and T. L. Bailey, "T-Gene: Improved Target Gene Prediction," *Bioinformatics,* vol. 36, no. 12, pp. 3902-3904, 2020.

[205] J. Salvatier, T. V. Wiecki, and C. Fonnesbeck, "Probabilistic Programming in Python Using Pymc3," *PeerJ Computer Science,* vol. 2, p. e55, 2016.

[206] R. Kumar, C. Carroll, A. Hartikainen, and O. A. Martín, "Arviz a Unified Library for Exploratory Analysis of Bayesian Models in Python," 2019.

[207] G. Wang, "Bayesian Regression Models for Ecological Count Data in Pymc3," *Ecological Informatics,* vol. 63, p. 101301, 2021.

[208] D. Van Ravenzwaaij, P. Cassey, and S. D. Brown, "A Simple Introduction to Markov Chain Monte–Carlo Sampling," *Psychonomic bulletin & review,* vol. 25, no. 1, pp. 143-154, 2018.

[209] M. Hecker, S. Lambeck, S. Toepfer, E. Van Someren, and R. Guthke, "Gene Regulatory Network Inference: Data Integration in Dynamic Models—a Review," *Biosystems,* vol. 96, no. 1, pp. 86-103, 2009.

[210] I. Shmulevich and J. D. Aitchison, "Deterministic and Stochastic Models of Genetic Regulatory Networks," *Methods in enzymology,* vol. 467, pp. 335-356, 2009.

[211] M. Altaf-Ul-Amin, T. Katsuragi, T. Sato, and S. Kanaya, "A Glimpse to Background and Characteristics of Major Molecular Biological Networks," *BioMed Research International,* vol. 2015, 2015.

[212] G. Box, "Signal-to-Noise Ratios, Performance Criteria, and Transformations," *Technometrics,* vol. 30, no. 1, pp. 1-17, 1988.

[213] T. Subkhankulova, F. Naumenko, O. E. Tolmachov, and Y. L. Orlov, "Novel Chip-Seq Simulating Program with Superior Versatility: Ischip," *Briefings in Bioinformatics,* vol. 22, no. 4, p. bbaa352, 2021.

[214] G. Sanguinetti, "Gene Regulatory Network Inference: An Introductory Survey," in *Gene Regulatory Networks*: Springer, 2019, pp. 1-23.

[215] R. Edgar, M. Domrachev, and A. E. Lash, "Gene Expression Omnibus: Ncbi Gene Expression and Hybridization Array Data Repository," *Nucleic acids research,* vol. 30, no. 1, pp. 207-210, 2002.

[216] H. Hu, Y.-R. Miao, L.-H. Jia, Q.-Y. Yu, Q. Zhang, and A.-Y. Guo, "Animaltfdb 3.0: A Comprehensive Resource for Annotation and Prediction of Animal Transcription Factors," *Nucleic acids research,* vol. 47, no. D1, pp. D33-D38, 2019.

[217] A. D. Yates *et al.*, "Ensembl 2020," *Nucleic acids research,* vol. 48, no. D1, pp. D682-D688, 2020.

[218] Y. Zhang, G. Parmigiani, and W. E. Johnson, "Combat-Seq: Batch Effect Adjustment for Rna-Seq Count Data," *NAR genomics and bioinformatics,* vol. 2, no. 3, p. lqaa078, 2020.

[219] J. Feng, T. Liu, and Y. Zhang, "Using Macs to Identify Peaks from Chip ‐ Seq Data," *Current protocols in bioinformatics,* vol. 34, no. 1, pp. 2.14. 1-2.14. 14, 2011.

[220] Q. Li, J. B. Brown, H. Huang, and P. J. Bickel, "Measuring Reproducibility of High-Throughput Experiments," *The annals of applied statistics,* vol. 5, no. 3, pp. 1752-1779, 2011.

[221] C. S. Wagner *et al.*, "Approaches to Understanding and Measuring Interdisciplinary Scientific Research (Idr): A Review of the Literature," *Journal of informetrics,* vol. 5, no. 1, pp. 14-26, 2011.

[222] A. R. Quinlan and I. M. Hall, "Bedtools: A Flexible Suite of Utilities for Comparing Genomic Features," *Bioinformatics,* vol. 26, no. 6, pp. 841-842, 2010.

[223] H. M. Amemiya, A. Kundaje, and A. P. Boyle, "The Encode Blacklist: Identification of Problematic Regions of the Genome," *Scientific reports,* vol. 9, no. 1, pp. 1-5, 2019.

[224] V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan, "Finding a" Kneedle" in a Haystack: Detecting Knee Points in System Behavior," in *2011 31st international conference on distributed computing systems workshops*, 2011: IEEE, pp. 166-171.

[225] H. Yao, K. Brick, Y. Evrard, T. Xiao, R. D. Camerini-Otero, and G. Felsenfeld, "Mediation of Ctcf Transcriptional Insulation by Dead-Box Rna-Binding Protein P68 and Steroid Receptor Rna Activator Sra," *Genes & development,* vol. 24, no. 22, pp. 2543-2555, 2010.

[226] X. Zhang *et al.*, "Pax6 Is a Human Neuroectoderm Cell Fate Determinant," *Cell stem cell,* vol. 7, no. 1, pp. 90-100, 2010.

[227] K. Takayama *et al.*, "Efficient and Directive Generation of Two Distinct Endoderm Lineages from Human Escs and Ipscs by Differentiation Stage-Specific Sox17 Transduction," *PloS one,* vol. 6, no. 7, p. e21780, 2011.

[228] M. Shimoda *et al.*, "Sox17 Plays a Substantial Role in Late-Stage Differentiation of the Extraembryonic Endoderm in Vitro," *Journal of cell science,* vol. 120, no. 21, pp. 3859-3869, 2007.

[229] C. Li, Y.-P. Li, X.-Y. Fu, and C.-X. Deng, "Anterior Visceral Endoderm Smad4 Signaling Specifies Anterior Embryonic Patterning and Head Induction in Mice," *International journal of biological sciences,* vol. 6, no. 6, p. 569, 2010.

[230] H. Xu *et al.*, "Escape: Database for Integrating High-Content Published Data Collected from Human and Mouse Embryonic Stem Cells," *Database,* vol. 2013, 2013.

[231] L. J. Jensen *et al.*, "String 8—a Global View on Proteins and Their Functional Interactions in 630 Organisms," *Nucleic acids research,* vol. 37, no. suppl_1, pp. D412-D416, 2009.

[232] R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef, "Deep Generative Modeling for Single-Cell Transcriptomics," *Nature methods,* vol. 15, no. 12, pp. 1053-1058, 2018.

[233] A. Gregorieff and H. Clevers, "Wnt Signaling in the Intestinal Epithelium: From Endoderm to Cancer," *Genes & development,* vol. 19, no. 8, pp. 877-890, 2005.

[234] W. Jiang, D. Zhang, N. Bursac, and Y. Zhang, "Wnt3 Is a Biomarker Capable of Predicting the Definitive Endoderm Differentiation Potential of Hescs," *Stem cell reports,* vol. 1, no. 1, pp. 46-52, 2013.

[235] S. Ma and Y. Zhang, "Profiling Chromatin Regulatory Landscape: Insights into the Development of Chip-Seq and Atac-Seq," *Molecular Biomedicine,* vol. 1, no. 1, pp. 1-13, 2020.

[236] C. Jansen *et al.*, "Uncovering the Mesendoderm Gene Regulatory Network through Multi-Omic Data Integration," *Biorxiv,* 2020.

[237] S. Li, H. Yan, and J. Lee, "Identification of Gene Regulatory Networks from Single-Cell Expression Data," in *Modeling Transcriptional Regulation*: Springer, 2021, pp. 153-170.

[238] T. Salimans, D. Kingma, and M. Welling, "Markov Chain Monte Carlo and Variational Inference: Bridging the Gap," in *International Conference on Machine Learning*, 2015: PMLR, pp. 1218-1226.

[239] H. Okae *et al.*, "Derivation of Human Trophoblast Stem Cells," *Cell stem cell,* vol. 22, no. 1, pp. 50-63. e6, 2018.

[240] X. XIE, "Principal Component Analysis," 2019.