Integrative Analysis of Modular Structure of Genes in High-throughput Tumor Profiles

by

Lifan Liang

B.S., Huazhong University of Science and Technology, 2013

M.S. University of Pittsburgh, 2018

Submitted to the Graduate Faculty of the

School of Medicine in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2021

UNIVERSITY OF PITTSBURGH

SCHOOL OF MEDICINE

This dissertation was presented

by

Lifan Liang

It was defended on

September 30, 2021

and approved by

Xinghua Lu, Professor, Department of Biomedical Informatics

Greg Cooper, Professor, Department of Biomedical Informatics

George Tseng, Professor, Department of Biostatistics

Dissertation Director: Songjian Lu, Assistant Professor, Department of Biomedical Informatics

Integrative Analysis of Modular Structure of Genes in High-throughput Tumor Profiles

Lifan Liang, PhD.

University of Pittsburgh, 2021

Cellular functions, such as signal transduction, transportation, cell cycle, and various metabolism, require cooperation of many gene products. Following the central dogma, such large-scale cooperation within and across cells often leave traces on different omics profiles. One major clue would be the strong correlation among genes in genomics, epigenetics, transcriptomics, and proteomics. Based on this premise, we started to identify functional modules by integrating pairwise correlation among genes from different information sources into the form of multiplex networks. Although all the layers of the multiplex shared the same protein interactome as the skeleton, edge weights in each layer represents pairwise correlation from a different type of information sources. This formation allows information flow from one data source to another. We also designed a novel graph clustering algorithm to detect gene sets with strong correlations inside.

However, the multiplex integration only yields marginal improvement against single omics. We turn to the mutual exclusivity patterns in cancer genomics. This pattern suggests that a single somatic alteration event may be sufficient to promote tumorigenesis. We pushed the assumption further to state that disruption of a single pathway could lead to differential expression of a large set of genes, which is supported by our work on Boolean matrix factorization. Then we proposed the OR-gate network (ORN) to model the causal mechanism from somatic alterations to transcriptomics. Results showed that it is able to recover the heterogeneity among cancer samples and functional modules responsible for certain dysregulation in cancer transcriptomics.

Still, ORN has two major limitations. One is the issue of co-amplification. ORN cannot distinguish passengers in the same copy number variation hotspot as the drivers. To this end, we applied the word2vec model to extract gene embedding from biomedical literature. Another issue is the transcriptional regulation module may not be accurate. To this end, we developed a novel algorithm (peak2vec) to uncover transcriptional motif patterns and coregulation from the chromatic accessibility profiles.

In the future, we will integrate gene embedding and peak2vec into the ORN framework to better understand the causal impact of somatic alteration as functional modules.

Table of Contents

List of Tables

List of Figures

Preface

To reach the point of drafting a dissertation thesis, I am grateful to many people here at Pitt or Pittsburgh. First of all, my research advisor, Dr. Songjian Lu, always encourages me to try out new ideas and new directions. He and Dr. Xinghua Lu have provided much helpful feedback. I also appreciated the help and company of many lab members, including Kunju Zhu, Xiaonan Fan, and Junyan Tao.

I have learned a lot from the courses at Pitt, especially those provided by Dr. Vanathi Gopalakrishnan, Dr. Roger Day, Dr. George Tseng, and Dr. Douglas Landsittel. Knowledge and skills from these classes are foundational to my work described in this thesis.

The strong sense of community in DBMI provides great support to me, especially in the first two years when I struggled the most. Dr. Gregory Cooper and Dr. Harry Hocheisser have provided valuable help and guidance on my research projects. Special thanks to Toni Porterfield who can handle everything with a smile. Rob Cecchetti and Genine Bartolotta have helped me deal with many problems with great patience.

Finally, I am grateful to my family for being supportive of my career decision at every step of my life.

Abbreviations

BMF: Boolean matrix factorization

BEM: Boolean matrix factorization with EM algorithm

ORN: OR-gate network

TCGA: the Cancer Genome Atlas

SGA: somatic genomic alteration

CNV: copy number variation

SNP: single nucleotide polymorphism

ME: mutual exclusivity

DEG: differentially expressed genes

TFBS: transcription factor binding site

PFM: position frequency matrix

PPM: position probability matrix

1.0 Introduction

Cancer is a disease caused by genetic changes leading to uncontrolled cell growth and invasion into nearby tissues. Hence biological research has focused on characterizing molecular mechanisms of somatic mutations in cancer cells. Such efforts have facilitated the development of precision oncology and greatly improved clinical outcomes of cancer patients. For example, the discovery of somatic mutations in the BRAF gene in 66% of melanoma cell line (H. Davies et al. 2002) has led to the development of Vemurafenib, a new standard of care for patients with BRAF-V600 mutant melanoma (Fisher and Larkin 2012). The success of targeted therapy demonstrated the importance of pathway-level understanding in cancer biology.

However, despite the improving survival rates, cancer remained the second-leading cause of death in the world (GBD 2018). This is because most cancer patients cannot be cured with targeted therapies. For example, the percentage of patients estimated to respond to immunotherapy targeting PD1 and CTLA4 was 12.46% (Haslam and Prasad 2019). And the reason why patients differ in responsiveness is still unknown, indicating our lack of understanding of how genomic changes perturbed signaling pathways that underlie cancer phenotypes.

Researchers on the computational side have proposed various approaches to investigate pathway dysregulation in cancer since the emergence of large high-throughput data repositories related to cancer. The Cancer Genome Atlas (TCGA) is the earliest effort to provide molecular profiles of patients from most cancer types, including genomic profile, transcriptomic profile, epigenomic profile, and proteomic profiles. Other initiatives, such as METABRIC (Curtis et al. 2012) and TARGET (Gerhard et al. 2018), have generated similar omics profiles of patients from certain cancer types. With emerging datasets and computational methods, precision oncology has

made tremendous progress. However, these computational approaches have limitations yet to be resolved. As illustrated in detail in the background section, single-omics approach may fail to capture data patterns that are truly related to disease mechanism. On the other hand, integrative analysis either relies too much on current knowledge or imposes assumptions not plausible in biology.

In this project, we developed several novel algorithms for pattern recognition problems in genomic, transcriptomic, and proteomic. Isolation clustering (Chapter 3) identified functional modules over the multiplex in a greedy way. BMF (Chapter 4) identified coregulation patterns in transcriptomics. ORN (Chapter 5) links genomics and transcriptomics together to identify latent pathways of somatic mutations that cause differential expression in transcriptomics. Gene embedding (Chapter 6) utilized knowledge curated in literature to summarize functional similarities among genes. Peak2vec (Chapter 7) learns motif mixture in the ATAC-seq to infer gene regulation. In the future, we aim to integrate all these models from different data sources into the framework of ORN, so we may conduct de novo inference of signaling pathways in cancer reliably.

1.1 Hypothesis and Specific Aims

The major hypothesis of this dissertation is that different data sources provide different clues to the functional interaction among genes. Specifically, we hypothesize that: (1) integrating multiple information sources improves the performance of functional module identification; (2) coexpression among genes hints at their functional similarity and their coregulation; (3) functional relationships among SGAs are implicated by their impact on transcriptomics; (4) biomedical

literature is a reliable data source to infer functional relationships among genes; (5) chromatin accessibility combined with DNA sequence information can reveal gene regulation.

To examine these hypotheses, this dissertation explored multiple omics data with the following specific aims:

1. **Isolation clustering**: As a preliminary attempt of integrative analysis of gene functions, we directly integrated various information into the form of a multiplex network. Information sources include protein-protein interaction, biomedical topic modeling, and gene expression profiles. To verify the performance gain of such network formation, we developed isolation clustering to identify functional modules and protein complexes in this network.
2. **BEM**: Within a set of similar tumor samples, a set of genes express with strong correlation. This set of genes may be regulated by the same pathway or participate in the same biological function. By modeling this coexpression pattern with the AND-OR product, we developed a novel Boolean matrix factorization method to systematically discover the coregulated genes sets and heterogeneous tumor situations.
3. **ORN**: The phenomenon of mutual exclusivity indicates that mutation of a single gene may be sufficient to disrupt one or more pathways. By integrating genomic and transcriptomic profiles of cancer samples with OR-gate relationship, it is possible to characterize the functioning role of somatic mutations in cancer and infer patient-specific pathway activities de novo.
4. **Gene embedding**: Neighboring passenger mutations are often amplified together with the driver mutations. In this case, we cannot distinguish driver mutations according to their functional impact. Hence, we applied word2vec to biomedical literature to

generate embedding vectors for genes. To avoid overrepresentation of well-known genes, we learned gene embedding from their semantic representation instead.

5. **Peak2vec**: BMF identified coexpression patterns with gene-level read count information. To complement this approach, we explored the usefulness of DNA sequence information by adapting deep learning models to recognize motif patterns in ATAC-seq peak sequences.

1.2 Outline of the Dissertation

Chapter 2 provides the research background and datasets in use for the dissertation so as to identify the gap in current research and motivate our research projects. Chapter 3 describes our initial attempt to integrate transcriptomics, proteomics, and literature to identify functional modules. Chapter 4 describes the Boolean matrix factorization applied to coexpression module identification, which naturally extends to the OR-gate network (ORN) illustrated in Chapter 5. One major limitation of ORN is that it cannot distinguish passenger mutation in the same CNV hotspot as the driver. We demonstrated our solution to address the issue of co-amplification in somatic mutation (Chapter 6). In Chapter 7, we introduce the sequence embedding techniques that may integrate sequence information into the framework. Finally, we will summarize the contribution and limitations of this dissertation project and discuss future research directions in Chapter 8. Our main content has been summarized in Figure 1.

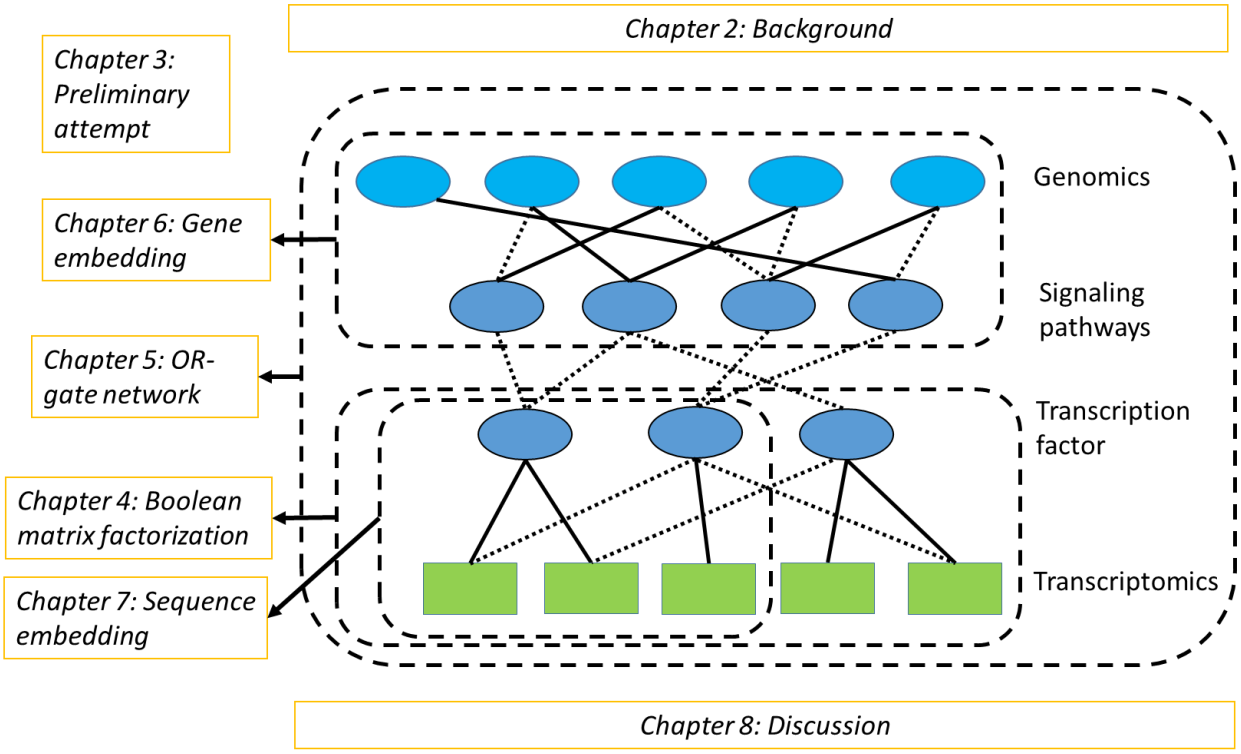


Figure 1 Outline of the dissertation

2.0 Background

This chapter introduced the datasets and methods we used in the subsequent project. Datasets are mostly from the TCGA project. As for the methods, we will describe the background knowledge about Boolean matrix factorization (BMF), word2vec, deep learning, and motif analysis. Since each of these methods represents a vast research field, this chapter only covers the basic knowledge of the computational techniques employed in the dissertation.

2.1 The Cancer Genome Atlas (TCGA)

TCGA (Tomczak, Czerwińska, and Wiznerowicz 2015) is the largest initiative to provide molecular profiles for over 30 types of human tumors. Over the past decade, TCGA has grown in both sample sizes and omics measurements. Currently, TCGA has collected 11315 samples, including from both normal and cancer conditions. Besides whole exome sequencing and RNA-seq data, TCGA has adopted methylation array, whole genome sequencing (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium 2020), ATAC-seq (Corces et al. 2018), and miRNA-seq (Chu et al. 2016), providing a much more comprehensive molecular profile of cancer samples.

In this project, we utilized somatic mutation, copy number variations (CNVs), RNA-seq, and ATAC-seq. These four data types will be described in this section.

Somatic mutations: Somatic mutations consist of single nucleotide polymorphism (SNP) and insertion-deletion mutations (INDEL). Somatic mutations have the potential to alter all steps

of gene expression depending on their genomic location. When present within transcriptional regulatory elements, they can affect mRNA expression. When present within coding regions, mutations can impact mRNA splicing, nucleo-cytoplasmic export, stability, and translation. When present within a coding sequence and leading to an amino acid change (referred to as a non-synonymous SNP or mutation), they can modify the protein's activity. If the mutation is synonymous (i.e., does not change the nature of the amino acid), then translation rates or mRNA half-life may be affected.(Robert and Pelletier 2018) Since this project is linking differential expression to somatic mutation, we only used non-silent gene-level mutations that have been sequenced and processed(Ellrott et al. 2018). The raw input for our algorithm would be a matrix with each row as a gene and each column as a sample. Value "1" in the *i*th row and *j*th column means the *i*th gene mutated in the *j*th sample.

Copy number variation: CNVs are a type of structural variant involving alterations in the number of copies of specific regions of DNA, which can either be deleted or duplicated. In TCGA, the copy number of DNA regions was segmented and estimated from the SNP array data. This project combined CNVs and somatic mutations to generate the somatic genomic alterations (SGAs) profiles of tumor samples. SGAs would be the actual input for the ORN algorithm. Please note that one region with CNV may cover multiple genes, making it difficult to distinguish driver mutations from passengers. We will try to address this issue in the Planned works with gene embedding methods.

Gene expression: RNA-seq measures the volume of gene expression. TCGA has completed over 11,000 RNA-seq assays. RNA expression represented cellular phenotypes of cancer. Molecular subtypes of breast cancer were identified based on gene expression data (Wallden et al. 2015). Another major application of RNA-seq is differential expression analysis.

Differential expression analysis can identify differentially expressed genes (DEGs), which indicate abnormal cellular functions leading to cancer. In this project, we performed differential expression analysis and used the DEG status as the end results for somatic mutations.

ATAC-seq: As the name suggests, Assay of Transposase Accessible Chromatin sequencing (ATAC-seq) aims at locating chromatin open region by sequencing DNA fragments from Tn5 enzyme cleavage (Buenrostro et al. 2013). While other technologies (e.g., DNase-seq, FAIRE-seq, and ChIP-seq) require millions of cells as input materials (Buenrostro et al. 2013), ATAC-seq only requires 500-50000 cells. Its simplicity results in an exponential increase of curated ATAC-seq datasets and publications, indicating its value in a wide spectrum of biological questions (Yan et al. 2020). Computational analysis of ATAC-seq often involves peak calling (Tarbell and Liu 2019), peak-gene linking (Corces et al. 2018; Wang, Jiang, and Wong 2016), nucleosome positioning (Schep et al. 2015), and footprinting analysis (Wang, Jiang, and Wong 2016). In TCGA, ATAC-seq was performed on 410 samples from 404 donors in 23 types of cancer. Analysis revealed over 100K novel regulatory elements and novel cancer subtypes (Corces et al. 2018).

2.2 Boolean Matrix Factorization

Boolean matrix factorization (BMF) is a variant of the standard matrix factorization problem in the Boolean semiring (Pauli Miettinen and Neumann 2020): given a binary matrix, the task is to find lower rank binary matrices so that their AND-OR product (described in Chapter 4) is as close to the original matrix as possible. Since the AND-OR product is a different operation

than the operation in linear algebra, standard matrix factorization techniques, such as singular value decomposition (SVD) and nonnegative matrix factorization (NMF), fail to work. In addition, finding the exact solution is NP-hard (P Miettinen, Mielikäinen, and Gionis 2008).

Therefore, over the past decades, researchers have proposed various techniques to find approximate solutions. The earliest approach is combinatorial optimization exemplified by ASSO (P Miettinen, Mielikäinen, and Gionis 2008), PANDA+ (Lucchese, Orlando, and Perego 2014) and GRECOND+ (Belohlavek and Trnecka 2018). ASSO alternates between the column factor and row factor to optimize the reconstruction error. PANDA+ first finds dense components as the core and extends from them to cover more ones while the error is not increasing. Similarly, GRECOND starts by generating a decomposition that does not over-cover and then iteratively updates to commit over-covering to decrease the error.

A more recent approach tries to find the solution with maximum likelihood, which our work in Chapter 4 belongs to. In order to compute the likelihoods of the lower rank factors, all the methods (Frolov, Husek, and Polyakov 2014; T Rukat, Holmes, and Titsias 2017; S Ravanbakhsh and Póczos 2016) need to make prior assumptions on the data generation process. The fully Bayesian approach (Tammo Rukat et al. 2017) assumed that Boolean factors were generated by a uniform Bernoulli distribution. A more recent method (Neumann 2018) assumed that one side of the matrix only contains clusters with relatively large size. Unfortunately, prior knowledge on factors' sizes is usually unavailable in transcriptomic data. Previous studies have shown that the number of genes within a coexpression module can be less than 100 (Padilha and Campello 2017) or close to 1000 (van Dam et al. 2018). The number of samples within a cluster is even more unpredictable. Thus, assumptions about bicluster sizes may a impose strong bias in gene expression analysis.

In Chapter 4, we presented a new model that introduced a more hierarchical framework that is free of assumptions about the factor size while retaining the advantages of previous algorithms.

2.3 Deep Learning Models

Some models described in this dissertation are inferred by the backpropagation algorithm. Particularly, the ORN (Chapter 5) is structured like the fully connected neural network and the multinomial convolution (Chapter 7) is a variant of the convolutional kernel. This section will briefly describe the structure of fully connected neural network and the kernel operation in convolutional neural network.

The fully connected neural network is a fundamental architecture of deep learning models. Namely, all the nodes in one layer are connected to all the nodes in the next. The connection between node i in layer L and all the nodes in the previous layer can be characterized by two steps: (1) inner product between node values and edge weights; (2) activation function of the product. The nonlinear activation function sets neural networks apart from linear models, enabling deep learning to approximate any function (Hornik, Stinchcombe, and White 1989).

Convolutional neural network (CNN) is usually applied to images with RGB channels. In this case, convolution is applying a cube to scan through the image with multiple kernels. Each kernel is essentially a tensor. When scanning a certain region, the pixel values will be summarized to a scalar that is the sum of the product between the kernel and the region. Therefore, convolution is effectively scanning through the image for the visual pattern defined by the kernel. Note that in order to accelerate the computation, CNN does not scan through the image. Rather, it generates many duplicate kernels to apply to the full image in parallel and ensure they share the same

parameters. Although this parameter sharing scheme requires much more memory space, it significantly reduces the computational complexity of time.

Deep learning has been well known for its prediction performance and computational efficiency. However, deep learning models are difficult to interpret. Although many research methods attempt to open the black box by approximating the trained black box with an interpretable model, we have not seen much success so far. In this dissertation, we took advantage of the deep learning's computational efficiency and force it to learn interpretable patterns from high-throughput data.

2.4 Word Embedding

Word embedding approach originated from the distributional hypothesis, i.e., words that occur in the same contexts tend to purport similar meanings. Researchers in computational linguistics aimed at representing words with vector space, such that words with similar meanings should appear close together in the vector space. The breakthrough in this direction is Word2Vec. It outperformed all the previous approaches, whether neural network methods or matrix factorization methods. Its simplicity also demonstrated superior interpretability. For example, the vector for "king" minus the vector "man" would approximate the vector for "queen". Following word2vec's success is a variety of improved approaches, including factorization of point-wise mutual information matrix (Levy and Goldberg 2014), GloVe (Pennington, Socher, and Manning 2014), and ConceptNet Numberbatch (Speer, Chin, and Havasi 2016).

Recently, the word2vec algorithm has been applied to several bioinformatics studies (Du et al. 2019; S. Kim et al. 2018). These studies used coexpression (Du et al. 2019; Xiangyu Li et al. 2017), co-mutation (S. Kim et al. 2018) or protein-protein interaction (S. Kim et al. 2018) to define the context of genes. These embedding vectors have been applied to tissue deconvolution or tumor driver identification. However, high-throughput data are often confounded with technical variance, especially batch effects (Soneson, Gerster, and Delorenzi 2014). More importantly, the correlation observed in high-throughput data may not indicate functional relations. For example, gene expression of tissue-specific house-keeping genes often fluctuates according to the tissue proportions in samples. In this case, coexpression merely imply that genes express within the same tissue. Gene2vec (Du et al. 2019) also showed that gene embedding constructed from coexpression reflected tissue specificity.

Hence it may be better directly apply word2vec to biomedical literature. However, Biomedical literature may focus on well-known genes and neglect others. In Chapter 6, we constructed semantic representation for each gene before applying the word2vec model. In this way, we may achieve higher order inference in the embedding space so that less well-known genes will also be modeled accurately.

2.5 Transcription Factor Motif Analysis

Sequence motifs are short, recurring patterns in DNA that are presumed to have a biological function (D'haeseleer 2006). Throughout this thesis, motifs indicate sequence-specific binding sites for transcription factors (TF). Although it is well known that TF recognizes specific sequences

of nucleotides, the binding sequence is not identical across binding sites. Hence a popular approach to describe the motifs is through a position frequency matrix (PFM) or a position probability matrix (PPM), which is the format for most databases of TF motifs.

However, the format of PFM or PPM may be too restricted to capture the regularity of the motif recognition mechanism. Early research has suggested that site dependency and the characteristics of protein domain need to be considered in the task of de novo motif detection (Xing and Karp 2004). Currently, computational researchers address this issue with multiple PPM or multiple kmers. For example, Deepbind (Alipanahi et al. 2015; Guo et al. 2018) used multiple convolution kernels to detect the binding affinity of one TF. When performing motif matching, research also showed that using a set of kmers instead of a PPM may improve performance (Guo et al. 2018). As a result, this over-kill fashion of motif detection can only be applied to ChIP-seq where only one putative TF is investigated. When multiple TFs are present in an assay, it is difficult to handle the heterogeneity of TF representation. When dealing with chromatin accessibility profiles where multiple TF activities can be interrogated, researchers often turn to footprinting analysis, despite the fact that only one fifth of TFs exhibit the footprint patterns (Baek, Goldstein, and Hager 2017).

In Chapter 7, we revisit the idea of motif detection in chromatin accessibility, particularly ATAC-seq. We showed that with slight modification of the convolution kernels, it is possible to infer cis-regulatory elements from ATAC-seq.

3.0 Preliminary Attempt: a Multilayer Approach to Identify Functional Modules by Integrating PPI, Gene Expression and Literature

This chapter presented our preliminary attempt at analyzing the modular structure of genes by integrating multi-omics data and literature knowledge with a multiplex network.

3.1 Background of Functional Module Identification

Understanding the mechanisms of pathway perturbations underlying complex human diseases remains a difficult problem, hindering the development of targeted therapeutics. Complex diseases involve many genes and molecules that interact within context-specific cellular networks, such as signaling networks, physical interaction networks, and coexpression networks (Choobdar et al. 2018). For example, cancer was often viewed as the disruption of cellular signaling networks. Such complex networks are inherently modular(Hartwell et al. 1999), meaning that genes usually perform certain biological functions in separate groups. Therefore, to investigate complicated cellular mechanisms, it is necessary to characterize the modular structure of cellular networks.

A functional module is defined as a group of genes or their products that are related by one or more genetic or cellular interactions, e.g., coregulation, coexpression or membership of a protein complex, of a metabolic or signaling pathway, or of a cellular aggregate (e.g. chaperone, ribosome, protein transport facilitator) (Tornow and Mewes 2003). Since physical protein-protein interactions directly indicate the cooperation of gene products to drive a biological process, a variety of clustering methods were developed to identify functional modules from protein-protein

interaction networks (Ji et al. 2014). Zinman, et al. (Zinman, Zhong, and Bar-Joseph 2011) have found that functional interactions that are part of functional modules are conserved at a much higher rate, further supporting the advantage of using protein interaction networks. Unfortunately, the computational methods for functional module identification are clearly limited by the poor quality of the underlying PPI data, which is noisy with high rates of false positive and false negative (Bader et al. 2004; Xiaoli Li et al. 2010).

Another popular approach is to identify functional modules from coexpression network. Unlike protein interaction networks, edges in coexpression networks indicate differential expression of two genes within the same sample or condition. It assumes that tightly interacting and functionally dependent proteins are co-expressed across most conditions. This assumption is a reliable heuristic for functional module identification, despite that coexpression is not direct evidence for functional relation. Studies had successfully identified stable functional modules from coexpression networks across species (Stuart et al. 2003). Therefore, the status of coexpression modules should be highly related to the activities or behavior of cells. Many biological studies have identified active functional modules related to certain diseases from coexpression networks (Shi et al. 2014; You et al. 2016).

However, in the case of coexpression network, identifying functional modules at the appropriate granularity is a big challenge. As each experimental condition usually has perturbed multiple signaling pathways, differentially expressed genes in each condition usually correspond to multiple dysregulated biological processes (Bader et al. 2004). This could result in predicted functional modules being a superset of several real functional modules.

In addition, high-throughput expression data also has its own data quality issues. For example, RNAseq data still suffered from technical issues, such as batch effects and

contamination. Recent studies have developed different methods to improve the accuracy of module identification by integrating coexpression networks and protein interaction networks (Tornow and Mewes 2003; Bader et al. 2004; Huang and Fraenkel 2009; Suthram et al. 2010; Dey, Hsiao, and Stephens 2017). However, data quality issues common in high-throughput data remain unresolved.

Besides high-throughput data, decades of research efforts have obtained and validated vast amounts of biological knowledge through wet-lab experiments, which are valuable resources for further research. Such knowledge should contain much fewer errors compared to high-throughput data. A few studies have attempted to utilized the literature for functional module identification (Y. Liu, Liang, and Wishart 2015; Chen, Paisley, and Lu 2017; J. Kim, Kim, and Lee 2017; Z. Yang et al. 2014). However, relying on literature alone may lead to findings biased towards well studied genes, providing less novel insights.

Since high-throughput data is less biased towards well-known genes and literature has fewer data quality issues, integrating these two information sources seems promising. This study has developed a multiplex clustering method to integrate data extracted from high-throughput experiments and biomedical literature for the purpose of functional module identification. Multiplex is a natural way to represent interactions in a complex system from multiple perspectives ("Networks - Mark Newman - Oxford University Press" n.d.). Random walks on multiplex can induce congestion even when every single layer remains decongested (Solé-Ribalta, Gómez, and Arenas 2016). Also, the fraction of nodes a random walker can travel has increased, owing to their resilience to uniformly random failures (De Domenico et al., n.d.). Thus, the dynamics of diffusion has changed in multiplex. Functional module identification on multiplex is likely to yield different results than each of its single layer.

Two major hypotheses were tested in this project (chapter): (1) gene-topic associations extracted from literature is able to reveal functional relations of genes and provide information complementary to high-throughput data; (2) integration of multiple information sources with multiplex approach can improve the accuracy of functional module identification.

3.2 Construction of the Multiplex Network

3.2.1 Topic modeling of genes

Title and abstract information of biomedical articles were downloaded from PubMed on April 10, 2013. First, by treating each gene as a document, tf-idf scores were calculated to identify words most pertinent to a certain gene. For yeast literature, words with tf-idf scores lower than 53 were removed; and the vocabulary was restricted to 6000. For human literature, the thresholds were 167 and 13000 for tf-idf scores and vocabulary size respectively. Second, a word vector was then created for each gene by going through its list of 200 words with the highest tf-idf scores and including only the ones that occur in the vocabulary. For each sample, whether collected from a yeast perturbation experiment or a cancer patient sample, word vectors for its differentially expressed genes were combined. nHDP (Paisley et al. 2015) was used to identify the latent topics in the set of combined word vectors.

Topic-document associations and topic-word associations generated from nHDP were further utilized to calculate the gene-topic association scores used in this study. Association strength between a certain gene g and a certain topic t was calculated by the total sum of products of: (1) a specific word w's count in g's word vector, (2) t's probability in document d, (3) the word

w's probability in t. We refer to Chen's work (Chen, Paisley, and Lu 2017) for a detailed description of this section.

3.2.2 Similarity measure

Functional similarity among genes was calculated with topic-gene association matrix and transcriptomic profiles respectively. For the topic-gene association matrix, association scores less than one were set to zero. The similarity measure was computed based on Simrank(Jeh and Widom 2002):

$$T_i = c_1(S^T G_i S)(1)$$

$$G_i = c_2(S^T T_{i-1} S)(2)$$

where S was a *g by n* matrix containing the association score between n topics and g genes, G_i was the g by g matrix containing the similarity among genes in the *i*th iteration, T_i was the n by n matrix containing the similarity among topics in the *i*th iteration, and c_1 and c_2 were the hyper-parameters controlling the impact of later iterations. In this study, both c_1 and c_2 were set to 0.8. The equation (1) and (2) were iterated until T and G reached convergence. Note that only the similarity matrix G was used in the next section.

For the transcriptomic profile data, expression values were dichotomized. Gene expressions higher or lower than 95% interval of the distribution was encoded as one, otherwise zero. Cosine similarity was used to compute the similarity among genes, which is:

$$sim_{ij} = \frac{exp_i \cdot exp_j}{\sqrt{||exp_i|| \cdot ||exp_j||}}(3)$$

where exp_i was the vector of expression values of the ith gene across all the experiments, $||exp_i||$ is the L2 norm of that vector.

3.2.3 Computation of similarity matrix

Protein-protein interaction (PPI) networks were used as the base network. The similarity measures computed in the last section were used as the edge weights for these PPI networks. Thus, the topic-based interactome consisted of the topology of a PPI network with edge weights from the topic-gene association matrix; and the expression-based interactome consisted of the topology of a PPI network with edge weights from transcriptomic profile data. For PPI curated in BioGrid for yeast, we only selected interactions supported by at least two studies.

These two interactomes were further combined into one network by treating each interactome as a layer and connecting the same gene across different layers, as demonstrated in Figure 2.

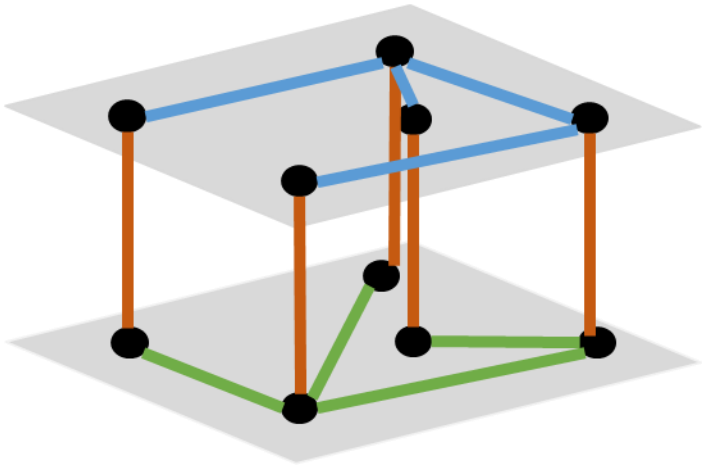


Figure 2 Illustration of the combined interactome, brown edges were artificial edges added to connect these two layers.

3.2.4 Network integration

For all the networks described above, self-loops were removed. Edges with zero similarity and nodes with zero weighted degrees were removed. The combined network is represented by a supra-adjacency matrix (Boccaletti et al. 2014):

$$A = \begin{bmatrix} A_1 & I_N \\ I_N & A_2 \end{bmatrix}$$

where A_i is the adjacency matrix for the ith layer, I_N is an N by N identify matrix, N is the number of nodes in a single layer.

3.3 Isolation Clustering on the Multiplex

The algorithm developed in this study consist of two steps: (1) transform the adjacency matrix into a matrix representing k-step walks visiting probability; (2) enumerate each node to identify clusters with locally optimal isolation.

3.3.1 Network transformation

With the network constructed from previous steps, the Markov transition matrix, M, should be computed next, which is:

$$M_{ij} = A_{ij}/A_{i.} (4)$$

where $A_{i.}$ is the sum of the ith row of A.

From M, we further computed a matrix C, where C_{ij} is the probability that node j is visited if a walk of K steps starts from node i. In this study, K is always set to 10. Since C_{ij} is complementary to the probability that node j never shows up in the path, it can be computed as:

$$C_{ij} = 1 - \mathbf{1}_i^{\boldsymbol{T}} (MI_{-j})^K \mathbf{1} \qquad (5)$$

where $\mathbf{1}_i$ is the vector with only the ith element as one, others zero, I_{-j} is an identity matrix with the jth diagonal value zero, $\mathbf{1}$ is the vector of 1.

As the vectorization of the operation above, the matrix C can be computed by the procedure below:

$C_1 = A \cdot (\mathbf{1} - I)$
  for i in (2: K):
    $diag(C_{i-1}) = \mathbf{0}$
    $C_i = A \cdot C_{i-1}$
    $C = \mathbf{1} - C_i$

Box 1. Algorithm for computing the matrix C

3.3.2 Objective function

Let us denote t_{ij} as the number of times node j is present in the path started from node i, then t_{ij} is sampled from a Bernoulli distribution with probability C_{ij}. Thus, C_{ij} can also be viewed as the expected number of times node j is present if a k-step walk is started from node i, which is:

$$C_{ij} = Pr(t_{ij} = 1) = E(t_{ij}) \qquad (6)$$

We further denote R as a subset of nodes and t_{iR} as the total number of nodes of R present in the walk:

$$t_{iR} = \sum_{j \in R} t_{ij} \qquad (7)$$

We can derive that:

$$E(t_{iR}) = E\left(\sum_{j\in R} b_{ij}\right) = \sum_{j\in R} E\left(b_{ij}\right) = \sum_{j\in R} C_{ij} \qquad (8)$$

We can further generalize the equation by denoting t_{QR} as the total number of nodes in R present in a walk started from a node in Q. A walk is started from node i in R for W_i times. From the law of total expectation, we can derive that:

$$E\left(t_{QR}\right) = \sum_{i\in R}\sum_{j\in Q} W_j C_{ij} \qquad (9)$$

Assuming W_j=1 for every j, we developed two measures to capture the degree of isolation of a subset R. One is retention:

$$retention = \frac{E(t_{RR})}{E(t_{RG})} = \frac{\sum_{i\in R}\sum_{j\in R} C_{ij}}{\sum_{i\in R}\sum_{j\in G} C_{ij}} \qquad (10)$$

where G is the subset for all the nodes within the graph, t_{RR} is the expected number of nodes of R visited in the k-step walks started from each node in R once, t_{RU} is the expected total number of nodes of G visited in the k-step walks started from each node in R once. The higher retention, the more likely walkers started in R will stay in R.

The other is:

$$exclusion = \frac{E(t_{RR})}{E(t_{GR})} = \frac{\sum_{i\in R}\sum_{j\in R} C_{ij}}{\sum_{i\in G}\sum_{j\in R} C_{ij}} \qquad (11)$$

where t_{RU} is the expected total number of nodes of R visited in the k-step walks started from all the nodes in G once. The higher exclusion, the less likely walkers outside R will get in.

Combining these two measures, the objective function, named isolation in this study, is:

$$isolation_{RR} = \frac{E(t_{RR})}{E(t_{RG})+E(t_{GR})} = \frac{\sum_{i\in R}\sum_{j\in R} C_{ij}}{\sum_{i\in R}\sum_{j\in G} C_{ij}+\sum_{i\in G}\sum_{j\in R} C_{ij}} \qquad (12)$$

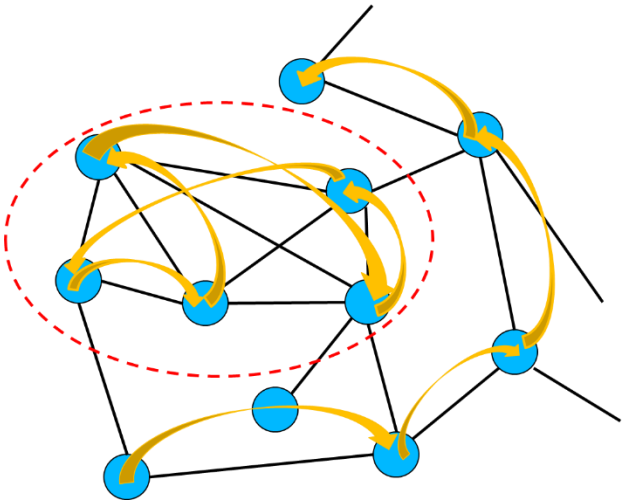


Figure 3 Illustration of the intuition of the objective function. Nodes within the red dotted circle would be a region with high isolation since walkers inside are likely to stay within and walkers outside are unlikely to get in.

3.3.3 Optimization procedures

To identify clusters with maximal isolation, we adopted a greedy approach iterating between two phases. One is expansion. In the expansion phase, isolation is calculated for each individual node outside the cluster:

$$isolation_{iR} = \frac{\sum_{j\in R} c_{ij} + \sum_{j\in R} c_{ji}}{\sum_{j\in G} c_{ij} + \sum_{j\in G} c_{ji}} \quad (13)$$

Note that the top 10 nodes with $isolation_{iR}$ higher than the original cluster are added into the cluster.

The other is shrinking. In this phase, isolation is calculated for each individual node within the cluster. All the nodes with $isolation_{iR}$ lower than the original cluster are removed from the cluster. The algorithm keeps iterating between expansion and shrinking until there are no more qualified nodes for expansion.

3.3.4 Proof of convergence

For expansion, let us denote the set of qualified nodes as X and the resulting cluster as R'. For each node i within X, $isolation_{iR} > isolation_{RR}$. In other words:

$$\frac{\sum_{i\in X}\sum_{j\in R} C_{ij} + \sum_{i\in X}\sum_{j\in R} C_{ji}}{\sum_{i\in X}\sum_{j\in G} C_{ij} + \sum_{i\in G}\sum_{j\in X} C_{ij}} > \frac{E(t_{RR})}{E(t_{RG}) + E(t_{GR})}$$

Thus:

$$\frac{E(t_{RR}) + \sum_{i\in X}\sum_{j\in R} C_{ij} + \sum_{i\in X}\sum_{j\in R} C_{ji}}{E(t_{RG}) + E(t_{GR}) + \sum_{i\in X}\sum_{j\in G} C_{ij} + \sum_{i\in G}\sum_{j\in X} C_{ij}} > \frac{E(t_{RR})}{E(t_{RG}) + E(t_{GR})}$$

On the other hand,

$$isolation_{R'R'} = \frac{E(t_{RR}) + \sum_{i\in X}\sum_{j\in R} C_{ij} + \sum_{i\in X}\sum_{j\in R} C_{ji} + E(t_{XX})}{E(t_{RG}) + E(t_{GR}) + \sum_{i\in X}\sum_{j\in G} C_{ij} + \sum_{i\in G}\sum_{j\in X} C_{ij}}$$

$$> \frac{E(t_{RR}) + \sum_{i\in X}\sum_{j\in R} C_{ij} + \sum_{i\in X}\sum_{j\in R} C_{ji}}{E(t_{RG}) + E(t_{GR}) + \sum_{i\in X}\sum_{j\in G} C_{ij} + \sum_{i\in G}\sum_{j\in X} C_{ij}}$$

Hence $isolation_{R'R'} > isolation_{RR}$ after expansion.

Similarly, an increase of isolation after shrinking can be proved. Thus, our objective function, isolation, is always increasing during iterations, and convergence is guaranteed.

```
function IsolationOptimization(C);
Input: The matrix C
Output: The list of tuples of index
let R be an empty list
let S be a set of indexes of all the nodes in C
Sort S in the descending order of RowSum(C)
while S != Null do
  region = S.pop()
  candidates = expand(C, region)
  while candidates != Null do
    region = region.add(candidates)
    region = shrink(C, region)
    candidates = expand(C,region)
  end while
  R.append(region)
  S = SetDifference(S, region)
end while
return (R)
end
```

Box 2. Clustering Algorithm

3.3.5 Merging overlapped clusters

Highly overlapping clusters are likely to exist for this method. Additionally, for integrated networks, duplicate gene IDs in the same cluster need to be removed. Therefore, overlapping among clusters were evaluated by Jaccard coefficients:

$$overlap(C_i, C_j) = \frac{|C_i \cap C_j|}{|C_i \cup C_j|} \quad (14)$$

where C_i was the *i*th cluster, $|C_i|$ was the number of genes in C_i. $C_i \cap C_j$ was the intersection of C_i and C_j, and $C_i \cap C_j$ is the union of C_i and C_j. A graph with clusters as nodes was constructed. There is an edge between cluster i and j if $overlap(C_i, C_j)$>0.8. Sets of highly overlapping clusters is identified as a connected component of the graph, union and intersection of all the clusters within

a set is computed and added into the set. For each sets, the cluster with the maximal isolation will remain while all the others will be removed.

3.4 Experimental Results

We first identified differentially expressed genes from RNA expression data. Then we calculated topic-gene association from Pubmed titles and abstracts. These two types of data were used to calculate functional similarity among genes used as edge weights for protein interaction networks respectively. The two weighted PPI networks were further connected with the multiplex approach. Finally, we developed a clustering algorithm to identify functional modules with locally maximum isolation from the two-layer protein interaction network. Our clustering algorithm on multiplex was compared with itself on single layer network. Then it was compared against other methods in terms of protein coverage and accuracy.

3.4.1 Descriptive Statistics

BioGrid curation of PPI for saccharomyces cerevisiae contained 32353 interactions among 4518 gene products. The transcriptomic profile of yeast perturbation experiments contained expression values of 5980 genes under 1525 knockout conditions. The topic-gene association matrix contained 216 topics and 5348 genes.

After network construction, the yeast interactome based on topic modeling had 4187 genes and 30989 interactions; the yeast interactome based on transcriptomic profiles contained 4179

genes and 30887 interactions; the interactome based on the combination of the transcriptomic interactome and the topic-gene associations contained 8302 genes and 65793 interactions.

The protein interaction network contained 10945 nodes and 56471 edges. The transcriptomic profile of breast cancer patients in TCGA contained 1218 samples and 20252 genes. The topic-gene association matrix contained 209 topics and 16712 genes.

After network construction, the human interactome based on transcriptomic profiles contained 10029 genes and 49909 edges. The human interactome based on topic modeling contained 10368 genes and 48806 edges. The combined interactome contained 19266 genes and 212292 edges.

3.4.2 Single-layer versus multiplex

We first checked if a method using both knowledge and expression data can obtain better performance than those using only protein interaction networks or combined with topic association. As shown in Figure 4, Figure 5, Figure 6, and Figure 7, after being weighted by topic association, PPV was improved across all evaluations. It was further improved when information about topic association and coexpression were combined with the multiplex approach. This suggests that our clustering algorithm tends to identify clusters with fewer false positives.

However, sensitivity has remained mostly unchanged or slightly worse. In particular, sensitivity has decreased when evaluated against CYC2008 in the species of yeast. This suggested that, while false positives were reduced, true functional relations may be more likely to be ignored, resulting in lower sensitivity.

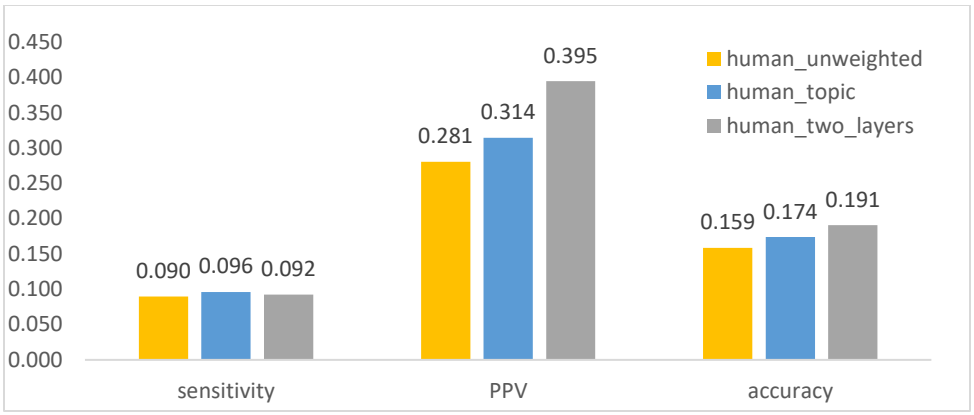


Figure 4 Performance of isolation clustering on three different human interactomes, using Gene Ontology as gold standard.

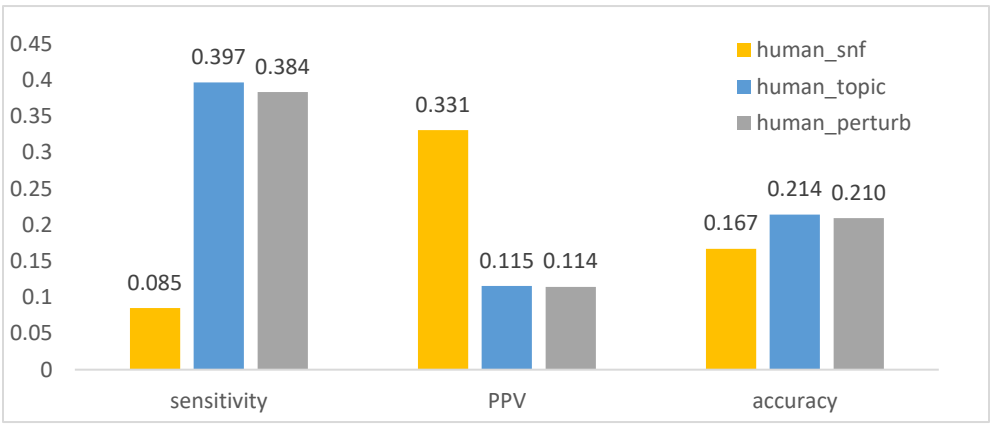


Figure 5 Performance of isolation clustering on three different human interactomes, using CORUM as the gold standard.

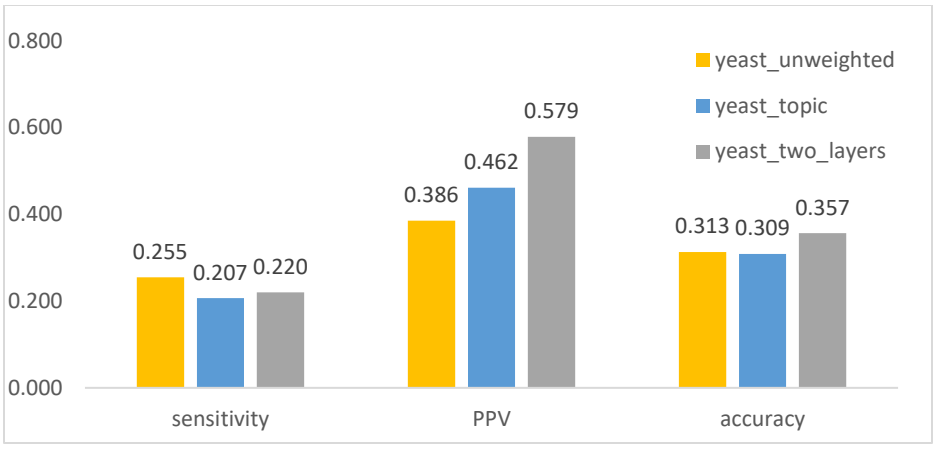


Figure 6 Performance of isolation clustering on three different yeast interactomes, using Gene Ontology as the gold standard.

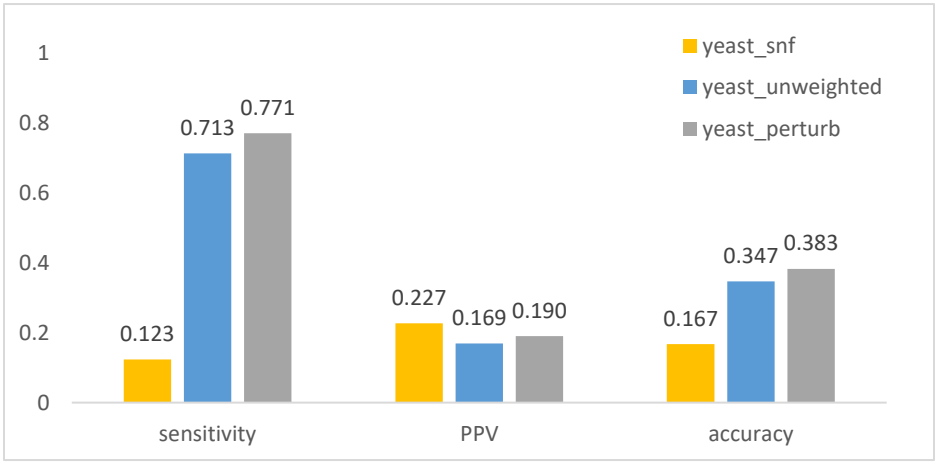


Figure 7 Performance of isolation clustering on three different human interactomes, using CYC2008 as the gold standard.

3.4.3 Comparison with other methods

We then compared our clustering method with some other well-known methods in terms of solution sizes, protein coverage, and accuracy. All the clusters with less than 3 proteins or larger than 200 proteins were removed. As shown in Table 1 and Table 2, the distribution of cluster size

for our method (isolation) is more skewed towards size 3-10. For the species of yeast, CYC2008 has over 83.3% of proteins with size less than 10, while the percentage of MCL, Infomap, Isolation was 73.8%, 64.4%, and 92.3% respectively. For the species of human, 89.5% of proteins complexes in CORUM contain less than or equal to 10 gene products, while 88.9% of functional modules generated by isolation clustering has such small size. Assuming that this distribution of CORUM and CYC2008 represents the true distribution of protein complexes, it indicated that the modular structure characterized by Isolation clustering was similar to that within real cells.

Table 1 The distribution of cluster size by different methods on yeast interactomes. The rightmost column is the gold standard used in this study.

Size	*MCL*	*Walktrap*	*Infomap*	*MCODE*	*ClusterOne*	*Isolation*	*CYC2008*
3 – 10	342	158	275	135	426	995	198
11 – 50	107	44	140	21	86	82	36
51 - 100	13	5	7	11	10	1	2
100-200	0	5	5	2	4	0	0
>200	1	0	0	0	0	0	0

Table 2 The distribution of cluster size by different methods on human interactomes. The rightmost column is the gold standard used in this study.

Size	*MCL*	*Walktrap*	*Infomap*	*MCODE*	*ClusterOne*	*Isolation*	*CORUM*
3 – 10	1008	323	506	319	1322	2131	1562
11 – 50	353	83	241	47	108	260	176
51 - 100	15	13	16	8	0	4	5
100-200	0	3	3	1	0	1	2
>200	0	0	0	3	0	0	0

3.4.3.1 Protein coverage

As shown in Figure 8, clusters generated by ClusterOne, MCODE, and Walktrap can only cover around half of the interactome. MCL, Infomap, and Isolation had covered over 90% of the interactome. Significantly higher coverages indicated that clustering methods based on random walks (i.e., MCL, Infomap, and Isolation) might provide more information about novel proteins so as to generate more biological insights. In the next section, only MCL, Infomap, and Isolation were compared against each other in terms of accuracy.

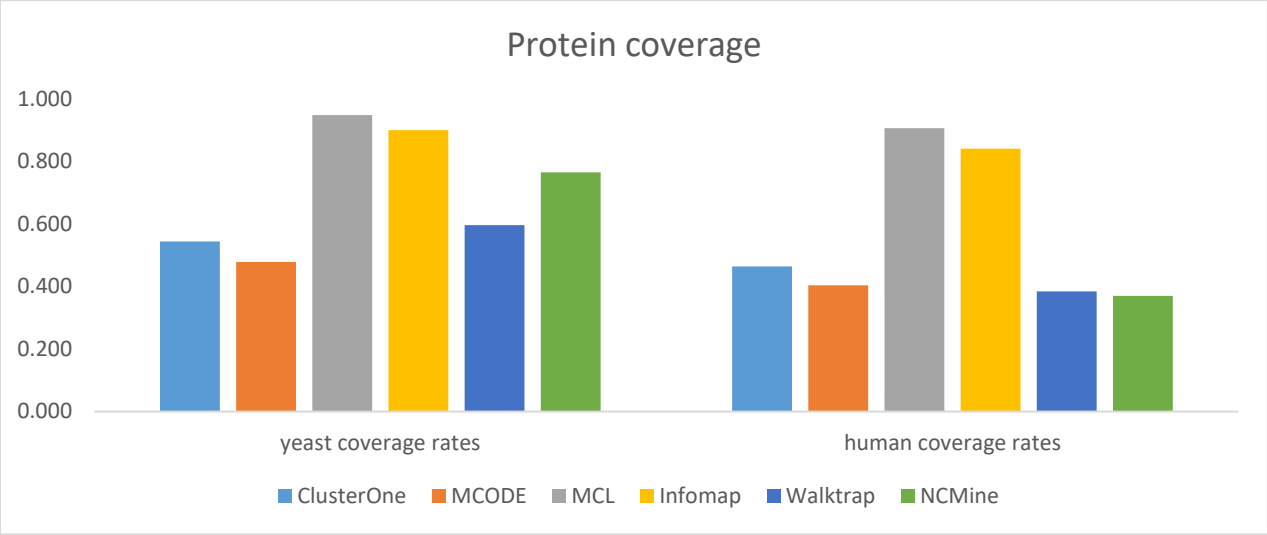


Figure 8 In clustering for both yeast and human interactomes, clustering based on random walks has covered most proteins, while density-based clustering discarded around half the proteins.

3.4.3.2 Geometric accuracy

As shown in Figure 9 and Figure 10, Isolation has outperformed MCL and Infomap in yeast interactome in terms of geometric accuracy. The accuracy of our method is slightly higher than other methods. However, in the case of human interactomes, these three methods yielded very similar performance.

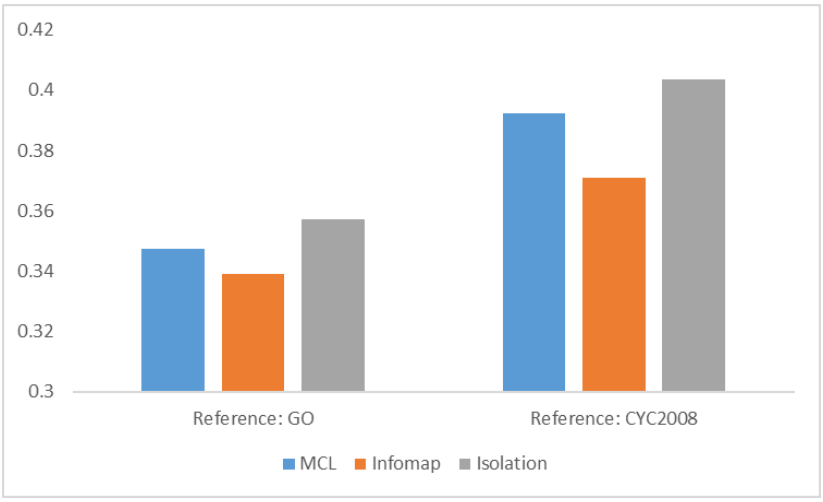


Figure 9 Comparison of geometric accuracy of MCL, Infomap, and Isolation on yeast interactomes

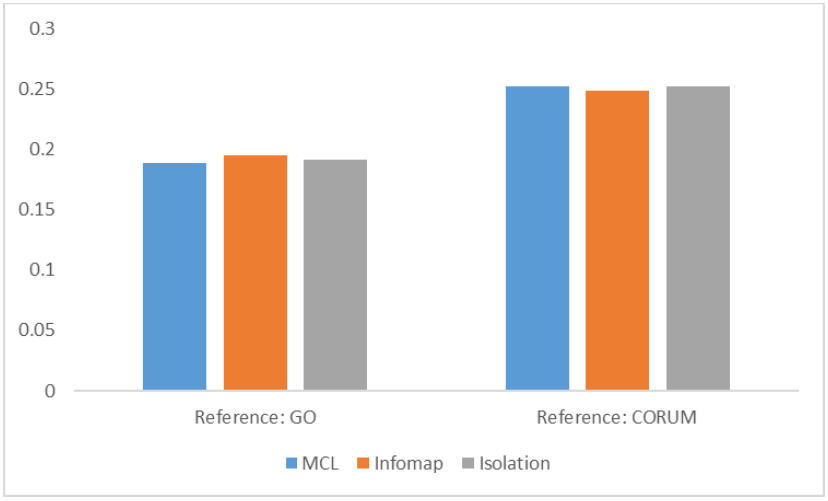


Figure 10 Comparison of geometric accuracy of MCL, Infomap, and Isolation on human interactomes.

3.4.4 Examples of clusters

Our clustering results have found many overlaps with known complexes. Two of them were perfect matches (Figure 11). For some genes misclassified to a complex, we are able to identify close functional relations from literature. For example, our methods had grouped PINX1

with TRF-Rap1 complex I (Figure 12). Although PINX1 is not part of the complex, it is well studied that PINX1 can mediate TRF1 (or TERF1) and TERT accumulation in the nucleus and enhances TERF1 binding to telomeres(X. Z. Zhou and Lu 2001; Yonekawa, Yang, and Counter 2012), thus affecting the function of the complex.

Furthermore, "misclassified" genes without direct evidence may be more interesting since they could provide new insights for current knowledge. For example, C18orf21 was grouped with Rnase/Mrp complex by our method (Figure 13). Several studies have found genetic associations between variants in C18orf21 and human phenotypes. Besides the high-throughput data (BioPlex (Huttlin et al. 2017)) used in this study, no further experiments have been conducted to investigate the functions of C18orf21. Our results suggested that C18orf21 could function by regulating Rnase/Mrp complex. Another example was shown in Figure 14, where PNMA6A, DRAP1, PTCD3, AURKAIP1, and DDX55 were grouped with the 28S ribosomal subunit. Through literature we found that these misclassified genes, except PNMA6A, have a significant impact on mitochondrial ribosome though detailed mechanisms are not clear (Koc et al. 2013; S. M. K. Davies et al. 2009; Schmid and Linder 1992).

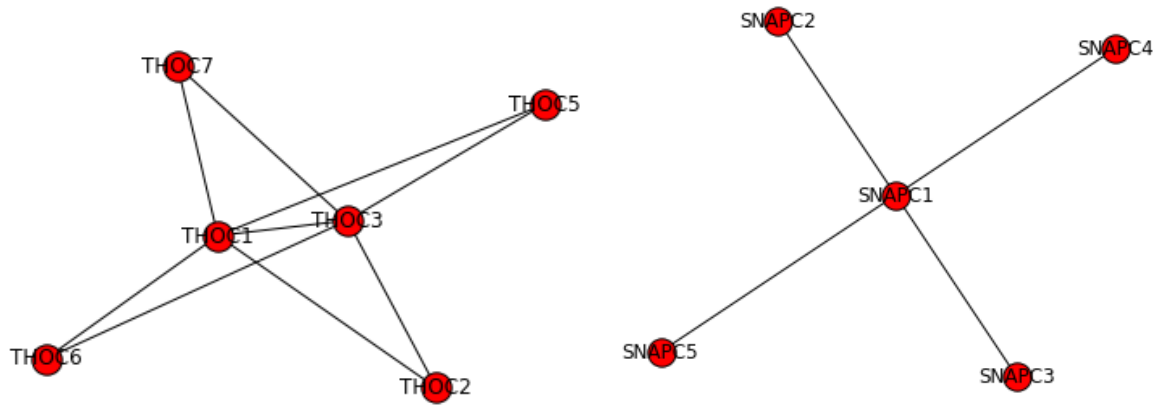


Figure 11 The two predicted complexes perfectly matched to CORUM complexes. On the left is matched to hTREX84 complex. On the right is matched to SNAPc complex.

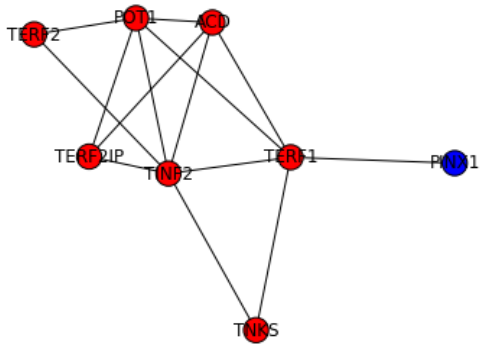


Figure 12 Predicted complex matched to telomere-associated protein complex and TRF-Rap1 complex I, 2MD. Blue nodes were genes predicted but absent in the gold standard.

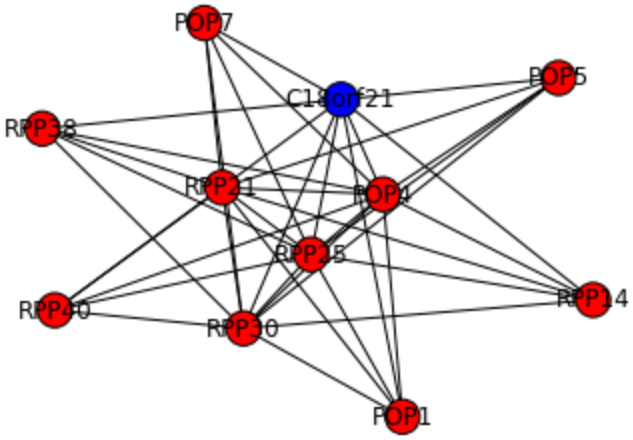


Figure 13 Predicted complex matched to Rnase/Mrp complex. Blue nodes were genes predicted but absent in the gold standard.

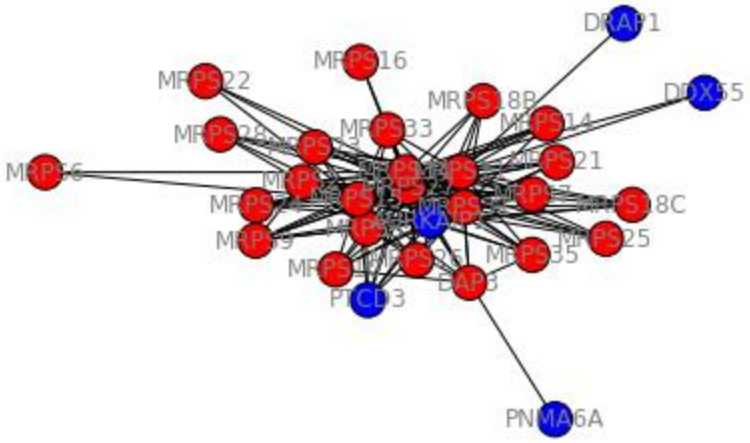


Figure 14 Predicted complex matched to 39S ribosomal subunit, mitochondrial. Blue nodes were genes predicted but absent in the gold standard.

3.5 Contribution and Limitations

As illustrated in the results, isolation clustering tends to identify isolated regions supported by both layers in the multiplex. Such tendency reduces false positives while inducing more false negatives. As shown in the results (Section 3.4), our new clustering algorithm, Isolation, has achieved better, or at least comparable, performance with other well-known clustering algorithms based on random walk. Particularly, subnetworks with locally maximal isolation exhibited higher confidence of being true positive when compared with MCL and Infomap. When compared with clustering algorithms such as ClusterOne, our algorithm has covered over 90% of proteins, while density-based clustering can only cover around 50%. This leads to higher PPV from density-based clustering algorithms.

In addition, end-users usually prefer PPV to sensitivity. In other words, biomedical researchers may care more about whether the predicted modules reveal functional relationships among genes rather than whether all the closely related genes are included in a module. Thus it is natural for users to focus on positive predictive value or precision rather than composite scores used by most methodological studies. From this perspective, our integrative approach provides practical values.

Selected examples in the result section has shown that false positive genes could be functionally related in a way other that protein complexes. This illustrated one fundamental limitation for functional module identification and its evaluation. Biological experiments should be conducted to further verify the predicted modules.

This project also demonstrated that topic modeling of biomedical literature is an effective complementary source of information. Knowledge validated and curated in the form of literature are generally more reliable than high-throughput data. By integrating knowledge into the functional module identification process, false positives caused by data quality issues can be reduced. Thus, functional modules are identified with higher confidence.

4.0 Identifying Coexpression Patterns with Boolean Matrix Factorization

This chapter describes the motivation, assumptions, and modeling of our improved Boolean matrix factorization. It also includes experimental results that applied BMF to bulk RNA-seq, scRNA-seq, and spatial transcriptomic profiles.

4.1 Need for Biclustering Algorithm in Gene Expression Analysis

Grouping genes or samples according to their shared expression patterns was an important task. On the genes' side, similar expression profiles across conditions indicated coregulation of gene expression, which can be used to infer upstream pathway activities (Tai et al. 2018) and the regulatory relationship between transcription regulators and target genes (Paul et al. 2015). On the samples' side, clusters of samples help reveal the heterogeneity in the disease population. For example, (Sørlie et al. 2001) has identified clinically relevant breast cancer subtypes from expression profiles alone. This task has become ever more prevalent since the emergence of new technologies such as single cell RNAseq (A. P. Patel et al. 2014) and spatial transcriptomics (Berglund et al. 2018), which enable us to interrogate tumor heterogeneity with finer granularity.

However, clustering directly on only one side (either on the sample side or the gene side) yields limited performance. That is because computational distance between objects is contaminated by the noise in irrelevant features. As illustrated in Fig. 1a, given that coregulation mechanism is prevalent in expression profile, the similarity between samples/conditions should only depend on a small group of genes with common upstream factors. Since at most several

hundred genes can be coregulated, the distinctive expression profiles for a cluster of samples is no more than several hundred genes. This means all the other genes acted as random noise for the identification of this one cluster. Most studies handled such issues with feature selection. This approach requires prior knowledge or external information, which potentially hinders the identification of novel and interesting features. Moreover, the issue of contamination cannot be resolved even with perfect feature selection. A selected subset of genes can be informative features for one cluster while being random noise to another.

Thus, biclustering should be a natural choice when it comes to high dimensional gene expressions analysis. By finding clusters and their corresponding features simultaneously, biclustering directly resolved the contamination issues above. Since first proposed by (Cheng and Church 2000), various biclustering algorithms have been developed and applied to gene expression data (Xie et al. 2019). However, most biclustering algorithms (Tanay, Sharan, and Shamir 2002; Bergmann, Ihmels, and Barkai 2003; G. Li et al. 2009) are heuristic-based with local iterative search. Thus these algorithms are mostly used to identify subtle gene-sample substructure, rather than tasks requiring systematic analysis of sample heterogeneity such as subtype classification or cell type deconvolution.

Another popular approach is to factorize the gene-sample matrix (Stein-O'Brien et al. 2018). As an example, nonnegative matrix factorization (NMF) has been widely used in gene expression analysis in the past decade (Brunet et al. 2004). By performing dimension reduction on both columns and rows, the matrix factorization approach provides more information about global heterogeneity. Recently, a comprehensive evaluation (Saelens, Cannoodt, and Saeys 2018) showed that matrix factorization outperforms clustering and biclustering in terms of identifying coexpression modules. However, this approach does not explicitly provide clustering structures.

Even for methods that provide bicluster structure (Hochreiter et al. 2010), the assumption of linear combination may not be sufficient to capture the coregulation patterns in transcriptomic data.

4.2 Biclustering Formulation with Boolean Matrix Factorization

Besides the differential expression matrix X. We need to define two latent variables, U and Z. U_{nl} indicates that pathway *l* is perturbed in sample *n*. Z_{ml} indicates that pathway *l* regulates gene *m*. Note that the term "pathway" used in this chapter denotes any biological mechanism that may play a regulatory role in gene expression, such as transcription factors, signaling pathways, cell population, and disease subtypes. Therefore, the matrix U represents the abnormality status of transcriptional regulation mechanisms.

First, we need to impose two assumptions about gene expression regulation: (1) differential expression takes place when its regulating pathway is perturbed; (2) if a gene is regulated by multiple pathways, perturbation of one pathway is sufficient to cause differential expression.

With these assumptions, we state that X is the outcome of the AND-OR product rule:

$$X_{nm} = \vee_{l \leq L} (U_{nl} \wedge Z_{ml}) \quad (1)$$

where $\vee$ is the OR operator and $\wedge$ is the AND operator. In this study, however, a different formulation was adopted. We assume that each element of X, X_{nm}, is sampled from a different Bernoulli distribution. Similarly, every element in the latent factors is sampled from different Bernoulli distributions. The generative process of X can be described as follows:

$$U_{nl} \sim Bernoulli(\mu_{nl})$$

$$Z_{ml} \sim Bernoulli(\zeta_{ml})$$

where μ is a N×L matrix with values in [0, 1], ζ is a M×L matrix with values in [0, 1]. Clearly, by forcing μ and ζ to be binary, our formulation will be identical to previous Bayesian approaches. Thus, our formulation is a generalized version of the Bernoulli model. With this approach, our goal for Boolean matrix factorization is to estimate the parameters μ and ζ instead of their samples U and Z.

Since the AND-OR product is a logical Boolean operation, X_{nm} can be seen as the output of a function of U and Z. Hence, we can derive how $P(X_{nm} = 1|\mu_{nl}, \zeta_{ml})$ can be computed with $P(U_{nl} = 1|\mu_{nl})$ and $P(Z_{nl} = 1|\zeta_{ml})$ given their logical relationship.

$$P_{nm} = P(X_{nm} = 1) = 1 - P(X_{nm} = 0) = 1 - \prod_{l \le L}[1 - P(U_{nl} = 1, Z_{nl} = 1)]$$

where the conditional parameters are ignored for convenience.

Assuming U and Z are independent, in other words, whether a pathway is perturbed is not related to which gene it regulates, we have:

$$P(U_{nl} = 1, Z_{nl} = 1|\mu_{nl}, \zeta_{ml}) = P(U_{nl} = 1|\mu_{nl})P(Z_{nl} = 1|\zeta_{ml}) = \mu_{nl} * \zeta_{ml}$$

Therefore, $P(X_{nm} = 1|\mu_{nl}, \zeta_{ml})$ can be expressed as:

$$P_{nm} = 1 - \prod\nolimits_{l \le L} (1 - \mu_{nl} * \zeta_{ml}) \quad (2)$$

4.3 Model Inference

And the model likelihood, the objective function we maximize regarding μ and ζ, can be computed as:

$$LL(\mu, \zeta; X) = \sum_{n \leq N, m \leq M} [X_{nm} log P_{nm} + (1 - X_{nm}) \log (1 - P_{nm})] \quad (3)$$

Conventional gradient descent is not applicable because μ and ζ need to be within the interval [0, 1]. Thus, μ and ζ are reparameterized as σ(A) and σ(B) elementwise:

$$\mu_{nl} = 1/(1 + e^{-A_{nl}})$$
$$\zeta_{nl} = 1/(1 + e^{-B_{nl}})$$

With reparameterization, it becomes a problem of unconstrained nonlinear programming. A simple gradient ascent algorithm is sufficient to jointly optimize the estimators of A and B. The partial likelihood gradients regarding A are:

$$\frac{\partial LL}{\partial A_{il}} = \sum_{m \leq M} \left[\frac{\mu_{nl} \zeta_{ml} (1 - \mu_{nl}) \left(1 - \frac{X_{nm}}{P_{nm}}\right)}{1 - \mu_{nl} \zeta_{ml}} \right] \quad (4)$$

Note that A and B are symmetric, thus the partial gradient of B can be computed similarly as A. In the subsequent description, equations related to B and Z were also neglected due to this symmetry.

We further introduce a parameter, ϵ, to explicitly model the probability that elements in X is contaminated by noise (flipped from 1 to 0 or vice versa). In this scenario, the observed data, X*, is generated as:

$$C_{nm} \sim Bernoulli(\epsilon)$$

$$X^*_{nm} = ABS(X_{nm} - C_{nm})$$

where C_{nm} is a N×M binary matrix with every element as a i.i.d sample from a Bernoulli distribution parameterized by a scalar ϵ. ABS is the function of taking absolute values. To reflect the addition of noise in the model, we need to add one step in the generative process:

$$P^* = (1 - \epsilon)P + \epsilon(1 - P)$$

The noisy observation, X^*, is sampled from P* instead of P:

$$X^*_{nm} \sim Bernoulli(P^*)$$

Thus, the model likelihood becomes:

$$LL(\mu, \zeta; X^*) = \sum\nolimits_{n \le N, m \le M} [X^*_{nm} log P^*_{nm} + (1 - X^*_{nm}) \log (1 - P^*_{nm})] \quad 5$$

To optimize the likelihood function regarding μ, ζ and ϵ, we applied the expectation maximization algorithm. In M step, μ and ζ are estimated with the gradient based method. The difference is the presence of a fixed ϵ, leading to a different equation for likelihood gradients:

$$\frac{\partial LL}{\partial A_{il}} = \sum_{m \leq M} \frac{\mu_{nl}\zeta_{ml}(1-\mu_{nl})(1-P_{nm})(1-2\epsilon)\left(P^*_{nm} - X'_{ij}\right)}{(1-P^*_{nm})P^*_{nm}(1-\mu_{nl}\zeta_{ml})} \tag{6}$$

In E step, based on the modified generative process described in the beginning of this section, the expected value of ϵ is equivalent to the difference between the noisy observation, X*, and the reconstructed data without noise, $\hat{X}$:

$$\epsilon = \frac{|C|}{NM} = \frac{\left|\hat{X} - X^*\right|}{NM} \tag{7}$$

The estimate above is only approximate. The exact estimate should be the average difference between X* and P*. However, the exact estimate requires a much more stringent convergence criterion in the M step. During synthetic experiments, the performance of the approximate estimate is not significantly different from the exact one. Thus, the approximate estimate of ϵ was adopted.

We further impose prior distribution on μ and ζ:

$$\mu_{ml} \sim Beta(\alpha, \beta)$$

$$\zeta_{ml} \sim Beta(\alpha, \beta)$$

In practice, μ and ζ can comply with different Beta distributions. For the convenience of notation, we simply assume they have a common prior distribution. Thus μ and ζ are estimated based on Maximum a Posteriori (MAP) estimator. The posterior probability function of μ and ζ is:

$$\Pr(X|\mu,\zeta,\epsilon) = LL + (\alpha-1)\left[\sum_{m\le M,l\le L} log\mu_{ml} + \sum_{n\le N,l\le L} log\zeta_{nl}\right] + (\beta-1)[\sum_{m\le M,l\le L}\log(1-\mu_{ml}) + \sum_{n\le N,l\le L}\log(1-\zeta_{nl})]$$

where LL is described in Section 2.3. We applied gradient ascent to the objective function. The partial gradient for $\Pr(X|\mu,\zeta,\epsilon)$ is:

$$\partial \Pr(X|\mu,\zeta,\epsilon)/\partial A_{nl} = \partial LL/\partial A_{nl} + (\alpha-1)(1-\mu_{nl}) - (\beta-1)\mu_{nl}$$

Clearly, when α and β are set to 1, the MAP estimator will be identical to the maximum likelihood estimator. When α and β are larger than 1, latent factors will be skewed towards 0.5; when α and β are less than 1, latent factors are pushed towards 0 or 1. Alternatively, the entropy of μ and ζ can be used as penalty, and the objective becomes minimizing KL divergence. However, users can push the sparsity of latent factors by making α and β asymmetric, which is not available with entropy.

Our approach to matrix completion is simple. During training, parameters are only updated based on the gradients from the observed data points. When convergence is reached, missing data are imputed by the reconstructed data without noise.

4.4 Experimental Results

Our algorithm was compared with the message passing approach (Siamak Ravanbakhsh, Póczos, and Greiner 2016) and the full Bayesian approach (Tammo Rukat et al. 2017), referred to as LoM/OrM below. The Bernoulli prior for the two algorithms were estimated using the empirical Bayes approach described in (Tammo Rukat et al. 2017). During synthetic experiments, we evaluated the three algorithms on two tasks: noisy matrix factorization and noisy matrix completion. In real data experiment, the three algorithms were compared by the subtype classification accuracy on RNAseq datasets from TCGA (Tomczak, Czerwińska, and Wiznerowicz 2015). Finally, we demonstrated our algorithm's real-world application to three datasets generated by bulk RNAseq, scRNAseq, and in situ hybridization, respectively.

4.4.1 Simulation experiment

The observed matrices with noise, X*, were synthesized based on the same sampling scheme as our probabilistic problem formulation, except that each scalar value in latent factors was sampled from a uniform distribution on interval [P-0.3, P+0.3]. P was determined by the preset matrix density P(X=1):

$$\mathrm{P}(X = 1) = 1 - (1 - P^2)^L \tag{8}$$

4.4.1.1 The task of matrix factorization

We evaluated the three algorithms on four different noise levels (flip probability): 0.0, 0.1, 0.2, 0.3. The sampling scheme was repeated 10 times for each noise level. The performance was measured by the reconstruction error rates, which is comparing the reconstructed matrix with the synthesized matrix without noise:

$$err = \left|\hat{X} - X\right|/(NM)$$

As shown in Figure 15, although EM algorithm is likely to reach a local optimum, the performance of BEM is more stable across different noise levels compared with other probabilistic approaches. BEM has achieved zero error in 9 out of 10 synthetic datasets with lower noise levels (flip probability ≤ 0.3), while the other two can only perfectly reconstruct the noiseless matrix in 6 to 9 synthetic samples. Statistical analysis showed that BEM was significantly better than LoM when there is no noise and outperformed MP when the noise level reached 0.3. However, when the flip probability is above 0.3, LoM performed slightly better than message passing and BEM. Such comparison results remained the same when matrix density was 0.3 (Appendix Figure 1) and 0.7 (Appendix Figure 2).

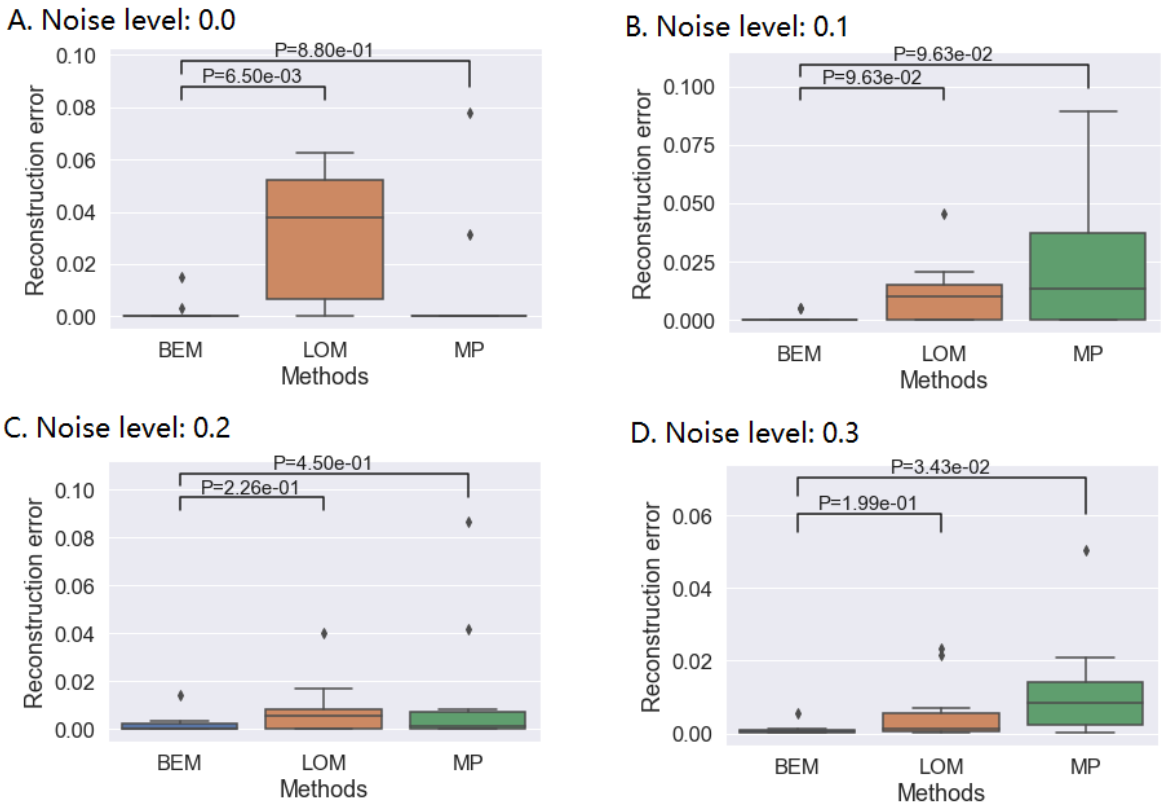


Figure 15 Reconstruction error (8% max) of synthetic data when Bernoulli priors varied. Synthetic matrices were 1000×1000 with rank 5. BEM (Left) is the algorithm proposed in this study; MP (right) is short for message passing; LOM (middle) is the Logical factorization machine.

When tested against various matrix sizes and Boolean ranks, the degree of freedom versus sample size, $(N + M)L/(NM)$, is important for the relative performance of BEM. As shown in Appendix Figure 3 and Appendix Figure 4, when Boolean rank was increased from 5 to 10, LoM achieved the best performance across different noise levels. However, when matrix sizes increased from 1000 to 2500, LoM's performance has a much greater variance than message passing and BEM.

4.4.1.2 The task of matrix completion

We evaluated the three methods with various observed fraction (i.e. 30%, 50%, 70%, 95%). The matrices were generated with the same sampling scheme as above. The noise was set at 20%. The performance was measured by the fraction of correctly inferred values. As shown in Figure 16, although BEM only significantly outperformed LoM when the observed fraction is 50%, its performance is favorable across different settings except when observed fraction is 30%.

In summary, BEM outperformed other probabilistic Boolean matrix factorization methods when the noise level is less than or equal to 30%. Since most gene expression datasets satisfy these conditions, BEM is more suitable to transcriptomic data than other Boolean matrix factorization methods.

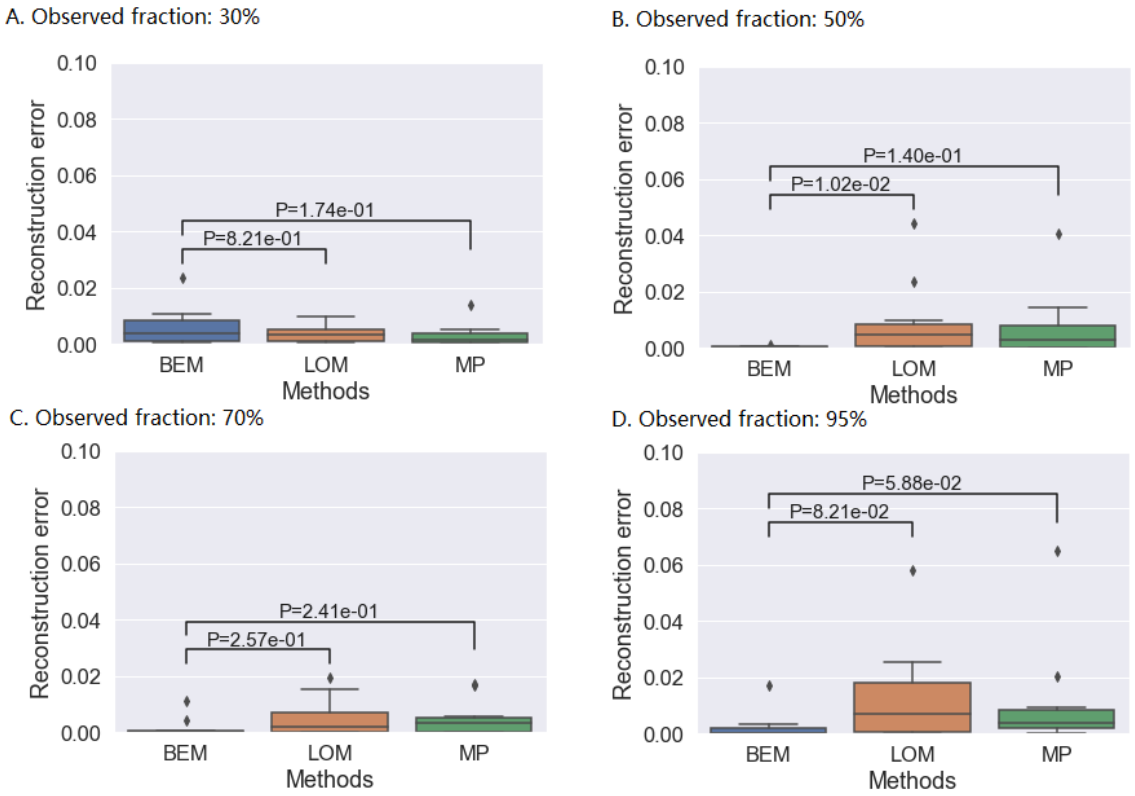


Figure 16 Reconstruction error on synthetic data with varying observed fractions. BEM (left) is the algorithm proposed in this study; MP (right) is short for message passing; LoM (middle) is the Bayesian sampling approach.

4.4.2 Real data experiments

4.4.2.1 Classification of breast cancer subtypes

We downloaded transcriptomic profiles of breast cancer patients from TCGA (Tomczak, Czerwińska, and Wiznerowicz 2015). The data was dichotomized to encode differential expression. The criteria for differential expression are: (1) absolute log fold change > 0.23; (2) adjusted p value ≤ 0.05. Differential expression was encoded as 1, otherwise 0.

From this binary matrix, 15 factors were extracted with our algorithm and others for comparison. The number of factors was determined by Akaike information criteria (AIC). To

examine the effectiveness of BEM, we investigated the proportion of patient subtypes in the factors on the samples' side.

To compare the performance of these factorization methods, we used factors about the samples (or meta-samples) as features for tumor subtype classification. It was conducted with Multinomial logistic regression. Logistic regression come from a python package named "scikit-learn". We randomly sampled 80% of the expression data to train the logistic regression. The rest were used as the test set to evaluate classification accuracy. This procedure was repeated 20 times to evaluate whether the performance difference was stable. As shown in Figure 17, our algorithm had achieved the highest classification performance among algorithms for Boolean matrix factorization.

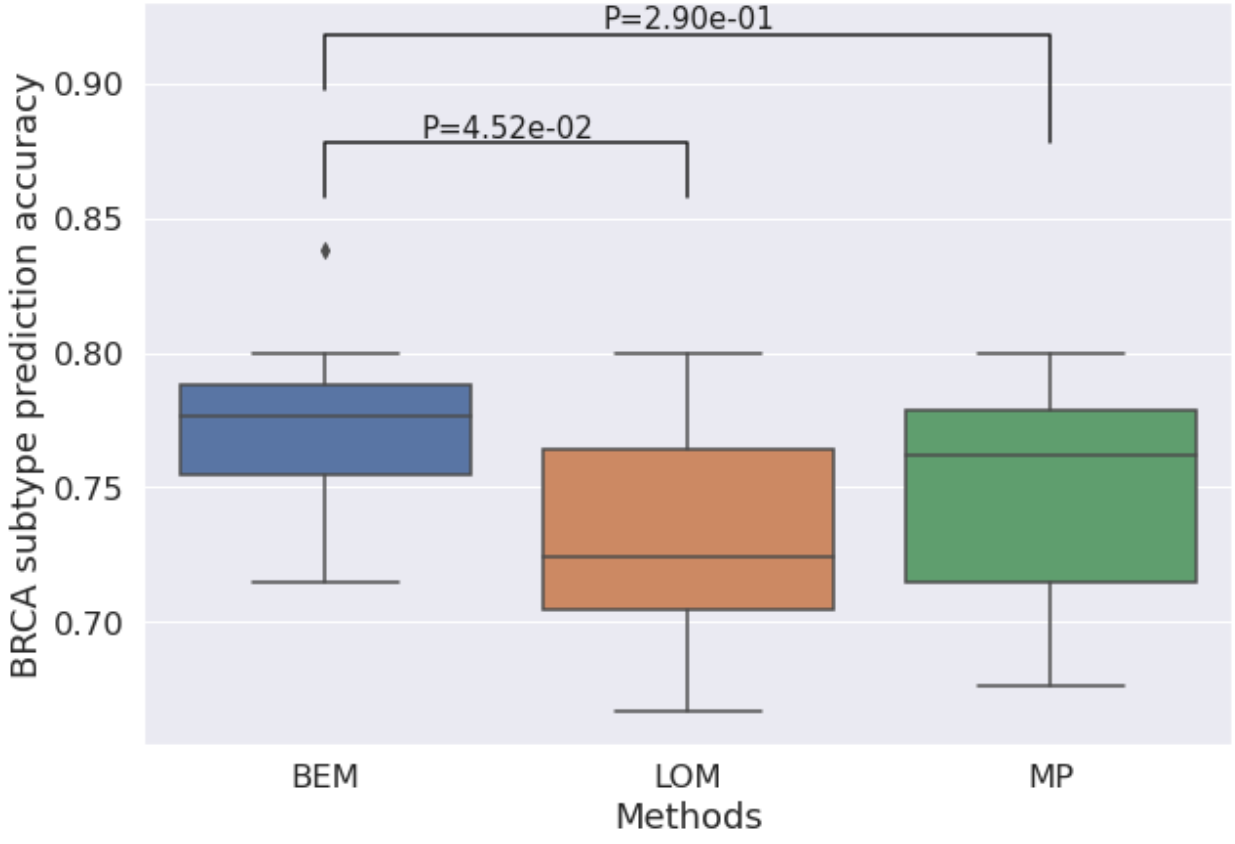


Figure 17 Breast cancer subtype classification accuracy

We further investigated classification accuracy with other Boolean matrix factorization methods in each tumor subtype. As shown in Figure 17, all the Boolean matrix factorization methods achieved high accuracy in the subtype of LumA and Basal. It indicated the genes expression data and the subsequent differential expression analysis had provided abundant discriminative information about the two subtypes. However, LoM and Message Passing were less effective in discerning the Her2-enriched subtype, which has the smallest sample size. This result showed that by getting rid of assumptions about factors' sizes, BEM was more likely to capture subtle patterns that have greater variance on factor sizes.

Table 3 Accuracy (%) with each subtype with 15 factors

Subtype	Normal	LumA	LumB	Her2	Basal
# of Samples	119	434	194	67	143
LoM	12.5	**87.9**	52.8	53.4	90.8
MP	**37.5**	83.5	**66.1**	62.1	89.7
BEM	0.0	86.1	59.8	**72.4**	**94.9**

4.4.2.2 Cell type deconvolution from single cell RNA-seq

The single cell RNAseq data about melanoma patients was collected from Gene Expression Omnibus (GSE120575). This dataset contained 55737 genes on 16291 cells across 48 samples. Patients were administered with CTLA4 therapy, PD1 therapy, or both. 19 out of 48 samples were measured before immunotherapy. The rest are measured afterward. At least 163 cells were measured within a patient sample. The expression values were encoded as 1 if the gene had nonzero expression values, otherwise 0. Genes that expressed in less than 1% of the cells or over 99% of the cells were removed. Only 10474 genes remained. We chose 10 factors based on the Akaike information criteria (AIC), which is close to the choice of 11 clusters in the original study.

We constructed the gold standard for major cell types from the marker gene sets provided in the original study (Sade-Feldman et al. 2018). In addition, due to high overlap in the gold standard, CD4 T cells and CD8 T cells were merged into T cells (87.8% of CD8 T cells were also CD4 T cells); cDCs dendritic cells, pDCs, macrophage, neutrophils, and myeloid were merged as myeloid cells. (Over 50% of each of these cell types were also classified as myeloid cells). 3764 cells that could not be classified by the gold standard marker genes were discarded. Cell types with small cell counts and little overlap with each other were simply denoted as "Others" in Figure 18.

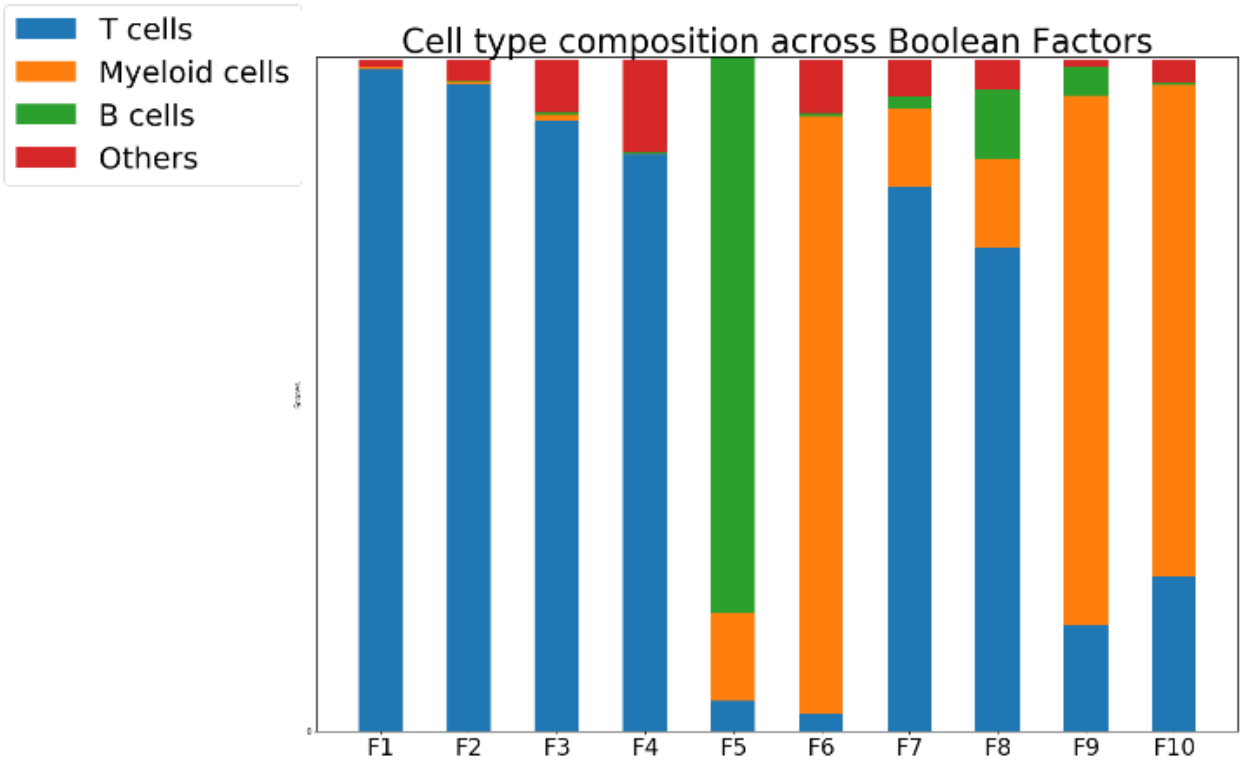


Figure 18 Each column shows the proportions of each cell type in one factor.

The cell-side factors were dichotomized with 0.5 as the cutoff. After dichotomization, if the factor value of the ith cell in the jth factor was 1, then the ith cluster contained the ith cell. Clearly, the clusters were not mutually exclusive. As shown in Figure 18, the first four factors

corresponded to T cells exclusively (from 98.7% to 85.9%). 94.2% of T cells belonged to at least one of these four factors. The fifth factor corresponded to B cells mostly (82.9%). And 97.7% of B cells belonged to this factor. The sixth factor corresponded to myeloid cells mostly (89.1%). Also, 77.6% of myeloid cells belong to this factor. The other four factors were mostly a mixture of T cells and myeloid cells. They might capture cellular functions across cell types. It showed our algorithm was able to identify high levels of expression patterns accurately.

We also aggregated the factors on the cell side into sample level features by taking an average of all the factor values in cells belonging to the same sample. These aggregated values for 19 samples before therapy were used as features in logistic regression to predict responsiveness of patients. The target variable was binary, either responsive or nonresponsive. Accuracy was evaluated with leave-one-out cross validation. As shown in Table 4, using the 10 features from our algorithm was significantly better than the original cell type information. Our algorithm had probably extracted information related to therapy responsiveness beyond merely cell types. Since the prior was set to encourage extreme values, sample features aggregated from binarized Boolean factors has achieved the best performance (78.9%).

Table 4 Prediction accuracy of immunotherapy responsiveness

Features	Accuracy (%)
Gold standard cell type	42.1
Aggregated Boolean factors (continuous)	63.2
Aggregated Boolean factors (0.5 cutoff)	78.9

Table 5 GO enrichment analysis of single cell gene factors

Factors	Enriched GO
1	cellular response to interferon-gamma (GO:0071346) cytokine-mediated signaling pathway (GO:0019221)
2	T cell activation (GO:0042110) interleukin-21-mediated signaling pathway (GO:0038114)
3	T cell receptor signaling pathway (GO:0050852) cytokine-mediated signaling pathway (GO:0019221)
4	transcription regulation in response to hypoxia (GO:0061418) neutrophil degranulation (GO:0043312)
3(unique)	T cell activation (GO:0042110) cytokine-mediated signaling pathway (GO:0019221)

Table 6 TF enrichment analysis of single cell gene factors

Genes' factors	Size of enriched genes	Enriched TFs
1	407	EZH1; FOXP3
2	260	IKZF4; XBP1; FOXP3
3	834	EZH1; NKX25; MEIS2
4	316	ZBTB7B; STAT1; XBP1
3(unique)	492	NKX25

We further performed transcription factor (TF) enrichment analysis and gene ontology (GO) enrichment analysis on factors on the gene side to investigate therapy related gene regulation mechanisms. More specifically, input for the analysis were genes with ones after binarization of the four factors that consist of T cells. Shown in Table 5 and Table 6 were the top TFs and GOs significantly enriched ($P<0.001$). Factor 1 and factor 3 were similar as they share 1 factor (EZH1) and 1 Biological process (cytokine-related). (Abdalkader et al. 2016) suggested that the absence of EZH1 was important in controlling proliferation/resting of lymphoid cells. The disruption of EZH1 / EZH2 ratio signified an abnormal immune cell state. Factor 1 might represent T cells in response to INF-gamma, the viral response. We further analyzed genes uniquely activated in factor 3 to distinguish the two. Factor 3 might capture expression patterns of T cells in hypoxia. Actually,

patients with high aggregated values in these two factors tended to be nonresponders. Thus, we suspected that the presence of imbalanced EZH1/EZH2 ratio and hypoxia had negative effects on patient responsiveness. Factor 2 is characterized by IL-21 response, which activates T cell. Patients with high Factor 2 values tend to be responders. This is consistent with current knowledge ((Santegoets et al. 2013). Two enriched TFs in factor 2, IKZF4 and FOXP3, often collaborated in immunosuppressive activities (Jia et al. 2019). Since the two TFs were enriched for knockout experiment, the expression pattern of factor 2 indicated the immunostimulatory state of immune cells. The enrichment analysis, combined with Figure 19, showed that these three T cell dominant factors described above had captured distinct expression patterns capable of discriminating immunotherapy responsiveness.

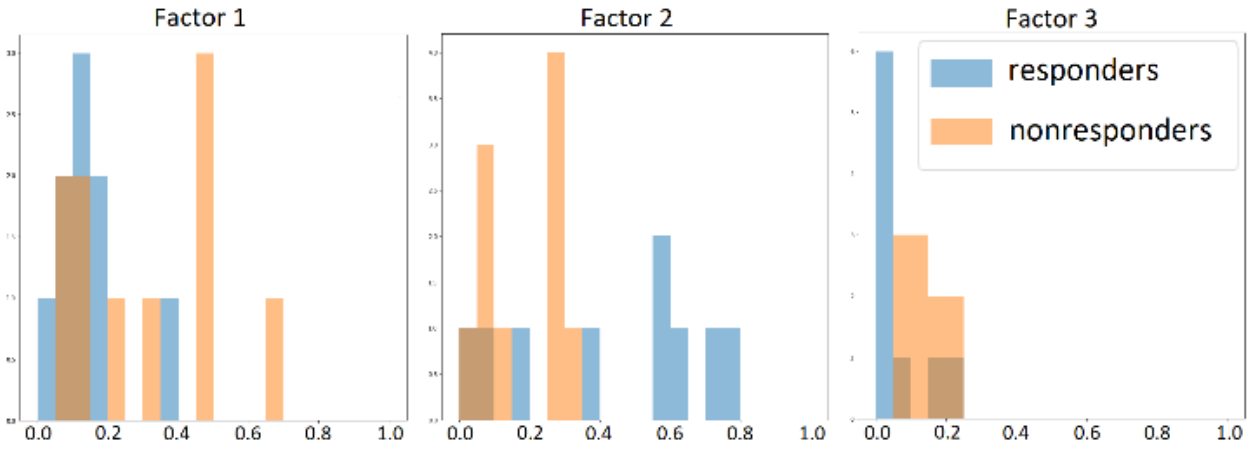


Figure 19 The distribution of responders (blue) and nonresponders (brown) over aggregated Boolean factor values. Bins in deep brown is the overlapping proportions. Nonresponders tend to have higher aggregated values in factor 1 and factor 3, while responders have higher values in factor 2. However, this is not statistically significant due to limited sample size (19 samples).

4.4.2.3 Segmentation of Spatial Transcriptomics

Spatial transcriptomic data about hippocampal formation in adult mouse brain was downloaded from Allen Brain Atlas (Lein et al. 2007). Our selected region had ~ 5000 voxels. Each voxel contained an expression profile of ~ 20000 genes. Gene expression values were measured within situ hybridization (ISH) technology. As shown in Appendix Figure 5, the number of non-expressed genes per voxel was consistent within the same Sagittal section. Thus, we believed that most non-expressed genes were actually missing values and masked them as is. Sagittal sections with less than 3000 expressed genes were removed. Above zero expressions were dichotomized based on the individual average of each gene. Clearly, this dataset contained both missing values and noisy measurements, which was suitable to test our algorithm's performance.

Different numbers of latent factors were attempted, including 2, 5, 10, 15, and 30 factors. As shown in Figure 20 and Figure 21, our algorithm produced spatially tight clusters without the aids of spatial information. We also tried 5 factors, 15 factors, and 30 factors, the voxels assigned to each factor were still close together (see Appendix Figure 6, Appendix Figure 7, and Appendix Figure 8).

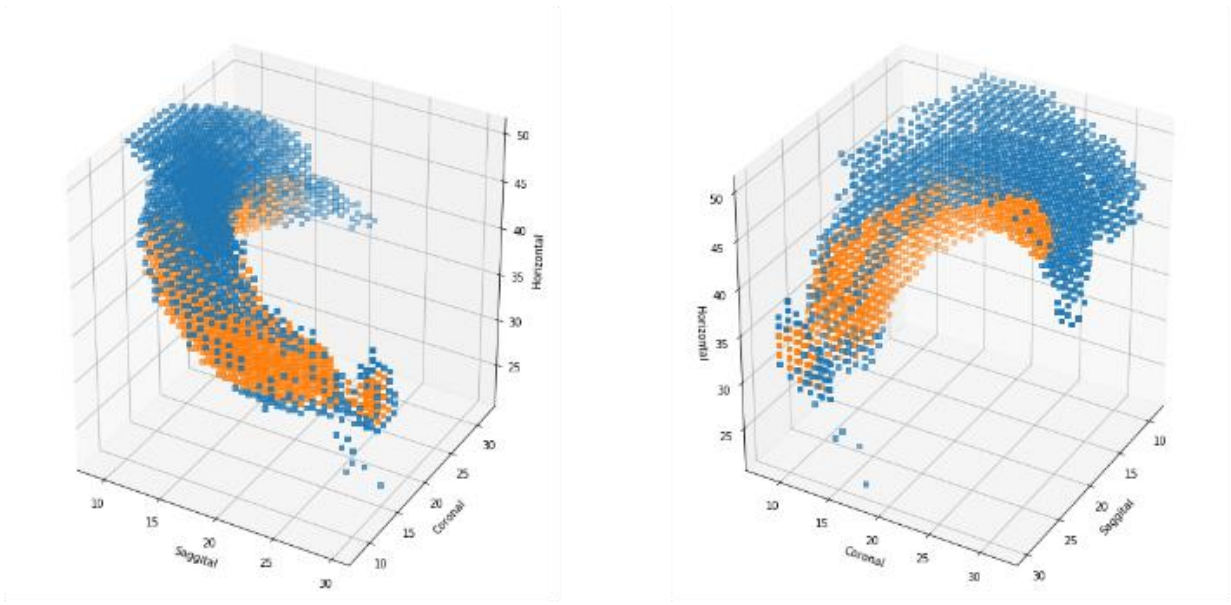


Figure 20 2-factorization of spatial transcriptomics in mouse hippocampal formation

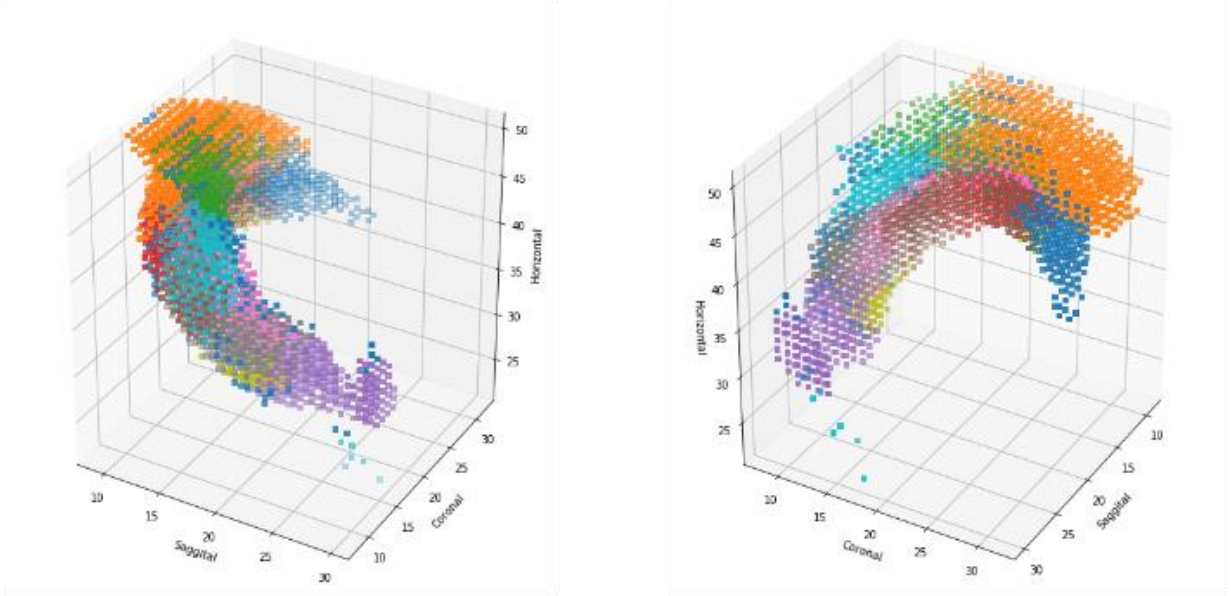


Figure 21 10-factorization of spatial transcriptomics in mouse hippocampal formations

We also investigated the alignment of our voxel factors with the anatomical labels. On a high-level anatomical category (Figure 22), most factors were dominant in either hippocampal region or retro hippocampal region, except factor 1 and factor 10. These two factors may represent the area bridging the two parts or expression patterns unrelated to anatomy labels. With finer granularity (Figure 23), factor 1 and factor 10 were indeed a mixture of many different areas. Other factors seemed to be somewhat aligned with anatomical structure. Only two out of seven labels were dominant in most of those factors.

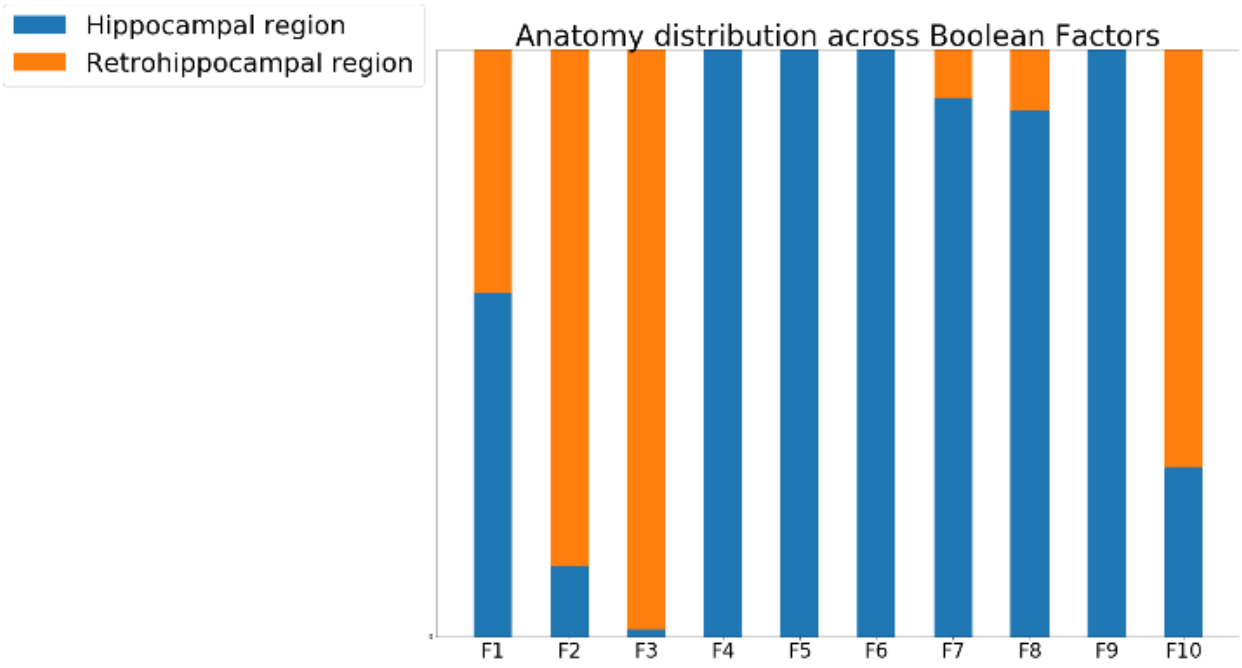


Figure 22 10-factorization of spatial transcriptomics mapped to high-level anatomical labels

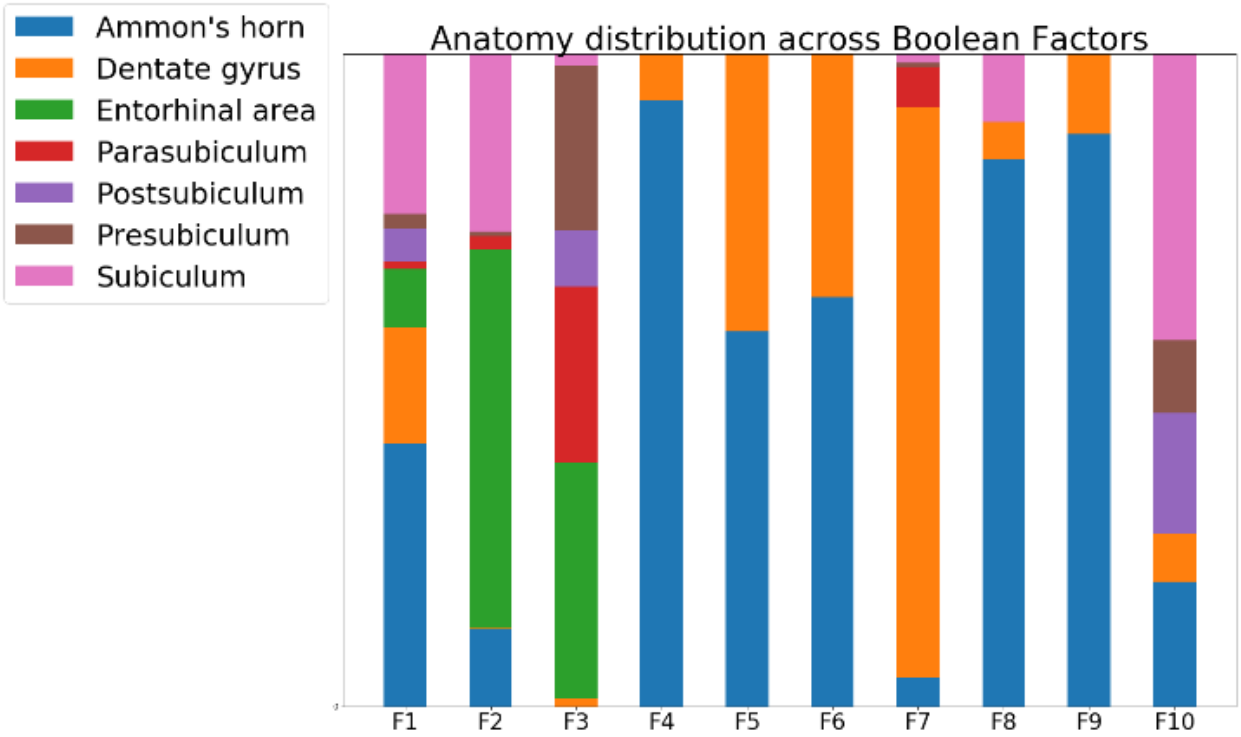


Figure 23 10-factorization of spatial transcriptomics mapped to 7 low level anatomical labels

4.5 Contribution and Limitations

In this chapter, we presented a new algorithm for Boolean matrix factorization via Expectation Maximization (BEM). Synthetic results showed that BEM could recover latent factors accurately even with varying bicluster sizes. We applied BEM to three transcriptomic datasets generated with bulk RNAseq, single cell RNAseq, and ISH respectively. Given appropriate dichotomization, results in Bulk RNAseq and single cell RNAseq showed that BEM was able to extract information related to expression patterns such as disease subtypes and cell types. Results in ISH expression data also indicated that our algorithm could extract neuron expression patterns

without the aid of spatial information. Our algorithm is more suitable for systematic analysis of coregulation patterns, such as subtype classification and gene signature identification.

When analyzing gene expression data with BEM, users should be aware of several practical details. One is hyperparameter tuning. In the step of noise estimation, we assume that noise was symmetric. That is, the probability of 1 flipped to 0 is the same as that of 0 flipped to 1. Future research could further look into ways to alleviate such assumptions.

Our algorithm is applicable to other high-throughput data as long as the tasks of latent variable inference can be represented as dense bipartite subgraph problem or the tiling problem. The hyperparameters in this algorithm, the Beta prior, should be set to 0.95 when binary factor values are preferred. If users need to estimate the uncertainty of the output, the Beta prior should be set to (1, 1). Hence the probabilistic estimates returned by the algorithm are not biased towards 0 or 1.

Boolean matrix factorization extracted clusters from genes and samples simultaneously. However, Boolean factors on the genes' side often consisted of more than a thousand genes. This hinders the interrogation of genes' contribution for phenotype shared within sample clusters. Future research may need to utilize external information to further decompose the gene factors, or identify the minimally representative gene sets for each factor.

5.0 Modeling the Impact of Somatic Mutations on Transcriptomic Profiles by Extending BMF to OR-gate Network

In this chapter, we first described the pattern of mutual exclusivity, which supports that the impact of somatic mutations on signaling pathways can be modeled by the AND-OR product. Then we connect the AND-OR product from somatic mutations to pathways to gene expression to formulate the OR-gate network (ORN). Comparison with fully connected neural network and application to real data show that ORN is capable of identifying patient factors related to survival and novel mutations related to tumorigenesis.

5.1 Modeling Mutual Exclusivity in Somatic Mutation Profiles with the AND-OR Product

Mutual exclusivity (ME) is a phenomenon that mutation events of genes participating in the same pathway often avoid occurring in the same sample. Computational researchers have utilized this pattern to identify mutations affecting the same pathway (Szczurek and Beerenwinkel 2014; Yulan Deng et al. 2019; Leiserson, Reyna, and Raphael 2016).

The mutual exclusivity pattern can be explained by the collider in the Bayesian network. As shown in Figure 24, CDKN2A and RB1 are one of the across-ME pairs identified by MEMCover (Y.-A. Kim et al. 2015), they cooperate to regulate cell cycle (Hatzistergos et al. 2019). Mutation on one of them is sufficient to dysregulate cell cycle and differentiation, significantly increasing the risk of cancer. Therefore, when conditioned on the observation of cancer, mutation events of RB1 and CDKN2A have negative interactions.

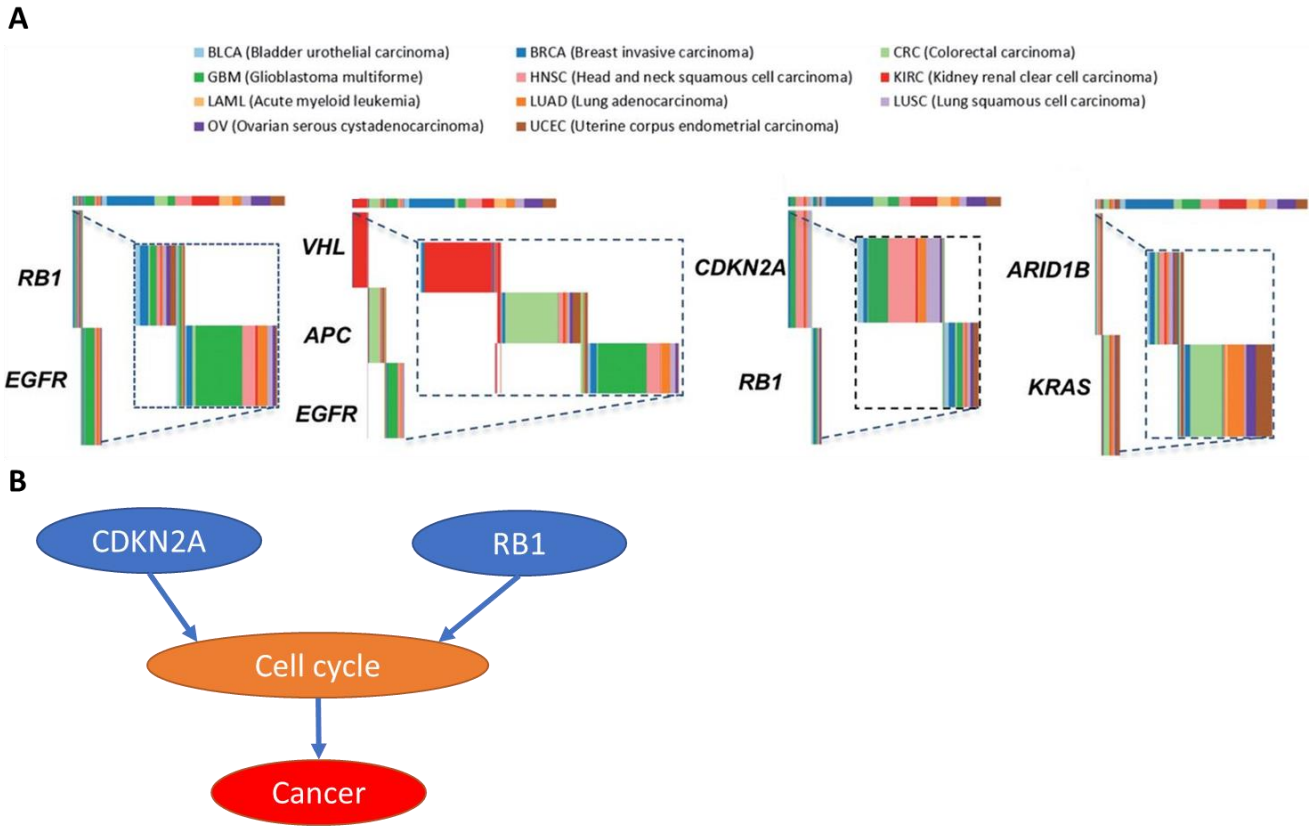


Figure 24 Fig A is a mutual exclusivity plot from Kim Yoo-Ah's study (Y.-A. Kim et al. 2015). Fig B explained the mutual exclusivity among VHL, APC, and EGFR by the collider shape in Bayesian network.

From the analysis above, mutual exclusivity implied that when a group of genes is affecting the same biological functions, mutation on one of them is sufficient to perturb the function, which can be modeled by the AND-OR product defined in Equation 1. Thus, we have

$$Path_{sp} = \vee_{m \leq M} \left(Mut_{sm} \wedge U_{mp} \right)$$

where S is the number of samples, M is the number of genes in genomics, P the number of pathways, and $Path$ is a S×P matrix, Mut is the S×M binary event matrix of somatic mutation, U is the M×P causal relationship matrix between SGA and pathways.

For the convenience of notation, we rewrote the AND-OR product as a vectorized function:

$$Path = U \otimes Mut$$

where $Path$ is the output of the AND-OR product ($\otimes$) of U and Mut.

5.2 Extending BMF to OR-gate Network

As established in Section 4.2, the relationship between differential expression and pathways can be modeled as

$$Expr = Z \otimes Path$$

where $Expr$ is the S×G RNA expression matrix, G is the number of genes in transcriptomic profiles, and Z is the P×G causal relationship matrix between pathways and DEG. Combining the equation we derived in Section 5.1, we have

$$\widehat{Expr} = Z \otimes (U \otimes Mut) \tag{3}$$

where $\widehat{Expr}$ is the estimated differential expression matrix. This notation is necessary because inference of ORN relies on minimizing the difference between $\widehat{Expr}$ and $Expr$ (described in Section 5.3.2). We need to emphasize that the elementwise operation of $\otimes$ is identical to Equation 2 (Section 4.2). Therefore, it is clear how to derive the details of the model from the overall formulation (Equation 3) here.

In addition, we need to emphasize that the matrix $Path$ is equivalent to the matrix U in Chapter 4. They both serve as the indicator for the abnormality status of latent transcriptional regulation mechanism. The term "pathway" here is also similar to the "pathway" in Chapter 4. Usage of this term in the two chapters is not related to the signal transduction order. To be more specific, this chapter contains two concepts that can be referred to as "pathway". One is the "pathway module" that consists of a group of somatic mutations participating in the same function. The other is the "pathway status", indicating whether the pathway is functioning normally or not.

As illustrated in Figure 25, pathway modules are identified by the edge weights (the matrix U). The pathway status is represented by the matrix $Path$.

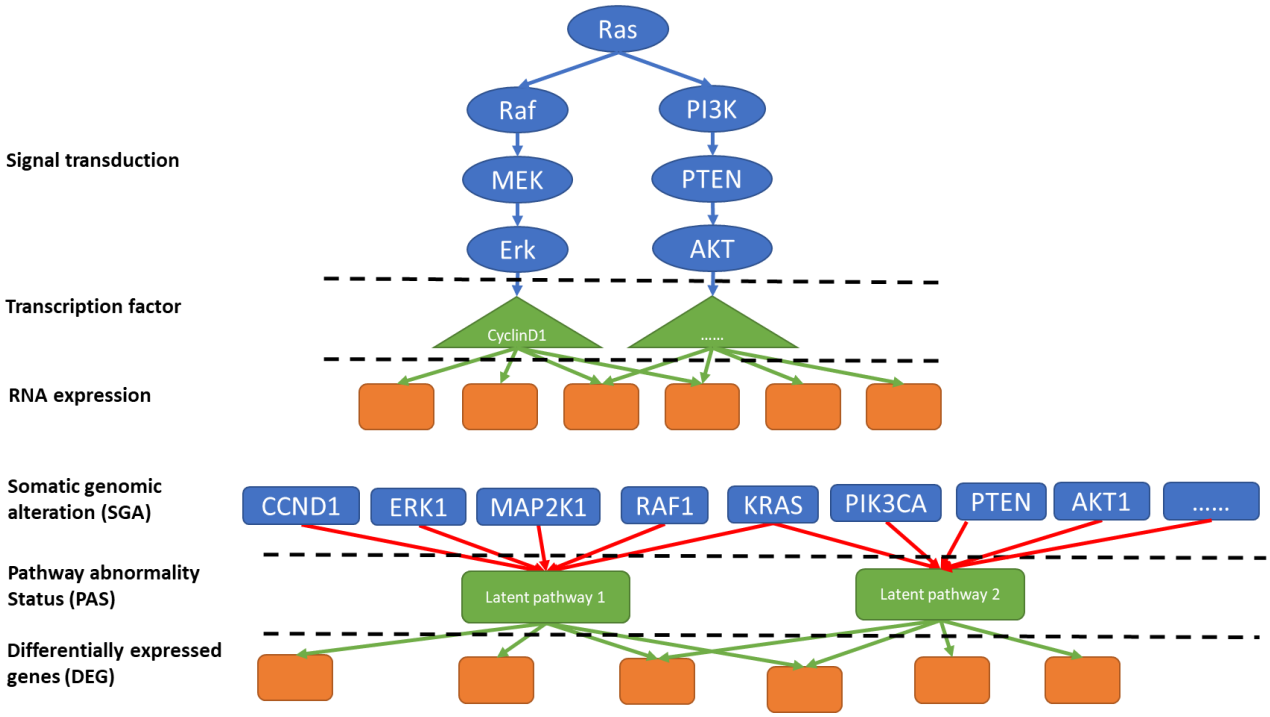


Figure 25 Illustration of pathway representation by ORN. Figure on the upper part is the biologically plausible representation of a signalling network of the gene products. An SGA event in one of the gene products can disrupt the normal signal cascade. Within the ORN framework, we replaced the realistic representation by connecting all the possible SGA events of genes to an OR gate indicating the pathway status. After parameter estimation and causal relation extraction, edges with larger weights remain, resulting in the figure in the lower part. Gene-level SGAs connecting to the same pathway status produce the functional module. Pathway status is then connected to transcriptomics with the same logical OR relationships. Instead of signal transduction or transcription regulation, ORN edges are more abstract, representing noisy logical induction.

5.3 Model Implementation

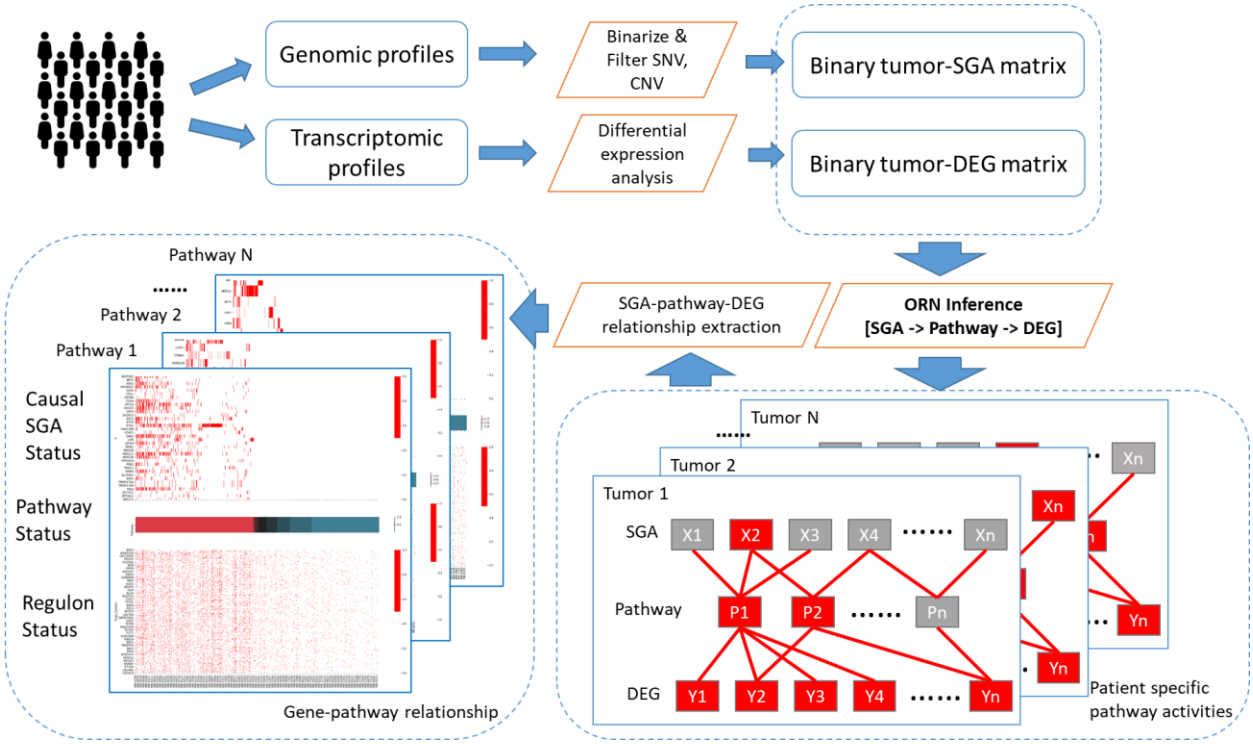


Figure 26 Workflow overview of ORN. The input for ORN consists of quantified matrices of single nucleotide variation (SNV), copy number variation (CNV), and gene expression (RNAseq). SNV and CNV were combined, binarized, and filtered on the genes' side to produce a binary event matrix. As for RNAseq, we calculate robust Z score for each gene in each sample. We assumed Logical OR relations when binary events led to pathway dysregulation and, in turn, led to differential expression. ORN algorithm aimed to infer: (1) the relationship between somatic mutations and signalling pathways; (2) the relationship between signalling pathways and differential expressions. With the ORN output, we can recover pathways that were perturbed by somatic mutations and caused differential expression.

This section covers implementation issues when ORN is applied in simulation experiments and real datasets. As illustrated in Figure 26, we first describe how to transform Genomic profiles and transcriptomic profiles to binary matrices in Section 5.3.1. Then we illustrate how to estimate

model parameters in Section 5.3.2. Section 5.3.3 illustrates how to interpret the results from ORN. Finally, Section 5.3.4 described the evaluation of ORN in Simulation experiments.

5.3.1 Data preprocessing

The input and output of a probabilistic OR-gate function are required to be Boolean variables. Therefore, before applying ORN to real-world datasets, we need to transform genomic profiles and transcriptomic profiles into binary matrices.

For the somatic alteration profiles, binary values dictate whether a gene has somatically mutated within a sample. We used two types of data: (1) non-silent gene-level single nucleotide variation (SNV) dataset; (2) gene-level copy number variation (CNV). An element in the CNV matrix was set to 1 if its original value was 2/-2. The cutoff for CNV was decided because a looser cutoff, such as 1/-1, does not have strong correlations with gene expression. We then combined CNV and SNV data into a binary event matrix, that is, if the alteration of gene i was observed in either SNV or CNV in sample j, then the ijth element in the binary event matrix was set to 1.

For the transcriptomic profiles, a binary value dictates whether a gene has differentially expressed within a sample. We first removed genes with median expression counts lower than 10 across all samples to avoid insufficient statistical power. Then the Z scores provided by the CBioPortal (Cerami et al. 2012) platform were binarized. A gene has differentially expressed if its Z score exceeds the range of P-value 0.05. More specifically, an element in the Z score matrix was set to 1 if its absolute value was greater than 1.96, otherwise 0.

To further filter the genes in the SGA level, we applied Multitask Lasso implemented in Scikit-learn (Pedregosa et al. 2011). The genomic profiles were used as independent variables and

the status of differential expression as targets. Somatic mutations with nonzero coefficients were retained as the input for ORN.

5.3.2 Gradient-based parameter estimation

Given the relationship matrix U and Z, the aggregate pathway status can be computed with the OR gate function. Therefore, we only need to estimate U and Z in the model. These two types of parameters are estimated by maximizing the likelihood of observed gene expression given the generative process of ORN. That is, the objective function to optimize for ORN is the overall log likelihood of the observed Expr given estimated Expr:

$$LL(Expr) = \sum_{s \le S, g \le G} [Expr_{sg} log \widehat{Expr_{sg}} + \left(1 - Expr_{sg}\right) log\left(1 - \widehat{Expr_{sg}}\right)]$$

where $\widehat{Expr_{sg}}$ is the probability of differential expression of the *g*th gene in the *s*th sample computed by the ORN. Similar latent variable models, such as LDA (Blei, Ng, and Jordan 2003), are usually computationally expensive with MCMC or variational inference. However, we found that the layered structure of ORN is similar with the neural network architect. Thus, the learning algorithm essential to all deep learning models, back propagation, can be used to estimate ORN parameters.

First, we need to reparameterize U and Z such that the parameters are not bounded within [0,1]. That is, every element in the matrix U and Z is regarded as an output of a scalar within $[-\infty, +\infty]$:

$$U_{mp} = sigmoid(\mu_{mp})$$

$$Z_{pg} = sigmoid(\zeta_{pg})$$

where μ and ζ are matrices with the same shape as U and Z, but their values are unconstrained. This enables us to apply gradient-based methods to identify maximum likelihood of ORN regarding μ and ζ.

During implementation, we adopted the Rprop algorithm (Braun 1992) to learn μ and ζ. This algorithm requires the gradient of ζ with respect to $LL(Expr)$:

$$\frac{\partial LL}{\partial \zeta_{pg}} = Z_{pg}(1 - Z_{pg}) \sum\nolimits_{n \leq N} [\frac{Path_{sp}(1 - \frac{Expr_{sg}}{\widehat{Expr_{sg}}})}{1 - Path_{sp}Z_{pg}}]$$

Using the chain rule, we can also derive the gradient of μ with respect to $LL(Expr)$:

$$\frac{\partial LL}{\partial \mu_{mp}} = \sum\nolimits_{s \leq S} \frac{\partial LL}{\partial Path_{sp}} \cdot \frac{\partial Path_{sp}}{\partial U_{mp}} = \sum\nolimits_{s \leq S} \frac{\partial LL}{\partial Path_{sp}} \cdot \frac{Mut_{sm} U_{mp}(1 - U_{mp})(1 - Path_{sp})}{1 - Mut_{sm} U_{mp}}$$

where the computation of $\partial LL / \partial Path_{sp}$ is symmetric to $\partial LL / \partial Z_{sp}$.

To control the sparsity of parameters, we assume U and Z are samples from the Beta distribution. Thus, the gradients above need to be modified. For example, the partial derivative of ζ should be modified as:

$$\frac{\partial LL}{\partial \zeta^*_{pg}} = \frac{\partial LL}{\partial \zeta_{pg}} + (\alpha - 1)(1 - \zeta) + (\beta - 1)\zeta$$

where α and β are the hyperparameters for Beta distribution. For all the experiments in this study, we set $\alpha = \beta = 0.95$. The procedure for model estimation has been summarized in Figure 27.

Algorithm 1: ORN inference

Input : Mut, a $S \times M$ binary matrix; Expr, a $S \times G$ binary matrix; P number of latent pathways; α, β, Beta priors; MaxIter, maximum iterations for the gradient descent

Output: U, a $M \times P$ binary matrix; Z, a $P \times G$ binary matrix; $Path$, a $S \times P$ binary matrix

$\mu^{M \times P} \leftarrow Gaussian(mean = 0, std = 0.1)$;
$\zeta^{P \times G} \leftarrow Gaussian(mean = 0, std = 0.1)$;
$i \leftarrow 0$;
while $i < MaxIter$ **do**
    $U \leftarrow sigmoid(\mu)$;
    $Z \leftarrow sigmoid(\zeta)$;
    $Path \leftarrow OR(Mut, U)$;
    $\hat{X} \leftarrow OR(Path, Z)$;
    $G_{\mu}^{M \times L}, G_{\zeta}^{N \times L} \leftarrow ComputeGradient(Mut, Expr, U, Z, \alpha, \beta)$;
    $\mu, \zeta \leftarrow RPROP(\mu, \zeta, G_{\mu}, G_{\zeta})$;
    $i \leftarrow i + 1$;
end
$U \leftarrow sigmoid(\mu)$;
$Z \leftarrow sigmoid(\zeta)$;
$Path \leftarrow OR(Mut, U)$;
return U, Z, $Path$

Figure 27 The pseudo code to compute the two relationship matrices U and Z, and the pathway activities

5.3.3 Causal relation extraction

After learning the model parameters, μ and ζ, we can recover U and Z through the element-wise sigmoid function of μ and ζ. The matrix of patient-specific pathway status, *Path*, can be recovered by computing the OR gate function given SGAs and U.

When applied to real-world datasets, we also need to extract pathway modules and regulons corresponding a latent pathway. The pathway module was determined by the matrix U. If $U_{mp} >$

0.5, then we concluded that mutation of gene *m* could disrupt pathway *p*. From an operational perspective, a pathway module corresponding to pathway *p*' would be genes with $U_{mp'} > 0.5$.

Regulon related to a latent pathway is extracted similarly. If $Z_{pg} > 0.1$, then we conclude that the disruption of pathway p can cause differential expression of gene g. In this way, the set of genes regulated by the same pathway are grouped into a coexpression module. Note that for real data analysis in this study, the cutoff for elements in Z was the top 5% value among all genes in the pathway *p*.

Please note that for convenience, the module of SGAs and the modules of DEGs corresponding to a latent pathway will simply be called "upstream module" and "downstream module" respectively in following sections.

Since there is no ground truth for real data analysis, we performed Gene Ontology (GO) enrichment analysis on the downstream modules to characterize the functional impacts of the upstream modules

5.3.4 Simulation and evaluation

Synthetic data was generated by following the probabilistic OR gate mechanism. First, somatic mutations and the two relationship matrices were generated with Bernoulli distribution. Then *Path* and *DEG* were generated by performing noisy OR-gate computation with $P_0 = 0$. To simulate the mutual exclusivity patterns observed in real data, we performed post pruning. When several mutations belonging to the same pathway took place in the same sample, all but one of them were removed.

The artificial neural network (NN) was used as a baseline to evaluate ORN's efficacy of inferring pathway activities. To ensure the neural network model was comparable with ORN, it

had one hidden layer, and the activation function was sigmoid. In this way, the values of hidden neurons were also within [0,1]. We ran NN and ORN on 20 synthetic datasets and computed their reconstruction error and the average cosine similarity of pathways. Reconstruction error was computed as:

$$\text{error} = \frac{\sum_{s \le S, g \le G} |\widehat{Expr_{sg}} - Expr_{sg}|}{S \times G}$$

We further propose Jaccard score to evaluate how ORN's performance changes in various settings. Unlike reconstruction error, this criterion measures the similarity between the inferred relationship matrix and the ground truth. To compute Jaccard score, we first need to compute Jaccard similarity for U and Z respectively:

$$sim(U^*_{\cdot p}) = max_{p' \le P}\left[Jaccard(U^*_{\cdot p}, U_{\cdot p'})\right]$$

$$sim(Z^*_{p\cdot}) = max_{p' \le P}\left[Jaccard(Z^*_{p\cdot}, Z_{p'\cdot})\right]$$

where U* and Z* are the true relationship matrix in synthetic data. The function Jaccard(A, B) takes the form:

$$Jaccard(A, B) = A \cdot B/(A + B - A \cdot B)$$

Then the Jaccard score for one dataset is:

$$Jaccard\ score = \sum_{p \le P} sim(U^*_{\cdot p}) sim(Z^*_{p\cdot})/P$$

5.4 Experimental Results

5.4.1 ORN was effective in recovering OR-gate relationships

Synthetic data were generated according to the generative process described in "Simulation and evaluation". The number of pathways was set to 5; The number of samples, SGAs, and DEGs were all set to 1000. This was referred to as the standard setting.

We proposed Jaccard score and reconstruction error to evaluate the performance. Jaccard score can measure the concordance between inferred relationship matrices and the ground truth. Details of calculation was described in "Simulation and evaluation".

From the standard setting, each condition was changed separately to see how they affected the performance. As shown in Figure 28, ORN has achieved almost perfect recovery (>99%) in the standard setting. However, ORN's performance dropped over 20% when the number of samples dropped from 1000 to 500, or the number of mutations increased from 1000 to 3000 mutations. Note that when the number of DEGs were reduced to 500, the performance remained the same. Interestingly, the performance of ORN also decreased and became unstable when more pathway modules were needed.

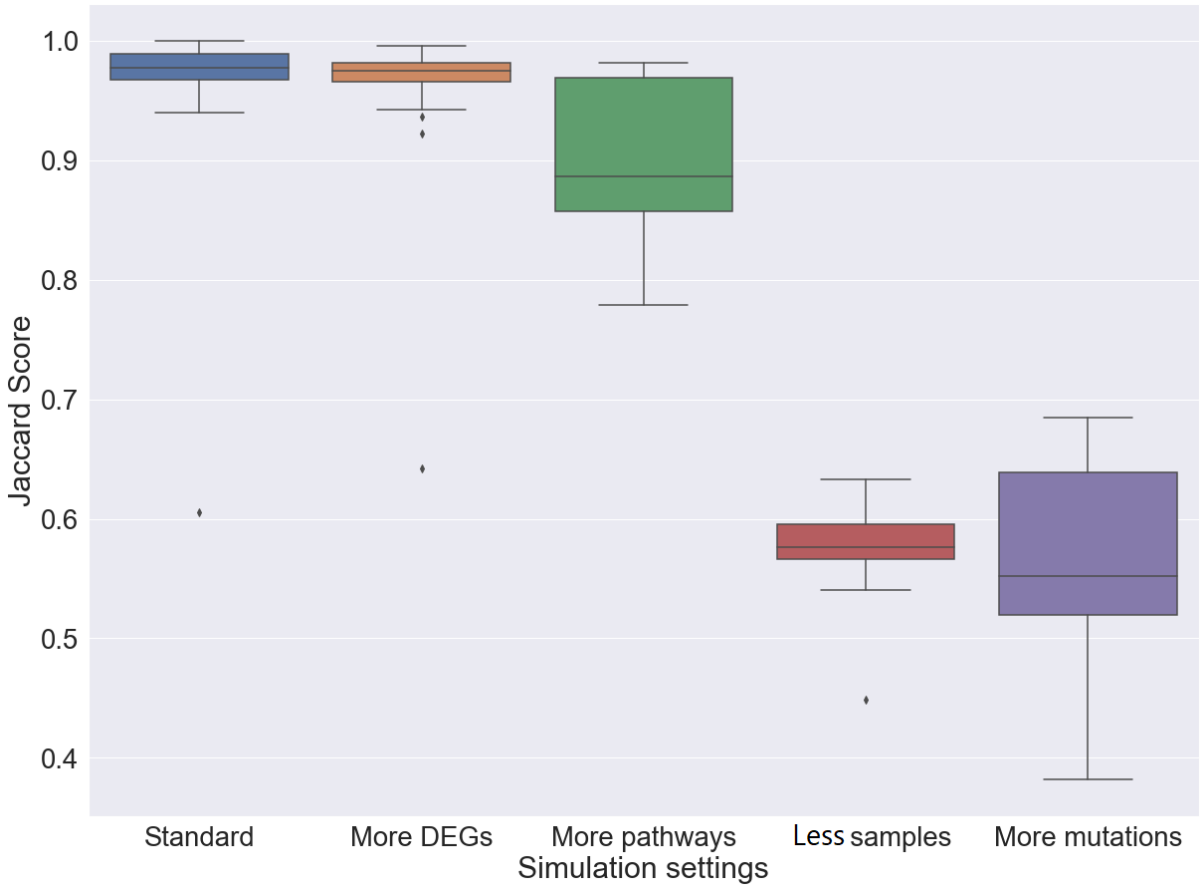


Figure 28 Performance of ORN across different settings. In the standard setting, ORN has recovered pathway modules with almost perfect accuracy. Reducing the number of DEGs did not affect the performance of ORN. Adding more pathway modules would introduce more variation to ORN's performance. When the number of samples decreased to 500 or the number of mutations increase to 3000, the median Jaccard score has decreased to 73% and 81% respectively.

5.4.2 ORN provided more insights than the neural network

We cannot identify similar algorithms that only used high-throughput data to infer pathway activities. However, we found that artificial neural networks (NN) with sigmoid activation function can also produce binary values in the hidden layer. In addition, both ORN and NN used backward

propagation to optimize parameters. Thus, we designed a neural network architecture similar with ORN and used it as a baseline.

In the synthetic experiment, although NN converged to comparable reconstruction error as ORN, its accuracy in pathway recovery was only around 50% (Figure 29). This showed that NN is less capable of capturing all the signals in the data.

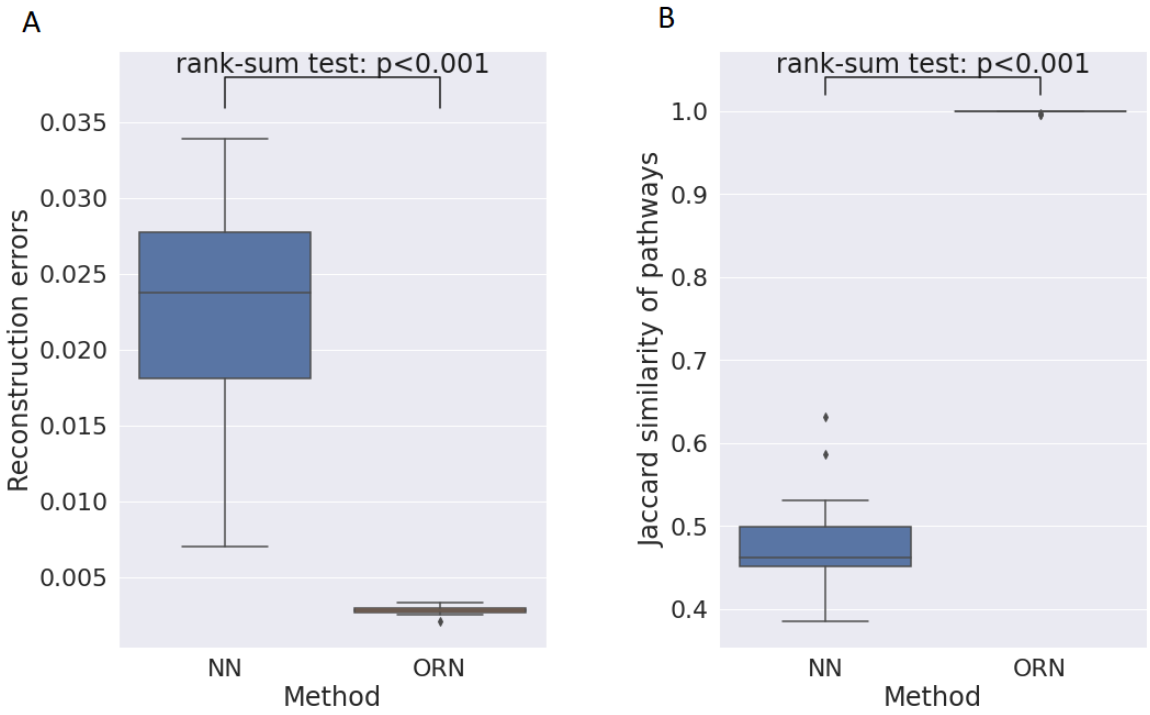


Figure 29 Comparison of ORN and NN on synthetic datasets. Boxplot on the left showed the distribution of prediction error of NN and ORN across 20 synthetic experiments. Boxplot on the right showed the distribution of cosine similarity between the inferred pathways and ground truth

Similar results were observed when we applied NN to the glioma dataset (described in Section 5.4.3). The relationship matrix estimated with NN is much more redundant than ORN. GO enrichment analysis of the downstream modules (see Appendix Table 2) also showed that NN

could only capture less than 5 major aspects with 10 hidden neurons, while ORN can cover different biological aspects of glioma with each pathway module. As shown in Figure 30, the relationship matrix learned by NN contains much redundancy, while each pathway in ORN regulated different sets of genes with few overlaps. This indicated that the biological mechanism from somatic mutations to transcriptomic profiles could be more accurately characterized by the OR-gate logic imposed by ORN rather than conventional non-linear relationships.

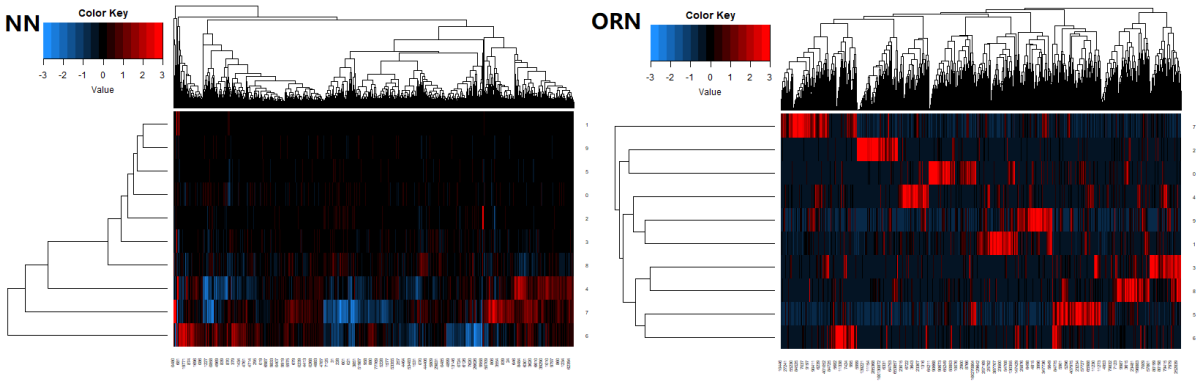


Figure 30 Heatmap representation of the relationship matrix between pathways and differential expression after row normalization. Relationship matrix generated by a neural network (NN) contains many redundant signals, while ORN automatically pushes for sparsity. Each pathway module in ORN has uniquely caused a subset of genes to express differentially.

5.4.3 ORN detected pathways closely related to patient survival

After applying ORN to the lower grade glioma dataset. Pathway activities showed that pathway 6 and pathway 7 had significant impacts on patient survival (Figure 31). We performed Gene Ontology (GO) enrichment analysis on the top DEGs in these pathways (see Appendix Table 2). The downstream module of pathway 6 is mostly related to DNA processing activities. Cancer

samples with this pathway dysregulated probably have compromised genomic instability (Negrini, Gorgoulis, and Halazonetis 2010), leading to worse survival. Its upstream module includes CDK13 (Blazek et al. 2011), H3F3A (Tagami et al. 2004), IDH1 (Wu et al. 2019), PTEN (Ho et al. 2020), SNRPE (Z Li and Pützer 2008) that are closely related to DNA repair or DNA replication.

As for pathway 7, we found that PTEN, H3F3A, and POM121L12 were shared by the upstream modules in both pathways. However, the top 300 DEGs caused by the two pathways have no genes in common. GO enrichment analysis showed that downstream modules are related to neutrophil activities, Ras signal transduction, and viral genome replication. We conjectured that cancer samples with pathway 7 dysregulated exhibited viral infection and its immune response. Since virus infection can drive glioma formation (McFaline-Figueroa and Wen 2017), This subgroup of patients may be more likely to progress to malignancy and worse survival.

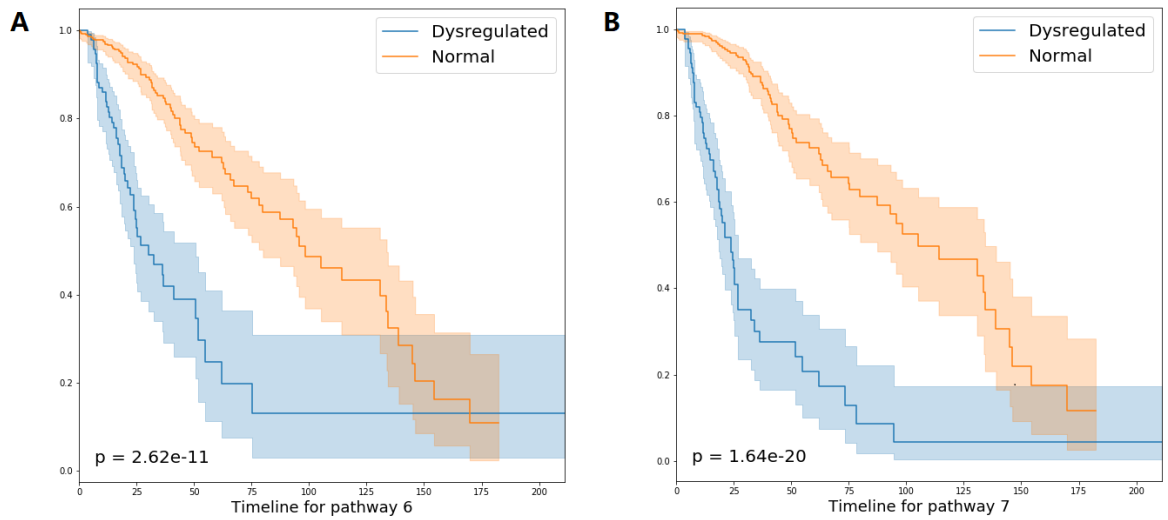


Figure 31 LGG patients with pathway 6 (A) and pathway 7 (B) dysregulated have worse overall survival. X-axis is in the unit of month; Y-axis represents the proportion of each subgroup. 102 patients' pathway 6 were dysregulated, 100 patients' pathway 7 were dysregulated. Both groups have 62 samples in common.

When applying ORN to liver cancer samples, we also identified two pathways related to patient survival (Figure 32). One is the pathway 2. GO enrichment analysis of the upstream alterations showed that pathway 2 mainly affected epithelial tube formation (GO:0072175) and other activities located on membrane (cytoskeletal anchoring and protein localization). Epithelial tube formation contributed to epithelial to mesenchymal transition (EMT) and mesenchymal to epithelial transition (MET) that played a vital role in liver cancer development and metastasis (Xia et al. 2015). the overlapped gene Podocalyxin (PODXL) was found to be overexpressed in HCC cell line and could be used as a biomarker to predict the prognosis of HCC due to participating in HCC migration and invasion processes (Amantini et al. 2016). Fibrosis growth factor receptor 2 (FGFR2) (also in neuron projection morphogenesis, GO:0048812) and its partner driver gene were frequently found in intrahepatic cholangiocarcinoma (ICC) (F. Li, Peiris, and Donoghue 2020; Sia et al. 2017). This pathway module was also enriched for regulation of cardiac conduction (GO:1903779). One gene related to cardiac conduction, ASPH, was found to be is highly overexpressed in cholangiocarcinoma(CCA) and HCC (Lavaissiere et al. 1996). Inhibition of ASPH could decrease CCA development (Nagaoka et al. 2020). Another related gene, ITPR2, is the major intracellular calcium release channel in hepatocyte to regulate Calcium (Ca2+) signaling, resulting in regulating lots of function of hepatocytes, including glucose and lipid metabolism, apoptosis, gene transcription, bile secretion, and cell proliferation (Kruglov et al. 2011). ITPR2 was also found to be decreased in fatty liver with impaired liver regeneration (Kruglov et al. 2011).

The other one, pathway 3, is related to progress free interval. Enrichment analysis showed that both upstream mutations and downstream regulons participate in the lectin pathway. Few studies have investigated the association between lectin and liver disease. A recent study (Schierwagen et al. 2020) showed that expression c type lectin played an important role in different

stages of chronic liver disease. It is possible that lectin in important for the immune response within tumor environment. Several genes related to lectin pathway were shown to be important in liver cancer. For example, KRAS is usually found mutated in CCA patients (Mahipal et al. 2018; Ross et al. 2014). The proto-oncogene tyrosine kinase Src is usually aberrant expressed in HCC with an effect on cell proliferation, differentiation (Zhu et al. 2020; Walker et al. 2019; El Sayed, Helmy, and El-Abhar 2018; Lau et al. 2009). Meanwhile, mucin 5AC (MUC5AC) as a secreted Mucins, It was upregulated in ICC and CHC patient and inflammation (L. Yang et al. 2013), it is a good diagnostic marker in CCA and be used a biomarker to differentiate CCA from benign biliary disease (Xuan et al. 2016; Cuenco et al. 2018). The upstream module in pathway 3 is also enriched for cation transport (GO:0006812). Genes related to cation transport were also important for liver cancer. For example, LRP2 is involved in fusion in HCC patients (Fernandez-Banet et al. 2014), and some research found complement C3 concentration changes occurred at very early stage of tumorigenesis in serum proteins of diethylnitrosamine (DEN). 2-AAF induced Wistar rats tumor model (Malik et al. 2013), it might be a novel therapeutic approach for liver cancer (Xu et al. 2020).

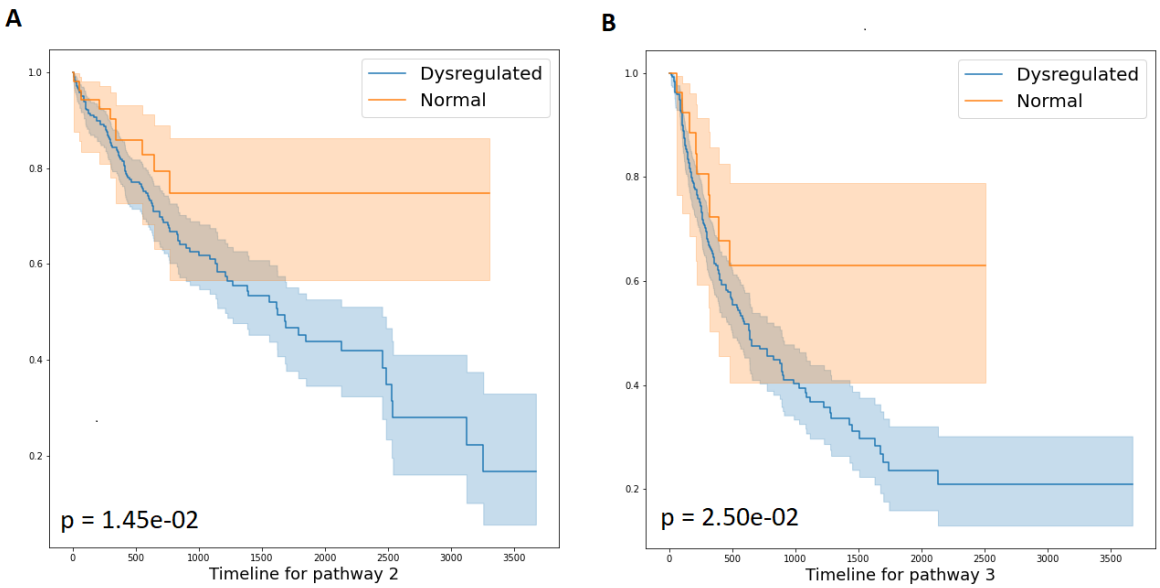


Figure 32 Liver cancer patients with pathway 2 (A) dysregulated have worse overall survival. Dysregulation of pathway 3 results in worse progression free interval.

5.4.4 ORN characterized common mechanisms in cancer

As for the results in the METABRIC dataset, upstream modules of six pathway modules were almost identical and hence merged into one union (pathway 0, 1, 2, 7, 9, 14 in Appendix Table 3). This module contained well-known oncogenes such as TP53, PTEN, PIK3CA, and MAP3K1. The corresponding pathway was dysregulated across all samples (probability above 0.5). This pathway likely represented the common cause of breast cancer. GO enrichment analysis (see Appendix Table 4) showed that its corresponding downstream module was mostly involved in mitosis, such as G1/S transition of mitotic cell cycle phase transition (GO:0044772). Pathway 5 also exhibited similar downstream effects but had different SGAs. Most notably, MIR604 was in the upstream modules of pathways. Studies have shown that polymorphism of MIR604 was

related to the development of hepatocellular carcinoma (Cheong et al. 2014) and the metastasis of colorectal cancer (Boni et al. 2011). MIR604 was differentially expressed in breast cancer (S. Zhang et al. 2018). Still, to our knowledge, the impact of MIR604 mutation in breast cancer has not been investigated. In the case of LIHC, both the upstream and downstream modules in pathway module 1 were significantly enriched with nucleotide-excision repair (GO:0006289) and DNA repair (GO:0006281). This pathway contained TP53BP1, ERCC3, BRCA2, which are well known for the DNA maintenance activities.

In the glioma dataset, we found pathway 4 to be closely related to immune response. Downstream modules of pathway 4 were enriched for various immune responses, including cytokine-mediated signalling and toll signalling. This subgroup of patients may not be responsive to immune therapy. Within this pathway was the mutation of interferon alpha 21 (IFNA21), which played an important role in inflammatory response and toll signalling. IFNs were also identified as major factors of patient response to various cancer therapies (Budhwani, Mazzieri, and Dolcetti 2018). Moreover, we found that PTK6 and SRMS were within the same upstream module. The products of two genes work closely together as intracellular kinases (Serfas and Tyner 2003) and promotes invasive prostate cancer (Wozniak et al. 2017). However, they are rarely studied in the context of glioma and immune response.

Like glioma, one particular pathway in breast cancer captured abnormal immune response in a subgroup of cancer samples. The downstream module in pathway 3 is related to immune response, including T cell activation (GO:0042110), regulation of immune response (GO:00507006), inflammatory response (GO:0006954). The upstream module included CDC20, COLEC12, MED8, MPL, SOX5, and OTUD1. CDC20 was known to be related to T cell activation. COLEC12's protein product is associated with innate immunity (Ma et al. 2015). SOX5

was shown to be related to B cell proliferation (Rakhmanov et al. 2014). Another interesting gene was MED8. Studies showed that MED8 was important to regulate resistance against bacteria in plants (An and Mou 2013). Meanwhile, MED8 was implicated in renal cell carcinoma. However, it is rarely investigated in the case of breast cancer and innate immunity. As for OTUD1, a recent study (L. Zhang et al. 2018) has shown that its induction by RNA virus may inhibit innate immune response.

5.4.5 ORN detected pathway dysregulation specific to cancer types

Although not related to patient survival, other pathways in glioma samples also captured different aspects of molecular characteristics. For example, pathway 0 is closely related to the biosynthesis of cholesterol, steroid, and alcohol, while cholesterol metabolism has recently been studied as a potential therapeutic target (Pirmoradi et al. 2019). In addition, downstream modules of pathway 0 contained differentially expressed genes enriched for central nervous system development. In the corresponding upstream module, we identified SZT2 (Tsuchida et al. 2018) and TIAM1 (M. B. Miller et al. 2013) to be closely related to nervous system development. Other mutations, such as CPAMD8 and RUBP1, exhibited mutual exclusivity and similar expression patterns. Yet, these two genes have not been studied in terms of central nervous system development.

As for the breast cancer samples, the upstream module of pathway 8 contained several well-known driver mutations, including KRAS, APC, and ARID1A. As shown in Figure 33, most genes in this upstream module exhibited mutual exclusivity, while these mutations caused differential expression of a similar set of genes. The downstream module was enriched for telomere and t-circle formation, which were well-known factors for cancer initiation and tumor survival

(Jafri et al. 2016). In the upstream module, the relation between telomere and APC (Yibin Deng, Chan, and Chang 2008), KRAS (W. Liu et al. 2017), ARID1A (Zhao et al. 2019), PRKG1 (Lee et al. 2013) were reported. Although mutually exclusive to the four genes above, we have not found research linking BAP1, MIR604, and MICAL3 to telomere activities.

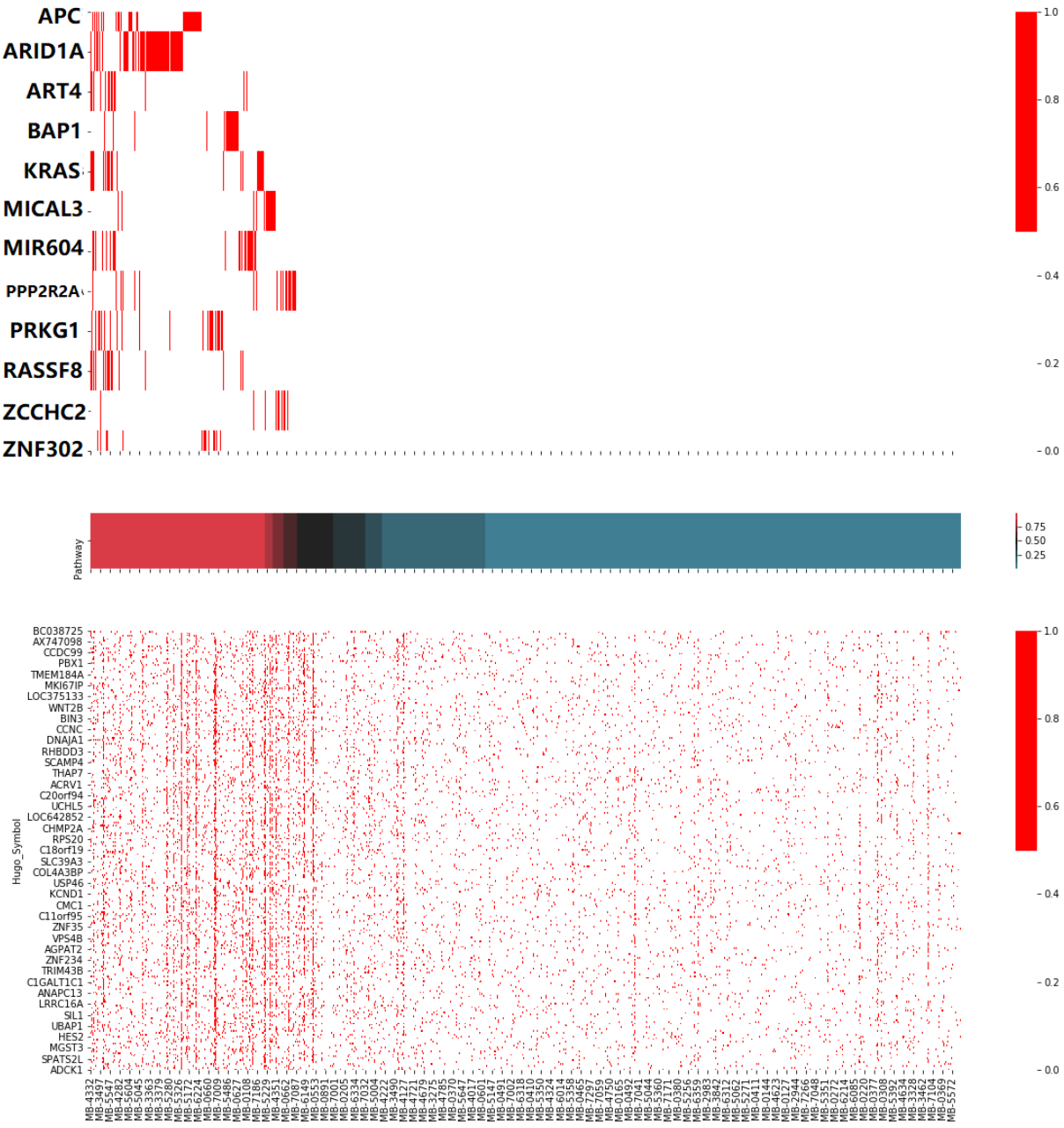


Figure 33 Illustration of pathway 8 in breast cancer samples. Cancer samples sorted by pathway activities (middle figure) was the X-axis shared by the three subplots. The upper figure showed the mutation event of the upstream module (cutoff 0.5), while the bottom figure showed the heatmap of differential expression of the downstream module (roughly top 1% related). The top figure showed patterns of mutual exclusivity, while the bottom showed a strong correlation between pathway activities and differential expression.

5.5 Contribution and Limitations of ORN

To our knowledge, ORN is the first *de novo* method to infer patient-specific pathway activities from genomic profiles to transcriptomic profiles of cancer patients. Compared with the traditional neural network method, ORN provided much more insight into how the somatic mutations function together. In a traditional neural network, the activation of a node is decided by the accumulation of all inputs. In contrast, the output of a node in ORN would be true if any input to this node is true. Hence, the ORN agrees with the premise of mutual exclusivity of somatic alteration in tumors.

Meanwhile, OR-gate also allows the co-occurrence of genes within the same pathway. This flexibility enables the model to handle the rare situation where genes within the same pathway mutated together. Still, when the pathway modules were known, the relations between SGA and pathways established the "collider" shape well known in Bayesian network. Thus, when learned with backward propagation, ORN tends to identify the mutual exclusivity patterns.

Besides mutual exclusivity, there are many other patterns or biological mechanisms in cancer biology. For example, if two mutations co-occurred in most samples, then they are likely to disrupt two different pathways causing tumor. This co-occurrence pattern is best captured by AND-gate instead of OR-gate. However, the output of AND-gate should be the occurrence/progression of tumor. Such information is difficult to integrate into our model. In the future, we may consider modifying the model to integrate other relevant information.

Please note that the "pathway module" formulated in this study may not exactly correspond to a known pathway. In the real data analysis, we showed that a pathway module may represent the impact of somatic alterations on immune response and microenvironment (i.e. pathway 4 in glioma and pathway 3 in breast cancer). In addition, a pathway module may also encapsulate the

joint status of multiple pathways. For example, pathway 0, 1, 2, 7, 9, and 14 in breast cancer probably captured a common set of dysregulated pathways. That is because ORN is performed on the sample level. If several biological pathways were dysregulated on the same sets of patients, ORN could not distinguish them.

In the real data analysis with lower grade glioma and breast cancer, ORN has recovered major mechanisms consistent with current knowledge, such as abnormal DNA repair ability and immune response. Glioma patients with these dysregulated pathways had lower survival rates. ORN further revealed mechanisms specific to cancer types, such as steroid metabolism and nervous system development in glioma. We identified several somatic mutations that might be related to certain malfunctions in cancer cells, worthy of further biological investigation. However, ORN requires in-depth analysis to obtain useful insights. In the future, we will try to develop statistical tests to automatically return meaningful genes in both upstream and downstream modules.

For the METABRIC dataset, none of the aggregate pathway status inferred by ORN is significantly related to patient survival. We conjectured that there are two reasons: (1) Somatic mutations may not be the only source of variance of RNA expression. Sharma, et al (Sharma, Jiang, and De 2018) showed that copy number alterations, epigenetic changes, transcription factors, and microRNAs collectively explain, on average, only 31–38% and 18–26% expression variation; (2) compared with glioma samples, cellular constitution in breast cancer samples was probably more diverse. To handle the first issue, we need to include more data sources in a principled way. In the future, epigenetic profiles may be included to inform the coregulation of RNA expression. To deal with the second issue, future research needs to incorporate reliable complete deconvolution algorithms in the data preprocessing step.

For some genes in the pathway module, we failed to find evidence supporting their functional associations to the affected DEGs. Although some of them were likely to provide novel molecular insights, many were false positive. Upon closer investigation, we believe there are two major sources of false positives: (1) passenger mutations that exclusively occur in highly mutated samples. For example, ACSS1 was in most upstream modules of glioma because it only occurred in highly mutated samples, which had most pathways dysregulated. (2) passenger mutations within the same copy number variation event as driver mutations. When a set of genes mutated in almost the same set of patients, it is likely that only one of them contributed to the pathway dysregulation.

During analysis, we also found that different pathway modules may share a subset of somatic mutations. For example, pathway 6 and pathway 7 in glioma have PTEN in common. Thus, it is possible that hierarchical structures of SGA functions can be inferred from the overlapping pathway modules. In the future, we may provide more convenient visualization utilities to analyze the hierarchies among pathway modules. In addition, so far we have not found an effective measure to identify the appropriate number of pathway modules. We encourage future research on the balance between model complexity and model likelihood of ORN.

We proposed ORN to infer pathway modules and their dysregulation status from high-throughput profiles of cancer samples. Application of ORN in lower grade glioma and breast cancer detected pathway modules closely related to patient survival. ORN also connected somatic mutations to key mechanisms of cancer, such as DNA repair and innate immune response. Although some mutations' function (e.g., MIR604) was not supported by literature, they were mutually exclusive to well-known driver mutations and caused differential expression in a similar subset of genes. We encouraged biological researchers to use ORN to infer personalized pathway activities and generate novel hypotheses for targeted therapy.

6.0 Distinguishing Cancer Drivers from Coamplification with Gene Embedding Learned from Biomedical Literature

This chapter describes the motivation, rationale, methods, and results of the construction of gene embedding. More specifically, it is about how we managed to extract knowledge from biomedical literature to address the co-amplification issues observed in ovarian cancer.

6.1 The Problem of Coamplification

As a type of structural variation, copy number variation is a common somatic mutation event that drives tumorigenesis. When amplified, certain DNA sections, as long as millions of base pairs, are duplicated on the same chromosome. As a result, it is difficult to pinpoint the driver genes from other passengers if they are located in the same copy number variation hotspot. For example, as shown in Figure 34, *NOTCH3* and *PIK3CA* co-amplified with several other genes in ovarian tumors simply because these genes are located near each other in the genome. Among all the genes located in the same copy number variation hotspot, only one or two genes contribute to tumorigenesis. Conventional statistical analysis (e.g., mutual exclusivity analysis) cannot distinguish the driver gene from its neighboring passengers. However, this challenge can be addressed easily with our gene embedding vectors.

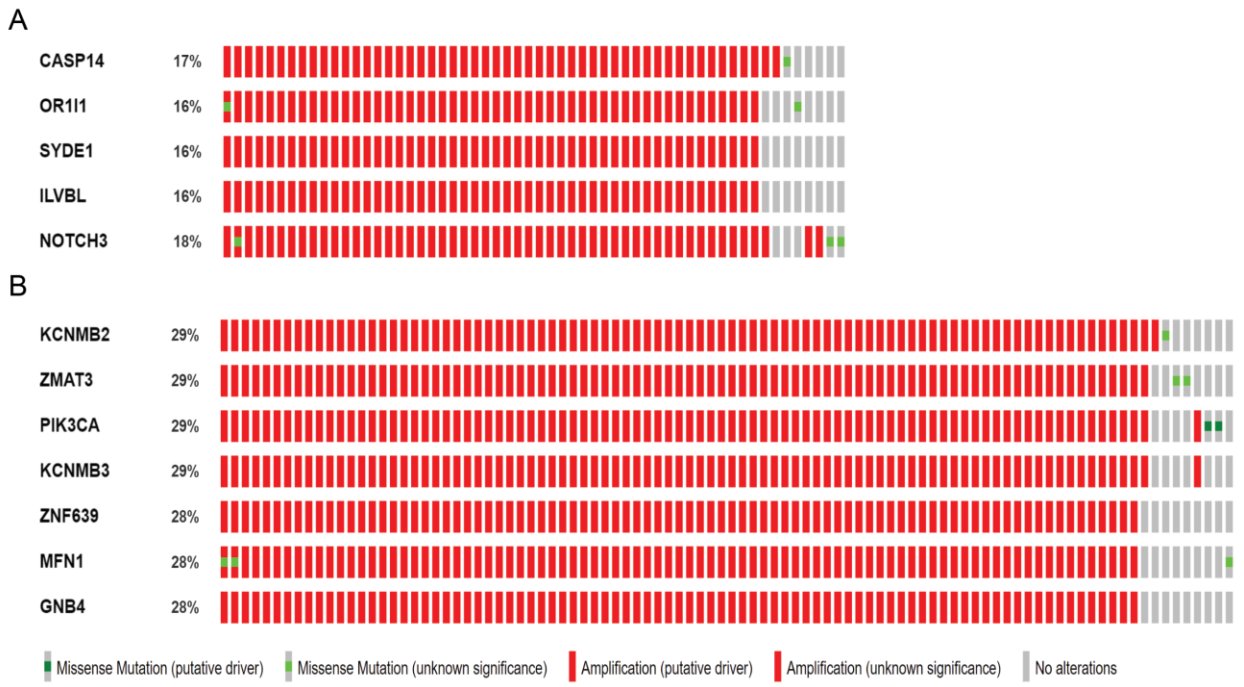


Figure 34 Visualization of somatic mutation profiles by OncoPrint on CbioPortal (Gao et al. 2013). Both *NOTCH3* (A) and *PIK3CA* (B) have almost identical copy number alteration profiles with their neighboring genes. It is difficult to distinguish *NOTCH3* and *PIK3CA* (drivers) from their neighbors (passengers) with data-driven approaches.

6.2 Material and Method

6.2.1 Corpus collection and text preprocessing

Four types of biomedical texts were collected: (1) Title and abstract of biomedical articles from PubMed; (2) Gene summary provided by RefSeq (updated in 2012); (3) Gene Reference into Function (GeneRIF) from NCBI where each gene must be covered with more than 5 literature excerpts; (4) Description of Gene Ontology (GO) terms in the category of biological process.

For all the text corpus, we removed numbers, multiple white spaces, words with less than 2 characters, and special symbols. All the words were changed to lower case and Porter-stemmed.

6.2.2 Semantic representation

We combined the corpus of GeneRIF and RefSeq to compute tf-idf. A score was obtained for each word by multiplying tf-idf with the total frequency of the corresponding word. Words with a top 30000 score remain in the corpus. Since each entry in GeneRIF and RefSeq corresponds to a gene, the corpus was reorganized such that each gene has a document containing all the unique words from both GeneRIF and RefSeq. Finally, we constructed the <gene, word> pairs if a word appears in the gene document. These <gene, word> pairs were used as the semantic representation for genes.

6.2.3 Word2Vec

We used Word2vec implemented in genism (Rehurek and Sojka 2010). Among all the variants of word2vec, we chose the skip-gram approach with negative sampling. Given a word in a sentence, the skip-gram model predicted its context. Each word was encoded as a one-hot vector and then linearly transformed into an embedding vector. Let the probability of word *i* in the context of word *j* be:

$$\Pr(w_i|w_j) = \frac{\exp\,(v_i^T v_j)}{\sum_{j' \in window} \exp\,(v_i^T v_j)}$$

where w_i is the ith word in the corpus, and v_i is its corresponding embedding vector. Embedding vectors were learned by optimizing $\Pr(w_i|w_j)$ for pairs of co-occurring w_i and w_j in the literature or <gene, word> pairs. To minimize $\Pr(w_i|w_j)$ when w_i and w_j do not co-occur, random samples from the vocabulary will be used as the negative samples instead of going through the whole vocabulary in each iteration.

6.2.4 Evaluation

We evaluated whether our gene embedding can improve functional module identification. To do this, we first constructed the mutual exclusivity network. It is well known that genes perturbing the same pathway often avoid co-mutation in tumor samples. This phenomenon is called mutual exclusivity (Yulan Deng et al. 2019; Remy et al. 2015). Mutual exclusivity has been widely used to determine whether genes belong to the same pathway (C. A. Miller et al. 2011; Ciriello et al. 2013; Canisius, Martens, and Wessels 2016; J. Zhang and Zhang 2018). We performed mutual exclusivity analysis on 579 somatic mutation profiles of TCGA (Tomczak, Czerwińska, and Wiznerowicz 2015) ovarian tumor samples downloaded from the Xena browser (Goldman et al. 2020). Genes that mutated in less than 5% of samples or absent in the embedding space were removed. 4718 genes remained. We adopted the classic one-sided Fisher exact test to compute pairwise mutual exclusivity among genes. An edge is added between two genes if their p value is less than 0.1. We also calculated the odd ratios as the edge weight:

$$OR = (A + 0.5)(D + 0.5)/(B + 0.5)(C + 0.5)$$

where A, B, C, and D are the four elements in the contingency table:

# of samples in different conditions	Gene j mutated	Gene j not mutated
Gene i mutated	A	C
Gene i not mutated	B	D

After constructing the mutual exclusivity network, we used cosine similarity of gene embedding as the edge weights for the network. Then we perform clustering on the network with and without the edge weights so as to evaluate whether gene embedding can improve functional module identification.

As for the clustering algorithm, we used isolation clustering proposed in our previous work (Liang et al. 2019). This algorithm first transforms the network into a Markov transition matrix. Then it computes the probability of node *i* visiting node *j* in 5 steps as the connectivity matrix *C*. Finally, clusters were identified with locally maximal isolation:

$$isolation = \frac{\sum_{i \in R} \sum_{j \in R} C_{ij}}{\sum_{i \in R} \sum_{j \in G} C_{ij} + \sum_{i \in G} \sum_{j \in R} C_{ij}}$$

where C_{ij} is the element at the *i*th row, the *j*th column of *C*, *R* is the subset of nodes in the cluster, and *G* is all the nodes in the graph.

6.3 Experimental Results

After preprocessing, our corpus contains (1) Title and abstract of 8,514,630 biomedical articles; (2) RefSeq gene summary that covered 15764 genes; (3) 1,198,717 GeneRIF excerpts covering 23837 genes; (4) 30835 Gene Ontology (GO) term definition in the category of biological

process. Corpus from RefSeq and GeneRIF was further transformed to around 1 million <gene, word> pairs. The gene embedding vectors were obtained after training the Word2vec model and converted to gene symbols. These 300-dimensional embedding vectors covered 31845 genes.

6.3.1 Word embedding has captured similarity of concepts from literature

Since the accuracy of gene embedding depends on the word embedding, we first explored the word embedding directly learned from literature. We computed cosine similarity between a query (e.g. “PTEN”) and all the other words in the embedding space. Top 300 most similar words for each query were visualized with the wordcloud package in Python. As shown in Figure 35, gene symbols were most similar with their closely related genes, pathways, or synonyms. For example, “pi3k” and “akt” were among the most similar words when querying “PTEN”. It is well known that these three proteins collaborated in *PI3K/AKT* signaling pathway (Castellino and Durden 2007).

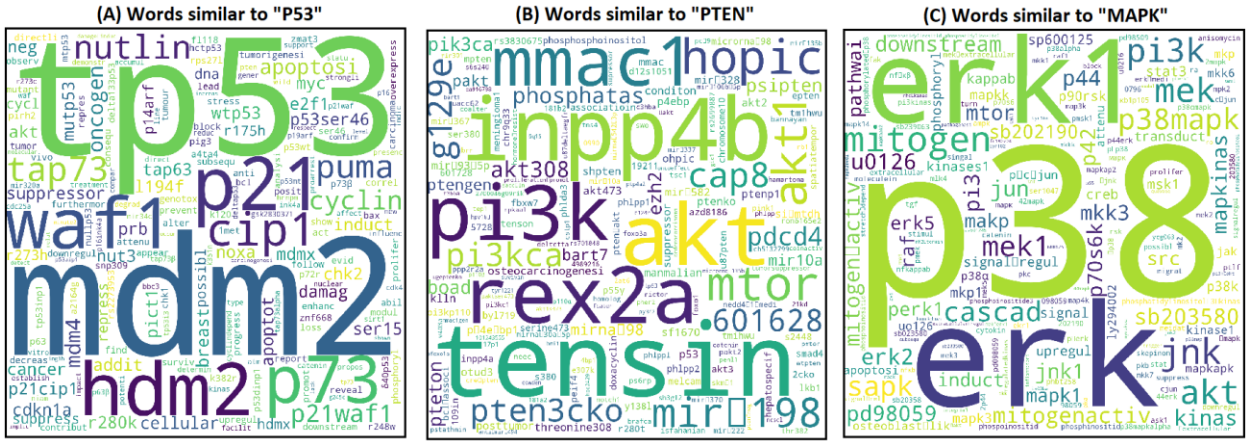


Figure 35 Wordcloud visualization of word query to the word embedding. The bigger the font size, the more similar the word is to the query word.

6.3.2 Gene embedding was consistent with current knowledge of biological pathways

We downloaded gene sets of biological pathways from WikiPathways (Slenter et al. 2018). Gene sets with 3 genes or less were removed. For each gene in a pathway, we computed:

$$Ratio_i = \frac{\sum_{j \in P} cosine(g_i, g_j) / |P|}{\sum_{k \in R} cosine(g_i, g_k) / |R|}$$

where *i* is a certain index of genes, g_i is the embedding vector for the *i*th gene, *P* is a set of gene indices within the same pathway except the *i*th gene, *R* is a set of 10 randomly selected gene indices, and |*P*| is the number of indices in *P*. Note that, the *i*th gene was excluded from the pathway *P*. Clearly, the higher the ratio, the closer pathway members are located in the embedding space. We computed this ratio for every gene and generated the distribution shown in Figure 36. 91.81% of genes have a ratio larger than 1, while a similar approach, mut2vec (S. Kim et al. 2018), has 82.78% larger than 1. It implied that neighboring genes in the embedding space were more likely to collaborate in biological processes.

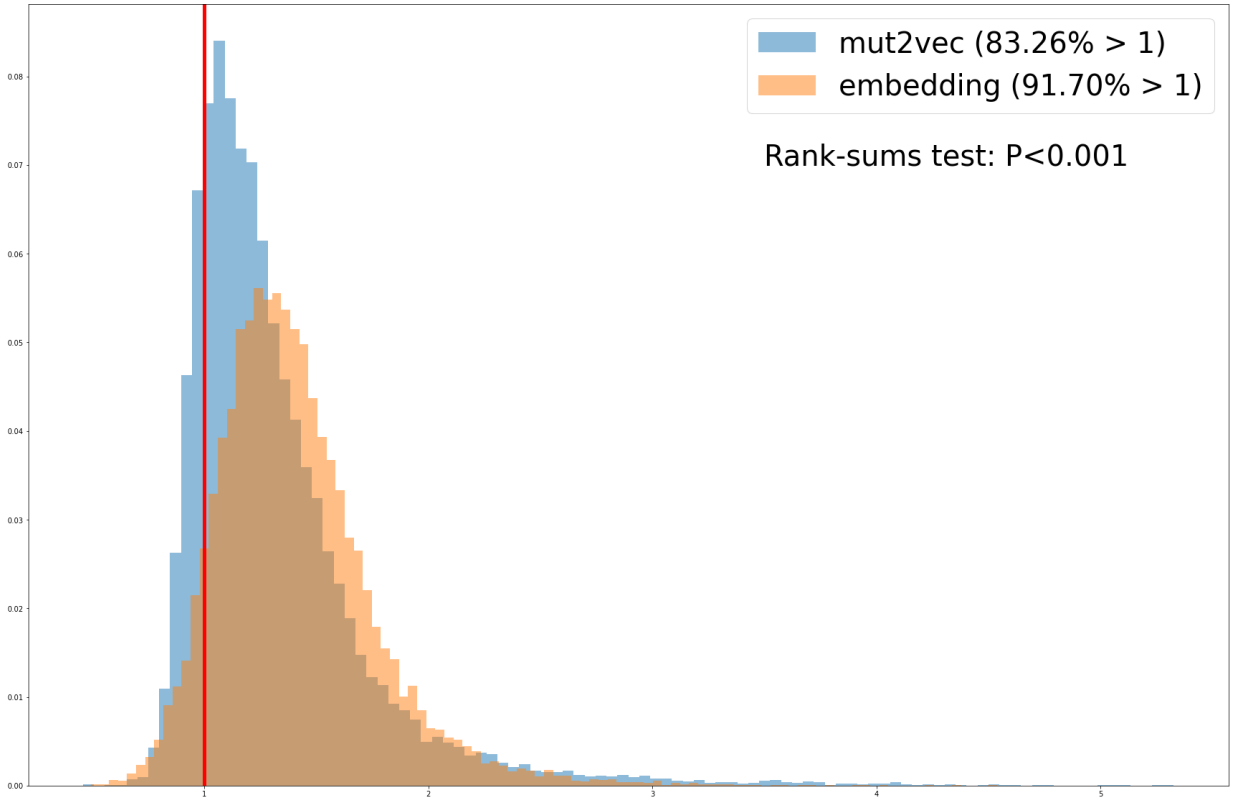


Figure 36 The distribution of cosine similarity ratios for all the genes based on Wikipathway. The embedding vectors of 91% of genes are more similar with its pathway members than random genes, while mut2vec has 83%.

Take the example of NOTCH3 and *PIK3CA* illustrated before (Figure 34), though it is difficult for statistical methods to distinguish driver genes from the genomics profiles, our gene embedding showed that *PIK3CA* was much closer to *PTEN* signaling members than its neighboring genes (Table 7 and Table 8). Although not as obvious, *NOTCH3* is also distinguishable from its neighbors.

Table 7 Cosine similarities of our embedding vectors between *PTEN-PIK3CA* pathway members and hotspot mutation neighbor of *PIK3CA*. Each row is a neighbor gene. Each column is a pathway member. Compared with its neighbors, *PIK3CA* is the most similar gene with all the pathway members.

	PTEN	*STK11*	*AKT1*	*AKT2*	*TSC1*	*MTOR*	*EGFR*	*ERBB2*	*ERBB3*	*FGFR1*	*KRAS*	*NRAS*	*BRAF*
KCNMB2	0.388	0.145	0.314	0.237	0.190	0.358	0.280	0.213	0.236	0.381	0.105	0.183	0.150
ZMAT3	0.499	0.344	0.411	0.284	0.301	0.426	0.371	0.333	0.333	0.367	0.351	0.362	0.262
PIK3CA	**0.627**	**0.627**	**0.690**	**0.366**	**0.541**	**0.650**	**0.688**	**0.656**	**0.590**	**0.539**	**0.694**	**0.617**	**0.490**
KCNMB3	0.475	0.222	0.370	0.214	0.209	0.422	0.350	0.288	0.240	0.402	0.183	0.189	0.176
ZNF639	0.603	0.279	0.300	0.353	0.309	0.330	0.370	0.409	0.333	0.525	0.249	0.283	0.279
MFN1	0.347	0.359	0.392	0.172	0.353	0.392	0.317	0.299	0.276	0.275	0.334	0.275	0.191
GNB4	0.462	0.307	0.292	0.297	0.335	0.317	0.347	0.388	0.368	0.412	0.260	0.306	0.220

Table 8 Cosine similarities of our embedding vectors between *NOTCH* signaling pathway members and hotspot mutation neighbor of *NOTCH3*. Each row is a neighbor gene. Each column is a pathway member. Compared with its neighbors, *NOTCH3* is the most similar gene with all the pathway members.

	JAG1	*JAG2*	*MAML2*	*MAML3*	*DLL1*	*DLL3*	*DLL4*	*HES1*
CASP14	0.498766	0.455305	0.304931	0.419004	0.348857	0.416664	0.332073	0.371251
OR1I1	0.325886	0.321224	0.201185	0.346904	0.199607	0.350854	0.205338	0.187183
SYDE1	0.440563	0.36717	0.194305	0.389681	0.315622	0.381245	0.375707	0.254521
ILVBL	0.274814	0.23507	0.150193	0.205635	0.132573	0.276393	0.096576	0.215826
NOTCH3	**0.501024**	**0.480401**	**0.385019**	**0.425642**	**0.426736**	**0.444915**	**0.429611**	**0.479549**

6.3.3 Gene embedding can distinguish cancer drivers from passengers

We further evaluated whether cancer drivers from the same type of cancer will be closer in the embedding space. We collected the list of drivers for different types of cancer from IntOGen (Martínez-Jiménez et al. 2020). As shown in Figure 37, cancer drivers are located closer to each other in our embedding space than mut2vec.

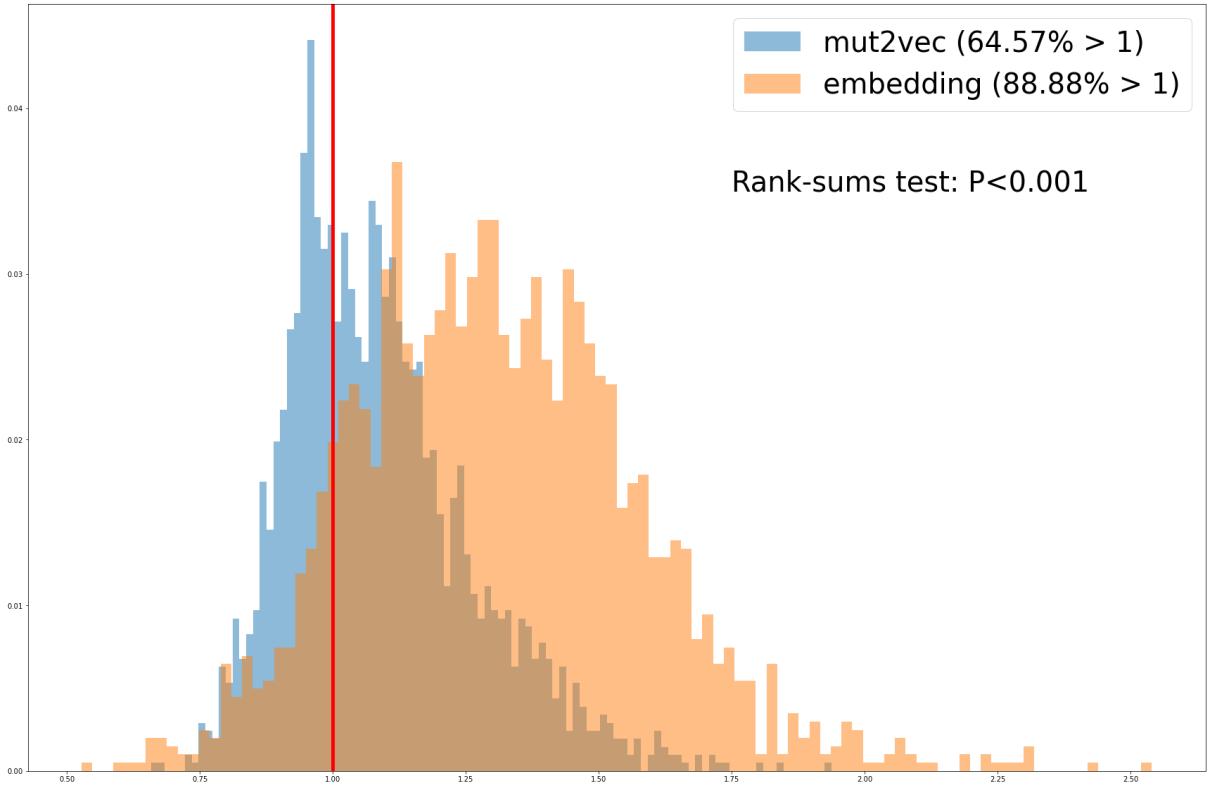


Figure 37 The distribution of cosine similarity ratios for all the genes based on Intogen. Embedding of 88.88% of genes are more similar with its pathway members than random genes, while mut2vec only has 64.57%.

6.3.4 Gene embedding improved functional module identification

A mutual exclusivity network with edge weights of odd ratios was constructed. We used isolation clustering to identify functional modules in this mutual exclusivity network. Then we followed our previous approach (Liang et al. 2019) to construct a two-layer multiplex by adding edge weights of cosine similarity from the gene embedding. In the gene embedding layer, each gene is only connected to neighbors with top 1% cosine similarity. Overall, biological pathways were more enriched in the modules identified from the multiplex (Figure 38). For example, a module from the multiplex overlaps with the MAPK signaling pathway over 13 genes, while the best hit from the single layer network overlaps with 4 genes (Table 9). We also performed

enrichment analysis on Gene Ontology (GO) annotation in the category of biological process. The result (Fig. 6) was consistent with pathway enrichment analysis.

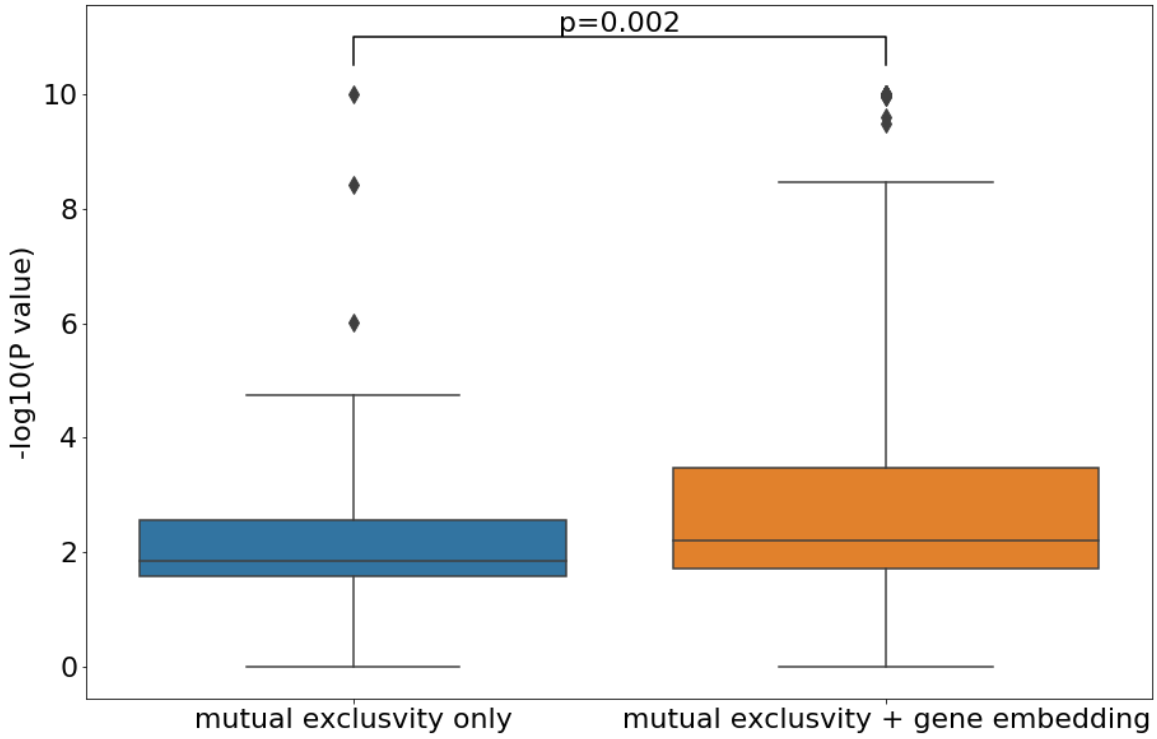


Figure 38 Distribution of log P values from pathway enrichment analysis of functional modules. The smallest p values across all pathways were selected for each functional module. Ranksums test showed that gene embedding significantly improved the performance (p=0.002).

Table 9 Examples of biological pathways where the best hits from multiplex were much more accurate than that from mutual exclusivity alone.

	pathway size	Mutual exclusivity		Multiplex with gene embedding	
		overlap/module size	-log10(P value)	overlap/module size	-log10(P value)
DNA IR-damage and cellular response via ATR	29	1/11	1.14	19/82	26.32
Globo Sphingolipid Metabolism	6	2/41	2.88	6/45	12.01
MAPK Signaling Pathway	70	4/118	0.89	13/23	17.7
Metapathway biotransformation Phase I and II	46	6/53	4.75	21/68	26.49
Novel intracellular components of RIG-I-like receptor (RLR) pathway	16	2/67	1.60	8/35	12.97
Nuclear Receptors Meta-Pathway	86	4/37	2.21	15/68	11.69
Oxidation by Cytochrome P450	20	3/53	2.75	10/68	13.05

6.4 Contribution and Limitations

This study proposed to use semantic representations of genes to learn gene embedding from biomedical literature. To our knowledge, no studies have attempted to construct gene embedding from literature alone. A previous study (S. Kim et al. 2018) indicated that a combination of mutation profiles, literature, and PPI is required to identify cancer driver mutations. We conjected there are two major obstacles. First, it is difficult to extract the concept of genes from text. Although genes may be directly mentioned by their symbols (e.g. *PTEN*, *MAPK*), authors in different articles may choose different terms or different levels of concepts. For example, researchers may refer to the fibroblast growth factor family instead of enumerating relevant members in that family (N. S. Patel et al. 2013). The other issue is that biomedical literature payed more attention to already well-known genes. It leads to difficulty in learning embedding for less well-known genes directly and accurately. However, the example of *PIK3CA* in Introduction

showed that our approach, with literature alone, can distinguish driver mutations from its passengers. Experiment results also showed that our gene embedding space was consistent with current knowledge of pathways. These results indicated our gene embedding had captured the latent knowledge about genes' functional relationships from biomedical literature.

In the future, we need to refine the semantic representation of genes. Although this study showed that GeneRIF and RefSeq provided effective representations for genes, their maintenance requires manual summary and annotation. Furthermore, they are still limited by our current knowledge of genes. We need to devise a way to integrate literature with high-throughput data before or during the learning process of gene embedding.

Another issue is that one single similarity metric is insufficient to capture the multifaceted nature of functional relationships among genes (e.g., protein moonlighting (Espinosa-Cantú et al. 2018)). Similarly, researchers in the NLP community have proposed various approaches of multi-sense word embedding (Tian et al., n.d.; Jain et al. 2019) to handle context-dependent semantics. In the future, we may need to develop a multi-sense model to capture different functional contexts of genes.

7.0 Peak2vec Enables Inference of Transcriptional Regulation from ATAC-seq

This chapter focus on the development of a novel algorithm to detect TF motifs from chromatin accessibility profiles. The resulting method can identify the motif and its corresponding binding sequences. Application to the consensus peak sets in TCGA liver cancer samples showed that we are able to identify several transcription factors related to liver tissues.

7.1 The Difficulty of Identifying Cis-regulatory Elements from Chromatin Accessibility Profiles.

Since first described in 2013 (Buenrostro et al. 2013), ATAC-seq has gained particular popularity. The exponential increase of ATAC-seq curated datasets indicates its value in a wide spectrum of biological studies (Yan et al. 2020). In particular, the project of TCGA has profiled 410 tumor samples with ATAC-seq to interrogate the transcriptional regulation (Corces et al. 2018). Unlike TF ChIP-seq, activities of multiple transcription factors are captured in ATAC-seq, offering a great opportunity to systematically analyze gene regulation in different conditions.

Currently, the major approach to analyze TF activities in ATAC-seq is to utilize the footprinting pattern (Zhijian Li et al. 2019; Karabacak Calviello et al. 2019). Although these methods have yielded insights into transcriptional regulation, only one fifth of TF motifs show protection from the DNA cleavage (Baek, Goldstein, and Hager 2017). The results can be limited by the range of applicable TFs.

Computational researchers have utilized the motif information to complement the footprinting pattern, such as TFEA (Rubin et al. 2021) and MEME-centrimo (Bailey and Machanick 2012). TFEA performed TF enrichment analysis on the peak regions in the differential analysis to identify TFs causally responsible for biological differences between samples. MEME-centrimo compared the cleavage counts of predicted binding sites with the cleavage events around them. However, these methods relied on curated TF motifs, which may be incomplete, inaccurate, and inconsistent across tissues and conditions.

On the other hand, deep learning has been successfully applied to de novo identification of TF motifs in TF ChIP-seq experiments, with the best performance among different approaches (Alipanahi et al. 2015). However, despite various improvements (Quang and Xie 2019; Park et al. 2020; Tareen and Kinney 2019) and expanded application (Ghanbari and Ohler 2020; J. Zhou and Troyanskaya 2015) over the past five years (Koo and Ploenzke 2020), we have rarely seen any deep learning applications to identify TF binding activities in the chromatin accessibility profiles. This is probably because a conventional deep learning algorithm may not be a suitable tool when there are multiple TFs mixed in the set of sequences. This will be demonstrated in the simulation experiment.

Here we presented a novel variant of convolutional neural network (CNN), Peak2vec, to perform de novo inference of the coregulation among enriched regions by constructing the embedding space from the peak sequences in ATAC-seq. We hypothesized that by modifying existing deep learning algorithms, it is possible to uncover various TF binding specificities and downstream regulon within the chromatin accessibility profiles without relying on current knowledge.

7.2 Inference of Sliding Window Multinomial Mixture via Modified Convolution Neural Network

The overall architecture of Peak2Vec is illustrated in Figure 39. For each fix-length DNA sequence, we generated the one-hot encoding matrix for both strands. The binary matrix is scanned by multinomial convolution of different sizes. Normalized convolution output is concatenated together as the embedding vector. The embedding space of regulatory regions is learned by classifying whether a sequence was a peak region or a random sequence. Features with positive correlation to the real peak regions were selected as the embedding vector for enriched sequences.

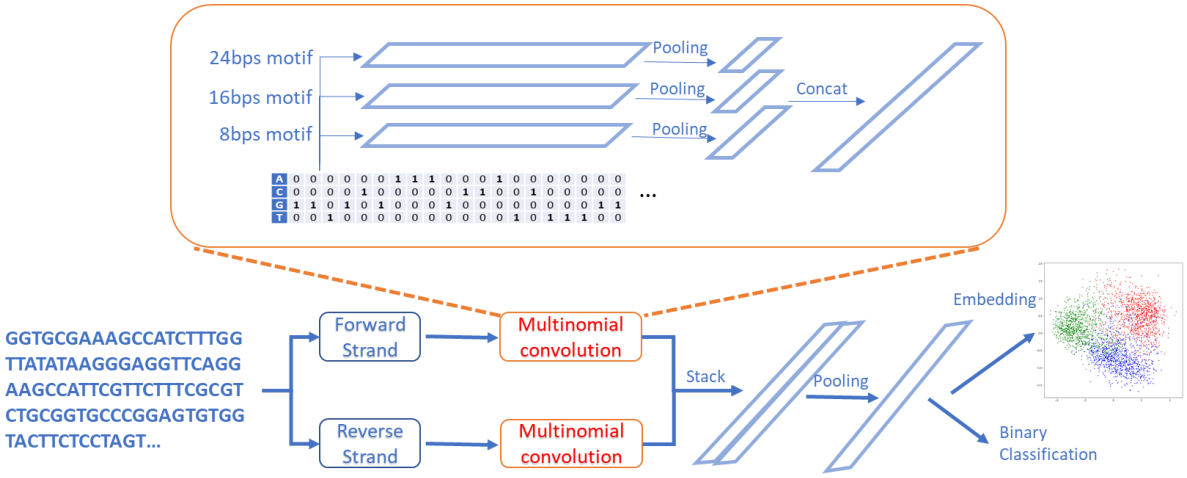


Figure 39 The architecture and workflow of Peak2vec. DNA sequences are transformed into one-hot encoding array. Three layers of multinomial convolution are applied to the array. Three max pooling layer are applied to each convolution output to generate feature representation that are concatenated for binary classification. After training the model for binary classification, features with positive coefficients towards peak region prediction were selected to construct the embedding space of peak sequences.

7.2.1 Multinomial convolution kernel

In this section, we show how one dimensional (1D) convolution connects to multivariate multinomial mixture in theory and introduce the basic convolution unit in Peak2Vec. First of all, 1D convolution is basically applying a sliding window (named "kernel") over an "image" along the first dimension. Suppose the input "image", *C*, is an N by M matrix, the kernel, *W*, is an N by K matrix, then the output (named "feature map"), C^*, would be a vector of length $(M - K + 1)$. The relationship between the input and output is:

$$C_i^* = b + \sum_{k=1}^{K} W_{.k} \cdot C_{.i+k-1}$$

where $W_{.k} \cdot C_{.k+i-1}$ is the inner product between the *k*th column of W and the *k+i-1*th column of C, *b* is the scaler for bias terms. This is the same one-dimensional convolution as in DeepBind or DeepSEA. Usually, multiple sliding windows will be applied to the "image". Hence the feature map, C^*, expands to a matrix of L by $(M - K + 1)$, given that L is the number of kernels. And the bias term, *b*, expands to a vector of length L.

However, in this study, we apply the Softmax function to each column of W such that each column sums up to one. In this way, the *l*th kernel, W_l, can be interpreted as the parameterization of a multivariate multinomial distribution Z_l. Hence *l*th component of multinomial distribution Z, the *k*th column of W, $W_{.k}$, represents that

$$W_{.k} = P(X_k|Z_l)$$

In addition, we also perform Softmax on the bias term. So b_l represent the prior probability of Z_l. Furthermore, we take the log of W and b elementwise before applying convolution. Considering the computation of C_i^*, we have

$$C_i^* = \log\left(P(Z_l)\right) + \sum_{k=1}^{K} \log\left(P(X_k|Z_l)\right) \cdot C_{.i+k-1}$$

$$= \log\left[(P(Z_l) \prod_{k=1}^{K} (C_{.i+k-1}|Z_l)\right]$$

$$= logP(C_{.[i,i+K-1]}, Z_l)$$

The output C* would become the log joint probability of the i to i+K-1 columns of C and the *l*th kernel. Then we further perform the Softmax function over each column of C*. Clearly, C_i^* after Softmax dictates the posterior probability of Z_l.

$$Softmax(C_i^*) = \frac{P(C_{.[i,i+K-1]}, Z_l)}{\sum_{l \le L} P(C_{.[i,i+K-1]}, Z_l)} = P(Z_l|C_{.[i,i+K-1]})$$

At this point, it is clear that the specialized convolution kernel here is analogous to performing the EM algorithm for multivariate multinomial mixture on each sliding window. The forward computation is inferring the membership of samples, the E step. The backward propagation is learning distribution parameters and priors, the M step. This is only an analogy because of two major differences: (1) we do not identify the maximum likelihood in each iteration of backpropagation as in the real M step; (2) the distribution parameters are shared across all sliding windows. Still, this theoretical connection may provoke insights in theoretical analysis and deep learning models development.

In summary, we applied log Softmax function to the kernel weights and biases before convolution. After convolution, we applied Softmax to the feature map. In this way, kernel weights (W) and biases (b) can be interpreted as components of multivariate multinomial mixture and corresponding priors. The feature map can be regarded as the posterior probability that the scanned section of C is sampled from the *l*th kernel.

In addition, as shown in Figure 39, the feature map of the first and the second convolution layer directly feed to the next convolution layer without pooling. This can reserve position

information for the next convolution. However, only the first convolution layer is interpretable because it is directly connected to the one-hot encoding matrix of DNA sequences. Although the other convolution layers cannot be interpreted as position weight matrix (PWM), they enable the model to capture complicated regulatory sequence patterns.

7.2.2 Max pooling layer

The max pooling layer outputs the maximum value of every row of C^*. This means that a kernel is activated if it has high posterior probability in any subsequence of the sample. Output of this layer become features for the classification task. The feature size is the number of kernels in the corresponding convolution layer.

7.2.3 Model training

All the outputs from pooling layers are concatenated as an embedding vector U. The embedding was trained by binary classification on whether the sequence contains regulatory elements or not. Suppose the binary label is a vector Y of length S (sample size), then the objective is the cross entropy between Y and the estimated Y.

$$cross\ entropy = \sum_{s \le S} \left[Y log \hat{Y} + (1 - Y) \log\left(1 - \hat{Y}\right) \right]$$

$$\hat{Y} = Sigmoid(U \cdot P + b')$$

where Sigmoid is the sigmoid function, P is the weights, b' is the bias term. The model is trained with gradient-based optimization, specifically RMSprop.

7.2.4 Handling the reverse complement strand

The reverse complement also went through the variable length multinomial convolution and produced a concatenated embedding vector. As shown in Figure 39, we selected the bigger value from both strands for each feature and constructed the final embedding vector for training.

7.2.5 Data preprocessing

The set of positive sequences were constructed by extending certain length upstream and downstream to the peak summits of ATAC-seq. In the simulation experiment, the extension length was 100 base pairs. In real data analysis, the extension length was 250. We generated a negative sequence with the same dinucleotide frequency from the positive sequence. A sequence of length M is then transformed into 4 by M matrix via one-hot encoding.

7.2.6 Embedding vector interpretation

We used a Gaussian mixture to identify the clusters of the embedding vectors. For each cluster, samples are divided into two groups, samples in the cluster and samples in other clusters. We then perform Wilcoxon test to identify motif features with significantly higher values in the cluster than outside the cluster. The feature with the smallest p value is regarded as the signature motif for the corresponding cluster.

7.3 Results

7.3.1 Simulation experiment

To validate Peak2Vec's capability of identifying different structural regularities from enriched DNA sequences, we collected TF ChIP-seq of JUN, CTCF, and POLR2A from the HepG2 cell line in the ENCODE project. We selected the top 1000 peaks from each TF in terms of q values. 6000 negative sequences were also generated. Peak2Vec was trained on 9000 samples in total. Peak2vec has three layers of multinomial convolution. Each layer has 256 kernels.

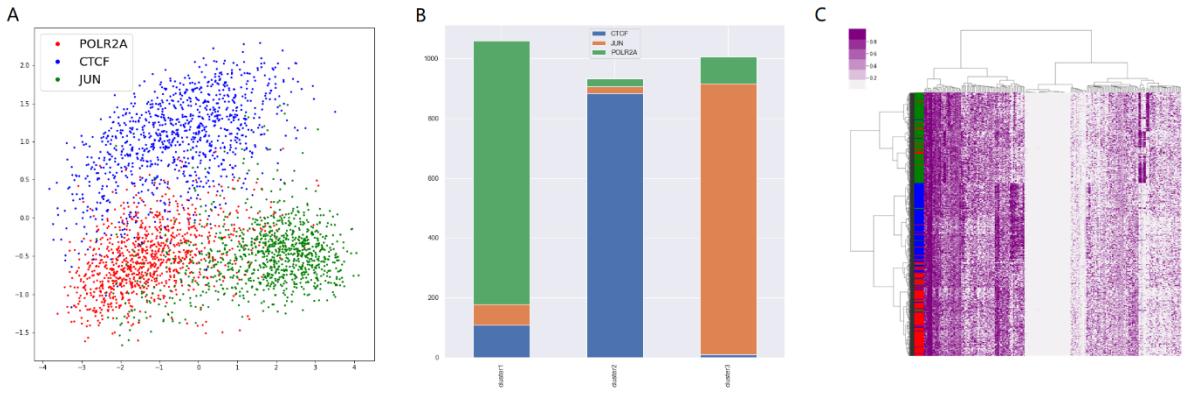


Figure 40 PCA visualization of the embedding space (A), Gaussian mixture (B), and hierarchical clustering (C) of the embedding vectors.

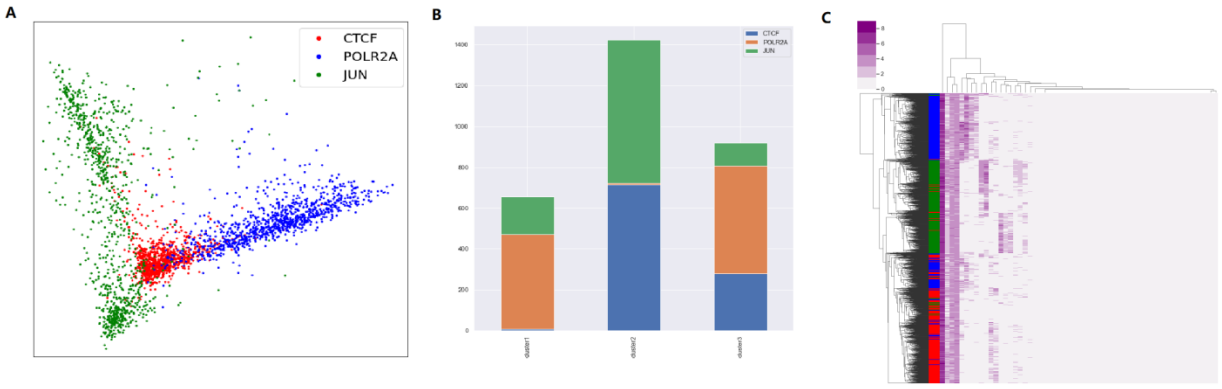


Figure 41 Embedding space generated by conventional convolutional neural network (CNN). PCA visualization of the embedding space (A), Gaussian mixture (B), and hierarchical clustering (C) of the embedding vectors.

We then performed conventional clustering algorithms on the embedding space. As shown in Figure 40, the three TF can be identified in different clusters despite the clustering algorithm. In subsequent real data analysis, we continued to use Gaussian mixture to identify coregulation modules in ATAC-seq profiles.

To demonstrate the necessity of multinomial convolution, we also implemented a convolutional neural network (CNN) model with conventional convolutional kernels and ReLu activation as described in DeepBind (Alipanahi et al. 2015). Other than the kernels, our implemented CNN has the same architecture and the same training procedure as the peak2vec. The results are shown in Figure 41. Clearly, CNN fails to distinguish different TFs in the embedding space.

We further extract the signature motif for the three clusters. As shown in Figure 42, the motif signature is similar to the motif in JASPAR. In the case of CTCF, “CCTCC” is similar to “CCACC” in JASPAR. As for JUN, the first four base pairs match nicely to the motifs in JASPAR. We cannot find motifs for POLR2A in any database. JASPAR also cannot find any high-quality

motifs for POLR2A. Still, Figure 42 showed that the embedding of POLR2A peak sequences are distinguishable from the other two.

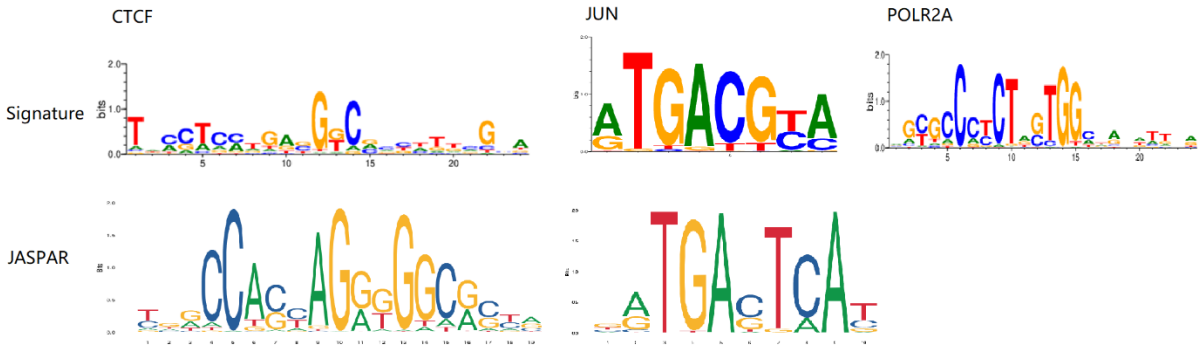


Figure 42 Comparison between signature motif from Peak2vec (above) and motifs from JASPAR (below). There is no curated motifs for POLR2A.

7.3.2 Application to ATAC-seq profiles of liver cancer

We downloaded liver cancer type-specific peak signals from Xena Browser. 26513 peaks with signal values above 10 have been selected. All the sequences have a fixed length of 500 bps. We then generated 53026 negative sequences following the same dinucleotide frequency of the peaks. The model is significantly larger than that in the simulation experiment, with 1024 kernels in each layer.

After extracting the embedding vectors from Peak2vec, we conducted Gaussian mixture on the embedding vectors. The number of clusters was set to 80, as determined by the Akaike information criterion (AIC). The signature motifs were extracted to compare with the motifs curated in JASPAR vertebrate dataset. Signatures extracted from clusters seem to be redundant for

many unique clusters. As shown in Figure 43, the signature motif of some clusters can be matched to known motifs quite nicely.

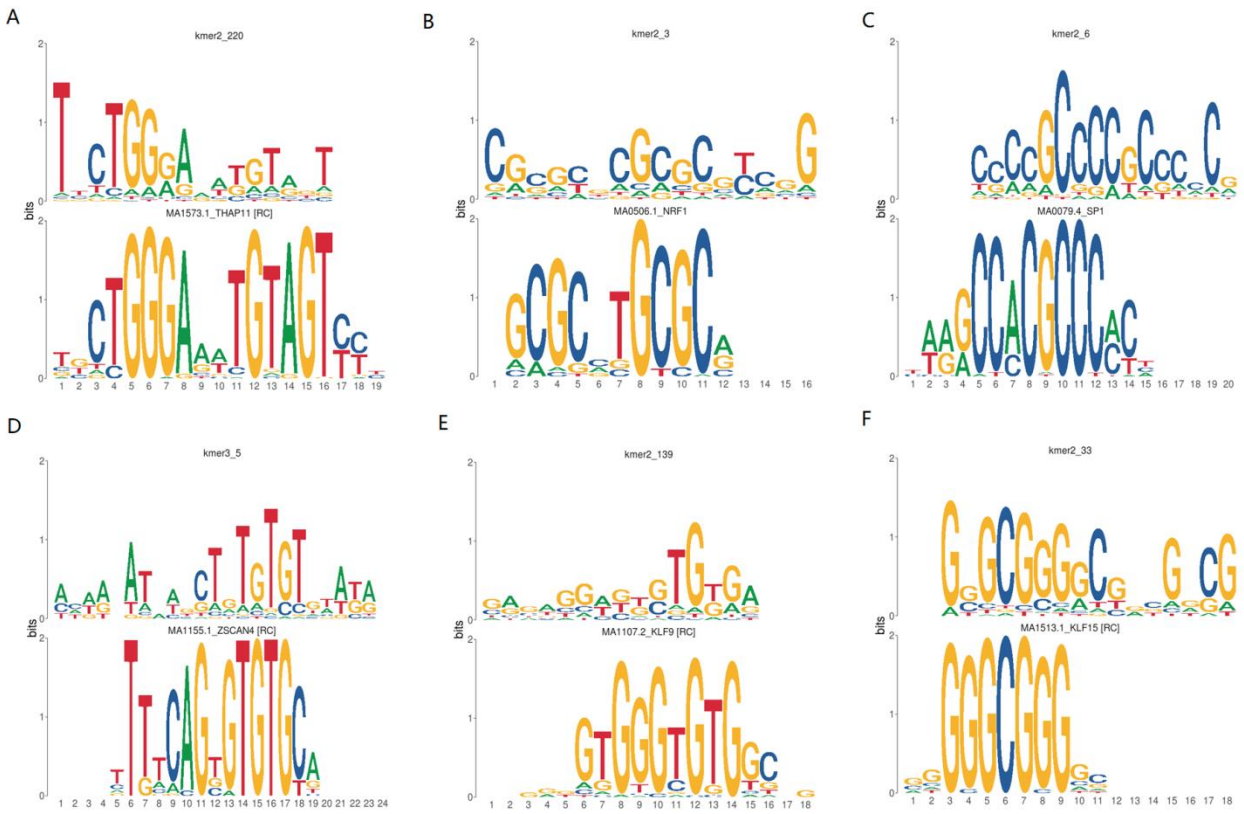


Figure 43 Signature motifs that can be matched to known motif in JASPAR. For each subplot, the upper part is the signature and the bottom part is the known motif. (A) is matched to the transcription THAP11; (B) is matched to the transcription factor NRF1; (C) is matched to SP1; (D) is matched to ZSCAN4; (E) is matched to KLF9; (F) is matched to KLF15.

In addition, we searched through the multinomial convolution kernels other than cluster signatures to identify meaningful motifs. The search was taking the intersection of motif matching results and gene set enrichment analysis results. Motif matching was performed in the same way as above. As for gene set enrichment analysis, we extracted peaks with top 1000 score for each

kernel and feed them to ChEA3 for enrichment analysis. As shown in Figure 44, brute force searching also revealed other motifs missed by the clustering analysis.

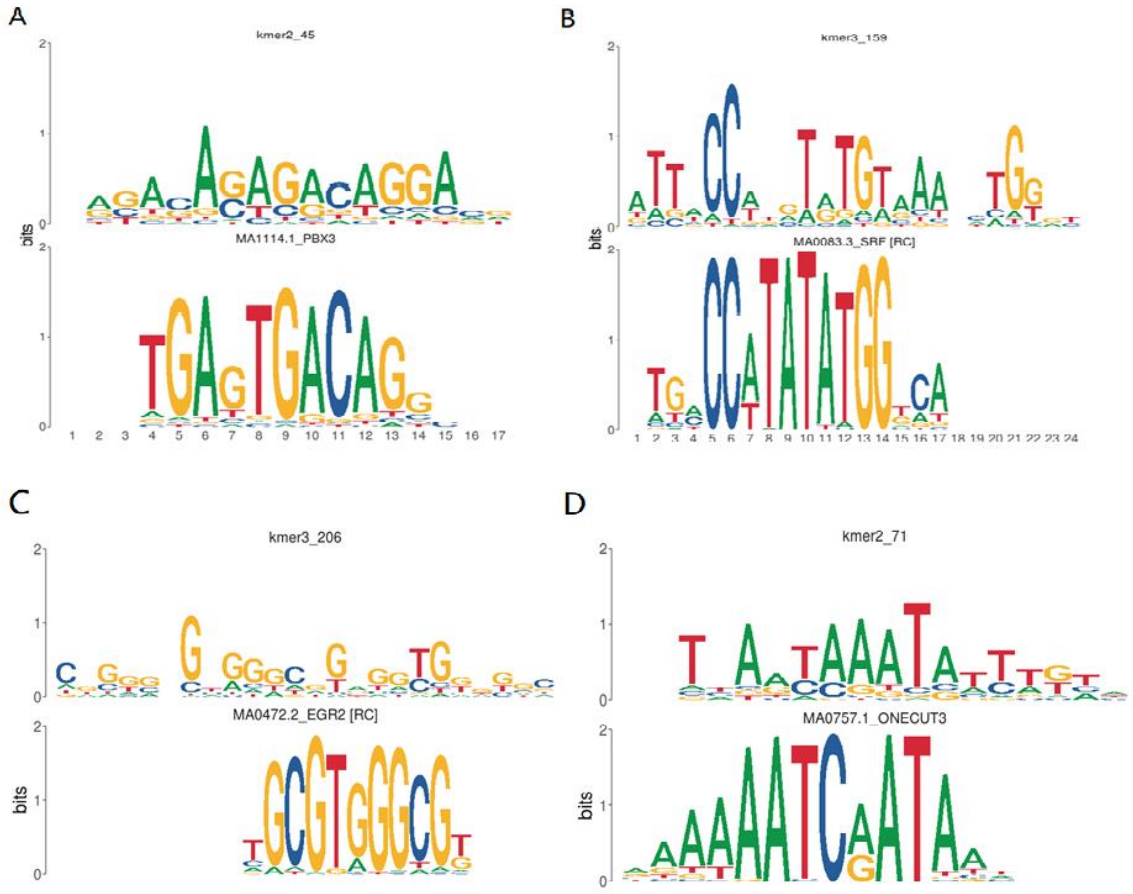


Figure 44 Motif identified through searching all the kernels. (A) was matched to PBX3 with enrichment analysis FDR=5.37e-6; (B) was matched to SRF with ENCODE enrichment analysis FDR=0.014; (C) was matched to EGR2 with ENCODE enrichment analysis FDR=0.0475; (D) was matched to ONECUT family with ReMAP enrichment analysis FDR=8.34e-6.

We also found that although some convolution kernels had not found a significant match to the curated motifs, they may still be biologically meaningful. Examples in Figure 45 showed

that some kernels might capture a pair of TFs at work (Figure 45A) or were not distinguishable enough (Figure 45B).

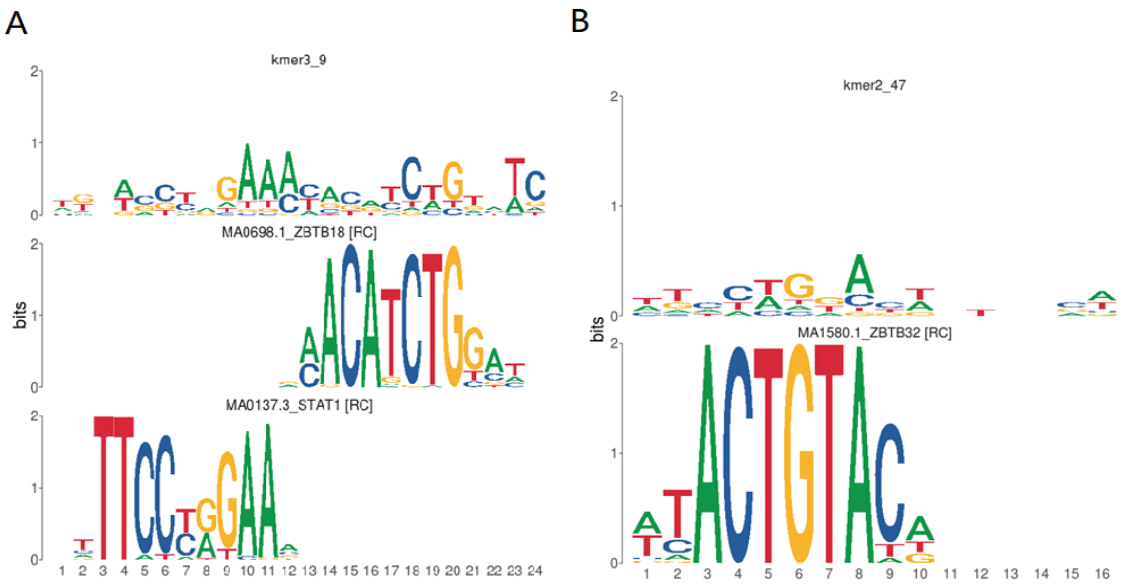


Figure 45 Learned motifs that only found a weak match in the JASPAR database. Kernel in (A) consists of two known TF motifs, ZBTB18 and STAT1. Kernel in (B) can be matched to ZBTB32 if only the top nucleotide was considered.

7.4 Contribution and Limitations

In this study, we presented Peak2vec, a novel algorithm that can identify ATAC-seq peaks regulated with the same TF while providing the corresponding signature motif. To our knowledge, this is the first model to infer transcriptional regulation relying solely on ATAC-seq. The idea of employing similar deep learning algorithms to study epigenomics has been quite popular along the past decade (Alipanahi et al. 2015; J. Zhou and Troyanskaya 2015; Lal et al. 2021). We contributed to the methodology by adapting the convolution kernels so that it can identify multiple TFs within

the chromatin accessibility profiles. Our results show that peak2vec is effective in analyzing sequences with diverse binding specificity.

Peak2vec is also easier to interpret since a multinomial convolution kernel directly represents a position weight matrix (PWM). When applying deep learning to extract motifs from TF ChIP-seq, previous research (Alipanahi et al. 2015) needs to align the sequences according to activated position in the feature map.

Still, there is much room for improvement. One particular issue is that enrichment analysis for coexpression module did not overlap with motif comparison much. It is probably due to two reasons: (1) peaks were mapped to the nearest genes. This may not be accurate, especially for peaks in the remote regions. They may be trans-regulatory elements; (2) TFs may have different binding sites in different tissues. In the future, we need to utilize the RNA expression information to determine the correlation between peaks and genes, which has been partially employed in the TCGA project (Corces et al. 2018).

Another limitation is that although Peak2vec is capable of finding novel TF motifs, it is difficult to determine the identity of the TF whose motif is not curated. Future research may develop methods to ascertain TF identity given the motif and corresponding regulon inferred from Peak2vec.

In addition, we performed Gaussian mixture on the embedding vector. This method performed well when the number of clusters was known. Such is the case with the simulation experiment. However, in real data experiment, the number of TFs involved are unknown. When determined by AIC, the number tends to be too large and led to redundant clusters. In the future, we need to refine the workflow on how to extract the TF motifs and their corresponding regulon.

Please note that application of peak2vec is not restricted to ATAC-seq. Any set of fixed size sequences can be the input. For example, peak2vec may also be applied to TF ChIP-seq experiment in case multiple motifs exist as cofactors. In the future, we may expand the application of peak2vec to other types of high-throughput technologies.

This study presented a novel algorithm named peak2vec. We believe peak2vec would serve as a valuable tool for biologists to analyze transcriptional regulation from ATAC-seq profiles. It also deserves the attention of computational researchers due to its general utility to extract sequence information from sequencing data.

8.0 Discussion

This dissertation project explores various topics of computational analysis of high-throughput data. While some topics have vast research body, such as network clustering and biclustering, others, such as inferring latent pathways and ATAC-seq deconvolution, have rarely been attempted with a data-driven approach. In the former topics, this dissertation has developed methods with performance gain. As for the latter, we developed new models to address the question with promising results. Overall, the major hypothesis was supported by our investigation throughout this dissertation. That is, each omics profile provides different clues to the big picture of gene functions.

Although this dissertation presented research projects with diverse methodologies, they are connected in terms of the biological questions to be addressed. First, we directly integrated transcriptomics, proteomics, and literature knowledge into the multiplex network formulation. We developed a novel clustering algorithm to identify functional modules from this integrative network. With all the integrated information, our method has improved greatly on precision. This is because our method is prone to identify functional modules supported by multiple knowledge sources, namely the intersection. This integrative approach improved the reliability and reproducibility of the computational discovery. The downside is that functional modules strongly supported by a single data source can be neglected, reducing the possibility of revealing novel biological insight. With the obvious improvement of precision and minor loss of recall, the overall accuracy was improved. Thus, our hypothesis about the performance gain from multiplex integration is supported by the multiplex approach. Still, we need to develop better heuristics from specific biological mechanisms so we may address the downside of the multiplex approach.

We then looked into the bicluster problem of transcriptomics. We assumed that when a gene is regulated by multiple pathways, dysregulation of one of them is sufficient to cause differential expression of the gene. With this assumption, we are able to model the relationship between gene expression and latent pathways with the AND-OR product. The biclusters identified by our method reveal information about cancer subtypes, tumor microenvironment, and cellular functions related to patient survival. Therefore, investigation in Chapter 4 supports our hypothesis that coexpression among genes hints at their functional similarity and their coregulation.

After developing an improved BMF algorithm for various transcriptomic data, we realized that the OR-gate mechanism underlying the BMF model could be generalized to capture the causal mechanism about how somatic mutations disturb biological pathways, particularly the mutual exclusivity pattern. Hence, we expanded the BMF model to integrate the genomic profiles, leading to the model of ORN. This model assumes that somatic mutations affect pathways and then cause differential expression following the same OR-gate logic. Experimental results show that ORN can identify somatic mutations affecting the same pathway and causing similar sets of genes to differentially express. The pathway genes identified by ORN also exhibit mutual exclusivity patterns. Our work on ORN indicated that functional relationships among SGAs can be identified by their impact on the transcriptomic profiles, which is one of our major hypotheses to examine in this thesis.

However, ORN has notable limitations. One is the size of somatic mutations. Since the OR-gate only needs one input unit to activate, the output will almost always become activated if the input size is over 10,000. In the case of ORN, the size of candidate mutations easily goes beyond 20,000. Therefore, we have filtered the somatic mutations with penalized regression before feeding ORN. Still, this approach may include too many passengers while excluding drivers with

moderate effects. To refine the input for ORN, we developed gene embedding from biomedical literature. By constructing semantic representations for each gene, we improve information coverage for less well-known genes. Results showed that our gene embedding captures the functional space of genes better than previous models. Our study showed that biomedical literature is an important information source for the functional relationships of high-throughput omics data. Although unlikely to generate novel insights, knowledge from literature is able to complement the data quality issues of high-throughput data. Therefore, our hypothesis about the value of this data source is supported by our evaluation of gene embedding.

Another limitation of ORN is the size of pathways. Despite the promising results of BMF, its performance decreases dramatically when too many latent pathways are involved. This is limited by the sample size as spurious correlation arises when the sample size (<1,000) is much smaller than the gene size (> 10,000). To this end, we attempted to complement the gene level read count information with DNA sequence information. It leads to a new model, Peak2Vec. Essentially, Peak2Vec is using multinomial mixture to project the DNA sequences into an embedding space, such that sequences with similar motifs patterns are closer in the embedding space. Our results show Peak2Vec is able to identify different TF motifs from the sequences sets of ATAC-seq. This supports our hypothesis that transcriptional regulation can be inferred from chromatin accessibility profiles.

9.0 Future Work

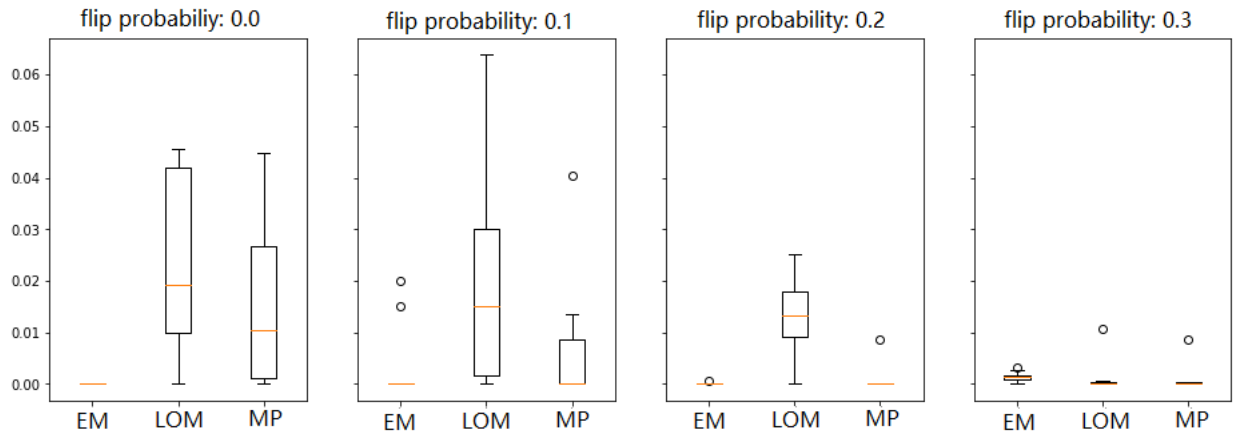
The results of this dissertation support that different types of omics data provide different information about gene functions. These differing data types can complement each other in discovering novel functioning roles of genes. We have not yet developed an integrated framework that utilizes all types of data at once. It is a longer-term goal, beyond this dissertation.

In the shorter term, we plan to investigate combining two methods together to see if doing so improves the performance of inferring regulatory mechanism. For example, BMF could be combined with Peak2vec to identify co-regulation modules with higher resolution and higher confidence. This is because BMF utilizes the correlation of gene expression while Peak2vec utilizes the similarity of DNA sequence in regulatory regions. The two types of information are complementary to each other. Another possible integration is to use gene embedding to filter the driver mutations as input for ORN. We can examine whether ORN provides more reliable insights when most neighboring passengers are removed.

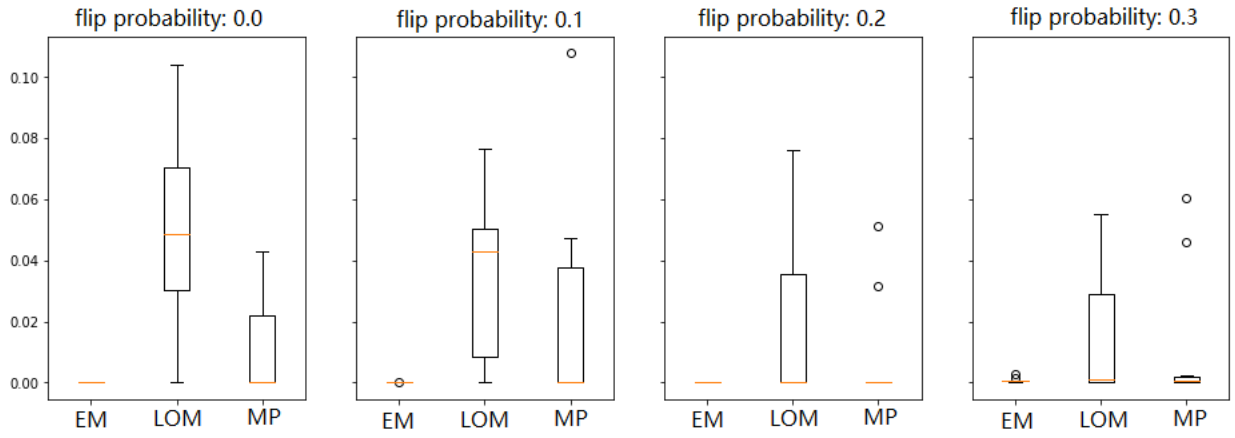
The model assumptions of ORN could be too strong. Although the OR-gate relationship is supported by the mutual exclusivity pattern, it cannot explain pathway crosstalk or the co-occurrence pattern. We may consider adopting the AND-gate relationship into the framework. In principle, AND-gate requires all the input to be one for the output to turn one. This property is aligned with the co-occurrence pattern and pathway crosstalk, which suggests that tumorigenesis requires a certain combination of pathway dysregulation. However, integration of AND-gate may substantially increase model complexity, hence adding difficulty to inference and interpretability. This could be the major issue we need to address in the future.

Furthermore, we need to expand the applicability of ORN to other kinds of diseases. Since most diseases are not caused by somatic mutations, we may start investigating how germline variants lead to increased susceptibility or prognosis. For the problem of elucidating the impact of germline variants, various approaches have already been proposed, including GWAS, PheWAS, eQTL, and Mendelian randomization. We need to learn more from the existing methods and consider how our work demonstrated in this dissertation may fit into this developed body of research.

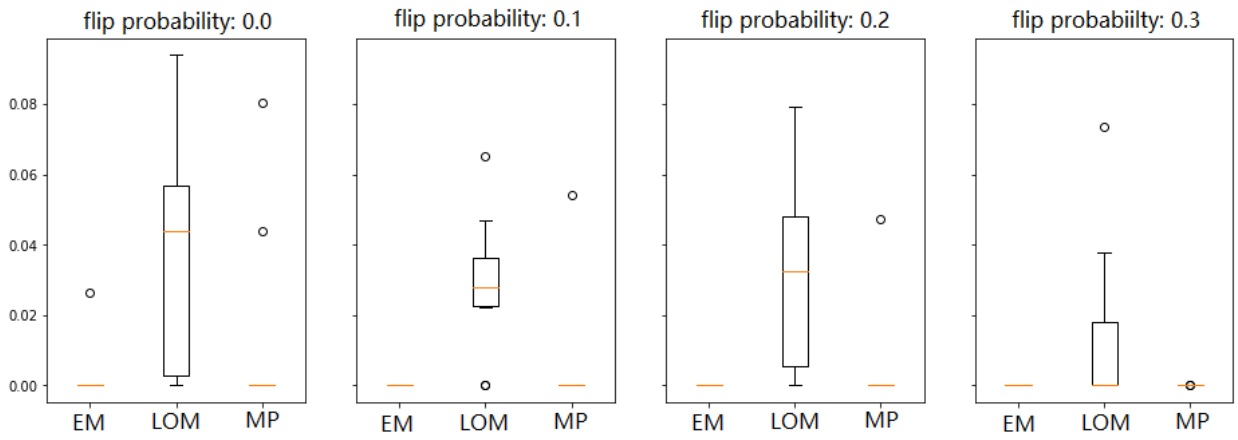
Appendix A Supplementary Information for Chapter 4 (BMF)



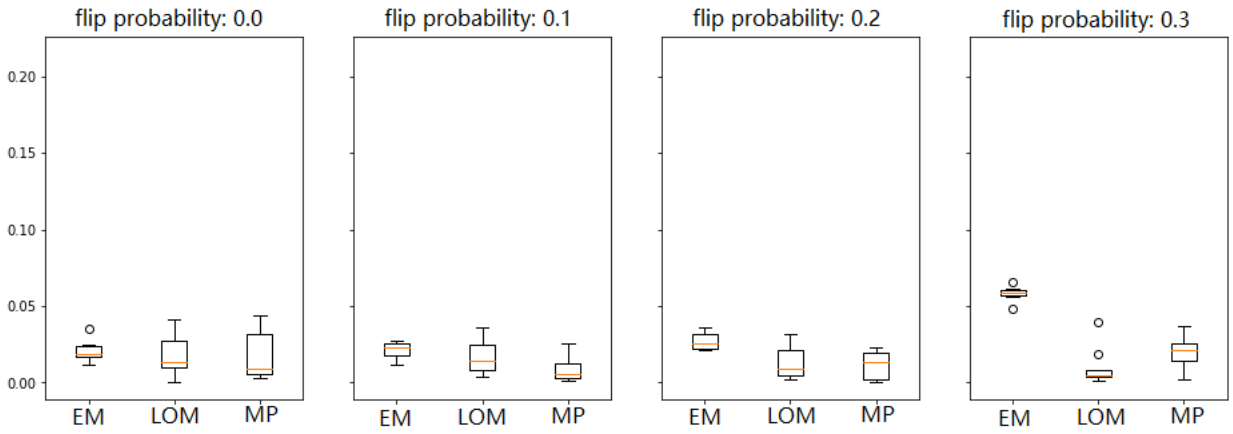
Appendix Figure 1 Reconstruction error of synthetic data when Bernoulli is different for all latent factors. Sample matrix is 1000 by 1000, rank 5, density 0.3



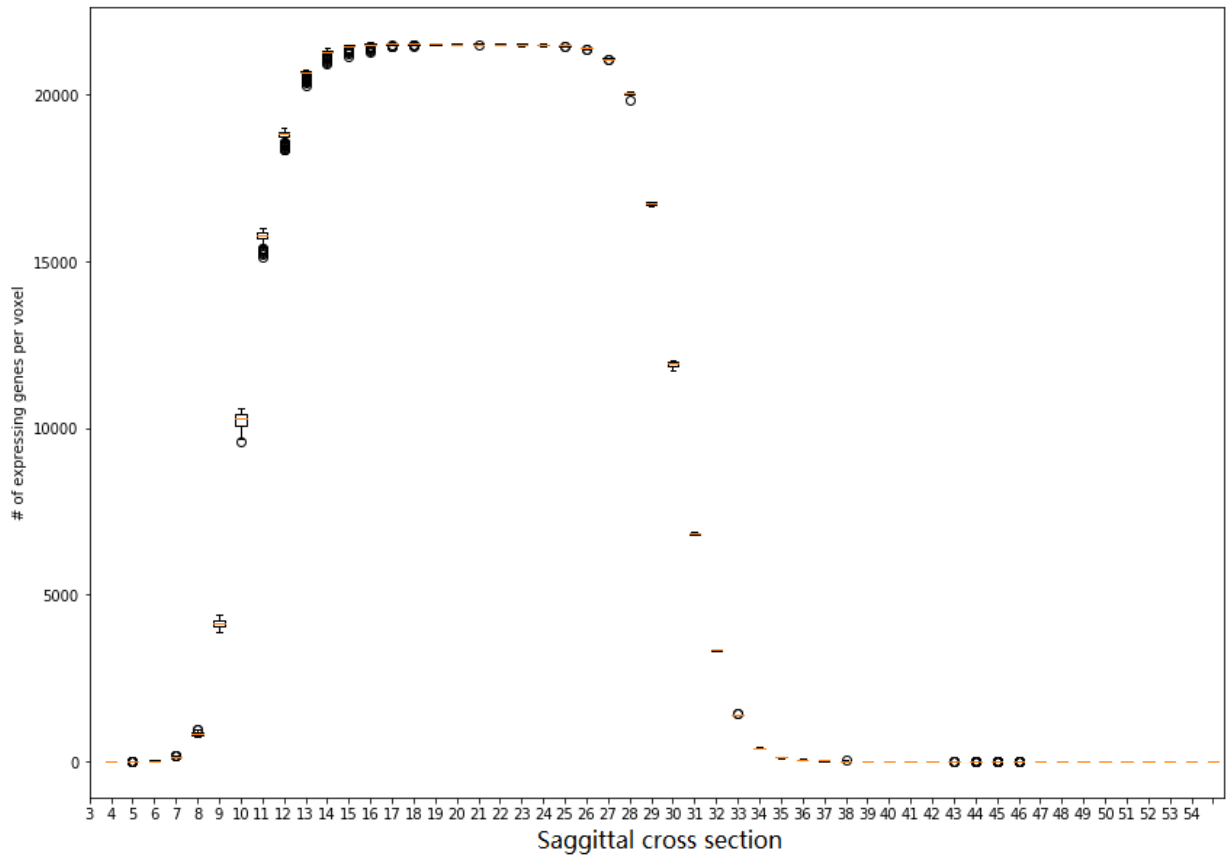
Appendix Figure 2 Reconstruction error of synthetic data when Bernoulli is different for all latent factors. Sample matrix is 1000 by 1000, rank 5, density 0.7



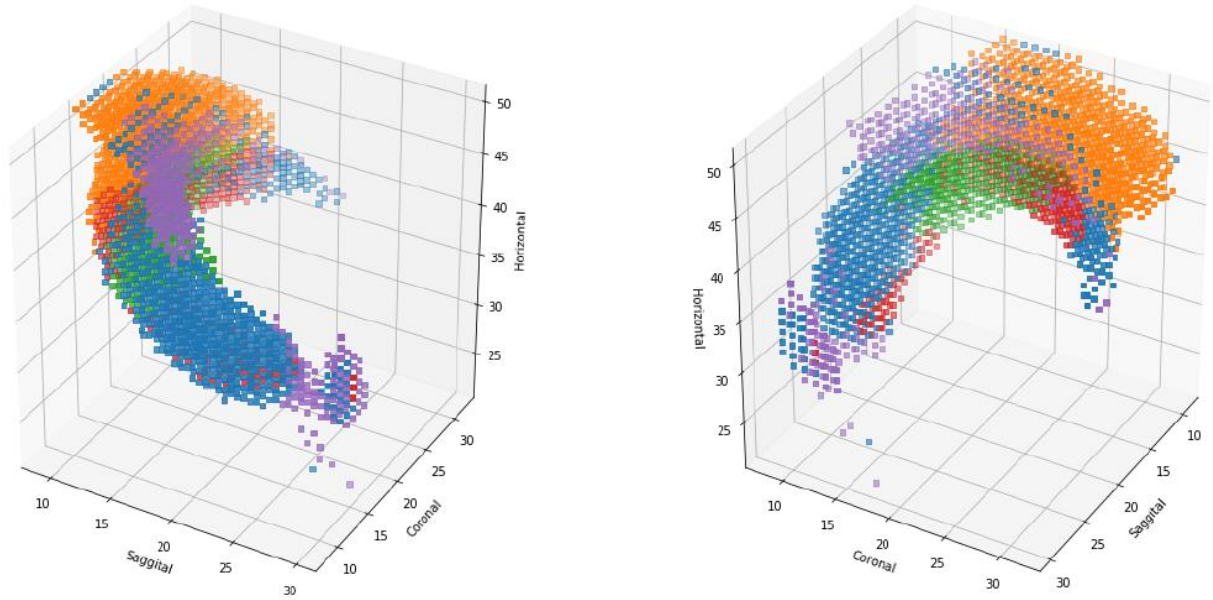
Appendix Figure 3 Reconstruction error of synthetic data when Bernoulli is different for all latent factors. Sample matrix is 2500 by 2500, rank 5, density 0.5



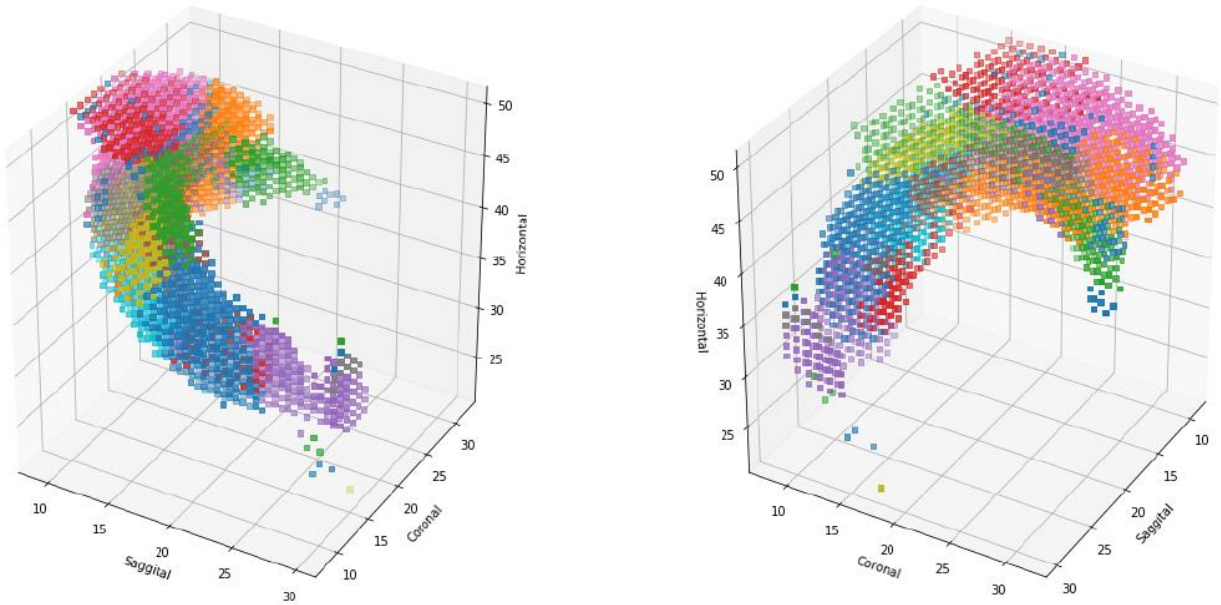
Appendix Figure 4 Reconstruction error of synthetic data when Bernoulli is different for all latent factors. Sample matrix is 1000 by 1000, rank 10, density 0.5



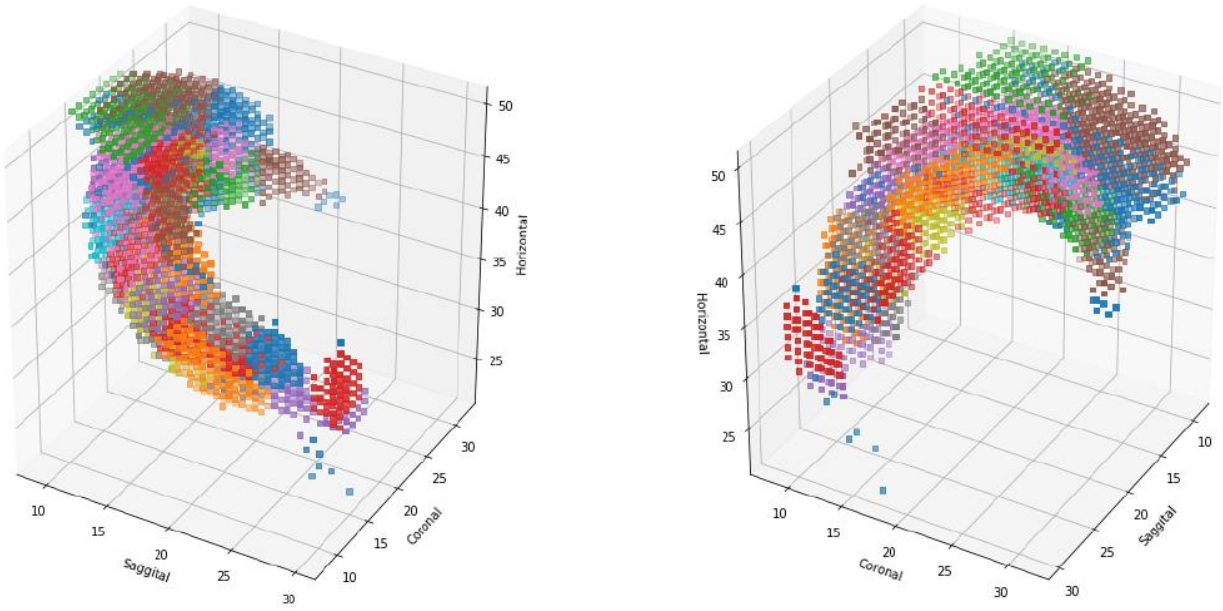
Appendix Figure 5 The x axis is the ordered saggital section of Hippocampal formation in mouse brain. The y axis is the number of expressed genes. Each box is the voxel distribution of the number of expressed genes in each slides. The number of expressed genes is unusually consistent across voxels within a section. Thus we assumed that "non-expressed" genes are actually not measured.



Appendix Figure 6 5-factorization of spatial transcriptomics in mouse hippocampal formation



Appendix Figure 7 15-factorization of spatial transcriptomics in mouse hippocampal formation



Appendix Figure 8 30-factorization of spatial transcriptomics in mouse hippocampal formation

Appendix B Supplementary Information for Chapter 5 (ORN)

Appendix Table 1 Comparison of the GO enrichment analysis results of ORN and NN on the dataset of glioblastoma. Note that the pathways are not aligned between ORN and NN.

	ORN	NN
0	cell-cell adhesion via plasma-membrane adhesion molecules, cholesterol biosynthetic process, secondary alcohol biosynthetic process	chromosome segregation, mitotic nuclear division, DNA replication
1	NA	mitotic nuclear division, chromosome segregation, mitotic sister chromatid segregation
2	synaptic vesicle cycle, vesicle-mediated transport in synapse, neurotransmitter secretion	cilium movement, axoneme assembly, microtubule bundle formation
3	oxidative phosphorylation, regulation of cellular amino acid metabolic process, anaphase-promoting complex-dependent catabolic process	axoneme assembly, microtubule bundle formation, cilium movement
4	positive regulation of cell activation, adaptive immune response, leukocyte proliferation	mitochondrial electron transport, NADH to ubiquinone, mitochondrial respiratory chain complex assembly, mitochondrial translational elongation
5	ribosome biogenesis, rRNA metabolic process, ncRNA transcription	chromosome segregation, sister chromatid segregation, mitotic nuclear division
6	chromosome segregation, DNA replication, mitotic nuclear division	NA
7	glycosyl compound catabolic process, negative regulation of viral genome replication, nucleoside catabolic process	vesicle-mediated transport in synapse, synaptic vesicle cycle, synaptic vesicle localization
8	establishment of protein localization to endoplasmic reticulum, SRP-dependent cotranslational protein targeting to membrane, protein targeting to ER	adaptive immune response, lymphocyte mediated immunity, positive regulation of cell activation
9	microtubule-based movement, intraciliary transport, intraciliary transport involved in cilium assembly	axoneme assembly, microtubule bundle formation, cilium movement

Appendix Table 2 GO enrichment analysis of the gene sets regulated by the 10 latent pathways in glioblastoma.

pathways	enriched GO	Adjusted P values
0	regulation of cholesterol biosynthetic process (GO:0045540)	0.006887475
	regulation of steroid biosynthetic process (GO:0050810)	0.006743178
	cholesterol metabolic process (GO:0008203)	0.009005787
1	sulfur compound biosynthetic process (GO:0044272)	0.879586995
	linoleic acid metabolic process (GO:0043651)	1
	unsaturated fatty acid metabolic process (GO:0033559)	1
2	chemical synaptic transmission (GO:0007268)	4.59E-04
	synaptic vesicle endocytosis (GO:0048488)	6.05E-04
	potassium ion transport (GO:0006813)	0.002609833
3	mitochondrial ATP synthesis coupled electron transport (GO:0042775)	2.82E-04
	positive regulation of protein ubiquitination involved in ubiquitin-dependent protein catabolic process (GO:2000060)	1.77E-04
	negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle (GO:0051436)	2.29E-04
4	cytokine-mediated signaling pathway (GO:0019221)	1.40E-09
	neutrophil activation involved in immune response (GO:0002283)	7.74E-09
	cellular response to cytokine stimulus (GO:0071345)	1.37E-08
5	mRNA splicing, via spliceosome (GO:0000398)	8.04E-10
	mRNA processing (GO:0006397)	3.35E-09
	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile (GO:0000377)	3.05E-08
6	DNA metabolic process (GO:0006259)	2.90E-30
	DNA replication (GO:0006260)	8.13E-23
	DNA repair (GO:0006281)	2.90E-18
7	negative regulation of Ras protein signal transduction (GO:0046580)	0.012246249
	neutrophil degranulation (GO:0043312)	0.010323356
	neutrophil activation involved in immune response (GO:0002283)	0.007721001
8	translation (GO:0006412)	4.91E-05
	peptide biosynthetic process (GO:0043043)	1.66E-04
	translational termination (GO:0006415)	1.38E-04
9	cilium assembly (GO:0060271)	5.55E-05
	organelle assembly (GO:0070925)	3.30E-05
	cilium organization (GO:0044782)	4.58E-05

Appendix Table 3 Enrichement analysis of the pathway modules (SGA) from each latent pathway in the METABRIC dataset.

pathway	Enriched pathways	adjusted pvalues
0,1,2,7,9,14	Endometrial cancer	4.48E-07
	Melanoma	1.08E-06
	Human T-cell leukemia virus 1 infection	3.40E-06
3	Cell cycle	0.511572572
	Viral carcinogenesis	0.660471068
	Human T-cell leukemia virus 1 infection	0.520416156
4	Thermogenesis	0.234719355
	Hippo signaling pathway	1
	Nicotinate and nicotinamide metabolism	1
5	Cell cycle	1
	Pathways in cancer	0.834110529
	Cellular senescence	0.557848337
6	Tyrosine metabolism	0.239695777
	Dopaminergic synapse	1
	Phenylalanine metabolism	1
8	Hepatocellular carcinoma	0.037311118
	Thermogenesis	0.047707284
	Endometrial cancer	0.054969018
10	Dopaminergic synapse	1
	Tyrosine metabolism	1
	ABC transporters	1
11	AMPK signaling pathway	0.019275147
	Adipocytokine signaling pathway	0.039825758
	Transcriptional misregulation in cancer	0.462991613
12	Cocaine addiction	1
	Amphetamine addiction	1
	Synaptic vesicle cycle	1
13	PI3K-Akt signaling pathway	1

Appendix Table 4 GO enrichment analysis of the downstream modules (DEG) of each latent pathways extracted from the METABRIC dataset.

pathways	enriched GO	Adjusted P values
0,1,2,7,9,14	metaphase plate congression (GO:0051310)	0.00605092
	mitotic sister chromatid segregation (GO:0000070)	0.006466205
	mitotic cell cycle phase transition (GO:0044772)	0.005949139
3	T cell activation (GO:0042110)	7.65E-05
	regulation of immune response (GO:0050776)	6.00E-05
	cellular defense response (GO:0006968)	1.88E-04
4	protein targeting to ER (GO:0045047)	6.71E-06
	cotranslational protein targeting to membrane (GO:0006613)	1.72E-05
	negative regulation of mitotic cell cycle phase transition (GO:1901991)	1.31E-05
5	mitotic cell cycle phase transition (GO:0044772)	0.002754501
	cytoskeleton-dependent cytokinesis (GO:0061640)	0.004108474
	cell cycle G1/S phase transition (GO:0044843)	0.003473781
6	viral gene expression (GO:0019080)	0.09646972
	viral transcription (GO:0019083)	0.056216791
	viral process (GO:0016032)	0.202794165
8	formation of extrachromosomal circular DNA (GO:0001325)	1
	t-circle formation (GO:0090656)	0.660897824
	telomere maintenance via telomere trimming (GO:0090737)	0.440598549
10	regulation of cellular component movement (GO:0051270)	1
	regulation of transcription from RNA polymerase II promoter in response to hypoxia (GO:0061418)	1
	regulation of epithelial cell migration (GO:0010632)	1
11	positive regulation of protein glycosylation (GO:0060050)	1
	nucleotide-excision repair (GO:0006289)	1
	nuclear-transcribed mRNA catabolic process, exonucleolytic (GO:0000291)	1
12	nucleosome organization (GO:0034728)	0.213337743
	mitochondrial translational elongation (GO:0070125)	0.213620717
	mitochondrial translational termination (GO:0070126)	0.167382305
13	mRNA processing (GO:0006397)	0.010155928
	mRNA splicing, via spliceosome (GO:0000398)	0.011027941
	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile (GO:0000377)	0.014034429

Bibliography

Abdalkader, Lamia, Takashi Oka, Katsuyoshi Takata, Hiaki Sato, Ichiro Murakami, Arie P Otte, and Tadashi Yoshino. 2016. “Aberrant differential expression of EZH1 and EZH2 in Polycomb repressive complex 2 among B- and T/NK-cell neoplasms.” *Pathology* 48 (5): 467–482. https://doi.org/10.1016/j.pathol.2016.05.002.

Alipanahi, Babak, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. 2015. “Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning.” *Nature Biotechnology* 33 (8): 831–838. https://doi.org/10.1038/nbt.3300.

Amantini, Consuelo, Maria Beatrice Morelli, Massimo Nabissi, Claudio Cardinali, Matteo Santoni, Angela Gismondi, and Giorgio Santoni. 2016. “Capsaicin triggers autophagic cell survival which drives epithelial mesenchymal transition and chemoresistance in bladder cancer cells in an Hedgehog-dependent manner.” *Oncotarget* 7 (31): 50180–50194. https://doi.org/10.18632/oncotarget.10326.

An, Chuanfu, and Zhonglin Mou. 2013. “The function of the Mediator complex in plant immunity.” *Plant Signaling & Behavior* 8 (3): e23182. https://doi.org/10.4161/psb.23182.

Bader, Joel S, Amitabha Chaudhuri, Jonathan M Rothberg, and John Chant. 2004. “Gaining confidence in high-throughput protein interaction networks.” *Nature Biotechnology* 22 (1): 78–85. https://doi.org/10.1038/nbt924.

Baek, Songjoon, Ido Goldstein, and Gordon L Hager. 2017. “Bivariate genomic footprinting detects changes in transcription factor activity.” *Cell reports* 19 (8): 1710–1722. https://doi.org/10.1016/j.celrep.2017.05.003.

Bailey, Timothy L, and Philip Machanick. 2012. “Inferring direct DNA binding from ChIP-seq.” *Nucleic Acids Research* 40 (17): e128. https://doi.org/10.1093/nar/gks433.

Belohlavek, Radim, and Martin Trnecka. 2018. “A new algorithm for Boolean matrix factorization which admits overcovering.” *Discrete Applied Mathematics* 249 (November): 36–52. https://doi.org/10.1016/j.dam.2017.12.044.

Berglund, Emelie, Jonas Maaskola, Niklas Schultz, Stefanie Friedrich, Maja Marklund, Joseph Bergenstråhle, Firas Tarish, et al. 2018. “Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity.” *Nature Communications* 9 (1): 2419. https://doi.org/10.1038/s41467-018-04724-5.

Bergmann, Sven, Jan Ihmels, and Naama Barkai. 2003. “Iterative signature algorithm for the analysis of large-scale gene expression data.” *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics* 67 (3 Pt 1): 031902. https://doi.org/10.1103/PhysRevE.67.031902.

Blazek, Dalibor, Jiri Kohoutek, Koen Bartholomeeusen, Eric Johansen, Petra Hulinkova, Zeping Luo, Peter Cimermancic, Jernej Ule, and B Matija Peterlin. 2011. “The Cyclin K/Cdk12 complex maintains genomic stability via regulation of expression of DNA damage response genes.” *Genes & Development* 25 (20): 2158–2172. https://doi.org/10.1101/gad.16962311.

Blei, D M, A Y Ng, and M I Jordan. 2003. “Latent dirichlet allocation.” *Journal of machine Learning research*.

Boccaletti, S, G Bianconi, R Criado, C I Del Genio, J Gómez-Gardeñes, M Romance, I Sendiña-Nadal, Z Wang, and M Zanin. 2014. “The structure and dynamics of multilayer networks.” *Physics reports* 544 (1): 1–122. https://doi.org/10.1016/j.physrep.2014.07.001.

Boni, V, R Zarate, J C Villa, E Bandrés, M A Gomez, E Maiello, J Garcia-Foncillas, and E Aranda. 2011. “Role of primary miRNA polymorphic variants in metastatic colon cancer patients treated with 5-fluorouracil and irinotecan.” *The Pharmacogenomics Journal* 11 (6): 429–436. https://doi.org/10.1038/tpj.2010.58.

Braun, Heinrich. 1992. “RPROP - A Fast Adaptive Learning Algorithm.” *PROC. OF ISCIS VII), UNIVERSITAT*.

Brunet, Jean-Philippe, Pablo Tamayo, Todd R Golub, and Jill P Mesirov. 2004. “Metagenes and molecular pattern discovery using matrix factorization.” *Proceedings of the National Academy of Sciences of the United States of America* 101 (12): 4164–4169. https://doi.org/10.1073/pnas.0308531101.

Budhwani, Megha, Roberta Mazzieri, and Riccardo Dolcetti. 2018. “Plasticity of Type I Interferon-Mediated Responses in Cancer Therapy: From Anti-tumor Immunity to Resistance.” *Frontiers in oncology* 8 (August): 322. https://doi.org/10.3389/fonc.2018.00322.

Buenrostro, Jason D, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf. 2013. “Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position.” *Nature Methods* 10 (12): 1213–1218. https://doi.org/10.1038/nmeth.2688.

Canisius, Sander, John W M Martens, and Lodewyk F A Wessels. 2016. “A novel independence test for somatic alterations in cancer shows that biology drives mutual exclusivity but chance explains most co-occurrence.” *Genome Biology* 17 (1): 261. https://doi.org/10.1186/s13059-016-1114-x.

Castellino, Robert C, and Donald L Durden. 2007. “Mechanisms of disease: the PI3K-Akt-PTEN signaling node--an intercept point for the control of angiogenesis in brain tumors.” *Nature Clinical Practice. Neurology* 3 (12): 682–693. https://doi.org/10.1038/ncpneuro0661.

Cerami, Ethan, Jianjiong Gao, Ugur Dogrusoz, Benjamin E Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, et al. 2012. “The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data.” *Cancer discovery* 2 (5): 401–404. https://doi.org/10.1158/2159-8290.CD-12-0095.

Chen, Vicky, John Paisley, and Xinghua Lu. 2017. “Revealing common disease mechanisms shared by tumors of different tissues of origin through semantic representation of genomic alterations and topic modeling.” *BMC Genomics* 18 (Suppl 2): 105. https://doi.org/10.1186/s12864-017-3494-z.

Cheng, Y, and G M Church. 2000. “Biclustering of expression data.” *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology* 8: 93–103.

Cheong, Jae Youn, Hyoung Doo Shin, Sung Won Cho, and Yoon Jun Kim. 2014. “Association of polymorphism in microRNA 604 with susceptibility to persistent hepatitis B virus infection and development of hepatocellular carcinoma.” *Journal of Korean Medical Science* 29 (11): 1523–1527. https://doi.org/10.3346/jkms.2014.29.11.1523.

Choobdar, Sarvenaz, Mehmet E. Ahsen, Jake Crawford, Mattia Tomasoni, David Lamparter, Junyuan Lin, Benjamin Hescott, et al. 2018. “Open Community Challenge Reveals Molecular Network Modules with Key Roles in Diseases.” *BioRxiv*, February. https://doi.org/10.1101/265553.

Chu, Andy, Gordon Robertson, Denise Brooks, Andrew J Mungall, Inanc Birol, Robin Coope, Yussanne Ma, Steven Jones, and Marco A Marra. 2016. “Large-scale profiling of microRNAs for The Cancer Genome Atlas.” *Nucleic Acids Research* 44 (1): e3. https://doi.org/10.1093/nar/gkv808.

Ciriello, Giovanni, Ethan Cerami, Bulent Arman Aksoy, Chris Sander, and Nikolaus Schultz. 2013. “Using MEMo to discover mutual exclusivity modules in cancer.” *Current Protocols in Bioinformatics* Chapter 8 (March): Unit 8.17. https://doi.org/10.1002/0471250953.bi0817s41.

Corces, M Ryan, Jeffrey M Granja, Shadi Shams, Bryan H Louie, Jose A Seoane, Wanding Zhou, Tiago C Silva, et al. 2018. “The chromatin accessibility landscape of primary human cancers.” *Science* 362 (6413). https://doi.org/10.1126/science.aav1898.

Cuenco, Joy, Natascha Wehnert, Oleg Blyuss, Anna Kazarian, Harry J Whitwell, Usha Menon, Anne Dawnay, Michael P Manns, Stephen P Pereira, and John F Timms. 2018. “Identification of a serum biomarker panel for the differential diagnosis of cholangiocarcinoma and primary sclerosing cholangitis.” *Oncotarget* 9 (25): 17430–17442. https://doi.org/10.18632/oncotarget.24732.

Curtis, Christina, Sohrab P Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M Rueda, Mark J Dunning, Doug Speed, et al. 2012. “The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.” *Nature* 486 (7403): 346–352. https://doi.org/10.1038/nature10983.

D’haeseleer, Patrik. 2006. “What are DNA sequence motifs?” *Nature Biotechnology* 24 (4): 423–425. https://doi.org/10.1038/nbt0406-423.

Davies, Helen, Graham R Bignell, Charles Cox, Philip Stephens, Sarah Edkins, Sheila Clegg, Jon Teague, et al. 2002. “Mutations of the BRAF gene in human cancer.” *Nature* 417 (6892): 949–954. https://doi.org/10.1038/nature00766.

Davies, Stefan M K, Oliver Rackham, Anne-Marie J Shearwood, Kristina L Hamilton, Reena Narsai, James Whelan, and Aleksandra Filipovska. 2009. “Pentatricopeptide repeat domain protein 3 associates with the mitochondrial small ribosomal subunit and regulates translation.” *FEBS Letters* 583 (12): 1853–1858. https://doi.org/10.1016/j.febslet.2009.04.048.

Deng, Yibin, Suzanne S Chan, and Sandy Chang. 2008. “Telomere dysfunction and tumour suppression: the senescence connection.” *Nature Reviews. Cancer* 8 (6): 450–458. https://doi.org/10.1038/nrc2393.

Deng, Yulan, Shangyi Luo, Chunyu Deng, Tao Luo, Wenkang Yin, Hongyi Zhang, Yong Zhang, et al. 2019. “Identifying mutual exclusivity across cancer genomes: computational approaches to discover genetic interaction and reveal tumor vulnerability.” *Briefings in Bioinformatics* 20 (1): 254–266. https://doi.org/10.1093/bib/bbx109.

Dey, Kushal K, Chiaowen Joyce Hsiao, and Matthew Stephens. 2017. “Visualizing the structure of RNA-seq expression data using grade of membership models.” *PLoS Genetics* 13 (3): e1006599. https://doi.org/10.1371/journal.pgen.1006599.

Domenico, Manlio De, Clara Granell, Mason Porter, and Alex Arenas. n.d. “The physics of spreading processes in multilayer net- works.”

Du, Jingcheng, Peilin Jia, Yulin Dai, Cui Tao, Zhongming Zhao, and Degui Zhi. 2019. “Gene2vec: distributed representation of genes based on coexpression.” *BMC Genomics* 20 (Suppl 1): 82. https://doi.org/10.1186/s12864-018-5370-x.

El Sayed, Ibrahim, Maged W Helmy, and Hanan S El-Abhar. 2018. “Inhibition of SRC/FAK cue: A novel pathway for the synergistic effect of rosuvastatin on the anti-cancer effect of dasatinib in hepatocellular carcinoma.” *Life Sciences* 213 (November): 248–257. https://doi.org/10.1016/j.lfs.2018.10.002.

Ellrott, Kyle, Matthew H Bailey, Gordon Saksena, Kyle R Covington, Cyriac Kandoth, Chip Stewart, Julian Hess, et al. 2018. “Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines.” *Cell Systems* 6 (3): 271–281.e7. https://doi.org/10.1016/j.cels.2018.03.002.

Espinosa-Cantú, Adriana, Diana Ascencio, Selene Herrera-Basurto, Jiewei Xu, Assen Roguev, Nevan J Krogan, and Alexander DeLuna. 2018. “Protein moonlighting revealed by noncatalytic phenotypes of yeast enzymes.” *Genetics* 208 (1): 419–431. https://doi.org/10.1534/genetics.117.300377.

Fernandez-Banet, Julio, Nikki P Lee, Kin Tak Chan, Huan Gao, Xiao Liu, Wing-Kin Sung, Winnie Tan, et al. 2014. “Decoding complex patterns of genomic rearrangement in hepatocellular carcinoma.” *Genomics* 103 (2-3): 189–203. https://doi.org/10.1016/j.ygeno.2014.01.003.

Fisher, Rosalie, and James Larkin. 2012. “Vemurafenib: a new treatment for BRAF-V600 mutated advanced melanoma.” *Cancer management and research* 4 (August): 243–252. https://doi.org/10.2147/CMAR.S25284.

Frolov, Alexander A., Dusan Husek, and Pavel Y. Polyakov. 2014. “Two Expectation-Maximization algorithms for Boolean Factor Analysis.” *Neurocomputing* 130 (April): 83–97. https://doi.org/10.1016/j.neucom.2012.02.055.

Gao, Jianjiong, Bülent Arman Aksoy, Ugur Dogrusoz, Gideon Dresdner, Benjamin Gross, S Onur Sumer, Yichao Sun, et al. 2013. “Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal.” *Science Signaling* 6 (269): pl1. https://doi.org/10.1126/scisignal.2004088.

GBD. 2018. “Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980-2017: a systematic analysis for the Global Burden of Disease Study 2017.” *The Lancet* 392 (10159): 1736–1788. https://doi.org/10.1016/S0140-6736(18)32203-7.

Gerhard, D S, S Hunger, C Lau, J Maris, P Meltzer, S Meshinchi, E Perlman, J Zhang, J Guidry-Auvil, and M Smith. 2018. “Therapeutically Applicable Research to Generate Effective Treatments (TARGET) Project: Half of Pediatric Cancers Have Their Own‘ Driver’ Genes.” *PEDIATRIC BLOOD & CANCER* 65.

Ghanbari, Mahsa, and Uwe Ohler. 2020. “Deep neural networks for interpreting RNA-binding protein target preferences.” *Genome Research* 30 (2): 214–226. https://doi.org/10.1101/gr.247494.118.

Goldman, Mary J, Brian Craft, Mim Hastie, Kristupas Repečka, Fran McDade, Akhil Kamath, Ayan Banerjee, et al. 2020. “Visualizing and interpreting cancer genomics data via the Xena platform.” *Nature Biotechnology* 38 (6): 675–678. https://doi.org/10.1038/s41587-020-0546-8.

Guo, Yuchun, Kevin Tian, Haoyang Zeng, Xiaoyun Guo, and David Kenneth Gifford. 2018. “A novel k-mer set memory (KSM) motif representation improves regulatory variant prediction.” *Genome Research* 28 (6): 891–900. https://doi.org/10.1101/gr.226852.117.

Hartwell, L H, J J Hopfield, S Leibler, and A W Murray. 1999. “From molecular to modular cell biology.” *Nature* 402 (6761 Suppl): C47–52. https://doi.org/10.1038/35011540.

Haslam, Alyson, and Vinay Prasad. 2019. “Estimation of the percentage of US patients with cancer who are eligible for and respond to checkpoint inhibitor immunotherapy drugs.” *JAMA network open* 2 (5): e192535. https://doi.org/10.1001/jamanetworkopen.2019.2535.

Hatzistergos, Konstantinos E, Adam R Williams, Derek Dykxhoorn, Michael A Bellio, Wendou Yu, and Joshua M Hare. 2019. “Tumor Suppressors RB1 and CDKN2a Cooperatively Regulate Cell-Cycle Progression and Differentiation During Cardiomyocyte Development and Repair.” *Circulation Research* 124 (8): 1184–1197. https://doi.org/10.1161/CIRCRESAHA.118.314063.

Ho, Jason, Edward S Cruise, Ryan J O Dowling, and Vuk Stambolic. 2020. “PTEN Nuclear Functions.” *Cold Spring Harbor perspectives in medicine* 10 (5). https://doi.org/10.1101/cshperspect.a036079.

Hochreiter, Sepp, Ulrich Bodenhofer, Martin Heusel, Andreas Mayr, Andreas Mitterecker, Adetayo Kasim, Tatsiana Khamiakova, et al. 2010. “FABIA: factor analysis for bicluster acquisition.” *Bioinformatics* 26 (12): 1520–1527. https://doi.org/10.1093/bioinformatics/btq227.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White. 1989. “Multilayer feedforward networks are universal approximators.” *Neural Networks* 2 (5): 359–366. https://doi.org/10.1016/0893-6080(89)90020-8.

Huang, Shao-Shan Carol, and Ernest Fraenkel. 2009. “Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks.” *Science Signaling* 2 (81): ra40. https://doi.org/10.1126/scisignal.2000350.

Huttlin, Edward L, Raphael J Bruckner, Joao A Paulo, Joe R Cannon, Lily Ting, Kurt Baltier, Greg Colby, et al. 2017. “Architecture of the human interactome defines protein communities and disease networks.” *Nature* 545 (7655): 505–509. https://doi.org/10.1038/nature22366.

ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. 2020. “Pan-cancer analysis of whole genomes.” *Nature* 578 (7793): 82–93. https://doi.org/10.1038/s41586-020-1969-6.

Jafri, Mohammad A, Shakeel A Ansari, Mohammed H Alqahtani, and Jerry W Shay. 2016. “Roles of telomeres and telomerase in cancer, and advances in telomerase-targeted therapies.” *Genome Medicine* 8 (1): 69. https://doi.org/10.1186/s13073-016-0324-x.

Jain, Shobhit, Sravan Babu Bodapati, Ramesh Nallapati, and Anima Anandkumar. 2019. “Multi Sense Embeddings from Topic Models.” *arXiv*, September.

Jeh, Glen, and Jennifer Widom. 2002. “SimRank: A measure of structural-context similarity.” In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining'' - KDD ’02*, 538. New York, New York, USA: ACM Press. https://doi.org/10.1145/775047.775126.

Ji, Junzhong, Aidong Zhang, Chunnian Liu, Xiaomei Quan, and Zhijun Liu. 2014. “Survey: Functional Module Detection from Protein-Protein Interaction Networks.” *IEEE transactions on knowledge and data engineering* 26 (2): 261–277. https://doi.org/10.1109/TKDE.2012.225.

Jia, Hao, Haolong Qi, Zhongqin Gong, Shucai Yang, Jianwei Ren, Yi Liu, Mingyue Li, and George Gong Chen. 2019. “The expression of FOXP3 and its role in human cancers.” *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*.

Karabacak Calviello, Aslıhan, Antje Hirsekorn, Ricardo Wurmus, Dilmurat Yusuf, and Uwe Ohler. 2019. “Reproducible inference of transcription factor footprints in ATAC-seq and DNase-seq datasets using protocol-specific bias modeling.” *Genome Biology* 20 (1): 42. https://doi.org/10.1186/s13059-019-1654-y.

Kim, Jeongkyun, Jung-Jae Kim, and Hyunju Lee. 2017. “An analysis of disease-gene relationship from Medline abstracts by DigSee.” *Scientific Reports* 7 (January): 40154. https://doi.org/10.1038/srep40154.

Kim, Sunkyu, Heewon Lee, Keonwoo Kim, and Jaewoo Kang. 2018. “Mut2Vec: distributed representation of cancerous mutations.” *BMC Medical Genomics* 11 (Suppl 2): 33. https://doi.org/10.1186/s12920-018-0349-7.

Kim, Yoo-Ah, Dong-Yeon Cho, Phuong Dao, and Teresa M Przytycka. 2015. “MEMCover: integrated analysis of mutual exclusivity and functional network reveals dysregulated pathways across multiple cancer types.” *Bioinformatics* 31 (12): i284–92. https://doi.org/10.1093/bioinformatics/btv247.

Koc, Emine C, Huseyin Cimen, Beril Kumcuoglu, Nadiah Abu, Gurler Akpinar, Md Emdadul Haque, Linda L Spremulli, and Hasan Koc. 2013. “Identification and characterization of CHCHD1, AURKAIP1, and CRIF1 as new members of the mammalian mitochondrial ribosome.” *Frontiers in physiology* 4 (July): 183. https://doi.org/10.3389/fphys.2013.00183.

Koo, Peter K, and Matt Ploenzke. 2020. “Deep learning for inferring transcription factor binding sites.” *Current Opinion in Systems Biology* 19 (February): 16–23. https://doi.org/10.1016/j.coisb.2020.04.001.

Kruglov, Emma A, Samir Gautam, Mateus T Guerra, and Michael H Nathanson. 2011. “Type 2 inositol 1,4,5-trisphosphate receptor modulates bile salt export pump activity in rat hepatocytes.” *Hepatology* 54 (5): 1790–1799. https://doi.org/10.1002/hep.24548.

Lal, Avantika, Zachary D Chiang, Nikolai Yakovenko, Fabiana M Duarte, Johnny Israeli, and Jason D Buenrostro. 2021. “Deep learning-based enhancement of epigenomics data with AtacWorks.” *Nature Communications* 12 (1): 1507. https://doi.org/10.1038/s41467-021-21765-5.

Lau, Grace M, Gillian M Lau, Guo-Liang Yu, Irwin H Gelman, Alan Gutowski, David Hangauer, and Jane W S Fang. 2009. “Expression of Src and FAK in hepatocellular carcinoma and the effect of Src inhibitors on hepatocellular carcinoma in vitro.” *Digestive Diseases and Sciences* 54 (7): 1465–1474. https://doi.org/10.1007/s10620-008-0519-0.

Lavaissiere, L, S Jia, M Nishiyama, S de la Monte, A M Stern, J R Wands, and P A Friedman. 1996. “Overexpression of human aspartyl(asparaginyl)beta-hydroxylase in hepatocellular carcinoma and cholangiocarcinoma.” *The Journal of Clinical Investigation* 98 (6): 1313–1323. https://doi.org/10.1172/JCI118918.

Lee, Joseph H, Rong Cheng, Lawrence S Honig, Mary Feitosa, Candace M Kammerer, Min S Kang, Nicole Schupf, et al. 2013. “Genome wide association and linkage analyses identified three loci-4q25, 17q23.2, and 10q11.21-associated with variation in leukocyte telomere length: the Long Life Family Study.” *Frontiers in genetics* 4: 310. https://doi.org/10.3389/fgene.2013.00310.

Lein, Ed S, Michael J Hawrylycz, Nancy Ao, Mikael Ayres, Amy Bensinger, Amy Bernard, Andrew F Boe, et al. 2007. “Genome-wide atlas of gene expression in the adult mouse brain.” *Nature* 445 (7124): 168–176. https://doi.org/10.1038/nature05453.

Leiserson, Mark D M, Matthew A Reyna, and Benjamin J Raphael. 2016. “A weighted exact test for mutually exclusive mutations in cancer.” *Bioinformatics* 32 (17): i736–i745. https://doi.org/10.1093/bioinformatics/btw462.

Levy, O, and Y Goldberg. 2014. “Neural word embedding as implicit matrix factorization.” *Advances in neural information processing …*.

Li, Fangda, Malalage N Peiris, and Daniel J Donoghue. 2020. “Functions of FGFR2 corrupted by translocations in intrahepatic cholangiocarcinoma.” *Cytokine & growth factor reviews* 52: 56–67. https://doi.org/10.1016/j.cytogfr.2019.12.005.

Li, Guojun, Qin Ma, Haibao Tang, Andrew H Paterson, and Ying Xu. 2009. “QUBIC: a qualitative biclustering algorithm for analyses of gene expression data.” *Nucleic Acids Research* 37 (15): e101. https://doi.org/10.1093/nar/gkp491.

Li, Xiangyu, Weizheng Chen, Yang Chen, Xuegong Zhang, Jin Gu, and Michael Q Zhang. 2017. “Network embedding-based representation learning for single cell RNA-seq data.” *Nucleic Acids Research* 45 (19): e166. https://doi.org/10.1093/nar/gkx750.

Li, Xiaoli, Min Wu, Chee-Keong Kwoh, and See-Kiong Ng. 2010. “Computational approaches for detecting protein complexes from protein interaction networks: a survey.” *BMC Genomics* 11 Suppl 1 (February): S3. https://doi.org/10.1186/1471-2164-11-S1-S3.

Li, Z, and B M Pützer. 2008. “Spliceosomal protein E regulates neoplastic cell growth by modulating expression of cyclin E/CDK2 and G2/M checkpoint proteins.” *Journal of Cellular and Molecular Medicine* 12 (6A): 2427–2438. https://doi.org/10.1111/j.1582-4934.2008.00244.x.

Li, Zhijian, Marcel H Schulz, Thomas Look, Matthias Begemann, Martin Zenke, and Ivan G Costa. 2019. “Identification of transcription factor binding sites using ATAC-seq.” *Genome Biology* 20 (1): 45. https://doi.org/10.1186/s13059-019-1642-2.

Liang, Lifan, Vicky Chen, Kunju Zhu, Xiaonan Fan, Xinghua Lu, and Songjian Lu. 2019. “Integrating data and knowledge to identify functional modules of genes: a multilayer approach.” *BMC Bioinformatics* 20 (1): 225. https://doi.org/10.1186/s12859-019-2800-y.

Liu, Weiran, Yuesong Yin, Jun Wang, Bowen Shi, Lianmin Zhang, Dong Qian, Chenguang Li, et al. 2017. “Kras mutations increase telomerase activity and targeting telomerase is a promising therapeutic strategy for Kras-mutant NSCLC.” *Oncotarget* 8 (1): 179–190. https://doi.org/10.18632/oncotarget.10162.

Liu, Yifeng, Yongjie Liang, and David Wishart. 2015. “PolySearch2: a significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more.” *Nucleic Acids Research* 43 (W1): W535–42. https://doi.org/10.1093/nar/gkv383.

Lucchese, Claudio, Salvatore Orlando, and Raffaele Perego. 2014. “A Unifying Framework for Mining Approximate Top- k' ' Binary Patterns.” *IEEE transactions on knowledge and data engineering* 26 (12): 2900–2913. https://doi.org/10.1109/TKDE.2013.181.

Ma, Ying Jie, Estrid Hein, Lea Munthe-Fog, Mikkel-Ole Skjoedt, Rafael Bayarri-Olmos, Luigina Romani, and Peter Garred. 2015. “Soluble Collectin-12 (CL-12) Is a Pattern Recognition Molecule Initiating Complement Activation via the Alternative Pathway.” *Journal of Immunology* 195 (7): 3365–3373. https://doi.org/10.4049/jimmunol.1500493.

Mahipal, Amit, Anuhya Kommalapati, Sri Harsha Tella, Alexander Lim, and Richard Kim. 2018. “Novel targeted treatment options for advanced cholangiocarcinoma.” *Expert Opinion on Investigational Drugs* 27 (9): 709–720. https://doi.org/10.1080/13543784.2018.1512581.

Malik, Shabnam, Shilpa Bhatnagar, Naveen Chaudhary, Deepshikha Pande Katare, and S K Jain. 2013. “Elevated expression of complement C3 protein in chemically induced hepatotumorogenesis in Wistar rats: a correlative proteomics and histopathological study.” *Experimental and Toxicologic Pathology* 65 (6): 767–773. https://doi.org/10.1016/j.etp.2012.11.003.

Martínez-Jiménez, Francisco, Ferran Muiños, Inés Sentís, Jordi Deu-Pons, Iker Reyes-Salazar, Claudia Arnedo-Pac, Loris Mularoni, et al. 2020. “A compendium of mutational cancer driver genes.” *Nature Reviews. Cancer* 20 (10): 555–572. https://doi.org/10.1038/s41568-020-0290-x.

McFaline-Figueroa, J Ricardo, and Patrick Y Wen. 2017. “The viral connection to glioblastoma.” *Current infectious disease reports* 19 (2): 5. https://doi.org/10.1007/s11908-017-0563-z.

Miettinen, P, T Mielikäinen, and A Gionis. 2008. “The discrete basis problem.” *IEEE transactions on …*.

Miettinen, Pauli, and Stefan Neumann. 2020. “Recent Developments in Boolean Matrix Factorization.” *arXiv*, December.

Miller, Christopher A, Stephen H Settle, Erik P Sulman, Kenneth D Aldape, and Aleksandar Milosavljevic. 2011. “Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors.” *BMC Medical Genomics* 4 (April): 34. https://doi.org/10.1186/1755-8794-4-34.

Miller, Megan B, Yan Yan, Betty A Eipper, and Richard E Mains. 2013. “Neuronal Rho GEFs in synaptic physiology and behavior.” *The Neuroscientist* 19 (3): 255–273. https://doi.org/10.1177/1073858413475486.

Nagaoka, Katsuya, Kousuke Ogawa, Chengcheng Ji, Kevin Y Cao, Xuewei Bai, Joud Mulla, Zhixiang Cheng, Jack R Wands, and Chiung-Kuei Huang. 2020. “Targeting Aspartate Beta-Hydroxylase with the Small Molecule Inhibitor MO-I-1182 Suppresses Cholangiocarcinoma Metastasis.” *Digestive Diseases and Sciences*, May. https://doi.org/10.1007/s10620-020-06330-2.

Negrini, Simona, Vassilis G Gorgoulis, and Thanos D Halazonetis. 2010. “Genomic instability--an evolving hallmark of cancer.” *Nature Reviews. Molecular Cell Biology* 11 (3): 220–228. https://doi.org/10.1038/nrm2858.

“Networks - Mark Newman - Oxford University Press.” n.d. Accessed April 10, 2018. https://global.oup.com/academic/product/networks-9780199206650?cc=us&lang=en&.

Neumann, Stefan. 2018. “Bipartite Stochastic Block Models with Tiny Clusters.” In *Advances in Neural Information Processing Systems*, 3867–3877.

Padilha, Victor A, and Ricardo J G B Campello. 2017. “A systematic comparative evaluation of biclustering techniques.” *BMC Bioinformatics* 18 (1): 55. https://doi.org/10.1186/s12859-017-1487-1.

Paisley, John, Chong Wang, David M Blei, and Michael I Jordan. 2015. “Nested hierarchical dirichlet processes.” *IEEE transactions on pattern analysis and machine intelligence* 37 (2): 256–270. https://doi.org/10.1109/TPAMI.2014.2318728.

Park, Sungjoon, Yookyung Koh, Hwisang Jeon, Hyunjae Kim, Yoonsun Yeo, and Jaewoo Kang. 2020. “Enhancing the interpretability of transcription factor binding site prediction using attention mechanism.” *Scientific Reports* 10 (1): 13413. https://doi.org/10.1038/s41598-020-70218-4.

Patel, Anoop P, Itay Tirosh, John J Trombetta, Alex K Shalek, Shawn M Gillespie, Hiroaki Wakimoto, Daniel P Cahill, et al. 2014. “Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma.” *Science* 344 (6190): 1396–1401. https://doi.org/10.1126/science.1254257.

Patel, Nishal S, Muriel Rhinn, Claudia I Semprich, Pamela A Halley, Pascal Dollé, Wendy A Bickmore, and Kate G Storey. 2013. “FGF signalling regulates chromatin organisation during neural differentiation via mechanisms that can be uncoupled from transcription.” *PLoS Genetics* 9 (7): e1003614. https://doi.org/10.1371/journal.pgen.1003614.

Paul, Franziska, Ya’ara Arkin, Amir Giladi, Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Deborah Winter, et al. 2015. “Transcriptional heterogeneity and lineage commitment in myeloid progenitors.” *Cell* 163 (7): 1663–1677. https://doi.org/10.1016/j.cell.2015.11.013.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. “Scikit-learn: Machine Learning in Python.” *Journal of Machine Learning Research*.

Pennington, J, R Socher, and C D Manning. 2014. “Glove: Global vectors for word representation.” *Proceedings of the 2014 …*.

Pirmoradi, Leila, Nayer Seyfizadeh, Saeid Ghavami, Amir A Zeki, and Shahla Shojaei. 2019. “Targeting cholesterol metabolism in glioblastoma: a new therapeutic approach in cancer therapy.” *Journal of investigative medicine : the official publication of the American Federation for Clinical Research* 67 (4): 715–719. https://doi.org/10.1136/jim-2018-000962.

Quang, Daniel, and Xiaohui Xie. 2019. “FactorNet: A deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data.” *Methods* 166 (August): 40–47. https://doi.org/10.1016/j.ymeth.2019.03.020.

Rakhmanov, Mirzokhid, Heiko Sic, Anne-Kathrin Kienzler, Beate Fischer, Marta Rizzi, Maximilian Seidl, Kerstina Melkaoui, et al. 2014. “High levels of SOX5 decrease proliferative

capacity of human B cells, but permit plasmablast differentiation." *Plos One* 9 (6): e100328. https://doi.org/10.1371/journal.pone.0100328.

Ravanbakhsh, S, and B Póczos. 2016. "Boolean matrix factorization and noisy completion via message passing." *... Conference on Machine ...*.

Ravanbakhsh, Siamak, Barnabás Póczos, and Russell Greiner. 2016. "Boolean Matrix Factorization and Noisy Completion via Message Passing." In *ICML*, 945–954.

Rehurek, R, and P Sojka. 2010. "Software framework for topic modelling with large corpora." *In Proceedings of the LREC 2010 Workshop on New ...*.

Remy, Elisabeth, Sandra Rebouissou, Claudine Chaouiya, Andrei Zinovyev, François Radvanyi, and Laurence Calzone. 2015. "A Modeling Approach to Explain Mutually Exclusive and Co-Occurring Genetic Alterations in Bladder Tumorigenesis." *Cancer Research* 75 (19): 4042–4052. https://doi.org/10.1158/0008-5472.CAN-15-0602.

Robert, Francis, and Jerry Pelletier. 2018. "Exploring the Impact of Single-Nucleotide Polymorphisms on Translation." *Frontiers in genetics* 9 (October): 507. https://doi.org/10.3389/fgene.2018.00507.

Ross, Jeffrey S, Kai Wang, Laurie Gay, Rami Al-Rohil, Janne V Rand, David M Jones, Hwa J Lee, et al. 2014. "New routes to targeted therapy of intrahepatic cholangiocarcinomas revealed by next-generation sequencing." *The Oncologist* 19 (3): 235–242. https://doi.org/10.1634/theoncologist.2013-0352.

Rubin, Jonathan D, Jacob T Stanley, Rutendo F Sigauke, Cecilia B Levandowski, Zachary L Maas, Jessica Westfall, Dylan J Taatjes, and Robin D Dowell. 2021. "Transcription factor enrichment analysis (TFEA) quantifies the activity of multiple transcription factors from a single experiment." *Communications Biology* 4 (1): 661. https://doi.org/10.1038/s42003-021-02153-7.

Rukat, T, C C Holmes, and M K Titsias. 2017. "Bayesian boolean matrix factorisation." *... conference on machine ...*.

Rukat, Tammo, Chris C Holmes, Michalis K Titsias, and Christopher Yau. 2017. "Bayesian Boolean matrix factorisation." In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2969–2978.

Sade-Feldman, Moshe, Keren Yizhak, Stacey L Bjorgaard, John P Ray, Carl G de Boer, Russell W Jenkins, David J Lieb, et al. 2018. "Defining T Cell States Associated with Response to Checkpoint Immunotherapy in Melanoma." *Cell* 175 (4): 998–1013.e20. https://doi.org/10.1016/j.cell.2018.10.038.

Saelens, Wouter, Robrecht Cannoodt, and Yvan Saeys. 2018. "A comprehensive evaluation of module detection methods for gene expression data." *Nature Communications* 9 (1): 1090. https://doi.org/10.1038/s41467-018-03424-4.

Santegoets, Saskia Jam, Annelies W Turksma, Daniel J Powell, Erik Hooijberg, and Tanja D de Gruijl. 2013. “IL-21 in cancer immunotherapy: At the right place at the right time.” *Oncoimmunology* 2 (6): e24522. https://doi.org/10.4161/onci.24522.

Schep, Alicia N, Jason D Buenrostro, Sarah K Denny, Katja Schwartz, Gavin Sherlock, and William J Greenleaf. 2015. “Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions.” *Genome Research* 25 (11): 1757–1770. https://doi.org/10.1101/gr.192294.115.

Schierwagen, Robert, Frank E Uschner, Cristina Ortiz, Sandra Torres, Max J Brol, Olaf Tyc, Wenyi Gu, et al. 2020. “The Role of Macrophage-Inducible C-Type Lectin in Different Stages of Chronic Liver Disease.” *Frontiers in immunology* 11 (July): 1352. https://doi.org/10.3389/fimmu.2020.01352.

Schmid, S R, and P Linder. 1992. “D-E-A-D protein family of putative RNA helicases.” *Molecular Microbiology* 6 (3): 283–291. https://doi.org/10.1111/j.1365-2958.1992.tb01470.x.

Serfas, Michael S, and Angela L Tyner. 2003. “Brk, Srm, Frk, and Src42A form a distinct family of intracellular Src-like tyrosine kinases.” *Oncology Research* 13 (6-10): 409–419.

Sharma, Anchal, Chuan Jiang, and Subhajyoti De. 2018. “Dissecting the sources of gene expression variation in a pan-cancer analysis identifies novel regulatory mutations.” *Nucleic Acids Research* 46 (9): 4370–4381. https://doi.org/10.1093/nar/gky271.

Shi, Baomin, Xiuyan Wang, Xujie Han, Pengfei Liu, Weiwei Wei, and Yan Li. 2014. “Functional modules analysis based on coexpression network in pancreatic ductal adenocarcinoma.” *Pathology Oncology Research* 20 (2): 293–299. https://doi.org/10.1007/s12253-013-9694-1.

Sia, Daniela, Augusto Villanueva, Scott L Friedman, and Josep M Llovet. 2017. “Liver Cancer Cell of Origin, Molecular Class, and Effects on Patient Prognosis.” *Gastroenterology* 152 (4): 745–761. https://doi.org/10.1053/j.gastro.2016.11.048.

Slenter, Denise N, Martina Kutmon, Kristina Hanspers, Anders Riutta, Jacob Windsor, Nuno Nunes, Jonathan Mélius, et al. 2018. “WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research.” *Nucleic Acids Research* 46 (D1): D661–D667. https://doi.org/10.1093/nar/gkx1064.

Solé-Ribalta, Albert, Sergio Gómez, and Alex Arenas. 2016. “Congestion induced by the structure of multiplex networks.” *Physical Review Letters* 116 (10): 108701. https://doi.org/10.1103/PhysRevLett.116.108701.

Soneson, Charlotte, Sarah Gerster, and Mauro Delorenzi. 2014. “Batch effect confounding leads to strong bias in performance estimates obtained by cross-validation.” *Plos One* 9 (6): e100335. https://doi.org/10.1371/journal.pone.0100335.

Sørlie, T, C M Perou, R Tibshirani, T Aas, S Geisler, H Johnsen, T Hastie, et al. 2001. “Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.”

Proceedings of the National Academy of Sciences of the United States of America 98 (19): 10869–10874. https://doi.org/10.1073/pnas.191367098.

Speer, Robyn, Joshua Chin, and Catherine Havasi. 2016. "ConceptNet 5.5: An Open Multilingual Graph of General Knowledge." *arXiv,* December.

Stein-O'Brien, Genevieve L, Raman Arora, Aedin C Culhane, Alexander V Favorov, Lana X Garmire, Casey S Greene, Loyal A Goff, et al. 2018. "Enter the Matrix: Factorization Uncovers Knowledge from Omics." *Trends in Genetics* 34 (10): 790–805. https://doi.org/10.1016/j.tig.2018.07.003.

Stuart, Joshua M, Eran Segal, Daphne Koller, and Stuart K Kim. 2003. "A gene-coexpression network for global discovery of conserved genetic modules." *Science* 302 (5643): 249–255. https://doi.org/10.1126/science.1087447.

Suthram, Silpa, Joel T Dudley, Annie P Chiang, Rong Chen, Trevor J Hastie, and Atul J Butte. 2010. "Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets." *PLoS Computational Biology* 6 (2): e1000662. https://doi.org/10.1371/journal.pcbi.1000662.

Szczurek, Ewa, and Niko Beerenwinkel. 2014. "Modeling mutual exclusivity of cancer mutations." *PLoS Computational Biology* 10 (3): e1003503. https://doi.org/10.1371/journal.pcbi.1003503.

Tagami, Hideaki, Dominique Ray-Gallet, Geneviève Almouzni, and Yoshihiro Nakatani. 2004. "Histone H3.1 and H3.3 complexes mediate nucleosome assembly pathways dependent or independent of DNA synthesis." *Cell* 116 (1): 51–61. https://doi.org/10.1016/s0092-8674(03)01064-x.

Tai, Yuling, Chun Liu, Shuwei Yu, Hua Yang, Jiameng Sun, Chunxiao Guo, Bei Huang, et al. 2018. "Gene coexpression network analysis reveals coordinated regulation of three characteristic secondary biosynthetic pathways in tea plant (Camellia sinensis)." *BMC Genomics* 19 (1): 616. https://doi.org/10.1186/s12864-018-4999-9.

Tanay, Amos, Roded Sharan, and Ron Shamir. 2002. "Discovering statistically significant biclusters in gene expression data." *Bioinformatics* 18 Suppl 1: S136–44. https://doi.org/10.1093/bioinformatics/18.suppl_1.s136.

Tarbell, Evan D, and Tao Liu. 2019. "HMMRATAC: a Hidden Markov ModeleR for ATAC-seq." *Nucleic Acids Research* 47 (16): e91. https://doi.org/10.1093/nar/gkz533.

Tareen, Ammar, and Justin Block Kinney. 2019. "Biophysical models of cis-regulation as interpretable neural networks." *BioRxiv,* November. https://doi.org/10.1101/835942.

Tian, F, H Dai, J Bian, B Gao, R Zhang, and E Chen. n.d. "A probabilistic model for learning multi-prototype word embeddings."

Tomczak, Katarzyna, Patrycja Czerwińska, and Maciej Wiznerowicz. 2015. “The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge.” *Contemporary oncology (Poznan, Poland)* 19 (1A): A68–77. https://doi.org/10.5114/wo.2014.47136.

Tornow, Sabine, and H W Mewes. 2003. “Functional modules by relating protein interaction networks and gene expression.” *Nucleic Acids Research* 31 (21): 6283–6289. https://doi.org/10.1093/nar/gkg838.

Tsuchida, N, M Nakashima, A Miyauchi, S Yoshitomi, T Kimizu, V Ganesan, K W Teik, et al. 2018. “Novel biallelic SZT2 mutations in 3 cases of early-onset epileptic encephalopathy.” *Clinical Genetics* 93 (2): 266–274. https://doi.org/10.1111/cge.13061.

van Dam, Sipko, Urmo Võsa, Adriaan van der Graaf, Lude Franke, and João Pedro de Magalhães. 2018. “Gene coexpression analysis for functional classification and gene-disease predictions.” *Briefings in Bioinformatics* 19 (4): 575–592. https://doi.org/10.1093/bib/bbw139.

Walker, Sarah, Miriam Wankell, Vikki Ho, Rose White, Nikita Deo, Carol Devine, Brittany Dewdney, et al. 2019. “Targeting mTOR and Src restricts hepatocellular carcinoma growth in a novel murine liver cancer model.” *Plos One* 14 (2): e0212860. https://doi.org/10.1371/journal.pone.0212860.

Wallden, Brett, James Storhoff, Torsten Nielsen, Naeem Dowidar, Carl Schaper, Sean Ferree, Shuzhen Liu, et al. 2015. “Development and verification of the PAM50-based Prosigna breast cancer gene signature assay.” *BMC Medical Genomics* 8 (August): 54. https://doi.org/10.1186/s12920-015-0129-6.

Wang, Yong, Rui Jiang, and Wing Hung Wong. 2016. “Modeling the causal regulatory network by integrating chromatin accessibility and transcriptome data.” *National science review* 3 (2): 240–251. https://doi.org/10.1093/nsr/nww025.

Wozniak, Darren J, Andre Kajdacsy-Balla, Virgilia Macias, Susan Ball-Kell, Morgan L Zenner, Wenjun Bie, and Angela L Tyner. 2017. “PTEN is a protein phosphatase that targets active PTK6 and inhibits PTK6 oncogenic signaling in prostate cancer.” *Nature Communications* 8 (1): 1508. https://doi.org/10.1038/s41467-017-01574-5.

Wu, Fan, Rui-Chao Chai, Zhiliang Wang, Yu-Qing Liu, Zheng Zhao, Guan-Zhang Li, and Hao-Yu Jiang. 2019. “Molecular classification of IDH-mutant glioblastomas based on gene expression profiles.” *Carcinogenesis* 40 (7): 853–860. https://doi.org/10.1093/carcin/bgz032.

Xia, Hongping, Jianxiang Chen, Ming Shi, Hengjun Gao, Karthik Sekar, Veerabrahma Pratap Seshachalam, London Lucien P J Ooi, and Kam M Hui. 2015. “EDIL3 is a novel regulator of epithelial-mesenchymal transition controlling early recurrence of hepatocellular carcinoma.” *Journal of Hepatology* 63 (4): 863–873. https://doi.org/10.1016/j.jhep.2015.05.005.

Xie, Juan, Anjun Ma, Anne Fennell, Qin Ma, and Jing Zhao. 2019. “It is time to apply biclustering: a comprehensive review of biclustering applications in biological and biomedical data.” *Briefings in Bioinformatics* 20 (4): 1449–1464. https://doi.org/10.1093/bib/bby014.

Xing, Eric P, and Richard M Karp. 2004. “MotifPrototyper: a Bayesian profile model for motif families.” *Proceedings of the National Academy of Sciences of the United States of America* 101 (29): 10523–10528. https://doi.org/10.1073/pnas.0403564101.

Xu, Yaping, Yihao Huang, Wanqiong Xu, Xiaohui Zheng, Xue Yi, Liyue Huang, Yuxiao Wang, and Kangni Wu. 2020. “Activated hepatic stellate cells (hscs) exert immunosuppressive effects in hepatocellular carcinoma by producing complement C3.” *OncoTargets and therapy* 13 (February): 1497–1505. https://doi.org/10.2147/OTT.S234920.

Xuan, Ji, Jing Li, Zhirui Zhou, Renrong Zhou, Huabing Xu, and Wei Wen. 2016. “The diagnostic performance of serum MUC5AC for cholangiocarcinoma: A systematic review and meta-analysis.” *Medicine* 95 (24): e3513. https://doi.org/10.1097/MD.0000000000003513.

Yan, Feng, David R Powell, David J Curtis, and Nicholas C Wong. 2020. “From reads to insight: a hitchhiker’s guide to ATAC-seq data analysis.” *Genome Biology* 21 (1): 22. https://doi.org/10.1186/s13059-020-1929-3.

Yang, Li, Song Junmin, Yu Hong, and Wu Shuodong. 2013. “PGE(2) induces MUC2 and MUC5AC expression in human intrahepatic biliary epithelial cells via EP4/p38MAPK activation.” *Annals of Hepatology* 12 (3): 479–486.

Yang, Zhi, Feng Yu, Hong Lin, and Jian Wang. 2014. “Integrating PPI datasets with the PPI data from biomedical literature for protein complex detection.” *BMC Medical Genomics* 7 Suppl 2 (October): S3. https://doi.org/10.1186/1755-8794-7-S2-S3.

Yonekawa, Tohru, Shuqun Yang, and Christopher M Counter. 2012. “PinX1 localizes to telomeres and stabilizes TRF1 at mitosis.” *Molecular and Cellular Biology* 32 (8): 1387–1395. https://doi.org/10.1128/MCB.05641-11.

You, Qi, Liwei Zhang, Xin Yi, Kang Zhang, Dongxia Yao, Xueyan Zhang, Qianhua Wang, et al. 2016. “Coexpression network analyses identify functional modules associated with development and stress response in Gossypium arboreum.” *Scientific Reports* 6 (December): 38436. https://doi.org/10.1038/srep38436.

Zhang, Junhua, and Shihua Zhang. 2018. “The discovery of mutated driver pathways in cancer: models and algorithms.” *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 15 (3): 988–998. https://doi.org/10.1109/TCBB.2016.2640963.

Zhang, Liting, Jin Liu, Liping Qian, Qian Feng, Xiaofang Wang, Yukang Yuan, Yibo Zuo, et al. 2018. “Induction of OTUD1 by RNA viruses potently inhibits innate immune responses by promoting degradation of the MAVS/TRAF3/TRAF6 signalosome.” *PLoS Pathogens* 14 (5): e1007067. https://doi.org/10.1371/journal.ppat.1007067.

Zhang, Shitao, Xiaoping Zhang, Xueqi Fu, Wannan Li, Shu Xing, and Yiling Yang. 2018. “Identification of common differentially-expressed miRNAs in ovarian cancer cells and their exosomes compared with normal ovarian surface epithelial cell cells.” *Oncology letters* 16 (2): 2391–2401. https://doi.org/10.3892/ol.2018.8954.

Zhao, Bo, Jianhuang Lin, Lijie Rong, Shuai Wu, Zhong Deng, Nail Fatkhutdinov, Joseph Zundell, et al. 2019. “ARID1A promotes genomic stability through protecting telomere cohesion.” *Nature Communications* 10 (1): 4067. https://doi.org/10.1038/s41467-019-12037-4.

Zhou, Jian, and Olga G Troyanskaya. 2015. “Predicting effects of noncoding variants with deep learning-based sequence model.” *Nature Methods* 12 (10): 931–934. https://doi.org/10.1038/nmeth.3547.

Zhou, X Z, and K P Lu. 2001. “The Pin2/TRF1-interacting protein PinX1 is a potent telomerase inhibitor.” *Cell* 107 (3): 347–359.

Zhu, Lihui, Chengyong Qin, Tao Li, Xiaomin Ma, Yumin Qiu, Yueke Lin, Dapeng Ma, et al. 2020. “The E3 ubiquitin ligase TRIM7 suppressed hepatocellular carcinoma progression by directly targeting Src protein.” *Cell Death and Differentiation* 27 (6): 1819–1831. https://doi.org/10.1038/s41418-019-0464-9.

Zinman, Guy E, Shan Zhong, and Ziv Bar-Joseph. 2011. “Biological interaction networks are conserved at the module level.” *BMC Systems Biology* 5 (August): 134. https://doi.org/10.1186/1752-0509-5-134.