Differential Dependency Network and Data Integration for Detecting Network Rewiring and Biomarkers

Yi Fu

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Electrical Engineering

Yue Wang, Chair
Guoqiang Yu
Zhen Zhang
Charles Chancy
Alireza Haghighat

December 5th, 2019
Arlington, Virginia

Keywords: molecular data integration, differential network analysis, biomarker.

Differential Dependency Network and Data Integration for Detecting Network Rewiring and Biomarkers

Yi Fu

ABSTRACT

Rapid advances in high-throughput molecular profiling techniques enabled large-scale genomics, transcriptomics, and proteomics-based biomedical studies, generating an enormous amount of multi-omics data. Processing and summarizing multi-omics data, modeling interactions among biomolecules, and detecting condition-specific dysregulation using multi-omics data are some of the most important yet challenging analytics tasks.

In the case of detecting somatic DNA copy number aberrations using bulk tumor samples in cancer research, normal cell contamination becomes one significant confounding factor that weakens the power regardless of whichever methods used for detection. To address this problem, we propose a computational approach – BACOM 2.0 to more accurately estimate normal cell fraction and accordingly reconstruct DNA copy number signals in cancer cells. Specifically, by introducing allele-specific absolute normalization, BACOM 2.0 can accurately detect deletion types and aneuploidy in cancer cells directly from DNA copy number data.

Genes work through complex networks to support cellular processes. Dysregulated genes can cause structural changes in biological networks, also known as network rewiring. Genes with a large number of rewired edges are more likely to be associated with functional alteration leading phenotype transitions, and hence are potential biomarkers in diseases such as cancers. Differential dependency network (DDN) method was proposed to detect such network rewiring and biomarkers.

However, the existing DDN method and software tool has two major drawbacks. Firstly, in imbalanced sample groups, DDN suffers from systematic bias and produces false positive differential dependencies. Secondly, the computational time of the block coordinate descent algorithm in DDN increases rapidly with the number of involved samples and molecular entities. To address the imbalanced sample group problem, we propose a sample-scale-wide normalized formulation to correct systematic bias and design a simulation study for testing the performance. To address high computational complexity, we propose several strategies to accelerate DDN

learning, including two reformulated algorithms for block-wise coefficient updating in the DDN optimization problem. Specifically, one strategy on discarding predictors and one strategy on accelerating parallel computing. More importantly, experimental results show that new DDN learning speed with combined accelerating strategies is hundreds of times faster than that of the original method on medium-sized data.

We applied the DDN method on several biomedical datasets of omics data and detected significant phenotype-specific network rewiring. With a random-graph-based detection strategy, we discovered the hub node defined biomarkers that helped to generate or validate several novel scientific hypotheses in collaborative research projects. For example, the hub genes detected by the DDN methods in proteomics data from artery samples are significantly enriched in the citric acid cycle pathway that plays a critical role in the development of atherosclerosis.

To detect intra-omics and inter-omics network rewirings, we propose a method called multiDDN that uses a multi-layer signaling model to integrate multi-omics data. We adapt the block coordinate descent algorithm to solve the multiDDN optimization problem with accelerating strategies. The simulation study shows that, compared with the DDN method on single omics, the multiDDN method has considerable advantage on higher accuracy of detecting network rewiring. We applied the multiDDN method on the real multi-omics data from CPTAC ovarian cancer dataset, and detected multiple hub genes associated with histone protein deacetylation and were previously reported in independent ovarian cancer data analysis.

Differential Dependency Network and Data Integration for Detecting Network Rewiring and Biomarkers

Yi Fu

GENERAL AUDIENCE ABSTRACT

We witnessed the start of the human genome project decades ago and stepped into the era of omics since then. Omics are comprehensive approaches for analyzing genome-wide biomolecular profiles. The rapid development of high-throughput technologies enables us to produce an enormous amount of omics data such as genomics, transcriptomics, and proteomics data, which makes researchers swim in a sea of omics information that once never imagined. Yet, the era of omics brings new challenges to us: to process the huge volumes of data, to summarize the data, to reveal the interactions between entities, to link various types of omics data, and to discover mechanisms hidden behind omics data.

In processing omics data, one factor that weakens the strengths of follow up data analysis is sample impurity. We call impure tumor samples contaminated by normal cells as heterogeneous samples. The genomic signals measured from heterogeneous samples are a mixture of signals from both tumor cells and normal cells. To correct the mixed signals and get true signals from pure tumor cells, we propose a computational approach called BACOM 2.0 to estimate normal cell fraction and corrected genomics signals accordingly. By introducing a novel normalization method that identifies the neutral component in mixed signals of genomic copy number data, BACOM 2.0 could accurately detect genes' deletion types and abnormal chromosome numbers in tumor cells.

In cells, genes connect to other genes and form complex biological networks to perform their functions. Dysregulated genes can cause structural change in biological networks, also known as network rewiring. In a biological network with network rewiring events, a large quantity of network rewiring linking to a single hub gene suggests concentrated gene dysregulation. This hub gene has more impact on the network and hence is more likely to associate with the functional change of the network, which ultimately leads to abnormal phenotypes such as cancer diseases. Therefore, the hub genes linked with network rewiring are potential indicators of disease status or known as biomarkers. Differential dependency network (DDN) method was proposed to detect network rewiring events and biomarkers from omics data.

However, the DDN method still has a few drawbacks. Firstly, for two groups of data with unequal sample sizes, DDN consistently detects false targets of network rewiring. The permutation test, which uses the same method on randomly shuffled samples is supposed to distinguish the true targets from random effects, however, is also suffered from the same reason and could let pass those false targets. We propose a new formulation that corrects the mistakes brought by unequal group size and design a simulation study to test the new formulation's correctness. Secondly, the time used for computing in solving DDN problems is unbearably long when processing omics data with a large number of samples scale or a large number of genes. We propose several strategies to increase DDN's computation speed, including three redesigned formulas for efficiently updating the results, one rule to preselect predictor variables, and one accelerating skill of utilizing multiple CPU cores simultaneously. In the timing test, the DDN method with increased computing speed is much faster than the original method.

To detect network rewirings within the same omics data or between different types of omics, we propose a method called multiDDN that uses an integrated model to process multiple types of omics data. We solve the new problem by adapting the block coordinate descending algorithm. The test on simulated data shows multiDDN is better than single omics DDN.

We applied DDN or multiDDN method on several datasets of omics data and detected significant network rewiring associated with diseases. We detected hub nodes from the network rewiring events. These hub genes as potential biomarkers help us to ask new meaningful questions in related researches.

Acknowledgements

I would like to express my most sincere gratitude to my advisor, Dr. Yue Wang, who not only kindly offers his academic guidance and support through my entire PhD study but also shows me an excellent example of a leader with confidence and optimism, a scholar with humbleness and endless curiosity, a teacher with responsibility and a warm heart. I am also deeply grateful to Dr. Zhen Zhang at Johns Hopkins Medical Institute, who offered me a precise opportunity to research at Johns Hopkins Medical Institute with an enjoyable working environment, and also guided me with his talented insights of statistics and machine learning.

My sincere gratitude also goes to the rest of my committee members, Dr. Guoqiang Yu, Dr. Charles Chancy and Dr. Alireza Haghighat, for their patience and feedback to my dissertation work. I would thank Dr. Yu in particular, for his genius ideas and academic suggestions and for all his help through my study in CBIL.

I would like to give my deepest gratitude to my wife, Yuanyuan Tang. We met and get married in my third year of PhD study. Her love is the most precise thing in my life. Her encouragement and support have been instrumental in my dissertation work.

My gratitude is extended to my colleagues and collaborators: Dr. Bai Zhang, Dr. Ye Tian, Dr. Niya Wang, Dr. Xu Shi, Dr. Xiao Wang, Lulu Chen, Yizhi Wang, and Yingzhou Lu from CBIL at Virginia Tech; Dr. Felix Ma from Johns Hopkins Medical Institute; Dr. David Herrington from Wake Forest University.

My last but not least thanks to my parents Guanxiao Fu and Rongfang Meng, and to my brother Haocheng Meng, for their encouragement and support, for their love throughout my life.

Table of Contents

List of Figures

List of Tables

List of Abbreviation

ANOVA	Analysis of variance
AUC	Area under the curve
BCD	Block coordinate descent
BN	Bayesian network
CPTAC	Clinical Proteomic Tumor Analysis Consortium
DEA	Differential expression analysis
DDN	Differential dependency network
DN	Differential network
FDR	False Discovery Rate
FN	False negative
FP	False positive
FDR	False Discovery Rate
FN	False negative
FP	False positive
GGM	Gaussian graphical model
gLASSO	Graphical LASSO
GO	Gene ontology
GPAA	Genomic and Proteomic Architecture of Atherosclerosis
LASSO	Least absolute shrinkage and selection operator
LC-MS	Liquid chromatography coupled with mass spectrometry
kDDN	Knowledge-fused differential dependency network
KSR	Kinase-substrate interaction
OV	Ovarian cancer
mRNA	Message RNA
miRNA	Micro RNA
NCI	National Cancer Institute
NS	Neighborhood selection
PPI	Protein-protein interaction
PTM	Post translational modification

RNA-seq	RNA sequencing
ROC	Receiver operating characteristic
TCGA	The cancer genome atlas

ATTRIBUTION

Several colleagues aided in the writing and research behind one of my chapters presented as part of the dissertation. A brief description of their contributions is included here.

Chapter 4: BACOM 2.0 facilitates quantification of somatic copy number alterations and estimation of tumor purity

Guoqiang Yu and Yue Wang, served as co-authors and help to develop the framework and wrote the manuscript.

Niya Wang, served as a co-author and helped to apply the UNDO method on the TCGA datasets.

Zhen Zhang, served as a co-author and provided biostatistics expertise.

Douglas A. Levine and Robert Clarke, served as co-authors and provided biomedical expertise to interpreting the results.

All co-authors contributed to editing the manuscript.

Chapter 1. Introduction

1.1. Background

We witnessed the start of the human genome project decades ago and stepped into the era of omics since then. Omics are comprehensive approaches for analyzing genome-wide biomolecular profiles. The rapid development of high-throughput technologies enables us to produce an enormous amount of omics data such as genomics, transcriptomics, and proteomics data, which makes researchers swim in a sea of omics information that once never imagined. Yet, the era of omics brings new challenges to us: to process the vast volumes of data, to summarize the data, to reveal the interactions between entities, to integrate various types of omics data, and to discover mechanisms hidden behind omics data.

For each type of biomolecules involved in cell activity, there is a corresponding type of omics: genomics for DNA molecules, transcriptomics for RNA molecules, proteomics for proteins, and. These three omics are the central parts of molecular biology researches. In the meantime, there are also other types of omics including metabolomics, epigenomics, glycoproteomics, etc.

In the genomics study, DNA copy number is defined as the number of copies of a specific DNA sequence within a cell. Human cells are diploid cells which carry two complete sets of chromosomes, hence the neutral copy number of gene in human somatic cells is two. Deletion or duplication of DNA regions can cause the change of copy numbers, also known as copy number variation (CNV). The change of copy number is either inherited as germline change, or caused by somatic DNA structural changes. Some researchers prefer to call the somatic copy number changes as copy number alteration (CNA) to distinguish from the germline CNV. The somatic copy number alteration is the key feature of many cancer diseases. For example, the complete loss of copy

numbers of tumor suppressor genes leads to the absence of gene functions of tumor-suppressing; and the copy number gains of oncogenes are associated with the development of cancer cells. Some types of cancer are highly associated with copy number changes. For example, the human high-grade serous cancer (HGSC) is the most common type of ovarian cancer and distinguishes from other subtypes by a large burden of copy number gains or losses.

Network data analysis is one of the most powerful tools for studying biomolecular interactions from omics data (Hu, et al., 2016; Yan, et al., 2016). Network analysis could better discover the interactions between entities and present the underlying data distributions and is also very helpful in visualizing the data interactions(Mateos, et al., 2019). Interactions between biomolecules, including gene transcription, protein translation, gene regulation, protein-protein interaction, and many other types, is crucial to every living cell. Discovering biomolecular interactions and is one of the central concerns of modern molecular biology(Kitano, 2002).

For many real-world networks and biological networks, a static network structure is not able to depict the dynamic nature of these networks. For example, the regulatory networks could vary across different cell types, stages of the disease development, or statuses of driver gene mutation (Ideker and Krogan, 2012). In response to DNA damage, structural change of genetic networks could lead to the activation of cancer (Bandyopadhyay, et al., 2010). Differential network (DN) analysis which focuses on the changes between different network topologies is gradually used by more and more data analysts and biomedical researchers who are interested in comparing between disparate groups at a network level (Califano, 2011; Creixell, et al., 2012; Ha, et al., 2015; Hudson, et al., 2009; Zhang, et al., 2016). Network rewiring, defined as the topological change of a network, is one of the critical features that DN analysis seeks to discover. Network rewiring could reveal critical and essential changes in a network. To detect significant network rewiring,

the DDN method, as one of the earliest DN analysis methods, is proposed and has been tested via several independent studies (Zhang, et al., 2009). It was further improved with joint regression solving (Zhang, et al., 2011; Zhang and Wang, 2010), and was also enhanced with integrated prior-knowledge known as kDDN (Tian, et al., 2014; Tian, et al., 2011). It is implemented in both R and Java language as a DN analysis tool (Tian, et al., 2015).

1.2. Motivation

In genomics studies, somatic copy number alteration (CNA) is one of the important features studied in biomedical researches especially in cancer researches (da Cruz, et al., 2018; Kuo, et al., 2010). The accurate detection of CNA from genomic data relies on the normalization procedure to calibrate the neutral copy number. However, the simple normalization methods like mean or median calibration would fail in the case of high genomic instability. Driven by the need for processing genomic data and accurately detecting DNA copy number variations, we are encouraged to develop a computation tool called BACOM 2.0.

Though DDN is a powerful differential network analysis tool (Zhang, et al., 2011), some drawbacks hinder the DDN tool from being more widely used in network study or biomedical data analysis. Firstly, the initial design of the DDN method is for comparing two groups with equal sizes of samples. In practice, we notice that for imbalance data, i.e., two groups have different sample sizes, DDN seems to detect network rewiring favoring one condition over the other. A systematic bias caused by data imbalance may exist and need to be correct. Secondly, when limited by the algorithm running time, DDN can handle only dozens of features and samples. The actual computation time grows more than cubically with a feature scale. In biological network analysis, the feature size in gene regulatory network inference could be as large as tens of thousands. The development of accelerated DDN learning algorithms is an urgent need for extending DDN's

application to broader fields. Thirdly, for omics data analysis, the current DDN method is designed for analyzing a single type of omics data. Though a simple method of merging multiple data matrices into a single matrix could theoretically extend the DDN method to multi-omic data, this kind of integration method would double to triple the feature scale and take impractically long computation time. This simple merging method also ignores inter-omics regulation knowledge. We believe a multi-omics integration method could take advantage of the additional knowledge such as directional inter-omics interactions (Buescher, et al., 2016), to effectively reduce the feature scale and increase the accuracy in differential network learning. With the advent of more are more multi-omics research projects, DDN needs such an integration method to discover novel network rewirings facilitated by multi-omics data.

Network analysis would not only help researchers of bioinformatics and biologists but also in various fields like transportation, economics, social relation study, etc. (Dong, et al., 2019). We aim to improve DDN's utility in both accuracy and speed to make it applicable to more broad fields.

Biomarkers can be effective indicators used in the early detection of disease (Yan, et al., 2016). With DDN as a powerful tool, we would like to perform differential network analysis to different fields of omics study and find significant network rewiring for various omics data researches to reveal unique condition-specific interaction and to help biologists generate novel scientific hypothesis. Notably, we believe that hub genes in scale-free bio-networks play essential roles and server as biomarkers for early disease diagnosis or potential therapeutic targets.

While the kDDN tool can integrate condition-specific gene expression data and prior knowledge (Tian, et al., 2014), we have long known that biological networks involve more players(Hecker, et al., 2009; Madhamshettiwar, et al., 2012). For example, many different

regulatory factors such as cis-factors of promoters, gene enhancers, and transcription factors regulate gene expression(Kitano, 2002; Song, et al., 2011). Though some work like genotype-expression correlation analysis has been done to include two types of omics data(Yuan, et al., 2011), many existing network studies for omics studies are still targeting to find intra-omics interactions from single-omics data(Ahmed and Xing, 2009; de Chassey, et al., 2008; Dong, et al., 2015; Friedman, et al., 2000; Zhao, et al., 2006), and hence incapable to capture inter-omics regulations. Therefore, we are motivated to make better use of available multi-omics data and peek into the black box of inter-omics interaction by integrating multiple types of regulators into network analysis.

Designing an integrated data model could help us to develop DDN into a new tool that capable of detecting network rewiring events from multi-omics data. With such a tool, we can significantly improve our understanding of the biological gene regulatory networks and cancer networks.

1.3. Objectives

Motivated by the needs mentioned above, we set our goals in this dissertation work as follows:

1. Design a novel approach to accurately quantify the copy number signals from genomics data to copy number signals and estimate the tumor purity;
2. Improve the methodology of DDN method:
    a. Correct the systematic bias brought by data imbalance
    b. Accelerate the core algorithm of differential network learning in the DDN method
3. Apply the DDN method to various single omics data sets

a. Detect network rewiring and gain unique insights of specific biomedical phenotypes

b. Discover potential biomarkers in detected differential networks

4. Design a novel differential network analysis tool (multiDDN) to integrate multi-omics data:

    a. Design an integrated multi-layer data model, incorporate prior knowledge of directional inter-omics regulation; validate the model's effectiveness

    b. Adapt the DDN method with the design data model to process and integrate multi-omics data; detect network rewiring events in piloting works on multi-omics data.

1.4. Organization of the dissertation

We organize the remainder of this dissertation as follows:

In Chapter 2, we overview the history of biological omics data, introduce three types of omics data that widely used in biology and biomedical researches. We also briefly overview some a few public databases of omics data and knowledge.

In Chapter 3, we overview the methodology development of network analysis in biology, introducing a few network analysis methods and tools used in omics data analysis. We also focus on two particular categories of network analysis: sparse networks and differential networks.

In Chapter 4, we introduce two proposed methods of processing omics data. The first method, BACOM 2.0, can accurately estimate tumor purity from genomics data and generate purity-corrected copy number signals. We propose a novel absolute normalization method to correct the sample heterogeneity brought by normal cell contamination. Experiment results on both realistic simulation data and benchmarking real data sets are included to support our claims. The

second method is designed to detect significant aberrant in methylation data statistically. We introduce the method's principles and one application on epigenomics data from mouse models.

In Chapter 5, we introduce the differential networks analysis tool of DDN, overview its methodologies. We discuss the issue of imbalanced group size and redesign the formulas for the DDN method, with supporting evidence from a simulation study. We introduce block coordinate descent algorithm for fast solving optimization in DDN by proposing several accelerate strategies. The newly proposed strategies significantly increase the computing speed of solving the DDN optimization problem. In the final part of this chapter, we report the pilot DDN analyses on the omics data generated from three different biomedical domains, including atherosclerosis, ovarian cancer, and psychological disorders.

In Chapter 6, we introduce our proposed method of multiDDN to integrate multiple types of biology omics data. We incorporate directional inter-omics regulation into gene regulatory networks and design a multi-layer network model for integrating at most three types of omics data. We report a simulation study and a real data experiment at the end of this chapter.

In Chapter 7, we propose a hub node detecting method to discover biomarkers in differential networks. We overview the characteristics of biology networks and the importance of hub nodes in the gene regulatory networks. Based on random graph theory, we design an approach to detect hub nodes in differential networks. With the proposed approach, we discovered a few biomarkers that show high potential as key regulators in the pathways associated with the phenotype of interest.

In Chapter 8, we summarize the author's contributions in this dissertation and point out some directions for future work. We list the author's biography and relevant publication records in the appendix.

Chapter 2. Overview of Omics Data

2.1. The era of omics

We have witnessed the burst development of high-throughput measuring technologies of biomolecules. Marked by the initial sequencing and analysis of the human genome by the Human Genome Project (HGP)(nature, 2001) at the beginning of the new millennium, we humans entered the new era of omics. Numerous omics platforms and technologies have been developed since then and put into large scale practice. Genomic testing tools even step into everyone's life. In this era of omics, researchers can perform large-scale experiments to get high-throughput data from genes, transcripts, and proteins of their interest, to detect biomolecules' existence, measuring the abundance, discovering the composition, etc.(Patti, et al., 2012).

Omics are comprehensive approaches for analyzing genome-wide genetic or other biomolecular profiles(Cancer Genome Atlas Research, 2011). For example, while genetics study is about single genes, genomics study focuses on a genome-wide set of genes and their inter-relationships. The omics data allows researchers to study complex interactions between biomolecules that influence the phenotype, such as the disease status in patients(Barabasi, et al., 2011; de Chassey, et al., 2008).

Ever since the era of omics, the rapid development of omics related technology provides numerous platforms to acquire high-throughput omics data from various types of biomolecules. These technologies include microarray chips, next-generation sequencing, mass spectrometry, single-cell sequencing, and many more others. Meanwhile, a large number of omics data processing software for data pre-processing, quality control, quantification, data analysis, functional annotation, and many others are available for biology researches and data analysts.

Numerous high-quality online data resources and databases are open to the public domain. Hundreds of thousands of computational approaches and analytical tools are developed every year for downstream omics data analysis.

2.2. Types of omics data

There are three major types of biomolecules involved in cell activities: DNA, RNA, and protein. The corresponding omics studies of genomics, transcriptomics, and proteomics are the central part of molecular biology researches. In the meantime, analogous to genomics and proteomics, comprehensive studies on the complete set of metabolites, epigenetic modifications (e.g., methylation, chromatin openness), glycoprotein established the new fields of metabolomics, epigenomics, and glycoproteomics. In this section, we will briefly overview the three types of omics data used in this dissertation: genomics data, transcriptomics data, and proteomics data.

2.2.1. Genomics Data

Genomics is the study of the genomes encoded in DNA sequences on the genetic material of a cell. DNA sequences include gene coding regions and non-coding regions. For different research interests, researchers may choose to test on the whole DNA sequences or coding regions only. Genomics study focuses on deciphering DNA sequences, analyzing the structure and function of genetic regions, detecting polymorphisms of the genome.

With the advent of polymerase chain reaction (PCR), scientists can probe the much smaller amount of DNA materials, even within a single cell. There are two categories of approaches to detect DNA molecules' structural variations. The first one is microarray chip technology based on the principle of DNA chain matching. It uses pre-designed probes to detect the existence of a specific DNA sequence and quantify its relative amount. The second one is sequencing technology,

or in most time the next-generation sequencing (NGS). NGS breaks DNA chains into smaller fragments, and read the fragments' sequences base by base.

DNA mutation, especially the single nucleotide polymorphisms (frequently called SNPs), is critical in determining the genotypes that heavily associated with genetic diseases and various types of cancer. Another one of the important DNA quantitative measures is DNA copy number, namely the number of copies of one DNA sequence in a cell(Hartwell, et al., 2008). Human cells are diploidy, meaning that most genes in a human cell have precisely two copies located on paternal and maternal chromosomes. DNA regions with copy numbers more than or less than the neutral number of 2 are called DNA copy number variations (CNV) of amplification or deletion. Some researchers prefer to call somatic changes of copy number as copy number alteration(CNA)(Li, et al., 2009) or somatic copy number alteration(SCNA)(Zack, et al., 2013). Due to the integer nature of copy number, there are only two possible types of DNA deletions: homo-deletion with the copy number of 0, and hetero-deletion with the copy number of 1.

The gene dosage effect is described as the significant correlation between gene dosage, which is the copy number of a particular gene present in a genome, and its gene expression level (Gardiner, 2004; Hartwell, et al., 2008; Nussbaum, et al., 2015). The mechanism behind this mostly positive correlation is straightforward: the more copies of one gene exist in a cell, the larger the number of gene regions can be transcribed simultaneously, and the higher the abundance level of its transcripts, i.e., gene expression. The effect of cis-correlation also exists to genes of the same chromosome. One type of cis-correlation is caused by the DNA structural changes which frequently occur on a long and continuous section in a chromosome. The continuous DNA structural variations, either DNA deletion or amplification, could cover multiple adjacent gene regions (Hartwell, et al., 2008). Therefore, genes with a short distance of genomic locations are

more likely to have a positive correlation between their gene copy numbers. This positive correlation further passes via the gene dosage effect to the correlation between their gene expressions. Due to the existence of gene dosage effect and cis-correlation, additional information on gene copy numbers could bring extra help to gene expression analysis(Yuan, et al., 2011).

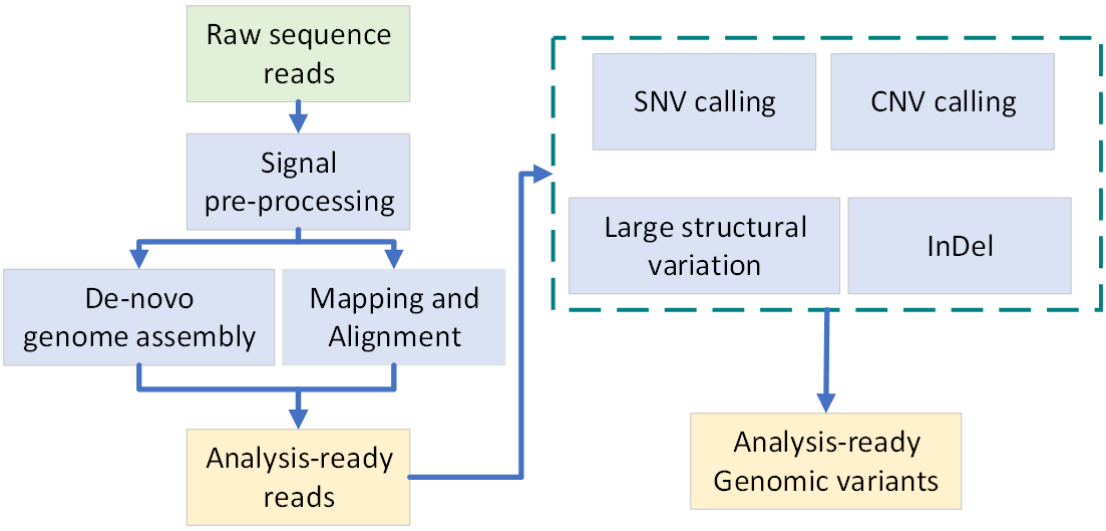


Figure 1- workflow for processing DNA sequencing data

Figure 1 shows a typical workflow for processing DNA sequencing data. Firstly, the raw sequence reads are acquired and passed through signal pre-processing and quality control. The reads are then assembled either with a de-novo approach or mapped to an existing reference genome sequence for alignment. After that, the analysis-ready reads are put to variant calling procedure to summarize genomic information of single-nucleotide polymorphism (SNP), copy number variant (CNV), large structural variations (e.g., chromosome dislocation), inserts/deletions (InDel) and other DNA structural variations. These genomic variations are the final product of genomic data quantification and summarization pipeline and are ready for downstream omics studies.

2.2.2. Transcriptomics Data

Transcriptomics is the study of the complete set of RNA transcripts that are produced by the genome. By function, RNA molecules could be further classified to three categories(Hegde, et al., 2003): 1. message RNA (mRNA) which is the majority family of RNA molecules, carry the coding sequence information from transcribed genes and could further translate into proteins (amino acid sequence). 2. micro RNA (miRNA), which is an essential part of gene regulation. 3. non-coding RNA (ncRNA). Unlike the double helix structure of DNA molecules, RNA molecules have only one strand of nucleotide chains, and hence could attach to other single-chain molecules to activate/deactivate or inhibit the biological function. The abundance of mRNA largely determines its downstream translation product of protein. And by RNA splicing mechanism, one gene's exons can be combined in various ways and transcribed into multiple isoforms of mRNA which could translate to different end products involved in different signaling pathways. RNA sequencing (RNA-Seq) technique could identify the isoforms of the gene transcript and quantify the expression level(Wang, et al., 2009).

Transcriptomics data have been widely used in large scale cancer research projects such as TCGA glioblastoma multiforme(TCGA, 2008) project, ovarian cancer project(Cancer Genome Atlas Research, 2011), breast cancer project(Mertins, et al., 2016) to characterize cancer subtypes, and discover gene expression bio-markers defining these cancer subtypes and subtype-specific deregulation between genes.

2.2.3. Proteomics Data

Proteomics researches focus on comprehensively studying the function and structure of protein molecules(Hutchins, 2014). The first level of protein structure, also known as the primary structure of a protein, is the sequence of amino acids in the chain of a polypeptide molecule. Unlike

DNA sequencing or RNA sequencing, there is no pairing "code" to amino acids. Therefore, in the proteomics study, researchers usually use mass spectrometry (MS) platforms to identify the fragments of proteins, also known as peptides. Peptides are chains of amino acids and are formed by digesting proteins into fragments of small-size chains. If a peptide contains unique sequence information that only occurs in one protein molecule, it then could be used as an indicator of the existence of that protein. In a typical untargeted proteomic study, researchers enzymatically digest proteins into smaller peptides in the first stage, then analyze the digested peptides by liquid chromatography coupled with mass spectrometry (MS) or tandem MS (MS/MS)(Tsai, et al., 2016). The MS/MS spectra data contains quantitative information of the detected peptides(Ranjbar, et al., 2014). Proteins are then identified by de novo sequencing or database searching and are associated with the peptides(Chen, et al., 2018). Labeling technologies can attach chemical labels to protein samples to allow relative expression comparison between samples. Widely used labeling methods include isobaric tagging for relative and absolute quantification (iTRAQ), and tandem mass tags (TMT) multi-plex labeling technology(Chen, et al., 2018). Multiple (4 or 8 samples for iTRAQ; 6, 8 or 10 samples for TMT) differentially labeled samples could be pooled in one run and analyzed by LC-MS/MS simultaneously for protein identification and quantification,

Due to the nature of proteins, proteomics data obtained by MS/MS technology usually contain a large portion of missing data. There are two significant resources of missing data: one is due to the low expression level of the tested protein in the cell, and the other is due to improper sample preparation or platform measuring bias(Lazar, et al., 2016). Missing data has haunted proteomics data analysts for decades, and various strategies were proposed to solve this problem(Pedreschi, et al., 2008; Webb-Robertson, et al., 2015).

Post-translational modification (PTM) such as glycosylation, phosphorylation, and acetylation may alter protein's structure and hence activate/deactivate or modify its function(Zhang, et al., 2016). Some researchers also proposed to establish new omics field for PTM, like glycoproteomics(Tian and Zhang, 2010; Wuhrer, et al., 2007).

2.3. Databases and resources of omics data

Numerous databases have emerged in this era of omics(Hutchins, 2014). There is an increasing diversity of data sources in the public domain, providing gateways to access the information about genes, transcripts, proteins, metabolites, pathways, and many others. The provided information set includes nucleotide sequence and location of genes, the amino acid sequence of proteins, functional annotation to biomolecules, association with phenotypes, regulatory relation with other biomolecules, etc. Depending on the information they provide, we can roughly categorize these databases into the following groups(Hutchins, 2014): 1. DNA/RNA/protein sequence database; 2. DNA mutations, DNA structural variations; 3. Gene functions; 4. Quantitative omics data: DNA copy number, transcript expression, protein expression; 5. Gene regulatory pathway and ontology resources; 6. Protein-protein interaction database; 6. PTM database. We will briefly overview a few databases used in this dissertation.

TCGA, short for The Cancer Genome Atlas (TCGA), is a landmark cancer genomics program held by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI). TCGA projects molecularly characterized over twenty thousand primary cancer samples and matched normal samples covering 33 different types of cancer. TCGA generated and stored over 2.5 petabytes of genomic, epigenomic, transcriptomic, and proteomic data, and is publicly available for anyone in the research community to use.

KEGG, short for the Kyoto Encyclopedia of Genes and Genomes, is a database for knowledge about functions and utilities of the biological systems. The KEGG pathway database makes the core of the KEGG resource. The wiring diagram database, as a "computer representation" of the biological system, is a collection of pathway maps that integrate building blocks of genes, proteins, RNAs, and many other chemical compounds, along with wiring diagrams of the bio-systems.

ENCODE, short for The Encyclopedia of DNA Elements, is a comprehensive resource of human genomic data and knowledge. The project is launched by the US National Human Genome Research Institute (NHGRI) aims to identify functional elements in the human genome. Its primary goal is to determine the role of the noncoding part of the genome, including genes' promoter regions, which could bind to multiple regulating transcription factors.

Ingenuity Pathway Analysis (IPA) is a commercial database and a web-based software application for comprehensive analysis, integration, and interpretation of omics data. It provides multiple types of pathway analysis, including identifying key regulators, predicting downstream effects on biological processes and providing targeted data on genes, proteins, and other molecules.

Chapter 3. Overview of Network Analysis in Biology

3.1. Introduction

The rapid advance in high-throughput genomic technologies (Hartwell, et al., 2008) and the growing number of large-scale public biological datasets (Hutchins, 2014) provide ample opportunities for bioinformatics researchers to study cellular activities at the individual gene level and also at a higher level of biological networks. The scientific research society has made significant progress in discovering new genes, transcripts, proteins and their functions. Uncovering the regulatory network structure between the genes and proteins still remains one of the systems biology's key goals (Klipp, et al., 2016).

Modern data analysis usually deals with large chunks of structured data. The structure buried in data usually carries critical information about the nature of data. Networks is a powerful data presentation tool for describing the structures hidden in data. For biology study, especially in the modern field of molecular biology, network analysis providing a flexible way to depict and present the interacting relationships between genes and entities of biomolecules.

The problem of network construction, or graph learning, can be summarized as follows: given N observations of P variables, represented in a data matrix X, with or without some prior knowledge (for example, data distribution, interaction model, etc.) about the data, the goal of graph learning is to infer the relationship between the variables which take the form of a graph G. Each column in the data matrix X becomes a graph signal defined on the node set of G; and the whole observations could be represented as X = F(G), where F represents a generative function on the graph.

Graph learning is the central part of network analysis. For structured data, graphs can capture the underlying geometry of interaction between entities, which is essential to data processing and analysis. The inferred graph, representing the network structure, could help in predicting future network status from data that share the same structure. For example, with a suitable perturbation model, researchers could use the inferred biological network to predict cell activity, metabolite level, immune system response, disease progression, etc.

3.2. Graphical model for network construction

A meaningful data model, or accurate prior knowledge on network structure, could significantly help to guide the process of inferring the graph, and lead to a graph topology that more accurately represents the intrinsic relationship between entities. The main challenge in graph leaning is to define such a model function F which converts the relationship between observed data matrix X into the inferred graph G. Inferring graph topologies from observations in most cases is an ill-posed problem, which means there may exist multiple ways to associate the graph topology with the observed data. A major category of approaches for learning graph topologies is statistical graph models. For statistical models, the generative function F is viewed as a function which draws realization over variables from the assumed probability distribution.

One of the most widely used statistical models is the probabilistic graphical model, in which the vertices in the graph represent the variables, and the graph topology encodes the conditional dependency or independency between these variables. The learning of graph topology in graphical models is then equivalent to learning the probability distribution of the random variables.

There are two main types of graphical models, directed networks also known as Bayesian networks or belief networks (BN), and undirected networks also known as Markov random fields (MRFs)(Dong, et al., 2019). In MRF, local neighborhoods of the graph capture variables' independence structure.

MRF admits a more straightforward representation of conditional dependence or independence. An MRF is an undirected graph with its nodes as a set of random variables that have a Markov property. We are particularly interested in the pairwise Markov property which states that, if and only if two nodes in the graph are not linked by an edge, their corresponding two variables are conditionally independent. In other words, the node-to-node dependency determines the network structure in a network with F Markov property.

For continuous variables, Gaussian MRF (GMRF), also known as the Gaussian graphical model (GGM), is the main class of MRF. There are various methods proposed to estimation the node dependency in GGM. For example, correlation networks using sample correlation or some other functions like Gaussian radius basis kernel function to measure the similarity between samples. These similarity-function-based methods mostly are purely based on observations, without adapting any prior knowledge or data models, and hence could be sensitive to noises.

Since a GGM graph is a representation of node-to-node relationships, it is self-evident that we can learn the graph by learning the pair-wise relationship for each node in the graph. In other words, graph learning is equivalent to finding the connecting nodes which are also called neighbors, for each node in the graph. Therefore, it is natural to assume that, for any node in the graph, its observed distribution could be represented by its connecting neighbors. This assumption is the base of the graph learning approach of neighborhood selection. The neighborhood selection

approach is intuitive for transforming GGM learning into node-wise similarity estimation, and certain theoretical guarantees prove its effectiveness (Meinshausen and Bühlmann, 2006).

Methods belong to the neighborhood selection family mainly vary on the ways of representation from neighboring nodes. For a sparse network, a sparse linear combination would be a natural and simple choice for presentation. For example, Yuan & Lin (2007) and Banerjee et al. (2007) use the LASSO method(Tibshirani, 1996) of sparse linear regression to approximate the observation at each variable.

3.3. Sparse network and the LASSO optimization problem

For most real-world networks, such as social and computer networks, the number of connected edges is smaller than the possible maximum number of connections within that network. Such networks are called sparse networks, opposite to complete or dense networks. Notably, we consider gene regulatory networks (GRN) used in omics study as sparse networks, for the reasons that genes are regulated not by all but only a subset of biomolecules in specific signaling pathways. The sparse networks are usually also scale-free networks that have a power-low distribution of node degrees.

The LASSO optimization problem was proposed by Tibshirani (1996) who laid the foundation of sparse variable selection. Since then, numerous types of LASSO-like optimization problems, such as elastic net (Zou and Hastie, 2005), group LASSO (Meier, et al., 2008), fused LASSO (Tibshirani, et al., 2005), multi-response LASSO (Hadfield, 2010), etc., were developed and put into practice. The basic LASSO optimization is:

$$\boldsymbol{\beta} = \arg\min_{\boldsymbol{\beta}} \left\{ f(\boldsymbol{\beta}) = \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda |\boldsymbol{\beta}| \right\}$$

where y is the response variable vector, X is the data matrix of predictors, β is the LASSO regression coefficient vector, and $\lambda \geqslant 0$ is a tuning parameter of controlling sparsity in β.

A primary reason for using LASSO regression is that the L1-norm penalty tends to make some entries of regression coefficients to exactly zero, and therefore it produces a sparse solution that performs variable selection. This sparsity property encouraged researchers to estimate sparse undirected graphical models using LASSO regularization. Meinshausen and Bühlmann (2006) proposed a simple approach of neighborhood selection: estimate the network by fitting a lasso regression model to each node and select the neighboring nodes. Friedman, et al. (2008) proposed a graphical LASSO method that uses LASSO regression to estimation the whole precision matrix.

3.4. Network rewiring and differential networks

Genes do not act in isolation but instead work as part of complex networks to perform various cellular processes (Isalan, et al., 2008). Dysregulated genes cause many human diseases including various types of cancer. The DNA and epigenetic mutations within the gene region or its regulatory elements could cause topological changes in the signaling pathway or in the regulatory network structure. Network structural change can ultimately impair normal cell physiology and cause diseases. For example, cancer driver mutations on a transcription factor can alter its interactions with many of the target genes that are important in cell proliferation.

Genes in living cells have complicated interactions with each other, keep the cells well functioned, and make responses to various environmental stress or stimulation. In a word, the interaction of genes is one of the fundamental parts of life. Gene regulatory networks are context-specific and dynamic. Under different conditions, different components and mechanisms in gene regulatory network are selectively activated or deactivated (Califano, 2011). In response to internal

or external stimuli, the topology of the underlying biological network may change and the cellular components exert their functions through interactions with other elements in the network. At different time points, the signaling pathway in a biological network system could also change periodically. Comparing to static structure, a condition-depend structure of differential networks would better depict the dynamic nature of such networks.

More evidence shows that cancer is not merely a disease with genetic mutations, but is one driven by dysregulated genes in the signaling network with perturbations. Genetic lesion or abnormality cause alteration of protein function or expression level, and then lead to a higher-level change of signaling networks' dynamics structure, and ultimately, the cellular phenotype. The importance of genetic mutations hence should be assessed on their effect on genetic networks. The concept of network rewiring is introduced to describe the network structural changes between different states, such as between the normal status and the cancer status.

Network rewiring in biological networks is a general phenomenon between different cell types disease conditions and in evolution. The biological networks are not static but dynamic in nature: their structure may vary in different cell types or different developmental stages; they may lose particular connection or build new by-pass in some diseases or disorders. These inflect the networks' structural changes, or more preferred term "network rewiring", in biological networks. Network rewiring also has the potential to act as network attractors, which could lead to cancer status. Some biological evidence also showed that certain viruses or DNA damage could cause rewiring in genetic networks(Bandyopadhyay, et al., 2010).

Differential network analysis is used to detect network rewiring events by comparing individual networks from different groups and identify group-specific connections.

Chapter 4. BACOM 2.0 facilitates quantification of somatic copy number alterations and estimation of tumor purity

Fu Y, Yu G, Levine DA, Wang N, Shih IM, Zhang Z, Clarke R, Wang Y. BACOM2. 0 facilitates absolute normalization and quantification of somatic copy number alterations in heterogeneous tumor. Scientific Reports. 2015 Sep 9;5:13955.

4.1. Introduction

Processing the omics data is usually the first step in dry lab omics data studies that pave the road to downstream high-level analysis. In this chapter, we focus on a specific target of processing the omics data: the accurate estimation of the copy number signals from raw genomics data. We proposed a statistical approach called BACOM 2.0.

Sample heterogeneity describes the observation that a single biological sample contains different types of cells that show distinct genetic or phenotypic profiles. The different types of cells consist of subpopulations in the biological sample. These subpopulations of cells could be tumor cells from various cancer subtypes, or could be normal cells such as stromal cells or blood cells. The most common type of sample heterogeneity in tumor samples is normal cell contamination, which could be either brought by during bulk sample collection or from the intratumor micro-environment.

Copy number alterations (CNAs) associated with cancer are known to contribute to genomic instability and gene deregulation. However, due to sample heterogeneity, copy number signal of cancer cells is usually a mix of signals from both tumor cells and normal cells in the same sample. Yu, et al. (2011) proposed a Bayesian analysis of copy number mixtures (BACOM) method to detect genomic deletion types and to correct normal cell contamination in copy number data. We test the BACOM method on two simulated and two prostate cancer datasets with

promising results. However, in a subsequent analysis of the TCGA ovarian cancer dataset, the average normal cell fraction estimated by BACOM was found to be much higher than expected. Further inspection shows that the high genetic instability in tumor samples can cause inaccurate identification of the neutral copy number, and the existence of loss of heterogeneity (LOH) regions can bias the estimation of correlation coefficients between allele signals. These factors lead to inaccurate quantification of CNA and biased estimation of tumor purity. We are motivated to improve the method by correcting the aforementioned biases.

4.2. BACOM methodology and unresolved problem

BACOM is a statistically principled and unsupervised method that detects copy number deletion types (homozygous versus heterozygous), estimates normal cell fraction, and recovers cancer-specific copy number profiles, using allele-specific copy number signals. In a heterogeneous tumor sample, the measured copy number intensity is a mixture of the signals from both normal and cancer cells:

$$X_i = \alpha \times X_{\text{normal},i} + (1-\alpha) \times X_{\text{cancer},i}$$

where X_i is the observed copy number signal at the locus i, α is the unknown fraction of normal cells, $X_{\text{normal},i}$ and $X_{\text{cancer},i}$ are the latent copy number signals in normal and cancer cells at the locus ^{i}, respectively. Let $X_{A,i}$ and $X_{B,i}$ be the allele-specific copy number signals, $X_i = X_{A,i} + X_{B,i}$ are assumed to be independently and identically distributed random variables following a normal distribution $\mathcal{N}(\mu_{A+B}, \sigma^2_{A+B})$ whose mean μ_{A+B} and variance σ^2_{A+B} can be readily estimated by the sample averages. Allele-specific analyses focus on the deletion regions with distinct genotypes and detect the types of deletions by a model-based Bayesian hypothesis testing. Specifically, BACOM uses a novel summary statistic,

$$Y = \sigma_{A-B}^{-2} \sum_{i=1}^{L} (X_{A,i} - X_{B,i})^2$$

where σ_{A-B}^2 is the variance of $X_{A,i} - X_{B,i}$ in a length-L deletion region.

It has been shown that under homo-deletion, Y follows an L degree of freedom standard χ^2 distribution, given by

$$\chi^2(y;L) = \begin{cases} \frac{1}{2^{L/2}\Gamma(L/2)} y^{(L/2)-1} e^{-y/2} & \text{for } y > 0, \\ 0 & \text{for } y \le 0, \end{cases}$$

and under hemi-deletion, Y follows an L degree-of-freedom non-central χ^2 distribution, given by

$$\chi^2(y;L,\lambda) = \begin{cases} \frac{e^{-(y+\lambda)/2}}{2^{L/2}} \sum_{k=0}^{\infty} \frac{y^{L/2+k-1}\lambda^k}{\Gamma(k+L/2)2^{2k}k!} & \text{for } y > 0, \\ 0 & \text{for } y \le 0, \end{cases}$$

where $\lambda = L(2-\mu_{A+B})^2 \sigma_{A+B}^{-2}(1+\rho)/(1-\rho)$, ρ is the genuine correlation coefficient between $X_{A,i}$ and $X_{B,i}$, and Γ denotes the Gamma function. Since for a deletion region, we have

$$\begin{cases} E[X_i] = \alpha \times 2 + (1-\alpha) \times 0 = 2\alpha, & \text{if homo-deletion,} \\ E[X_i] = \alpha \times 2 + (1-\alpha) \times 1 = 1 + \alpha, & \text{if hemi-deletion,} \end{cases}$$

then, the average normal cell fraction $\bar{\alpha}$ across the whole genome can be estimated, as well as cancer-specific copy number profiles, given by

$$\hat{X}_{\text{cancer},i} = \frac{X_i - 2\bar{\alpha}}{1-\bar{\alpha}}, \text{ with } \alpha_{\text{homo}} = \frac{E[X_i]}{2}, \ \alpha_{\text{hemi}} = E[X_i] - 1$$

In our independent analyses of TCGA samples with BACOM, we confirmed unexpectedly higher average normal cell fractions. Upon closer examination of the interim results of the entire BACOM analytic pipeline, we found that many normal/amplified copy regions and hemi-deletions were misclassified as homo-deletions. This observation explains well the suspected overestimation of normal cell fraction, since α will be overestimated when non-deletion regions are wrongly used, or α_{homo} is applied to hemi-deletions ($\alpha_{homo} > \alpha_{hemi}$). Thus, we propose that this elevated estimate is the combined result of imprecise signal modeling and normalization, particularly in the presence of copy-neutral loss of heterozygosity (LOH) and aneuploidy. For example, if a non-deletion region is firstly misclassified as deletion due to imprecise signal normalization, it can be further misclassified as homo-deletion in the cases of allelic balance. Moreover, if the value of ρ is firstly underestimated due to copy-neutral LOH (allelic-imbalance) contamination in normal/allelic-balanced regions, hemi-deletion will then be misclassified as homo-deletion caused by a much-reduced signal-to-noise ratio.

To accurately estimate somatic copy number aberrations (SCNAs) in cancer research, we established a generative model of DNA copy number signals as shown in Figure 2. The signal modeling is pursued through three major levels: a mixture of tumor and normal cells, signal intensity, and observed copy number signal.

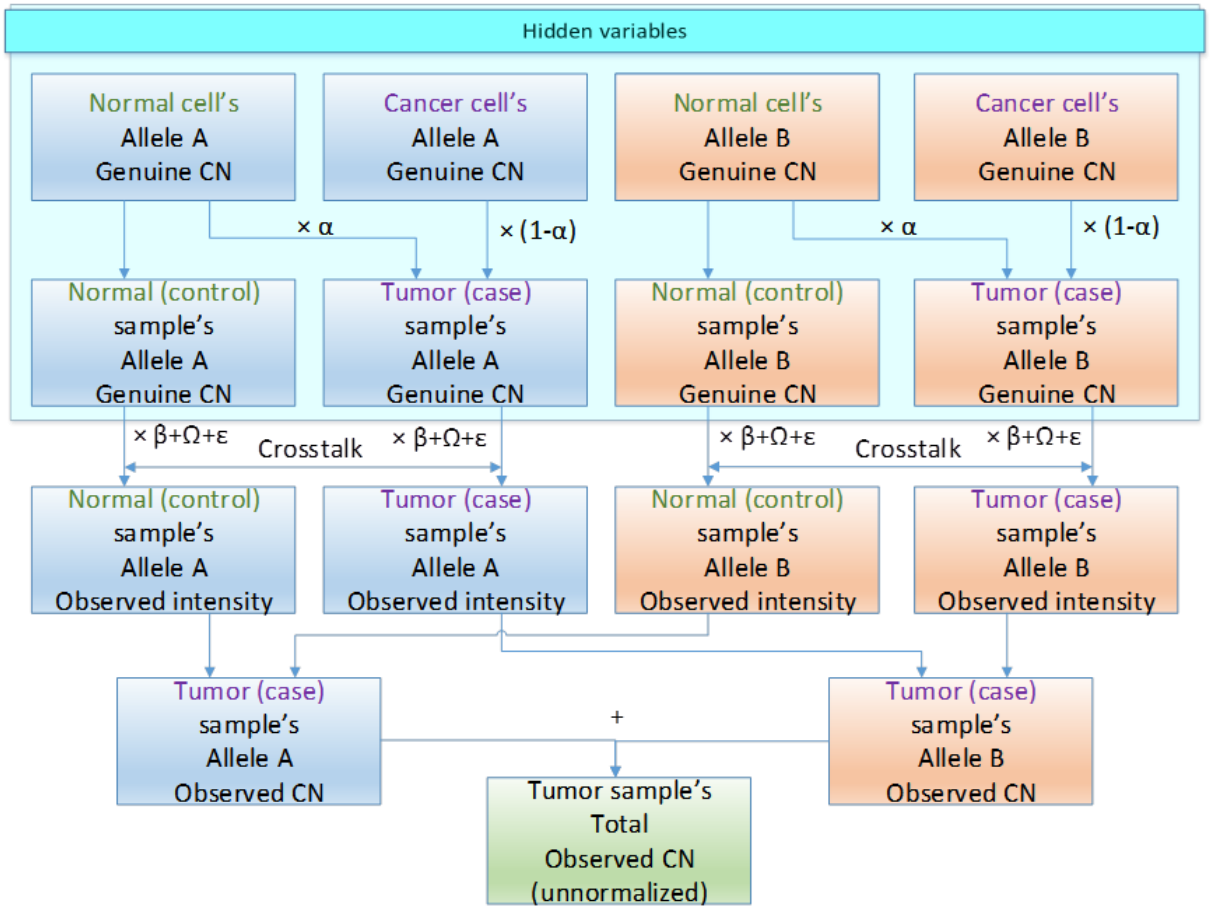


Figure 2- Signal model of allelic intensity and observed copy number signal

Through this signal model, we proved that the original BACOM method mainly suffered from three biases: 1. Inaccurate normalization of copy number signals; 2. Inaccurate quantification of CNV signals; 3. Misclassification of deletion type. All these biased could be corrected with an accurate quantification and normalization method, and this motivates us to propose an improved method BACOM 2.0

4.3. BACOM 2.0 methodology and workflow

Accurate signal normalization essentially rescales the relative signal intensities on the basis of normal copy regions (diploid reference loci), here termed as absolute normalization (Attiyeh, et al., 2009; Popova, et al., 2009). As the intertwined result of normal cell contamination, copy

number aberrations, and tumor aneuploidy, the average ploidy of tumor cells cannot be assumed to be 2N or integer (Rasmussen, et al., 2011). Though absolute normalization is critical to inferring absolute copy numbers in a tumor sample, the classic normalization procedure based on median-centering of the total probe intensities is problematic (Carter, et al., 2012; Wang, et al., 2002; Yu, et al., 2011), since the dominant component of the intensity mixture distribution rarely coincides with the normal copy number '2' (Rasmussen, et al., 2011).

Let us consider histogram modeling of genome-wide copy number signals. Based on the underlying signal characteristics, we adopt a mixture of K Gaussian distributions (Attiyeh, et al., 2009), given by

$$f(x) = \sum_{k=1}^{K} \pi_k g(x|\mu_k, \sigma_k^2)$$

where is the relative proportion of the k-th copy number component and g(.|.) is the Gaussian kernel with being the mean and variance. Such mixtures can be estimated from observed histograms using soft clustering or the maximum likelihood method. However, our experimental studies on real tumor data confirmed that the component means with the largest does not always correspond to the mean of normal copy regions, probably due to the aforementioned factors, and thus cannot serve as the baseline for absolute normalization. While we have also observed that the largest component(s) often resides within the neighborhood of the normal copy components.

Thus, we first develop an effective scheme to eliminate the loci be-longing to the hemi-deletions (with copy number '1') and the allelic-imbalanced regions (with copy number '3', '5', etc.). Specifically, we use a sliding window centered at a locus to estimate the inter-allele correlation coefficient and remove those loci whose correlation coefficients are lower than an automatically-determined threshold value, since the imbalanced allele signals associated with odd

copy numbers would produce a sufficiently negative value of ρ, given by (in the case of copy number '3')

$$\rho_{\text{allelic-imbalanced}} \cong \frac{4(1+\rho)\sigma^2}{(1-\alpha)^2 + 4\sigma^2} - 1$$

where σ2 is the variance of noise and ρ is the genuine inter-allele correlation coefficient. This procedure also eliminates copy-neutral LOH loci and thus can improve the accuracy of estimating ρ by using only normal copy loci. It can be shown that copy-neutral LOH contamination will result in an inaccurate estimate of ρ, given by

$$\rho_{\text{LOH-contaminated}} \cong \frac{(1+\rho)\sigma^2}{\eta(1-\alpha)^2 + \sigma^2} - 1$$

where η is the percentage of copy-neutral LOH contamination.

Subsequently, a revised Gaussian mixture model (7) is derived solely from the remaining allelic-balanced loci. Tested on many real copy number datasets, we found that the dominant component of the revised Gaussian mixture distribution now corresponds to the normal copy number regions in most cancer types. Thus, we propose to rescale the measured copy number signal intensities using the mode of the dominant component. Since such signal normalization is performed in each individual sample and based on the signals of normal copy number regions, BACOM 2.0 implements an accurate and absolute normalization (Attiyeh, et al., 2009).

Moreover, BACOM 2.0 includes an accurate estimation of allelic correlation coefficient ρ (related to model parameter λ in defining hemi-deletion summary statistic) that was often underestimated due to copy-neutral LOH contamination. Again, by excluding copy-neutral LOH loci and identifying dominant normal copy regions via the aforementioned scheme, we can now

obtain a more accurate estimate of allelic correlation coefficient ρ and subsequently differentiate be-tween hemi- and homo- deletions.

Also, we tried to calibrate allele signal crosstalk and saturation effects. Theoretically, signal crosstalk from the probes that differ only in one SNP adds positive bias to the copy number estimate that could lead to an overestimation of normal cell fraction by (6). As aforementioned, the allelic crosstalk effect also biases the estimate of the allele correlation coefficient. Concerning copy number signal saturation using SNP arrays, we adopted a similar linearization strategy used by ABSOLUTE (Carter, et al., 2012).

Lastly, we exploited a mathematically-justified scheme to correct for the confounding impact of intratumor heterogeneity on estimating tumor purity (Oesper, et al., 2013; Rasmussen, et al., 2011). Though normal fraction α can hypothetically be estimated using any deletion segments, it can be experimentally and theoretically shown that the value of α will highly likely be overestimated when intratumor heterogeneity occurs in the deletion segment being used. Thus, in the presence of suspected intratumor heterogeneity, only the 'pure' deletion segments with homogeneous tumor genotypes should be used to estimate the normal fraction. Based on the distribution of α estimates across the whole genome, BACOM 2.0 calculates the final value of the normal fraction using the 9-percentile of α estimates.

In relation to previous work, the concept of using allele-specific information for analyzing copy number data is shared by others (Van Loo, et al., 2010; Yuan, et al., 2012)for exploratory data visualization in conjunction with a visual inspection of aneuploidy and tumor heterogeneity. There is also some similarity between our objectives and others in cancer copy number restoration and tumor purity estimation. The major limitations of the approach by Yuan, et al. (2012)are that it requires matched genomic and histopathological image data and relies heavily on image quality

(coarse H&E staining, artifacts, batch effects). ABSOLUTE, which was developed by Carter, et al. (2012)is supported by an elegant yet complex mathematical framework and can select the most likely combination of estimated tumor purity and ploidy by integrating copy number data and supervised learning. It is acknowledged that the cornerstone system of equations is underdetermined and various heuristics cannot guarantee a unique and correct solution(Oesper, et al., 2013). For example, in the presence of intratumor heterogeneity, the restored copy number signals are not necessarily all integer values, thus using the highest likelihood of producing all integer signals to select the most likely solution may be problematic(Oesper, et al., 2013). Su, et al. (2012)estimates normal cell fraction using single-nucleotide variants but not original sequence reads. The formulation does not explicitly consider the effects of copy number gains/losses thus may bias tumor purity estimation. Moreover, PurityEst(Su, et al., 2012), TheTa(Oesper, et al., 2013) and AbsCN-seq(Bao, et al., 2014) rely on next-generation sequencing data, therefore they may not be applicable to existing copy number data acquired using more classic methods.

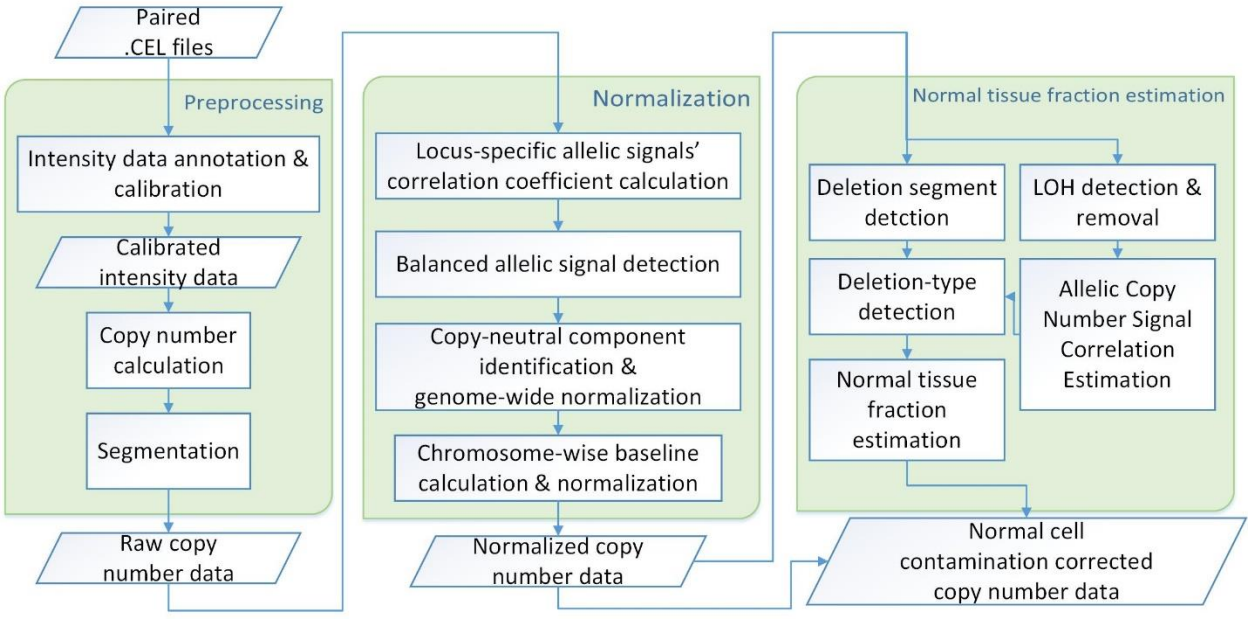


Figure 3- Analytic pipeline of BACOM 2.0: schematic flowchart

4.4. Simulation study and experimental results

We first considered numerical mixtures of simulated normal and cancer copy number profiles across a chromosome region, a situation in which all factors are known and the use of a linear mixture model is valid. We reconstituted mixed copy number signals by multiplying the simulated cancer copy number profile by the tumor purity percentage in a given heterogeneous sample. The realistic simulations were generated using a specifically selected pair of matched tumor-normal ovarian cancer samples in TCGA, where the tumor somatic copy number profile is approximately normal, i.e., allelic-balanced, summed copy number '2', and no LOH contamination. After variably dividing the whole region into eight segments, we assigned allelic-specific copy number status to each of the segments ranging from 0 to 3, as specified in Fig. 2. The raw copy number signals (the sum of the two alleles) were produced by mixing 1-α fraction of simulated tumor copy number profile with α fraction of normal copy number profile. This simulation represents a highly challenging scenario in which the majority of probe sets were not 'normal' but amplified, yet also contained both hemi-deletion and copy-neutral LOH segments.

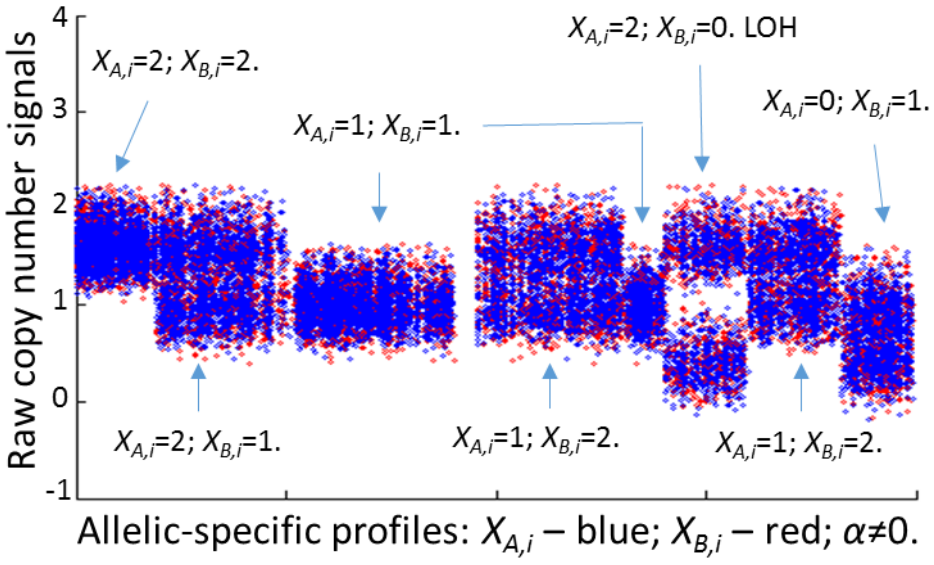


Figure 4- Realistic simulated allelic-specific copy number signals

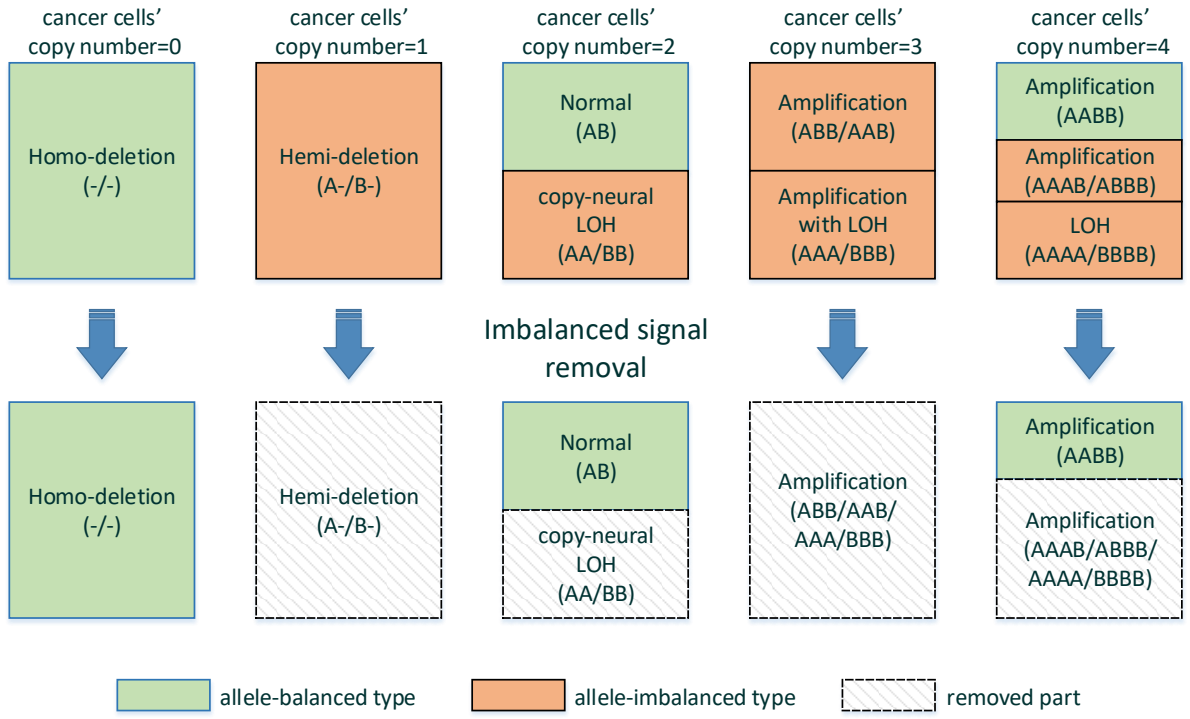
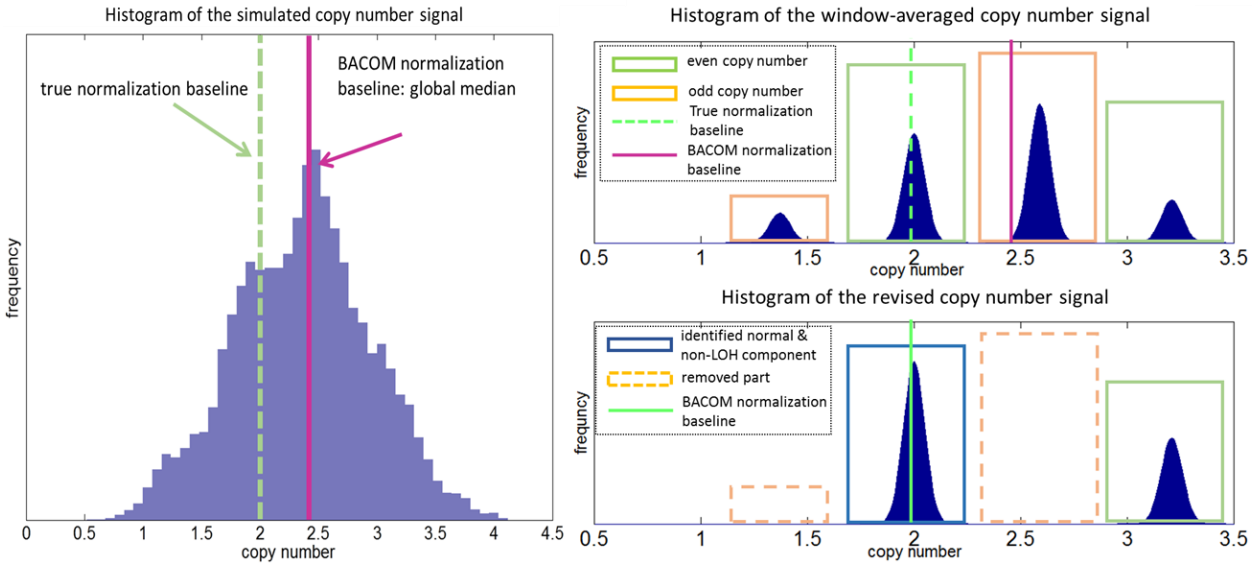


Figure 5- Brief illustration of the principles of removing allele-imbalanced loci to revise the signal histogram



(a) Histogram of simulated copy number signals; (b) Histogram of preprocessed copy number signals after moving-average; (c) 'revised' histogram of copy numbers after eliminating allelic-

Figure 6- Histogram of revised copy number signals

Using the BACOM 2.0 analytic pipeline, we first calculated the histogram of the raw copy number signals; then we preprocessed the raw copy number signals by a moving-average low-pass filter that significantly reduced the noise effect, and re-calculated the histogram; lastly we eliminated all allelic-imbalanced loci and generated a revised histogram whose dominant peak correctly coincided with the normal copy number '2' component (Figure 6).

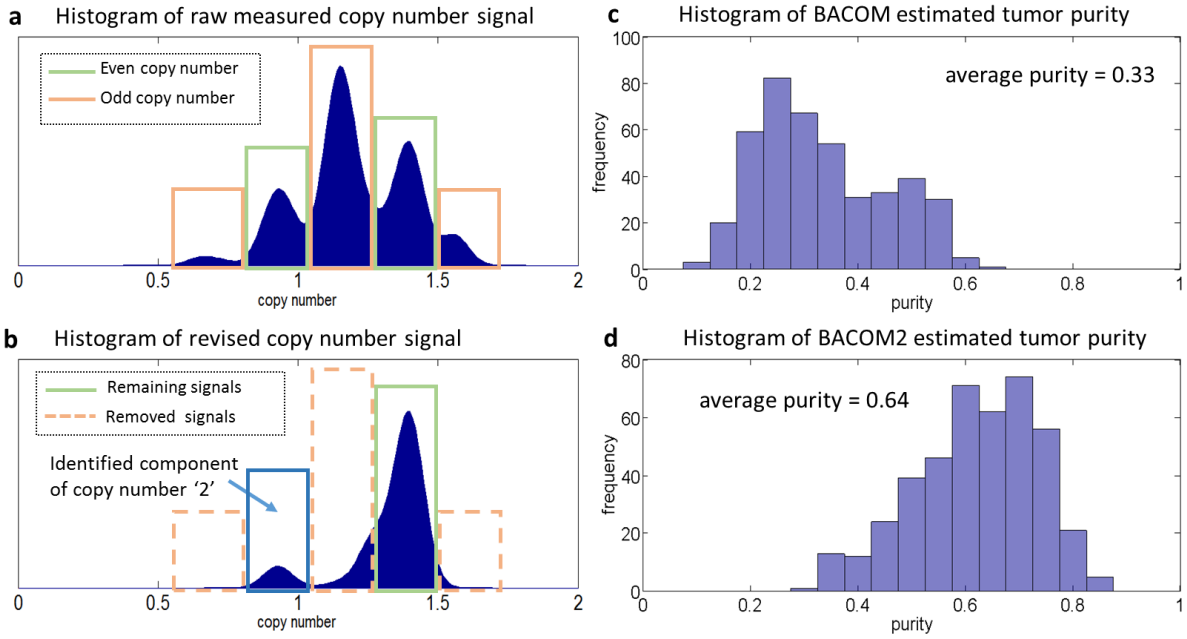
Table 1. Comparative parameter estimates by BACOM and BACOM 2.0

Parameter	Ground truth	BACOM	BACOM 2.0
ρ	-0.042	-0.714	-0.063
α	40%	79%	39%

With a successful absolute normalization, we first checked the estimated value of between-allele correlation coefficient ρ, and then recalculated the normal cell fraction α. Based on the comparative estimates given in Table 1, the power of BACOM 2.0 is evident since the model parameter estimates were very close to the ground truth as compared to what we obtained using the original BACOM.

4.5. Benchmark analysis

We applied BACOM 2.0 to the challenging case of the TCGA ovarian cancer dataset (466 samples), where a high genomic instability has been well-documented in high-grade ovarian cancers. We have observed that, in a large number of tumor samples, the dominant component of raw measured copy number histogram does not correspond to the normal copy number '2' but rather the allele-imbalanced loci (Figure 7). This observation suggests the widely existed partial aneuploidy in these samples, and highlights the improper use of global mean/median as the normalization baseline.

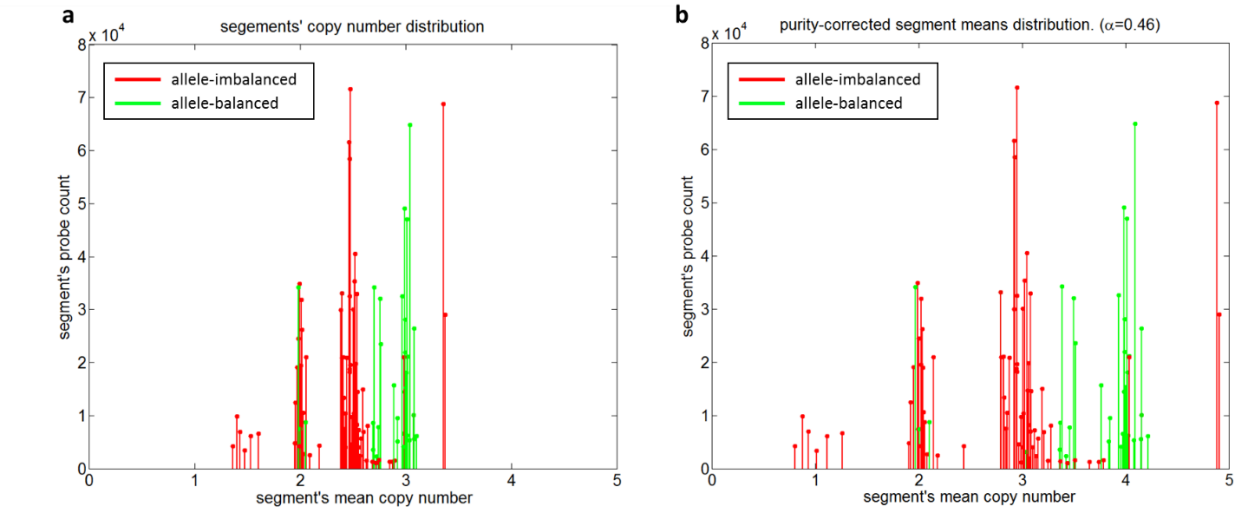


(a) Histogram of copy number signals after moving-average preprocessing; (b) Histogram of 'revised' copy number signals after eliminating allelic-imbalanced loci; (c) Histogram of tumor purity estimated by original BACOM; (d) histogram of tumor purity estimated by BACOM 2.0.

Figure 7- Analysis by BACOM 2.0 on TCGA ovarian cancer samples.

Using the BACOM 2.0 analytic pipeline, we preprocessed the raw measured copy number signals by a moving-average low-pass filter, eliminated all allelic-imbalanced loci, generated a revised histogram, and identified the component of normal copy number '2' (Figure 7b). With a successful absolute normalization, we estimated tumor purity and tumor-specific copy number

profile on each sample. From a comparison between the histogram of tumor purities likely underestimated by the original BACOM (Figure 7c) and the histogram of tumor purities newly estimated by BACOM 2.0 (Figure 7d), we can see that BACOM 2.0 has now produced much higher tumor purity estimates (average purity of 64% versus 33%) that are theoretically expected and consistent with the protocol baseline adopted in independent studies (using 50% purity as the threshold to differentiate between high and low tumor purity in three cancer types)(Downey, et al., 2014; Huijbers, et al., 2013; Su, et al., 2012).



(a) normalized signals by BACOM 2.0; and (b) copy number signals in tumor cells, where the height of the bins is the locus counts in the segment.

Figure 8- Distribution of mean of copy number on genomic segments

Using the same dataset, we further compared the estimates generated by BACOM 2.0 with those produced by ABSOLUTE. As a closely relevant method, ABSOLUTE reports the estimates of tumor purity and average ploidy on two TCGA datasets, ovarian cancer (OV) and brain cancer (GBM). With a quality control selection on paired tumor and normal samples, ABSOLUTE analyzed 392 tumor samples in the OV dataset. The average tumor purity estimates by BACOM 2.0 and ABSOLUTE are 64% and 78%, respectively; and the average tumor ploidy estimates by BACOM 2.0 and ABSOLUTE are 2.33 and 2.73, respectively. The sample-wise correlation

coefficients show that both tumor purity and tumor ploidy estimated by BACOM 2.0 correlate well with the estimates by ABSOLUTE (Figure 9), achieving a high correlation coefficient of r = 0.74 on purity and a high correlation coefficient of r = 0.71 on ploidy. On the GBM dataset, the average tumor purity estimates by BACOM 2.0 and ABSOLUTE are 59% and 71%, respectively; and the average tumor ploidy estimates by BACOM 2.0 and ABSOLUTE are 2.09 and 2.17, respectively.

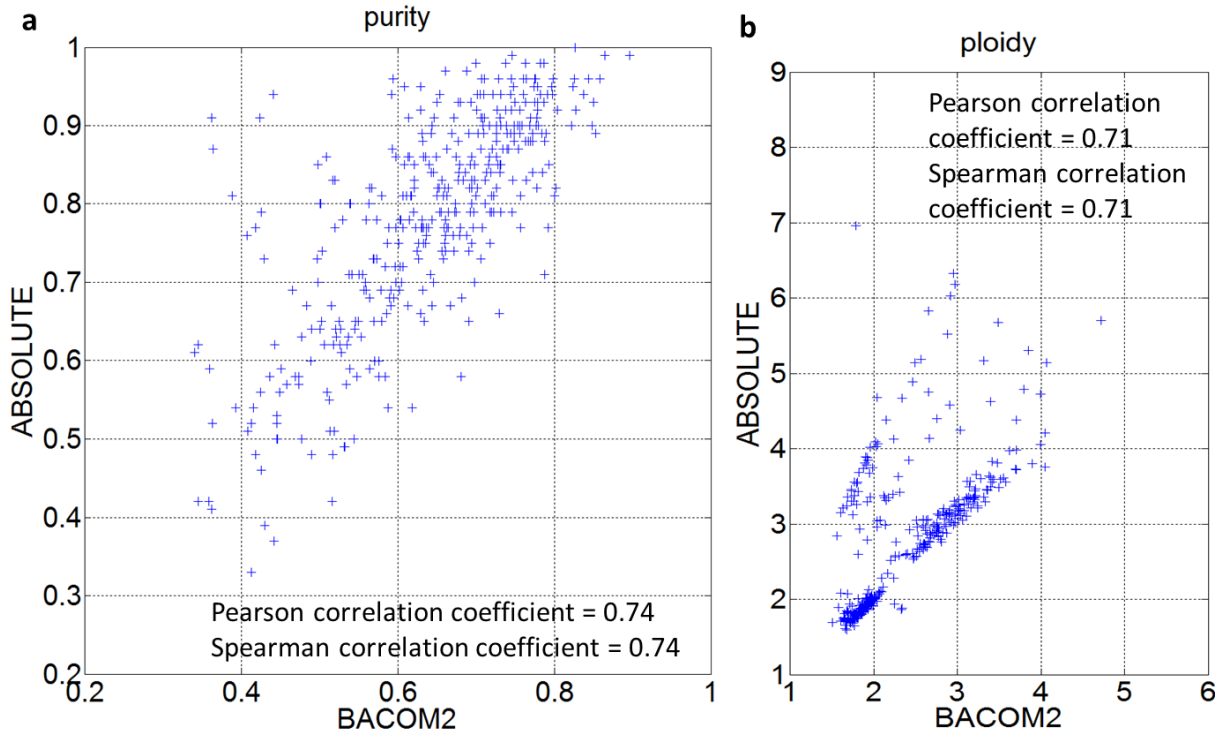


Figure 9- Sample-wise comparison between BACOM 2.0 and ABSOLUTE on TCGA-OV samples.

In the absence of definite ground truth about the tumor purities in real samples, the validation of a new method for quantifying absolute copy numbers is always problematic. A reasonable alternative is to perform some form of 'cross' affirmation by exploiting the 'orthogonal' information structures provided by the independent sources related to a common set of nature states. We lastly compared the tumor purity estimates by BACOM 2.0 with the estimates by an independent method (called UNDO)(Wang, et al., 2015) that de-convoluted the mixed gene expression profiles of tumor and stroma cells acquired from the same TCGA OV samples. Using

the UNDO software, we analyzed the tumor samples with purity estimate by BACOM 2.0. The experimental result shows that the tumor purity estimates by BACOM 2.0 (based on copy number data) correlate well with the estimates by UNDO (based on gene expression data), consistently achieving a strong average 'cross' correlation coefficient of 0.5~0.6 in multiple runs. We performed the same comparison on the TCGA GBM samples and obtained consistent results.

TCGA_OV (n=392)	BACOM2	ABSOLUTE	Sample-wise correlation r
purity	0.64	0.78	0.74
ploidy	2.33	2.73	0.71

Table 2. Comparison between BACOM 2.0 and ABSOLUTE on TCGA_OV dataset

TCGA_GBM (n= 79)	BACOM2	ABSOLUTE	Sample-wise correlation *r*
Purity	0.59	0.71	0.56
Ploidy	2.09	2.17	0.78

Table 3. Comparison between BACOM 2.0 and ABSOLUTE on TCGA_GBM dataset

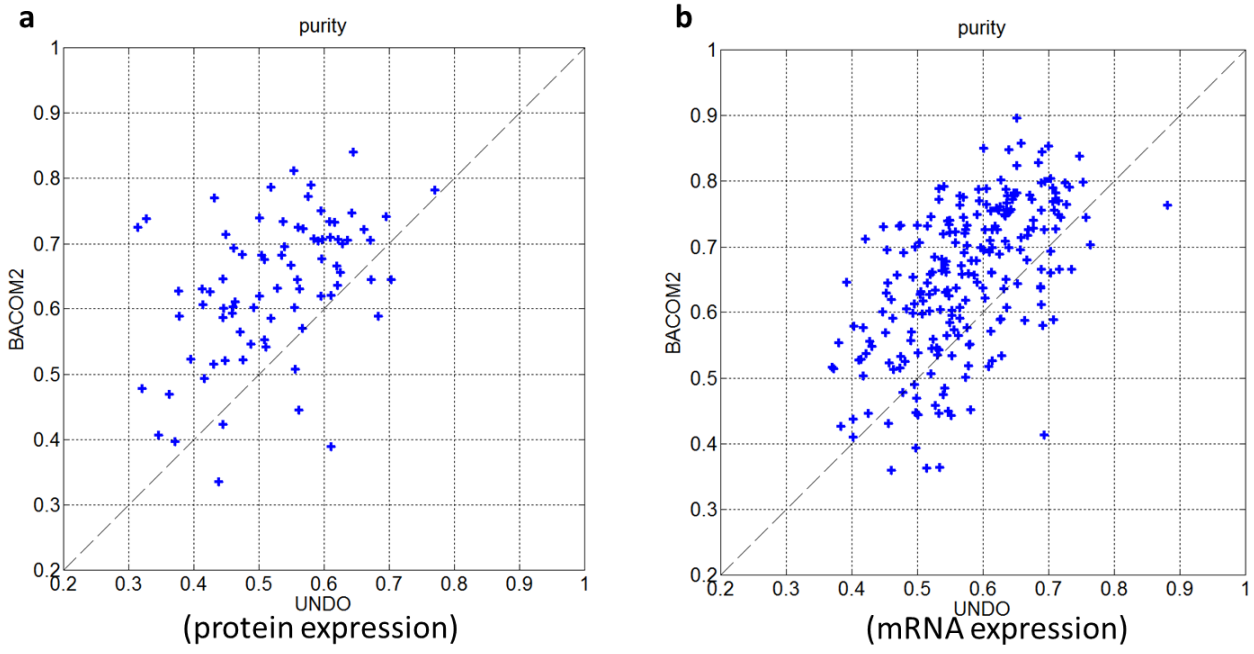


Figure 10- Comparison between tumor purity estimated by BACOM 2.0 and UNDO on TCGA_OV samples.

The imperfect ‘cross’ correlation between the tumor purity estimates made by UNDO, and BACOM 2.0 may be expected and well justified due to the following. First, the estimate by UNDO was based on gene expression values, while the estimate by BACOM 2.0 was based on copy number values. Second, though the two information sources are related to a common set of states of nature, they are also ‘orthogonal’ in various aspects. For example, copy number values are always ‘2’ across all normal cells (e.g., stroma, T-cells, monocytes), while gene expression values are cell type specific. In fact, there are multiple gene expression profiles corresponding to various normal cells. Third, copy number values are generally ‘static’, while gene expression values are intrinsically ‘dynamic’. Such differences can confound the correlation analysis. Fourth, the degree of technical variability, e.g., noise levels, can be significantly different in acquiring copy number versus gene expression signals. For example, in the recent supervised deconvolution work, called ESTIMATE, by Yoshihara, et al. (2013), in order to obtain a high correlation between the tumor purity estimates derived from copy number and gene expression data, a nonlinear regression function was used to map the ‘score’ by to the estimate by ABSOLUTE. Though a higher correlation was obtained and validated on multiple datasets after such nonlinear mapping, it is somewhat ‘indirect’.

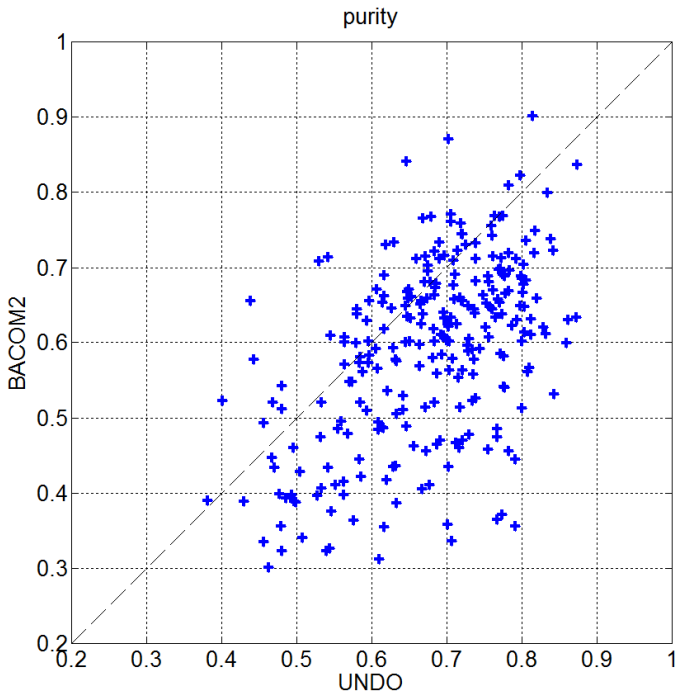


Figure 11- Comparison between tumor purity estimated by BACOM 2.0 and UNDO on TCGA_GBM samples.

Chapter 5. DDN analysis for detecting network rewirings on single-omics data

5.1. Differential network analysis and DDN methods

To explicitly address differential network analysis and detect the group-specific connections under different conditions, Zhang, et al. (2009) proposed the framework of the DDN method that compares the network topologies between two conditions. We denote this initial version of the DDN method as DDN1. The philosophy behind DDN1 is simple: construct sparse networks separately under each of the two different conditions, and then compare them to detect changes between two networks. LASSO regression is used to fit each node to construct a sparse network under each condition. The DDN1 objective is, for each node, finding the LASSO regression coefficients of the following optimization problem:

$$\begin{cases} \boldsymbol{\beta}_i^{(1)} = \arg\min_{\boldsymbol{\beta}_i^{(1)}} f\left(\boldsymbol{\beta}_i^{(1)}\right) = \frac{1}{2}\left\|\mathbf{y}^{(1)} - \mathbf{X}^{(1)}\boldsymbol{\beta}_i^{(1)}\right\|_2^2 + \lambda\left|\boldsymbol{\beta}_i^{(1)}\right| \\ \boldsymbol{\beta}_i^{(2)} = \arg\min_{\boldsymbol{\beta}_i^{(2)}} f\left(\boldsymbol{\beta}_i^{(2)}\right) = \frac{1}{2}\left\|\mathbf{y}^{(2)} - \mathbf{X}^{(2)}\boldsymbol{\beta}_i^{(2)}\right\|_2^2 + \lambda\left|\boldsymbol{\beta}_i^{(2)}\right| \end{cases}$$

where $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ are the observed data matrix under condition 1 and condition 2; $\mathbf{y}^{(1)}$ and $\mathbf{y}^{(2)}$ are the observed data vector of the *i*-th node under two conditions; $\boldsymbol{\beta}_i^{(1)}$ and $\boldsymbol{\beta}_i^{(2)}$ are the LASSO regression coefficients, also denoted as local neighborhood structure in which the non-zero elements represent the connection between the corresponding node and the *i*-th node. After the local neighborhood structures $\boldsymbol{\beta}_i$ are learned for all nodes, they are merged to form the adjacency matrix which represents the network structure.

In detecting network rewiring between two conditions, we expect that network rewiring events are sparse. In other words, we assume that the networks under two conditions will share a large portion of common network structures. Based on this assumption, Zhang and Wang (2010) further improved the initial DDN method by jointly solving LASSO regressions while introducing a penalty term on the structural difference. We denote this version of the DDN method as DDN2. The DDN2 optimization problem is:

$$\boldsymbol{\beta}_i = \left(\boldsymbol{\beta}_i^{(1)}, \boldsymbol{\beta}_i^{(2)}\right) = \arg\min_{\boldsymbol{\beta}_i} f\left(\boldsymbol{\beta}_i\right) = \frac{1}{2}\left\|\mathbf{x}_i^{(1)} - \mathbf{X}^{(1)}\boldsymbol{\beta}_i\right\|_2^2 + \lambda_1 \sum_{j=1}^{p}\left(\left|\beta_{ji}^{(1)}\right| + \left|\beta_{ji}^{(2)}\right|\right) + \lambda_2 \left\|\boldsymbol{\beta}_i^{(1)} - \boldsymbol{\beta}_i^{(2)}\right\|_1$$

where i is the node index; β_{ji} is the regression coefficient from node i to node j under a specific condition; $\boldsymbol{y}_i$ and $\boldsymbol{X}$ are the expression values of dependent and input variables, respectively; P is the number of nodes; λ_1 and λ_2 are the parameters on the two penalty terms that are used to assure both a sparse common network structure and sparse differential network rewiring.

Tian, et al. (2011) proposed to integrate prior-knowledge into DDN2 framework. This knowledge-fused DDN method is developed into an open-source Cytoscape app called kDDN (Tian, et al., 2015). It has a better graphical user interface for user-computer interaction, and is very convenient for data visualization. The kDDN optimization problem is:

$$\boldsymbol{\beta}_i = \left(\boldsymbol{\beta}_i^{(1)}, \boldsymbol{\beta}_i^{(2)}\right) = \arg\min_{\beta_i} f\left(\boldsymbol{\beta}_i\right) = \frac{1}{2}\left\|\mathrm{y}_i - \mathbf{X}\boldsymbol{\beta}_i\right\|_2^2 + \lambda_1 \sum_{j=1}^{P}\left(1 - W_{ji}\theta\right)\left(\left|\beta_{ji}^{(1)}\right| + \left|\beta_{ji}^{(2)}\right|\right) + \lambda_2 \left\|\boldsymbol{\beta}_{ji}^{(1)} - \boldsymbol{\beta}_{ji}^{(2)}\right\|_1$$

where W is the matrix of fused prior-knowledge of node-to-node connections, such as gene interactions in pathways; and θ is a weighting parameter.

If no prior knowledge is used in kDDN, it degrades to the DDN2 method. Since DDN2 is better in joint-regression, and we are not discussing knowledge fusing in this dissertation. In the remaining part of the dissertation the term "DDN" refers to the DDN2 method by default.

DDN infers molecular networks by estimating conditional dependencies among genes. Conditional dependency is a key type of probabilistic relationship that is distinct from the basic correlation relationship. If two genes are conditionally dependent, their expression levels are still correlated even after accounting for all other genes' expressions. Conditional dependence relationship hence is less likely to reflect transitive effects than basic correlation relationship, and provides stronger evidence of functional relationships between genes. These functional relationships could be regulatory or other molecular interactions that cause the two genes' expressions to be tightly coupled.

DDN uses local dependency models to characterize the dependencies among genes in the network and represent local network structures. Unlike pairwise correlation, the conditional dependence between two genes cannot be measured solely based on these two genes' expressions. Instead, the whole collection of possible links to all other genes should be considered, in order to find the network that best explains the expression data. DDN adapts an efficient neighborhood selection strategy based on a LASSO regression to enable such inference. Mathematically, we formulate DDN's network structure learning by solving the optimization of:

$$\boldsymbol{\beta}_i = \left(\boldsymbol{\beta}_i^{(1)}, \boldsymbol{\beta}_i^{(2)}\right) = \arg\min_{\boldsymbol{\beta}_i} f\left(\boldsymbol{\beta}_i\right) = \frac{1}{2}\left\|\mathbf{x}_i^{(1)} - \mathbf{X}^{(1)}\boldsymbol{\beta}_i\right\|_2^2 + \lambda_1 \sum_{j=1}^{p}\left(\left|\beta_{ji}^{(1)}\right| + \left|\beta_{ji}^{(2)}\right|\right) + \lambda_2 \left\|\boldsymbol{\beta}_i^{(1)} - \boldsymbol{\beta}_i^{(2)}\right\|_1$$

$$\text{s.t. } \beta_{ii}^{(1)} = \beta_{ii}^{(2)} = 0$$

where i is the node index; β_{ji} is the regression coefficient from node i to node j under a specific condition; $\boldsymbol{y}_i$ and $\boldsymbol{X}$ are the expression values of dependent and input variables, respectively; P is the number of nodes; λ_1 and λ_2 are the parameters on the two penalty terms that are used to assure both a sparse common network structure and sparse differential network rewiring.

The differences between $\beta_{ji}^{(1)}$ and $\beta_{ji}^{(2)}$ indicate the differential dependence edges, while common dependence edges in the network are inferred by the consistent coefficients. Note that the differential dependences are of particular interest, because such network rewiring may reveal pivotal information on how the biological system responds to different biological conditions.

The permutation test is introduced in DDN to evaluate empirical p-values of the detected network rewirings (Tian, et al., 2011). The detected differential edges with multi-test corrected p-values less than the preset significance threshold (e.g., p-value<0.05) are marked as significant network rewirings.

5.2. Improved DDN method for imbalanced data

5.2.1. Problem diagnosis

When applying the DDN method to various data sets, we noticed a potential systematic bias: in imbalanced data which is defined as data with unequal numbers of observations in each class, the DDN method consistently detects more differential edges in one condition of the smaller sample size than in the other condition of the larger samples size. As shown in Figure 12 in section 5.2.3, the simulation study we designed confirms our suspicion of this systematic bias in the DDN method. Therefore, we are motivated to correct this bias and improve the DDN methodology.

There are two directions to minimize or eliminate the bias brought by imbalanced data. The first one is to make the data balanced again, by over-sampling the minority group, or under-sampling the majority group, or generating additional synthetic samples. It is a simple and universal approach to all problems caused by imbalanced data, but it also suffers from reduced statistical power or addition bias from synthetic samples. The second direction is changing the performance metric or the algorithm so that the solution will be irrelevant to sample scales. It is

not always viable, but will efficiently correct the bias brought by imbalanced data. We choose the second direction and aim to redesign DDN's mathematical formulation.

The objective function in LASSO regressions is basically the sum of squared residuals and the penalty term on parameters. Take the basic LASSO form (Tibshirani, 1996) for example, we expand the objective function in the form of each observation as follows:

$$f(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda|\boldsymbol{\beta}| = \frac{1}{2}\sum_{i=1}^{N}(y_i - \mathbf{x}_i\boldsymbol{\beta})^2 + \lambda|\boldsymbol{\beta}| = \frac{1}{2}\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{P} x_{i,j}\beta_j\right)^2 + \lambda\sum_{j=1}^{P}|\beta_j|$$

in which, N denotes the number of samples (or called the sample scale), and P denotes the number of features (or called the feature scale); y_i denotes the response variable for *i*-th sample and $x_{i,j}$ denotes the *j*-th predictor variable for *i*-th sample; β_j denotes the LASSO regression coefficient for *j*-th feature; and λ is the parameter controlling the overall sparsity.

In the DDN method implemented as a Cytoscape plugin (Tian, et al., 2015), the predictor variable $x_{i,j}$ is by default standardized to zero-mean and unit-variance: $z_{ij} = \left(x_{ij} - \overline{x}_i\right)/\sigma_{x_i}$, in which $\overline{x}_i$ and σ_{x_i} denote the sample mean and the sample standard deviation of the predictor variable x_i, respectively. The standardization procedure is equivalent to calculate the z-scores for each variable. Without loss of generality, we assume all predictor variables $\mathbf{x}_i$ are standardized in the remaining part of this dissertation.

From the formula, it is clear that LASSO regression's objective function contains two parts: the first part $\frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$ is the squared residuals; the second part is the penalty part, and in the case of LASSO, it contains only one term which is L-1 norm of the coefficient. For the sake of

discussion, in the remaining sections we call the two parts as the error part and the penalty part, respectively.

With a further look into the objective function, we can find the dependency of the two parts with the sample size N and/or the feature size P. The penalty part is merely the sum of absolute values of all regression coefficients, hence is independent to the sample size N. On the other hand, suppose the predictor variables are standardized to z-scores, the error part is positively correlated with the measurement scale N. In a special case, when y is zero vector and beta is unit vector:

$$\frac{1}{2}\left\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right\|_2^2=\frac{1}{2}\sum_{i=1}^{P}\left\|\mathbf{X}_i\right\|_2^2=\frac{1}{2}\sum_{i=1}^{P}\sum_{j=1}^{N}x_{ij}^2=\frac{1}{2}PN,\lambda\left|\boldsymbol{\beta}\right|=\lambda P,\ \textit{when}\ \mathbf{y=0},\ \boldsymbol{\beta}=\mathbf{1}$$

In the design of the DDN framework, the objective function does not explicitly assign a weight to each group, which equivalently gives equal weights to both case and control groups. The underlying assumption of assign equal weights is that, the sample sizes of case and control groups are similar and hence equally contribute to the objective function. This assumption holds true for the datasets that were designed as pair-matched, for example, both diseased tissue samples and normal tissue samples (usually blood sample) are collected from the same patient, and there are total 2N samples consist of N case samples plus N control samples. However, many research projects are designed without paired samples available(Mertins, et al., 2016; Zhang, et al., 2016), and the disparate biological groups are instead defined on disease subtypes, genetic mutation status, or other sub-groups. In these studies, the group sizes (i.e., the number of samples in the group) in the case group and control group are usually no longer equal; sometimes, one group size could be several times larger than the other. We call groups with unequal group sizes as imbalanced groups, or groups with imbalanced group sizes.

If we expand the objective function of DDN, we have:

$$
\begin{aligned}
f(\boldsymbol{\beta}_i) &= \frac{1}{2}\left\|\mathbf{y_i} - \mathbf{X}\boldsymbol{\beta}_i\right\|_2^2 + \lambda_1\left(\left|\boldsymbol{\beta}_i^{(1)}\right| + \left|\boldsymbol{\beta}_i^{(2)}\right|\right) + \lambda_2\left\|\boldsymbol{\beta}_i^{(1)} - \boldsymbol{\beta}_i^{(2)}\right\|_1 \\
&= \frac{1}{2}\left[\sum_{l=1}^{n_1}\left(x_{i,l} - \sum_{k=1}^{P} x_{k,l}\beta_{ik}^{(1)}\right)^2 + \sum_{l=n_1+1}^{n_1+n_2}\left(x_{i,l} - \sum_{k=1}^{P} x_{x,l}\beta_{ik}^{(2)}\right)^2\right] + \left[\lambda_1\left|\boldsymbol{\beta}_i^{(1)}\right| + \lambda_1\left|\boldsymbol{\beta}_i^{(2)}\right|\right] + \lambda_2\left\|\boldsymbol{\beta}_i^{(1)} - \boldsymbol{\beta}_i^{(2)}\right\|_1 \\
&= \left[\frac{1}{2}\sum_{l=1}^{n_1}\left(x_{i,l} - \sum_{k=1}^{P} x_{k,l}\beta_{ik}^{(1)}\right)^2 + \lambda_1\left|\boldsymbol{\beta}_i^{(1)}\right|\right] + \left[\frac{1}{2}\sum_{l=n_1+1}^{n_1+n_2}\left(x_{i,l} - \sum_{k=1}^{P} x_{x,l}\beta_{ik}^{(2)}\right)^2 + \lambda_1\left|\boldsymbol{\beta}_i^{(2)}\right|\right] + \lambda_2\left\|\boldsymbol{\beta}_i^{(1)} - \boldsymbol{\beta}_i^{(2)}\right\|_1 \\
&= \sum_{l=1}^{n_1}\left[\frac{1}{2}\left(x_{i,l} - \sum_{k=1}^{P} x_{k,l}\beta_{ik}^{(1)}\right)^2 + \frac{\lambda_1}{n_1}\left|\boldsymbol{\beta}_i^{(1)}\right|\right] + \sum_{l=n_1+1}^{n_1+n_2}\left[\frac{1}{2}\left(x_{i,l} - \sum_{k=1}^{P} x_{x,l}\beta_{ik}^{(2)}\right)^2 + \frac{\lambda_1}{n_2}\left|\boldsymbol{\beta}_i^{(2)}\right|\right] + \lambda_2\left\|\boldsymbol{\beta}_i^{(1)} - \boldsymbol{\beta}_i^{(2)}\right\|_1
\end{aligned}
$$

We can see that, although the weights for each group (case vs. control) are equal, and the sample-wise squared residual is at a comparable level, the actual weighted penalty of beta added to each sample differs. Suppose one group's sample scale is ten times as the other group, the actual sample-wise weight penalty term is as small as one-tenth of the other group. When minimizing the total object function with the L-1 penalty, the samples in the group of smaller sizes would be more likely to have smaller sparsity in network structure.

5.2.2. Reformulated DDN objective function

To redesign the weights applied to each group, we first define two LASSO objective functions(Friedman, et al., 2017) for each group, and rewrite the DDN objective function as follows:

$$
\begin{cases}
f_1\left(\boldsymbol{\beta}^{(1)}\right) = \frac{1}{2}\left\|\mathbf{y}^{(1)} - \mathbf{X}^{(1)}\boldsymbol{\beta}^{(1)}\right\|_2^2 + \lambda_1\left|\boldsymbol{\beta}^{(1)}\right| \\
f_2\left(\boldsymbol{\beta}^{(2)}\right) = \frac{1}{2}\left\|\mathbf{y}^{(2)} - \mathbf{X}^{(2)}\boldsymbol{\beta}^{(2)}\right\|_2^2 + \lambda_1\left|\boldsymbol{\beta}^{(2)}\right|
\end{cases}
$$

$$
\min f(\boldsymbol{\beta}_i) = f\left(\boldsymbol{\beta}_i^{(1)}, \boldsymbol{\beta}_i^{(2)}\right) = f_1\left(\boldsymbol{\beta}_i^{(1)}\right) + f_2\left(\boldsymbol{\beta}_i^{(2)}\right) + \lambda_2\left\|\boldsymbol{\beta}_i^{(1)} - \boldsymbol{\beta}_i^{(2)}\right\|_1
$$

in which f_1 only contains measurement and regression coefficients of condition 1, and so does f_2 to condition 2.

For standardized data, the LASSO objective functions f_1 and f_2 is positively dependent on the sample size N. To make the two group's objective function values at a comparable level, we simply add sample scale normalizer to the basic form of LASSO objective function. Define the LASSO objective function with the sample scale normalizer as:

$$\tilde{f}\left(\boldsymbol{\beta},\tilde{\lambda}\right)=\frac{1}{2N}\left\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right\|_2^2+\tilde{\lambda}\left|\boldsymbol{\beta}\right|$$

where $\tilde{\lambda}$ is the parameter for controlling the sparsity. Basically, we use mean squared error to replace the square error term in the original LASSO objective function. This normalized LASSO objective function is the scaled version of basic LASSO form, with $\lambda = N\tilde{\lambda}$:

$$\tilde{f}\left(\beta,\tilde{\lambda}\right)=\frac{1}{N}\left(\frac{1}{2}\left\|\mathrm{y}-\mathrm{X}\beta\right\|_2^2+N\tilde{\lambda}\left|\beta\right|\right)=\frac{1}{N}f\left(\beta,N\tilde{\lambda}\right)$$

Therefore, the solutions to these two forms of LASSO objective functions are identical, with the condition of $\lambda = N\tilde{\lambda}$. When the value $\tilde{\lambda}$ is given, the normalized LASSO objective function is independent of the sample scale N.

We now adjust DDN's formulation accordingly to the normalized LASSO objective function, and hence make the new DDN objective function also independent of the sample scales. The new DDN objective function is:

$$\begin{aligned}\tilde{f}\left(\boldsymbol{\beta}\right)&=\tilde{f}_1\left(\boldsymbol{\beta}_i^{(1)}\right)+\tilde{f}_2\left(\boldsymbol{\beta}_i^{(2)}\right)+\lambda_2\left\|\boldsymbol{\beta}_i^{(1)}-\boldsymbol{\beta}_i^{(2)}\right\|_1\\&=\frac{1}{2}\left[\frac{1}{n_1}\sum_{l=1}^{n_1}\left(x_{i,l}-\sum_{k=1}^{P}x_{k,l}\beta_{ik}^{(1)}\right)^2+\frac{1}{n_2}\sum_{l=n_1+1}^{n_1+n_2}\left(x_{i,l}-\sum_{k=1}^{P}x_{x,l}\beta_{ik}^{(2)}\right)^2\right]+\lambda_1\sum_{j=1}^{P}\left(\left|\beta_{ji}^{(1)}\right|+\left|\beta_{ji}^{(2)}\right|\right)+\lambda_2\left\|\boldsymbol{\beta}_i^{(1)}-\boldsymbol{\beta}_i^{(2)}\right\|_1\end{aligned}$$

5.2.3. Simulation Studies

We designed a simulation experiment to confirm the bias brought by imbalanced data in the original DDN method, and also to show our proposed reformulated DDN objective function is capable of handling imbalanced data correctly.

To illustrate the systematic bias, we need to identify which differential edges are the false positive detections caused by the bias. We purposely designed the data that contains no groud truth positives of differential edges, and hence all detected differential edges will be false positives. The simulated data is actually single condition data that follow i.i.d multi-variate Gaussian distribution, and is manually divided into two groups to form the pseudo two conditions. We generate data with the total sample scale N=500 and the feature scale P=30, and then divided into two groups with three settings of sample ratios: 1:1, 1:10, and 10:1. The first sample ratio of 1:1 represents the perfect balance of sample scale, which corresponds to balanced data; the rest two sample ratios correspond to imbalanced data wither the larger sample scale in either the first or the second group.

The generated data are standardized to zero-mean and unit variance, and then test by the original DDN method and the reformulated DDN method. The original DDN method is performed by kDDN plugin in Cytoscape without knowledge input, which is mathematically equivalent to the DDN2 method. The reformulated DDN method is implemented by the R language. Its results are visualized by Cytoscape to compare with results from the original DDN method.

The simulation results are shown in Figure 12. The condition-specific differential edges are colored as red or green for condition 1 or 2. The edge width is mapped from p-values evaluated by a permutation test, and wider edges have smaller p-values. For data group with a sample size ratio of 1:1, which is balanced data, the original DDN method and the sample-scale-normalized DDN method gave identical results: only two network rewirings events with insignificant p-values

are detected. These results fit the expectation since there should be no network rewiring in the simulated data. For imbalanced data groups with sample size ratios of 1:10 and 10:1, original DDN gives network rewiring detection results of overwhelmingly one-sided condition-specific network rewirings. Some of the detected differential edges have significant p-values evaluated from the permutation test. Since there is no network rewiring in the designed ground truth network, all these condition-specific network rewirings detected by the original DDN method are false positives, and they confirm the existence of systematic bias brought by imbalanced data. On the other hand, the sample-scale-normalized DDN method detected only one or zero network rewiring events in these imbalanced data. The detected differential edges have insignificant p-values and could be further filtered. The results show the low false positive rate of the sample-scale-normalized DDN method in the simulation data. The systematic bias in the original DDN method, i.e., the condition-specific false positives in the imbalanced data, has been successfully corrected.

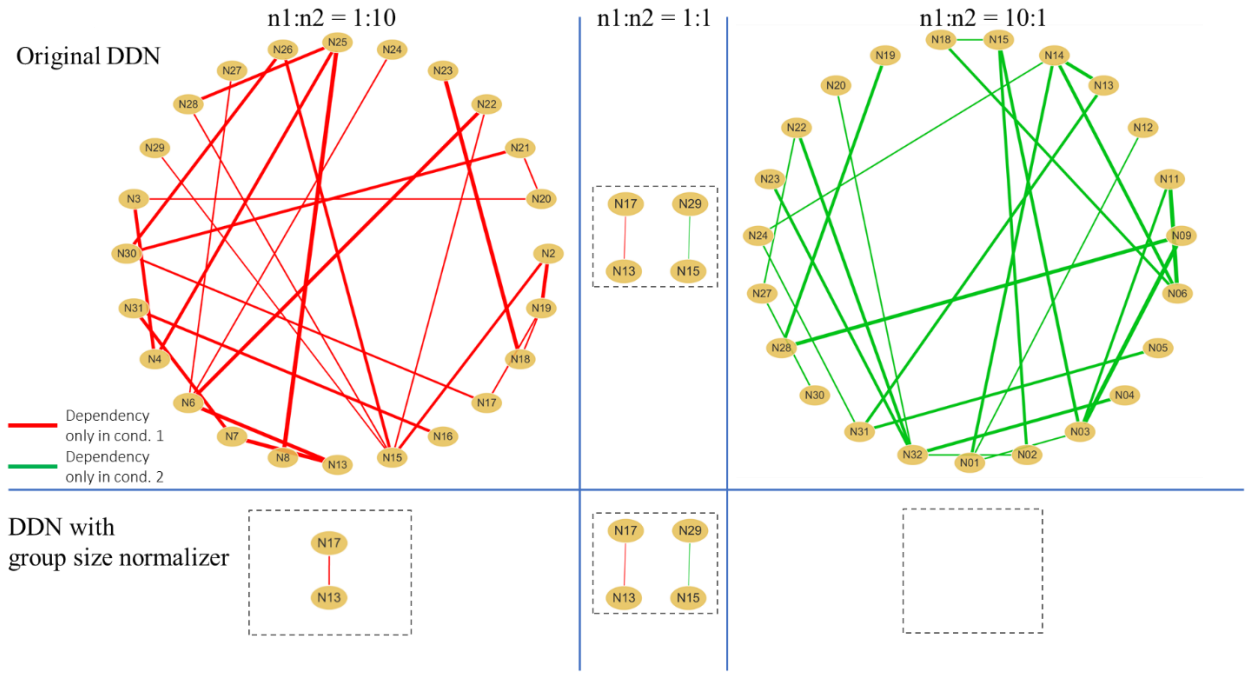


Figure 12- Comparison of DDN detected network rewirings from data with different group size ratios.

5.3. Solving DDN with accelerated algorithms

In this section, we will first introduce the block coordinate descent (BCD) algorithm used in solving the DDN optimization problem. We then proposed four accelerating strategies for : two proposed methods to accelerate the BCD algorithm by reformulating the computation procedure, the third strategy of integrating the Strong Rules into solving DDN optimization problem to reduce the feature scale, and the fourth one of implementing paralleling computing.

5.3.1. Solving DDN optimization with the BCD algorithm

The differential network inference is achieved by solving the optimization problem of minimizing DDN's objective function. In this section, we discuss how to achieve the minimum with the block coordinate descent (BCD) algorithm. The block coordinate descent approach was proposed by Banerjee, et al. (2008) for fast solving LASSO problem. Friedman, et al. (2008)showed with experiments that a coordinate descent procedure for the graphical lasso problem, is 30-4000 times faster than competing methods, making it a computationally attractive method. So far, block-wise descent algorithm is still among the fastest methods in state-of-the-art(Simon, et al., 2013). We use it as a starting point to solve DDN's optimization problem.

Firstly, DDN's objective function is convex. It is easy to prove that all terms in DDN's objective function are convex, and therefore the objective function itself as the sum of these functions is also convex. If a local minimum is achieved, by the convex property it will also be the global minimum value of the objective function.

Secondly, DDN's objective function has block-wise separability. It could be expressed as the sum of two parts: the first part is convex and differentiable; and the second part is convex and

non-differentiable, but is the summation of several non-overlapping members. We rewrite the DDN's objective function into two parts as follows:

$$f(\boldsymbol{\beta}_i)=\left(\frac{1}{2n_1}\left\|\mathbf{y}_i^{(1)}-\mathbf{X}^{(1)}\boldsymbol{\beta}_i^{(1)}\right\|_2^2+\frac{1}{2n_2}\left\|\mathbf{y}_i^{(2)}-\mathbf{X}^{(2)}\boldsymbol{\beta}_i^{(2)}\right\|_2^2\right)+\sum_{j=1}^{p}\left(\lambda_1\left|\beta_{ji}^{(1)}\right|+\lambda_1\left|\beta_{ji}^{(2)}\right|+\lambda_2\left|\beta_{ji}^{(1)}-\beta_{ji}^{(2)}\right|\right)$$

The second part of the objective function can be written as the sum of p terms with non-overlapping members $\left(\beta_j^{(1)},\beta_j^{(2)}\right), j=1,\cdots,p$. Each $\left(\beta_j^{(1)},\beta_j^{(2)}\right), j=1,\cdots,p$ is a coordinate block.

Tseng (2001)proved that, if one function is block-wise separable convex function, its global minimum could be reached by repeatedly block-wise optimization with the cyclic rule until convergence, and the convergence is guaranteed.

The essence of the BCD algorithm is "one-block-at-a-time". At iteration r + 1, only one coordinate block, $\left(\beta_k^{(1)},\beta_k^{(2)}\right)$ is updated, with the remaining $\left(\beta_j^{(1)},\beta_j^{(2)}\right), j\in\{1,\cdots k-1,k+1,\cdots,p\}$ fixed at their values. The cyclic rule is used to update parameter estimation iteratively, i.e., update parameter pair $\left(\beta_j^{(1)},\beta_j^{(2)}\right)$ for each $j=1,\cdots,p$, one by one in a circular way in the iterations.

In the remaining of this section, we will derive the solution of $\left(\beta_k^{(1)},\beta_k^{(2)}\right)$ that updated at the end of each iteration. Take DDN objective function's partial derivative to $\beta_k^{(1),r}$ and $\beta_k^{(2),r}$:

$$\begin{cases}\dfrac{\partial f}{\partial\beta_k^{(1),r}}=\dfrac{1}{n_1}\left(-\mathbf{y}_i^{(1)}\bullet\mathbf{x}_k^{(1)}+\sum_{l\neq i,k}\mathbf{x}_l^{(1)}\bullet\mathbf{x}_k^{(1)}+\mathbf{x}_k^{(1)}\bullet\mathbf{x}_k^{(1)}\beta_k^{(1),r}\right)+\lambda_1\operatorname{sgn}\left(\beta_k^{(1),r}\right)+\lambda_2\operatorname{sgn}\left(\beta_k^{(1),r}-\beta_k^{(2),r}\right)=0\\ \dfrac{\partial f}{\partial\beta_k^{(2),r}}=\dfrac{1}{n_2}\left(-\mathbf{y}_i^{(2)}\bullet\mathbf{x}_k^{(2)}+\sum_{l\neq i,k}\mathbf{x}_l^{(2)}\bullet\mathbf{x}_k^{(2)}+\mathbf{x}_k^{(2)}\bullet\mathbf{x}_k^{(2)}\beta_k^{(2),r}\right)+\lambda_1\operatorname{sgn}\left(\beta_k^{(2),r}\right)-\lambda_2\operatorname{sgn}\left(\beta_k^{(1),r}-\beta_k^{(2),r}\right)=0\end{cases}$$

Define the residual as:

$$\mathbf{y}_{i,-k}^{(1),r}=\mathbf{y}_i^{(1)}-\sum_{l\neq i,k}\mathbf{x}_l^{(1)}\beta_l^{(1),r}\quad \mathbf{y}_{i,-k}^{(2),r}=\mathbf{y}_i^{(2)}-\sum_{l\neq i,k}\mathbf{x}_l^{(2)}\beta_l^{(2),r},$$

And define the inner products between the residuals and current node's observation:

$$\rho^{(1),r}=\frac{1}{n_1}\mathbf{y}_{i,-k}^{(1),r}\bullet\mathbf{x}_k^{(1)},\ \rho^{(2),r}=\frac{1}{n_2}\mathbf{y}_{i,-k}^{(2),r}\bullet\mathbf{x}_k^{(2)}$$

Recalling that the standardized data have unit-variance, the partial derivative equations could be further written as:

$$\begin{cases}\dfrac{\partial f}{\partial\beta_k^{(1),r}}=\beta_k^{(1),r}-\rho^{(1),r}+\lambda_1\operatorname{sgn}\left(\beta_k^{(1),r}\right)+\lambda_2\operatorname{sgn}\left(\beta_k^{(1),r}-\beta_k^{(2),r}\right)=0\\ \dfrac{\partial f}{\partial\beta_k^{(2),r}}=\beta_k^{(2),r}-\rho^{(2),r}+\lambda_1\operatorname{sgn}\left(\beta_k^{(2),r}\right)-\lambda_2\operatorname{sgn}\left(\beta_k^{(1),r}-\beta_k^{(2),r}\right)=0\end{cases}$$

Giving the sign of $\beta_k^{(1),r}$, $\beta_k^{(2),r}$ and $\beta_k^{(1),r}-\beta_k^{(2),r}$, we can get closed-form solutions of $\beta_k^{(1),r}$ and $\beta_k^{(2),r}$. For example, when the conditions are $\beta_k^{(1),r}>0$ $\beta_k^{(1),r}>0$ and $\beta_k^{(1),r}-\beta_k^{(2),r}>0$, the solution is

$$\begin{cases}\beta_k^{(1),r}=\rho^{(1),r}-\lambda_1-\lambda_2\\ \beta_k^{(2),r}=\rho^{(2),r}-\lambda_1+\lambda_2\end{cases}$$

And the condition could then be converted to a sub-region in the plane of $\left(\rho^{(1),r},\rho^{(2),r}\right)$, which is:

$$\begin{cases}\rho^{(1),r}>\rho^{(2),r}+2\lambda_2\\ \rho^{(2),r}>\lambda_1-\lambda_2\end{cases}$$

Similarly, we can get all other closed-form solutions for possible conditions, and convert the conditions accordingly to sub-regions on the plane of $\left(\rho^{(1),r},\rho^{(2),r}\right)$. We list all solutions and corresponding subregions as follows

$$\boldsymbol{\beta}_k^r \triangleq \left(\beta_k^{(1),r},\beta_k^{(2),r}\right)$$

$$\begin{cases}
\boldsymbol{\beta}_k^r=\left(\rho^{(1),r}-\lambda_1-\lambda_2,\rho^{(2),r}-\lambda_1+\lambda_2\right), & \textbf{for } \rho^{(1),r}\geq\rho^{(2),r}+2\lambda_2,\rho^{(2),r}\geq\lambda_1-\lambda_2\\
\boldsymbol{\beta}_k^r=\left(\rho^{(1),r}+\lambda_1+\lambda_2,\rho^{(2),r}+\lambda_1-\lambda_2\right), & \textbf{for } \rho^{(1),r}\leq\rho^{(2),r}-2\lambda_2,\rho^{(2),r}\leq-\left(\lambda_1-\lambda_2\right)\\
\boldsymbol{\beta}_k^r=\left(\rho^{(1),r}-\lambda_1+\lambda_2,\rho^{(2),r}-\lambda_1-\lambda_2\right), & \textbf{for } \rho^{(1),r}\geq\lambda_1-\lambda_2,\rho^{(2),r}\geq\rho^{(1),r}+2\lambda_2\\
\boldsymbol{\beta}_k^r=\left(\rho^{(1),r}+\lambda_1-\lambda_2,\rho^{(2),r}+\lambda_1+\lambda_2\right), & \textbf{for } \rho^{(1),r}\leq-\left(\lambda_1-\lambda_2\right),\rho^{(2),r}\leq\rho^{(1),r}-2\lambda_2\\
\boldsymbol{\beta}_k^r=\left(\frac{1}{2}\left(\rho^{(1),r}+\rho^{(2),r}\right)-\lambda_1,\frac{1}{2}\left(\rho^{(1),r}+\rho^{(2),r}\right)-\lambda_1\right), & \textbf{for } \rho^{(1),r}<\rho^{(2),r}+2\lambda_2,\ \rho^{(2),r}<\rho^{(1),r}+2\lambda_2,\rho^{(2),r}>-\rho^{(1),r}+2\lambda_1\\
\boldsymbol{\beta}_k^r=\left(\frac{1}{2}\left(\rho^{(1),r}+\rho^{(2),r}\right)+\lambda_1,\frac{1}{2}\left(\rho^{(1),r}+\rho^{(2),r}\right)+\lambda_1\right), & \textbf{for } \rho^{(1),r}>\rho^{(2),r}-2\lambda_2,\ \rho^{(2),r}>\rho^{(1),r}-2\lambda_2,\rho^{(2),r}<-\rho^{(1),r}-2\lambda_1\\
\boldsymbol{\beta}_k^r=\left(0,\rho^{(2),r}-\lambda_1-\lambda_2\right), & \textbf{for } \rho^{(1),r}<\lambda_1-\lambda_2,\rho^{(1),r}>-\lambda_1-\lambda_2,\rho^{(2),r}>\lambda_1+\lambda_2\\
\boldsymbol{\beta}_k^r=\left(0,\rho^{(2),r}+\lambda_1+\lambda_2\right), & \textbf{for } \rho^{(1),r}>-\lambda_1+\lambda_2,\rho^{(1),r}<\lambda_1+\lambda_2,\rho^{(2),r}<-\lambda_1-\lambda_2\\
\boldsymbol{\beta}_k^r=\left(\rho^{(1),r}-\lambda_1-\lambda_2,0\right), & \textbf{for } \rho^{(1),r}>\lambda_1+\lambda_2,\rho^{(2),r}>-\lambda_1-\lambda_2,\rho^{(2),r}<\lambda_1-\lambda_2\\
\boldsymbol{\beta}_k^r=\left(\rho^{(1),r}+\lambda_1+\lambda_2,0\right), & \textbf{for } \rho^{(1),r}<-\lambda_1-\lambda_2,\rho^{(2),r}<\lambda_1+\lambda_2,\rho^{(2),r}>-\lambda_1+\lambda_2\\
\boldsymbol{\beta}_k^r=\left(\rho^{(1),r}-\lambda_1-\lambda,\rho^{(2),r}+\lambda_1+\lambda_2\right), & \textbf{for } \rho^{(1),r}\geq\lambda_1+\lambda_2,\rho^{(2),r}\leq-\lambda_1-\lambda_2\\
\boldsymbol{\beta}_k^r=\left(\rho^{(1),r}+\lambda_1+\lambda,\rho^{(2),r}-\lambda_1-\lambda_2\right), & \textbf{for } \rho^{(1),r}\leq-\lambda_1-\lambda_2,\rho^{(2),r}\geq\lambda_1+\lambda_2\\
\boldsymbol{\beta}_k^r=(0,0), & \textbf{for others}
\end{cases}$$

In Figure 13, we illustrate the solution subregions of $\beta_k^{(1),r}$ and $\beta_k^{(2),r}$ on the plane of $\left(\rho^{(1),r},\rho^{(2),r}\right)$.

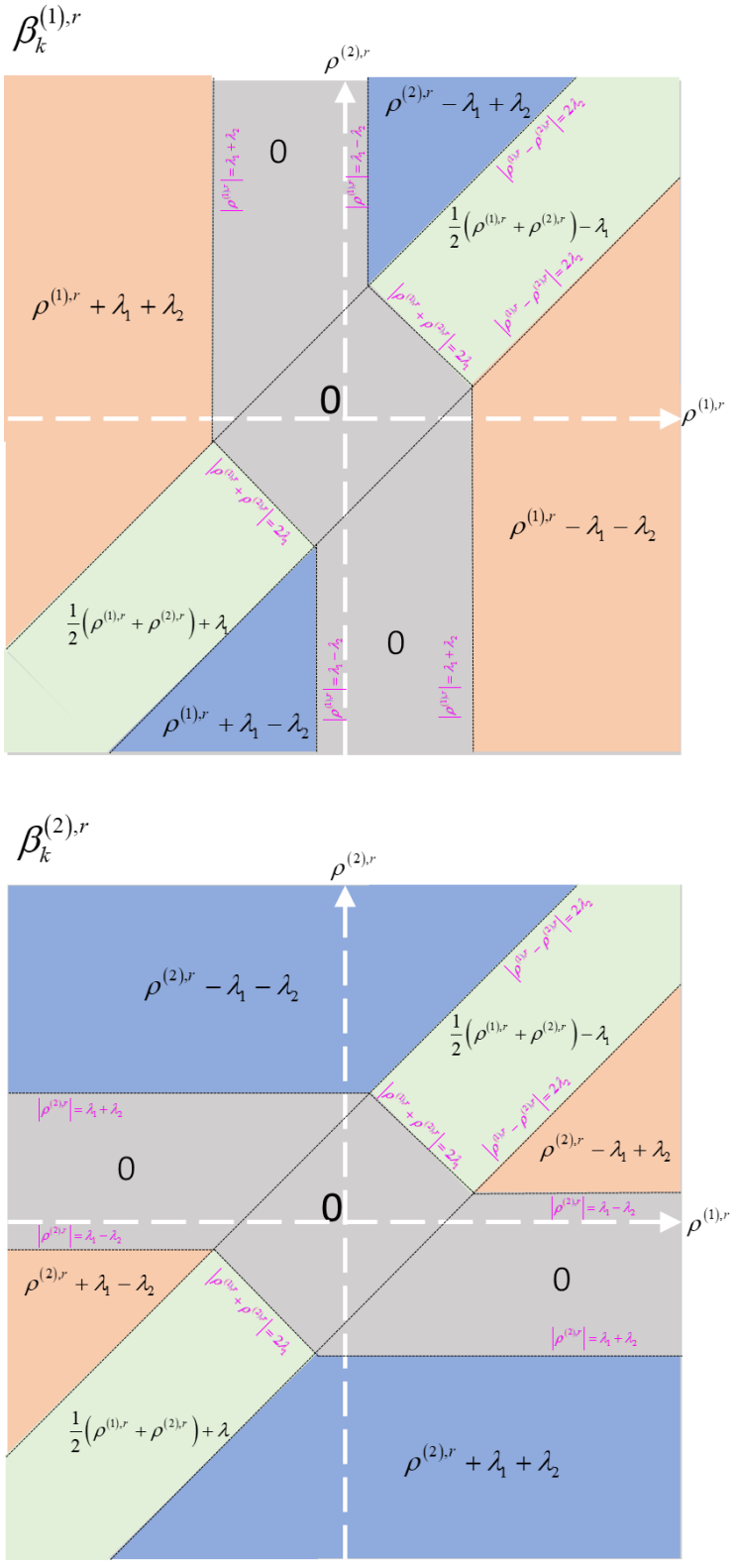


Figure 13- BCD solution subregions on the plane of rho1 and rho2

5.3.2. Accelerated BCD algorithm using the correlation matrix

The first task for accelerating the DDN method is estimating the computation complexity of the original DDN with the BCD algorithm. In a network of P nodes, DDN uses the neighborhood selection strategy to detect each node's neighbors by applying the BCD algorithm on the optimization problem. BCD uses the cyclic rule to update the coefficient for each coordinate block once in a cyclic round, and repeat the cyclic round iteratively until convergence. Assuming the convergence is reached after T times of cyclic rounds on average, DDN method in total needs TP^2 times of coefficient updating.

In each coordinate block's coefficient updating, we only count the times of multiplication operation on floating numbers as the computation complexity. For each update, the values of $\rho^{(1),r}$ and $\rho^{(2),r}$ which decide the solution of $\left(\beta_k^{(1)}, \beta_k^{(2)}\right)$ need to be calculated. In the original form of $\rho^{(1),r}$ and $\rho^{(2),r}$ defined in the last section, it firstly needs to calculate the residuals $\mathbf{y}_{i,-k}^r$ which are the summation of P weighted observation vectors, and takes about PN times of multiplication to complete. Second, the inner product of the residual vector and the current response variable vector requires another N times of multiplication operation. Therefore, calculation of $\rho^{(1),r}$ and $\rho^{(2),r}$ in each BCD update needs about $2\times\left(PN+N\right)\cong 2PN$ times of multiplication. And in conclusion, the computation complexity of DDN is about $O\left(TP^2\times 2PN\right)=O\left(2P^3N\right)$. For observations with large sample scale N or feature scale P, it will be a heavy burden of computation.

Now we look back at the solution formula of $\boldsymbol{\beta}_k^r$ in *r*-th iteration, it is clear that the updating of $\boldsymbol{\beta}_k^r$ only depends on the values of λ_1, λ_2, $\rho^{(1),r}$, $\rho^{(2),r}$. Therefore, the computation load of solving DDN by the BCD algorithm is mainly on computing the values of $\rho^{(1),r}$ and $\rho^{(2),r}$. The two

parameters by original definition take the form of an inner product between residual vector and the current node's observed data vector. Consider the fact of zero-mean and unit-variance for the standardized data, the inner products between two observed data vectors are actually linear to the elements in their covariance matrix, or equivalently, the Pearson's correlation matrix R. Therefore, we can replace the inner product of observed data by the pre-calculated correlation coefficients: $\mathbf{x}_l^{(1)} \bullet \mathbf{x}_k^{(1)} = n_1 \operatorname{cov}\left(\mathbf{x}_l^{(1)}, \mathbf{x}_k^{(1)}\right) = n_1 \mathbf{R}_{ik}^{(1)}$ and $\mathbf{x}_l^{(2)} \bullet \mathbf{x}_k^{(2)} = n_2 \mathbf{R}_{ik}^{(2)}$. In the neighborhood selection approach, the response variable is one of the nodes: $\mathbf{y}_i^{(1)} = \mathbf{x}_i^{(1)}$. Recalling the definition of $\rho^{(1),r}$ and $\rho^{(2),r}$, we rewrite them in the form of correlation matrix elements:

$$
\begin{aligned}
\rho^{(1),r} &= \frac{1}{n_1} \mathbf{y}_{i,-k}^{(1),r} \bullet \mathbf{x}_k^{(1)} \\
&= \frac{1}{n_1} \left(\mathbf{x}_i^{(1)} \bullet \mathbf{x}_k^{(1)} - \sum_{l \neq i,k} \beta_l^{(1),r} \mathbf{x}_l^{(1)} \bullet \mathbf{x}_k^{(1)} \right) \\
&= \mathbf{R}_{ik}^{(1)} - \sum_{l \neq i,k} \beta_l^{(1),r} \mathbf{R}_{lk}^{(1)} \\
&= \mathbf{R}_{ik}^{(1)} - \sum_{l \neq i,k} \beta_l^{(1),r-1} \mathbf{R}_{lk}^{(1)} \\
&= -\tilde{\boldsymbol{\beta}}^{(1),r-1} \bullet \mathbf{R}_{\cdot k}^{(1)} \\
\rho^{(2),r} &= -\tilde{\boldsymbol{\beta}}^{(2),r-1} \bullet \mathbf{R}_{\cdot k}^{(2)}
\end{aligned}
$$

In which:

$$
\begin{aligned}
&\tilde{\boldsymbol{\beta}}^{(1),r-1} = \left(\tilde{\beta}_1^{(1),r-1}, \ldots, \tilde{\beta}_p^{(1),r-1} \right), \tilde{\boldsymbol{\beta}}^{(2),r-1} = \left(\tilde{\beta}_1^{(2),r-1}, \ldots, \tilde{\beta}_p^{(2),r-1} \right) \\
&\tilde{\beta}_i^{(1),r-1} = \tilde{\beta}_i^{(2),r-1} = -1, \tilde{\beta}_k^{(1),r-1} = \tilde{\beta}_k^{(2),r-1} = 0; \tilde{\beta}_l^{(1),r-1} = \beta_l^{(1),r-1}, \tilde{\beta}_l^{(2),r-1} = \beta_l^{(2),r-1}, \textbf{for } l \neq i,k
\end{aligned}
$$

Therefore, $\rho^{(1),r}$ and $\rho^{(2),r}$ could be directly calculated from $\boldsymbol{\beta}^{r-1}$ and the observation's correlation matrix $\mathrm{R}^{(1)}$ and $\mathrm{R}^{(2)}$, without the need for direct computing on the original data matrix X. We call this computation skill as the BCD algorithm using the correlation matrix, or BCD-

CorrMtx for short, and notated the DDN method utilizing this BCD-CorrMtx algorithm as DDN-CorrMtx.

The new form of $\rho^{(1),r}$ or $\rho^{(2),r}$ has only one inner product of two P-element vectors, and thus needs only 2P times of multiplication operations in each updating iteration. Therefore, the computation complexity of DDN-Corr is $O\left(TP^2 \times 2P\right) = O\left(2TP^3\right)$, approximately N times faster than the procedure with the original definition.

5.3.3. Accelerated BCD algorithm using the residual updating strategy

Recalling that in the BCD algorithm we only update one coordinate at a time, the updated $\boldsymbol{\beta}^r$ will has most of its elements overlapped with those of $\boldsymbol{\beta}^{r-1}$ from the previous iteration. Reinspection on the residuals $\mathbf{y}_{i,-k}^r$ used in the original form of $\rho^{(1),r}$ or $\rho^{(2),r}$ definition reveals its relationship with the previous residual:

$$
\begin{aligned}
\mathbf{y}_{i,-(k+1)}^{(1),r+1} &= \mathbf{y}_i^{(1)} - \sum_{l \neq i,k+1} \mathbf{x}_l^{(1)} \beta_l^{(1),r+1} \\
&= \mathbf{y}_i^{(1)} - \sum_{l \neq i,k,k+1} \mathbf{x}_l^{(1)} \beta_l^{(1),r+1} - \mathbf{x}_k^{(1)} \beta_k^{(1),r+1} \\
&= \mathbf{y}_i^{(1)} - \sum_{l \neq i,k} \mathbf{x}_l^{(1)} \beta_l^{(1),r} + \mathbf{x}_{k+1}^{(1)} \beta_{k+1}^{(1),r} - \mathbf{x}_k^{(1)} \beta_k^{(1),r} \\
&= \mathbf{y}_{i,-k}^{(1),r} + \mathbf{x}_{k+1}^{(1)} \beta_{k+1}^{(1),r} - \mathbf{x}_k^{(1)} \beta_k^{(1),r} \textbf{, for } r \geq 1
\end{aligned}
$$

Instead of directly calculating residuals from P weighted observed vectors, we can update the residuals in a new iteration from the previous one, with a small computation load on adding 2 weighted observation vectors. We call this computation skill as the method of BCD with residual updating, or BCD-ResiUpd algorithm, and notated the DDN method utilizing this BCD-ResiUpd algorithm as DDN-ResiUpd.

For calculating $\rho^{(1),r}$ and $\rho^{(2),r}$ in each iteration, the residual updating requires $2\times\left(2N_1+2N_2\right)=4N$ times of multiplication, while the direct computing of residual requires N times of multiplication. Therefore, the computation complexity of DDN-ResiUpd is $O\left(TP^2\times 5N\right)=O\left(5TNP^2\right)$ which is about 0.4P times faster than the original DDN method.

The inner product form observed data vectors is the result of the least square's quadric form as a likelihood function. For general linear regression models, the likelihood function could take other forms of functions. Unlike the BCD-CorrMtx method, the BCD-ResiUpd method keeps the flexibility of adopting different forms of the likelihood function in solving DDN or other LASSO problems with the BCD algorithm.

For omics data studies, especially for those cover the whole genome, the feature scale P usually could reach over 20k, and the sample scale N is often far smaller than P. In these cases, the BCD-ResiUpd method could perform better than the BCD-CorrMtx method, and is considerably faster than original BCD algorithm used in DDN.

5.3.4. Accelerated BCD algorithm with integrating the Strong Rule

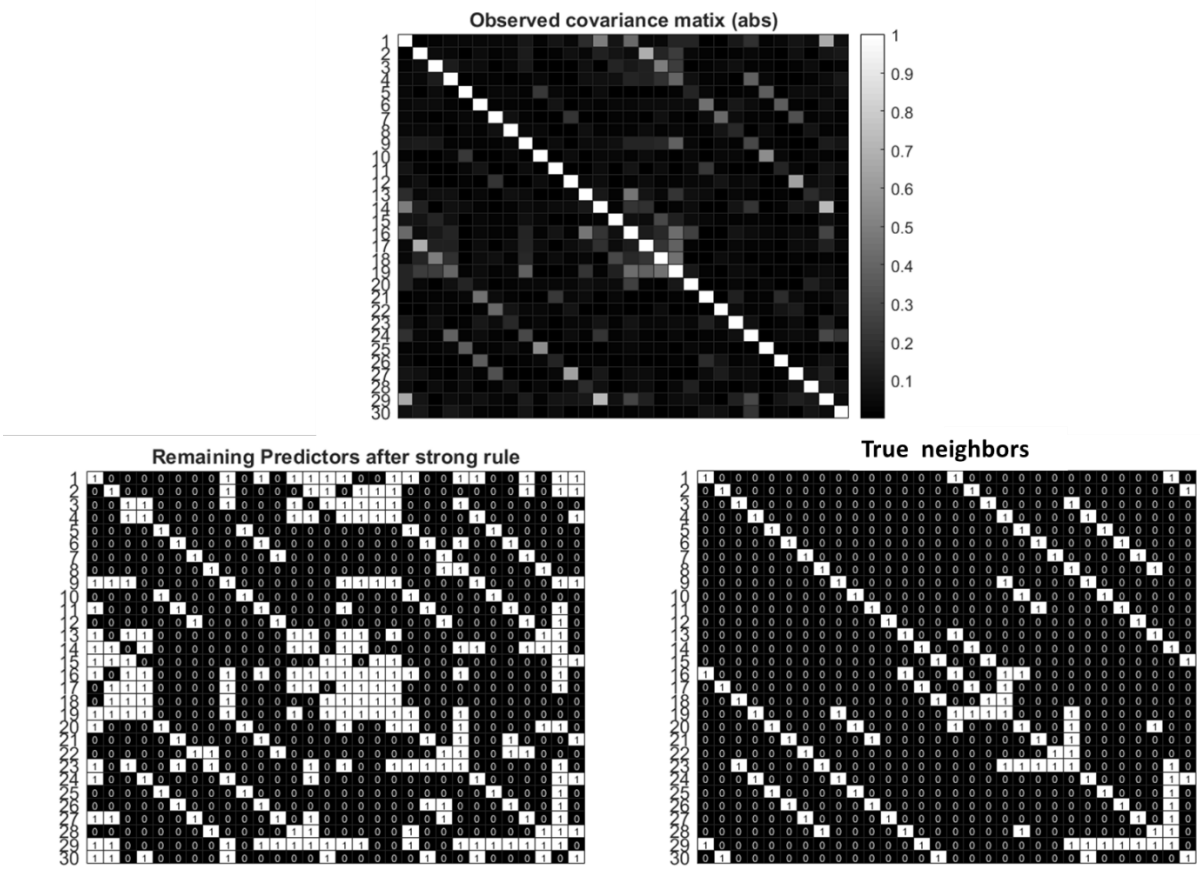
Tibshirani, et al. (2012) proposed "Strong" rules for discarding predictors in Lasso-type problems before for computational efficiency. The Strong rules are developed based on the "Safe" rules proposed by El Ghaoui et al. (2010), and are able to effectively reduce the actual number of predictors need to be solved in LASSO problems. Tibshirani, et al. (2012) also showed that, although in extremely rare cases the Strong rules may erroneously discarding predictors, the error could be amended by checking the KKT conditions.

The basic Strong rule is defined as follows: for the lasso problem, discard the j-th predictor from the optimization problem if:

$$\left|\mathrm{x}_j^{\mathrm{T}}\mathrm{y}\right| < 2\lambda - \lambda_{\max}$$

where $\lambda_{\max} = \max_j \left|\mathrm{x}_j^{\mathrm{T}}\mathrm{y}\right|$ is the smallest tuning parameter value such that $\hat{\beta}(\lambda_{\max}) = 0$.

For the BCD algorithm used in DDN, we use a cross-validation strategy to get λ_1 while setting $\lambda_2 = 0$. DDN optimization is degraded to basic LASSO problems in this case, and therefore we could apply the "Strong" rule to discard predictors before applying the BCD algorithm.



(a) the estimated covariance matrix from observed data, showing absolute value; (b) remaining predictors (white color) after applying the Strong rule to each column; C) true neighborhood matrix

Figure 14- An illustrative example of the basic Strong rule.

Figure 14 shows an illustrative example of the basic Strong rule on simulated data following multi-variate Gaussian distribution. The remaining predictors after applying the basic Strong rule still cover all true predictors, therefore the results will be identical to predictors without

applying the Strong rule. Since the number of remaining predictors is much smaller than the total number of matrix elements, the procedure of LASSO problem solving will take much less computation time(Tibshirani, et al., 2012).

5.3.5. Accelerated BCD algorithm with parallel computing

In the procedure of the neighborhood selection approach, the solution of one node's optimization is independent of the other nodes' solution. Therefore, the BCD optimization could work parallelly for each noded selected from the total P nodes in the network. In parallel computing, we assign one CPU core from a multi-core computer to independently solve one node's DDN optimization problem. The whole DDN network construction could be about *N_core* times fast, in which *N_core* is the number of available CPU cores.,

We developed an R package to implement parallel computing along with other accelerating methods of DDN-CorrMtx or DDN-ResiUpd. We denote the methods with parallel computing as DDN-CorrMtx-Parallel and DDN-ResiUpd-Parallel. In the test on simulated data discussed in the next section, we find the advantage of parallel computing in case of a large feature size P is huge, while in case of a moderate feature size the advantage is not obvious.

5.3.6. Computation time comparison on simulated data

We set a series of simulation studies to compare the actual computation time of DDN with the four proposed accelerating strategies: DDN-CorrMtx, DDN-ResiUpd, DDN-StrongRule and DDN-CorrMtx/ResiUpd-Parallel. To compare the methods on a fairground, we set the simulation conditions as follows: each of the simulated P nodes' observation data follows i.i.d standard Gaussian distribution; the true covariance matrix is set to the identity matrix, hence no edges exist in the graph and only the first round of iterations is needed for each node; λ_1 and λ_2 are also set to

big enough, therefore the convergence is expected to achieve after the first round of iterations. For each case of simulated data, all methods are using the same data as inputs.

The testing environment is listed as follows: CPU Intel® Core™ i5-8300H @2.30Ghz; RAM 16GB; R version 3.6.1. The computation time is recorded by the R function of *system.time()*.

Table 4 listed two simulated cases: one with high feature scale P and the other with high sample scale N. From the comparison of computation time used for each method we could see that, the original DDN method takes the longest time which is unbearable in large scale; DDN with correlation matrix takes little time for large sample scale N, but is slower than DDN with residual updating strategy in case of large feature scale P. Parallel computing offers great help when P is large, but the time saving is limited in case of moderate value of P.

Table 4. The computation time of accelerated DDN methods

	N=100, P=1600	**N=1600, P=100**
Original DDN	2695.48s	141.25s
DDN-CorrMtx	93.19s	0.28s
DDN-CorrMtx-Parallel (nCore=7)	1.48s	0.25s
DDN-ResiUpd	31.39s	2.94s

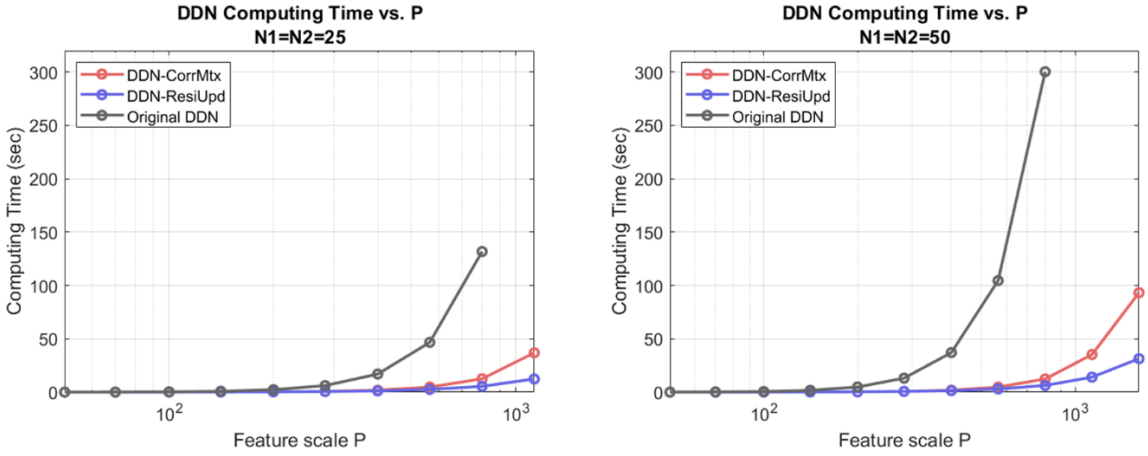


Figure 15- DDN computation time versus feature scale P

Figure 15 shows the computation time used by DDN-CorrMtx, DDN-ResiUpd and the original DDN versus the feature scale P. The DDN-ResiUpd method shows a large advantage when P grows. Figure 16 shows the computation time used by the three methods versus sample scale N. Since the size of the correlation matrix only relies on P, the DDN-CorrMtx method uses almost the same time for different sample scales, and hence is much faster than the other two methods when N is large.

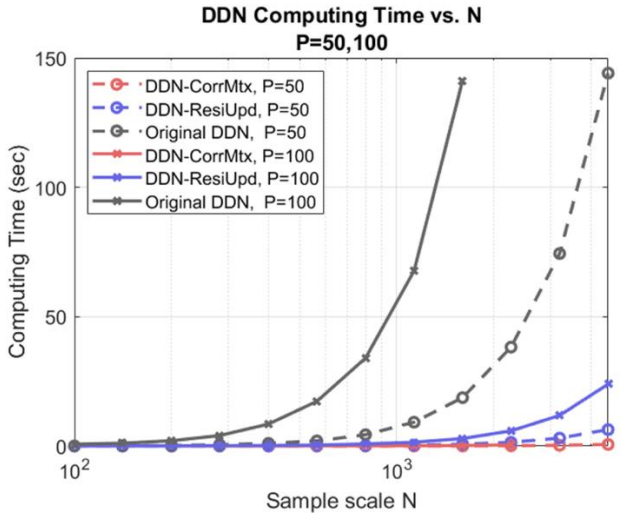


Figure 16- DDN computation time versus the sample scale N

We compare the computation time of the DDN-CorrMtx methods with and without parallel computing (Figure 17). Since parallel computing is done for each node selected from all P nodes, the parallel computing uses much leas computation time when the feature scale P is large. On the other hand, the time saved by parallel computing is limited when P is small.

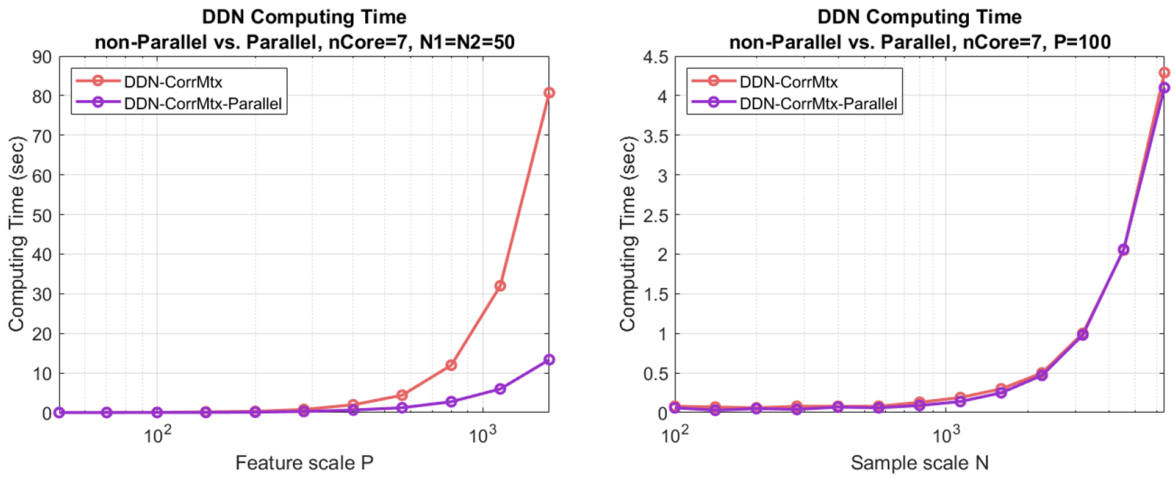


Figure 17- DDN computation time comparison between parallel and non-parallel computing

Table 5 shows the effectiveness of the Strong rule. In this simulation study, the ground truth network is designed as a single ring in which each node has exactly two neighbors. Therefore, the basic Strong rules could effectively discard most of the predictors. The results show that computation time could be significantly reduced by applying the Strong rule when the network has high sparsity.

	N=100, P=400	**N=100, P=800**
DDN-CorrMtx	57.86s	393.15s
DDN-CorrMtx+ Strong Rule	2.15s	5.91s

Table 5. Computation time comparison between DDN with and without Strong rule

In summary, we proposed three reformulated computing methods for the BCD algorithm used in solving optimization in DDN. We also propose accelerating strategies of discarding predictors by the Strong rule and by parallel computing. Depending on which one of the sample scale N and the feature scale P has the larger value, we may choose from the method of computation with correlation matrix, or residual updating strategy, or with combined effort in coefficient updating strategy. The results show a tremendous reduction in computation time comparing to the original DDN method. The proposed DDN-ResiUpd method with parallel

computing is now able to handle hundreds of genes in a reasonable time period (<1 hour), compared with the original DDN method which is only capable of handling dozens of genes.

5.4. DDN application on single-omics biomedical data

5.4.1. DDN application on discovering the proteomic architecture of human coronary and aortic atherosclerosis

Atherosclerosis is defined at the molecular level as an assembly of intra- and extra-cellular proteins which jointly alter the cellular processes and produce characteristic remodeling of the local vascular environment. In GPAA research projects, the human arterial proteomics data are collected from the coronary artery and aortic specimens from 200 arterial specimens(Herrington, et al., 2018). The proteins were identified and quantified using high-resolution mass spectrometry from the Thermo Scientific™ LTQ-Orbitrap platform using liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS). The DDN analysis compared the two groups, respectively the diseased tissue samples and the normal tissue samples. The differential network detected defined the composition of the protein networks and also the regulatory features likely associated with early atherosclerosis.

GPAA research project includes data from 100 autopsies included in the study(Herrington, et al., 2018). The samples are collected from two anatomic locations of the left anterior artery (LAD) and distal abdominal aorta (AA). Each sample will then be examined by pathologists to evaluate the proportions of four tissue types: complicated lesion (CO), fibrous plaque (FP), fatty streak (FS), and normal tissue (NL). According to pathologist' evaluation, no sample contains CO tissue, 66 samples were evaluated as completely normal, and the rest samples contain certain proportions of FP tissue and/or FS tissue. These evaluated proportions, along with CAM (a data-

driven deconvolution method) estimated proportions, are then analyzed with principal component analysis. The samples are eventually clustered by the first two principal components into 3 major groups, namely FP, FS and NL group. The comparison between the FP group and the NL group are of most interest, since the existence of FP tissue indicates more severe atherosclerosis status. Giving the fact that majority of samples are evaluated as completely normal, in order to reduce the impact of imbalance of group sizes, we further selected 30 samples from normal group with top principal component values so that we could have an accept group size ratio (2:1) with FP group while retaining as many samples as possible.

We then put differentially expressed proteins between FP and NL groups into IPA analysis. Several pathways are found significantly enriched with these proteins. Among them, LXR/RXR activation pathway, collagen pathway and complement system pathway are of high interest in the atherosclerosis research community. We also performed GO term enrichment with GO Term Finder and pathway enrichment analyses with IPA (Ingenuity Pathway Analysis, Ingenuity Systems). Comparing selected normal (NL) and atherosclerosis-enriched (FP) LAD samples identified 89 individual proteins with absolute fold change > 1.7 and a t-test q-value <0.05 for FP versus NL. Bioinformatic functional analysis of these atherosclerosis-associated proteins revealed a pattern consistent with activation of the tumor necrosis factor-α (TNF-α) pathway but also inhibition of insulin receptor, peroxisome proliferator-activated receptor-α (PPAR-α), and peroxisome proliferator-activated receptor-γ (PPAR-γ) pathways. We perform DDN analysis on LAD samples' protein expression data of these identified 89 genes between FP and NL groups, and on the top significantly enriched pathways and the gene groups regulated by top upstream regulators (TNF-α, INSR, PPAR-α, PPAR-γ). In a similar way 80 genes are identified by GO term enrichment and IPA analysis for the AA samples. Although in the AA sample proteomes the

enrichment analysis gives no significant call to inhibition of the insulin receptor, PPAR-α, or PPAR-γ pathways, DDN results show distinct network patterns and rewiring differences for the two anatomic locations (LAD and AA).

Each of the activated and inhibited regulatory pathways identified in the LAD samples (TNF, INSR, PPAR-α, and PPAR-γ) exhibited significant re-wiring among pathway proteins between FP and NL samples. The re-wiring models reveal selective coupling or uncoupling between specific protein pairs in these regulatory networks depending on the specific tissue phenotype (FP or NL). These differential networks suggest major disruptions in energy metabolism, extracellular matrix remodeling and PKA signal transduction when compared to normal tissue. Significant network rewiring events in the setting of fibrous plaque is also detected in five additional pathways that were over-represented in both the LAD and AA samples (TGFB1, TP53, MEGEA5, MYC, ERBB2).

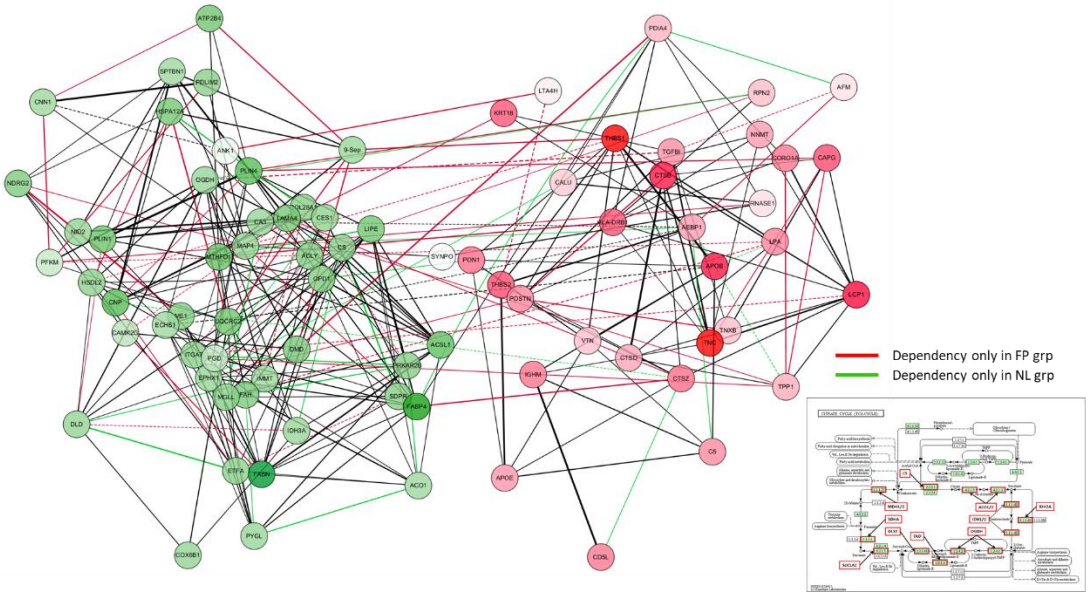


Figure 18- DDN results for 89 selected genes on proteomics data from LAD samples

With detected significant rewiring events both in FP-associated proteins and in candidate gene set from IPA picked pathways and master regulators, DDN analysis gives a comprehensive

understanding of arterial protein networks and how they change in early atherosclerosis. It also suggests that the human arterial proteome can be viewed as a complex network whose architectural features vary considerably as a function of anatomic location and the presence or absence of atherosclerosis. These divergent proteomic features allude to anatomic specificity could also have important implications for personalized treatment and prevention.

For the 89 identified FP-associated proteins, DDN detected a group of proteins pivotal in the rewiring of the network structure between NL and FP in the LAD samples (Figure 1). Further analysis of these 26 rewiring hub proteins revealed significant enrichment of tricarboxylic acid proteins (p-value=4.8E–6). Subsequent DIA-MS analysis of n=114 mitochondrial proteins was performed comparing FP versus NL samples from the LAD. The results document an average 60% reduction of a wide range of mitochondrial proteins in the FP samples in comparison with NL samples after adjustment for vascular smooth muscle-specific housekeeping proteins, age, and sex. This consistent reduction indicates divergent mitochondrial dynamics in the setting of atherosclerosis characterized by reduced mitochondrial mass in the LAD. In contrast, a similar analysis of the same proteins in the distal Abdominal Aorta (AA) samples revealed a much less consistent and statistically non-significant pattern of mitochondrial protein suppression. Collectively, these data demonstrate how DDN detected network rewiring could produce biologically coherent insights that may not be evident from conventional statistical or pathway enrichment analysis strategies.

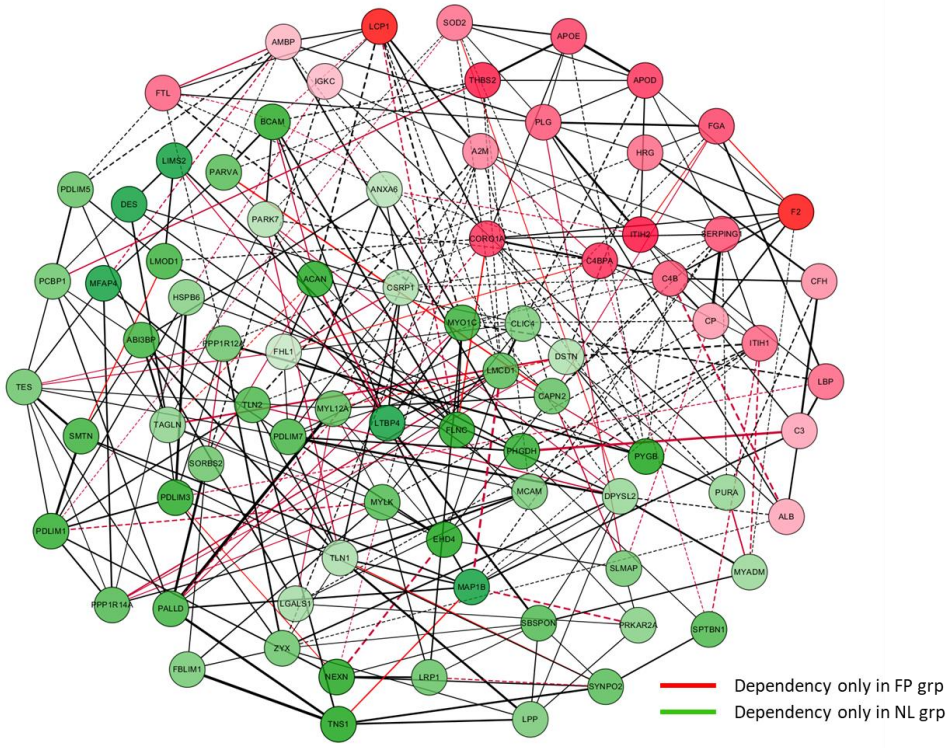


Figure 19- DDN results for selected genes on proteomics data from AA samples

We also found some protein groups play important roles in network changes from normal to disease. For example, low-density lipoproteins (LDL) and low-density lipoproteins (HDL) in the LXR/RXR activation pathway, and collagen proteins in the collagen system pathway are linked with many detected network rewiring.

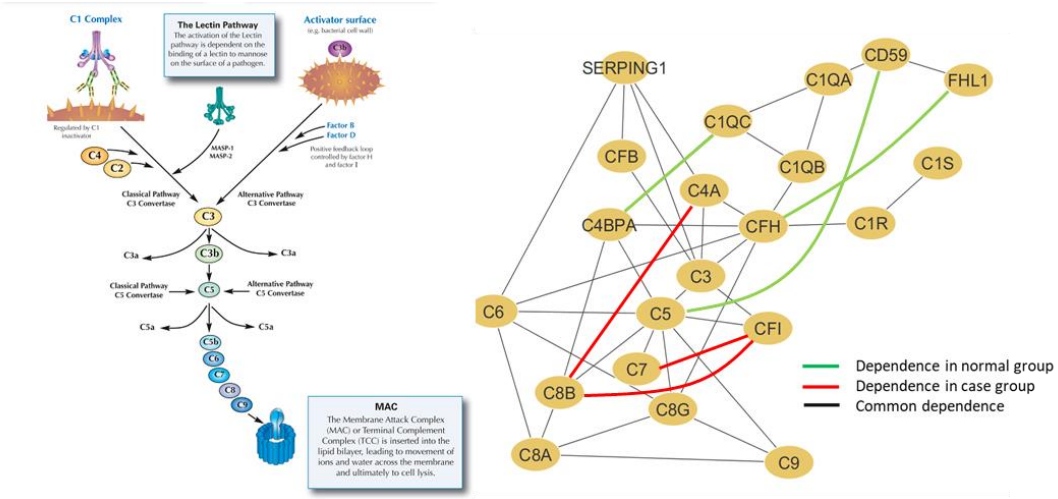


Figure 20- Illustration of complement system pathway and DDN results on this pathway

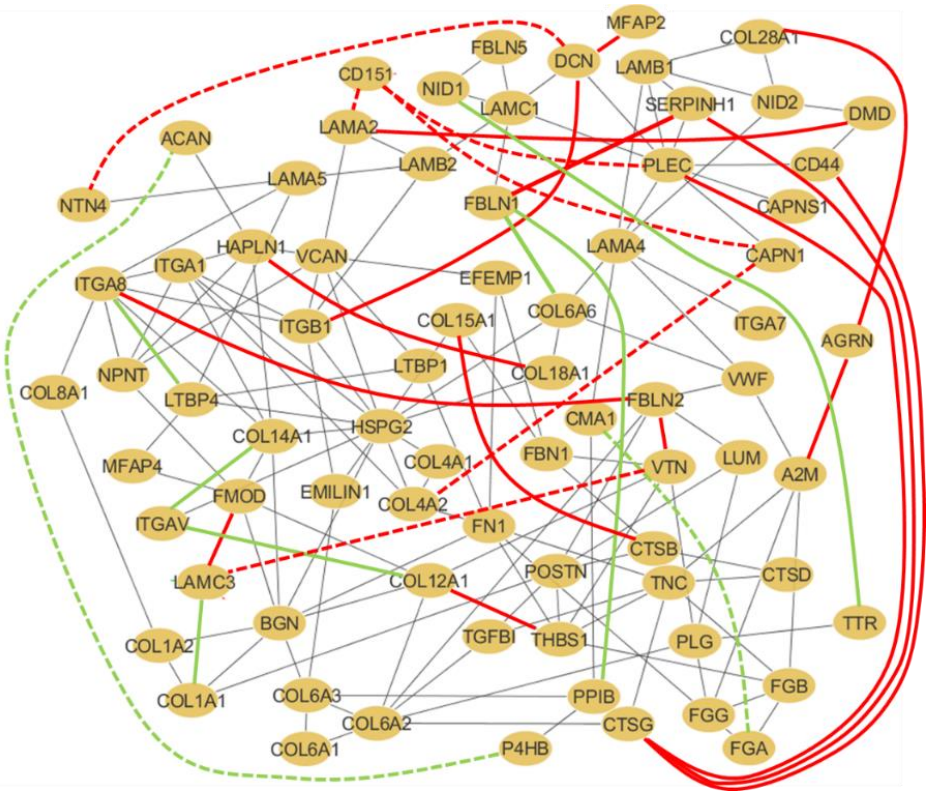


Figure 21- DDN network on genes from the GO term "extracellular matrix organization"

5.4.2. DDN application on the proteomic characterization of ovarian cancer

DDN tool is used in CPTAC (Clinical Proteomic Tumor Analysis Consortium) project of integrated proteogenomic characterization of human high grade serous ovarian cancer(Zhang, et al., 2016). CPTAC project ovarian cancer group performed a comprehensive mass-spectrometry-based proteomic characterization of 174 ovarian tumors previously analyzed by The Cancer Genome Atlas (Zhang, et al., 2016). DDN analysis was applied to 171 BRCA1/BRCA2-related proteins on global proteomics data, between homologous recombination deficiency (HRD) positive and HRD negative groups.

DDN detected results show distinct patterns of network rewirings between HRD positive and HRD negative patients (Figure 25). The quantitative level of lysine-acetylated peptides both

in iTRAQ global protein expression data and in the validating SWATH data showed that acetylation levels of two key peptides from histone protein H4 have a significant difference between HRD positive and HRD negative samples.

In the second stage of the CPTAC ovarian cancer research project, additional 100+ samples of which over 84 were HGSC tumors are included in the prospective sample set, and characterized in proteomics with the new TMT10 labeling technique. The global protein expression data of these prospective samples once again validated the statistically significant lower acetylation levels in both K12 and K16 peptides of H4 histone as shown in Figure 25.

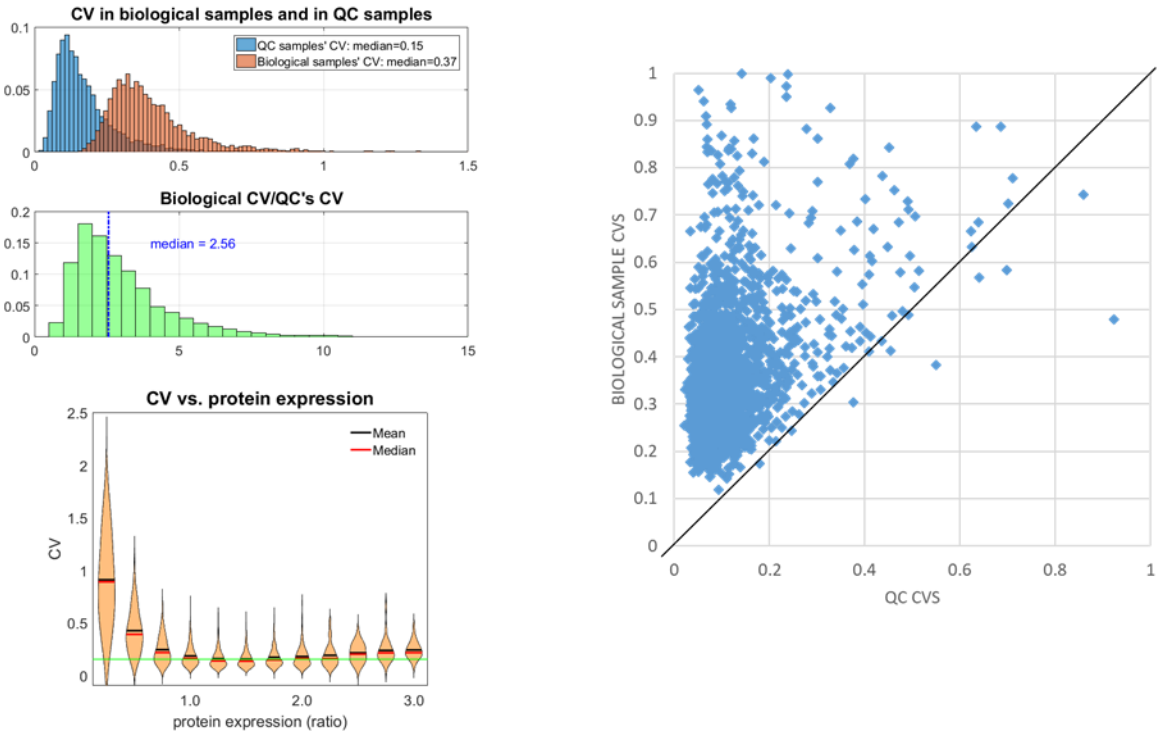


Figure 22- Data processing of CPTAC-OV prospective samples shows good sample quality

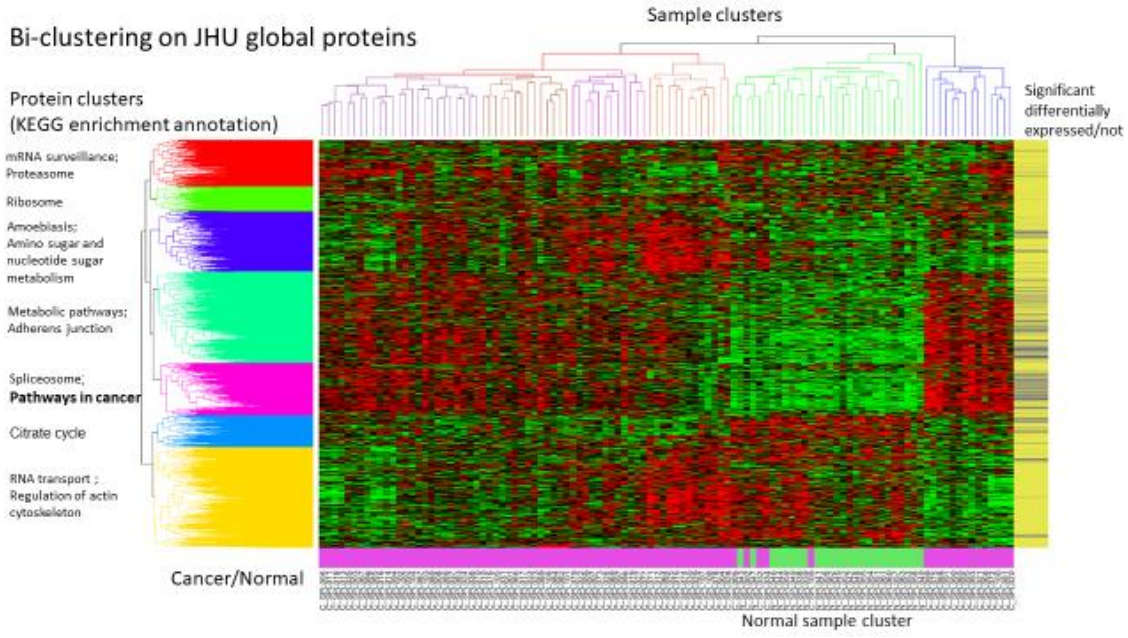


Figure 23- Bi-clustering on protein expression matrix shows a clear separation between tumor and normal samples

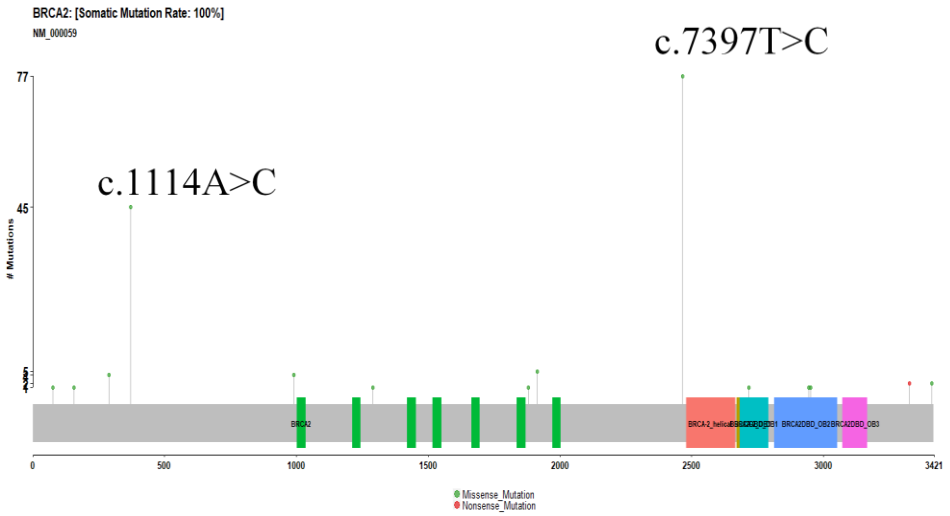


Figure 24- Mutation calling on CPTAC-OV genomics data gives a high mutation rate of BRCA2 gene

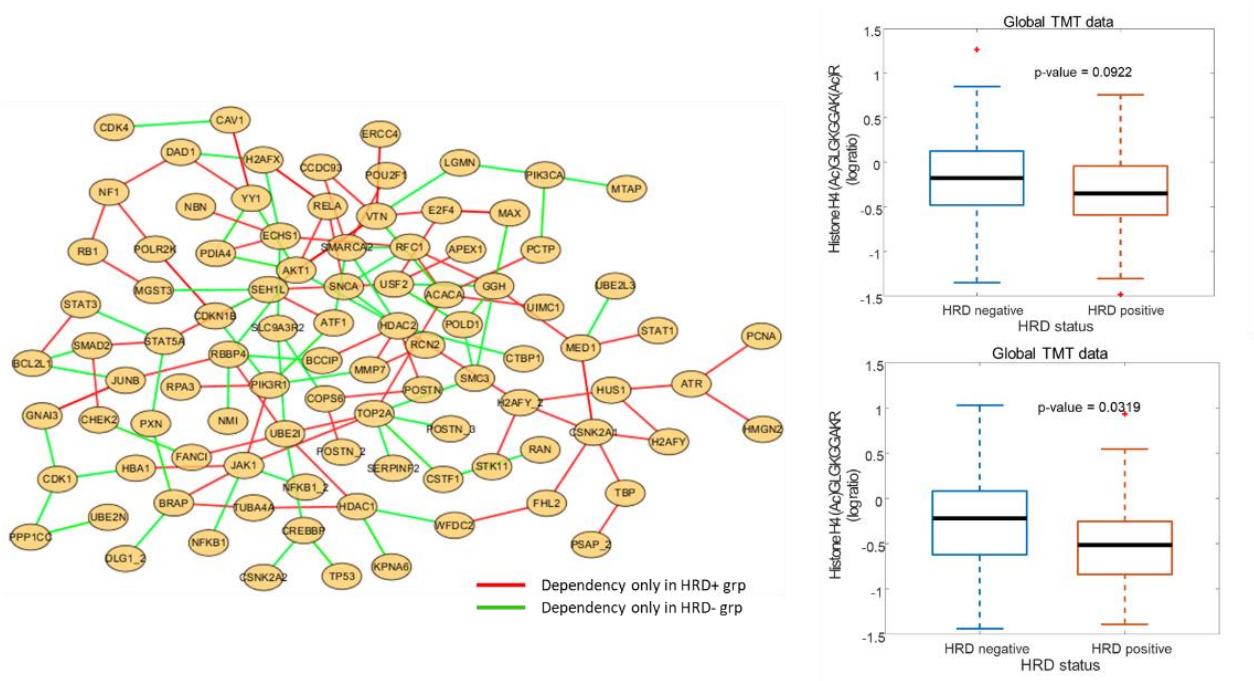


Figure 25- DDN results on CTPAC-OV data and comparison of the acetylation levels of histone H4 peptides

For CPTAC retrospective sample set, homologous recombination deficiency (HRD) is defined by the presence of germline or somatic BRCA1 or BRCA2 mutations, BRCA1 promoter methylation, or homozygous deletion of PTEN. For the CPTAC prospective sample, HRD samples are classified by germline or somatic BRCA1/BRCA2/PTEN mutation. The BRCA1/BRCA2-related target genes are collected and integrated from the literature.

To select the candidate genes from the whole collection of data, we choose some core pathways that have been well studied and proved to be highly related to cancer. DDN results on the MAPK signaling pathway highlighted the AKT gene as one of the hub genes pivotal in the network rewiring events. AKT is modulated by PTEN mutation which is one of the key definers of HRD status. Figure 26 shows this network constructed on the protein expression data from 122 samples of CPTAC retrospective dataset, with Genotype grouping of 66 HRD positive vs. 56 HRD negative.

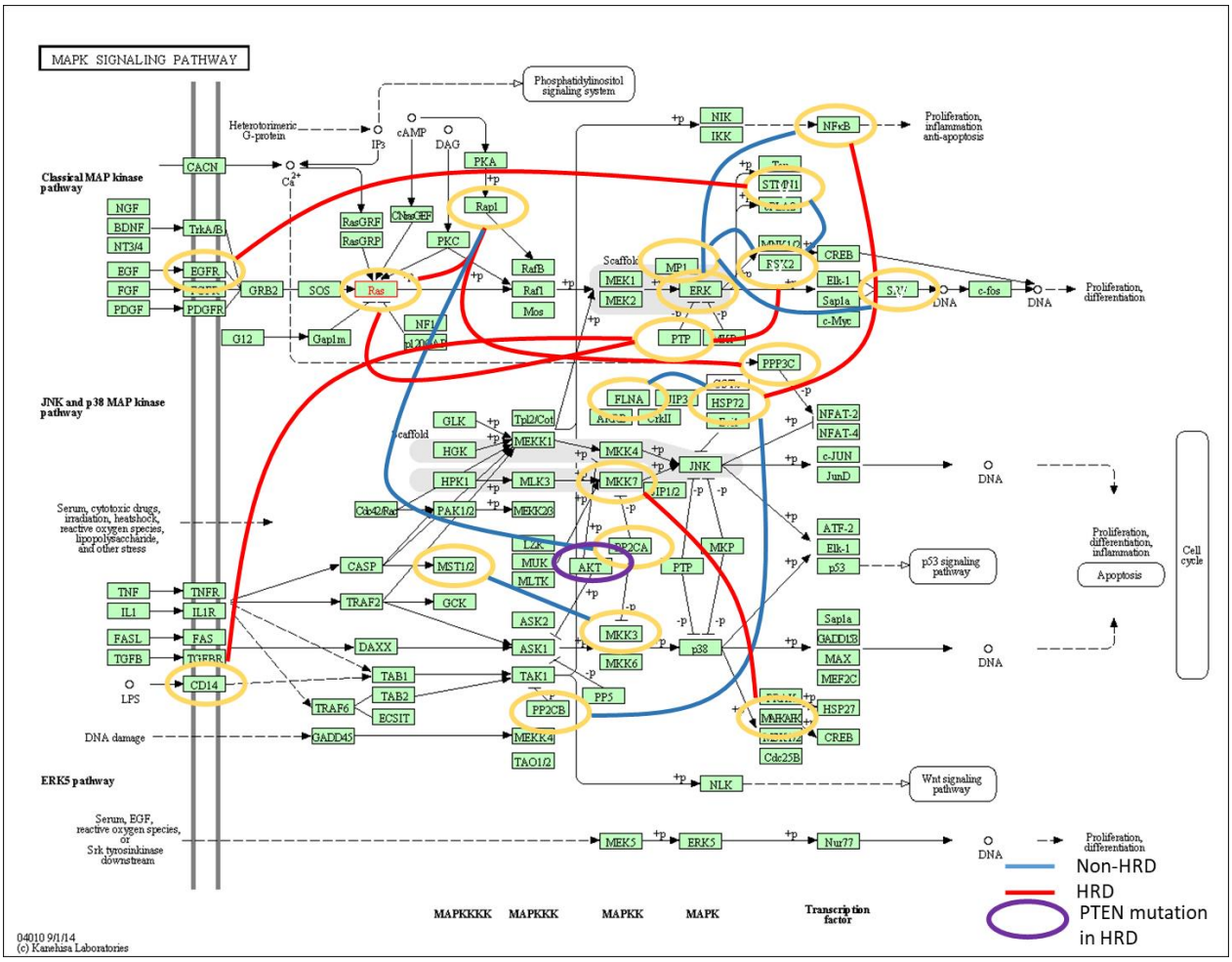


Figure 26- DDN network overlapped on the KEGG diagram of the MAPK signaling pathway.

5.4.3. DDN application on the transcriptomic characterization of psychotic disorders

Schizophrenia (Schiz.) and bipolar disorder (BP) are psychiatric disorders with relatively high heritability. Reports show transcriptome shows dynamic dysregulation in the diseased brain sections (Colantuoni, et al., 2011; Kang, et al., 2011; Kuhn, et al., 2011). The center of depression and resilience at the University of Illinois at Chicago conducted a research project investigating the deregulation changes in bipolar disorder (BP) and schizophrenia (SCH). Cerebellum (CB) and parietal cortex (PC) brain tissues(Chen, et al., 2014). The biological samples were obtained from Neuropathology Consortium and Array collections, and the genome-wide gene expression data

were tested by Affymetrix microarray chips. Samples include cerebellum (CB) and parietal cortex (PC) brain tissues collected from 150 subjects, including 50 bipolar disorder (BP) samples, 50 schizophrenia (Schiz.) samples and 50 unaffected (control) samples(Chen, et al., 2013) are collected from 150 subjects, including 50 bipolar disorder (BP) samples 50 schizophrenias (Schiz.) samples and 50 unaffected samples.

Expression data are from the NCBI Gene Expression Omnibus (GEO) database of GSE35978. One of the two PC data sets came from SMRI samples (PFC-SMRI), and the second came from the Victorian Brain Bank Network (PFC-VBBN) and was obtained from the GEO database of GSE21138.Candidate gene sets for DDN analysis are selected from the top pathways picked by the IPA tool and also from the genes in two co-expression modules from WGCNA analysis.

The DDN results (Figure 27) show there are multiple significant network rewirings in the cell adhesion pathway for each comparison scenario. These network rewirings highlight the fact that the two psychiatric disorders of interest likely involve numerous proteins and dynamic alterations in gene regulation networks rather than one or a few isolated proteins. We also observed some highly similar gene expression patterns between schizophrenia and bipolar disorder in selected pathways. For example, the gene ITGA8 is significantly differentially expressed between CB and PC samples. It shows a common network rewiring feature that this gene has disconnections in the diseased group with significant p-values for both BP and Schiz. diseases in CB samples and for Schiz. disease group in PC samples. On the other hand, DDN analysis revealed some in-depth differences of network rewiring events of the two psychiatric disorders, which could be hidden in traditional differential analysis of gene expression level.

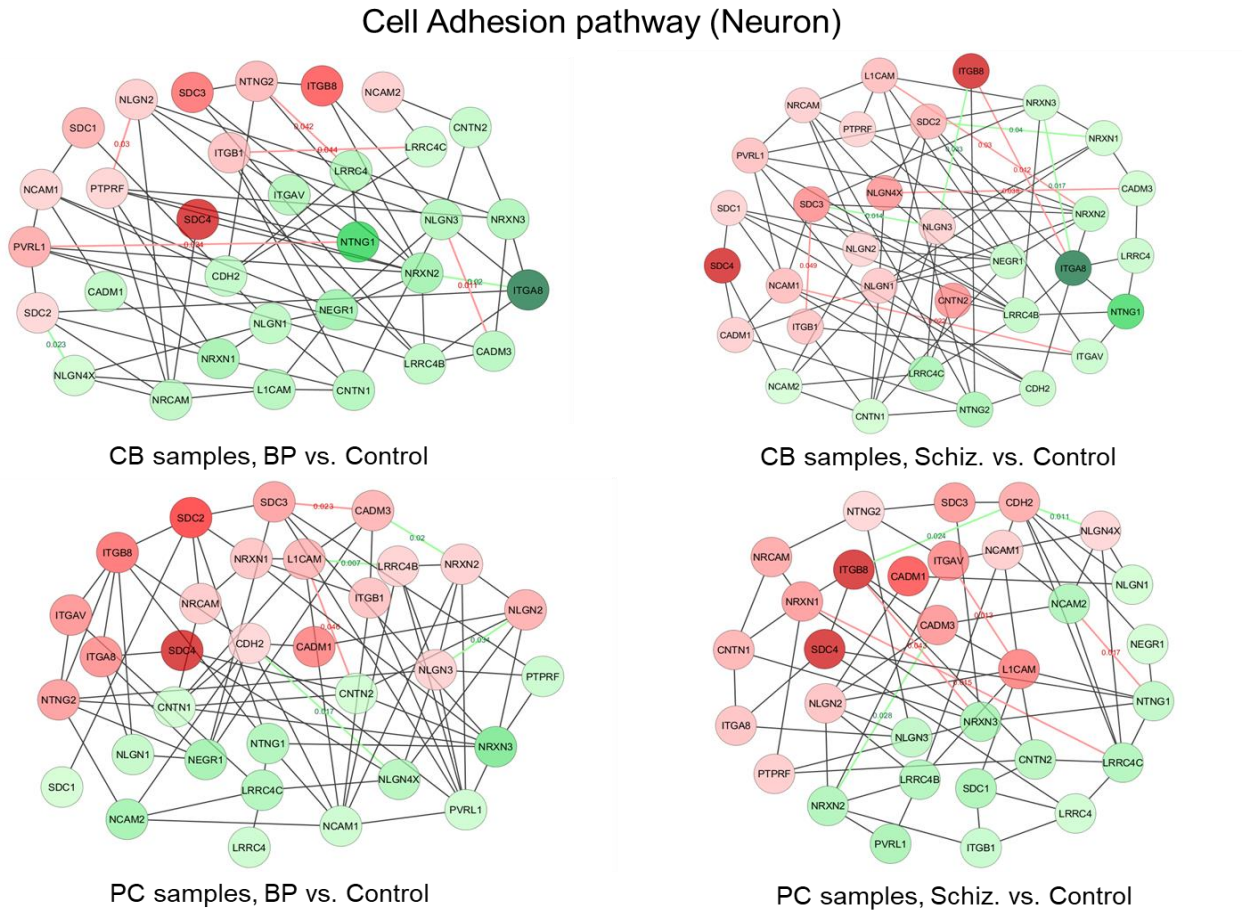


Figure 27- DDN detected rewiring events in gene expression dependency network from samples of the two psychiatric disorders (BP, Schiz.) group vs. the control group

Chapter 6. multiDDN detects intra-omics and inter-omics differential dependency from integrated multi-omics data

6.1. Introduction

Data integration has a long history in omics data study(Gatza, et al., 2014). There are different categories of data integration methods. Depending on the accessible types of omics data and the purpose of integration, we may use one or more kinds of integration methods and applied to different levels of data. For example, for raw data of single omics but are collected from multiple platforms, we could choose data cleaning and data summarization to combine different sources at a higher data level. For omics data and prior knowledge data, we may design a suitable analysis model to fuse the knowledge into omics data analysis. For single omics data and phenotype data, we may use differential expression analysis and many other bioinformatics tools to build the links and discover phenotype-specific features. For data of different omics types, in other words multi-omics data, one simple yet effective integrating method is to standardize multiple omics data into z-scores and naively merge into a single data matrix. For example, Barretina, et al. (2012)in the Cancer Cell Line Encyclopedia (CCLE) project adopt this simple multi-omics integration method to genomics and transcriptomics data from 947 human cell lines, to predict anticancer drug sensitivity.

Tough the strategy of this naïve merging of multi-omics data is simple, it suffers many drawbacks. It completely ignores the inter-omics interaction, and cannot benefit from the knowledge of regulation in the multi-omics data. The integrated feature scale which is simply the summation of the feature scales of all omics types, could be too large to handle. The data from

different omics types usually follow distinct distributions, and naïve merging which ignores distribution differences could bring bias in downstream analysis.

Some other multi-omics integration methods take the use of inter-omics associations. For example, expression quantitative trait loci (eQTL) analysis uses both genomics and transcriptomics data to detect genomic loci that explain the variation of gene expression levels(Shabalin, 2012). Similarly, methylation quantitative trait loci (mQTL) analysis tries to find loci associated with methylation levels(Volkov, et al., 2016). CPTAC-OV project explored the correlation between chromosome instability summarized from genomics data and the abundance of proteins(Zhang, et al., 2016). These multi-omics integration methods enable researchers to gain unique insights into inter-omics associations.

In integrating multi-omics data in differential networks analysis, naïve merging is inapplicable for the DDN method on genome-wide data due to DDN's limitation on computing complexity. We also believe that the integration of omics should be more than the sum of its parts. Therefore, in this dissertation we design a multi-layer integrated signaling model to reduce the total feature scale by incorporating knowledge of inter-omics interaction. On top of this signaling model, we propose a differential network analysis method called multiDDN that is capable of integrating multi-omics data and detecting both intra- and inter-omics network rewiring.

6.2. Integrating DNA and mRNA data with multiDDN

Our proposed data model starts from a simpler model of two types of omics data: DNA copy number and gene expression (mRNA expression).

Gene expression is one of the most important measures in transcriptomics study. It describes the abundance of one gene's transcribed mRNA molecules. There have been a large

number of available biological datasets and databases that contain gene expression measures. Gene regulation network has long been introduced to analyze gene expression data and discover the regulatory relationship between genes. However, gene expression is regulated by multiple factors, like transcription factors, methylation in the promoter region, etc. Using gene expression data alone is not enough to accurately reconstruct the gene regulatory networks.

Yuan, et al. (2011) proposed an integration method on copy number and mRNA expression to detect the differential association between the two omics data. In that method they use linear regression that treats mRNA expression solely as the response variable and all gene copy numbers as predictors. However, in molecular biology there is no direct interacting mechanism between these two molecules except the gene dosage effect which is the correlation between copy number and gene expression in gene transcription, hence the association detected by this method is indirect and difficult to interpret. We believe that direct regulation between genes is performed by DNA's downstream products like RNA molecules or proteins, and network built from RNA or protein expression may better detect these regulations than networks from DNA genetic variants. In other words, the correlation between gene copy number and gene expression, not including gene dosage effect, is more likely caused by confound factors like cis-correlation but not by direct DNA-RNA regulation(Bryois, et al., 2014).

One of the most import regulatory factors to gene expression level is the gene dosage effect from the transcribed gene itself (Gardiner, 2004). A gene's dosage which is quantified as the gene copy number determines the maximum number of copies of the gene that can be transcribed into mRNA simultaneously. Yang, et al. (2007) reported that the gene dosage effect could reach as high as r=0.98 in some genes such as HER2 and GRB7. In our study, we calculated Spearman's correlation coefficients between one gene's copy number and its own gene expression on a large

number of samples (N=109, TCGA-OV dataset), and found the gene dosage effects are universal and largely positive and across whole genome (Figure 28), with a median value as high as 0.43. In comparison, the correlation between one gene's copy number to another gene's mRNA expression is insignificant, with a median value of zero across the genome (Figure 28). The results confirm that gene copy number has a large contribution to its own gene expression's variation, and this direct interaction through transcription and should not be ignored when listing the regulators of gene expression.

Gene dosage effects could also help to discover hidden regulations. For example, the abnormal negative correlation coefficient suggests a higher chance of strong repression regulation from other genes. Dong et al (2015) discussed the possibility of exploring such conflicting correlations.

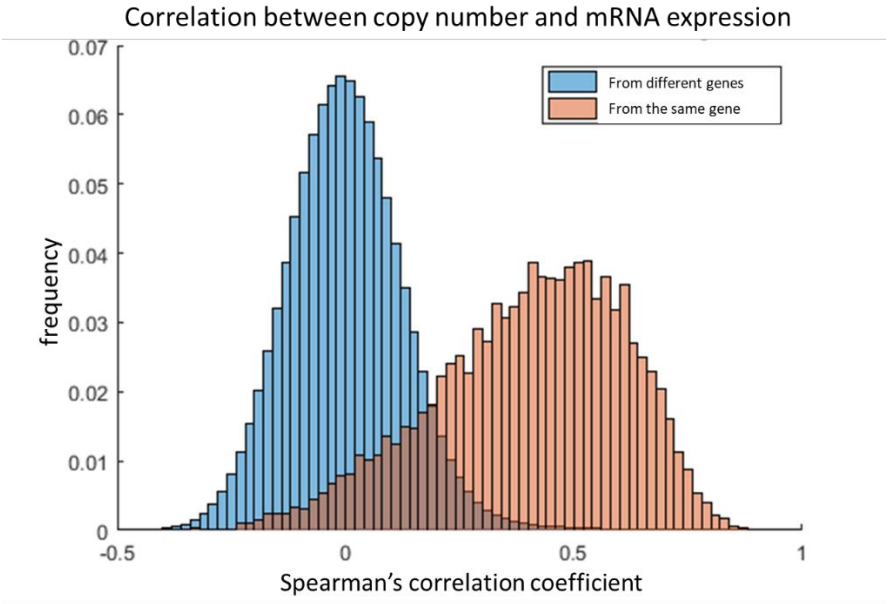


Figure 28 – Histogram of the correlation coefficients between copy number and mRNA expression. The blue ones are the correlation coefficients between one gene's copy number with another gene's expression, and the distribution is centered at 0. The orange ones are from the correlation between copy number and gene expression of the same gene. The distribution with the median value of 0.43 shows the gene dosage effect.

By sorting the significant correlations between gene's copy number variation and mRNA expression with their genomic locations, we also noticed that gene locations are highly likely to positively correlated with copy numbers from genes in nearby genomic locations, which appear like chromosome-wise squares along the antidiagonal direction in the plot (Figure 29). This phenomenon referred to as cis-correlation is likely caused by long-range structural variations in the DNA chain, and is confirmed by many reports(Bryois, et al., 2014; Yang, et al., 2007). It is one of the major confounding factors in gene regulatory network inference, and our integrated model aims to alleviate this confounding factor.

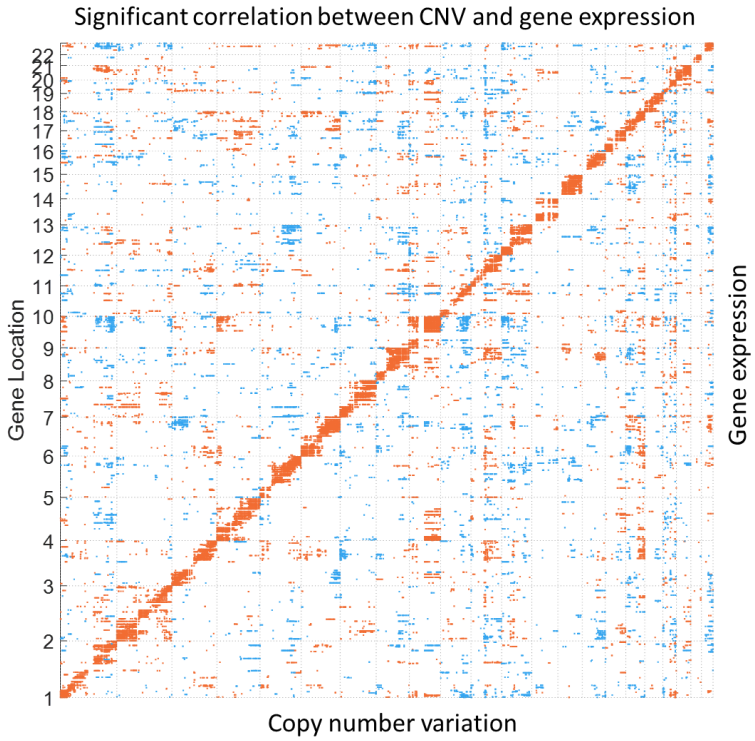


Figure 29 - Significant correlation between CNV and gene expressions, sorted by genomic location

Based on the facts of the gene dosage effect of copy numbers, we propose an integration method to add DNA copy number information as additional predictors to gene expression variation in the gene regulation network construction. Figure 30 shows an illustrative example of this model shows how the DNA copy number information interacts with the gene regulation network in the

mRNA level. Although DNA may carry germline or somatic copy number variations in living cells, it is commonly believed that the numbers of gene copies are not regulation targets and are not altered by other biomolecules. Therefore, in the network construction we may treat DNA copy number signals solely as input predictor variables. We also limit its response variables as its gene expression variable due to the strong gene dosage effect, so that only one additional predictor is added to each node in gene expression layer, and the total feature scale in each node's DDN optimization problem will be p+1 instead of 2p.

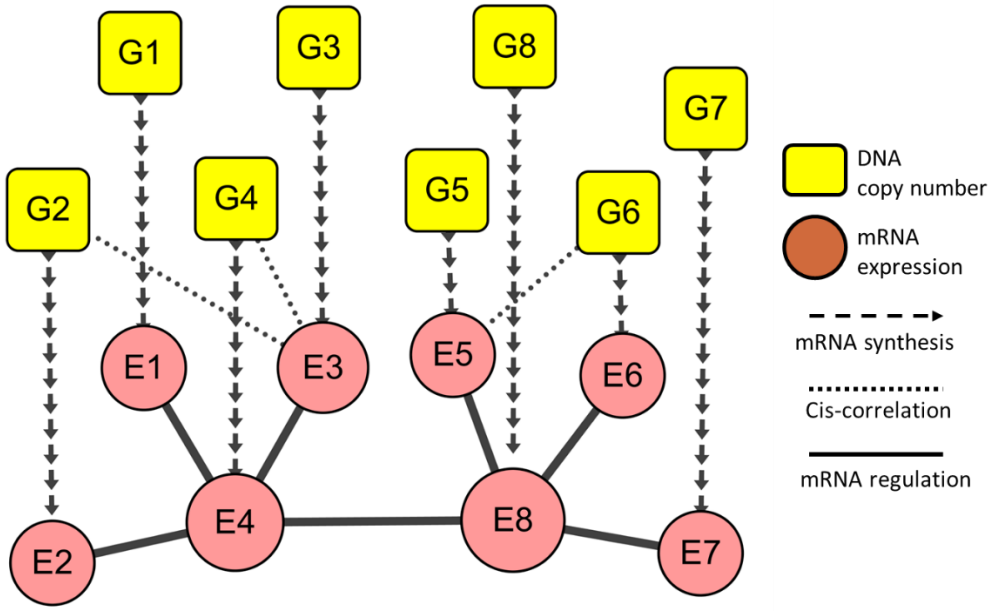


Figure 30- Integrated data model of gene copy number and mRNA expression

A pilot simulation study is designed to show the integrated model's effectiveness in inferring a gene regulatory network. We predesigned a gene regulatory network of 15 genes in the RNA layer, and associate the genes' expression to their own copy numbers in the DNA layer and some additional weaker links that represent the cis-correlation effect. The copy number and gene expression values are sampled from multivariate Gaussian distributions. We compare two methods to reconstruct the sparse networks: the first one is the baseline LASSO regression approach for the single-omics data (gene expression data) alone; and the second one is the integrated method of the

LASSO regression with an additional predictor of copy number of the same gene. We use the receiver operating characteristic curve (ROC) which is the curve of the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. We also use the bootstrapping method with 100 times of boosting to evaluate the confidence interval (CI) of the ROC. The simulation result shows the integrated method has a larger area under the curve (AUC) than the method on single-omics data alone. And for a given false positive rate (FPR), the integrated method has a significantly higher true positive rate than the single-omics method.

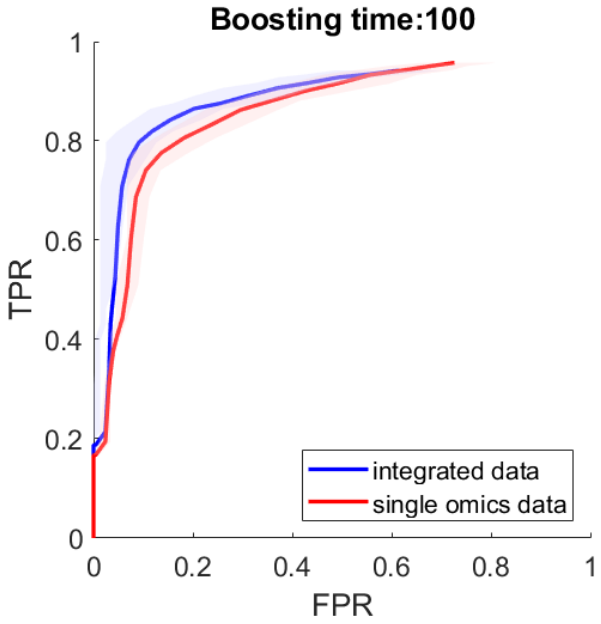


Figure 31- ROC curves of constructing sparse genetic networks from single omics data and from integrated data

6.3. The integrated data model of multiDDN

Encouraged by the success of introducing additional regulators from other omics layers in network reconstruction, we continued to add a new type of regulator: the transcription factor (TF) from proteomics data. TF is a subtype of proteins that actively involved in gene transcription by binding to the gene's promoter regions. The abundances of one gene's binding TFs are actively associated with this gene's expression level. Similarly, we also restrict the inter-omics association

between gene expression and TF expression limited to those TF binding with theoretical or experimental evidence, in order to reduce the total feature space and make the detected inter-omics dependency more plausible

Based on these principles of gene dosage effect and TF binding, we design a three-layer data signaling model to integrate the three types of omics data: copy number signals from genomics data in DNA layer, gene expression levels from transcriptomics data in RNA layer, and TF expression levels from proteomics data in protein layer. Figure 32 shows an illustrative sample of our designed multi-layer omics data model. On top of this model, we propose a novel method called multiDDN that constructs differential dependency networks on integrated multi-omics data instead of single omics.

The role of TFs in the integrated data signaling model is the regulators of the genes in the RNA layer. Gene regulatory network is inferred mainly from gene expression data and from inter-omics dependency between gene expression and TF expression. We are not detecting regulations between TFs in this model, but the method of multiDDN could be easily extended to include such dependencies between nodes in the protein layer.

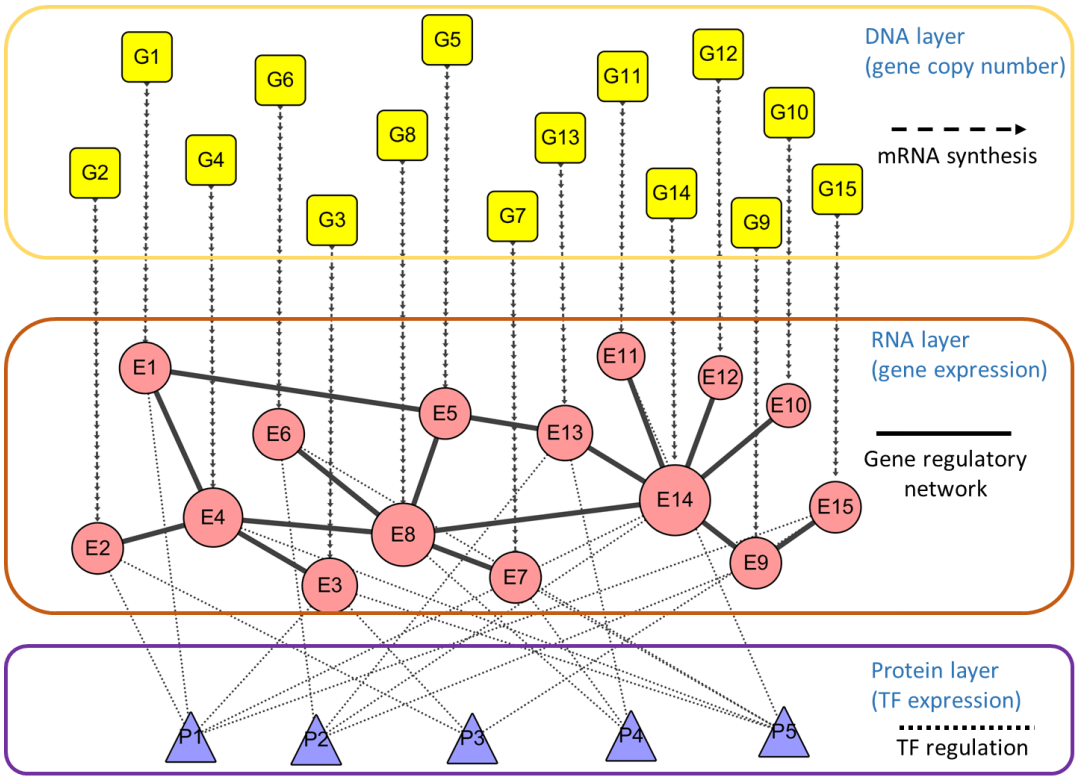


Figure 32- Multi-layer data signaling model for multiDDN

The multi-layer model could be summarized to dependency networks of gene entities. Gene entity in a multiDDN network is defined as one gene's expression combined with its own copy number regulator and with its regulating TFs. The differential dependency (i.e., the network rewiring) in the multiDDN network is in two tiers: the first is the intra-omics network rewiring between the gene entities; and the second is the inter-omics network rewiring within a gene entity. Figure 33 illustrates the conceptional multiDDN network between gene entities.

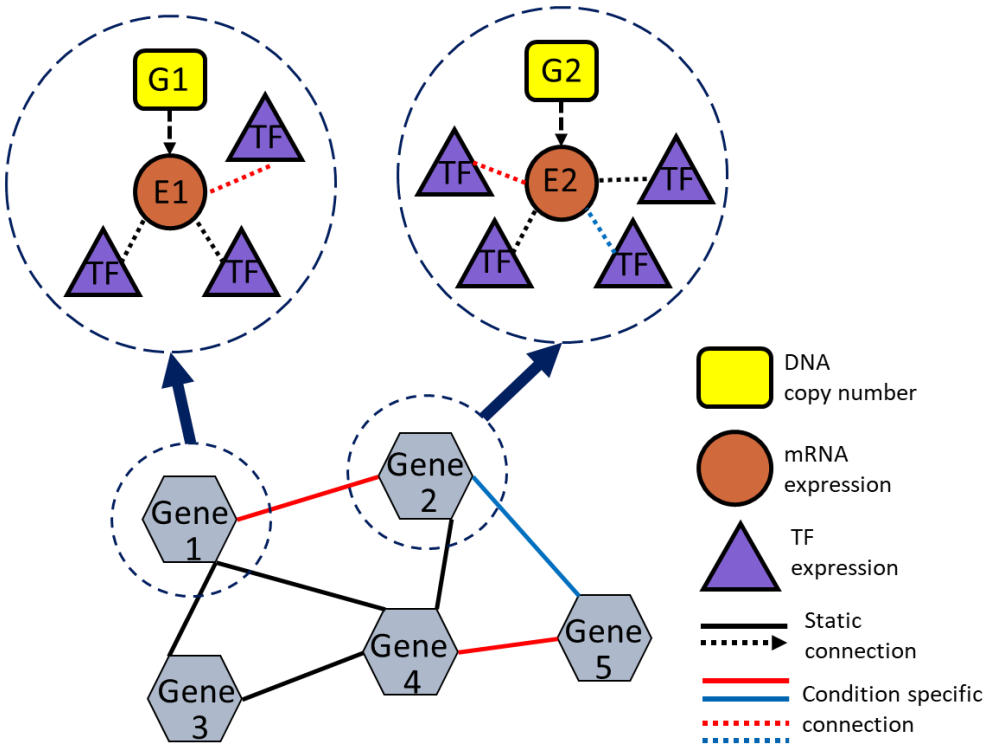


Figure 33- Graphical model of gene entities for multiDDN

There are various ways of selecting candidate nodes from the whole set of molecules in omics data. For multiDDN, we may choose genes from pathways of interest or curated gene list; or if pathway information is not available, we significantly differentially expressed genes from transcriptomics data.

We select the TF nodes for the integration model in the following way: The TF-gene binding information is retrieved from the TRRUST database(Han, et al., 2017), restricted to the Homo Sapiens and with experimental evidence. The genes in the RNA layer are used as input to find their regulating TFs. The retrieved list of TFs is further filtered to keep only those regulate at least three genes in the RNA layer and P-value less than 1E-3.

6.4. Problem formulation of multiDDN

Now consider the problem of learning graphical structure changes in the data model between two conditions. The problem is equivalent to estimate the conditional dependence or independence between a subset of random variables as gene entities under two conditions, with additional variables to each gene entity as entity-specific predictors. We have a set of $p_G = p_E = p$ genes of interest which are bind with a total of p_P TF proteins. We observed samples from n_1 objects under condition 1, and n_2 objects under condition 2. For each object, we collected three variables of copy number, gene expression and TF expression.

For convenience, we firstly define a few terms. Define the vectors of variables observed from the *i*-th sample under condition 1 as:

$$\begin{cases} \mathbf{x}^{(1)}_{G,i,\bullet} = \left[x^{(1)}_{G,i,1}, x^{(1)}_{G,i,2}, \cdots, x^{(1)}_{G,i,p} \right] \\ \mathbf{x}^{(1)}_{E,i,\bullet} = \left[x^{(1)}_{E,i,1}, x^{(1)}_{E,i,2}, \cdots, x^{(1)}_{E,i,p} \right], i \in \{1, \cdots, n_1\} \\ \mathbf{x}^{(1)}_{P,i,\bullet} = \left[x^{(1)}_{P,i,1}, x^{(1)}_{P,i,2}, \cdots, x^{(1)}_{P,i,p_P} \right] \end{cases}$$

in which The letter G, E, P represents data from genomics data, gene expression data and protein data, respectively. The vectors of variables under condition 2 are defined in a similar manner.

In the dimension of features, denote all observation on *j*-th gene or *j'*-th TF protein under condition 1 as:

$$\begin{cases} \mathbf{x}^{(1)}_{G,\bullet,j} = \left[x^{(1)}_{G,1,j}, x^{(1)}_{G,2,j}, \cdots, x^{(1)}_{G,n_1,j} \right]^T \\ \mathbf{x}^{(1)}_{E,\bullet,j} = \left[x^{(1)}_{E,1,j}, x^{(1)}_{E,2,j}, \cdots, x^{(1)}_{E,n_1,j} \right]^T \end{cases}, j \in \{1, \cdots, p\}$$

$$\mathbf{x}^{(1)}_{P,\bullet,j'} = \left[x^{(1)}_{P,1,j'}, x^{(1)}_{P,2,j'}, \cdots, x^{(1)}_{P,n_1,j'} \right]^T, j' \in \{1, \cdots, p_P\}$$

Similarly denote the observation vectors under condition 2. The vectors of variables are merged to data matrices of omics, either from sample dimension or feature dimension.

Define:

$$\mathbf{X}_G^{(1)} = \begin{bmatrix} \mathbf{x}_{G,1,\bullet}^{(1)} \\ \vdots \\ \mathbf{x}_{G,n_1,\bullet}^{(1)} \end{bmatrix} = \left[\mathbf{x}_{G,\bullet,1}^{(1)}, \mathbf{x}_{G,\bullet,2}^{(1)}, \cdots, \mathbf{x}_{G,\bullet,p}^{(1)} \right] = \left[x_{G,i,j}^{(1)} \right]_{n_1 \times p},$$

$$\mathbf{X}_E^{(1)} = \begin{bmatrix} \mathbf{x}_{E,1,\bullet}^{(1)} \\ \vdots \\ \mathbf{x}_{E,n_1,\bullet}^{(1)} \end{bmatrix} = \left[\mathbf{x}_{E,\bullet,1}^{(1)}, \mathbf{x}_{E,\bullet,2}^{(1)}, \cdots, \mathbf{x}_{E,\bullet,p}^{(1)} \right] = \left[x_{E,i,j}^{(1)} \right]_{n_1 \times p},$$

$$\mathbf{X}_P^{(1)} = \begin{bmatrix} \mathbf{x}_{P,1,\bullet}^{(1)} \\ \vdots \\ \mathbf{x}_{P,n_1,\bullet}^{(1)} \end{bmatrix} = \left[\mathbf{x}_{E,\bullet,1}^{(1)}, \mathbf{x}_{E,\bullet,2}^{(1)}, \cdots, \mathbf{x}_{E,\bullet,p_P}^{(1)} \right] = \left[x_{P,i,j'}^{(1)} \right]_{n_1 \times p_P},$$

as the three data matrices of three omics under condition 1, and similarly for condition 2. $\mathrm{X}_{\mathrm{G}}^{(c)}, \mathrm{X}_{\mathrm{E}}^{(c)}, \mathrm{X}_{\mathrm{P}}^{(c)}, c \in \{1,2\}$ are data matrices of gene copy number, mRNA expression and protein expression, respectively. Denote $\mathrm{X}^{(c)} = \left[\mathrm{X}_{\mathrm{G}}^{(c)}, \mathrm{X}_{\mathrm{E}}^{(c)}, \mathrm{X}_{\mathrm{P}}^{(c)} \right], c \in \{1,2\}$ as the entire observation data matrix under conditions 1 or 2.

For the *j*-th gene entity, omics under condition 1 or 2, define the coefficient β vector which is combined from three sub β vectors from each omics as:

$$\beta_{\mathrm{X},\bullet,j}^{(c)} = \begin{bmatrix} \beta_{G,\bullet,j}^{(c)} \\ \beta_{E,\bullet,j}^{(c)} \\ \beta_{P,\bullet,j}^{(c)} \end{bmatrix} = \begin{bmatrix} \left[\beta_{\mathrm{G},1,j}^{(c)}, \beta_{\mathrm{G},2,j}^{(c)} \cdots, \beta_{\mathrm{G},p,j}^{(c)} \right]^T \\ \left[\beta_{\mathrm{E},1,j}^{(c)}, \beta_{\mathrm{E},2,j}^{(c)} \cdots, \beta_{\mathrm{E},p,j}^{(c)} \right]^T \\ \left[\beta_{\mathrm{P},1,j}^{(c)}, \beta_{\mathrm{P},2,j}^{(c)} \cdots, \beta_{\mathrm{P},p_P,j}^{(c)} \right]^T \end{bmatrix}, c \in \{1,2\}$$

The β vector under condition 1 or 2 for all gene entities are merged into the β matrix which is the representation of the multiDDN network structure:

$$\mathrm{B}_{\mathrm{X}}^{(c)} = \left[\beta_{\mathrm{X},\bullet,1}^{(c)}, \beta_{\mathrm{X},\bullet,2}^{(c)}, \cdots \beta_{\mathrm{X},\bullet,p}^{(c)}\right] = \begin{bmatrix} \mathrm{B}_{G}^{(c)} \\ \mathrm{B}_{E}^{(c)} \\ \mathrm{B}_{P}^{(c)} \end{bmatrix},$$

Along the sample feature dimension, the β vector for the *j*-th gene entity under both conditions form the network dependency structure for the *j*-th gene entity, define as:

$$\beta_{\mathrm{X},\bullet,j} = \begin{bmatrix} \beta_{\mathrm{X},\bullet,j}^{(1)} \\ \beta_{\mathrm{X},\bullet,j}^{(2)} \end{bmatrix}$$

Finally, define two LASSO objective function for each condition and the multiDDN's objective function as:

$$f_c\left(\beta_{\mathrm{X},\bullet,j}^{(c)}\right) \triangleq \frac{1}{2n_c}\left\|\mathrm{x}_{\mathrm{E},\bullet,j}^{(c)} - \mathrm{X}^{(c)}\beta_{\mathrm{X},\bullet,j}^{(c)}\right\|_2^2 + \lambda_1\left|\beta_{\mathrm{E},\bullet,j}^{(c)}\right| + \lambda_3\left|\beta_{\mathrm{P},\bullet,j}^{(c)}\right|, c \in \{1,2\}$$

$$g\left(\beta_{\mathrm{X},\bullet,j}\right) \triangleq \lambda_2\left|\beta_{\mathrm{E},\bullet,j}^{(1)} - \beta_{\mathrm{E},\bullet,j}^{(2)}\right| + \lambda_4\left|\beta_{\mathrm{P},\bullet,j}^{(1)} - \beta_{\mathrm{P},\bullet,j}^{(2)}\right|$$

$$f\left(\beta_{\mathrm{X},\bullet,j}\right) \triangleq f_1\left(\beta_{\mathrm{X},\bullet,j}^{(1)}\right) + f_2\left(\beta_{\mathrm{X},\bullet,j}^{(2)}\right) + g\left(\beta_{\mathrm{X},\bullet,j}\right)$$

Mathematically, we formulate a multiDDN problem of learning structural changes of the multi-omics data model between two conditions as a convex optimization problem. We solve the following optimization problem for the *j*-th gene entity as follows (*j* = 1, 2, ..., *p*):

$$\beta_{X,\bullet,j} = \arg\min_{\beta_{X,\bullet,j}} f\left(\beta_{X,\bullet,j}\right)$$

$$== \arg\min_{\beta_{X,\bullet,j}} \left\{ f_1\left(\beta^{(1)}_{X,\bullet,j}\right) + f_2\left(\beta^{(2)}_{X,\bullet,j}\right) + g\left(\beta_{X,\bullet,j}\right)\right\}$$

$$= \arg\min_{\beta_{X,\bullet,j}} \{ \frac{1}{2n_1}\left\| x^{(1)}_{E,\bullet,j} - X^{(1)}\beta^{(1)}_{X,\bullet,j}\right\|_2^2 + \frac{1}{2n_2}\left\| x^{(2)}_{E,\bullet,j} - X^{(c)}\beta^{(2)}_{X,\bullet,j}\right\|_2^2$$

$$+ \lambda_1\left(\left|\beta^{(1)}_{E,\bullet,j}\right| + \left|\beta^{(2)}_{E,\bullet,j}\right|\right) + \lambda_2\left|\beta^{(1)}_{E,\bullet,j} - \beta^{(2)}_{E,\bullet,j}\right| + \lambda_3\left(\left|\beta^{(1)}_{P,\bullet,j}\right| + \left|\beta^{(2)}_{P,\bullet,j}\right|\right) + \lambda_4\left|\beta^{(1)}_{P,\bullet,j} - \beta^{(2)}_{P,\bullet,j}\right|\}$$

$$s.t.\ \ \beta^{(1)}_{G,i,j} = \beta^{(2)}_{G,i,j} = 0, i \in \{1,2,\cdots p\} \text{ and } i \neq j$$

$$\beta^{(1)}_{E,j,j} = \beta^{(2)}_{E,j,j} = 0$$

$$\beta^{(1)}_{P,l,j} = \beta^{(2)}_{P,l,j} = 0, \text{if no binding between } l\text{-th TF and } j\text{-th Gene}$$

In the multiDDN optimization problem, we learn the graphical structures of the integrated data model under two conditions jointly. The LASSO objective function $f_1\left(\beta^{(1)}_{X,\bullet,j}\right)$ and $f_2\left(\beta^{(2)}_{X,\bullet,j}\right)$ for each condition lead to the identification of a sparse graph structure. The penalty term $g\left(\beta_{X,\bullet,j}\right)$, encourages sparse changes in the network structure of both intra-omics and inter-omics interactions between two conditions, and thereby suppresses parametric inconsistencies due to noise or limited samples.

After solving the multiDDN optimization problem for each gene entity, the matrix of $B^{(c)}_X = \left[\beta^{(c)}_{X,\bullet,1}, \beta^{(c)}_{X,\bullet,2}, \cdots \beta^{(c)}_{X,\bullet,p}\right]$ are the parametric representation of the multiDDN network under each condition. For $\beta^{(c)}_{E,i,j}$ and $\beta^{(c)}_{E,j,i}$, we may replace them with the one with the larger absolute value to get a symmetric parametric structure for the intra-omics part ($B^{(c)}_E$) which could be converted to adjacency matrix for nodes in the RNA layer. The two parametric representations $B^{(1)}_X$ and $B^{(2)}_X$ are then compared to exact the network rewiring events from the differential matrix $\Delta B = B^{(1)}_X - B^{(2)}_X$. We may further separate the differential matrix to ΔB_G ΔB_E and ΔB_P, to categorize the network rewiring events into intra-omics network rewirings and inter-omics network rewirings.

6.5. Solving multiDDN optimization problem

6.5.1. BCD algorithm for multiDDN

As we discussed in Section 5.3.1, the original DDN method on single omics data uses the block coordinate descent (BCD) algorithm to solve the DDN optimization problem. The BCD algorithm updates each coordinate block cyclically and is very fast in solving sparse linear regressions in LASSO families (Friedman, et al., 2008). Similarly, we could modify and adapt the BCD algorithm in a multi-omics case for fast solving multiDDN optimization problems. Rewrite the objective function of multiDDN as follows

$$f\left(\beta_{\mathrm{X},\bullet,j}\right)=\left[\frac{1}{2n_1}\left\|\mathrm{x}^{(1)}_{\mathrm{E},\bullet,j}-\mathrm{X}^{(1)}\beta^{(1)}_{\mathrm{X},\bullet,j}\right\|_2^2+\frac{1}{2n_2}\left\|\mathrm{x}^{(2)}_{\mathrm{E},\bullet,j}-\mathrm{X}^{(c)}\beta^{(2)}_{\mathrm{X},\bullet,j}\right\|_2^2\right]$$
$$+\sum_{l=1}^{p}\left(\lambda_1\left|\beta^{(1)}_{\mathrm{E},l,j}\right|+\lambda_1\left|\beta^{(2)}_{\mathrm{E},l,j}\right|+\lambda_2\left|\beta^{(1)}_{\mathrm{E},l,j}-\beta^{(2)}_{\mathrm{E},l,j}\right|\right)+\sum_{k=1}^{p_P}\left(\lambda_3\left|\beta^{(1)}_{\mathrm{P},k,j}\right|+\lambda_3\left|\beta^{(2)}_{\mathrm{P},k,j}\right|+\lambda_4\left|\beta^{(1)}_{\mathrm{P},k,j}-\beta^{(2)}_{\mathrm{P},k,j}\right|\right)$$

The function is convex and continuous to $\beta^{(1)}_{\mathrm{G},i,j}$ and $\beta^{(2)}_{\mathrm{G},i,j}$, hence the block-wise minimum is reached when $\beta^{(c),r}_{G,k,j}=\rho^{(c),r}=\frac{1}{n_c}\mathbf{y}^{(c),r}_{j,-k}\bullet\mathbf{x}^{(c)}_k, c\in\{1,2\}$. For $\beta_{\mathrm{E},\bullet,j}$ and $\beta_{\mathrm{P},\bullet,j}$, the first part of L-2 loss functions in the formula is continuous and differentiable, and the penalty part is convex and block-wise separable for the coordinates pair $\left(\beta^{(1)}_{\mathrm{E},i,j},\beta^{(2)}_{\mathrm{E},i,j}\right)$ or $\left(\beta^{(1)}_{\mathrm{P},i,j},\beta^{(2)}_{\mathrm{P},i,j}\right)$. Therefore, as we discussed in Chapter 5, the global minimum could be achieved by iteratively updating with the block-wise minimum, and the convergence is guaranteed(Tseng, 2001). In the same section of Chapter 5, we have already detailly discussed how we determine the solutions of $\left(\beta^{(1)}_{\mathrm{E},i,j},\beta^{(2)}_{\mathrm{E},i,j}\right)$ in subregions on the plane of $\left(\rho_1,\rho_2\right)$, for each coordinate block' iteration in the BCD algorithm. Noticing the fact that the symbol sets of $\left(\beta^{(1)}_{\mathrm{E},i,j},\beta^{(2)}_{\mathrm{E},i,j},\lambda_1,\lambda_3\right)$ and $\left(\beta^{(1)}_{\mathrm{P},i,j},\beta^{(2)}_{\mathrm{P},i,j},\lambda_2,\lambda_4\right)$ are exchangeable in the formula

of multiDDN's objective function, we can get the solutions of $\left(\beta_{\mathrm{P},i,j}^{(1)}, \beta_{\mathrm{P},i,j}^{(2)}\right)$ in a similar manner by simply changing the corresponding symbols in the solutions. For *r*-th iteration, denote $\boldsymbol{\beta}_k^r \triangleq \left(\beta_{E,k,j}^{(1),r}, \beta_{E,k,j}^{(2),r}\right)$ and $\Lambda = \left(\Lambda_1, \Lambda_2\right) \triangleq \left(\lambda_1, \lambda_2\right)$ if updating *k*-th element in $\boldsymbol{\beta}_{E,\bullet,j}$; or denote $\boldsymbol{\beta}_{k'}^r \triangleq \left(\beta_{P,k',j}^{(1),r}, \beta_{P,k',j}^{(2),r}\right)$ and $\Lambda = \left(\Lambda_1, \Lambda_2\right) \triangleq \left(\lambda_3, \lambda_4\right)$ if updating *k'*-th element in $\boldsymbol{\beta}_{P,\bullet,j}$. The solutions are:

$$
\begin{cases}
\boldsymbol{\beta}_k^r = \left(\rho^{(1),r} - \Lambda_1 - \Lambda_2, \rho^{(2),r} - \Lambda_1 + \Lambda_2\right), & \textbf{for } \rho^{(1),r} \geq \rho^{(2),r} + 2\Lambda_2, \rho^{(2),r} \geq \Lambda_1 - \Lambda_2 \\
\boldsymbol{\beta}_k^r = \left(\rho^{(1),r} + \Lambda_1 + \Lambda_2, \rho^{(2),r} + \Lambda_1 - \Lambda_2\right), & \textbf{for } \rho^{(1),r} \leq \rho^{(2),r} - 2\Lambda_2, \rho^{(2),r} \leq -\left(\Lambda_1 - \Lambda_2\right) \\
\boldsymbol{\beta}_k^r = \left(\rho^{(1),r} - \Lambda_1 + \Lambda_2, \rho^{(2),r} - \Lambda_1 - \Lambda_2\right), & \textbf{for } \rho^{(1),r} \geq \Lambda_1 - \Lambda_2, \rho^{(2),r} \geq \rho^{(1),r} + 2\Lambda_2 \\
\boldsymbol{\beta}_k^r = \left(\rho^{(1),r} + \Lambda_1 - \Lambda_2, \rho^{(2),r} + \Lambda_1 + \Lambda_2\right), & \textbf{for } \rho^{(1),r} \leq -\left(\Lambda_1 - \Lambda_2\right), \rho^{(2),r} \leq \rho^{(1),r} - 2\Lambda_2 \\
\boldsymbol{\beta}_k^r = \left(\dfrac{1}{2}\left(\rho^{(1),r} + \rho^{(2),r}\right) - \Lambda_1, \dfrac{1}{2}\left(\rho^{(1),r} + \rho^{(2),r}\right) - \Lambda_1\right), & \textbf{for } \rho^{(1),r} < \rho^{(2),r} + 2\Lambda_2,\ \rho^{(2),r} < \rho^{(1),r} + 2\Lambda_2, \rho^{(2),r} > -\rho^{(1),r} + 2\Lambda_1 \\
\boldsymbol{\beta}_k^r = \left(\dfrac{1}{2}\left(\rho^{(1),r} + \rho^{(2),r}\right) + \Lambda_1, \dfrac{1}{2}\left(\rho^{(1),r} + \rho^{(2),r}\right) + \Lambda_1\right), & \textbf{for } \rho^{(1),r} > \rho^{(2),r} - 2\Lambda_2,\ \rho^{(2),r} > \rho^{(1),r} - 2\Lambda_2, \rho^{(2),r} < -\rho^{(1),r} - 2\Lambda_1 \\
\boldsymbol{\beta}_k^r = \left(0, \rho^{(2),r} - \Lambda_1 - \Lambda_2\right), & \textbf{for } \rho^{(1),r} < \Lambda_1 - \Lambda_2, \rho^{(1),r} > -\Lambda_1 - \Lambda_2, \rho^{(2),r} > \Lambda_1 + \Lambda_2 \\
\boldsymbol{\beta}_k^r = \left(0, \rho^{(2),r} + \Lambda_1 + \Lambda_2\right), & \textbf{for } \rho^{(1),r} > -\Lambda_1 + \Lambda_2, \rho^{(1),r} < \Lambda_1 + \Lambda_2, \rho^{(2),r} < -\Lambda_1 - \Lambda_2 \\
\boldsymbol{\beta}_k^r = \left(\rho^{(1),r} - \Lambda_1 - \Lambda_2, 0\right), & \textbf{for } \rho^{(1),r} > \Lambda_1 + \Lambda_2, \rho^{(2),r} > -\Lambda_1 - \Lambda_2, \rho^{(2),r} < \Lambda_1 - \Lambda_2 \\
\boldsymbol{\beta}_k^r = \left(\rho^{(1),r} + \Lambda_1 + \Lambda_2, 0\right), & \textbf{for } \rho^{(1),r} < -\Lambda_1 - \Lambda_2, \rho^{(2),r} < \Lambda_1 + \Lambda_2, \rho^{(2),r} > -\Lambda_1 + \Lambda_2 \\
\boldsymbol{\beta}_k^r = \left(\rho^{(1),r} - \Lambda_1 - \Lambda_2, \rho^{(2),r} + \Lambda_1 + \Lambda_2\right), & \textbf{for } \rho^{(1),r} \geq \Lambda_1 + \Lambda_2, \rho^{(2),r} \leq -\Lambda_1 - \Lambda_2 \\
\boldsymbol{\beta}_k^r = \left(\rho^{(1),r} + \Lambda_1 + \Lambda_2, \rho^{(2),r} - \Lambda_1 - \Lambda_2\right), & \textbf{for } \rho^{(1),r} \leq -\Lambda_1 - \Lambda_2, \rho^{(2),r} \geq \Lambda_1 + \Lambda_2 \\
\boldsymbol{\beta}_k^r = (0,0), & \textbf{for others}
\end{cases}
$$

The accelerating strategies we discussed in Chapter 5 are still applicable to the BCD algorithm for multiDDN. The algorithm is described as follows:

```
%% BCD algorithm for multiDDN
Initialize Beta matrices B1 and B2
Loop1: loop through the 1st to p-th gene entity
    Initialize residuals Y1 and Y2
    Loop2: loop until convergence of B1 and B2
        Determine the next coordinate block to update by cyclic rule
        Update rho1 and rho2 with CorrMtx or ResiUpd method
        Get the solutions from the subregions in which (rho1, rho2) location
        Update B1 and B2
    EndLoop2
EndLoop1
Return Beta matrices
```

6.5.2. Determining parameters

If there is no proteomics data available for multiDDN analysis, the corresponding parameters lambda3 and lambda4 are simply set to zero; if similar sparsity level for transcriptomics network and proteomics network is assumed, we may set lambda3 equals to lambda1, and lambda4 equals to lambda2; otherwise, we may use two-dimension grid searching on lamda1 and lamda3 to minimize cross-validation errors.

For the gene regulatory network in the transcriptomics layer, we use a cross-validation strategy to choose lambda1 firstly and then determine the value of lambda2. If we temporarily set lambda2 as zero, the differential network analysis is simplified to standard GGM network inference, and the neighborhood selection for each node in multiDDN is equivalent to a standard

LASSO regression, and hence we may use the same cross-validation strategy as discussed in (Friedman, et al., 2017) to choose lambda1 that minimizes the cross-validation error. One standard error rule (Friedman, et al., 2017) could be used as the rule of thumb to increase the robustness.

Secondly, we will determine the value of λ_2 for a given significance level. From the solution subregions in the plane of (ρ_1, ρ_2) we can see that $\beta_k^{(1),r}$ and $\beta_k^{(2),r}$ will be identical in the subregion of $|\rho_1 - \rho_2| < 2\lambda_2$. Therefore, the question becomes what value of $|\rho_1 - \rho_2| < 2\lambda_2$ is considered significantly large enough, at a given significance level. Since X1 and X2 are following multivariate Gaussian distribution, we could apply Fisher's transform to ρ_1 and ρ_2 as correlation coefficients for standardized. Define:

$$z_1 = \frac{1}{2}\ln\frac{1+\rho_1}{1-\rho_1}, z_2 = \frac{1}{2}\ln\frac{1+\rho_2}{1-\rho_2}$$

z_1 and z_2, as the results of Fisher's transform, well approximately follows the Gaussian distributions of $G\left(\frac{1}{2}\ln\frac{1+\bar{\rho}_1}{1-\bar{\rho}_1}, \frac{1}{n_1-3}\right)$ and $G\left(\frac{1}{2}\ln\frac{1+\bar{\rho}_2}{1-\bar{\rho}_2}, \frac{1}{n_2-3}\right)$. Since ρ_1 and ρ_2 are equal in the null hypothesis of no differential edges, z_1 and z_2 share the same mean. If independence is assumed between ρ_1 and ρ_2, we could infer that the variable $z = z_1 - z_2$ follows a Gaussian distribution with zero-mean and variance equal to the sum of two Gaussian random variable's variances: $1/(n_1-3)+1/(n_2-3)$. For a given significance level α (for example, 0.05), define the significance threshold for $|z| = |z_1 - z_2|$ as:

$$s(\alpha) = \sqrt{\frac{1}{n_1-3}+\frac{1}{n_2-3}}\Phi^{-1}\left(1-\frac{\alpha}{2}\right) < |z| = |z_1 - z_2|$$

Since group imbalance is not considered in original DDN's derivation, its significance threshold is calculated on a special case of $n_1 = n_2 = N$. In that case, $s(\alpha) = \frac{2}{\sqrt{N-3}} \Phi^{-1}\left(1-\frac{\alpha}{2}\right)$

Rewrite the significance condition in the form of ρ_1 and ρ_2, we have:

$$|z| = |z_1 - z_2| > s(\alpha) \Leftrightarrow |\rho_1 - \rho_2| > \frac{e^{2s(\alpha)}-1}{e^{2s(\alpha)}+1}\left(1-\rho_1\rho_2\right) = 2\lambda_2$$

Since λ_2 is applied for all nodes' optimization, we replace $\rho_1\rho_2$ with the sample mean values estimated from all the samples $\rho_1\rho_2 \leftarrow \overline{\rho_1\rho_2} = 2\sum_{1\le i<j\le p} R_{ij}^{(1)} R_{ij}^{(2)} / \, p(p-1)$.

And finally, we get the value of λ_2 under significance level α:

$$\lambda_2 = \frac{e^{2s(\alpha)}-1}{2\left(e^{2s(\alpha)}+1\right)}\left(1-\overline{\rho_1\rho_2}\right)$$

6.6. Simulation study

We design a simulation study to show that our proposed multiDDN method will have higher prediction precision in constructing differential networks from the multi-omic data, comparing with the DDN method on single omics data.

The ground truth of the multi-layer regulation network is designed as a scale-free network to mimic the gene regulatory network form real biological data, as shown in Figure 34. We then generate the adjacency network according to the ground truth. For the regulation strengths, we measure the intra-omics Pearson's correlation coefficient distribution from over 200 samples in TCGA ovarian cancer dataset(Cancer Genome Atlas Research, 2011; Zhang, et al., 2016), and take the mean and variance values as the guiding parameters for the simulated covariance matrix, giving

the fact that the correlation matrix is identical to the covariance matrix for standardized data. Similarly, we measure inter-omics Pearson's correlation coefficient distribution for interactions between DNA and mRNA, and interactions between TF with mRNA. Depending on the type, the non-zero elements in simulated covariance Σ are sampled from one of the three distributions, and the whole data matrix is then sampled from multivariate Gaussian distribution of $G(0,\Sigma)$. These steps are repeated once again with a slightly different network skeleton and covariance matrix, to generate the data matrix of condition 2. The network structure differences between conditions 1 and 2 are recorded as the ground truth of network rewiring, as shown in Figure 34 as colored edges.

We test the multiDDN method on four groups of simulated data: 1. Single omics data that contains only gene expression; 2. Two-omics data combined from copy number data and gene expression data; 3. Two-omics data combined from gene expression data and TF expression data; 4. Three-omics data combined from all three types of omics data. To compare the results on a fairground, only the common layer of four groups' networks which is the RNA layer is compared. The multiDDN performance in these four groups is shown in Figure 35- The ROC curves for multiDDN on multi-omics data with different integration levels. The black curve is for the multiDDN method with all three types of omics data as the input. The blue and red curves are for the multiDDN method with only two of the three types of omics data. The green curve is for the DDN method with single-omics data of mRNA expression.

. As expected, multiDDN on the integrated three omics data has the largest area under the ROC curve, and DDN on single omics data has the smallest. This confirms the benefit of integrating additional omics data into differential network analysis.

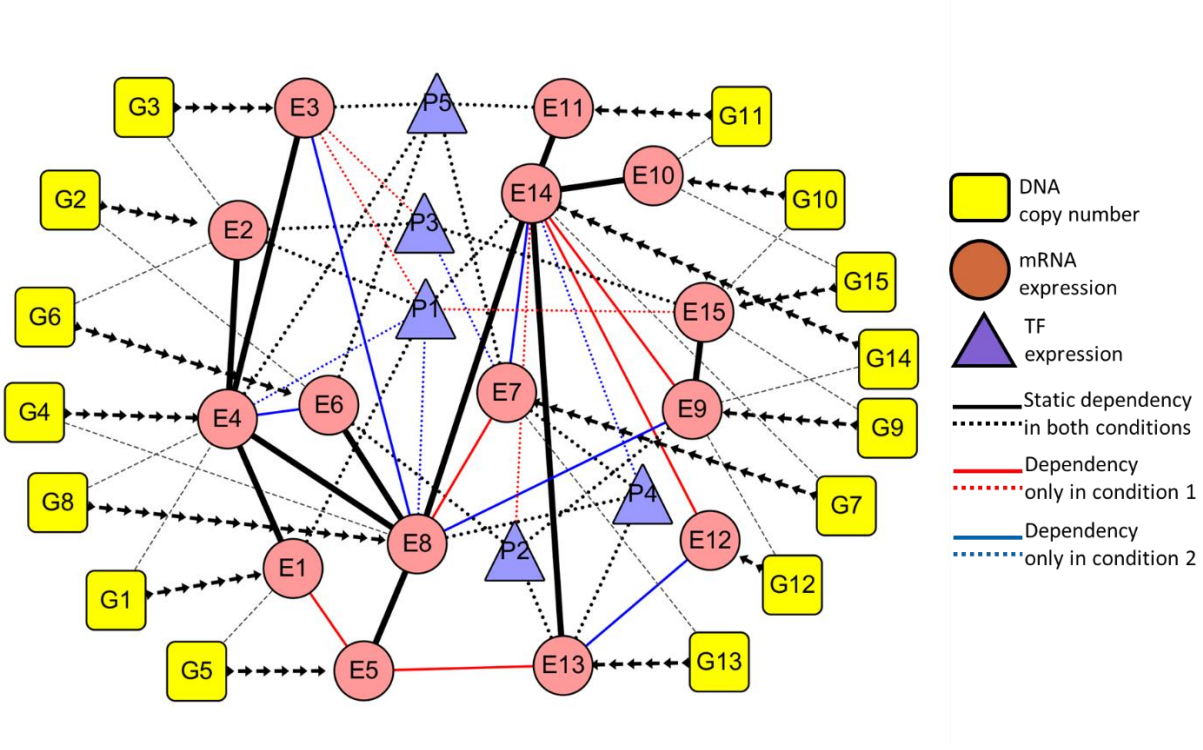


Figure 34- Synthesized multi-layer differential network used in multiDDN simulation

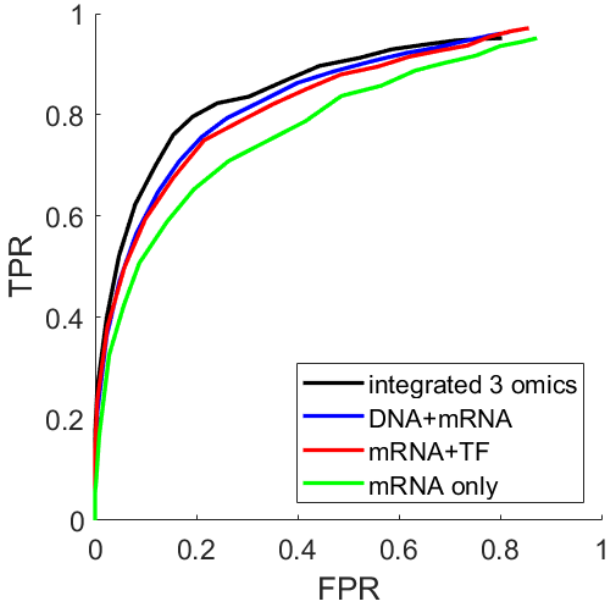


Figure 35- The ROC curves for multiDDN on multi-omics data with different integration levels. The black curve is for the multiDDN method with all three types of omics data as the input. The blue and red curves are for the multiDDN method with only two of the three types of omics data. The green curve is for the DDN method with single-omics data of mRNA expression.

6.7. Real data experiments

6.7.1. Gene regulatory network on CPTAC-OV dataset

We applied our proposed multiDDN method on real omics data to discover novel network rewiring and biomarkers. We use data from the CPTAC-OV2 prospective data set in this study. The projects aim to perform a comprehensive proteomics and genomics characterization of human ovarian high-grade serous carcinoma tumors. The global proteomics expression data are prepared by the PNNL lab. The digested proteins are tested by LC-MS/MS with 12 sets of TMT 10-plex labeling. The MS/MS data are then processed with the MS-PyCloud pipeline to identify and quantify proteins(Chen, et al., 2018). Among all 110 samples, we select 83 samples that are from ovarian cancer tumors in our study and exclude normal or control samples. The protein expression matrix is normalized by total expression quantity. Proteins with >10% missing values in the 83 tumor samples are removed, and the rest missing values imputed with the mean expressions. The RNA-seq data of the matching samples are from the NCI database, quantified into mRNA expression matrix with an in-house pipeline, 82 of 83 tumor samples have matching RNA-seq profiles. The DNA sequence data of the matching samples are processed to gene mutation call and copy number signals by Li Ding's lab at Washington University in St. Louis. Tumor samples with somatic mutations on the genes of BRCA1/BRCA2/PTEN are identified as homologous recombination deficiency (HRD), and the rest of tumor samples are grouped as none-HRD or HRD negative tumors. In the 82 tumor samples, 19 samples are marked as HRD+ and 63 samples are marked as HRD-.

The list of 171 HRD associated genes is from CPTAC-OV1 research. 120 genes are overlapped both with the gene list of copy number data and the gene list of mRNA expression data. These genes are further searched through TRRUST online database(Han, et al., 2017) to find the

top up-regulating TFs. 53 TFs are found with at least three regulated genes among the 120 genes and with p-value<0.005, FDR<0.01. In the proteomics expression data, we found 23 TFs are overlapped with the protein list.

The final data matrices contain DNA copy number data and mRNA expression data for 120 genes, and protein expression data for 23 up-regulating TFs. The total feature scale is P=263. The data are collected from 19 HRD+ ovarian tumor samples and 63 HRD- samples, with total sample scale N=82. The data matrices and TF-binding information are then sent to multiDDN to reconstruct the differential gene regulatory network. Due to the limitation of a small sample size of the HRD+ group, we perform 9-fold cross-validation to determine the multiDDN parameter. The minimum cross-validation error is achieved when $\lambda_1 = 0.069$, and we use the one-standard-error rule to select $\lambda_1 = 0.131$ as the chosen parameter, indicated by green broken lines in Figure 36. The rest parameters are set as default.

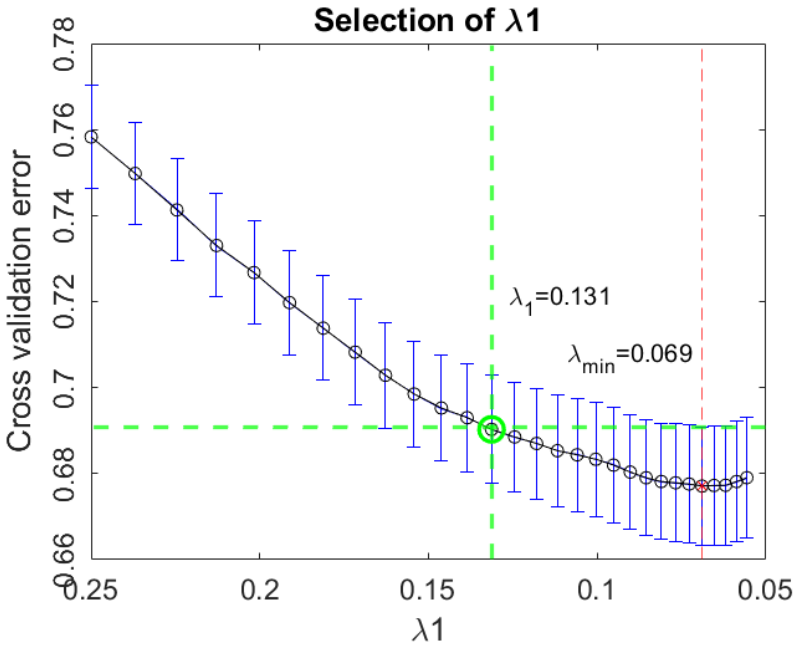


Figure 36- Parameter selection by cross-validation

Figure 37 shows the reconstructed differential network. The nodes in the DNA layer are not shown in order to give a better view of the entire network. Our method detects 632 static intra-omics connections between the 120 genes, and 107 differential intra-omics connections. For inter-omics links between TF nodes and RNA nodes, there are 68 static edges and 25 differential edges. Figure 38 shows only the network rewirings detected in the differential network. 11 gene and one TF is identified in this network as hubs with high connecting degrees. Among these hubs, gene RBBP4 is one of the five genes (HDAC1, RBBP4, RBBP7, EP300, HUS1) that reported involved in histone acetylation or deacetylation by CPTAC-OV1 project, and hub gene HDAC2 is known to be responsible to histone deacetylase.

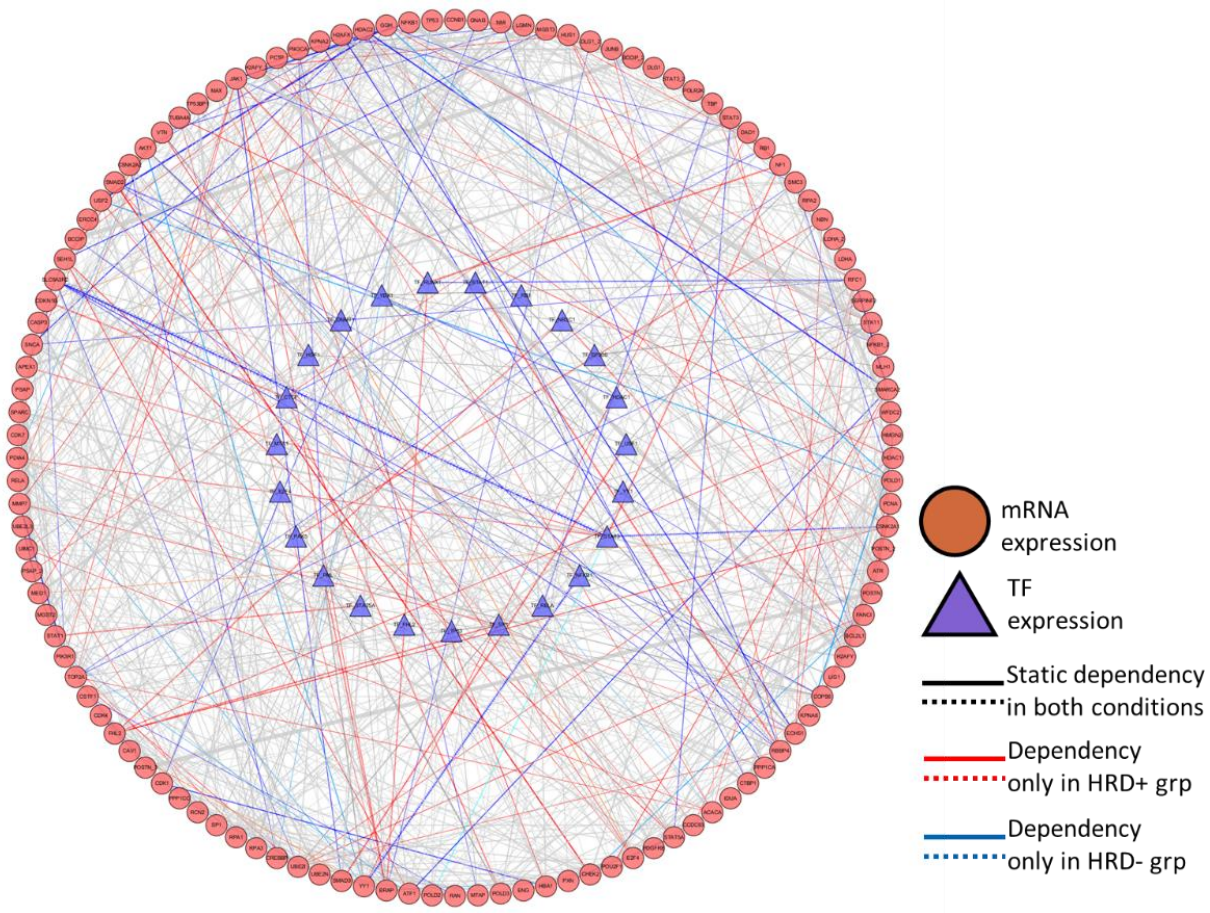


Figure 37- multiDDN constructed differential network on CTPAC-OV data

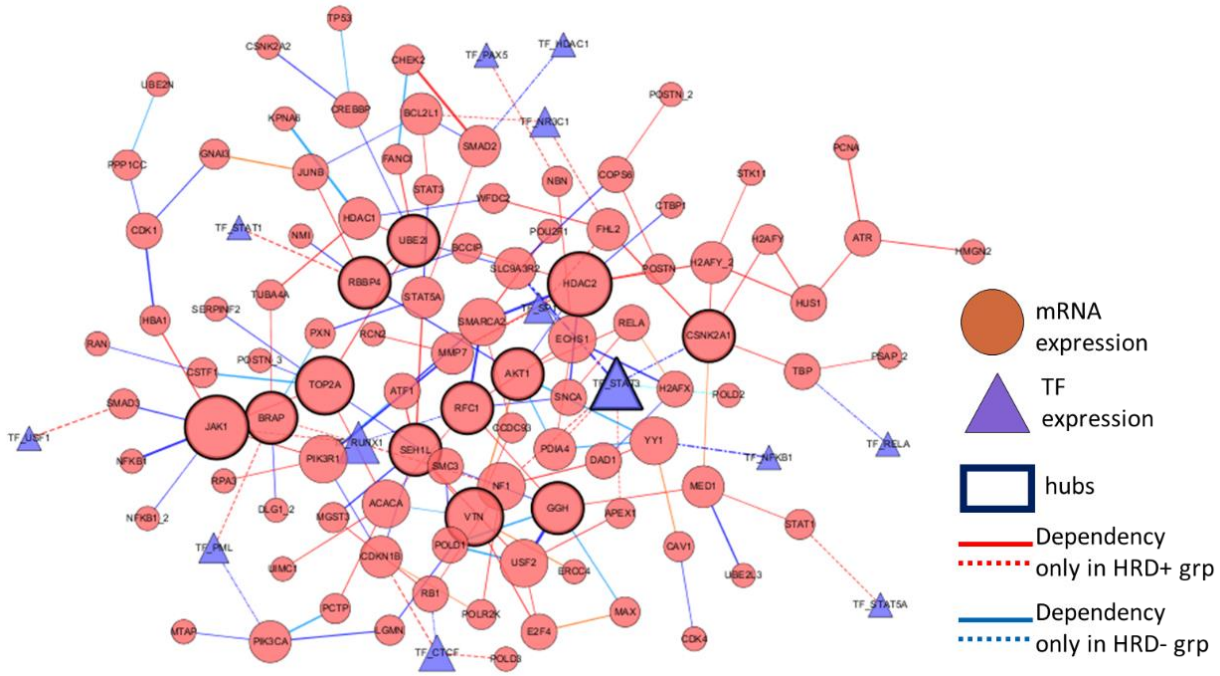


Figure 38- multiDDN detected network rewirings on CTPAC-OV data

6.7.2. Phosphorylation network on CPTAC-OV dataset

For the phosphorylation network study which is a part of CPTAC research, the ovarian tumor tissue samples were a subset of the TCGA high-grade serous ovarian carcinoma specimens. All the biospecimens were collected from newly diagnosed patients with HGSC. PTM signatures were profiled on HuProt™ arrays for 108 ovarian tumor samples. The tyrosine phosphorylation signals are tested on HuProt™ array using ovarian tumor lysates. For the observed Tyr phosphorylation, 54 kinases together with the corresponding 118 substrates are detected to form 245 active kinase-substrate interaction (KSR) in ovarian(Hu, et al., 2014).

We adapt the multiDDN method to reconstruct phosphorylation networks and detect the network rewirings. We specified the protein kinase as the regulators and replaced the protein layer in multiDDN with kinase expression, replaced the RNA layer with substrate expression as the regulating targets, and left the DNA layer empty. To detect network rewiring of KSR, we detect

only inter-omics rewirings between the regulators and their targets. The KSR is limited to those retrieved from the KSR database of PhosphoNetworks(Hu, et al., 2014).

multiDDN method detected multiple rewiring events in KSR between HRD and none-HRD groups on the LC-MS/MS quantitated expression of kinases and HuProt array-based activity quantitation of substrates. Figure 39 shows the multiDDN results of static and differential inter-omics network rewirings. Orange ovals represent the kinases; green hexagons represent their substrates with known kinases substrates relationships based on the KSR database; yellow and purple colored edges are the dependency edge only in HRD+ and HRD- groups respectively. Two kinases PTK2 and PTK2B linked with significant rewiring events were found dysregulated in ovarian cancer in cell line models.

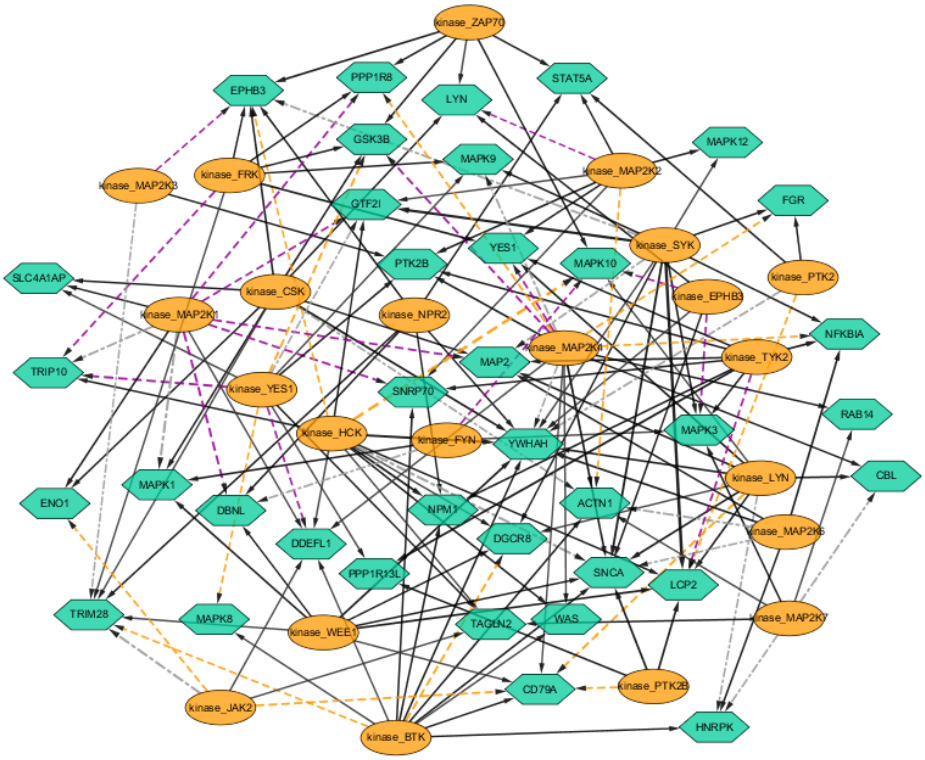


Figure 39- multiDDN detected network between protein kinases and substrates

Chapter 7. Biomarker discovery by hub detection in biological networks

7.1. Introduction

Biomarkers are indicators of the severity or presence of phenotypes like disease states(Klipp, et al., 2016). It could be anything measurable, for example, cell counts, mutation status, gene expression, protein abundance, metabolite level, etc. In the omics study, differential expression analysis (DEA) has long been used to find significant expression indicators between phenotypic states. The biomarkers found by DEA answer the question of what genes are differentially expressed between different conditions.

Differential network analysis provides a unique view of biological systems and answers the question of what genes are differentially connected between different conditions. For network analysis of omics data, especially networks inferred from transcriptomics or proteomics data, the dependencies between nodes in the reconstructed network are considered to reflecting the interactions between genes or proteins. Biomarkers carrying gene or protein interaction information could help to reveal the dysregulation mechanism of disease or phenotype of interest. Naturally, genes with a large number of network rewiring are good indicators of phenotypic changes, and hence are potential candidates of biomarkers. For example, ER regulated pathways behave very differently between subtypes of breast cancer and ovarian cancer. The researches on dysregulated genes in these pathways have developed ER targeting drugs that allow mutation-specific personal therapies.

7.2. Graphical characteristics of biological networks

In graph theory, we call an undirected network as a random graph or random network if its generation follows the Erdős–Rényi random graph model(Newman, et al., 2001). In brief, the Erdős–Rényi model requires the graph's every possible edge that occurs with fixed probability p, independently of the other edges. This model describes what the simplest type of random network when there is no prior knowledge of the property of the edges other than the independent probability of occurrence, and usually serves as a starting point for complex network analysis. The degree distribution of a random network of Erdős–Rényi model follows a binomial distribution. Most networks in the real world are not random networks and hence have very different degree distributions. If a network's degree distribution follows a power-low distribution, it is called a scale-free network. Biological networks, especially protein-protein interaction networks and gene regulatory networks, in many cases, could be considered as scale-free networks. A large number of real-world networks, including scale-free networks, have highly right-skewed distribution, in which a majority of nodes have low degrees while a few "hub" nodes have high degrees. In network science, we define a hub node as a node with the number of its connecting edges that significantly exceed the average. If a network is a random network whose edges are randomly connected between nodes, the hub nodes will not emerge; on the other hand, in scale-free networks, hub nodes are expected to emerge with a power-low distribution of $P(k) \sim k^{-\gamma}$.

The existence of hub nodes is one of the most significant differences between random networks and scale-free networks. If keeping the number of node and the number of edges in the network as constant, the degrees of hub nodes in a scale-free network is much higher than the largest node degrees in a random network. We show an illustrative in Figure 40.

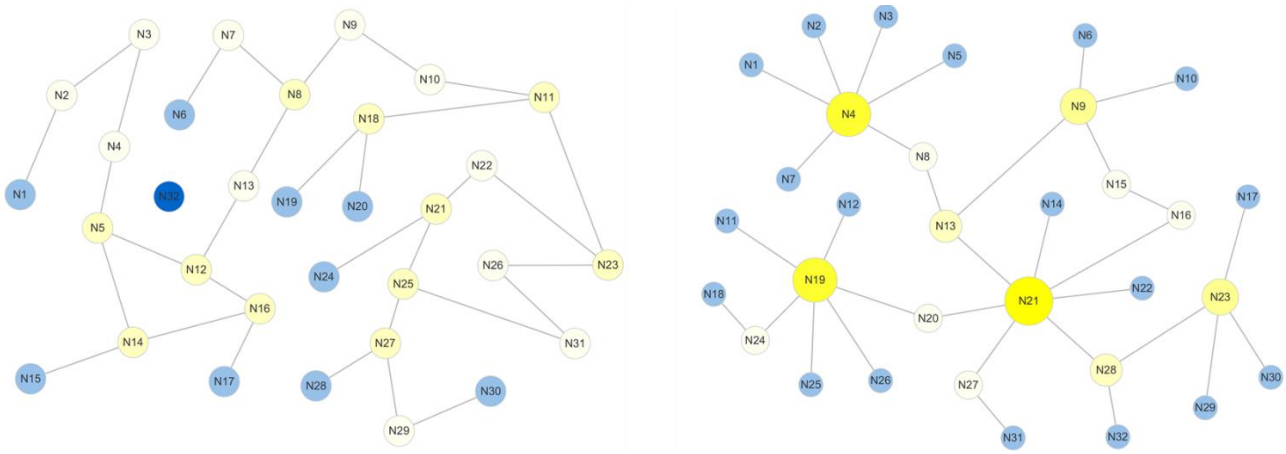


Figure 40-Examples of a random network and a scale-free network of the same scale

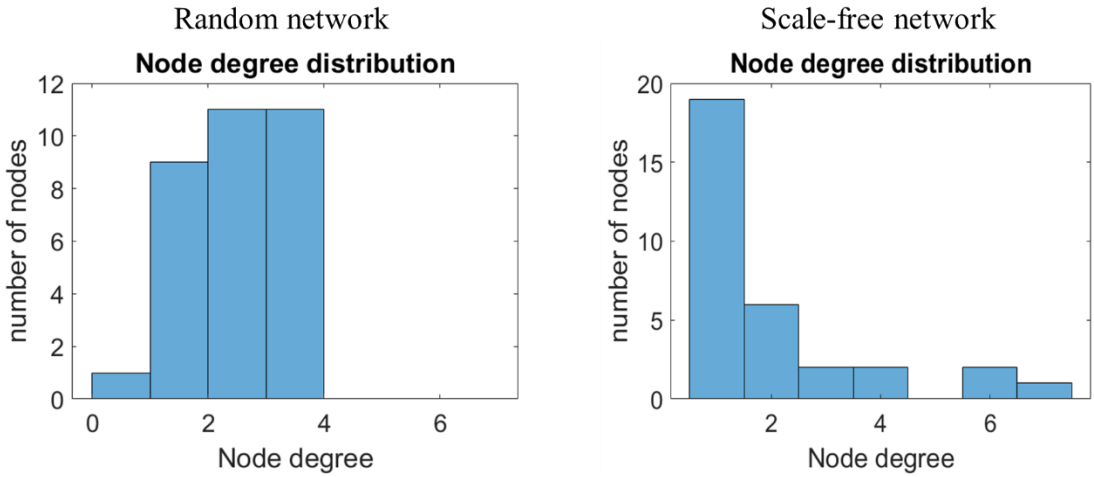


Figure 41-Distributions of node degrees in random network and scale-free network

Hub nodes have significant impacts on the topology of a network. Firstly, hub nodes are the pivotal nodes connecting sub-networks into a bigger connected network. If a hub gene lost its connecting to other genes in a gene pathway, the whole pathway as a network might change the topology from a single connected network to multiple isolated sub-networks, and as a result, the entire pathway loses its normal function. A large number of researches reported and validated such dysfunctional pathways causes by gene mutations. Some hub genes are so critical in these pathway dysfunctions that researchers recognize these gene mutations as hallmarks in cancer development. For example, the famous gene tumor protein p53, also known as the TP53 gene, in the P53 pathway and MAPK pathway.

Cells make decisions based on the integration of received cues. In multicellular organisms, decisions are made to benefit the whole, and the signaling networks help the cooperative behaviors between cells. Accumulation of multiple network rewiring connected to a single gene is more like to deviate the whole network's function and finally trigger the cell's abnormal status. Therefore, in a differential network, we focus on the hub nodes that link to a large number of network rewirings, and mark them as biomarkers that are potentially associated with the phenotype.

The hub nodes in scale-free networks such as biology networks play critical roles in linking the networking, preventing random attacks and shorten the distance between nodes. The existence of hub nodes in a network would increase the network robustness and attack tolerance to random failure of nodes. The hub nodes that function as critical components to maintain network connectivity are responsible for the exceptional robustness of the network. Since the hub nodes with vast connecting degrees coexist with a large number of small degree nodes in the network, the chance of a random node failure occurred on a hub node would be very small. Deleting a few small degree nodes does not have a considerable effect on the network's integrity, due to the remaining hub nodes would still keep the network mostly connected. In this way, hub nodes are the strength and key connectors of the network.

On the other hand, if a network was attacked on its hub nodes instead of random attack to all nodes, the integrity of the network will fall apart much faster than random networks. The living cells are facing numerous attacks from the environment, such as radiation, virus, chemicals that cause DNA mutation, etc. Some like ultraviolet radiation behave as random attacks and cause almost equal chances to DNA mutation to every gene; some like virus favors more on specific genes, could be more effectively change the network topologies to alter the functions of specific gene pathways.

For a connected network, the average distance is defined as the arithmetic average of the distances between any two nodes randomly drawn from the node set of the network. In practice, we could list all combinations of two-node pairs, calculate the distances between the two nodes in each pair, and take the mean of all calculated distances as the average distance of the given network.

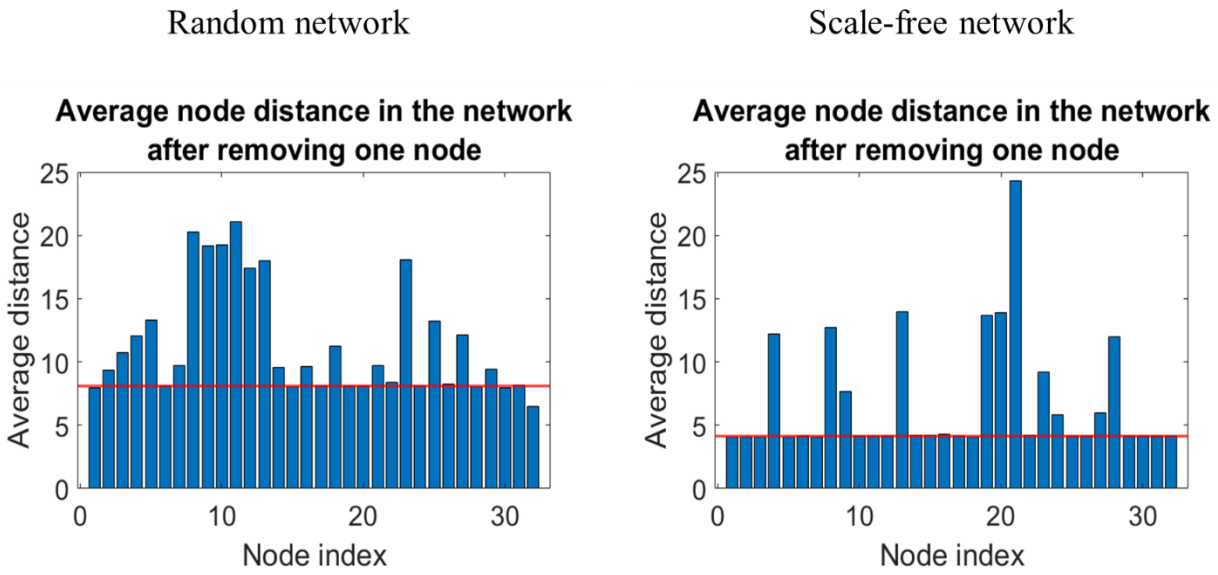


Figure 42- Average node distances after removing one node in the example random network and scale-free network

Extensively, we may define the average distance for any network. To avoid infinity for nodes that are not connected, we define such distance as a constant larger than the largest possible distance of a connected network. The number of nodes P is a suitable choice for this constant. We now calculate the average distance of the original network (red line in Figure 42) and compare it to the average distance of the network after removing one of its hub or non-hub nodes. The result in Figure 42 shows removing hub nodes from the network would significantly increase the average distance between nodes, comparing with removing non-hub nodes. It also confirms the property aforementioned: the robustness of the network to random attacks and the vulnerability to targeted attacks.

DNA repairing mechanism allows cells to repair damaged DNA regions and restore the functions of the genes on the damaged DNA regions. For scale-free gene regulation networks, the loss of function of a large number of genes/nodes with low connecting degrees could not substantially alter the overall functionality of the network. However, DNA damages on hub genes/nodes in the node are more likely to modify the network topology without the possibility of reversing the change and hence alter the overall functionality of the network. The mutation of oncogenes and tumor suppressor genes are considered as one of the hallmarks of cancer development.

7.3. Hub node detection in biological networks

We used the DDN method for analyzing gene pathways that include known oncogenes or tumor suppressors (e.g., TP53), on various types of cancers. The hub nodes are then detected on the DDN constructed gene regulation networks and differential networks. The results showed that the known oncogenes and tumor suppressors are more likely to be hub genes in the gene regulation network and more likely to reside significant network rewirings in the cells of the diseased group. For example, in the DDN detected network rewirings in the MAPK signaling pathway (Figure 26) for ovarian cancer tissue versus normal tissue, the RAS gene family, known to be related to cell growth and cell death, is classified in the pathway as a hub and also resides two significant rewirings.

As aforementioned, hub genes are the critical feature that distinguished scale-free networks from random networks. We proposed a method of detecting hub nodes in biology networks via statistical hypothesis testing. By identifying the top-ranked nodes with a significantly higher level of degrees, we have better confidence in claiming these as hub genes that differentiate the network from the null hypothesized random networks.

Consider a network under the null hypothesis of random networks, the distribution of its node degrees follows the binominal distribution(Bollobás and Béla, 2001) of

$$P(k)=\binom{n-1}{k}p^{k}(1-p)^{n-1-k}$$

in which k is the degree, p is the probability of connecting an edge.

Under the null hypothesis, the probability of a node with its node degree larger or equal to k_0 is:

$$P(K\geq k_0)=1-P(K<k_0)=1-\sum_{k=0}^{k_0-1}\binom{n-1}{k}p^{k}(1-p)^{n-1-k}$$

We proposed to detect the hub nodes in the network by a hypothesis testing method to identify the nodes that violate the null hypothesis of random networks. These nodes contribute the most to distinguish their belonging network from a random network. In other words, without these nodes the remaining network cannot be statistically separated from a random network. The proposed method works in two steps: in the estimation step, we estimate the parameter *p* which is the probability of connecting an edge from the given network; in the detection step, we identify the hub nodes that not fit the null distribution of binomial distribution with the parameter *p*.

For differential networks, we use a similar method of hypothesis testing. The null hypothesis assumes that the differential edges are randomly connected among all nodes. And the parameter *p* is calculated not from the static edges but from the differential edges. Correspondingly, the detected hubs are now nodes connect with significantly higher numbers of network rewiring.

We applied our proposed method on various differential networks detected by the DDN method. Some of the hubs detected have already been reported in Section 5.4. In the differential network detected from the GPAA proteomics data(Herrington, et al., 2018) on genes from the

LXR/RXR pathway, by using the proposed hub detection method we found two hub genes as shown in Figure 43, One of the two hubs, the gene APOE, is responsible for carrying cholesterol and other fat molecules through the bloodstream. It is reported as the most consistently identified genetic risk factor related to vascular diseases(Peila, et al., 2001).

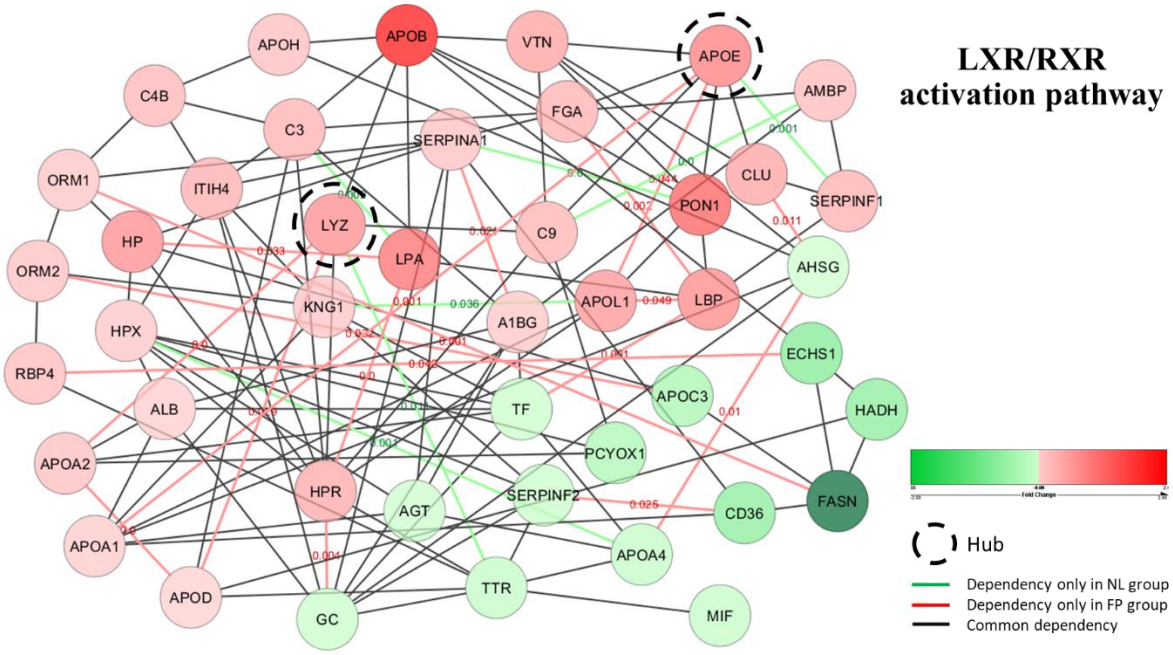


Figure 43- Hub nodes detected in differential networks of LXR/RXR pathway on GPAA proteomics data

Chapter 8. Contribution and Future work

8.1. Contribution

In this dissertation, we propose and develop novel methods for correcting normal cell contamination in tumor samples using DNA copy number data, inferring intra or inter omics differential dependency networks from single and multiple omics data, and detecting networked molecular biomarkers across different biological conditions. We demonstrate and assess the performance of these methods on both simulation and real omics data sets. We incorporate or develop several effective complexity-reduction strategies to accelerate the learning algorithms in case of big data, and adaptive normalization scheme to correct the potential bias due to imbalanced sample sizes. Experimental results show the expected performance and the potential to achieve the intended objectives, either technically or biologically. The specific contributions are summarized as follows:

- We propose and develop a novel computational approach, BACOM 2.0, that can accurately detect deletion types of DNA copy numbers in cancer cells and estimates normal cell fraction in bulk tumor samples. This method improves the detection of statistically significant somatic copy number aberrations. Specifically, we propose an allele-specific absolute normalization method and a systematic analytics pipeline to eliminate major confounding factors, evaluate normal cell fractions, and correct the DNA copy number signals.
- We found that the systematic bias in the original DDN method is caused by data imbalance through both the theoretical analysis and simulation studies. We proposed a reformulation to the DDN method by adding a sample scale normalizer. The test on simulated data proves that the bias brought by imbalanced data is corrected in the proposed reformulated DDN method.

- We improve the computational speed or efficiency of the DDN algorithm(s) by designing several accelerating strategies, including two reformulated calculation methods, one rule of discarding predictor variables, and one implementation of parallel computing.
- We applied the DDN method to various omics data, including transcriptomics and proteomics data. From the inferred differential dependency network, we detected network rewiring events associated with the phenotype of interest and discovered biomarkers with the potential for further analysis. Specifically, we identified a sub-network inferred from CPTAC-OV proteomics dataset which reveals a significant change in acetylation level of histone proteins; in the differential dependency network inferred from GPAA proteomics dataset, we identified a hub gene APOE which is implicated in vascular disease, and also detected a group of hub genes that significantly enriched in TCA pathway which controls energy activities in vessels; in transcriptomics data from the psychiatric disorder research, we detected similar network rewiring events associated to schizophrenia disease and bi-polar disease comparing to normal subjects
- We propose multiDDN for integrated DDN analysis using multi-omics data. By introducing directional restrictions on the regulation relationship between omics data, this method improves original differential network analysis in terms of both computational time and inference accuracy. The simulation study shows the accuracy advantage over the single-omics DDN method. We applied our proposed method on two large-scale multi-omics data and discovered potential cancer biomarkers.
- We propose a statistical approach of identifying hub genes in biological networks, based on the random graph theory. The detected hub genes could serve as biomarkers of concentrated network rewiring events in differential networks

8.2. Future work

Future work includes some further potential improvements or development of the methods proposed by this dissertation. For the BACOM 2.0 method, further software development to migrate the method from the MATLAB platform to an R or Python package can extend its users in the bioinformatics community. The adaptation to the next-generation DNA-sequencing data would definitely give it a new life, although it requires fundamental changes on both the signal model of next-generation sequencing data and the statistical model to differentiate copy number status.

For the DDN method, the proposed third accelerating strategy of integrating the Strong rule is currently limited to the LASSO optimization with a single parameter. Although this proposed strategy could help to accelerate the procedure of determining the first parameter used by DDN, it is not fully utilized in solving the DDN optimization. Extension of the Strong rule to the case of the two-parameter DDN optimization can further accelerate the computation of solving DDN optimization problems.

For multiDDN, we use cross-validation to select parameters at the current stage. An efficient way of selecting multiple optimal parameters can help to reduce the training time. Adapting the multi-response LASSO regression method (Simon, et al., 2013) into the current DDN framework is also a promising direction to construct differential networks between multiple groups.

For discovering biomarkers via hub node detection in differential networks, integration of both inter-omics and inter-omics network rewiring detected by the multiDDN method could help to extend the current method to multi-omics data. A further study on the association between

therapeutic biomarkers with active hub nodes in differential networks from multi-omics data would better link the hubs in networks to the actual biomarkers.

Appendix A. Personal Information

A.1. Biography

Yi Fu received his B.S. and M.S degrees in Electrical Engineering from Tsinghua University, Beijing, China, in the year of 2009 and 2012. Since August 2012, he has been a doctoral student and graduate research assistant in Computational Bioinformatics and Bio-imaging Laboratory (CBIL) in the Bradley Department of Electrical and Computer Engineering at Virginia Polytechnic Institute and State University (Virginia Tech), under the supervision of Dr. Yue Wang.

A.2. List of Publication

Journal publications

1. Zhang B, Hou X, Yuan X, Shih IM, Zhang Z, Clarke R, Wang RR, **Fu Y**, Madhavan S, Wang Y, Yu G. "AISAIC: a software suite for accurate identification of significant aberrations in cancers". Bioinformatics. 2013 Nov 29;30(3):431-3.
2. **Fu Y**, Yu G, Levine DA, Wang N, Shih IM, Zhang Z, Clarke R, Wang Y. BACOM2. 0 facilitates absolute normalization and quantification of somatic copy number alterations in heterogeneous tumor. Scientific Reports. 2015 Sep 9;5:13955.
3. Herrington DM*, Mao C, Parker S, Fu Z, Yu G, Chen L, Venkatraman V, **Fu Y**, Wang Y, Howard T, Goo J, Liu Y, Saylor G, Athas G, Troxclair D, Hixson J*, Vander Heide R*, Wang Y*, Van Eyk J, "Proteomic Architecture of Human Coronary and Aortic Atherosclerosis," Circulation, 2018.

4. da Cruz RS, Carney EJ, Clarke J, Cao H, Cruz MI, Benitez C, Jin L, **Fu Y**, Cheng Z, Wang Y, de Assis S. Paternal malnutrition programs breast cancer risk and tumor metabolism in offspring. Breast Cancer Research. 2018 Dec;20(1):99.

5. Clark D. J., Hu Y., Schnaubelt M., **Fu Y**., Ponce S., Chen S. Y., ... & Zhang H. Simple Tip-Based Sample Processing Method for Urinary Proteomic Analysis. Analytical Chemistry, 2019 91(9), 5517-5522.

6. McDermott J, Arshad O, Petyuk V, **Fu Y**, Gritsenko M, ... & the Clinical Proteomic Tumor Analysis Consortium Investigators. Proteogenomic characterization of ovarian high-grade serous cancer implicates mitotic kinase and replication stress. Cell Reports Medicine. Submitted.

Manuscripts in preparation

7. **Fu Y**, Lu Y, Herrington D, Zhang Z, Liu C, Yu G, Clarke R, Wang Y. Differential dependency network analysis reveals significant rewiring of molecular networks between biological conditions.

Conference publications and book chapter

8. Tian Y, Zhang B, **Fu Y**, Yu G, Wang Y. A statistical approach to identifying significant transgenerational methylation changes. In Signal and Information Processing (GlobalSIP), 2014 IEEE Global Conference on 2014 Dec 3 (pp. 1394-1397).

9. **Fu Y**, et al. “Biologic Computing”, Biomedical Information Technology 2nd Edition. David Feng. Academic Press, Massachusetts: Cambridge, 2019. ISBN 9780128160343. Print.

Bibliography

Ahmed, A. and Xing, E.P. Recovering time-varying networks of dependencies in social and biological studies. *Proc Natl Acad Sci U S A* 2009;106(29):11878-11883.

Bandyopadhyay, S., *et al.* Rewiring of genetic networks in response to DNA damage. *Science* 2010;330(6009):1385-1389.

Banerjee, O., Ghaoui, L.E. and d'Aspremont, A.J.J.o.M.l.r. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. 2008;9(Mar):485-516.

Bao, L., Pu, M. and Messer, K. AbsCN-seq: a statistical method to estimate tumor purity, ploidy and absolute copy numbers from next-generation sequencing data. *Bioinformatics* 2014;30(8):1056-1063.

Barabasi, A.L., Gulbahce, N. and Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat Rev Genet* 2011;12(1):56-68.

Barretina, J., *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. 2012;483(7391):603.

Bollobás, B. and Béla, B. Random graphs. Cambridge university press; 2001.

Bryois, J., *et al.* Cis and Trans Effects of Human Genomic Variants on Gene Expression. *PLOS Genetics* 2014;10(7):e1004461.

Buescher, J.M., Driggers, E.M.J.C. and metabolism. Integration of omics: more than the sum of its parts. 2016;4(1):4.

Califano, A. Rewiring makes the difference. *Mol Syst Biol* 2011;7:463.

Cancer Genome Atlas Research, N. Integrated genomic analyses of ovarian carcinoma. *Nature* 2011;474(7353):609-615.

Carter, S.L., *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol* 2012;30(5):413-421.

Chen, C., *et al.* Two gene co-expression modules differentiate psychotics and controls. *Mol Psychiatry* 2013;18(12):1308-1314.

Chen, L., *et al.* Unsupervised Deconvolution of Dynamic Imaging Reveals Intratumor Vascular Heterogeneity and Repopulation Dynamics. *PLoS One* 2014;9(11):e112143.

Chen, L., *et al.* MS-PyCloud: An open-source, cloud computing-based pipeline for LC-MS/MS data analysis. 2018:320887.

Colantuoni, C., *et al.* Temporal dynamics and genetic control of transcription in the human prefrontal cortex. *Nature* 2011;478(7370):519-523.

Creixell, P., *et al.* Navigating cancer network attractors for tumor-specific therapy. *Nat Biotechnol* 2012;30(9):842-848.

da Cruz, R.S., *et al.* Paternal malnutrition programs breast cancer risk and tumor metabolism in offspring. 2018;20(1):99.
de Chassey, B., *et al.* Hepatitis C virus infection protein network. *Mol Syst Biol* 2008;4:230.
Dong, X., *et al.* Learning graphs from data: A signal representation perspective. 2019;36(3):44-63.
Dong, X., *et al.* Reverse enGENEering of regulatory networks from big data: a roadmap for biologists. 2015;9:BBI. S12467.
Downey, C.L., *et al.* The prognostic significance of tumour-stroma ratio in oestrogen receptor-positive breast cancer. *British journal of cancer* 2014;110(7):1744-1747.
Friedman, J., Hastie, T. and Tibshirani, R. The elements of statistical learning. Springer series in statistics New York; 2017.
Friedman, J., Hastie, T. and Tibshirani, R.J.B. Sparse inverse covariance estimation with the graphical lasso. 2008;9(3):432-441.
Friedman, N., Linial, M. and Nachman, I. Using Bayesian networks to analyze expression data. *Journal of Computational Biology* 2000;7:601-620.
Gardiner, K.J.G.b. Gene-dosage effects in Down syndrome and trisomic mouse models. 2004;5(10):244.
Gatza, M.L., *et al.* An integrated genomics approach identifies drivers of proliferation in luminal-subtype human breast cancer. *Nature genetics* 2014;46(10):1051-1059.
Ha, M.J., Baladandayuthapani, V. and Do, K.-A. DINGO: differential network analysis in genomics. *Bioinformatics* 2015;31(21):3413-3420.
Hadfield, J.D.J.J.o.S.S. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. 2010;33(2):1-22.
Han, H., *et al.* TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. 2017;46(D1):D380-D386.
Hartwell, L., *et al.* Genetics: from genes to genomes. McGraw-Hill New York; 2008.
Hecker, M., *et al.* Gene regulatory network inference: data integration in dynamic models—a review. 2009;96(1):86-103.
Hegde, P.S., White, I.R. and Debouck, C.J.C.o.i.b. Interplay of transcriptomics and proteomics. 2003;14(6):647-651.
Herrington, D.M., *et al.* Proteomic architecture of human coronary and aortic atherosclerosis. 2018;137(25):2741-2756.
Hu, J., *et al.* PhosphoNetworks: a database for human phosphorylation networks. *Bioinformatics* 2014;30(1):141-142.
Hu, J.X., Thomas, C.E. and Brunak, S.J.N.R.G. Network biology concepts in complex disease comorbidities. 2016;17(10):615.

Hudson, N.J., Reverter, A. and Dalrymple, B.P. A differential wiring analysis of expression data correctly identifies the gene containing the causal mutation. *PLoS Comput Biol* 2009;5(5):e1000382.

Huijbers, A., *et al.* The proportion of tumor-stroma as a strong prognosticator for stage II and III colon cancer patients: validation in the VICTOR trial. *Annals of oncology* 2013;24(1):179-185.

Hutchins, J.R.J.M.b.o.t.c. What's that gene (or protein)? Online resources for exploring functions of genes, transcripts, and proteins. 2014;25(8):1187-1201.

Ideker, T. and Krogan, N.J. Differential network biology. *Mol Syst Biol* 2012;8.

Isalan, M., *et al.* Evolvability and hierarchy in rewired bacterial gene networks. *Nature* 2008;452(7189):840-845.

Kang, H.J., *et al.* Spatio-temporal transcriptome of the human brain. *Nature* 2011;478(7370):483-489.

Kitano, H. Systems Biology: A Brief Overview. 2002;295(5560):1662-1664.

Klipp, E., *et al.* Systems biology: a textbook. John Wiley & Sons; 2016.

Kuhn, A., *et al.* Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain. *Nature methods* 2011;8(11):945-947.

Kuo, K.T., *et al.* DNA copy numbers profiles in affinity-purified ovarian clear cell carcinoma. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2010;16(7):1997-2008.

Lazar, C., *et al.* Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies. 2016;15(4):1116-1125.

Li, W., Lee, A. and Gregersen, P.K.J.B.b. Copy-number-variation and copy-number-alteration region detection by cumulative plots. 2009;10(1):S67.

Madhamshettiwar, P.B., *et al.* Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets. *Genome Medicine* 2012;4(5):41.

Mateos, G., *et al.* Connecting the dots: Identifying network structure via graph signal processing. 2019;36(3):16-43.

Meier, L., van de Geer, S. and Bühlmann, P. The group lasso for logistic regression. *J. R. Statist. Soc. B* 2008;70:53-71.

Meinshausen, N. and Bühlmann, P. High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics* 2006;34(3):1436-1462.

Meinshausen, N. and Bühlmann, P.J.T.a.o.s. High-dimensional graphs and variable selection with the lasso. 2006;34(3):1436-1462.

Mertins, P., *et al.* Proteogenomics connects somatic mutations to signalling in breast cancer. 2016;534(7605):55.

nature, I.H.G.S.C.J. Initial sequencing and analysis of the human genome. 2001;409(6822):860.
Newman, M.E., Strogatz, S.H. and Watts, D.J.J.P.r.E. Random graphs with arbitrary degree distributions and their applications. 2001;64(2):026118.
Nussbaum, R.L., McInnes, R.R. and Willard, H.F. Thompson & Thompson genetics in medicine e-book. Elsevier Health Sciences; 2015.
Oesper, L., Mahmoody, A. and Raphael, B.J. THetA: Inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome biology* 2013;14(7):R80.
Patti, G.J., Yanes, O. and Siuzdak, G.J.N.r.M.c.b. Innovation: Metabolomics: the apogee of the omics trilogy. 2012;13(4):263.
Pedreschi, R., *et al.* Treatment of missing values for multivariate statistical analysis of gel-based proteomics data. 2008;8(7):1371-1383.
Peila, R., *et al.* Joint effect of the APOE gene and midlife systolic blood pressure on late-life cognitive impairment. 2001;32:2882-2889.
Ranjbar, M.R.N., *et al.* Bayesian normalization model for label-free quantitative analysis by LC-MS. 2014;12(4):914-927.
Shabalin, A.A.J.B. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. 2012;28(10):1353-1358.
Simon, N., Friedman, J. and Hastie, T.J.a.p.a. A blockwise descent algorithm for group-penalized multiresponse and multinomial regression. 2013.
Song, L., *et al.* Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome research* 2011;21(10):1757-1767.
Su, X., *et al.* PurityEst: estimating purity of human tumor samples using next-generation sequencing data. *Bioinformatics* 2012;28(17):2265-2266.
TCGA. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 2008;455(7216):1061.
Tian, Y., *et al.* KDDN: an open-source Cytoscape app for constructing differential dependency networks with significant rewiring. *Bioinformatics* 2015;31(2):287-289.
Tian, Y., *et al.* Knowledge-fused differential dependency network models for detecting significant rewiring in biological networks. *BMC Syst Biol* 2014;8(1):87.
Tian, Y., *et al.* Knowledge-guided differential dependency network learning for detecting structural changes in biological networks. In, *ACM International Conference on Bioinformatics and Computational Biology*. 2011. p. 254-263.
Tian, Y. and Zhang, H.J.P.C.A. Glycoproteomics and clinical applications. 2010;4(2):124-132.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* 1996;58:267-288.

Tibshirani, R., *et al.* Strong rules for discarding predictors in lasso-type problems. 2012;74(2):245-266.

Tibshirani, R., *et al.* Sparsity and smoothness via the fused lasso. 2005;67(1):91-108.

Tsai, T.-H., Wang, M. and Ressom, H.W. Preprocessing and analysis of LC-MS-based proteomic data. In, *Statistical Analysis in Proteomics*. Springer; 2016. p. 63-76.

Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory applications* 2001;109(3):475-494.

Van Loo, P., *et al.* Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A* 2010;107(39):16910-16915.

Volkov, P., *et al.* A genome-wide mQTL analysis in human adipose tissue identifies genetic variants associated with DNA methylation, gene expression and metabolic traits. 2016;11(6):e0157776.

Wang, N., *et al.* UNDO: a Bioconductor R package for unsupervised deconvolution of mixed gene expressions in tumor samples. *Bioinformatics* 2015;31(1):137-139.

Wang, Z., Gerstein, M. and Snyder, M.J.N.r.g. RNA-Seq: a revolutionary tool for transcriptomics. 2009;10(1):57.

Webb-Robertson, B.-J.M., *et al.* Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. 2015;14(5):1993-2001.

Wuhrer, M., *et al.* Glycoproteomics based on tandem mass spectrometry of glycopeptides. 2007;849(1-2):115-128.

Yan, W., *et al.* Biological networks for cancer candidate biomarkers discovery. 2016;15:CIN. S39458.

Yang, S., *et al.* Identification of genes with correlated patterns of variations in DNA copy number and gene expression level in gastric cancer. 2007;89(4):451-459.

Yoshihara, K., *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun* 2013;4:2612.

Yu, G., *et al.* BACOM: in silico detection of genomic deletion types and correction of normal cell contamination in copy number data. *Bioinformatics* 2011;27(11):1473-1480.

Yuan, X., *et al.* Genome-wide identification of significant aberrations in cancer genome. *BMC genomics* 2012;13:342.

Yuan, Y., *et al.* Penalized regression elucidates aberration hotspots mediating subtype-specific transcriptional responses in breast cancer. *Bioinformatics* 2011;27(19):2679-2685.

Yuan, Y., *et al.* Penalized regression elucidates aberration hotspots mediating subtype-specific transcriptional responses in breast cancer. 2011;27(19):2679-2685.

Zack, T.I., *et al.* Pan-cancer patterns of somatic copy number alteration. 2013;45(10):1134.

Zhang, B., *et al.* Differential Dependency Network Analysis to Identify Condition-Specific Topological Changes in Biological Networks. *Bioinformatics* 2009;25(4):526-532.

Zhang, B., *et al.* DDN: a caBIG(R) analytical tool for differential network analysis. *Bioinformatics* 2011;27(7):1036-1038.

Zhang, B. and Wang, Y. Learning structural changes of Gaussian graphical models in controlled experiments. In, *Uncertainty in Artificial Intelligence (UAI 2010)*. 2010.

Zhang, B. and Wang, Y. Learning structural changes of Gaussian graphical models in controlled experiments. In, *26th Conference on Uncertainty in Artificial Intelligence, UAI 2010*. 2010. p. 701-708.

Zhang, H., *et al.* Integrated proteogenomic characterization of human high-grade serous ovarian cancer. 2016;166(3):755-765.

Zhang, X.-F., *et al.* Differential network analysis from cross-platform gene expression data. 2016;6:34112.

Zhao, Z., *et al.* Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature genetics* 2006;38(11):1341-1347.

Zou, H. and Hastie, T. Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society B* 2005;67:301-320.