DECIPHERING TRANSCRIPTIONAL PATTERNS OF GENE REGULATION: A COMPUTATIONAL APPROACH

by
Princy Parsana

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
July, 2020

Abstract

With rapid advancements in sequencing technology, we now have the ability to sequence the entire human genome, and to quantify expression of tens of thousands of genes from hundreds of individuals. This provides an extraordinary opportunity to learn phenotype relevant genomic patterns that can improve our understanding of molecular and cellular processes underlying a trait. The high dimensional nature of genomic data presents a range of computational and statistical challenges. This dissertation presents a compilation of projects that were driven by the motivation to efficiently capture gene regulatory patterns in the human transcriptome, while addressing statistical and computational challenges that accompany this data. We attempt to address two major difficulties in this domain: a) artifacts and noise in transcriptomic data, and b) limited statistical power.

First, we present our work on investigating the effect of artifactual variation in gene expression data and its impact on trans-eQTL discovery. Here we performed an in-depth analysis of diverse pre-recorded covariates and latent confounders to understand their contribution to heterogeneity in gene expression measurements. Next, we discovered 673 trans-eQTLs across 16 human tissues using v6 data from the Genotype Tissue Expression (GTEx) project. Finally, we characterized two trait-associated trans-eQTLs; one in Skeletal Muscle and another in Thyroid.

Second, we present a principal component based residualization method to correct gene expression measurements prior to reconstruction of co-expression networks. In this work, we demonstrated theoretically, in simulation, and empirically, that principal component correction of gene expression measurements prior to network inference can reduce false positive edges. Using data from the GTEx project in multiple tissues, we showed that this approach reduced false discoveries beyond correcting only for known confounders.

Third, we present a multi-study integration approach to identify universal transcriptional patterns underlying epithelial to mesenchymal transition (EMT) across different cancer types. With informed statistical analysis and functional validation, we identified consensus ranked universal EMT genes. This gene list consisted of a) known EMT genes, b) genes studied in a subset of carcinomas, unknown in prostate cancer, and c) novel unknown EMT and cancer genes such as *C1orf116*.

Finally we present methods to integrate co-expression signals across multiple human RNA-seq data to reconstruct networks with increased power. First, we considered multiple aggregation strategies to build context-agnostic networks using data from recount2. These networks captured ubiquitous patterns of gene co-expression shared across tissues and cell types. Next, we briefly describe a hierarchical mixture model *groupNet* that leverages signal from multiple datasets to learn the structure of a Gaussian Markov random field (GRMF) to build context-specific co-expression networks.

Thesis Readers

Dr. Alexis Battle (Primary Advisor)
Associate Professor
Department of Biomedical Engineering
Department of Computer Science
Johns Hopkins University

Dr. Kenneth J. Pienta (Co-advisor)
Professor
Department of Urology
Johns Hopkins University

Dr. Michael C. Schatz
Associate Professor
Department of Computer Science
Department of Biology
Johns Hopkins University

For my mom Bharti, and my husband Anand

Acknowledgements

I'm deeply grateful to all those who have supported me through my Ph.D - in terms of science and in terms of life. You all have helped me be one whole person through this journey. I owe it to so many people.

I would like to express my deepest gratitude to my advisor Alexis Battle for her constant support, guidance, and mentorship throughout my Ph.D. I am grateful to her for taking a gamble on me as her first Ph.D. student, and bringing me to this day. Alexis welcomed me as the first member of Battle lab when I was relatively new to quantitative concepts. She provided me with a rich intellectual environment and showed tremendous patience while I learned statistical concepts and machine learning. Alexis has been a role model right from the beginning, and over the years I have only come to realize how hard it is to be as good as her. I continue to be amazed by her creativity, intelligence, and scientific rigor. In addition to science, I have valued the kindness and understanding that Alexis has shown in all these years. I could not have asked for a better Ph.D. advisor. Thank you for making this ride worthwhile; thank you for everything, Alexis. You will be *always* be my advisor.

I am thankful to my co-adviser Kenneth Pienta, for his patience and support through these years. Ken has continuously emphasized the importance of communicating scientific research in a way that is accessible to a wide audience. In addition to science, he has invested time into helping me develop my presentation skills. I have appreciated his understanding when projects did not work. I am extremely thankful to Ken for his selfless support in all endeavours that I took up during my Ph.D. Thank you for Ken, for always believing in me.

I am deeply grateful to both Alexis and Ken for moulding me into the scientist that I am today.

Michael Schatz, has been a source of encouragement. His scientific and career advice during different stages of graduate school have been invaluable. Thank you Mike, for always believing in me, and keeping my confidence high. I would like to thank Jeffrey Leek for the collaboration due to which I had the opportunity to work on networks correction project. I am thankful for his scientific advice on the project, and his time and mentorship when I was trying to make career decisions.

I would like to thank Michael Schatz, Jeffrey Leek, and Benjamin Langmead also for agreeing to be on my GBO committee.

In addition, I would also like to thank Scott Smith, the Graduate Program Director of Computer Science for his support for the founding of CS Graduate Student Council. Scott has always been a student supporter, and I am thankful to him and Zacchary Burwell for many instances when they helped and advised us (the council members) to navigate and address student concerns. I would like to thank staff members across Hopkins who helped me so much during these years - Deborah DeFord, Zachhary Burwell, Tracy Marshall, Cecelia Nina Jackson-Goode, Laura Graham, Shani McPherson, Cathy Thornton, Beverly Comegys, Diane Labuda, and Sarah Andersson. I am also thankful to Gregory Hager; it was through him that I found Alexis was going to join Johns Hopkins.

It has been a pleasure to work with many talented collaborators including Brian Jo, Barbara Englehardt, Meritxell Olivia, Barbara Stranger Tuuli Lappalainen, Stephen Montgomery, Christopher Brown, Andrew Jaffe, Kasper Hansen, Claire Ruberman, James Hernandez, Sarah Amend, Gonzalo Torga, Dorhyun Johng, and William Isaacs. I am also grateful to the members of the GTEx consortium; along with great scientific collaborations, I've also found friends by the virtue of being on the consortium. Working on GTEx papers has emphasized the importance and value of team work like nothing else.

I am so grateful to the tutors and organizers of the Leena Peltonen School of Human Genomics that I attended in August 2019 - Manolis Dermitzakis, Nancy Cox, Ceciliar Lindgren, Tuuli Lappalainen, Sara Pulit, Alexandre Reymond, Jacques Fellay, Ruth Loos, Len Pennacchio, Samuli Ripatti, Didier Trono, Gosia Trynka, and Cisca Wijmenga. It was the best meeting of my graduate school. I am also thankful to the organizers of the Rising Stars Biomedical workshop that I attended in 2019; particularly Archana Venkatraman, Nicholas Durr, and Sri Sarma for their valuable inputs during the workshop.

There have been many people in and outside of Hopkins who have helped me in my journey as a growing scientist.

I would first like to thank Giovanni Parmigianni for responding to a cold email, and giving me the opportunity to intern with his group at Dana Farber Cancer Institute in 2012. Thank you to Curtis Huttenhower for co-advising me during the internship. This internship marked the beginning of my research career. In particular, I am grateful to Levi Waldron for all his patience during the internship and teaching me everything from how to organize my code, how to interpret results, to research writing. I am amazed by his kindness and patience, and his dedication towards science. My positive experience working with Giovanni, Levi, and Curtis was the reason I decided to go to graduate school. Not having studied at one of the top schools in India, this internship was very critical for my career. Thank you Giovanni, that email indeed opened doors for nearly all the opportunities that have followed. I owe it to you.

I have been fortunate to be surrounded by individuals who have given their time to share thoughts about life and career - particularly thankful to Hariharan Easwaran, Sara Pulit, Cecilia Lindgren, Leonardo Collado-Torres, Abhinav Nellore, Stephanie

Hicks, Margaret Taub, and Shannon Ellis. They have been generous with sharing their experiences, while encouraging me to choose what worked best for me.

Waking up with the excitement to go to work is the sign of a good workplace and I have found that in my labmates - a group of smart, intelligent, and yet humble folks, who all strive to maintain a positive and friendly work environment. I have enjoyed long technical discussions that helped me grow as a researcher, and the tea/coffee breaks that helped me unwind and relax. Thank you - Benjamin Strober, Yuan He, Ashis Saha, Marios Arvanitis, Prashanthi Ravichandran, Amy He, Karl Tayeb, Diptavo Dutta, Benjamin Shapiro, Yungil Kim, Taibo Li, April Kim, Rebecca Keener, Jessica Bonnie, Surya Chhetri, Ashton Omdahl, Josh Popp, Victor Wang, and Matt Figdore. I am also thankful to my friends Katherine Henry, Narges Ahmidi, Christopher Wilks, Jonathan Jones, Robert DiPietro, Nikita Ivkin, Ayushi Sinha, Molly O'Brien, Rachel Sherman, Yassamin N, Eli Sherman, and Mark Jacob Pritt.

Around 9 years ago, I moved to Baltimore to pursue my Masters at Johns Hopkins. I could not have made it this far in the academic journey without the bunch of people whom I call my family away from home. I would like to thank my roommates Pooja Chaudhari and Amrita Kundu for always being there. We shared our groceries and household responsibilities, just like family. I still love to boast about our chemistry as roommates, and as friends. Amod Jog has been my go to person; I have benefited from his advice at several instances. Thank you for listening to my endless practice talks and lectures, providing detailed feedback, and asking insightful questions. I have enjoyed our discussions on PCA over wine. Aagam, Pooja, Kundu, Siddharth, and Gowtham - thank you for listening to me when I wanted to be heard and for advising me when I needed it. I am so glad to have shared the start and the end of my Ph.D. with two of my closest friends, Aagam and Kundu - from different continents we have shared the happy and difficult times of graduate life. Pooja, Gowtham, Vishwa, and Kesha - I can't thank you enough for voluntarily being a part of our life as parents and embracing us with our new normal. I would like to thank Vishnu, Mehul, Pratik, and Anasuya for always making a conscious effort to come visit us from Boston, and for sharing invaluable career advice especially during the current times. Pratik has been a great friend through these years, has patiently answered my questions on linear algebra and machine learning. I have cherished all the *chai pe charcha*, game nights, potlucks, and vacations with Siddharth M, Amod, Pooja, Gowtham, Kundu, Aagam, Vishwa, Anushka, Aditya, Tushar, Priyanka, Aditya, Neha, Ambhi, Rajita, Akanksha, Varun, Shourya, Sravya, Ashish, Amit, Vaibhav S, Piyush, Rujuta, Akshay, Shambhavi, Sonal, Hari, Madhuri, Farhad, Balaji, Michelle, Samata, Sahil, Priti, Tariq, and so many more names I am forgetting.

Finally, I'd like to thank my family. I am grateful to my parents Bharti and Navin for their never ending support. My mom is my support system and the pillar that keeps me strong during challenging times. Thank you Mummy, for always encouraging me to chase my dreams, for holding my hand when things were difficult with that smile which still fills me with energy and enthusiasm, and for teaching me to be kind in my actions and polite in my words. Thank you for making me the person I am today. I thank my siblings Radhika (and Lalit), Janki (Juhi), Chaitanya, and Ramkrishna for being so positive through these years, and for being my first set of students. They have advised me and supported me on important decisions of life. I also want to thank my Ba (Nirmalaben), my cousins Prasann and Nandini, my masi (Daksha), and my mama (Vipul) for their love. Thank you for being my happy place, you all mean the world to me!

A series of fortunate events led me to come to Hopkins, where I met my husband Anand, who grew up in the same neighborhood as me in Mumbai. No words can summarize the roles Anand has played in my life, and the kind of support he has brought. He has been an everlasting cheerleader, a patient teacher, a selfless mentor, caring friend, a loving husband, and an incredible father to our daughter, Eera. He taught me my first lessons on calculus and linear algebra, without which none of the quantitative learning would have happened. I have found my best friend in him. Thank you for dealing with my workaholic nature, reminding me of life outside of lab, for understanding me ALWAYS, and for all the happy times Anand; it has been such a joy to have traversed this journey with you. You make me a better person. I also thank Anand's family - Vaibhav, Akanksha, Mummy (Shobha), and Papa (Omprakash). Particularly Vaibhav who is my second champion in the Malpani family.

I want to thank my daughter Eera, our bundle of joy. She has redefined love, patience, and work for me. I have learned to be persistent while watching her take her first steps; to be patient from her calmness on days when our meals were delayed; and to be curious as she continues to learn new things everyday. Thank you for being an awesome baby.

I would like to end my acknowledgements by thanking my late grandparents Bacchubhai, Shantaben, and Nanubhai for their love and care throughout their lives. They would have been very proud today.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Innovations in engineering and technology have played a remarkable role in discovery and clinical translation in medicine (https://aimbe.org/milestones-of-innovation/). The completion of the Human Genome Project (HGP) in 2003 was a notable breakthrough, with its main goals of a) determining a draft sequence of the 3 billion nucleotides that make up the human genome and b) identifying genes that it contains. This was the culmination of a large-scale, interdisciplinary, thirteen-years-long effort involving scientists with diverse expertise ranging from biology, chemistry, and genetics to mathematics and computer science. The publication of the sequence of the human genome in April 2003 was a groundbreaking achievement with the potential to revolutionize medicine[1]. Since then, rapid advancements in sequencing technologies have made it possible to sequence the human genome and quantify expression measurements for every human gene – thereby presenting the opportunity to investigate patterns of gene expression and regulatory structure at the genome-wide level. Effective utilization of these data also presents a range of computational and statistical challenges.

Background

Cellular and tissue level organization of the human body

Humans are complex multi-cellular organisms, and this complexity increases with increasing levels of cellular organization. Cells are the smallest independent functional entities of the human body. For example cardiomyoctyes are cells that make up the cardiac muscle. A group of similar types of cells that work together to perform a specific function come together to form a tissue (e.g. cardiac muscle tissue). An organ is composed of a group of tissues that perform specific physiological functions. The heart, for instance, is the organ composed of cardiac muscle tissue that pumps blood throughout the body. Finally, an organ system constitutes multiple organs that together perform systemic physiological functions. The circulatory system consists of the organs that transport nutrients and oxygen to different parts of the body. These levels of organization increase in complexity starting from cells to tissues to organs to organ systems to a complete organism.[2]

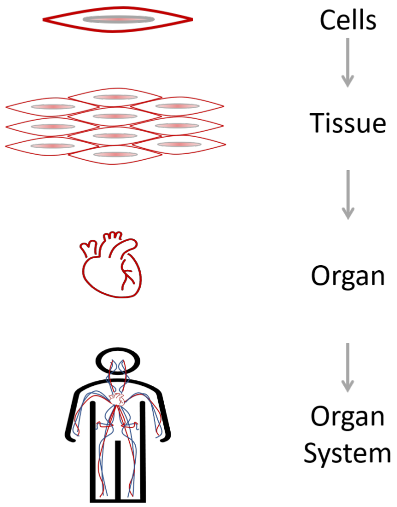


Figure 1-1. Structural organization of human body.

Central Dogma of molecular biology

Each cell of every multi-cellular organism, including humans, contains a copy of its genome. A genome carries the complete set of genetic information required for normal functioning and survival of an organism. It is often also referred as the ‘blueprint’ of an organism. It is made of Deoxyribonucleic Acids (DNA) that contains instructions to make proteins and other molecules required for normal cell functioning. The information in DNA is stored in the form of a code that contains four nucleic acid bases: adenine, cytosine, guanine, thymine. Parts of DNA that code for functional end products such as RNA (Ribonucleic Acid) and protein are called genes. The central dogma of molecular biology explains the information flow from DNA to its end products through gene expression. There are two key stages of gene expression:

- Transcription – the process during which information in DNA is converted to RNA

- Translation – the process during which information from RNA is converted to amino acid sequences, which are the building blocks of proteins [3]

Similar to DNA, the information in RNA is also stored as a code in the form of four nucleic acid bases: adenine, cytosine, guanine, and uracil. Historically, RNA was viewed as an intermediate messenger between genes and proteins. While messenger RNA, also called as mRNA serves as an intermediate molecule, there exists a diverse group of non-protein coding RNAs such as lincRNAs, miRNAs, piRNAs, snoRNAs, rRNA, and tRNA that are functional end products by themselves[4]. Despite containing the same exact copy of the genome, different cells and tissues in the human body have different phenotypes and perform very diverse functions. Regulatory programs recruit gene products in the form of proteins and RNAs to generate specific transcriptional patterns thus enabling cells, tissues, and organs to perform distinct functions. This

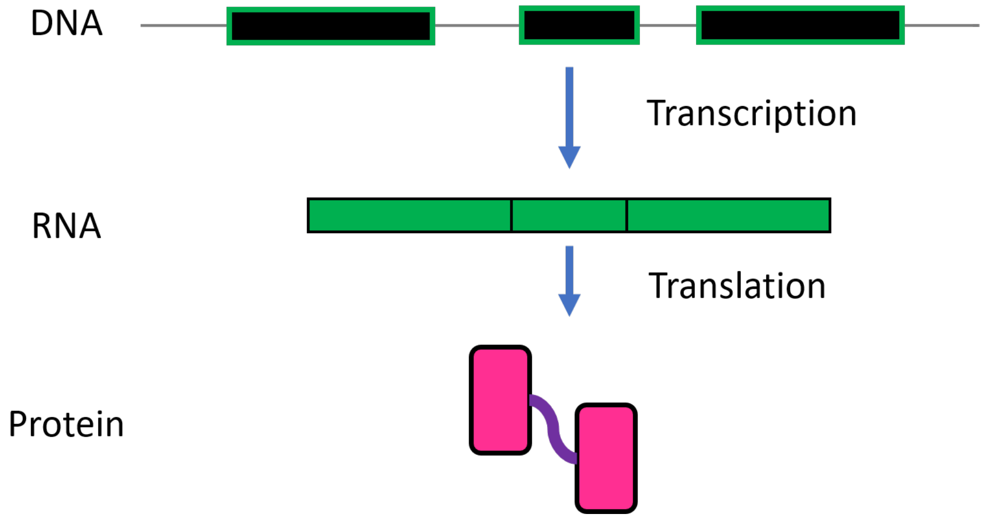


Figure 1-2. Central dogma of molecular biology.

control can be exerted at various stages of gene expression including genetic, epigenetic, transcriptional, and post-transcriptional regulation[5–7].

Gene expression and phenotypic variation

For many years after the discovery of the structure of DNA in 1953 scientists turned their focus to understanding the functions of protein coding genes and how mutations in these genes alter quantities and structures of the resulting proteins[8]. Defects in gene regulatory programs can alter expression patterns in cells and tissues which can eventually lead to clinically relevant phenotypes. It is critical to understand regulation of gene expression and how it changes or adapts in response to genetic, chemical, or environmental stimulus. Therefore, identifying genes or genetic variants associated with phenotypic variance can provide insight into understanding cellular and molecular processes involved in trait manifestation.

Differential gene expression analysis

Differential gene expression analysis (DGE) involves identifying genes that show significantly different expression patterns between conditions or are associated with a phenotype of interest such as diseased vs normal patients, treatment vs control, cell types, tissues, developmental stage, or other conditions.

Statistical hypothesis tests that are commonly used to identify differentially expressed genes can be grouped into two categories:

- Parametric: These tests are based on assumptions that the data being tested follow particular statistical distribution. For example, t-test and ANOVA assumes that the data follows a gaussian distribution.

- Non-parametric: These tests do not make any assumptions about the data. Wilcoxon rank-sum test is an example of a rank based non-parametric test

Usually in these tests, there are two hypotheses: a) the null hypothesis (H_0) that states that there is no statistical difference between the expression of the gene between the two groups and b) an alternate hypothesis (H_1) that states that there is a statistically significant difference in the expression of the gene between the two groups.

Example:

Given that we are interested in using a t-test to test if a gene g is differentially expressed between individuals with cancer (C) and normals (N), we test:

$$H_0 : \mu_C = \mu_N \tag{1.1}$$

$$H_1 : \mu_C \neq \mu_N \tag{1.2}$$

Here, μ_C and μ_N correspond to the mean expression of gene g in cancer patients and normal patients.

Identification of differentially expressed genes can extend our understanding of genetic and molecular processes involved in phenotypic variation. Further, genes identified can be used in downstream analysis such as building a relevant prediction model or identifying disease subtypes.

Genetic variation and transcriptomic regulation

The most frequent variants in the human genome are Single Nucleotide Variants (SNV). These are substitutions that affect a single base pair. A SNV that is present in a sufficiently large fraction of a population is called a Single Nucleotide Polymorphism (SNP). Genome wide association studies (GWAS) utilize large scale population data to identify germline SNPs associated with a trait. Since the first GWAS on age related macular degeneration in 2005, there has been an abundance of trait associated variants reported in literature. However, a traditional GWAS does not provide a framework to link these statistical associations to genes or functional biological mechanisms underlying a trait; this remains a daunting task. Genetic variants can have a series of cascading effects first relayed at the molecular level from mRNA to protein to pathways which can then impact cellular and physiological processes underlying traits and diseases.

The majority of trait-associated genetic variants discovered by GWAS have been found in the non-coding regions of the genome. In response, over the last 10 years there have been several large scale efforts to improve understanding of the functional effect of genetic variation on gene expression and its impact within and across tissues. Genetic variants that are quantitatively associated with amounts of gene expression

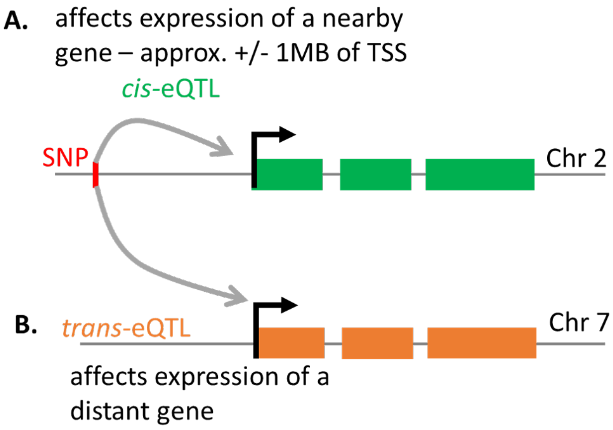

A.
affects expression of a nearby gene – approx. +/- 1MB of TSS
cis-eQTL
SNP
Chr 2
B.
trans-eQTL
Chr 7
affects expression of a distant gene

Figure 1-3. Cis-eQTLs and trans-eQTLs.

are called expression quantitative trait loci(eQTL)[7, 9, 10]. Two types of eQTLs are of common interest, as seen in Figure 1-3:

- cis-eQTLs: genetic variants that quantitatively affect expression of genes on the same molecule of the chromosome.
- trans-eQTLs: genetic variants that quantitatively affect expression of gene located on a different molecule of chromosome.

While there has been significant progress in understanding the mechanism of cis-genetic effect on gene expression, gene expression regulation mediated by trans-eQTLs is not completely understood[11]. Trans-regulatory effects can be mediated through cis-regulation affecting expression of a nearby gene, which would then in turn alter expression of a distant gene. Trans-eQTLs can provide a framework to understand the cascading effect of genetic and molecular signalling and implication in diseases1-4.

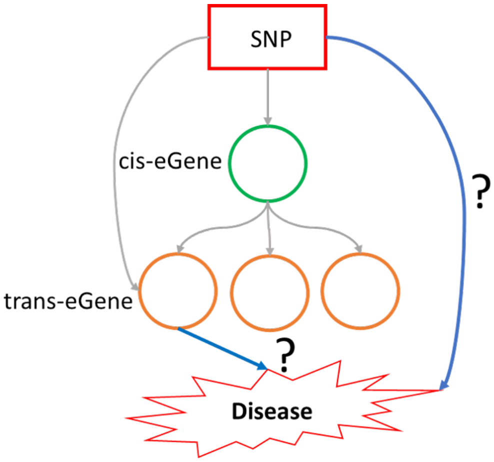


Figure 1-4. Trans-eQTLs.

Co-expression networks

Genes are known to interact with each other to relay signal across biological pathways and processes. This form of signal transduction is a critical piece in the foundation of human biology. [12–14]. This is evident by concerted expression patterns among genes observed in empirical data. Networks are often used to model such interactions among entities in complex systems.

A co-expression network is an undirected graph where genes are represented as nodes and a functional relationship between genes is represented as an edge between nodes. Gene networks can identify patterns of expression indicative of functional and regulatory relationships among genes. These can be used to determine critical pathways and genes underlying a trait or disease [15]. A complete understanding of in vivo functional interactions among genes is lacking for most cell types, tissues, or disease relevant contexts. Therefore, discovering functional relationships between genes can improve our understanding of genetic and molecular bases of gene regulation. These functional relationships can also provide insights into the cascade of molecular events critical for disease manifestation in humans under a variety of conditions.

Challenges

A primary concern for nearly all modalities of genomic measurements is that the data is inherently noisy. Transcriptomic measurements are routinely affected by a wide-range of artifacts and unwanted heterogeneity that is not the primary signal of interest [16–19]. Structured artifactual signal by covariates such as RNA integrity number of a sample, proportion of GC nucleotides in a gene, or batches in which samples are processed adds biased noise in gene expression measurements. These biases in the data are sometimes correlated with outcomes or variables of interest, that can lead to inaccurate conclusions[16]. Most scientific data also contains unbiased random (white) noise that does not have a pattern, but can yet affect statistical power to detect signal. In this dissertation, we attempt to address the impact of biased noise in gene expression data. Additionally, limited statistical power is another major hurdle in computational genomics. Low probability of finding true signal, overestimation of effect sizes leading to false positives, and lack of reproducibility are some of the major issues that result from under-powered studies. While we have the ability to measure expression for over 40,000 genes, the number of gene expression samples available is usually limited to a few hundred samples which is insufficient to make consistent and statistically robust conclusions.

Thesis outline

Amidst noisy, under-powered, and heterogeneous genomic data, this thesis presents carefully informed statistical and machine learning approaches to elucidate the underlying basis of transcriptional regulation along with some examples of its implication in human disease. Towards this, we describe: a) methods to analyze and address the impact of confounding on gene expression data in trans-eQTL mapping and co-expression networks, and b) multi-study integration based approaches to leverage power across multiple studies to improve signal estimation in differential expression analysis and gene co-expression networks.

- In chapter 2 using GTEx data, we perform an in-depth analyses of diverse confounding variables that contribute to gene expression heterogeneity and affect mapping of eQTLs. Next, we identify genetic variants with distal regulatory effects on gene expression (trans-eQTLs) and characterize some trait associated trans-eQTLs. This work was published in Nature and involved joint effort of multiple trainees and PIs in the GTEx consortium. Trans-eQTL analysis was joint work with Brian Jo, Yuan He, and Benjamin Strober [20].

- In chapter 3, we present a principal component based residualization approach to correct gene expression data prior to reconstruction of co-expression networks. This was joint work with Claire Ruberman and was published in Genome Biology [21].

- In chapter 4, we present a multi-study integration based approach to identify global transcriptional regulatory patterns underlying epithelial to mesenchymal transition (EMT) phenotype in cancer. In this work, with informed statistical analysis and functional validation, we identified global expression patterns in epithelial to mesenchymal transition (EMT) phenotype in cancer and discovered candidate regulatory genes. This work was published in BMC cancer [22].

- In chapter 5, we present an aggregation-based approach to reconstruct context-agnostic gene co-expression networks using large scale transcriptomic data from recount2[23]. We find influential genes and relationships involved in critical processes shared across different biological contexts such as cell cycle, mitosis, etc. This is joint work with Prashanthi Ravichandran.

- In chapter 6, we present a mixture model based probabilistic method to reconstruct context specific gene co-expression networks. This can be particularly useful when working with publicly available open source datasets, which often have missing context-specific meta-data. Using preliminary simulation analyses we show that our method can identify and group studies by relevant context.

- In chapter 7, we summarize the work in this dissertation and discuss future directions.

Chapter 2

Distant regulatory effects of genetic variation across human tissues

While we have successfully mapped cis-eQTLs for a majority of genes, discovery of replicable trans-eQTLs remains a challenging task. Trans-eQTL discovery is particularly impacted by: a) small effect sizes, b) artifact-induced false positives, and c) statistical power. Most studies have attempted trans-eQTL discovery in limited cells or tissue types [24–26]. It is known that genetic regulation of gene expression varies across tissues and cell types. Context-specific regulation of gene expression recruits specific transcriptional programs that allow different types of cells and tissues to perform distinct biological functions. In this work, we performed trans-eQTL mapping across 44 human tissues from the Genotype Tissue Expression (GTEx) Project to understand distant regulatory effects of genetic variation on gene expression.

Contributions

This chapter describes the trans-eQTL analyses from the GTEx project that I co-led along with Brian Jo, Yuan He, and Benjamin Strober. My main contributions to this work included:

- Investigation of latent factors anticipated to capture artifactual variation in gene expression that were used for eQTL mapping, and its impact on trans-eQTL discovery

- Replication of trans-eQTLs in thyroid with TCGA Thyroid cancer data

- Investigation of functional role of eVariants and eGenes in thyroid and skeletal muscle, and performed relevant analyses

This work was published in [20].

Introduction

The human genome encodes instructions for the regulation of gene expression, which varies both across cell types and across individuals. Recent large-scale studies have characterized the regulatory function of the genome across a diverse array of cell types each from a small number of samples[27–29]. Measuring how gene regulation and expression vary across individuals has further expanded our understanding of healthy tissue function and the molecular origins of complex traits and diseases[7, 9, 24, 25, 30, 31]. However, to date, these studies have been conducted in limited, accessible cell types, thus restricting the utility of these studies in informing regulatory biology and human health.

In this study, we associate genetic variants with gene expression levels from the GTEx v6p release. For the first time, we identify trans-eQTLs across 16 tissues and highlight their increased tissue specificity relative to cis-eQTLs. We evaluate trans-eQTLs to characterize their functional characteristics, genomic context, and relationship to disease-associated variation.

Study design

The GTEx project has created a reference resource of gene expression from ’normal’, non-diseased tissues. Every tissue sample was examined histologically. If the tissue was non-diseased and in the normal age-range of the donor, the sample was accepted. RNA was isolated from postmortem samples in an ongoing manner as donors were enrolled in the study. For this data release, 44 sampled regions or cell lines were considered, each from at least 70 donors and thereby considered suitable for eQTL analysis: 31 solid-organ tissues, ten brain subregions including duplicates of two regions (cortex and cerebellum), whole blood, and two cell lines derived from donor blood and skin samples. We hereafter refer to these tissues, regions, and cell lines as the *tissues* used in eQTL analysis. A total of 7,051 samples from 449 donors represent the GTEx v6p analysis freeze (Fig. 1a). This is 4.3 times more samples than reported in the GTEx pilot phase[32]. DNA was genotyped at 2.2 million sites and imputed to 12.5 million sites (11.5 million autosomal and 1 million X chromosome sites) using the multi-ethnic reference panel from 1000 Genomes Project Phase 1 v3[33]. Sampled donors were 83.7% European American and 15.1% African American. Whole genome sequencing was performed for 148 donors to a mean coverage greater than 30×, and all donors were exome-sequenced to a mean coverage over captured exons of 80×. The resulting data provide the deepest survey of individual and tissue-specific gene expression to date, enabling a comprehensive view of the impact of genetic variation on gene expression. All data are available from dbGaP (accession phs000424.v6.p1) with multiple publicly available data views available from the GTEx Portal (www.gtexportal.org).

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

Sample procurement

All human donors were deceased donors. Informed consent was obtained for all donors via next-of-kin consent to permit the collection and banking of de-identified tissue samples for scientific research. The research protocol was reviewed by Chesapeake Research Review Inc., Roswell Park Cancer Institute's Office of Research Subject Protection, and the institutional review board of the University of Pennsylvania. Complete descriptions of the donor enrollment and consent process, as well as biospecimen procurement, methods, sample fixation, and histo-pathological review procedures were previously described[32, 34]. Briefly, whole blood along with fresh skin samples was collected from each donor and shipped overnight to the GTEx Laboratory Data Analysis and Coordination Center (LDACC) at the Broad Institute for DNA genotyping, RNA expression, and culturing of lymphoblastoid and fibroblast cells. Two adjacent aliquots were then prepared from each sampled tissue and preserved in PAXgene tissue kits. One of each paired sample was embedded in paraffin (PFPE) for histopathological review. The second was shipped to the LDACC for processing and molecular analysis. Brains were collected from approximately 1/3rd of the donors and were shipped on ice to the brain bank at the University of Miami where 11 brain sub-regions were sampled and flash-frozen. These samples were also shipped to the LDACC for processing and analysis.

All DNA genotyping was performed on blood-derived DNA samples unless unavailable, in which case a tissue-derived DNA sample was substituted. RNA was extracted from all tissues and RNA sequencing was performed on all samples with a RIN score of 5.7 or higher and with at least 500 ng of total RNA. Nucleic acid isolation protocols and sample QC metrics applied are as described in the previous study[32] (Supplementary Information in [20]).

Data production

Non-strand specific, polyA+ selected RNA-seq libraries were generated using the Illumina TruSeq protocol. Libraries were sequenced to a median depth of 78 million 76-bp paired-end reads. RNA-seq reads were aligned to the human genome (hg19/GRCh37) using TopHat (v1.4) based on GENCODE v19 annotations. This annotation is available on the GTEx Portal (gencode.v19.genes.v6p model.patched contigs.gtf.gz). Gene-level expression was estimated as reads per kilobase of transcript per million mapped reads (RPKM) using RNA-SeQC on uniquely mapped, properly paired reads fully contained with exon boundaries and with alignment distances ≤ 6. Samples with less than 10 million mapped reads or with outlier expression measurements based on the D-statistic were removed.

DNA from 450 donors was genotyped using Illumina Human Omni 2.5M and 5M Beadchips. Genotypes were phased and imputed with SHAPEIT2[35] and IMPUTE2[36], respectively, using multi-ethnic panel reference from 1000 Genomes Project Phase 3[37]. Variants were excluded from analysis if they: (1) had a call rate $< 95\%$; (2) had minor allele frequencies $< 1\%$; (3) deviated from Hardy-Weinberg Equilibrium ($P < 1.0 \times 10^{-6}$); or (4) had an imputation info score less than 0.4. The final genotyped and imputed array VCF (file format v4.1) for autosomal variants contained genotype posterior probabilities for each of the three possible genotypes for 11,552,519 variants across 450 GTEx donors. The dosages of the alternative alleles relative to the human reference genome hg19 were used as the genotype measure for subsequent eQTL analysis. In addition to array-based genotyping, 148 and 524 donors were whole genome and exome-sequenced respectively. Additional details on genotyping, imputation, and sequencing can be found in the Supplementary Information of [20]

RNA-seq data processing and correction for technical confounders

We conducted trans-eQTL mapping within the 44 tissues with at least 70 samples each. Only genes with ten or more donors with expression estimates > 0.1 RPKM and an aligned read count of six or more within each tissue were considered significantly expressed and used for trans-eQTL mapping. Within each tissue, the distribution of RPKMs in each sample was quantile-transformed using the average empirical distribution observed across all samples. Expression measurements for each gene in each tissue were subsequently transformed to the quantiles of the standard normal distribution. The effects of unobserved confounding variables on gene expression were quantified with PEER[17], run independently for each tissue. Fifteen PEER factors were identified for tissues with fewer than 150 samples; 30 for tissues with sample sizes between 150 and 250; and 35 for tissues with more than 250 tissues. The covariates that were most consistently associated with PEER factors include factors related to parameters of donor death, ischaemic time, RIN, and sequencing quality control metrics. In addition, we have observed that little, if any, genetic signal is present in the PEER factors.

To further understand the effect of PEER correction on gene expression and trans-eQTL mapping in each tissue, we compared the PEER factors from each tissue to sample and donor specific covariates. First, we fit a linear model between gene expression data E and loadings PEER factors L. Using this model, we obtained the expression component E_f, that was removed by PEER correction as given below:

$$E = \mu + \beta \cdot L$$

$$E_f = E - E_r$$

We tested the association of E_f with different sample specific and donor specific covariates. In each tissue, we first selected covariates with more than one unique entry after excluding missing values. For covariates with categorical entries, we only considered categories with more than 20 observations. Samples that did not meet this criteria were considered as missing values. The covariates that were included in this analysis had at least 50 observations (without missing values) in at least 31 tissues based on the above criteria. Finally, for each selected covariate we fit a linear model with expression component removed by PEER factors, E_f as the dependent variable and the covariate C as:

$$E_f = \mu + \beta \cdot C$$

From this model, we computed the proportion of variance of E_f explained by the covariate as the adjusted R^2:

$$\text{Adjusted } R^2 = R^2 - \left[(1 - R^2)\frac{p}{p - n - 1}\right]$$

Trans-eQTL mapping

Matrix eQTL[38] was used to test association of all autosomal variants (MAF > 0.05) with all gene transcripts restricted to variants and genes lying on different chromosomes in each tissue independently using an additive linear model. For trans-eQTL mapping, we tested variants for association with expression of only protein coding or lincRNA genes. We included, as covariates, the three genotype PCs, genotyping platform, sex, and PEER factors estimated from expression data in Matrix eQTL when performing association testing. The correlation between variant and gene expression levels was evaluated using the estimated t-statistic from this model. The corresponding FDR was estimated using Benjamini-Hochberg FDR correction[38, 39] separately within each tissue and using permutation analysis. For all trans association tests, we applied stringent quality control to account for potential false positives due to RNA-seq read mapping errors, repeat elements, and population stratification (Supplementary Information at [20]).

Multi-tissue eQTL mapping

We quantified tissue-specificity and tissue-sharing of trans-eQTLs using Meta-Tissue[40] which extends Metasoft[41], a meta-analysis package, by using a mixed effects model for eQTL sharing that accounts for correlation of expression between tissues driven by overlapping donors.

All genotypes and gene expression quantification estimates were adjusted for covariates in accordance to the single tissue analysis as described in the previous sections. For each variant-gene pair, we calculated mixed model effect size estimates in each expressed tissue, thereby adjusting for partial sharing of signal between tissues. These effect size estimates were used in meta-analysis using Metasoft[41] to assess tissue-specificity of each variant-gene pair. For each variant-gene pair tested, Meta-Tissue estimates a global P-value of association and the posterior probability that an effect exists in a tissue (m-value). For computational feasibility, the MCMC method was used to approximate the exact solution.

To supplement this analysis, we also performed multi-tissue analysis using a hierarchical FDR control[42] for trans-eQTLs analysis (Supplementary Information at [20])

Co-localization of GWAS and eQTL associations

In order to assess the probability that molecular traits as estimated by cis- and trans-eQTLs, and physiological traits as estimated by GWAS share the same causal variant, we applied the `coloc` R package[43]. For each GWAS, we approximated the number of independent loci by extracting variants with at least genome-wide significance ($P < 5 \times 10^{-8}$) and farther than 1 MB away from all other variants of higher statistical significance. For each genome-wide significant variant, we extracted the list of all eGenes (q-value < 0.05 for cis-eGene) within 1 Mb for coloc analyses. For each eGene, we excluded any variants without either eQTL or GWAS association statistics (effect size estimate, standard error and P-value). We obtained reference information such as MAF, sample size, and case-to-control proportions (in case of binary traits) for each variant whenever available otherwise, study-wide estimate was used as a proxy. We defined a region or an eGene as having evidence of co-localization when region- or gene-based posterior probability of co-localization ($PP.H4.ABF$) $\frac{PP4}{PP3+PP4} > 0.9$.

TCGA thyroid RNA-seq analysis

To replicate trans-eVariants in thyroid, we used Thyroid Carcinoma (THCA) RNA-seq and genotype array data from The Cancer Genome Atlas (TCGA). Filtering out tumor normal and metastatic samples, we restricted our analysis to 498 primary tumor samples. Next, after log transforming RNA-seq RSEM measurements, we ensured that expression of each gene follows a Gaussian distribution by projecting each gene expression levels to the quantiles of a standard normal. To account for noise and confounding factors in RNA-seq measurements, we corrected the data by 35 PEER factors. Using a linear model while adjusting for 35 PEER factors with MatrixeQTL, we tested the effect of each variant on chr 9 position 100600000 - 100670000 on expression levels of all trans genes. We used the Benjamini-Hochberg method to correct for multiple hypotheses testing (assessed only among 24 variants tested). Genes with FDR ≤ 0.1 were called as trans-eGenes.

Data and biospecimen availability

Genotype data from the GTEx v6p release are available in dbGaP (study accession phs000424.v6.p1; www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424.v6.p1). The VCFs for the imputed array data are in phg000520.v2.GTEx MidPoint Imputation.genotype-calls-vcf.c1.GRU.tar (the archive contains a VCF for chromosomes 1-22 and a VCF for chromosome X). Allelic expression data is also available in dbGap. Expression data (read counts and RPKM) and eQTL input files (normalized expression data and covariates for 44 the tissues) from the GTEx v6p release are available from the GTEx Portal (http://gtexportal.org). eQTL results are available from the GTEx Portal.

Results

Trans-eQTL mapping 44 human tissues

To identify trans-eQTLs, we tested for association between every protein-coding or lincRNA gene and all autosomal variants where the gene and variant were on different chromosomes. To minimize false positives in trans-eQTL detection, we controlled for the same observed and inferred confounders as optimized for cis-eQTL discovery, and further removed genes with poor mappability, variants in repetitive regions, and trans-eQTLs between pairs of genomic loci that show evidence of RNA-seq read cross-mapping due to sequence similarity[44]. Applying this approach, we found 673 trans-eQTLs at a 10% genome-wide FDR. This includes 112 distinct loci ($R^2 \leq 0.2$) and 93 unique genes (94 total gene associations, including an eGene detected in both testis and thyroid) in 16 tissues (Table 2-I, Figure 2-1). An alternative approach to quantify FDR at the gene level identified 46 genes at 10% FDR, with estimated q-values less than 0.4 for all 94 gene associations identified using the genome-wide FDR (Table 2-I).

Testis had the most trans-eGenes, with 35 eGenes in 157 samples (Figure 2-1), reflecting the elevated number of expressed genes (16,853 protein-coding genes and 4,362 lincRNA genes) and cis-eGenes (6,796 genes). We found statistical power to detect additional associations in these restricted tests, such as the test restricted to cis-eVariants. Our results indicate that increases in sample sizes will continue to yield additional eQTLs, especially in the trans-eQTL setting where statistical power is the

		Genome wide		**Gene-level FDR**
Tissue	No. of samples	No. of trans-eGenes	No. of trans-eVariants	No. of trans-eGenes
Muscle – Skeletal	361	9	43	4
Whole Blood	338	1	2	1
Skin – Sun Exposed (Lower leg)	302	6	16	3
Adipose – Subcutaneous	298	2	7	0
Lung	278	2	2	2
Thyroid	278	21	181	3
Cells – Transformed fibroblasts	272	1	10	1
Nerve – Tibial	256	0	0	1
Esophagus – Mucosa	241	3	11	3
Artery – Aorta	197	1	1	1
Skin – Not Sun Exposed (Suprapubic)	196	1	1	2
Stomach	170	0	0	2
Colon – Transverse	169	2	10	2
Testis	157	35	267	16
Pancreas	149	2	12	1
Adrenal Gland	126	1	1	1
Brain – Putamen (Basal ganglia)	82	3	11	2
Vagina	79	4	27	1
Total unique		93	602	46

Table 2-I. Trans-eVariant and trans-eGene discoveries for genome-wide FDR control, and trans-eGene discoveries for gene-level FDR control. Each tissue with non-zero values is included as a row; the columns include the number of samples for that tissue, followed by the number of unique trans-eGenes and trans-eVariants identified in the genome-wide tests and the number of unique trans-eGenes found using gene-level FDR calibration. Ultimately the set of 673 trans-eQTLs identified in the genome-wide approach yielded 602 unique trans-eVariants.

major limitation.

Correction for technical confounders in trans-eQTL mapping

To account for hidden batch effects and other potential confounders in the gene expression data, we used the Probabilistic Estimation of Expression Residuals (PEER) [45] method to estimate a set of latent covariates for gene expression levels for each tissue type. The number of PEER factors was selected to maximize ciseGene discovery, and this optimization was performed for three sample size bins: tissues with fewer than 150 samples, tissues with ≤ 150 and < 250 samples, and tissues with ≥ 250 samples. Specifically, the eQTL discovery pipeline was run in increments of 5 PEER factors for 12 tissues spread across the sample size bins using a 100 permutations. Based on these results, and to avoid potential overfitting, 15, 30, and 35 PEER factors were selected, respectively for the three sample size bins(Figure 2-2). We did not

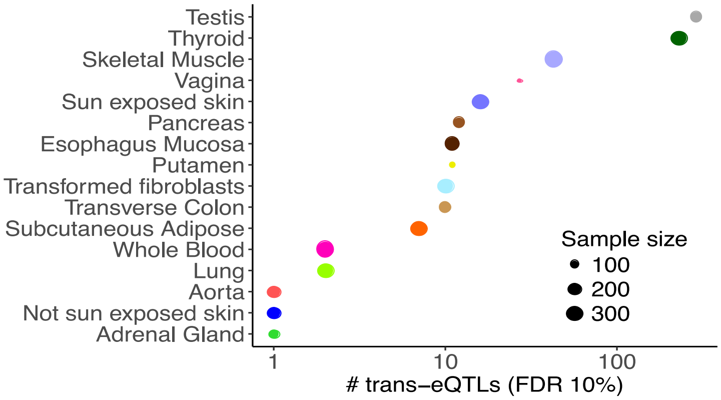


Figure 2-1. Trans-eQTL discovery. Number of trans-eQTLs (x-axis) per tissue (y-axis), with sample size indicated by point size.

have sufficient statistical power or sufficient numbers of trans-eQTLs to tune the number of PEER factors for trans-eQTL analysis without facing potential overfitting to spurious signal. Post-hoc analysis demonstrated no clear trend in number of trans-eQTL discoveries as we varied the number of PEER factors removed. Further, failure to remove confounding factors could result in false positive transeQTL associations. Therefore, we opted to use the settings determined by the analysis of cis-eQTLs for the trans-eQTL analysis as well. This aggressive correction, explained 59-78% of total variance in gene expression levels, however may lead to false negatives, reducing the signals for broad effect trans-eVariants with many target genes. Indeed, several loci with numerous associations that were found in uncorrected data disappeared after controlling for PEER factors. We also found that the *trans*-eVariants detected before PEER correction were enriched for association with known technical confounders (Figure 2-3).

We tested association of PEER factors from each tissue with known technical and biological covariates recorded for each sample and donor. PEER factors from each tissue were correlated with known technical and biological covariates recorded for each sample and donor (Figure 2-4,2-5). The covariates that were most consistently associated with PEER factors include factors related to parameters of donor death,

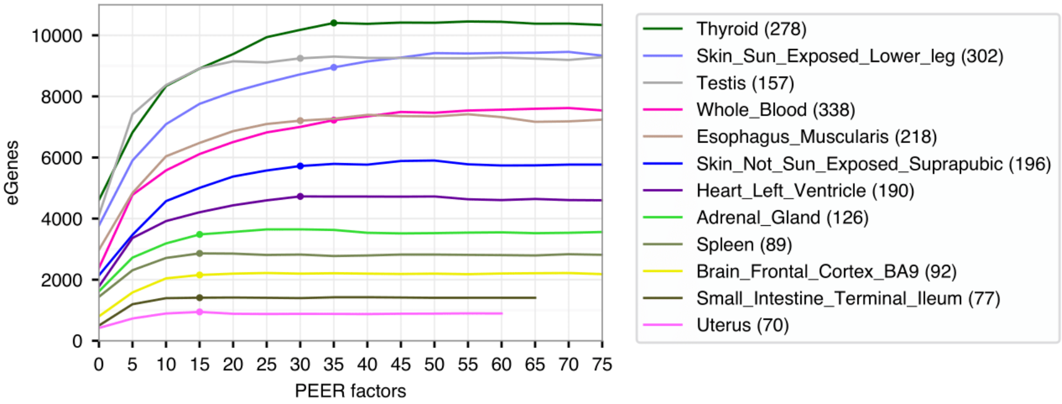


Figure 2-2. Identification of the optimal number of PEER factors for hidden covariate correction during eQTL analyses. The number of PEER factors was chosen to maximize eGene discovery and this optimization was performed for three sample size bins: tissues with < 150 samples, tissues with $\geq$150 and $<$250 samples, and tissues with $\geq$250 samples available. The eQTL discovery pipeline was run with increments of 5 PEER factors for the 12 tissues shown using a reduced number of permutations (100 instead of the adaptive 1000-10000 used for all other analyses). Based on these results and to avoid potential overfitting, 15, 30, and 35 PEER factors were selected respectively.

ischemic time, RIN, and sequencing quality control metrics. Nucleic acid isolation and library construction batches and total sequencing depth were also moderately associated. Across tissues, the median percent variance explained (PVE) by RIN of the set of PEER factors used for correction was 0.05, with a maximum PVE of 0.13 in heart $-$ left ventricle. The PVE by these covariates of the expression data after PEER correction was negligible-median 4×10^-3 for RIN. Similarly, after correction, the detected trans-eVariants show little association with known covariates. For example, the two tissues with the most trans-eQTLs, thyroid and testis, show no association between RIN and any trans-eVariant at FDR 50%.

Tissue-specific patterns of trans-eQTLs

We observed much greater tissue specificity for trans-eQTLs than a set of FDR-matched cis-eQTLs (Figure 2-6). This observation was robust to choices of m-value threshold and selection criteria for matching cis-eQTLs. While 3.8% of trans-eQTLs were shared

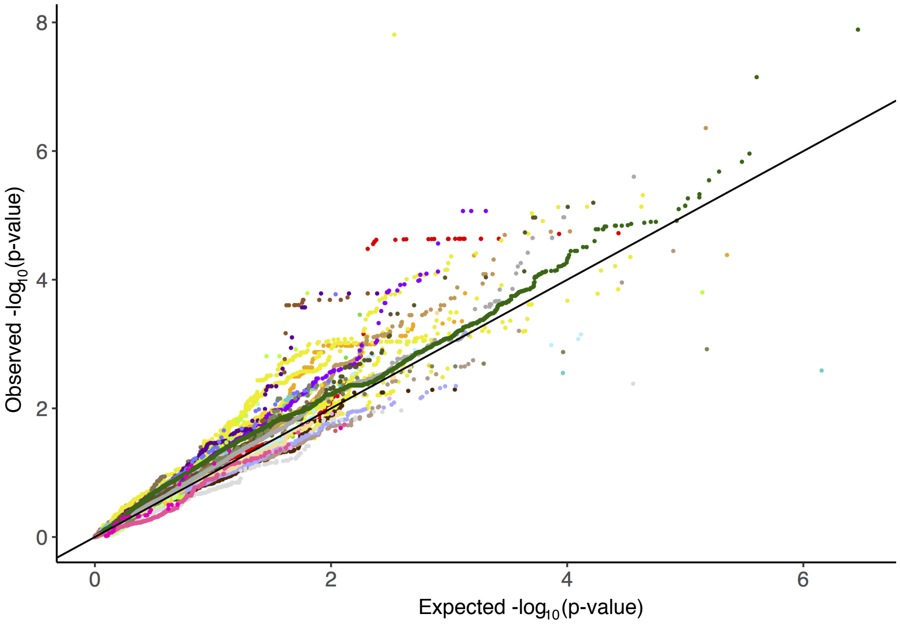


Figure 2-3. Trans-eVariants lost after PEER correction are enriched for association with known covariates Trans-eVariants that were detected in raw expression data but lost after PEER correction were tested for association with known sample covariates using a linear model. This quantile-quantile plot shows - log10(P-values) of trans-eVariants lost after PEER correction as compared to matched random variants, with each tissue shown as a distinct color. Combined across tissues, the association - log10(P-values) are significantly larger than random (Wilcoxon rank sum test; $P \leq 2.2 \times 10^{-16}$).

across three or more tissues at m-value > 0.9, 25.3% of FDR-matched cis-eQTLs were shared. Extensive tissue-specificity for trans-eQTLs was also observed based on a hierarchical approach for FDR control[46], where we found no trans-eQTLs shared across more than one tissue (Table2-II). Our estimate of increased tissue specificity for trans-eQTLs agreed with the minimal sharing of trans effects reported in previous eQTL studies with fewer tissues[25, 47, 48], and dramatically exceeds what would be expected based on replication between tissues for cis-eQTLs of matched minor allele frequency (MAF) and effect size (Wilcoxon rank sum test; P $\leq 2.2 \times 10^{-16}$ for all choices of replication FDR). Given greater tissue-specificity of trans-eQTLs, we note that heterogeneity in cellular composition of bulk tissue samples is one important confounder that may reduce power to detect trans-eQTLs, or even lead to false positive associations[24]. Despite high tissue-specificity, we did observe a small number of

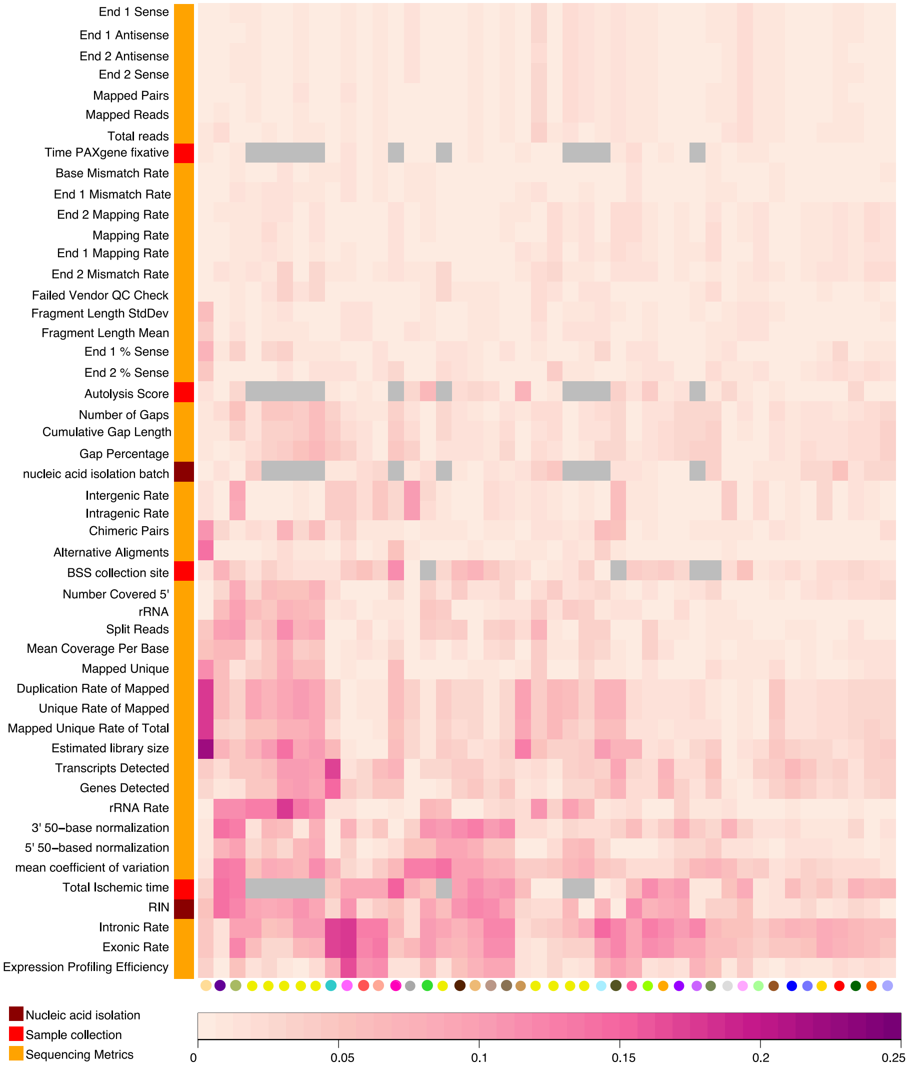


Figure 2-4. Sample covariates associated with PEER factors in each tissue. For each tissue, adjusted (R^2) reflecting the proportion of variance explained by each sample-specific covariate, for the entire PEER component removed from the expression data. Each cell reflects variance explained for a tissue/covariate pair (color scale at bottom). Grey cells represent pairs with insufficient data for estimation.

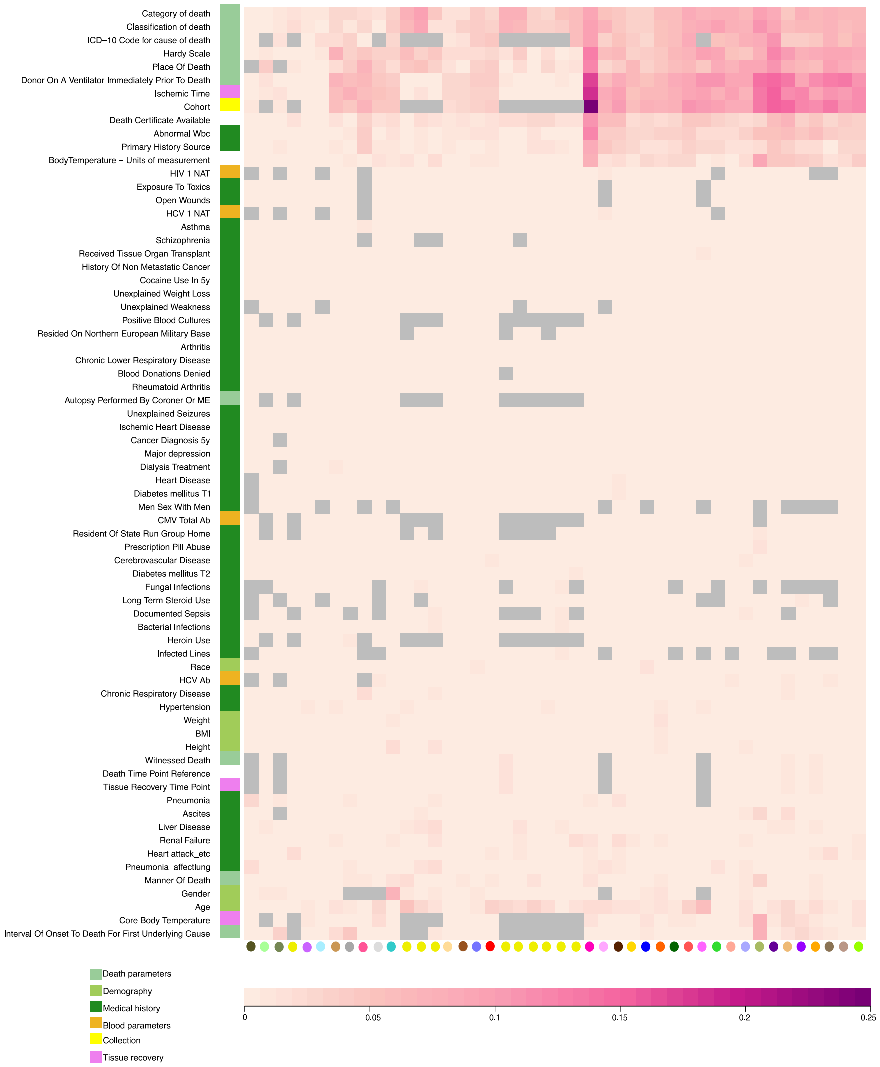


Figure 2-5. Donor covariates associated with PEER factors in each tissue. For each tissue, adjusted (R^2) reflecting the proportion of variance explained by each donor-specific covariate, for the entire PEER component removed from the expression data. Each cell reflects variance explained for a tissue/covariate pair, color scale at bottom. Grey cells represent pairs with insufficient data for estimation.

tissue-shared trans-eQTLs, including rs7683255, which was moderately associated in trans with *NUDT13* across most tested GTEx tissues with consistent direction of effect. We also found examples of trans-eQTLs shared across a subset of related tissues, such as an association between rs60413914 and *RMDN3*, a gene with increased expression in brain, and for which the trans-eQTL had moderate effects in all tested brain regions but no strong effect in other tissues.

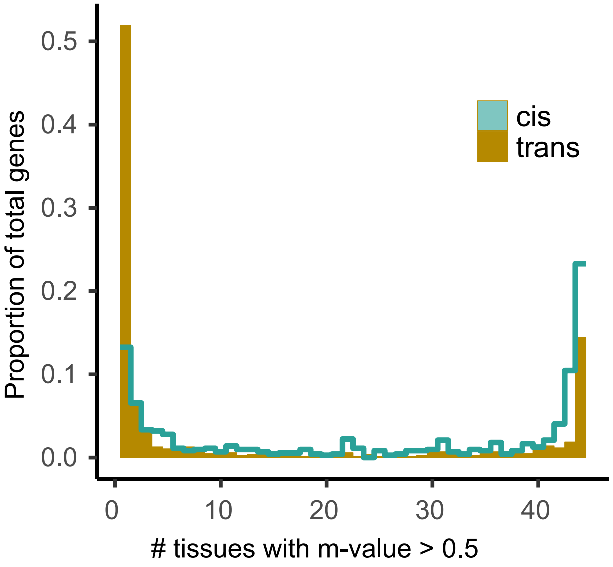


Figure 2-6. Tissue specific patterns of trans-eQTLs. Distribution of the number of tissues having Meta-Tissue $m < 0.5$ for the top variant for each trans-eGene at 50% FDR, and FDR-matched, randomly selected cis-eGenes (also 50% FDR). cis-eGenes were matched for discovery tissue to the trans-eGenes

Trans-eQTLs and complex disease associations

Overlaps between GWAS associations and eQTLs have provided important insights into regulatory genes and variants for a wide range of complex traits and diseases[30, 49]. Genetic variants associated with complex traits have been suggested to be enriched for trans-eQTLs[24, 50–52]. Accordingly, we performed trans-eQTL mapping restricting to variants associated with a complex trait in a GWAS. In this analysis, across the 44 tissues, we found 29 trans-eQTL associations involving 24 unique variants and 25 unique genes, each specific to a single tissue. There were more trans-eVariants at

Tissue	No. of samples	No. of trans-eGenes	No. of trans-eVariants
Whole Blood	338	1	1
Skin – Sun Exposed (Lower leg)	302	2	3
Lung	278	2	2
Thyroid	278	2	2
Esophagus – Mucosa	241	3	3
Artery – Aorta	197	1	1
Skin – Not Sun Exposed (Suprapubic)	196	1	1
Heart – Left Ventricle	190	1	1
Testis	157	4	5
Colon – Sigmoid	124	1	1
Brain – Cortex	96	1	1
Brain – Putamen (Basal ganglia)	82	1	1
Total unique		20	22

Table 2-II. Trans-eVariant and trans-eGene discoveries with hierarchical FDR control. Only tissues with non-zero discoveries are shown. The three-level hierarchical procedure (see Online Methods) performs FDR control across tissues. More specifically, it controls the FDR of eVariants, the average proportion of false variant-gene associations across all eVariants, and a weighted average of false tissue discoveries for the selected variant-gene pairs (weighted by the size of the eVariant and eGene sets). The procedure was applied after LD pruning.

FDR ≤ 0.5 with association in at least one tissue when testing was restricted to trait-associated variants compared with random variants matched by MAF and distance to TSS (Fisher's exact test, P $\leq 1.3 \times 10^{-3}$). Among trait-associated variants with trans-eQTL effects, we found two genome-wide significant trans-eVariants at the 9q22 locus (rs7037324 and rs1867277, $R^2 = 0.74$) with thyroid-specific associations in trans with *TMEM253* and *ARFGEF3* (P $\leq 2.2 \times 10^{-16}$ for both with rs1867277; Figure 2-7). The 9q22 locus has previously been linked to multiple thyroid-specific diseases including goiter, hypothyroidism, and thyroid cancer[53–55], and LoF mutations in a thyroid-specific TF at this locus, *FOXE1*, manifest as ectopic thyroid tissue or cleft palate in developing mice[56]. However, the mechanism of any cis-effects of these trans-eVariants remains uncertain from the GTEx data. A post-hoc analysis demonstrated that PEER correction removed broad regulatory signals from the 9q22 locus, particularly from cis- and trans-eQTL signals for *FOXE1*. In PEER corrected data, cis- and trans-eQTL signals co-localized for another cis-eGene in 9q22, *C9orf156*, for both trans-eGenes (posterior probability > 0.99)[43]. Mendelian randomization

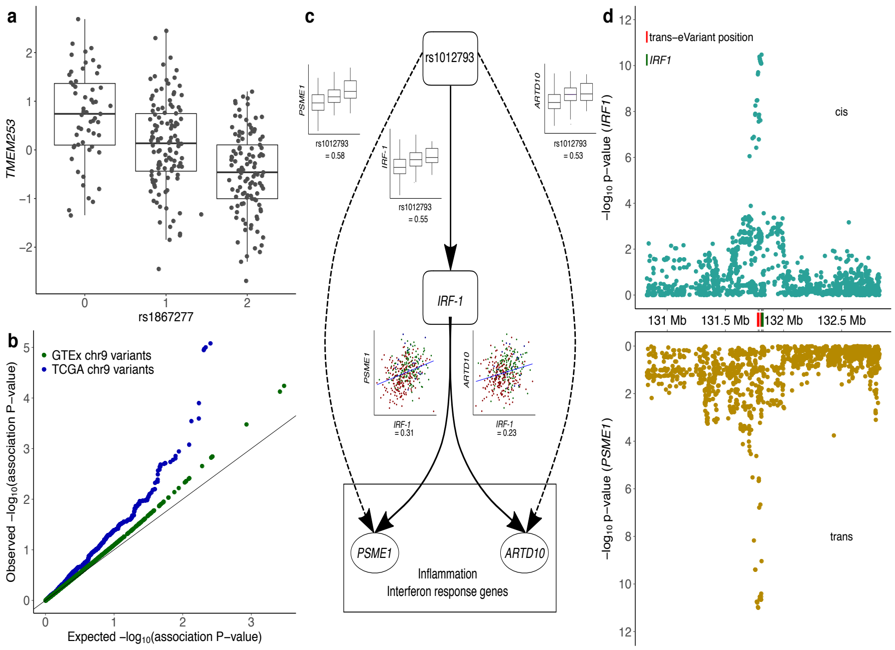


Figure 2-7. Characterization of complex trait-associated trans-eQTLs. (a) Association of rs1867277 with PEER corrected *TMEM253* expression levels (P $\leq 2.2 \times 10^{-16}$). (b) Quantile-quantile plot of associations between 19 variants in the 9q22 locus and all genes in GTEx thyroid gene expression levels, compared to 19 random variants from the same chromosome, and associations between 23 variants in the 9q22 locus and all genes in TCGA thyroid tumor expression data, compared to 23 random variants from the same chromosome. (c) Network depicting cis- and trans-regulatory effects of rs1012793 mediated through interferon regulatory factor 1 (*IRF1*). Rs1012793 affects expression of *IRF1* in cis and *PSME1* and *ARTD10* in trans (box plots). *IRF1* is significantly co-expressed with the trans-eGenes (scatter plots). (d) Cis and trans association significance of variants within 1 Mb of *IRF1* TSS in chromosome 5 locus with cis-eGene *IRF1* (blue) and trans-eGene *PSME1* (brown), showing concordant signal across the locus.

analysis of the PEER-corrected data supported *C9orf156* regulating *TMEM253* (P $\leq 1.3 \times 10^{-9}$) and *ARFGEF3* (P $\leq 2.1 \times 10^{-11}$) based on trans-eVariant rs1867277. In contrast, *FOXE1* had weak Mendelian randomization support in the PEER-corrected data. Despite the ambiguity of cis-mediation, the locus is one of the strongest trans-eQTL signals in GTEx. We further replicated both the broad regulatory effect and specific target genes of this locus in 498 primary thyroid cancer RNA-seq samples from The Cancer Genome Atlas (*TCGA*; Figure 2-7b)[57].

In a second example, two muscle-specific trans-eVariants at the 5q31 locus (rs2706381

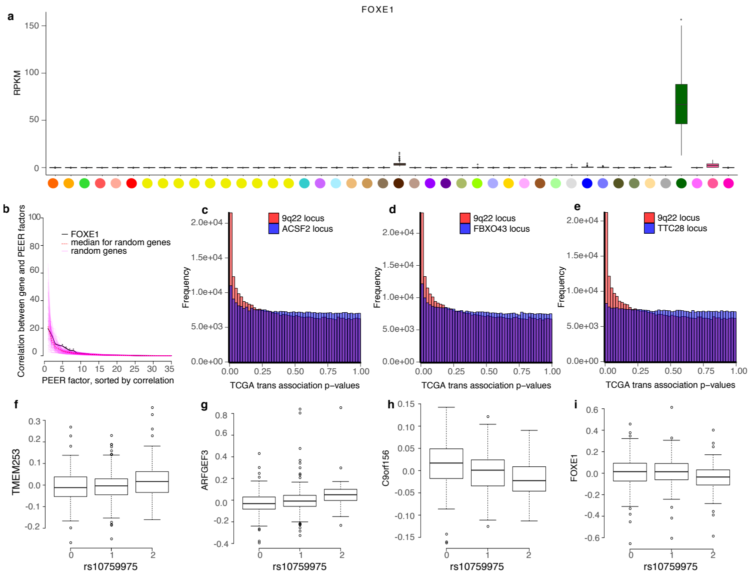
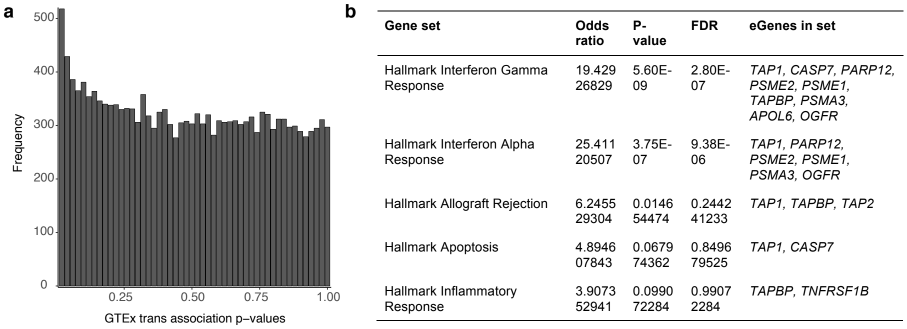


Figure 2-8. Broad trans-regulatory locus 9q22 in thyroid tissue. (a) *FOXE1* expression is thyroid-specific. **(b)** Correlation between *FOXE1* expression levels and thyroid PEER factors compared to 100 random genes. For every gene, absolute correlation was sorted in decreasing order. The correlation of *FOXE1* with the 5th, 6th, 7th, and 8th PEER factors was significantly higher than the correlation of random genes at those rank ordered PEER factors (empirical P ≤ 0.05). **(c-e)** Variants in the chr 9q22 locus were enriched for association with genes on other chromosomes in thyroid carcinomas compared to randomly selected variants nearby randomly selected genes. We used variants that were found within 35 Kb upstream or downstream of the gene TSS. **(f)** rs10759975 is associated with trans-eGene *TMEM253*. **(g)** rs10759975 is associated with trans-eGene *ARFGEF3*. **(h)** rs10759975 shows cis association with *C9orf156*. **(i)** rs10759975 is weakly associated in cis with *FOXE1*.

and rs1012793; $R^2 = 0.84$) were associated in trans with *PSME1* (P $\leq 1.1 \times 10^{-11}$) and *ARTD10* (P $\leq 7.8 \times 10^{-10}$), and in cis with *IRF1* (P $\leq 2.0 \times 10^{-10}$; Figure 2-7c), a transcription factor known to facilitate regulation of interferon-induced immune-response[58–61]. Both variants are associated with circulating fibrinogen levels[62] influencing muscle injury, Duchene muscular dystrophy (DMD), multiple sclerosis, and rheumatoid arthritis[63–66], and have been shown to drive fibrosis in DMD, where they promote expression of *IL-1β* and *TGF-β57*. These variants were moderately associated with numerous genes in skeletal muscle (50 trans-eGenes at 20% FDR, assessed only among the three variants; Figure 2-9a). Additional candidate target genes (at 20% FDR) were enriched in multiple immune pathways from MsigDB[67]



Gene set	Odds ratio	P-value	FDR	eGenes in set
Hallmark Interferon Gamma Response	19.42926829	5.60E-09	2.80E-07	*TAP1, CASP7, PARP12, PSME2, PSME1, TAPBP, PSMA3, APOL6, OGFR*
Hallmark Interferon Alpha Response	25.41120507	3.75E-07	9.38E-06	*TAP1, PARP12, PSME2, PSME1, PSMA3, OGFR*
Hallmark Allograft Rejection	6.245529304	0.014654474	0.244241233	*TAP1, TAPBP, TAP2*
Hallmark Apoptosis	4.894607843	0.067974362	0.849679525	*TAP1, CASP7*
Hallmark Inflammatory Response	3.907352941	0.099072284	0.99072284	*TAPBP, TNFRSF1B*

Figure 2-9. Trait-associated variants in skeletal muscle near interferon regulatory factor *IRF1*. (a) rs1012793 has broad regulatory impact in skeletal muscle. **(b)** Gene set enrichment for potential trans-eGene targets (identified at P ≤ 0.001) of skeletal muscle 5q31 locus.

(Figure 2-9b). Mendelian randomization analysis supported *IRF1* regulating *PSME1* (P $\leq 3.1 \times 10^{-8}$) and *ARTD10* (P $\leq 1.9 \times 10^{-7}$) through cis-eVariant rs2706381 with a consistent direction of effect (Figure 2-7c). Moreover, the cis-eQTL signal for *IRF1* co-localized with the trans-eQTL signals for both trans-eGenes (Figure 2-7c; posterior probability > 0.99)[43]. Together, these results suggest that cis-regulatory loci affecting *IRF1* are regulators of interferon-responsive inflammatory processes involving genes including *PSME1* and *ARTD10*, with implications for complex traits specific to muscle tissue.

Discussion

Since the initial sequencing of the human genome, extensive effort has been devoted to the characterization of genome function and phenotypic consequences of genetic variation. Describing the effects of genetic variation on gene expression levels across tissues is a critical but challenging component of this goal. Here, we describe advances enabled by the GTEx project v6p data, which provide a comprehensive survey of gene expression and the impact of genetic variation on gene expression across diverse human tissues. While considering the effect of artifacts and latent confounders, we report trans-eQTLs in 18 tissues and discovered that trans-eQTL effects tended to be tissue-specific and were correspondingly more enriched in enhancer regions. Further, we characterize two trait associated trans-eQTLs, one in skeletal muscle, and another in thyroid tissue - along with a potential mechanism of cis-mediated trans genetic regulation. Our work provides comprehensive characterization of trans-eQTLs across human tissues, which contribute to an improved understanding of tissue-specific cellular mechanisms of regulatory genetic variation.

Chapter 3

Addressing confounding artifacts in reconstruction of gene co-expression networks

Introduction

Gene co-expression networks seek to identify transcriptional patterns indicative of regulatory relationships between genes[12, 13, 15]. These are not yet fully characterized for most species, tissues, and disease-relevant contexts. Therefore reconstructing co-expression networks from high-throughput measurements is of common interest. However, accurate reconstruction of such networks remains a challenging problem. Though some specialized methods for reconstruction of co-expression networks do consider confounding signals within their model[68, 69], routinely used network learning methods [70, 71] do not directly account for technical and unwanted biological effects known to confound gene expression data. Despite this, many studies do not employ any form of data correction, or correct only for known confounders prior to network reconstruction (Table A-I). These artifacts influence gene expression measurements, often introducing spurious correlations between genes[18, 19, 19, 72]. These correlations are often inferred as relationships between genes, leading to inaccurate network structure and erroneous conclusions in downstream analyses[19, 68, 69, 73, 74]. Therefore, it is critical to correct gene expression data for unwanted biological and technical variation without eliminating signal of interest before applying standard network learning methods.

In this chapter, we present a principal component based residualization method to correct gene expression data prior to building co-expression networks. We demonstrate theoretically, in simulation, and empirically, that principal component correction of gene expression measurements prior to network inference can reduce false discoveries. Using data from the GTEx project in multiple tissues, we show that this approach reduces false discoveries beyond correcting only for known confounders.

Contributions

I co-led this project along with Claire Ruberman. My main contributions to this work include:

- Design and analysis of empirical experiments
- Design and analysis of simulation experiments along with Claire Ruberman

The work described in this chapter was published in [21]. The text of this chapter is a slight modification of the published work.

Methods

All analyses was performed using R and scripts are available on github [75]

Principal component based correction of gene expression

Using a permutation based approach as described in [76], we first determined the number of principal components p to correct the data for with the *num.sv* function in the Bioconductor package sva (Table 3-I). By permuting expression of each gene, *num.sv* identifies the number of top principal components that contribute to non-random expression variance in the data. Next we compute the principal component loadings L of the standardized expression matrix with singular value decomposition (SVD).Using a linear model, we regressed the top p principal components (p as determined by *num.sv*) on the each gene E_i, from the expression data and computed the residuals $\hat{E}_i$:

$$E_i = \mu_i + \beta_i \times L_{1:p} \tag{3.1}$$

$$\hat{E}_i = E_i - [\mu_i + (\beta_i \times L_{1:p})] \tag{3.2}$$

	Total # of PCs removed
Whole Blood	23
Lung	28
Skeletal Muscle	36
Tibial Artery	31
Sun-exposed skin	32
Tibial Nerve	31
Adipose Subcutaneous	37
Thyroid	36

Table 3-I. Number of principal components removed in each tissue.

Simulated example

We construct a network with eight nodes that represent genes and three edges that represent conditional dependencies between the genes. Next, we simulate 10,000 observations from a multivariate normal distribution encoding the conditional dependencies corresponding to three edges as non-zero entries in the precision matrix (Figure 1a). To introduce confounding in the data, we simulate a sample specific term by drawing a random vector of 10,000 observations from a standard normal distribution, and add a scalar multiple of that to genes 2 through 6 (Figure 1d). Finally, to correct the data, we regress out the first principal component from the confounded data (Figure 1g). We used graphical lasso to reconstruct networks using the three versions of the data. The code for this simulation example and network reconstruction can be found at: https://github.com/leekgroup/networks_correction/blob/master/publication_rmd/simulation_example_fig1/figure1.Rmd

Simulation with scale-free networks

We simulated 10,000 observations from a multivariate gaussian distribution that encodes conditional dependencies across 100 genes corresponding to a sale-free network. This was obtained with B-A algorithm implemented in 'huge.generator' in '*huge*' R package. Next to introduce confounding in the data, we simulated a sample specific term from a standard normal distribution, and added a scalar multiple of that to genes 20 genes in the data. To correct the data, we regressed out the first principal component from the confounded data. We used graphical lasso to reconstruct networks using the three versions of the data.

We also simulated 350 observations from a multivariate gaussian distribution that encodes conditional dependencies across 5000 genes - sample and gene numbers similar to those in our empirical experiments. We simulated two sample specific terms, and two gene specific terms to introduce weighted confounding to 1500 genes multiplied by a scalar constant. This confounding data was corrected by regressing 2 PCs (as estimated by the permutation procedure). We used graphical lasso to reconstruct networks with three versions of data.

The code for these simulation examples and network reconstruction can be found at: https://github.com/leekgroup/networks_correction/blob/master/publication_rmd/

Determining sample specific estimate of GC bias

Studies have shown that GC content of genes have significant impact on sequencing read coverage in DNA-seq and RNA-seq experiments. This eventually introduces sample specific biases in expression quantification. To quantify the effect of GC bias, using transcript level fasta files from Gencode v25 we first computed the GC% of each transcript by:

$$GC\%(T) = \frac{(\#G + \#C)}{(\#A + \#T + \#G + \#C)} \tag{3.3}$$

We summarized GC content of genes, by averaging over all transcripts belonging to the gene. Suppose k transcripts were transcribed from gene G_i then,

$$GC\%(G_i) = \frac{\sum_{j=1}^{k} GC\%(T_j)}{k} \tag{3.4}$$

Next using a linear model, we obtain sample specific estimates of GC content of genes:

$$E_i = \mu + \beta_i \times G \tag{3.5}$$

where, E_i is the vector of expression values of all genes in sample i, G is the GC content for each gene and β_i is the estimate of GC bias for sample i.

Network reconstruction using GTEx data

Based on sample size, we used gene expression RNA-seq data from eight tissues in the GTEx project[77] that included whole blood, lung, skeletal muscle, tibial artery, sun-exposed skin, tibial nerve, subcutaneous adipose, and thyroid (Table 3-II). In each tissue we filtered for non-overlapping protein coding genes that had scaled expression (counts scaled by the total coverage of the sample) of at least 0.1 25% of total number of observations. Next, we log2 transformed the scaled gene expression data, and performed the following steps to select the most variable 5000 genes across all tissues, correct gene expression data, and build co-expression networks.

	# of samples
Whole Blood	393
Lung	320
Skeletal Muscle	430
Tibial Artery	332
Sun-exposed skin	356
Tibial Nerve	304
Adipose Subcutaneous	349
Thyroid	323

Table 3-II. Tissue sample size.

(a) Select genes expressed in all five tissues.

(b) For each tissue, assign a rank to each gene by variance, such that the most variable gene is ranked first and least variable gene is ranked in last.

(c) Using the ranked list of genes from five tissues, assign an average rank to each gene across five tissues.

(d) Select the top 5000 genes based on average rank for network inference with WGCNA and graphical lasso.

We used multiple approaches to correct gene expression data from each tissue individually as described below:

- Residuals from RIN/Exonic Rate/ GC bias: Using a linear model, we regressed the RNA integrity number (RIN), exonic rate or sample specific estimate of GC bias on the expression data and computed the residuals

- Residuals from multiple covariate correction: In each tissue individually, we estimated expression percent variance $R2$ explained by the known technical confounders. Next, using a linear model we regressed the technical covariates with $R2 \geq 0.01$ in a tissue and computed the residuals. (Table 3-I)

- Residuals from principal components: For each tissue, principal component based gene expression residuals were computed as described in above.

Prior to reconstructing co-expression networks with WGCNA and graphical lasso, we transformed the uncorrected and corrected expression of each gene to a Gaussian distribution by projecting the expression of each gene to the quantiles of a standard normal.

To reconstruct unsigned weighted co-expression networks with WGCNA, we identified the lowest power for which scale-free fit R^2 between $\log(p(k))$ and $\log(k)$ exceeds 0.85. Here $p(k)$ is the fraction of nodes in the network with at least k neighbors. After that we used the 'blockwisemodules' function in the *WGCNA* CRAN package to perform co-expression module detection at varying cut-heights of hierarchical dendrogram ranging from 0.9 to 1.0. For networks reconstructed with WGCNA, we considered all genes in the same module to be a fully-connected subgraph.

For reconstruction of co-expression networks with graphical lasso, we first computed the gene covariance matrix and then used 'QUIC' function in the QUIC R package to infer co-expression networks with penalization parameter λ ranging from 0.3 to 1.0.

Evaluation of co-expression networks

To evaluate our correction method and its effect on reconstruction of co-expression networks, we used two methods to infer the structure of gene co-expression networks: a) weighted gene co-expression networks (WGCNA)[78] and b) graphical lasso[71] - as described above. Since the underlying network structure is generally unknown, we used a) genes known to be functional in the same pathways and b) known transcription factors and their targets as ground truth to assess these networks.

- ***Canonical pathway databases:*** We downloaded the latest pathway information (2016) from KEGG, Biocarta and Pathway Interaction Database from

Enrichr [79, 80], that were also annotated as canonical pathways by MSigDB [81]. The number of pathways/genesets in each of these databases were:

- KEGG - 293
- Biocarta - 237
- Reactome - 1530
- Pathway Interaction Database - 209

Any pair of genes that have at least one pathway in common were assumed as true functional relationship. An edge that was observed between a pair of genes in the inferred network (from WGCNA or graphical lasso) and was also present in the list of real connections was called as a true positive (TP). We defined false positive (FP) to be an edge that was observed between a pair of genes in the inferred network, however was absent in the list of real connections. Shared true positives: We obtained a refined list of real connections described above by restricting to pairs of genes that were present in at least two pathway databases. All TP, FP and FN were computed with genes restricted to the most variable 5000 genes that were used for reconstructing co-expression networks. We compute false discovery rate as given below:

$$FDR = \frac{FP}{(TP + FP)} \tag{3.6}$$

Results

In this study, we provide a framework for data correction leveraging the structure of scale-free networks. We show that for scale-free networks, principal components of a gene expression matrix can consistently identify components that reflect artifacts in the data rather than network relationships. It has been shown that real world networks including co-expression networks often have scale-free topology, i.e. the node degree distribution of these networks follow a power law[74, 82, 83]. Several studies have employed the assumption of scale-free topology to infer high-dimensional gene co-expression and splicing networks[14, 70].

Latent factor-based data correction has been successfully employed in many applications in genomics from genome-wide association studies, cis- and trans-eQTL mapping, to differential expression analysis[17, 84–87]. In genome-wide association studies investigating the association between genotype and complex traits, it has been shown that the top principal components explain the broad correlation between genotypes which generally reflects population structure rather than a desired functional biological signal of interest[87]. Co-expression analysis is more complicated because confounders affect sets of genes in ways that induce correlation or apparent co-expression. Here, we show mathematically, through simulation and through real data examples that similar to genetic association studies, the broad correlation between gene expression levels in uncorrected data appears to reflect artifacts. We expect that most real co-expression networks are sparse which means that most genes are only connected to a small subset of other genes. We prove that when such networks satisfy the scale-free property, the signals from the network will not be sufficiently broad across genes to influence the latent variable estimates from PCA. Thus, principal components will primarily capture latent confounders, which can then be regressed from the expression data before network reconstruction is performed.

Using a toy and scale-free simulation, we first showed that confounding can introduce false correlations between sets of genes that can mimic co-expression and can lead to false edge discovery during reconstruction of co-expression networks with graphical lasso - sometimes at the expense of losing true connections (Figure3-1). We corrected the confounded simulated data using our PC based approach and reconstructed the network using the residuals. Graphical lasso correctly estimated the network structure obtained from corrected data, which was same as the true network structure that was

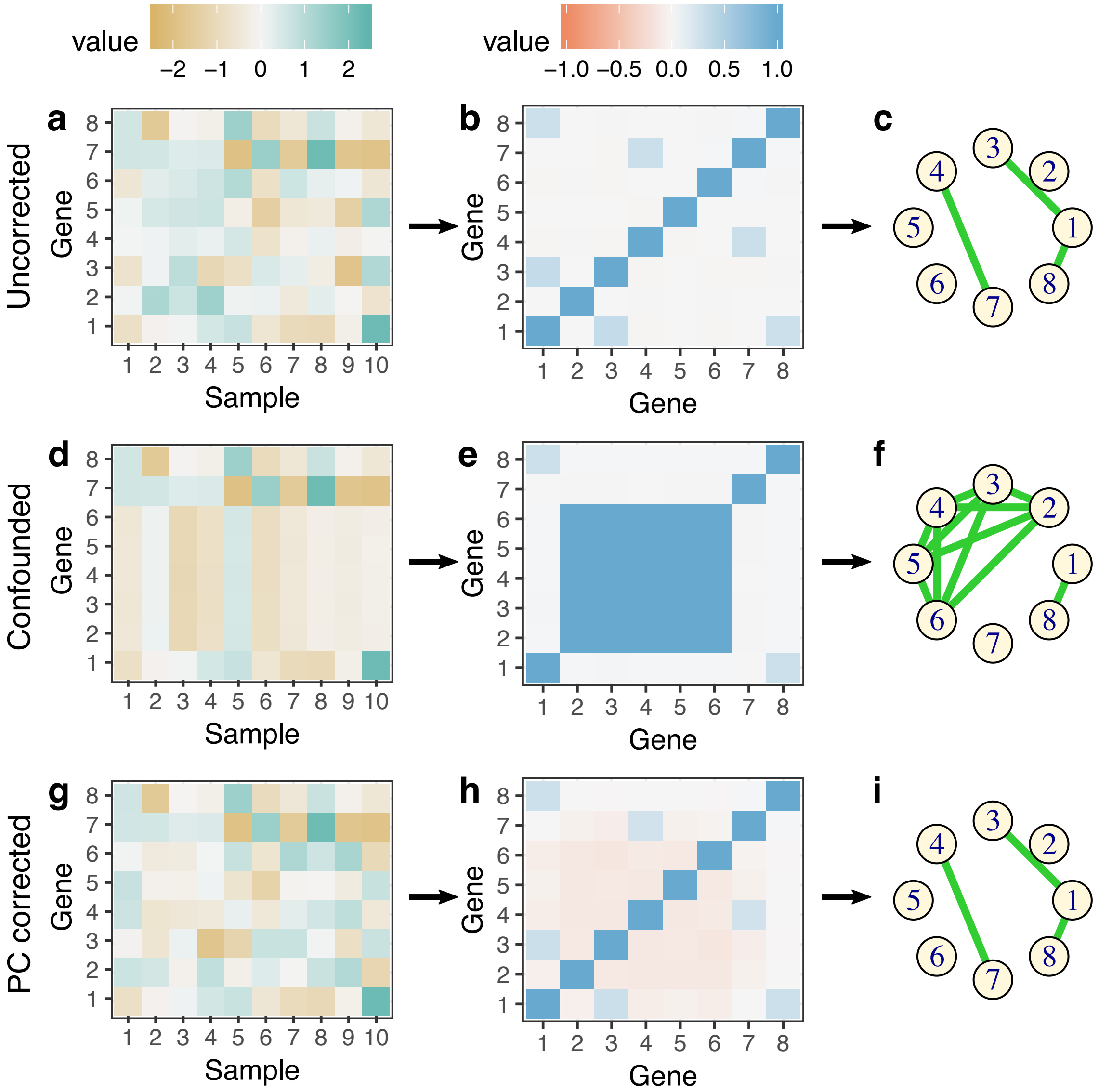


Figure 3-1. Toy Simulation Example. This toy simulation shows the reconstruction of gene co-expression networks is affected by confounders. (g-i) True underlying network structure can be reconstructed after principal component correction of gene expression data as described in the paper

obtained from the original simulated data (Figure3-1). We also simulated multivariate gaussian data with 350 samples and 5000 genes from an underlying scale-free network. Similar to previous simulation, we found that confounding in data can introduce many false positives in reconstructed co-expression networks. We also showed that networks reconstructed with PC corrected data in this setting were more similar to original simulated data compared to confounded data. Throughout our analysis, to estimate the number of principal components to be removed, we used a permutation based scheme[76] as implemented in the sva package[84].

To demonstrate the impact of latent confounders and principal component correction on reconstruction of co-expression networks from real large-scale human gene expression measurements, we applied our method to RNA-seq data from the GTEx project v6p release. We considered data from eight diverse tissues containing between 304 and 430 samples each (Table3-I): Subcutaneous adipose, Lung, Skeletal muscle, Thyroid, Whole blood, Tibial artery, Tibial nerve and Sun-exposed skin. Using the most variable 5000 genes, we reconstructed co-expression networks for each tissue with two popular methods: (a) weighted gene co-expression network analysis[70, 78], and (b) graphical lasso[71, 88]. Since the true underlying co-expression network structure is not known, we assessed the networks using gene pairs annotated to function in the same pathways[79, 80] as ground truth edges. We inferred networks obtained by using a) uncorrected expression data, the residuals after regressing out b) RNA integrity number (RIN)[89], c) exonic rate - a mapping covariate that corresponds to fraction of reads mapped to exons[20], d) sample specific estimate of GC bias, all known to be common confounders in mRNA gene expression data[90–92], and e) residuals from multiple regression model using covariates that explained at least 1 percent of expression variance (adjusted $R^{\geq}$0.01, TableA-II)[89, 93–95].

Co-expression gene modules obtained from weighted signed co-expression networks were interpreted as fully-connected subgraphs. For most tissues, networks obtained from data corrected for latent confounders showed fewer false discoveries compared to those obtained from uncorrected data, or from correcting for individual covariates including RIN, exonic rate (a quality metric from RNA-seq mapping), or sample-specific GC bias (Figure3-2(a-c), A-1,A-6, A-2). Improved performance of networks obtained from PC corrected data was more evident in whole blood, skeletal muscle, tibial artery, tibial nerve, subcutaneous adipose and thyroid. But for some tissues such as lung, PC correction only contributes to moderate improvement on false discovery rates in the reconstructed networks. It is possible that in these cases, the networks may violate the scale-free assumption, or that true signal was already sufficiently strong in the raw data. We also observed that correcting gene expression data with multiple technical

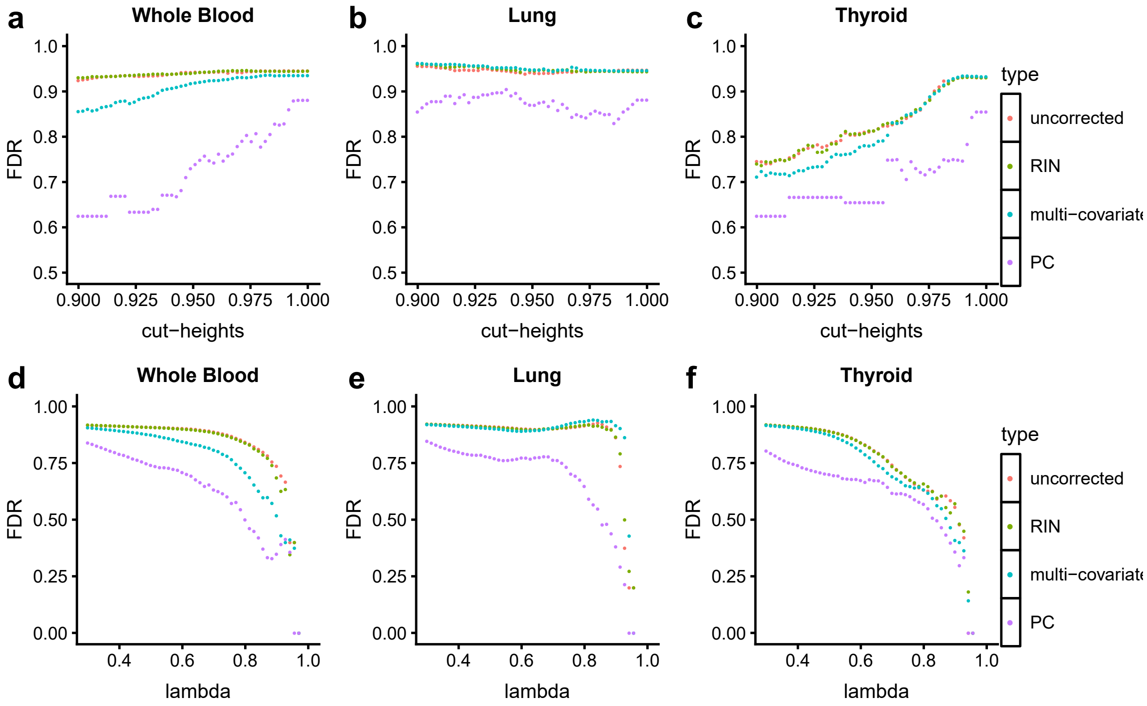


Figure 3-2. False discovery rate of WGCNA modules and graphical lasso networks based on canonical pathways a–c FDR of WGCNA networks obtained at varying cut heights. Each point corresponds to FDR of the network obtained at a specific cut height. Each color represents networks reconstructed with a specific correction approach. d–f Each point in the figure corresponds to false discovery rates of networks obtained at a specific L1 penalty parameter value (lambda) in the graphical lasso. Each color represents networks reconstructed with a specific correction approach—uncorrected, multi-covariate, RIN, and PC corrected.

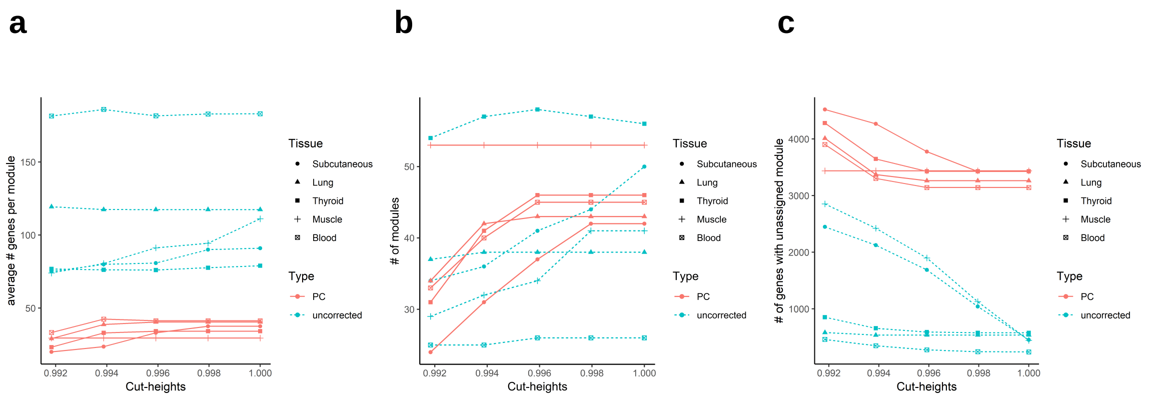


Figure 3-3. Module properties of WGCNA before and after PC correction of gene expression measurements. a) On average the number of genes per module are considerably smaller in WGCNA after PC correction of data b) The number of modules identified are different and varies across tissues. The pattern was inconclusive among PC corrected and uncorrected networks. c) The number of genes assigned to gray module is considerably higher upon PC correction.

covariates (approximately 9 - 17 were used per tissue, TableA-II) sometimes improved reconstruction of co-expression networks obtained by WGCNA (Figure3-2(a-c),A-1). The average WGCNA module size for networks with cut-height greater than 0.99 is smaller with PC corrected data compared to uncorrected counterparts (Figure 3-3). We also observed that the number of genes assigned to gray(unassigned) module in WGCNA was considerably higher in PC corrected networks (Figure 3-3). Finally, we repeated this analysis by varying multiple settings of WGCNA and found that PC corrected showed improvement in most tissues consistently.

In graphical lasso networks, we found that networks estimated with principal component corrected data showed fewer false discoveries compared to networks estimated with uncorrected, RIN corrected or multiple covariates corrected data (Figure 3-2, FigureA-3).

We observed that in general improved performance on false discoveries in PC corrected networks over raw data in whole blood, skeletal muscle, tibial artery and tibial nerve. Compared to raw data, jointly correcting the gene expression data for multiple technical covariates that affect expression measurements also improved reconstruction with graphical lasso in some tissues such as whole blood, thyroid, and tibial artery, while it showed little to no improvement over uncorrected data in lung, muscle, tibial nerve, and sun exposed skin (Figure 3-2, FigureA-3).

However we observed that across all tissues PC correction still shows fewer false discoveries compared to multiple technical covariate based correction. There was no visible improvement in network reconstruction between using uncorrected data and residuals from RIN or exonic rate; thereby suggesting that RIN, exonic rate or GC bias individually is not a sufficient alternative for the wide range of confounding variation found in gene expression data (Figure 3-2,A-3,A-7,A-4). We also found that there was no improvement on false negative rates upon PC correction in networks built with WGCNA or graphical lasso. With both WGCNA and graphical lasso, networks inferred from principal component corrected data were much more sparse than networks from uncorrected, and RIN, exonic rate or GC bias corrected counterparts (Figure 3-4). Further, PC corrected networks from graphical lasso also showed higher clustering coefficient, and fewer hubs compared to others.

Conclusion

Network reconstruction methods are vulnerable to latent confounders present in gene expression data. Co-expression networks obtained from data corrected for effects of RIN, exonic rate, or GC bias individually show little improvement on false discoveries compared to uncorrected data and are not a sufficient surrogate for the diverse sources of confounding variation in gene expression data. With empirical analysis supported by theoretical proof, we show that PC correction is a simple, yet effective approach to address confounding variation for reconstruction of gene co-expression networks. We do note for particularly dense or connected sub-graphs in the underlying biological system

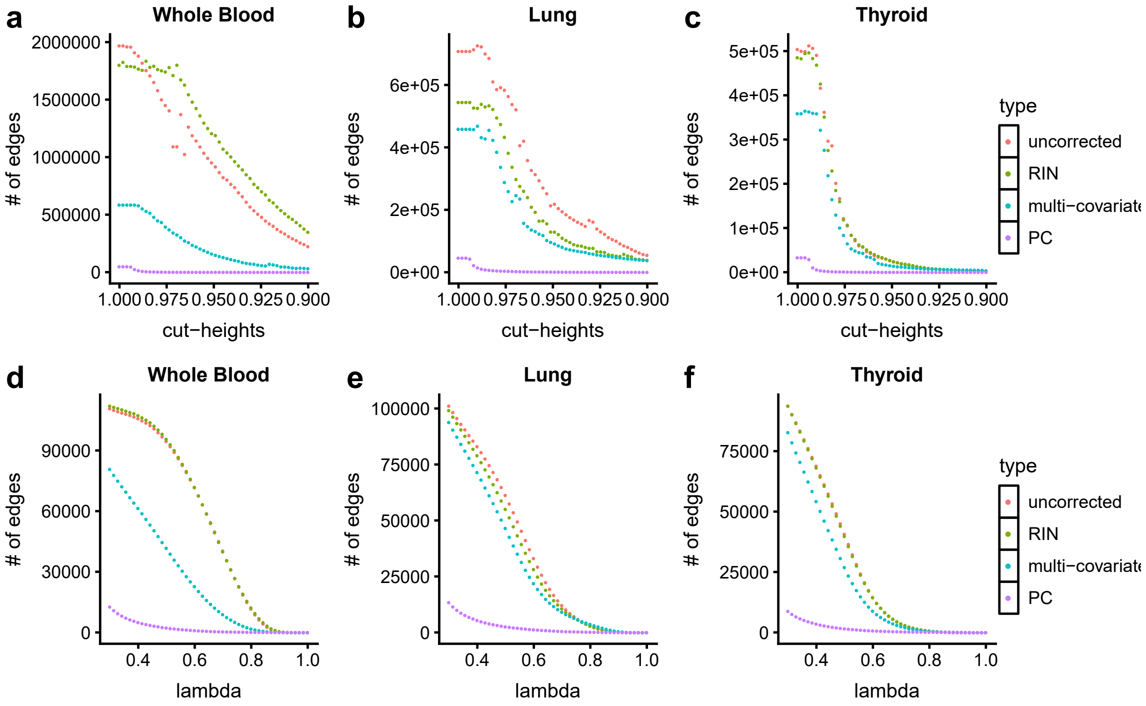


Figure 3-4. Density of the inferred co-expression networks. a-c Each point corresponds to a number of edges in networks inferred by WGCNA at a cut height. d-f Each point corresponds to a number of edges inferred by graphical lasso in networks obtained at a specific L1 penalty parameter value. Networks inferred by PC-corrected data have fewer edges compared to uncorrected or RIN-corrected data

that may not match the scale-free assumption, or when large differences in expression changes are expected (e.g. cancer vs normal), removing principal components may remove biological signal of interest and, as with any data cleaning methodology, should be used with caution. We have implemented our PC correction approach as a function - *sva_network* in sva Bioconductor package which can be used prior to network reconstruction with a range of methods.

In summary, this chapter shows that known and latent confounders introduce biases in the form of false correlation structure in gene expression measurements which leads to a large number of false discoveries during inference of co-expression networks. Therefore, it is critical to correct the gene expression data to remove patterns of artifactual variation. PC residualization of gene expression data can adjust for the effect of confounders, and can reduce false discoveries in reconstruction of gene co-expression networks.

Chapter 4

Multi-study integration to identify global expression patterns and key regulators of Epithelial to Mesenchymal transition (EMT) in cancer

Introduction

Cancer is the second leading cause of death in United States. Metastasis is the leading cause of cancer-related morbidity and mortality[96], but identifying tumors with metastatic potential remains a challenge[97]. Tumor metastasis is a multi-step process in which primary tumor cells disseminate from their site of origin to seed secondary tumors at a distant site[98]. It is believed that in a critical early event in cancer progression, metastatic cancer cells undergo an epithelial to mesenchymal transition (EMT). During EMT, stationary epithelial cells lose cell polarity and transdifferentiate to spindle-shaped motile mesenchymal cells. EMT is a crucial physiologic process involved in early development during embryogenesis and organogenesis. It also plays an important role in tissue regeneration and wound healing. However, in cancer, EMT may contribute to tumor progression and malignant transformation. Several epithelial cancer cells have been described to undergo EMT transform to a more malignant

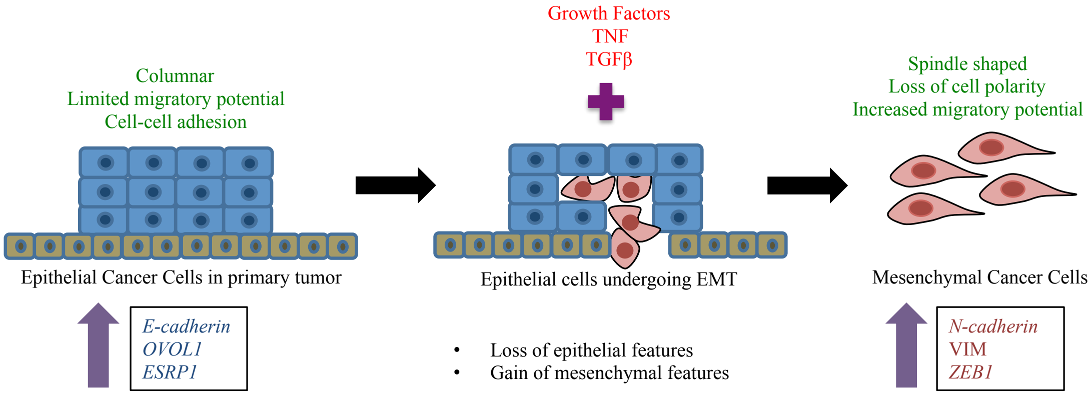

Growth Factors
TNF
TGFβ
Columnar
Limited migratory potential
Cell-cell adhesion
Spindle shaped
Loss of cell polarity
Increased migratory potential
Epithelial Cancer Cells in primary tumor
Epithelial cells undergoing EMT
Mesenchymal Cancer Cells
E-cadherin
OVOL1
ESRP1
• Loss of epithelial features
• Gain of mesenchymal features
N-cadherin
VIM
ZEB1

Figure 4-1. Epithelial to mesenchymal transition. During EMT, non-motile epithelial cells trans-differentiate to mesenchymal cells with increased migratory potential. During this, cells show decreased expression of epithelial specific genes that include *E-cadherin, OVOL1*, and *ESRP1*. At the same time, expression of mesenchymal genes such as *N-cadherin, VIM*, and *ZEB1* increases.

phenotype[99] that can further promote formation of secondary tumors[100]. The role of EMT has been frequently debated in clinical cancer metastasis[101]. However, several in vitro studies have shown that epithelial cancer cells can undergo EMT in response to a combination of signals from the tumor microenvironment [97]. During EMT, cells go through multiple morphological and biochemical changes resulting in loss of epithelial properties coupled with gain of mesenchymal characteristics[102–116]. Microarrays have been widely used to study gene expression patterns of cell populations under different experimental settings, including EMT-inducing conditions (Figure 4-1). While there have been many studies investigating the effect of a gene or pathway in EMT, none have explored the universal changes across multiple cancer tissue types or EMT induction methods. Several gene expression datasets examining EMT in a variety of different cell lines under different conditions are available on open access databases such as Gene Expression Omnibus (GEO)[117]. It has been demonstrated that re-use and aggregation of public gene expression data facilitates discovery of signals too weak to be detected in an individual experiment[118–121]. Gröger et al. performed meta-analysis of 18 EMT gene expression studies and identified 130 core-EMT genes, which were differentially expressed in at least 10 of the 18 studies[122].

Genes such as *TGFB, GNG11, TIMP1, ETS1, S100A14, DPYSL3*, and *C1orf116* that we discovered as differential EMT, were not found in their core EMT gene list. Furthermore, we experimentally validated some of these genes (*S100A14, DPYSL3*, and *C1orf116*) in PC3 epithelial, PC3-EMT and PC3-taxol resistant cell lines confirming their association in EMT. Also, each dataset in [122] was confined by small sample size per class (n<=6). The drawback with under-powered studies are: a) low probability of identifying true effects and b) overestimation of effect size[123, 124]. Therefore, genes that showed consistent moderate effects across datasets could be missed. In contrast, systematic integration of multiple studies promotes reliable detection of consistent gene expression changes that may otherwise be false negatives in results obtained from individual experiments[125]. At the same time, it helps avoid false discoveries that could result from intra-study variability resulting from single experiment. Batch effects and noise introduce spurious signal and correlations in microarray gene expression data [84, 91, 126]. Therefore, data normalization is crucial in order to correct the data for unwanted biological or non-biological effects. However, Groger et al. do not account for batch effects, cross-platform differences, or cross-tissue effects in their meta-analysis study that could potentially lead to false positive findings. In this study, to identify universal EMT genes common across multiple cancer types, we integrated 15 independent gene expression studies representing 12 cell lines (49 epithelial and 46 mesenchymal phenotypes) from 6 cancer tissue types and multiple EMT induction modalities (Table 4-I).

After correcting data to account for cross-study differences, cross-platform differences, and other sources of noise, we performed differential expression analysis and identified global changes in gene expression patterns between epithelial and mesenchymal states (Figure 4-2). Importantly, our candidate gene list was enriched for EMT-related genes and we identified known markers of EMT. In addition, we also identified EMT genes that had only been described in a subset of malignant disease states but were

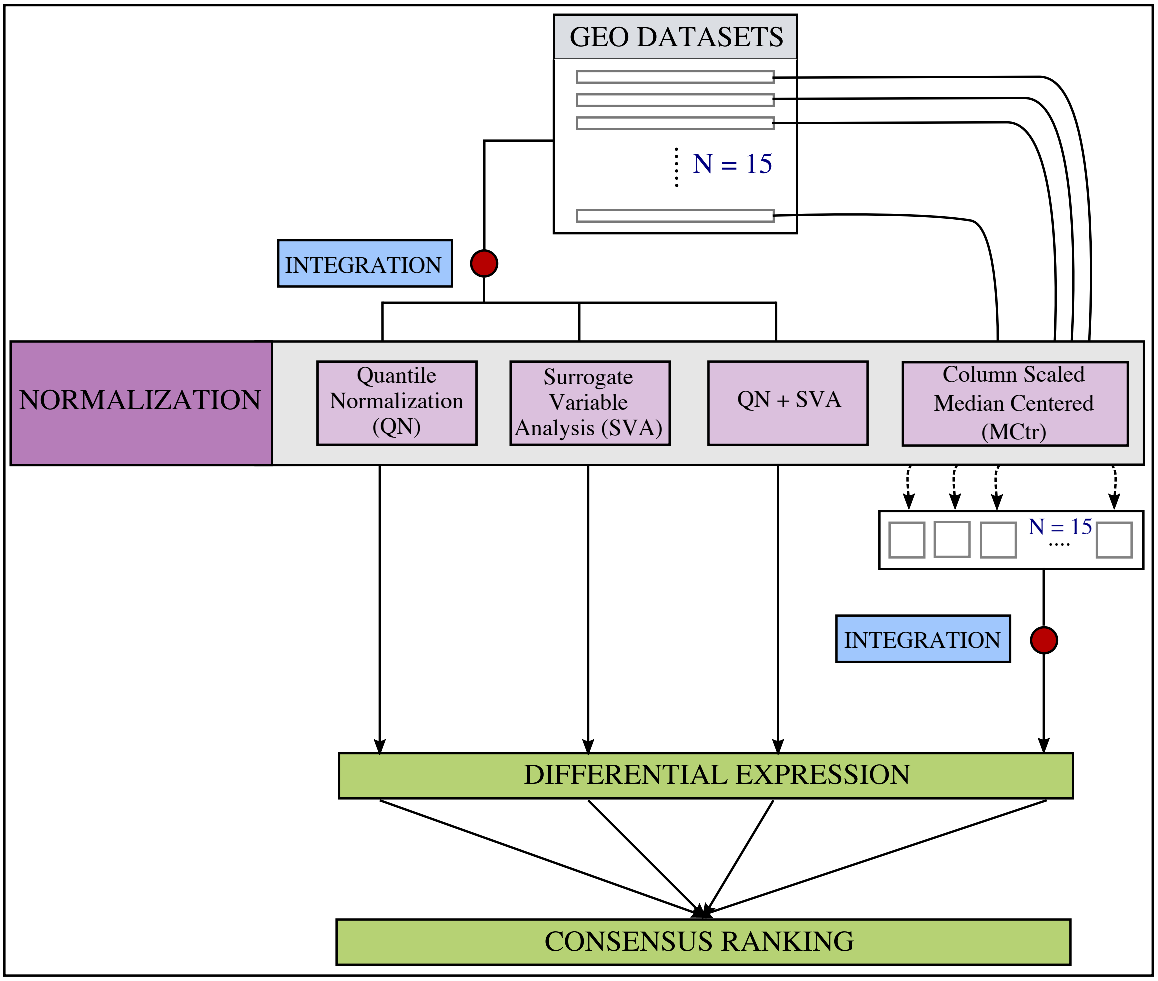


Figure 4-2. Workflow for multi-study data integration, normalization, and identification of candidate universal EMT genes.

previously unknown in prostate cancer (e.g. *LSR, S11A14, DPYSL3*), implying a common EMT program across multiple cancer types. We further identified genes that had not been previously characterized in EMT in any disease state including *C1orf116*, which we then experimentally validated using siRNA knockdown in PC3 epithelial cells. This approach of multi-study integration enabled identification of differential EMT genes universal across different types of cancer. Functional validations of these genes indicate manifestation of molecular mechanisms contributing to EMT shared across disease types. This study also identifies an uncharacterized candidate novel EMT regulator gene *C1orf116*. These findings thereby extend our knowledge and understanding of EMT biology.

Methods

Data Overview

We used 15 published EMT microarray gene expression datasets from GEO (Gene Expression Omnibus) (Table 4-I). This is comprised of 95 observations (45 unique samples and 50 replicates), 49 epithelial and 46 mesenchymal cell lines exposed to different treatment modalities. The cell lines come from 6 different tissue types including breast, prostate, colon, esophageal, liver and retinal pigment and 4 different microarray platforms (8 chips), Affymetrix, Agilent, Stanford Microarray Database (SMD) and Illumina. All the datasets were downloaded in the format they were submitted to GEO. We mapped platform specific probe IDs to Ensembl IDs and gene symbols. When multiple probes mapped to same gene, we used median values to represent expression of that gene. We used 7276 genes common across all datasets.

Data Normalization

This work combined data from multiple studies spanning diverse cell lines and different platforms. Batch effects and noise are inherent in gene expression data. To account for confounders in data as a result of cross-study and cross-platform effects, we used multiple correction methods, such as quantile normalization (QN), Surrogate Variable Analysis (SVA), quantile normalization followed by SVA, and Column Standardized Median Centered (MCtr). We merged all 15 datasets into one matrix prior to quantile normalization and SVA. For MCtr, we individually processed each study and combined them after normalization.

Quantile Normalization: Quantile normalization makes the gene expression distribution of each sample in the dataset the same. Given a dataset D with g genes and n samples, this process:

1. Sorts each column in D
2. Computes the mean for each row and assigns it to each element in the row giving D'
3. Rearranges columns in D' such that it has the same ordering as original D, thus giving normalized data, $D_normalized$

At the end of this, each column in D has the same distribution [127].

Surrogate Variable Analysis: Surrogate variable analysis allows us to preserve the phenotype signal of interest (epithelial and mesenchymal). It estimates known and hidden confounding factors using Singular Value Decomposition on residual variation matrix. We regress out estimated surrogate variables from gene expression data to get SVA normalized gene expression [84]. We also quantile normalize combined data followed by SVA to correct for hidden confounders.

Column standardized Median centered: Samples from each study are standardized and median centered by gene as described in [128] and combined them.

Differential Expression Analyses and concordance between normalization methods

With each of the normalized dataset, we used a two-sample t-test to identify differentially expressed genes between epithelial and mesenchymal states. Assuming equal variance, we compared the mean expression of a gene between the two populations. For each gene, we tested:

$$H_0 : \mu_e = \mu_m \tag{4.1}$$

$$H_1 : \mu_e \neq \mu_m \tag{4.2}$$

We ranked genes by raw p-values and applied Bonferroni correction for multiple hypothesis testing. To test concordance between normalization methods, we used Spearman's rank correlation to test association between gene ranks ($n = 7276$) obtained by different correction methods. Assuming equal probability of error for each normalization method, we computed average rank for each gene across the four methods that represented the consensus position of each gene according to the differential expression test statistic.

Cluster evaluation of normalized data

To evaluate if normalization improved overall grouping of epithelial and mesenchymal phenotypes together, we clustered each of the normalized data using hierarchical clustering (with all 7276 genes). Next, to evaluate grouping we used Baker-Hubert Gamma index for cluster evaluation. Baker Hubert's Index (BH) [129] is an adaptation of Goodman and Kruskal gamma statistic in the context of clustering.

$$BH = \frac{S^+ - S^-}{S^+ + S^-} \tag{4.3}$$

Here S^+ is the number of concordant quadruples and S^- is the number of disconcordant quadruples. To compute BH, it tests all possible quadruples in the input.

Suppose we were testing quadruple samples a, b, c, d. And $d(a, b)$ is the distance between samples a and b. A quadruple is concordant if it fulfills one of the following two conditions:

- $d(a, b) > d(c, d)$; And c and d are in same cluster and a and b are in different clusters

- $d(c, d) > d(a, b)$; And a and b are in same cluster and c and d are in different clusters

A quadruple is disconcordant if:

- $d(a, b) > d(c, d)$; And a and b are in same cluster and c and d are in different clusters
- $d(c, d) > d(a, b)$; And c and d are in same cluster and a and b are in different clusters

Since we were interesting in improvement in grouping of epithelial and mesenchymal samples, we used known phenotype vector as cluster assignment for evaluation.

Gene co-expression module detection using WGCNA

With 200 DE genes from QN+SVA data, unsigned co-expression network was constructed using the WGCNA package in R[78]. Since we used differentially expressed genes, prior to constructing networks, the effect of phenotype (epithelial and mesenchymal) from each gene was removed using a linear model.

RT-qPCR

RNA was isolated from cells at 80% confluency using RNeasy kit (Qiagen) and subsequent cDNA libraries were prepared using Bio-Rad cDNA synthesis kit. TaqMan gene expression assays were used to determine mRNA expression levels using the following probes: β -actin Hs_1060665_g1, LSR Hs01076319_g1, S100A14 Hs04189107, DPYSL3 Hs00181665_m1, C1orf116 Hs00539900_g1, OVOL1 Hs00970334, CDH1 Hs01023894, CDH2 Hs00983056_m1, ZEB1 Hs00232783_m1. Relative Expression Calculations: In the qPCR, the target of interest in each sample is measured using at least three biological replicates. The Ct value for each biological replicate is calculated as an average of three technical replicates. Then the Ct value of each biological replicate is normalized to β -actin by subtracting it from the corresponding Ct value of β -actin ($-\Delta$ Ct). The two groups of interest are compared using a Student's t-test.

The values plotted in the graph are the average of the base 2 anti-log transformations of $-\Delta$ Ct for the biological replicates of interest divided by the average of the base 2 anti-log of $-\Delta$ Ct for the reference group. The standard errors of the mean are determined from biological replicates.

Western Blot

Protein extracts were prepared using Frackleton-lysis buffer with protease inhibitors (Thermo Scientific 78410), and samples were electrophoresed on $4-15\%$ SDS-PAGE (Bio-Rad), transferred to a nitrocellulose membrane and blocked with casein blocking buffer (Sigma B6429). The list of antibodies used for western blotting is in Table B-I. The Licor Odyssey fluorescence scanner was used for visualizing the westerns.

siRNA knockdown of C1orf116

C1orf116 siRNA (ThermoFisher, cat#: 4392420) with RNAiMAX transfection reagent (ThermoFisher) was used for siRNA transfections. Some alterations were made to manufacturer's recommended protocol. Cells were seeded at a density result in 50% confluency the following day. Using a 6 well plate, 9 ul of RNAiMAX reagent and 3 ul (30 pmol) of siRNA (each diluted in 150 ul of Opti-MEM media) was added to each well the day after seeding. 72 hours later RNA was isolated (Qiagen, Rneasy mini kit) from plates and gene expression was analyzed.

C1orf116 expression in cancer patient data

We identified publicly available published cancer patient (breast, prostate, esophageal, liver, colorectal, and lung) gene expression studies with at least 150 patients on Oncomine[130]. Gene expression data for studies (GSE17536[131], GSE11121[132], GSE25066[133], GSE22358[134], GSE7390[135], GSE68465[136], GSE31210[137], and GSE21034[138]) available on GEO were obtained using the GEOquery R package[139].

Probeset IDs corresponding to *C1orf116* were used. Gene level expression was obtained by aggregating multiple probe expression values with median. Wilcoxon rank sum test was used to test association between expression of *C1orf116* and grade, smoking status and cancer sample site. We also looked at association between tumor grade and *C1orf116* expression in 4 breast cancer, 1 colorectal cancer and 1 lung cancer studies from Oncomine. We adjust Wilcoxon rank sum p-values with bonferroni correction for a total of 23 tests performed for clinical associations.

Results

We identified publically available gene expression microarray datasets that queried gene expression of cell lines induced to undergo EMT[102–116]. We confirmed the phenotype of the samples by referring to associated publications for immunohistochemistry staining and/or protein expression of known epithelial or mesenchymal markers (Table 4-I). 95 cell line observations (45 unique samples and 50 replicates) from 15 datasets that showed sufficient evidence of correct phenotypic labeling included 49 cell lines of epithelial phenotype and 46 cell lines of mesenchymal phenotype.

Normalization methods show consistency in signal

Technical variability in the form of noise and batch-effects is inherent in gene expression data. We performed rigorous confounding factor correction to make gene expression comparisons between epithelial and mesenchymal samples that came from different studies, platforms, and cell lines. We used standard normalization methods including column standardized mean centered (MCtr)[128] and Quantile Normalization (QN) [127] and more rigorous methods that included Surrogate Variable Analysis (SVA) [84] and combination of QN followed by SVA (QN+SVA). With each normalization method (MCtr, QN, SVA, QN+SVA), we compared the mean expression of epithelial and mesenchymal cell lines by a two-sample t-test for differential expression. We

GEO ID	Platform ID	Disease Type	Cell line	Samples*
GSE12811	GPL7319	Breast	MCF10A	3
GSE13915	GPL7785	Breast	BT549, EFM19	4
GSE18070	GPL570	Breast	MCF10CA1h	9
GSE28569	GPL6480	Breast	MCF10A	8
GSE39356	GPL6480	Breast	MCF-7	4
GSE8240	GPL3921	Breast	MCF10A	11
GSE12203	GPL2700	Colon	Caco-2	4
GSE14773	GPL570	Colon	HT29, SW480	8
GSE27424	GPL570	Esophageal	EPc2-hTERT	12
GSE26391	GPL6244	Liver	HCC-1.1, HCC-1.2	8
GSE14405	GPL570	Prostate	PC3, TEM4, TEM2	6
GSE22010	GPL6244	Prostate	PrEC-hTERT	2
GSE22764	GPL6884	Prostate	PC3	6
GSE43489	GPL570	Prostate	PC3	4
GSE12548	GPL570	Retinal pigment	ARPE19	6

Table 4-I. Dataset Information.

evaluated concordance among normalization methods to determine signal robustness – any individual method may be subject to false positives due to different patterns such as outliers, batch effects, etc. For this, we restricted our analysis to 7276 genes that were common across all studies. We used spearman correlation to test association between raw test statistics (n = 7276 genes) obtained from two-sample t-test from each of type of normalized data. Test-statistic distributions from individual normalization methods were significantly correlated with each other (p-value $< 2.2e{-}16$, n=7276). This indicates that signal produced by data normalized using a particular method is consistent with others (Figure 4-3). Next, to assess if normalization improved overall grouping of epithelial and mesenchymal phenotypes together, we clustered samples from each of the normalized datasets using hierarchical clustering (using all 7276 genes). Next, to evaluate this grouping we used the Baker Hubert Index (BH) with known phenotype vector as group assignments. Values of the BH index range from -1 to 1, with larger values indicating better grouping[140]. Table 4-II shows that grouping of samples by phenotype (epithelial or mesenchymal) is considerably

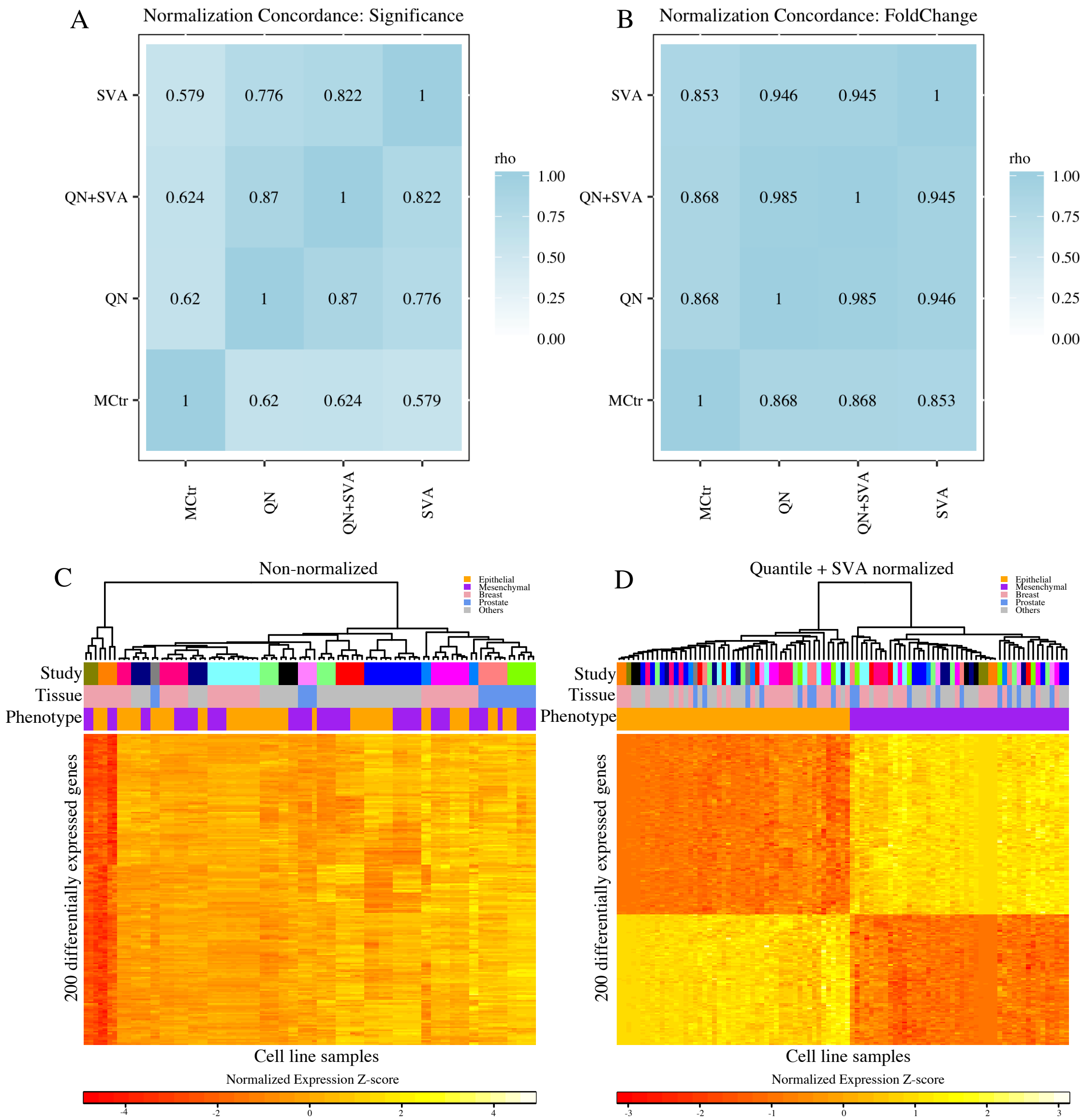


Figure 4-3. Consistency in differential expression signal across normalization methods. A. Correlation heatmap showing concordance (Spearman rho) among ranks of differentially expressed genes using the four normalization methods (n=7276). Genes were ranked by raw t-test p-values. B. Correlation heatmap showing concordance (Spearman rho) among fold-change of differentially expressed genes using the four normalization methods (n=7276). C. Hierarchical Clustering of top 200 differentially expressed genes with uncorrected data shows strong clustering of samples by study rather than by phenotype. D. Hierarchical Clustering of top 200 differentially expressed genes with QN + SVA (Quantile Normalized + SVA) corrected data clusters by epithelial and mesenchymal phenotype.

improved in normalized datasets in comparison to non-normalized data. QN + SVA performs the best, followed by SVA, MCtr and QN.

	No normalization	Quantile Normalization(QN)	Surrogate Variable Analysis (SVA)	QN + SVA	Median Centered Column Scaled
Baker Hubert Index	0.0001	0.047	0.864	0.7995	0.0705

Table 4-II. Evaluation of sample grouping (with 7276 genes) using Baker Hubert index and phenotype information.

Differential expression analyses reveal universal EMT genes across multiple carcinoma types

With every form of normalized data (MCtr, QN, SVA, QN+SVA), we determined differentially expressed genes between epithelial and mesenchymal cell phenotype by a two-sample t-test. A gene list ranked by raw p-values from the t-test was generated for each normalization method. Assuming equal likelihood of error in correction methods (Figure 4-2), for each gene we assigned a differential rank that was the average of p-value ranks from all four normalization methods. This was used to generate a final integrated ranked gene list. We defined a candidate universal EMT gene list by the top 200 genes from the integrated gene list (absolute fold change > 1.2 and FDR < 0.005 in SVA, QN + SVA and MCtr normalized data) (Table 4-IV). These genes are representative of global differential EMT patterns independent of cell line origin and treatment modality.

Cancer cells recruit developmental pathways and processes to acquire migratory and invasive properties. To determine if the candidate gene list contained groups of genes working together and shared common biological functions we tested enrichment it's enrichment for Hallmark genesets (MSigDB) defined and curated by the Broad Institute [141] using a right-tailed Fisher's exact test. The most significantly enriched gene set was epithelial to mesenchymal transition (Odds ratio = 18.3575636, FDR =

4.92E-31). Among the other hallmark gene sets, we found increased representation (FDR < 10%) of several EMT related pathways including estrogen responsive genes (early and late), genes upregulated in response to low oxygen levels (hypoxia) and others [5,49–56] (Table 4-III). We also found that specific estrogen responsive genes (early and late) were differentially expressed even when restricted just to the prostate cancer samples (Supplementary Figure B-3) indicating this enrichment was not due exclusively to breast cancer cell lines in our combined analysis. When tested for GO biological processes, we found enrichment (FDR < 10%) for several developmental terms including epidermis development, anatomical structure morphogenesis and organ development. This further confirms that our analyses capture comprehensive signals in identifying changes in gene expression patterns across cancer types during EMT.

Among genes on our candidate gene list, we found known epithelial- and mesenchymal-specific genes such as E-cadherin (CDH1), Zinc Finger E-Box Binding Homeobox 1 (ZEB1), Vimentin (VIM), Transforming Growth Factor, Beta 1 (TGFB1), Tissue

Geneset	p-values	oddsratio	FDR	Genes in set
HALLMARK Epithelial mesenchymal transition	9.84E-33	18.3575636	4.92E-31	*CD59, CDH11, CDH2, COL1A1, COL1A2, COL4A2, COL5A1, COL6A3, CTGF, CYR61, DAB2,DPYSL3, EDIL3, EMP3, ENO2, FAP, FBN1, FBN2, FERMT2, GEM, GJA1, GREM1, LGALS1, LOX, MMP14, MMP2, PCOLCE,PCOLCE2, PLAUR, PLOD1, PMP22, POSTN, SERPINE1, SERPINE2, SLIT2, SPARC,* *SPOCK1, TGFB1, TIMP1, VCAN, VIM, WNT5A*
HALLMARK Estrogen response late	9.36E-06	4.332224532	0.00019652	ALDH3A2, ASS1, CDH1, CELSR2, LLGL2, LSR, MAPK13, PLXNB1, RAPGEFL1, SCNN1A, SLC22A5, SLC27A2, ST14, TOB1, TRIM29
HALLMARK Apical junction	1.18E-05	4.516129032	0.00019652	AKT3, CDH1, CDH11, CLDN7, FBN1, GRB7, JAM3, JUP, MAPK13, MMP2, MPZL2, PVRL3, SLIT2, VCAN
HALLMARK UV response dn	8.16E-05	4.23768997	0.001019448	AKT3, COL1A1, COL1A2, CYR61, DAB2, FZD2, GJA1, HAS2, KCNMA1, MAP1B, PMP22, SERPINE1
HALLMARK Estrogen response early	0.000247578	3.495078664	0.002475779	AQP3, CELSR2, CLDN7, ELF3, GJA1, KRT15, PMAIP1, RAPGEFL1, SCNN1A, SLC22A5, SLC27A2, TOB1, WWC1
HALLMARK Hypoxia	0.000436298	3.276838008	0.003635818	AKAP12, CHST2, COL5A1, CTGF, CYR61, ENO2, ETS1, HMOX1, KDELR3, LOX, PLAUR, SERPINE1, SRPX
HALLMARK Inflammatory response	0.000679488	3.786760716	0.004246802	CD70, CHST2, EMP3, FZD5, HAS2, HRH1, MMP14, PLAUR, SERPINE1, TIMP1
HALLMARK KRAS signaling up	0.00061698	3.554348835	0.004246802	AKAP12, EPB41L3, ETS1, GFPT2, GNG11, JUP, MAP7, MPZL2, PLAUR, TMEM158, TRIB2
HALLMARK Angiogenesis	0.003822541	7.2	0.02123634	JAG2, POSTN, TIMP1, VCAN
HALLMARK Complement	0.00451196	3.068992514	0.022559801	CD59, COL4A2, CTSD, MMP14, PLAUR, SERPINE1, TIMP1, TIMP2, ZEB1
HALLMARK Myogenesis	0.00594623	2.929880329	0.027028319	COL1A1, COL4A2, COL6A3, ERBB3, MEF2C, NCAM1, PDLIM7, SPARC, TGFB1
HALLMARK TGF beta signaling	0.010673511	4.097902098	0.044472964	BCAR3, CDH1, SERPINE1, SMURF2, TGFB1

Table 4-III. Enriched MsigDB Hallmark genesets.

Inhibitor Of Metalloproteinase 1 (TIMP1)[100, 142], N-cadherin (CDH2) (Table 4-II). We also observed enrichment of collagen genes that are known to be associated with cell adhesion and migration amongst DE genes (Fisher's exact p-value 1.124e-05) [49]. In addition, we also found known EMT related transcription factors such as ZEB1, ETS1 and LSR in our candidate gene list. We also compared our list of genes to the core EMT gene signature described by Groger et. al. [122]. We found 43 common genes from their study (Supplementary Table 4-III). These included genes such as CDH1, CDH2, VIM, LSR and some collagen genes. Several known EMT genes such as TGFB, TIMP1, ETS1 that were found in universal EMT genes were missing from their list. Some other genes such as S100A14, DPYSL3 and C1orf116 (Supplementary Figure B-1,B-2) that we validate as differential EMT genes in our study, were also not found in their core gene list.

Candidate gene list identified genes previously unknown in prostate cancer EMT

In addition to genes well established in the process of EMT, we also identified genes that had only been described in EMT in a subset of cancer types, including two epithelial specific genes, lipolysis stimulated lipoprotein receptor (LSR) and S100 calcium binding protein A14 (S100A14), and one mesenchymal specific gene, dihydropyrimidinase-like 3 (DPYSL3). Previous studies have investigated role of LSR in breast cancer EMT [143], and S100A14 has been examined in pancreatic and cervical cancer [144, 145]. Previous studies have indicated involvement of DPYSL3 in malignant pancreatic and gastric tumors[146, 147]. We validated the expression of these genes in an in vitro model of prostate cancer EMT. mRNA and protein expression levels of these genes were determined in one epithelial and two mesenchymal prostate cancer cell line PC3 derivatives. PC3-Epi is an expansion of a highly epithelial clone from the parental PC3 population. The mesenchymal derivatives were generated from PC3 cells by M2

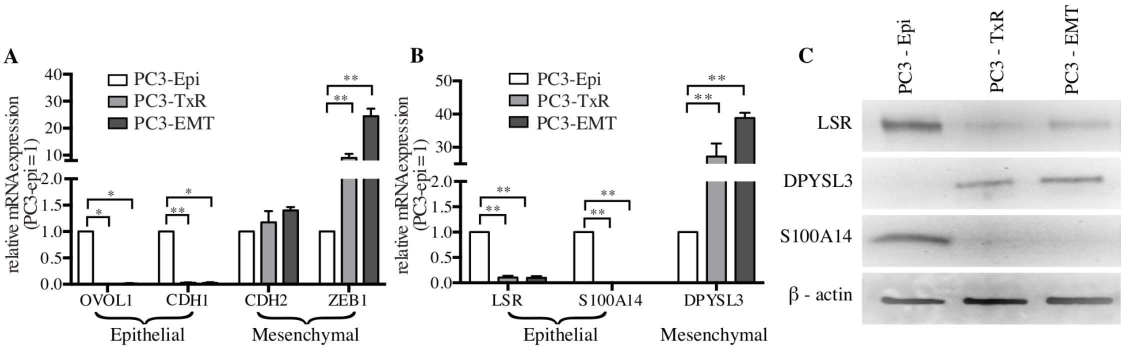


Figure 4-4. Expression of EMT associated genes in prostate cancer EMT.

macrophage co-cultures (PC3-EMT) and Taxol treatment and subsequent resistance (PC3-TxR) [115, 148]. RT-qPCR of canonical epithelial and mesenchymal genes, OVOL1, OVOL2, CDH1, ZEB1, and CDH2, confirmed the appropriate phenotypic states for these cells lines (Figure 4-4A). Elevated levels of S100A14 mRNA was observed in PC3-Epi compared to mesenchymal PC3-EMT and PC3-TxR. Similarly, mRNA expression of epithelial gene LSR was found to be higher in PC3-Epi than in its mesenchymal counterparts, PC3-EMT and PC3-TxR (Figure 4-4B). Conversely, the mesenchymal gene DPYSL3 was extremely upregulated in PC3-EMT and PC3-TxR than in PC3-Epi (Figure 4-4B). These results were supported by western blot analysis, which demonstrated protein levels mirrored the mRNA expression (Figure 4-4C).

C1orf116 was discovered to be a novel EMT regulator

Our candidate gene list also contained genes that have not been previously described as related to the EMT process in any cancer type or in any physiologic process. One of these novel candidate EMT genes, C1orf116 (also known as SARG), is a poorly characterized gene with only one PubMed listed publication[149]. We first validated our finding from microarray data using the PC3 in vitro model of EMT and found increased mRNA expression in PC3-Epi cells compared to PC3-emt (1.3 fold) and PC3-TxR (8.8 fold). These results were supported by elevated protein expression of

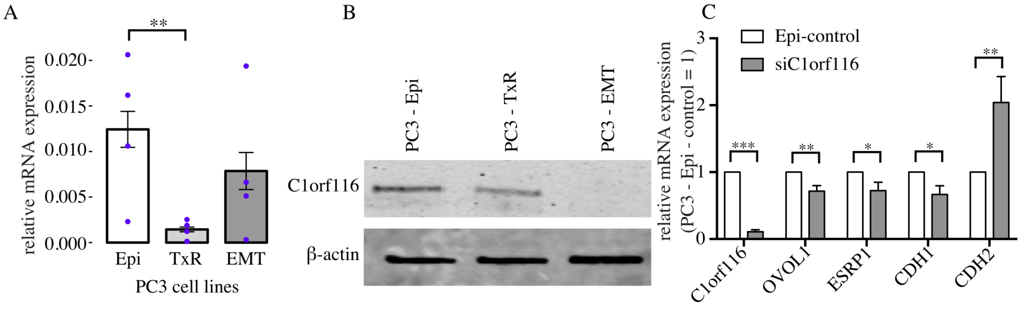


Figure 4-5. C1orf116: a novel EMT regulator. A. qPCR: mRNA expression of C1orf116 in EMT model prostate cancer cell lines PC3-Epi, PC3-EMT and PC3-TxR $*P < 0.1; **P < 0.05; ***P < 0.005$ B. Immunoblot: Protein expression of C1orf116 in EMT model prostate cancer cell lines PC3-Epi, PC3-EMT and PC3-TxR (LSR, DPYSL3, S100A14, C1orf116, and β-actin were all probed on the same blot, so the B-actin loading control is appropriate for both figure 4-4C (LSR, DPYSL2, S100A14) and figure 5B (C1orf116). Data were separated into two figures for clarity.) C. qPCR: mRNA expression of C1orf116 and other known epithelial (OVOL1, ESRP1 and CDH1) and mesenchymal (CDH2) gene in PC3-Epi cells transfected with C1orf116-siRNA relative to empty vector control $*P < 0.1; **P < 0.05; ***P < 0.005$

C1orf116 in PC3-epi cells (Figure 4-5A-B).

Increased expression *C1orf116* in epithelial cells confirmed of it as an epithelial marker gene. We applied gene network analysis [78], that revealed weighted coexpression gene modules (groups of co-expressed genes) and showed that *C1orf116* clustered with other epithelial genes including *CDH1*, *LSR*, *S100A14* and others (Figure 4-6). *LSR* and *S100A14* were among the known-unknown genes whose expression was validated in PC3 cell lines. This confirmed its association with other epithelial genes universal across other disease types. Through manual literature search, we identified that a subset of the *C1orf116* module gene list have been shown to be associated with multiple cancer types. Among other genes in the modules, *SH2D3A, AP1M2, CDS1* and *SCNN1A* haven't been previously studied in cancer biology. This shows that in addition to being a novel EMT regulator in prostate cancer, *C1orf116* could have broad effects across multiple cancer types. Next, we interrogated the possible role of C1orf116 in in vivo malignant progression. For this, we identified gene expression

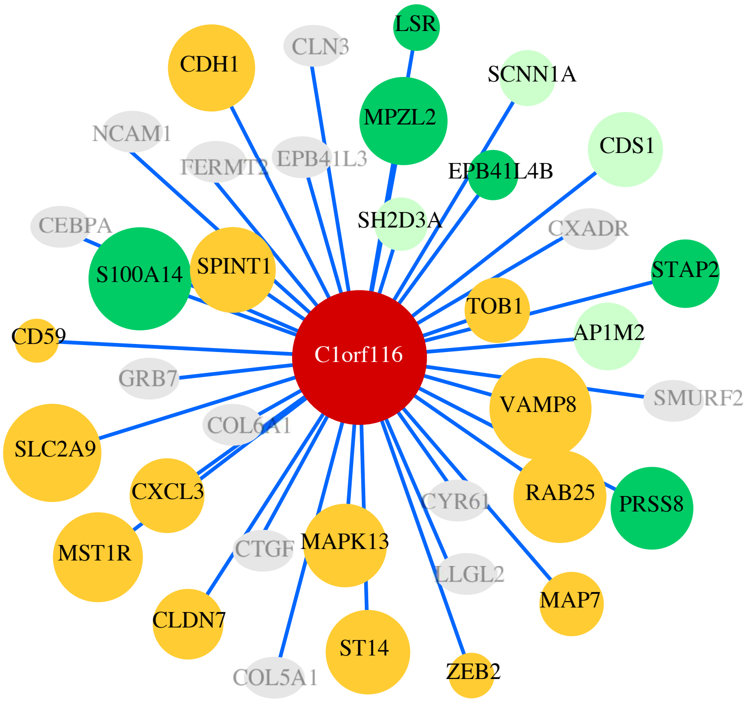


Figure 4-6. *C1orf116* associated genes in weighted gene correlation network module. This correlation network shows association of C1orf116 module genes obtained from WGCNA. Node size is a function of correlation with C1orf116 expression. Yellow nodes represent genes that have been previously studied in multiple (greater than 3) cancer types. Bright green nodes are the genes that have been studied in 3 or less cancer types. Light green nodes are genes that have not been specifically studied in cancer. Gray nodes were genes that were not significantly associated with expression of C1orf116.

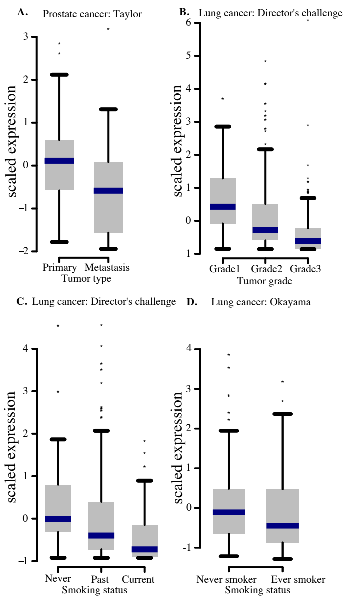


Figure 4-7. ***C1orf116*** **expression in cancer patient data.** A. Decreased expression of *C1orf116* is seen in metastatic tumor type compared to primary prostate cancer; unadjusted P = 0.0340, Bonferroni adjusted P = 0.51 B. Expression of C1orf116 decreases in high grade lung cancer; Bonferroni adjusted P < 0.0005 C. C1orf116 is downregulated in lung cancer patients with increased smoking habits; unadjusted P < 0.01, Bonferroni adjusted P < 0.1 D. C1orf116 is downregulated in lung cancer patients with smoking habits in comparison to non-smokers; unadjusted P = 0.0586, Bonferroni corrected P = 0.879

studies with at least 150 patients that also had information on tumor grade and expression data for C1orf116 and were able to find breast, prostate, colorectal and lung cohorts (Supplementary figure B-4). We found that C1orf116 expression is decreased in metastatic lesions compared to localized tumors in prostate cancer patients (Figure 4-7 A) [137]. Likewise, C1orf116 expression decreased with increasing cancer grade in patients with lung cancer (Figure 4-7 B) [135]. Studies have shown that lung cancer patients with history of smoking tobacco/cigarette exhibit lower expression levels of E-cadherin and higher levels of mesenchymal markers such as vimentin [150, 151]. Previous studies have also indicated that cigarette smoking can induce EMT in non-small cell lung cancer[152]. Analogous to these findings, we observed reduced expression of C1orf116 among lung cancer patients with smoking habits (Figure 4-7C-D)[135, 136]. In some breast cancer datasets expression of *C1orf116* increased with increasing cancer grade (Supplementary Figure B-4). This suggested that in addition to expression changes in in vitro cell line models, changes in *C1orf116* expression could potentially have a functional role in clinically-important disease progression in cancer patients.

To test the role of *C1orf116* as a driver of an epithelial phenotype, we used siRNA-mediated knockdown of the gene in PC3-Epi cells. We found that siRNA-mediated knockdown of *C1orf116* expression resulted in decreased expression of epithelial markers *OVOL1, ESRP1,* and *CDH1*, and increased expression of mesenchymal marker *CDH2* (Figure 4-5C). This suggests that C1orf116 plays a functional role in maintaining epithelial phenotype. Significant upregulation of mesenchymal genes in response to *C1orf116* knockdown indicates it as a novel regulator of EMT.

Test group	Wilcoxon rank sum p-value	Bonferroni adjusted p-value
Lung cancer (Director's Lung Challenge): grade [43]		
Grade1 vs Grade 2	1.4191e-06	3.27E-05
Grade 2 vs Grade 3	1.1481e-10	2.65E-09
Grade 1 vs Grade 3	2.6121e-17	6.00E-16
Lung cancer (Director's Lung Challenge): Smoking Status [43]		
Never vs Past	0.006	1.38E-01
Past vs Current	0.006	1.38E-01
Never vs Current	0.0002	4.60E-03
Lung cancer (Okayama): Smoking status [44]		
Never smoker vs ever smoker	0.0586	1E+00
Prostate cancer (Taylor): Tumor type [45]		
Primary vs Metastatic	0.0340	7.82E-01

Table 4-IV. textbfAssociation of *C1orf116* expression in lung and prostate cancer patients.

Discussion

EMT may be an early step in cancer metastasis and has been associated with chemoresistance and disease progression[153, 154]. Though EMT is common among all solid tumor types and is essential in early development, common drivers of EMT across multiple cancer types have not been described. Several studies have investigated EMT in cell lines from within a single disease type. However, most of these studies have been confined to very small sample size. To address this, we systematically integrate multiple EMT studies to increase power and identify novel drivers of EMT universal to all cancer types. A significant challenge in multi-study analysis comes from various sources of heterogeneity arising from study specific technical and biological variation. Biological variation interferes with analyses, especially when it is not the signal of interest. We employed two strategies to address various sources of heterogeneity and noise. First, we chose stringent normalization methods that have been shown to reduce the influence of such heterogeneity (SVA, quantile normalization, and scaled median centering). We recognize that these methods may have their failure modes and limitations. Therefore, we defined our final differentially expressed gene list from consensus ranking across all four normalization schemes. Thus even if a single method introduced an error or failed to account for a particular effect, the final gene list may be more robust than results from any individual method. However, technical variation and experimental heterogeneity may still influence the results of our analysis, as no method has been shown to fully remove such effects from expression data. Therefore, experimental validation and comparison with external functional annotation were important. Integrating across multiple studies did improve power and helped us detect novel genes that showed consistent effect across multiple studies, which could be concealed in a single study. We found three groups of genes in the EMT differentially expressed list: a) known EMT genes (e.g. CDH1, ZEB1, TGFB, CDH2, VIM, TIMP1), b) EMT genes previously unknown in prostate cancer (LSR, S100A14, DPYSL3) and c) novel EMT genes (including C1orf116). We confirmed our discovery of unknown EMT genes in prostate cancer by testing expression of LSR, S100A14, and DPYSL3 in a PC3 prostate cancer cell line model of EMT. Previous studies have shown that LSR suppresses EMT phenotype in claudin-low breast cancer cell lines[143]. S100A14 has been studied in breast cancer progression and is showed to be involved in EMT in human cervical and pancreatic cancer cells[144, 145, 155]. DPYSL3 is associated with malignant gastric and pancreatic tumors [146, 147]. Moreover studies suggest that mRNA expression of DPYSL3 is positively correlated with Vascular Endothelial

Growth Factor (VEGF), a gene thought to be involved in EMT [155]. This data indicates that our method bridged EMT cancer biology across different disease types and captures global expression patterns in EMT (Supplementary Figure B-1A-C). We confirmed discovery of C1orf116 as epithelial specific gene by testing its expression in PC3 in vitro model of EMT. siRNA knockdown of C1orf116 in PC3 epithelial cell lines showed loss of epithelial markers and gain of mesenchymal markers thereby confirming its functional role as a negative driver of EMT. Clinical data from breast, prostate cancer and lung cancer patients also suggested that changes in expression of C1orf116 could have functional implications in disease progression. Altogether, through this study we have found genes whose effects are represented by multiple cancer types (breast, prostate, liver, colon, esophagus and retinal pigment). We have also validated expression of some genes in an in vitro prostate cancer cell line model and potential relevance in vivo data from three tissues, including one (lung) that was not represented among our cell line data. However, these effects might not necessarily be extrapolated for cancer types not included in this study. As data become available for other tissues and cancers, further analysis can be performed.

Conclusion

Using multi-study integration approach, we identified consensus ranked universal EMT genes. This gene list comprised of a) known EMT genes that included CDH1, ZEB1 and CDH2 b) genes studied in a subset of carcinomas, unknown in prostate cancer: LSR, S100A14 and DPYSL3 and c) novel unknown EMT and cancer genes such as C1orf116. siRNA experiments indicate it to be a potential novel regulator of EMT. Patient gene expression data shows that reduced expression of C1orf116 is associated with poor prognosis in lung and prostate cancer (unadjusted Wilcoxon rank sum p-value < 0.05). In conclusion, our approach of statistical analysis and functional validation identified universal EMT genes and candidate global regulatory genes, thereby both extending current knowledge of EMT and showed preliminary evidence of disease progression in cancer.

This work demonstrates informed statistical modeling to integrate data from multiple independent small sample studies can improve power to detect ubiquitous phenotype associated signal in gene expression measurements. This can also enable discovery of novel genes and biological processes underlying a particular trait, particularly when large datasets are not available for the trait of interest.

Chapter 5

Leveraging large scale human RNAseq studies for reconstruction of context-agnostic gene co-expression networks

Introduction

Accurate reconstruction of gene co-expression networks continues to remain a difficult problem. Reconstruction of co-expression networks over a few thousand genes is of common interest, particularly to understand the gene regulatory landscape of the the human transcriptome. Most network learning methods that are based on pairwise association between genes are sensitive to artifactual variation in gene expression measurements, and introduce false positive edges in the networks [21, 156]. Further, with a typical RNA-seq study that contains a few hundred samples, we are highly underpowered for accurately estimating millions of parameters in a co-expression network over few thousand genes.

In this study, we leverage $> 24{,}000$ uniformly processed and quantified publicly available human RNA-seq samples spanning 236 studies and tissues from recount2 to build context-agnostic gene co-expression networks in order to discover shared biological processes across tissues and cell types in a well-powered analysis [23]. We formulate reconstruction of networks as a structure learning task for a Gaussian Markov Random Field (GMRF), and use the graphical lasso[71] algorithm for inference. Using empirical covariance C as the input, graphical lasso estimates a lasso penalized precision matrix Θ by maximizing the penalized log likelihood of a multivariate gaussian distribution.

$$\log \det \Theta - \mathrm{trace}(\mathrm{C}\Theta) - \Lambda||\Theta||_1$$

It has been demonstrated that aggregation of data across multiple studies can help improve inference and generalizability of models [157, 158]. While accounting for latent sources of intra-study and inter-study heterogeneity in these studies, we considered three strategies of aggregation to obtain an estimate of empirical covariance matrix that was used as an input to graphical lasso: a) compute empirical covariance by merging data from all studies, b) unweighted aggregation of individual study specific empirical covariance matrices, and c) weighted aggregation of individual study specific empirical covariance matrices. We demonstrate networks obtained by integrating studies/datasets shows improvement on held-out data likelihood across all aggregation strategies. Next, we evaluate biological and genetic relevance of topological properties of the context agnostic network. We find that genes with high degree centrality in this network are enriched for mitosis and cell cycle related pathways which are needed in all cell types. We assessed the biological significance of high centrality scores in 14 genesets that included transcription factors, eQTL deficient genes, genes strongly depleted for protein truncating variations (pLI >0.9), happloinsufficient genes, and others. Finally, we apply stratified LD score regression to quantify the contribution of measures of node centrality obtained from context agnostic networks to disease heritability [159–161].

Contributions

I co-led this project with Prashanthi Ravichandran. My contributions include:

- Mentoring Prashanthi on designing simulation and empirical experiments for aggregation strategies, and statistical and biological evaluation of networks
- Download and pre-processing of recount2 data with Prashanthi
- Reconstructing consensus cancer networks with TCGA data from recount2
- Model selection for context agnostic and TCGA networks
- Enrichment of hub genes in mitosis and cell cycle related pathways
- Stratified LD score regression

Methods

Data aquisition, pre-processing and quality control

Raw gene expression RNA-seq counts were downloaded from recount2[23] using the R package *recount*. This is comprised of data from Sequence Read Archive (SRA) including data from the Genotype Tissue Expression (GTEx) project [20] and The Cancer Genome Atlas (TCGA) project. Raw base level coverage counts for each gene were transformed to RPKM. Next, we selected for expressed genes by filtering for RPKM >= 0.1 in more than 25% of samples individually in SRA, GTEx, and TCGA samples. To overcome erroneous associations that may arise from genes that have overlapping genomic location, we restrict our analyses to 6871 non-overlapping protein coding genes with a corresponding gene symbol. We $\log_2(x+1)$ transformed RPKM values, and applied the following sample and study level processing:

- First we selected samples that had metadata information in SRAdb [162], an R package that has a compilation of metadata associated with datasets in SRA.
- Aggregated multiple runs from the same SRA experiment meaning replicates using median expression for each gene across the runs.

- Excluded samples with more than 50% of genes with 0 expression value.
- Excluded potential small RNA expression studies/samples by filtering out samples with size fractionation based library selection protocol.
- Excluded potential single cell RNA-seq studies based on list obtained by an abstract search of terms that included. We used the following search terms were used for the abstract search: "microRNA", "miRNA", and "single cell". The list of these studies can be found at: https://github.com/princyparsana/process_recount2_data/blob/master/projects_excluded.rds
- Selected studies or tissues with 30 or more samples.

At the end of this processing, we obtain 14685 observations from SRA studies, 9633 observations from the GTEx project, and 11284 observations from TCGA. An automated implementation of this processing pipeline can be found on github.

Aggregation of studies

To systematically integrate information from different studies, we employed aggregation strategies at different levels of the data as described below: (Figure 5-1):

Study level aggregation: In this approach, we merge all studies into a single dataset by concatenating the observations. While accounting for intra-study heterogeneity by applying principal component (PC) residualization as described in [21], we performed study level aggregation across multiple tissues in GTEx, and studies in SRA. Using GTEx, we employed an additional study level aggregation where, we accounted only for inter-study heterogeneity after merging data across tissues. Next, we quantile normalize each merged dataset such that every gene follows a Gaussian distribution. We standardize expression measurements so that every gene has zero mean and unit variance. Finally we compute covariance matrix which is used for network inference.

Covariance level aggregation: In this approach, we first corrected gene expression

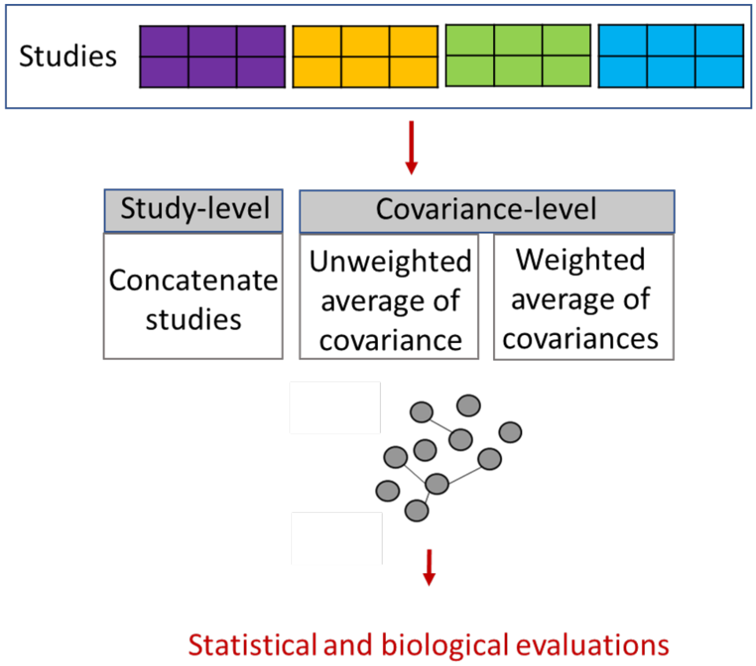

Studies
Study-level
Covariance-level
Concatenate studies
Unweighted average of covariance
Weighted average of covariances
Statistical and biological evaluations

Figure 5-1. Aggregation strategies for integrating co-expression signal across multiple studies.

measurements in each study or tissue using PC based residualization. Next we perform the following steps for aggregation:

- Quantile normalize each study such that every gene follows a Gaussian distribution
- Standardize gene expression measurements in each study such that every gene has zero mean and unit variance
- Compute gene-by-gene covariance matrix within each study S_k

Assuming equal likelihood of error from each study, we compute the unweighted average of covariance matrices $C_{unweighted}$ as:

$$C_{unweighted} = \frac{\sum_k S_k}{|K|} \tag{5.1}$$

where, S_k corresponds to empirical covariance matrix estimated from study k, and $|K|$ is the total number of studies. Next, assuming that studies with larger sample size would have a better estimate of individual covariances, we compute weighted average of covariance matrices $C_{weighted}$ weighed by sample size as:

$$w_k = \frac{n_k}{\sum_k n_k} \tag{5.2}$$

$$C_{weighted} = \sum_k w_k \cdot S_k \tag{5.3}$$

where, S_k corresponds to empirical covariance matrix estimated from study k, n_k is the number of samples in study k and w_k is the weighted contribution of study k

Co-expression network inference

We reconstructed gene co-expression networks using the empirical estimate of covariance matrix obtained from different aggregation approaches as described above. We formulate this as a structure learning problem for a Gaussian Markov Random Field (GMRF), and use graphical lasso as implemented in the *QUIC* [88] R package for inference. Assuming that our gene expression data contains N multivariate gaussian observations each of dimension p, i.e. for each observation, we have expression measurements for p genes, graphical lasso estimates the structure of the co-expression network over genes by maximizing L_1-penalized log likelihood of a multivariate gaussian:

$$\log\det\Theta - \mathrm{trace}(\mathrm{C}\Theta) - \Lambda||\Theta||_1 \tag{5.4}$$

Here C is the empirical covariance matrix obtained from one of the aggregation approaches, and $\Theta = C^{-1}$ is the inverse covariance matrix. The L_1 penalty on Θ induces and controls the amount of sparsity in the solution [71]. If an entry $\Theta_{i,j}$ is 0, then variable i is conditionally independent of variable j given other variables. We inferred networks with penalization parameter Λ ranging from 0.2 to 1.0.

Held-out data likelihood based evaluation of aggregated networks

We assess if networks inferred by integration of co-expression signal across multiple studies improves power to reconstruct reliable and robust co-expression networks using a held-out data likelihood based appraoch. First we applied aggregation strategies described above to two splits of data from recount2: a) aggregate across tissues in GTEx and b) aggregate across studies in SRA (excluding GTEx and TCGA). We compute held-out data likelihood for each version of SRA aggregated networks using five tissues with the largest sample size in GTEx. Similarly, each version of GTEx networks were evaluated using five studies with the largest sample size from SRA as given below:

$$L_i = n_i[\log\det\Theta - \mathrm{trace}(\mathrm{S_i}\Theta)] \tag{5.5}$$

$$\mathbf{L} = \frac{\sum_i L_i}{|I|} \tag{5.6}$$

In equation 5.5, L_i is the held out data likelihood for dataset i, S_i corresponds to empirical covariance matrix from dataset i, Θ is the precision matrix representing the co-expression network being evaluated, $|I|$ is the total number of datasets used to test held-out data likelihood. $\mathbf{L}$ is the average of held-out likelihood across $|I|$ datasets.

Pathway based evaluation of co-expression networks

We used genes known to be functional in the same pathways as ground truth to assess precision and recall of the networks. We downloaded the pathway information (2016) from KEGG, Biocarta and Pathway Interaction Database from Enrichr [79, 80], that were also annotated as canonical pathways by MSigDB [81]. Table 5-I shows the number of pathways in each of these database.

Any pair of genes that have at least one pathway in common were assumed to have a true functional relationship. An edge that was observed between a pair of

Database	# of Genesets
KEGG	293
Biocarta	237
Reactome	1530
Pathway Interaction Database	209

Table 5-I. Canonical pathway genesets.

genes in the inferred network and was also present in the list of real connections was called as a true positive (TP). We defined false positive (FP) to be an edge that was observed between a pair of genes in the inferred network, however was absent in the list of real connections. All TP, FP and FN were computed with genes restricted to the most variable 5000 genes that were used for reconstructing co-expression networks. We compute *precision* and *recall* as given below:

$$Precision = \frac{TP}{TP + FP} \tag{5.7}$$

$$Recall = \frac{TP}{TP + FN} \tag{5.8}$$

Computing gene centrality scores using network structure

Using measures of network connectivity, we compute centrality scores for each gene in the network. Given a weighted undirected graph G, first we normalize the graph by dividing the weight of each edge by the maximum of all edge weights in the network. If E is the list of all edges in the network (excluding diagonals) and $E_{v,j}$ is the weight of an edge connecting genes v and j, we get the normalized edge weight $\hat{E}_{v,j}$ for this edge as:

$$\hat{E}_{v,j} = E_{v,j}/max(E_{v,j})$$

Next, using normalized edge weights, we compute the following types of centrality scores:

- Degree(v): The degree centrality of a gene v corresponds to the number of neighbors connected to v

- Closeness(v): captures how close gene v is to all other genes in the network. For this, we first compute the weighted distance between gene v and gene j in the network as:

  $$d_{v,j} = \frac{1.00}{\hat{E}_{v,j}}$$

  If v and j are disconnected, then $d_{v,j}$ is set to 0. Using this, we can compute the closeness centrality of a gene v as:

  $$\frac{1}{\sum_{v,j \neq v} d_{v,j}} \tag{5.9}$$

- Betweenness(v): is the number of shortest paths in the network that pass through gene v. A shortest path between nodes v and j is a path where the total sum of the edge weights in the path is minimum.

- Max weight(v): is the maximum of weights of all edges connected to v

- Eccentric(v): is the shortest path distance from the farthest node in the graph

- Eigenvector(v): is proportional to the sum of centrality of neighbors of v. It is given by:

  $$\mathbf{A}x = \lambda x \tag{5.10}$$

  where where x is the eigenvector of the weighted adjacency matrix $\mathbf{A}$ with the largest eigenvalue λ.

Enrichment and overlap with genesets

We assess the biological significance of high centrality scores in 13 genesets that correspond to different gene importance related metrics and were obtained from [159]. These included:

1. All genes: This set includes all 19031 protein coding genes according to HGNC [159].

2. MGI essential genes: This includes genes for which homozygous knockout in mice resulted in pre-, peri-, or post-natal lethality.

3. Autosomal dominant genes: Genes among OMIM disease genes that are considered to follow autosomal-dominant inheritance.

4. Happloinsufficient genes: Genes of severe, moderate, and mild haploinsufficiency, where having only a single functioning copy of a gene is not enough to carry out normal functions.

5. High pLI genes: Genes with high probability for being loss-of-function intolerant (pLI). This list contains genes with pLI > 0.9, meaning that these genes are strongly depleted for protein truncating variants

6. High s_{het} genes: Genes with high selective effects for heterozygous protein truncating variants (s_{het}). This geneset contains genes with $s_{het} > 0.1$, reflecting strong selection against protein truncating variants

7. High Phi genes: This list included LoF-constrained genes with probability of haploinsufficiency (Phi) > 0.95.

8. High missense Z genes: These are genes that are strongly depleted for missense mutations.

9. ClinVar: Genes with a pathogenic or likely pathogenic variant.

10. OMIM genes: Genes obtained from Online Mendelian Inheritance in Man (OMIM) database.

11. GWAS genes: This list includes genes closest to the peak of a significant GWAS loci (p $\leq 5e-8$).

12. Transcription factors: This list contains the list of transcription factors.

13. High EDS: Genes with high score in the enhancer domain score.

include genes with high pLI scores, Genes with high S_{het} scores, and others. First, we divide genes with centrality score > 1 into 25 bins ordered by scores. Genes with centrality score of zero are binned together into one group. Next, using genes in each bin, we compute excess overlap (EO) as described in [159]:

$$EO(G_1, G_2) = \frac{P_d}{P_{tot}} \tag{5.11}$$

$$P_d = \frac{|G_1 \cap G_2|}{|G_2|} \tag{5.12}$$

$$P_{tot} = \frac{|G_1 \cap G_{net}|}{|G_{net}|} \tag{5.13}$$

where G_1 is geneset one of the 14 genesets, G_2 is the our test set corresponding to genes in one of the bins, and G_{net} is total number of genes in the network. We compute the standard error of this excess overlap as[159]:

$$SE = \frac{\sqrt{\frac{(P_d(1-P_d)}{|G_2|}}}{P_{tot}} \tag{5.14}$$

Heritablity enrichment with S-LDSC using annotations obtained from measures of network centrality

We applied stratified LD score regression to quantify the contributions of measures of node centrality obtained from context-agnostic networks in disease heritability. First we transform all centrality scores to lie between 0 and 1. Next we annotate SNPs within 100kb of a gene with the centrality score assigned to the gene. If a SNP was within a 100kb of more than one gene, we assigned the maximum centrality score to the SNP. We generate six network centrality based annotations and estimate their heritability enrichment and the standardized effect size ($\tau*$) of an annotation as described in [159–161]. Given β_j is the effect size of a trait associated SNP j, its variance is a linear additive contribution to the annotation c which is given by:

$$Var(\beta_j) = \sum_c a_{cj}\tau_c \tag{5.15}$$

here, τ_c is the per-SNP contribution of the annotation c to the heritability of the trait. S-LDSC esimtates τ_c by fitting the following regression[160, 161]:

$$\mathrm{E}[\chi_j^2] = N \sum_c \ell(j,c)\tau_c + 1 \tag{5.16}$$

here, N is the number of samples in the GWAS, $\mathrm{E}[\chi_j^2] = N\beta_j^2$, $\ell(j,c)$ LD score of SNP j to the annotation c in the pre-defined window-size (100kb in our analysis). We compute the standardized effect size ($\tau*$), i.e. the proportionate change in per-SNP heritability associated with a one standard deviation increase in the value of the annotation conditional on all the other annotations in the model as defined by [159, 161]:

$$\tau_c* = \frac{\tau_c sd(C)}{\frac{h_g^2}{M}} \tag{5.17}$$

here, $sd(C)$ is the standard deviation of annotation C, h_g^2 is the estimated SNP heritability, and M is the number of variants used to compute SNP heritability. Enrichment of an annotation is the proportion of heritability explained by SNPs in the given annotation divided by the proportion of SNPs in the annotation. In our analysis we used European samples from the 1000G as reference SNPs, and regression SNPs were obtained from HapMap3, and SNPs in the MHC regions were excluded. We conditioned all our analysis on the v2.2 baseline LD annotations. All analyses was done using hg38 build of the human genome, and relevant files were obtained from: https://data.broadinstitute.org/alkesgroup/LDSCORE/GRCh38/.

Results

Aggregation of data from multiple studies improves network reconstruction

To account for study-specific latent artifacts in gene expression measurements, we PC based residualization as described in [21] to each dataset in our analyses. Next, we considered three strategies to aggregate data from multiple studies and obtain empirical estimates of covariance matrices, which were used to build co-expression networks with graphical lasso: a) first, we performed a study level aggregation where we merged corrected expression matrices by concatenating samples from all studies to form one dataset which was used to compute a gene by gene covariance matrix, b) second, we computed an unweighted average of covariance matrices obtained from each dataset, c) third, we computed a weighted average of study-specific covariance matrices weighted by sample size. We applied this approach to two sets of data from recount2; one using tissues as a proxy for studies in GTEx with a total of 9633 samples, and second with individual projects in SRA (excluding GTEx and TCGA) with a total of 14685 samples. (Figure 5-1). We assessed if aggregation of data helps improve power for inferring gene co-expression networks using a held-out data likelihood approach. We used the estimated precision matrix (represents networks) from graphical lasso to evaluate likelihood of held-out studies for each aggregation-dataset pair using a multivariate Gaussian distribution. Five GTEx tissues with largest sample size were used as held-out data for networks obtained from SRA, and *vice versa*. Both with GTEx and SRA we observe that networks obtained by merging studies or covariances showed improvement in held-out data likelihood across all aggregation strategies (Figure 5-2). Reconstructed networks become sparser as we increase the number of studies that were included in aggregation. Next, using known biological pathways as ground truth, we compute precision, recall and F1-score for each aggregation-dataset pair. With aggregation we see consistent improvement on recall and F-1 score, while we find that precision shows improvement in aggregated networks with higher number of edges (Figure 5-3). Among the highest level of aggregated networks (50 tissues for GTEx and 186 studies for Recount), we found that merging studies and weighted aggregation of covariance matrices yielded very similar networks, which was also observed in their performance on held out data likelihood. This was consistent with

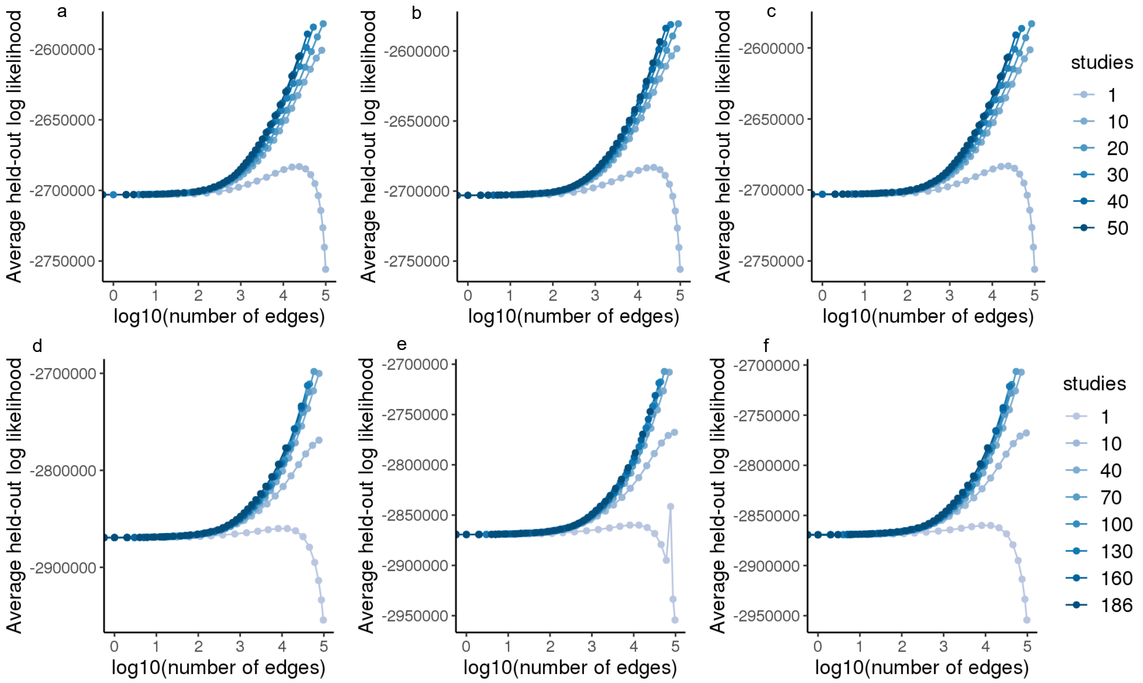


Figure 5-2. Aggregating studies improves recontruction of co-expression networks in GTEx (top row) and SRA (bottom row).

previous observations [163]. Both approaches outperformed unweighted aggregation of covariance matrices (Figure 5-4). For each aggregation strategy, we selected networks corresponding to a specific value of graphical lasso penalty such that the networks exhibited a scale-free topology, i.e. node degrees of the network followed a power law distribution. We adapted the the scale-free test as described in [164] that measures the variance explained (R^2) by fitting a linear model between $\log(p(d))$ and $\log(d)$, where $p(d)$ is represents the fraction of genes in the network with d neighbors. We selected penalty parameters for networks based on $R^2 \approx [0.75 - 0.85]$. We find that networks obtained by merging studies, and by weighted aggregation of covariance matrix showed high overlap among the top 500 hub genes (Figure 5-5). This was consistent with our previous observation where we find the two versions of aggregation were found to be highly similar.

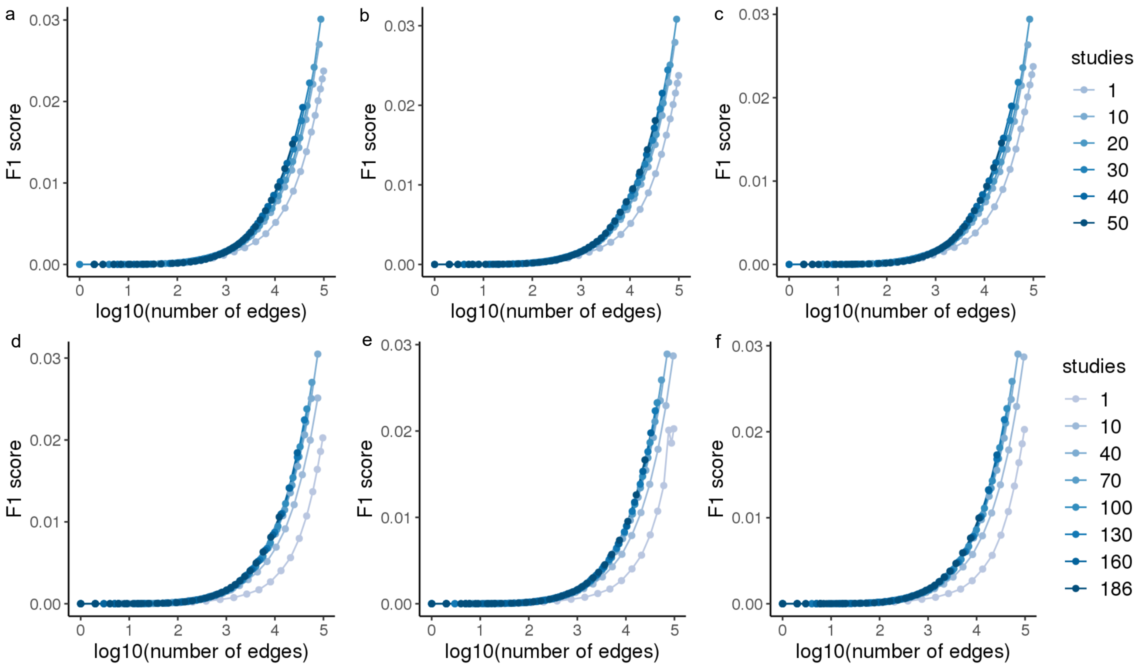


Figure 5-3. F1 score of aggregated networks evaluated using canonical pathways in GTEx(a-c) and SRA (d-f).

Context-agnostic gene co-expression networks: inference and network properties

While weighted aggregation and merging perform comparably in our analyses, it has been demonstrated that for effect size estimation, simple pooling of heterogenous datasets can lead to confounding results, and a aggregation based meta-analysis approach can protect against such effects [165]. Hence, we chose to reconstruct context-agnostic co-expression networks using all non-TCGA data from Recount2

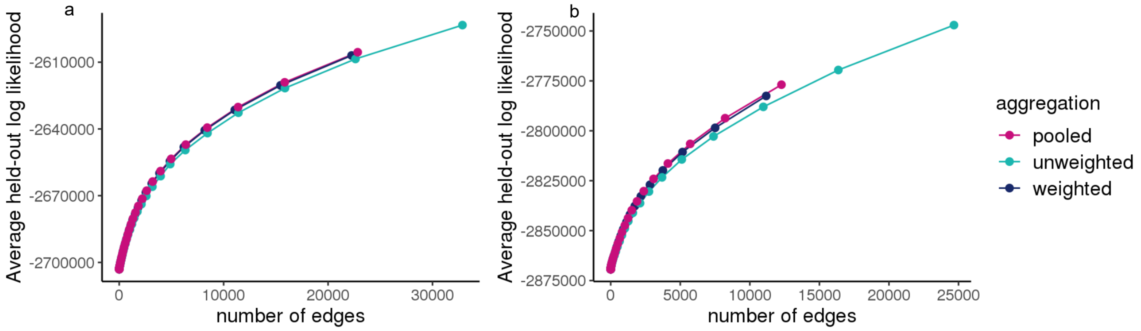


Figure 5-4. Comparing across aggregation strategies. a) GTEx, b) SRA

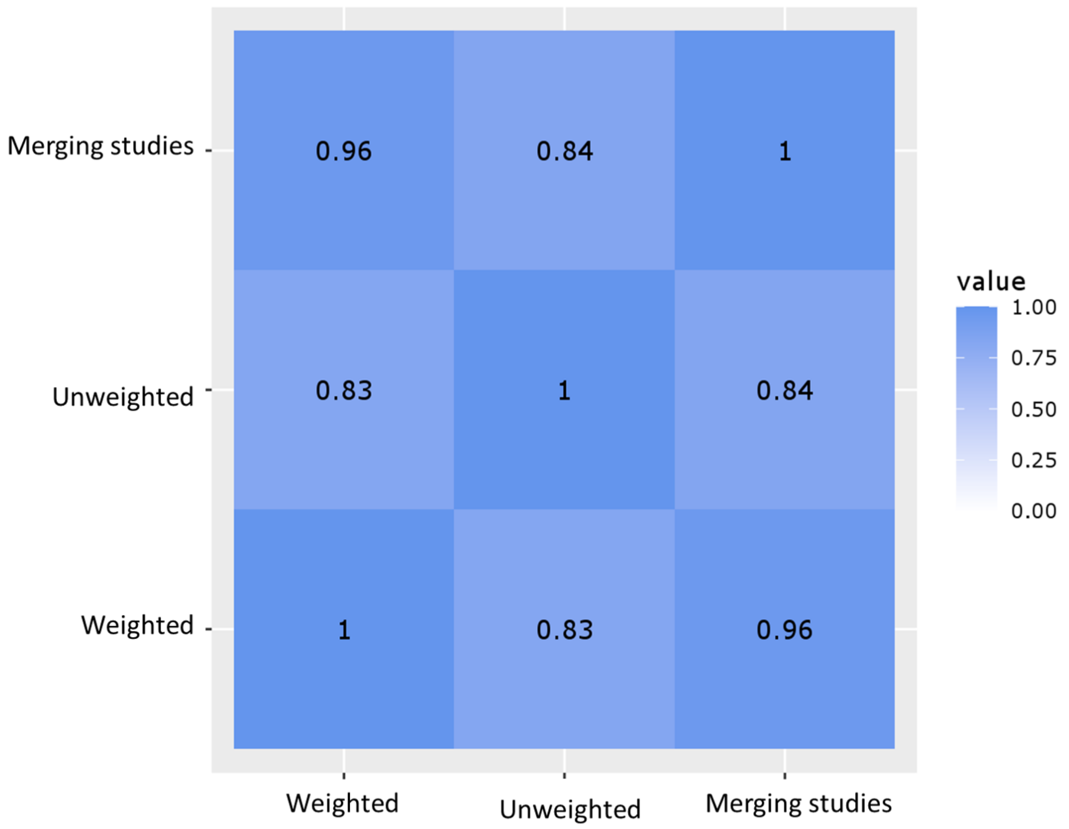


Figure 5-5. High overlap among the top 500 hub genes across aggregation. The entries in the heatmap correspond to proportion of shared hub genes. This figure was generated from SRA networks. Trend was similar in GTEx.

using a weighted aggregation of study-specific covariance matrix as the estimate of empirical covariance in graphical lasso. This included 186 studies from SRA with 14685 samples, and 50 tissues from GTEx with 9633 samples. Since we found that networks get sparser over aggregation, we expanded our range of $\Lambda = [0.1, 1]$ for building context-agnostic networks. We selected the network corresponding to $\Lambda = 0.2$, which had an $R^2 = 0.79$. This network had 12341 edges over 6871 genes, and had an average connectivity of 3.59 per node (Figure 5-6). We computed node importance for genes in the network using different measures of node centrality (See Methods). Across multiple centrality annotations, we found that few nodes have high scores (Figure 5-7). We also find that most centrality scores are highly correlated with each other (Figure 5-8). However, we would take these correlation estimates with caution,

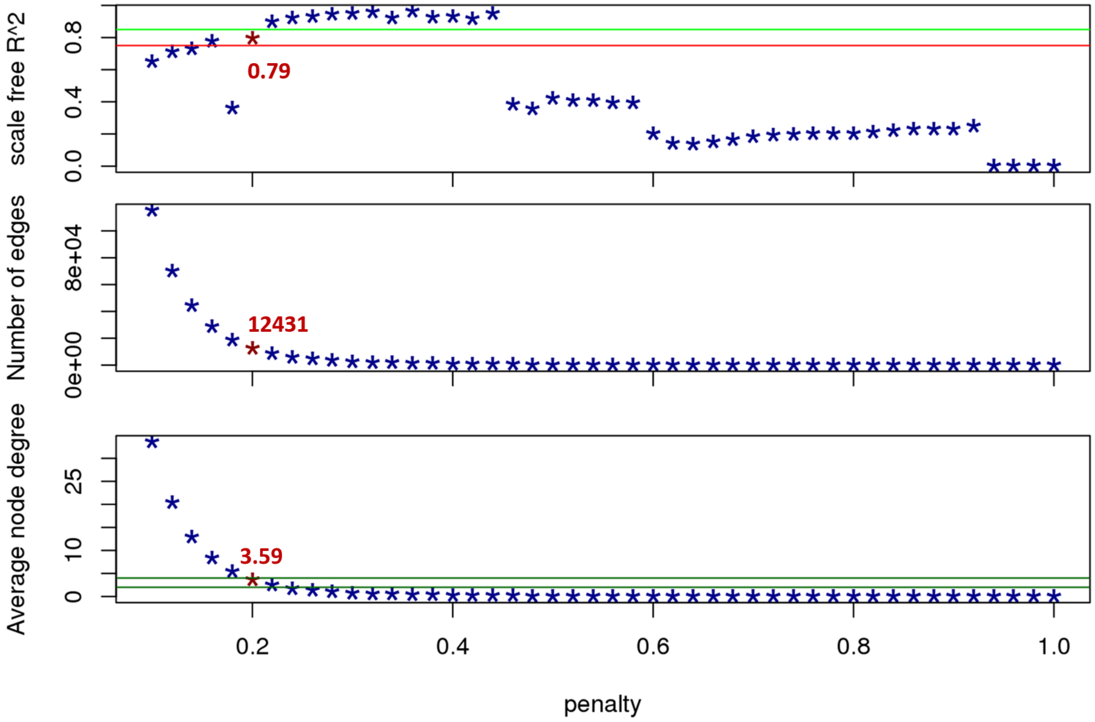


Figure 5-6. Context agnostic networks were selected to have scale-free topology.

since it could be inflated by the high number of genes with a 0 centrality score.

Genes with high connectivity in context-agnostic networks enriched for annotations of biological importance

Genes with high network centrality are assumed to have wider influence on the information transfer in the networks, and may be indicative of increased regulatory relevance. We used the 13 genesets [159] (described in Methods) that are indicative of increased biological significance to evaluate genes that had high measures of centrality. For this, we grouped genes into 25 equal bins of ordered centrality scores, and an additional bin for genes with a score of 0. Next, we compute excess overlap of genes in each bin with each of the 14 genesets. Genes corresponding to bins with high centrality scores were enriched for high pLI genes (pLI $>$ 0.9; these genes are strongly depleted for protein truncating variatns), high S_{het} genes ($S_{het} > 0.1$; these genes show strong selection against protein truncating variants) [159, 166], genes strongly

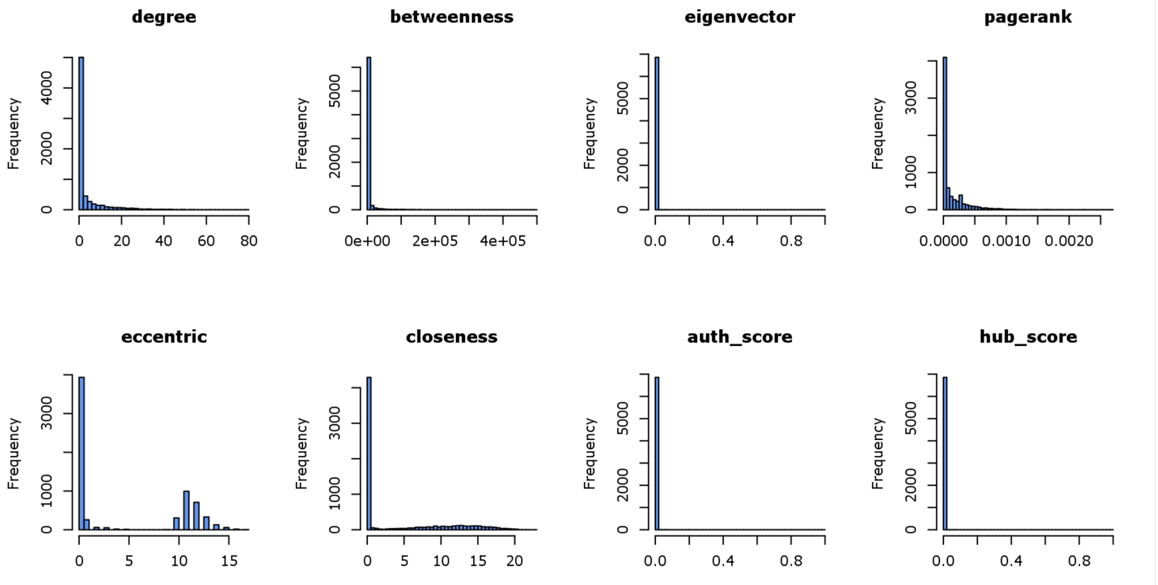


Figure 5-7. Distribution of node centrality scores.

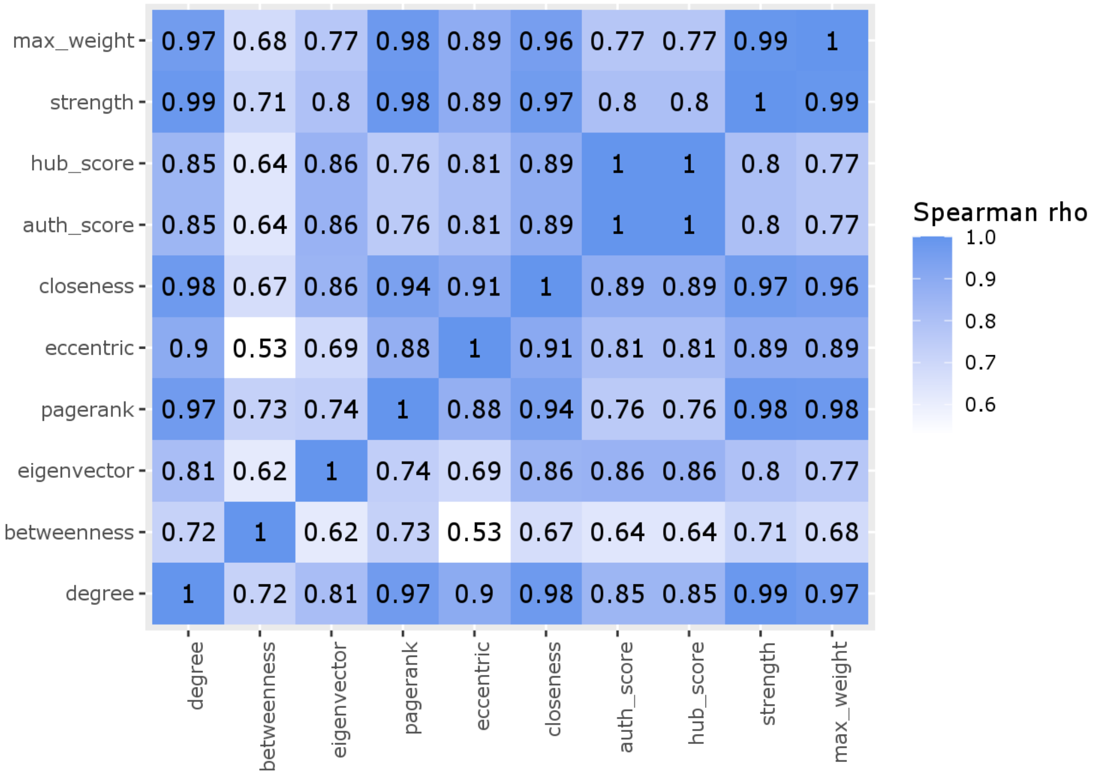


Figure 5-8. Spearman correlation between different node centralities.

depleted for missense mutations, and Mouse Genome Informatics essenetial genes (Figure 5-9). We do not find a conclusive pattern of enrichment for other genesets (described in Methods) we attempt to evalaute. Exploratory analyses revealed the hubs of context agnostic networks included multiple mitosis and cell cycle related genes such as *CDK2*, *TOP2A*, *CENPE*, and *CDC20* (Table 5-II). Genes in the top 98th percentile of degree centrality were significantly enriched for mitotic spindle assembly ($OR = 16.23, P = 2.97 \times 10^{-8}$) and G2/M transition of mitotic cell cycle genesets from GO biological processes($OR = 8.96, P = 8.90 \times 10^{-6}$). We compared this to a tissue-specific skeletal muscle network from GTEx, and do not similar trend of enrichment. This shows that tissue agnostic networks can capture critical biological processes that are shared across tissues and cell types.

Gene Symbol	**# of neighbors**
NCL	80
CDCA5	65
NCAPH	63
MELK	63
KIF11	62
CDK1	58
AURKAIP1	58
HUWE1	55
UBL5	55
TOP2A	54
ZWINT	53
SCAF1	53
NUSAP1	53
BUB1	53
KIF18B	52
ATP5B	51
BIRC5	50
GTSE1	49
RHOT2	49
PLK4	49

Table 5-II. Top 20 hub genes.

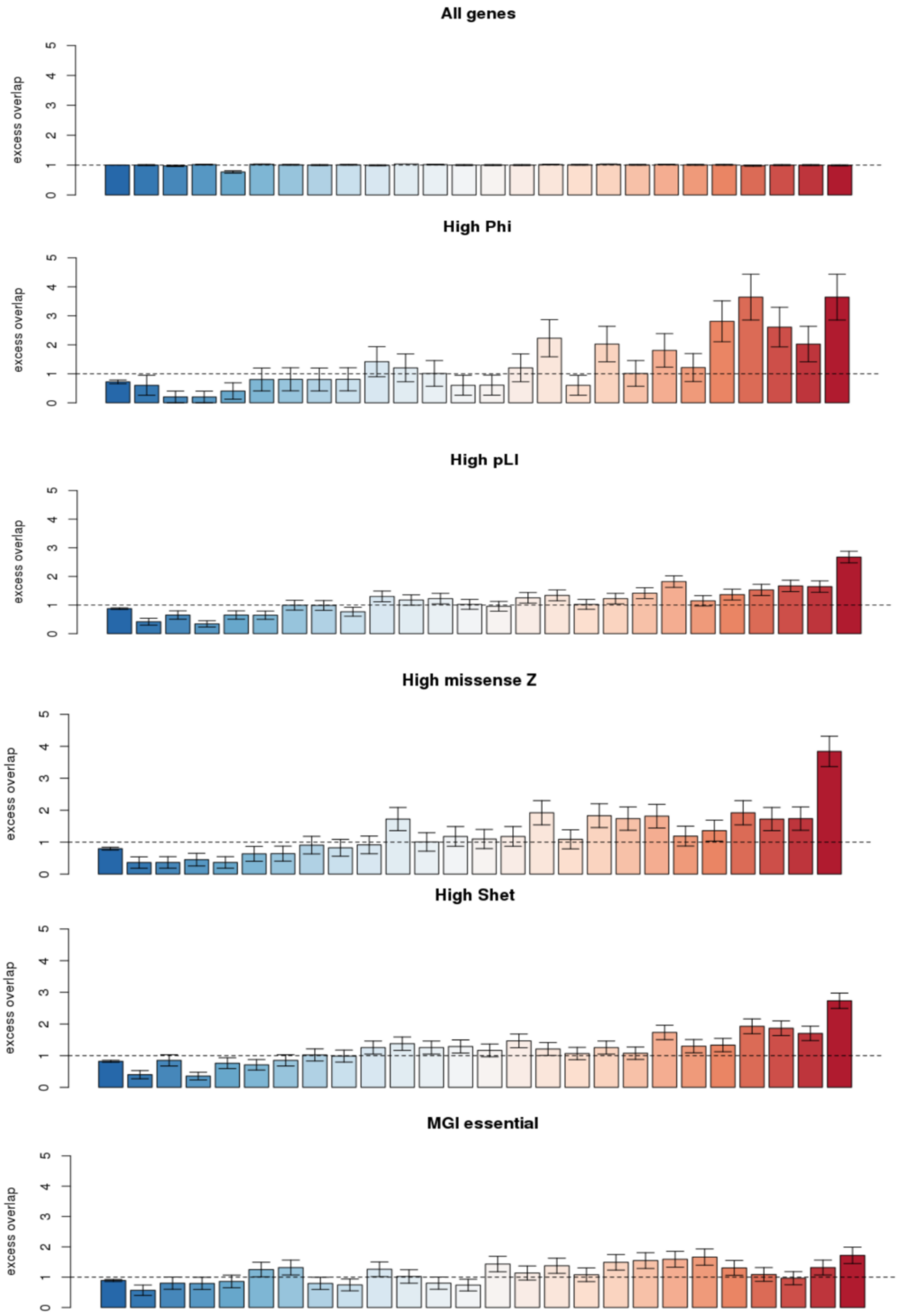


Figure 5-9. Genes with high closeness centrality enriched for genes under strong selection.

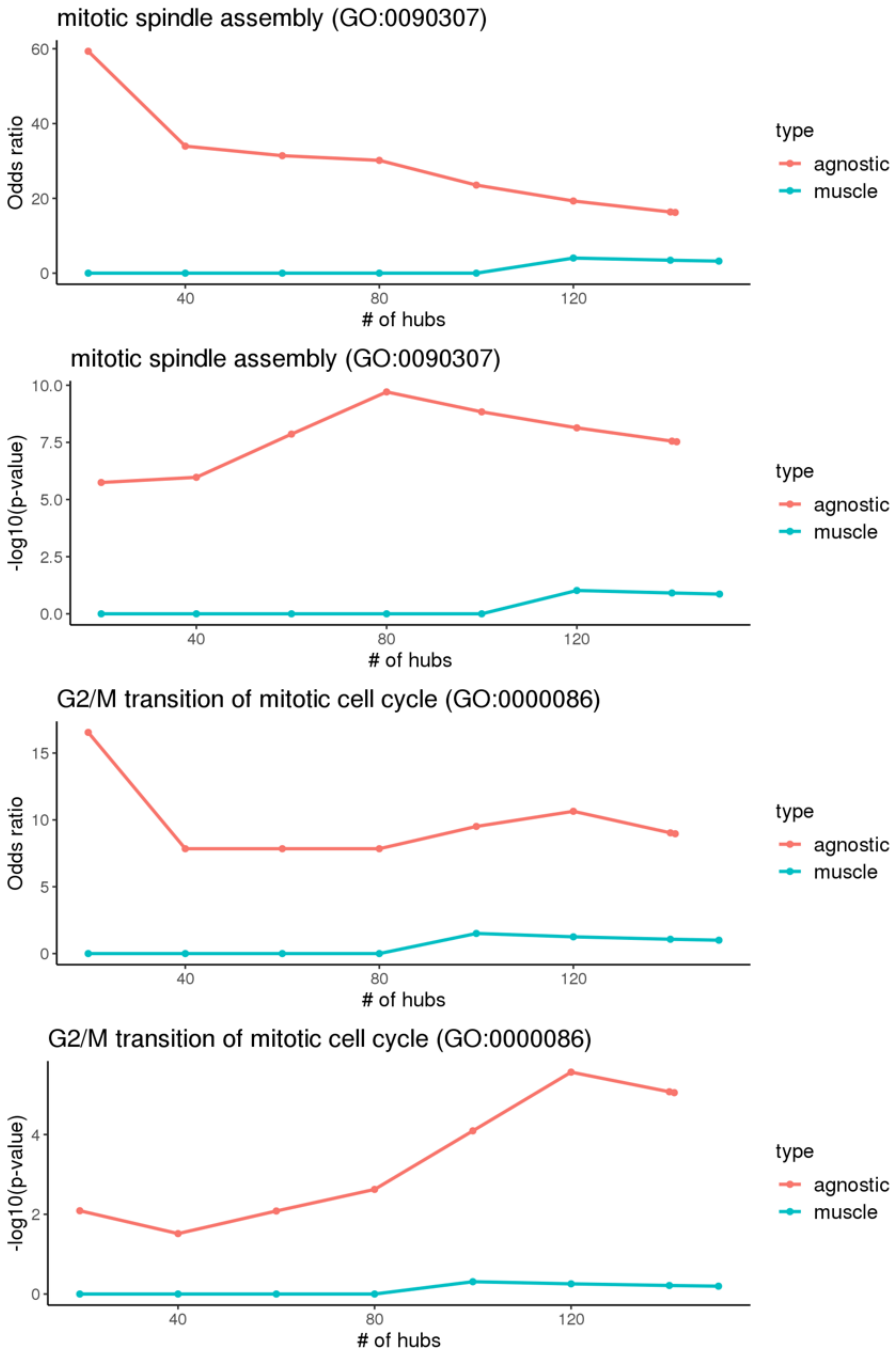


Figure 5-10. Genes with high degree centrality enriched for mitosis related cell cycle pathways.

Enrichment of phenotypic heritability with centrality based network annotations

Next we attempt to quantify the contribution of centrality based network annotations to phenotypic heritability. We annotate SNPs within +/- 100kb of a gene TSS using each of the six measures of gene level centralities. We applied stratified LDSC to 285 independent phenotypes from the UKBB cohort with heritability z-score >=7. Following guidelines in [160], we also excluded traits with genetic correlation > 0.9. We find significant enrichment of heritability with each network annotations across multiple traits. Annotations obtained from betweenness centrality ovreall showed high enrichment of disease heritability compared to others. (add heritability estimate and pvals) (Figure 5-11. Estimates of regrerssion coefficients $\tau*$ were not significant across individual traits. $\tau*$ quantifies the estimate of annotation based effect on heritability of a phenotype conditioned on the baseline LD annotations. This could imply that though network annotations were enriched for trait heritability, most of it was captured by the 97 baseline annotations we conditioned the model on.

Next, we meta-analyzed the results of heritailibity enrichments and estimate of standardized $\tau*$ across 285 phenotypes using a random effects meta-analysis. For all six network annotations we find strong enrichment of disease heritability ranging from $1.21(SE0.0213, P < 2.2 \times 10^-16)$ to 1.64 $(SE0.0323, P < 2.2 \times 10^-16)$ (Figure 5-12. While estimates of $\tau*$ did not show any significance across individual phenotypes, upon meta-analyses across phenotypes, it was significant for annotations based on betweenness centrality $(\tau* = 0.0123, SE = 0.00297, P = 0.0000166)$. However, we note that the effect size of esimates of $\tau*$, were small for all network annotations being evaluated, including betweenness centrality. Our results from this analyses indicate that while the six centrality annotations obtained from context-agnostic networks show enrichment in trait heritability, we only find significant effect of $\tau*$ with betweenness centrality with a small effect size.

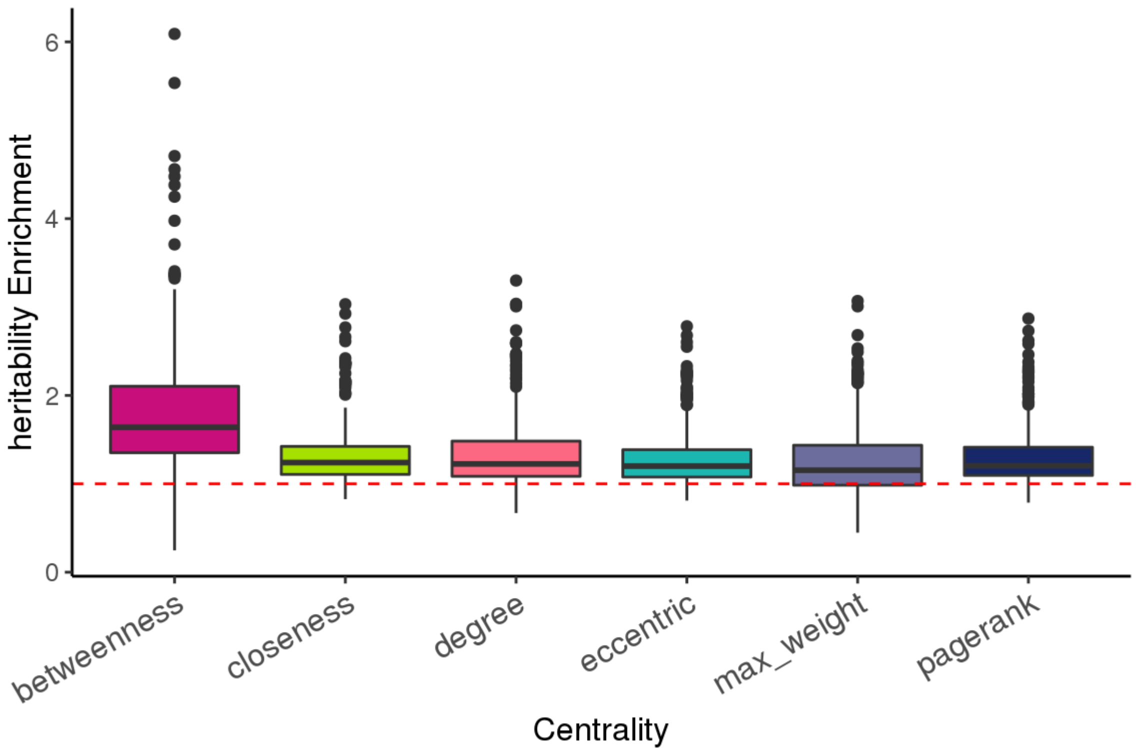


Figure 5-11. Distribution of heritability enrichment across 285 phenotypes from UKBB based on different measures of centarlity.

Conclusion

In a typical RNA-seq study with just a few hundred samples, we are highly under powered to estimate high dimensional gene co-expression networks. Inconsistent and missing metadata make it challenging to effectively utilize a large repertoire of publicly available gene expression studies. With informed probabilistic modeling, these data can be aggregated across multiple studies to discover shared co-expression patterns across different biological contexts. Our work shows that aggregation of data across studies helps improve reconstruction of context-agnostic co-expression networks. Held-out data likelihood using inferred networks show consistent improvement upon aggregation across all three strategies. After accounting for study-specific heterogeneity, networks obtained from merging data and from weighted aggregation of study-specific covariance matrices yield very similar networks. In this work, using data from recount2, we build

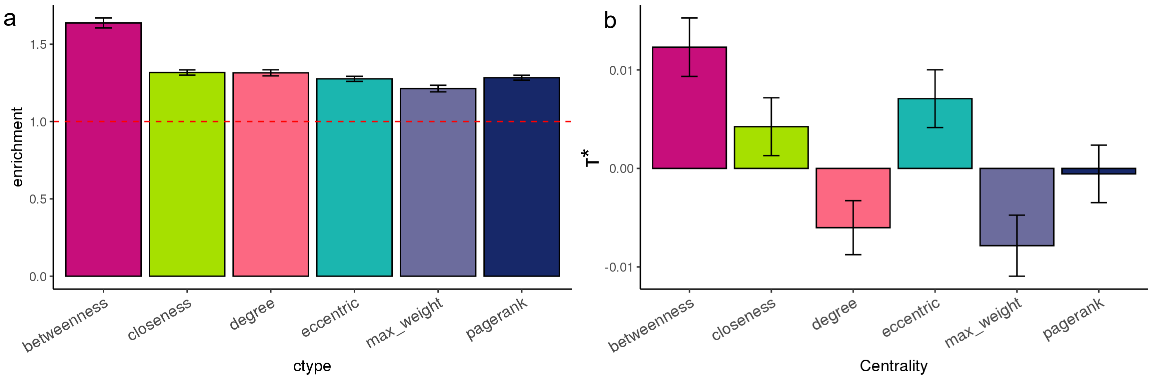


Figure 5-12. Heritability enrichment and $\tau*$ estimates meta-analyzed across traits.

context-agnostic co-expression networks using data from 50 GTEx tissues and 186 studies in SRA. Hub genes from our networks show significant enrichment of mitosis and cell cycle related pathways, implying that biological signal identified captures ubiquitous cellular processes. Further, we find genes with high closeness centrality are enriched for genesets that reflect strong evolutionary constraints such as those with high pLI, Phi, and S_{het} scores. Finally we assess if topological properties of genes in the network can explain phenotypic heritability using S-LDSC regression. Agnostic network based annotations show significant enrichment in trait heritability, however most of the heritability contribution by network annotations is explained by the 97 baselined LD annotations we conditioned our analyses on. Overall, in this work we show that: a) informed aggregation of public data can improve network inference, b) context-agnostic networks can provide insights on universal biological processes critical across tissues and cell types, and c) network central genes captures patterns of phenotypic heritability.

Chapter 6

Probabilistic mixture model to reconstruct context-specific gene co-expression networks

Introduction

The abundance of publicly accessible human gene expression studies available on databases such as Gene Expression Omnibus (GEO) and Sequence Read Archive (SRA) is constantly increasing. There have been several efforts to uniformly process this data, and quantify RNAseq measurements across all publicly available human RNAseq studies using a standardized pipeline[23]. These databases form an attractive resource for reconstructing gene co-expression networks. However, inconsistent and unreliable metadata continue to be a major hurdle in effectively leveraging public RNAseq studies to improve statistical power.

In this chapter, I describe a hierarchical mixture model *groupNet* that leverages multiple datasets to learn the structure of a Gaussian Markov random field (GRMF) to build context-specific co-expression networks. In absence of reliable meta-data, our model works by assigning a mixture weight to each study that defines it's relatedness to a context. A context can be a biological phenotype such as tissues, cell types, or disease states. The model borrows strength across studies via a mixture weight driven aggregation to estimate context-specific sparse co-expression networks. Preliminary evidence shows that at reasonable network density, *groupNet* can correctly identify study groupings, and shows moderate performance on reconstruction of context-specific networks

Methods

Background on graphical lasso

Estimating the structure of a sparse high dimensional Gaussian graphical model is of common interest across biology and statistics. Graphical lasso is one of the most widely used algorithm for this purpose. Assuming that our gene expression data contains N multivariate Gaussian observations each of dimension p, i.e. for each observation, we have expression measurements for p genes, graphical lasso estimates the structure of the co-expression network over genes by maximizing L_1-penalized log likelihood of a multivariate gaussian given by:

$$\log\det\Theta - \mathrm{trace}(\mathrm{S}\Theta) - \Lambda||\Theta||_1 \tag{6.1}$$

Here S is the empirical covariance matrix and $\Theta = \Sigma^{-1}$ is inverse covariance matrix. As described in the previous chapter, Λ corresponds to the L_1 penalty on Θ that controls the amount of sparsity in the solution [71], and captures patterns of conditional independence between genes. In our work, we only penalize the non-diagonal elements of the precision matrix, meaning that the diagonal entries of Λ are fixed to 0.

Next we describe our model *groupNet* which is a mixture model building on the network learning principles from graphical lasso.

Problem set up

Suppose we have a database of K independent studies from a publicly available human gene expression repository. E.g. *Recount2.* These studies belong to C different tissues, however we do not have tissue annotations for study k. Our goal is to first identify the most likely tissue assignment for study k. We do this by giving soft clustering assignment to each study-tissue pair. Next, we use these assignments for each tissue to obtain a context-specific covariance matrix, which is used as an estimate of an empirical covariance matrix in graphical lasso.

***groupNet*: model description and inference**

Here we describe the mathematical formulation of the model, and derive updates for parameter inference. We first introduce some notation that we use to describe *groupNet*. We use K to denote the set of independent datasets or studies in our analyses, and C to denote the number of classes that are specified by the group of K studies. X_k is a matrix of gene expression from dataset k with n samples and p genes.

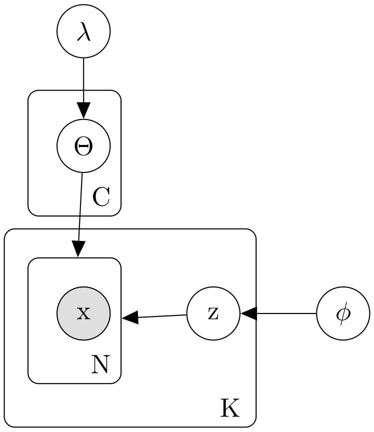


Figure 6-1. Soft-clustering based graphical lasso.

Figure 6-1 is the graphical model representation of our model. ; the variables in the model in figure 6-1 are described below:

- K is the number of studies/datasets
- each dataset contains N_k samples

- x_{nk} is p-dimensional vector of gene expression measurements from observation n from study K. It is drawn from a multivariate gaussian distribution given by:

$$p(x_{k,n}|z_k = c) \sim MVN(0, \Theta_c) \tag{6.2}$$

  Here Θ_c is a $p \times p$ precision matrix

- Θ_c has an elementwise Laplace prior:

$$p(\Theta_c|\lambda) \sim Laplace(0, \lambda) \tag{6.3}$$

- z_k is categorical variable of assignment of dataset K to a context

$$p(z_k = c) \sim Categorical(\phi_c) \tag{6.4}$$

The joint likelihood of the model in Figure 6-1 is given by:

$$p(X, Z|\Theta_c, \phi, \lambda) = \prod_k \left[\prod_n p(x_{k,n}|z_k = c, \Theta_c)\right] p(\Theta_c|\lambda)p(z_k = c|\phi) \tag{6.5}$$

Since $Z's$ are not observed, we marginalize it out, and hence the marginal probability is then given by:

$$p(X|\Theta, \phi) = \prod_k \sum_c \left(\left[\prod_n p(x_{k,n}|z_k = c, \Theta_c)\right] p(\Theta_c|\lambda)p(z_k = c|\phi)\right) \tag{6.6}$$

Next, we take the log of 6.6

$$\log p(X|\Theta, \phi) = \sum_k \log \sum_c \left(\left[\prod_n p(x_{k,n}|z_k = c, \Theta_c)\right] p(\Theta_c|\lambda)p(z_k = c|\phi)\right) \tag{6.7}$$

If $Z's$ were observed, this could be re-written as:

$$\log p(X, Z|\Theta_c, \phi) = \sum_k \sum_c 1(z_k = c)\Bigg(\left[\sum_n \log p(x_{k,n}|z_k = c)\right] + \log p(\Theta_c|\lambda) + \log p(z_k = c|\phi)\Bigg) \tag{6.8}$$

Since z_k is unobserved, we take an expectation maximization based approach for inference. Given a specific initialization of Θ_c, we compute the posterior probability

$p(z_k = c|X_k, \Theta_c, \lambda)$ in the E-step as:

E Step:

$$p(z_k = c|X_k, \Theta_c, \lambda, \phi) = \gamma_{k,c} = \frac{([\prod_n p(x_{k,n}|z_k = c, \Theta_c, \lambda)]\, p(z_k = c|\phi))}{\sum_c ([\prod_n \log p(x_{k,n}|z_k = c, \Theta_c, \lambda)]\, p(z_k = c|\phi))} \tag{6.9}$$

Once we have computed the posteriors, the expected log likelihood of the data is given by:

$$\begin{aligned} E_{Z|X,\Theta}\Big[p(X, Z|\Theta_c, \lambda, \phi)\Big] = \sum_k \sum_c \gamma_{k,c} \bigg(\Big[\sum_n \log p(x_{k,n}|z_k = c) \Big] \\ + \log p(\Theta_c|\lambda) + \log p(z_k = c|\phi) \bigg) \end{aligned} \tag{6.10}$$

M Step:

Next given the current estimates of $\gamma_{k,c}$, we estimate the parameter Θ_c as:

$$\arg\min_{\Theta^c} \sum_k \gamma_{k,c} \left([n_k \left(-\log|\Theta_c| + \mathrm{Tr}(S_k \Theta_c) \right)] + \log \phi_c \right) + \Lambda||\Theta_c||_1 \tag{6.11}$$

Here, $\Lambda||\Theta_c||_1$ is the lasso penalty on Θ_c that comes from the Laplacian prior and induces sparsity in our networks. Taking the derivative of 6.11 w.r.t. Θ_c we get:

$$= \left(\sum_k \gamma_{k,c} \left(\Big[n_k \Big(-(\Theta_c)^{-1} + S_k \Big) \Big] \right) \right) + \Lambda\Gamma \tag{6.12}$$

$$= \left(\sum_k -\gamma_{k,c} n_k (\Theta_c)^{-1} + \sum_k \gamma_{k,c} n_k S_k \right) + \Lambda\Gamma \tag{6.13}$$

Let $\gamma_{k,c} n_k = w_{kc}$, then we will have:

$$= \left(\sum_k -w_{kc}(\Theta_c)^{-1} + \sum_k w_{kc} S_k \right) + \Lambda\Gamma \tag{6.14}$$

$$= -\Theta_c^{-1} + \frac{\sum_k w_{kc} \cdot S_k}{\sum_k w_{kc}} + \frac{\Lambda}{\sum_k w_{kc}} \Gamma \tag{6.15}$$

The above equation can now be solved through graphical lasso. Using an empirical estimate of the covariance matrix as,

$$\frac{\sum_k w_{kc} \cdot S_k}{\sum_k w_{kc}} \tag{6.16}$$

we used R package *glasso* to infer the L1 penalized context specific precision matrix.

Simulation study

We consider a two class problem for this simulation. We first generate two scale-free networks with $p = 1500$ genes using *igraph* R package [167]. Next, given a network structure, we generate a corresponding covariance matrix as follows[168]:

- Create a $p \times p$ matrix with ones on the diagonal, and zeros on elements that do not correspond to edges
- assign entries corresponding to edges with values from a uniform distribution with support on $\{[-0.4, 0.1] \cup [-0.1, 0.4]\}$
- Divide each off-diagonal element of the matrix by 1.5 times the sum of the absolute values of off-diagonal elements in its row. This helps ensure positive definiteness
- average the matrix with its transpose to get a symmetric positive definite $\mathbf{B}$ matrix
- finally we compute the corresponding Σ as:

$$\frac{d_{ij}(\mathbf{B})^{-1}_{ij}}{\sqrt{(\mathbf{B})^{-1}_{ii}(\mathbf{B})^{-1}_{jj}}} \tag{6.17}$$

We used this as the estimate of the covariance matrix Σ to generate data from a multivariate gaussian distribution with mean 0. This simulation setup has been implemented as an R package *netsimulatR* and is available on github. We generate data from six studies drawn from two classes with this framework. Using this data we reconstruct co-expression networks using *groupNet*, context-specific average based graphical lasso, and context-agnostic average based graphical lasso.

Results

For this experiment, we used six studies drawn from two simulated covariance matrices as described above. We applied *groupNet* to reconstruct two context-specific networks. Preliminary evidence shows that *groupNet* captures context-specific co-expression patterns and correctly clusters related studies together into the same group. *groupNet* shows performance comparable to class-specific average based graphical lasso on the number of true positive and false positive edges. Upon evaluation on sum of squared error of edge values of the inferred network with ground truth, we find that if a context is known, a class-specific average based graphical lasso shows lowest squared error, followed by networks obtained from *groupNet*. Context-agnostic average based graphical lasso shows the worst performance on both true positives vs false positives, and squared error. This experiment provides preliminary evidence that: a) *groupNet* can identify patterns of context-specific co-expression and group studies by context, and b) it can reconstruct co-expression networks while capturing context-specific edge information.

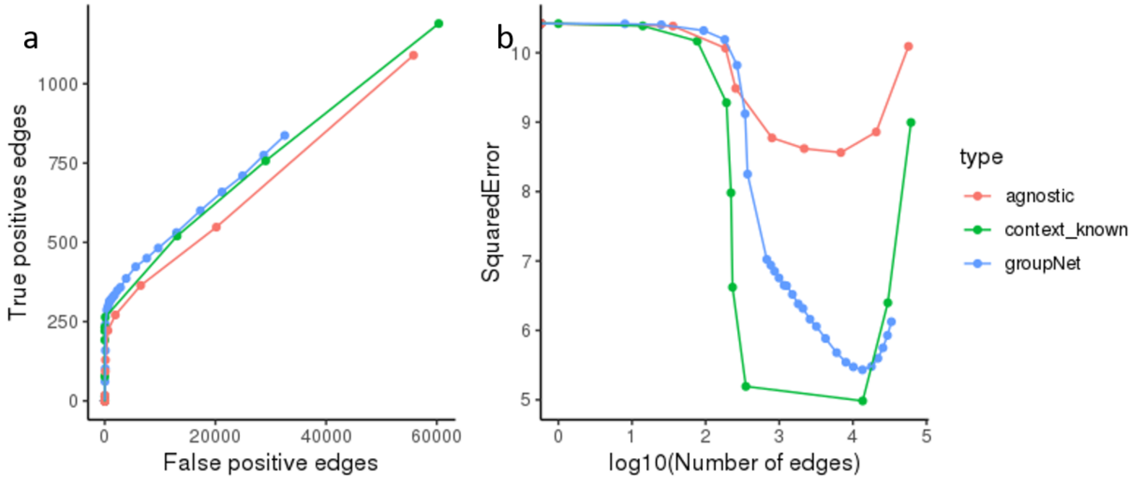


Figure 6-2. Performance of *groupNet* using simulated data. a. The number of edges correctly identified to be non-zero (TP) is plotted against the number of edges incorrectly classified as non-zero (FP) b. The sum of squared error in edge values is shown against $\log_{10}$(number of edges).

Conclusion

In this chapter we present a probabilistic method that in the absence of reliable metadata in publicly available RNAseq studies, can learn latent assignments for each study and jointly infer context-specific GCNs by sharing information across related studies. We describe the model, and derive updates for parameter inference. Finally, using a small simulation analyses, we provide preliminary evidence that that at a reasonable network density *groupNet* correctly clusters studies into context-specific groups, and shows moderate performance at recovering the structure of a context-specific co-expression network.

Chapter 7

Conclusion and future directions

In this chapter, we will first summarize the work presented in this thesis and next discuss some future direction and extension of this work.

Summary of contributions

The ability to sequence the entire human genome and to quantify expression of over 40,000 genes from hundreds of individuals provides an extraordinary opportunity to learn phenotype relevant genomic patterns that can expand our understanding of molecular and cellular processes underlying a trait. The nature of high-dimensional genomic data presents a range of computational and statistical challenges. The work in this thesis attempts to address two major difficulties in this domain: a) artifacts and noise in transcriptomic data, and b) limited statistical power.

Gene expression measurements are routinely affected by noise and artifacts that introduce spurious structure in the data that can lead to erroneous downstream conclusions. In chapter 2, we perform an extensive analysis to understand the contribution of known and latent confounders on gene expression and its effects on eQTL mapping. Next, while accounting for latent artifacts, we discovered 673 trans-eQTLs across 16 human tissues, characterized some trait associated trans-eQTLs, and hypothesize potential functional mechanisms. In chaper 3 we demonstrate that commonly used network learning methods are vulnerable to noise and artifacts in gene expression data thereby introducing a large number of false positive edges. We present a principal component based residualization method to address the effect of confounders in reconstruction of gene co-expression networks. Using empirical data, and in simulation we show that applying PC based correction prior to network learning reduces false discoveries in reconstructed networks.

In the next part of the thesis, we present methods and strategies to leverage multiple related studies to increase statistical power in transcriptomic analyes. In chapter 4 we present a multi-study integration based approach to identify global gene expression patterns underlying epithelial to mesenchymal transition (EMT) phenotype across different types of cancer cell lines. We demonstrate that leveraging data from multiple studies can enable identification of universal phenotype associated genes. Our comprehensive approach of statistical analysis and functional validation in this work identified global expression patterns in EMT and candidate regulatory genes, thereby both extending current knowledge and identifying novel drivers of EMT phenotype. In chapter 5 we present an aggregation based approach to build context-agnostic gene co-expression networks. Using data from > 250 datasets from Recount2, we sought to discover shared patterns of essential biological processes across tissues and cell types. We demonstrate that agnostic network central genes are enriched for evolutionarily constrainted genesets, and mitosis spindle formation related GO processes. We find that network central annotations also show strong enrichment for phenotypic heritability in multiple disease relevant traits such as blood and cardiovascular phenotypes in the UKBB. In chapter 6, we present a mixture model based probabilistic framework *groupNet* to reconstruct context specific gene co-expression networks by leveraging unstructed RNAseq data from public resources. Building on the hierarchical structure of biological phenotypes *groupNet* learns latent assignments for each study and jointly infers context-specific networks by sharing information across studies from related phenotypes. Using a small simulation, we show preliminary evidence that *groupNet* can capture patterns of similarity across datasets with moderate performance on network reconstruction. Overall this thesis presents a compilation of diverse projects that were driven by the motivation to efficiently capture gene regulatory patterns in the human transcriptome while addressing statistical and computational challenges that accompany this data.

Future directions

We are far from the dream of fully characterizing genetic and transcriptomic basis of gene regulation, and understanding the underpinnings trait manifestation in humans. While we have tried addressing some challenges in inferring high dimensional gene networks, it still continues to remain a daunting task. However, if one can accurately infer functional gene relationships in the human transcriptome, they can serve as powerful tools for: a) understanding the molecular and cellular basis of phenotypic variability, b) discover discover disease linked causal genes, c) identify relevant genes or groups of genes for drug discovery, d) discover interpretable biomarkers for patient stratification based on context specific patterns of differential co-expression. Next we describe three potential future directions based on the work described in this thesis.

Identifying causal gene regulatory circuits by modeling small variable sub-problem

Inferring directed gene-gene relationships can help discover causal regulatory mechanisms of gene expression. While learning the structure of a Gaussian undirected graphical model to construct co-expression network is a hard, learning the structure of the bayesian network is even more challenging. Artifacts in gene expression data, and statistical power continue to remain a hurdle, further the huge search space that grows exponentially with the number of nodes in a graph is an additional caveat with directed networks. There have been studies that propose several approaches to reduce the size of the search space for directed networks. Leveraging the topology of an undirected network can enable identification of a finite set of possible structures for a directed graph. One can utilize the topology of undirected graphs to identify small groups of genes This along with sparsity inducing structured priors can further assist in accurate inference of gene regulatory circuits as gaussian directed acyclic graphs. A potential extension to chapter 5 could be to identify causal relationships in context-agnostic processes identified in the study.

Integrating genomic annotations to build a multi-modal probabilistic model for co-expression

Our work in chapter 5 using stratified LD score regression demonstrates that while network based annotations were enriched for explaining trait heritability, their individual contribution was almost completely explained by the 97 baseline LD annotations. Several studies have shown the value of integrating genomic annotations to improve variant interpretation and to predict the deleterious impact of common and rare genetic variation. While there have been efforts to integrate multi-modal genomic data to construct co-expression networks, leveraging genomic annotations to construct structured priors in a probabilistic framework may have the potential to enable inference of more accurate gene-gene relationships.

Network based models for clinical genomics

Disease biomarkers play a vital role at several stages of clinical decision making such as: i) predict patient response to treatment or intervention (predictive), ii) predict patient outcomes (prognostic), and iii) identify if patient has a specific disease or subtype of a disease (diagnostic). Despite the abundance of nucleic acid biomarkers reported in literature, there has been little success in translating them to the clinic.

Further, most of the currently used Laboratory Developed Tests are not able to explain the underlying molecular mechanism disease manifestation. The probabilistic model that we described in chapter 6 can be extended to build a network based classifier for patient stratification.

References

[1] L. Hood and D. Galas, “The digital code of dna,” *Nature*, vol. 421, no. 6921, pp. 444–448, 2003.

[2] F. Martini, J. L. Nath, E. F. Bartholomew, W. C. Ober, C. E. Ober, K. Welch, and R. T. Hutchings, *Fundamentals of anatomy & physiology*, vol. 7. Pearson Benjamin Cummings San Francisco, CA, 2006.

[3] F. CRICK, “Central dogma of molecular biology,” *Nature*, vol. 227, pp. 561–563, 08 1970.

[4] K. R. Kukurba and S. B. Montgomery, “Rna sequencing and analysis,” *Cold Spring Harbor Protocols*, vol. 2015, no. 11, pp. pdb–top084970, 2015.

[5] X. Yu, J. Lin, D. J. Zack, and J. Qian, “Identification of tissue-specific cis-regulatory modules based on interactions between transcription factors,” *BMC bioinformatics*, vol. 8, no. 1, p. 437, 2007.

[6] E. Pierson, D. Koller, A. Battle, S. Mostafavi, G. Consortium, *et al.*, “Sharing and specificity of co-expression networks across 35 human tissues,” *PLoS computational biology*, vol. 11, no. 5, p. e1004220, 2015.

[7] A. Battle, S. Mostafavi, X. Zhu, J. B. Potash, M. M. Weissman, C. McCormick, C. D. Haudenschild, K. B. Beckman, J. Shi, R. Mei, A. E. Urban, S. B. Montgomery, D. F. Levinson, and D. Koller, “Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals,” *Genome Research*, vol. 24, pp. 14–24, jan 2014.

[8] M. D. Gallagher and A. S. Chen-Plotkin, “The post-gwas era: from association to function,” *The American Journal of Human Genetics*, vol. 102, no. 5, pp. 717–730, 2018.

[9] T. Lappalainen, M. Sammeth, M. R. Friedländer, P. AC‘t Hoen, J. Monlong, M. A. Rivas, M. Gonzalez-Porta, N. Kurbatova, T. Griebel, P. G. Ferreira, Others, P. A. C. ‘t Hoen, J. Monlong, M. A. Rivas, M. Gonzàlez-Porta, N. Kurbatova, T. Griebel, P. G. Ferreira, M. Barann, T. Wieland, L. Greger, M. van Iterson, J. Almlöf, P. Ribeca, I. Pulyakhina, D. Esser, T. Giger, A. Tikhonov, M. Sultan, G. Bertier, D. G. MacArthur, M. Lek, E. Lizano, H. P. J. Buermans, I. Padioleau, T. Schwarzmayr, O. Karlberg, H. Ongen, H. Kilpinen, S. Beltran, M. Gut, K. Kahlem, V. Amstislavskiy, O. Stegle, M. Pirinen, S. B. Montgomery, P. Donnelly, M. I. McCarthy, P. Flicek, T. M. Strom, The Geuvadis Consortium, H. Lehrach, S. Schreiber, R. Sudbrak, Á. Carracedo, S. E. Antonarakis, R. Häsler, A.-C. Syvänen, G.-J. van Ommen, A. Brazma, T. Meitinger, P. Rosenstiel, R. Guigó, I. G. Gut, X. Estivill, E. T. Dermitzakis, P. A. C. ’t Hoen, J. Monlong, M. A. Rivas, M. Gonzàlez-Porta, N. Kurbatova, T. Griebel, P. G. Ferreira, M. Barann, T. Wieland, L. Greger, M. van Iterson, J. Almlöf, P. Ribeca, I. Pulyakhina, D. Esser, T. Giger, A. Tikhonov, M. Sultan, G. Bertier, D. G. MacArthur, M. Lek, E. Lizano, H. P. J. Buermans, I. Padioleau, T. Schwarzmayr, O. Karlberg, H. Ongen, H. Kilpinen, S. Beltran, M. Gut, K. Kahlem, V. Amstislavskiy, O. Stegle, M. Pirinen, S. B. Montgomery, P. Donnelly, M. I. McCarthy, P. Flicek, T. M. Strom, Geuvadis Consortium, H. Lehrach, S. Schreiber, R. Sudbrak, A. Carracedo, S. E. Antonarakis, R. Häsler, A.-C. Syvänen, G.-J. van Ommen, A. Brazma, T. Meitinger, P. Rosenstiel, R. Guigó, I. G. Gut, X. Estivill, and E. T. Dermitzakis, “Transcriptome and genome sequencing uncovers functional variation in humans,” *Nature*, vol. 501, pp. 506–511, sep 2013.

[10] A. C. Nica and E. T. Dermitzakis, "Expression quantitative trait loci: present and future," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 368, no. 1620, p. 20120362, 2013.

[11] N. Shan, Z. Wang, and L. Hou, "Identification of trans-eqtls using mediation analysis with multiple mediators," *BMC bioinformatics*, vol. 20, no. 3, p. 126, 2019.

[12] L. I. Furlong, "Human diseases through the lens of network biology," *Trends in genetics*, vol. 29, no. 3, pp. 150–159, 2013.

[13] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," *Nature reviews genetics*, vol. 12, no. 1, pp. 56–68, 2011.

[14] A. Saha, Y. Kim, A. D. Gewirtz, B. Jo, C. Gao, I. C. McDowell, B. E. Engelhardt, A. Battle, F. Aguet, K. G. Ardlie, *et al.*, "Co-expression networks reveal the tissue-specific regulation of transcription and splicing," *Genome research*, vol. 27, no. 11, pp. 1843–1858, 2017.

[15] Y. Yang, L. Han, Y. Yuan, J. Li, N. Hei, and H. Liang, "Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types," *Nature communications*, vol. 5, no. 1, pp. 1–9, 2014.

[16] J. T. Leek and J. D. Storey, "Capturing heterogeneity in gene expression studies by surrogate variable analysis," *PLOS Genetics*, vol. 3, p. e161, sep 2007.

[17] O. Stegle, L. Parts, R. Durbin, and J. Winn, "A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies," *PLoS Computational Biology*, vol. 6, no. 5, p. e1000770, 2010.

[18] C. Chen, K. Grennan, J. Badner, D. Zhang, E. Gershon, L. Jin, and C. Liu, "Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods," *PloS one*, vol. 6, no. 2, 2011.

[19] J. T. Leek, R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, and R. A. Irizarry, "Tackling the widespread and critical impact of batch effects in high-throughput data," *Nature Reviews Genetics*, vol. 11, no. 10, pp. 733–739, 2010.

[20] G. Consortium *et al.*, "Genetic effects on gene expression across human tissues," *Nature*, vol. 550, no. 7675, pp. 204–213, 2017.

[21] P. Parsana, C. Ruberman, A. E. Jaffe, M. C. Schatz, A. Battle, and J. T. Leek, "Addressing confounding artifacts in reconstruction of gene co-expression networks," *Genome biology*, vol. 20, no. 1, pp. 1–6, 2019.

[22] P. Parsana, S. R. Amend, J. Hernandez, K. J. Pienta, and A. Battle, "Identifying global expression patterns and key regulators in epithelial to mesenchymal transition through multi-study integration," *BMC cancer*, vol. 17, no. 1, p. 447, 2017.

[23] L. Collado-Torres, A. Nellore, K. Kammers, S. E. Ellis, M. A. Taub, K. D. Hansen, A. E. Jaffe, B. Langmead, and J. T. Leek, "Reproducible rna-seq analysis using recount2," *Nature biotechnology*, vol. 35, no. 4, pp. 319–321, 2017.

[24] H.-J. Westra, M. J. Peters, T. Esko, H. Yaghootkar, C. Schurmann, J. Kettunen, M. W. Christiansen, B. P. Fairfax, K. Schramm, J. E. Powell, A. Zhernakova, D. V. Zhernakova, J. H. Veldink, L. H. Van den Berg, J. Karjalainen, S. Withoff, A. G. Uitterlinden, A. Hofman, F. Rivadeneira, P. A. C. 't Hoen, E. Reinmaa, K. Fischer, M. Nelis, L. Milani, D. Melzer, L. Ferrucci, A. B. Singleton, D. G. Hernandez, M. A. Nalls, G. Homuth, M. Nauck, D. Radke, U. Völker, M. Perola, V. Salomaa, J. Brody, A. Suchy-Dicey, S. A. Gharib, D. A. Enquobahrie, T. Lumley, G. W. Montgomery, S. Makino, H. Prokisch, C. Herder, M. Roden, H. Grallert, T. Meitinger, K. Strauch, Y. Li, R. C. Jansen, P. M. Visscher, J. C. Knight, B. M. Psaty, S. Ripatti, A. Teumer, T. M. Frayling, A. Metspalu, J. B. J.

van Meurs, and L. Franke, “Systematic identification of trans eQTLs as putative drivers of known disease associations,” *Nature Genetics*, vol. 45, pp. 1238–1243, oct 2013.

[25] E. Grundberg, K. S. Small, Å. K. Hedman, A. C. Nica, A. Buil, S. Keildson, J. T. Bell, T.-P. Yang, E. Meduri, A. Barrett, J. Nisbett, M. Sekowska, A. Wilk, S.-Y. Shin, D. Glass, M. Travers, J. L. Min, S. Ring, K. Ho, G. Thorleifsson, A. Kong, U. Thorsteindottir, C. Ainali, A. S. Dimas, N. Hassanali, C. Ingle, D. Knowles, M. Krestyaninova, C. E. Lowe, P. Di Meglio, S. B. Montgomery, L. Parts, S. Potter, G. Surdulescu, L. Tsaprouni, S. Tsoka, V. Bataille, R. Durbin, F. O. Nestle, S. O’Rahilly, N. Soranzo, C. M. Lindgren, K. T. Zondervan, K. R. Ahmadi, E. E. Schadt, K. Stefansson, G. D. Smith, M. I. McCarthy, P. Deloukas, E. T. Dermitzakis, T. D. Spector, and Multiple Tissue Human Expression Resource (MuTHER) Consortium, “Mapping cis- and trans-regulatory effects across multiple tissues in twins,” *Nature Genetics*, vol. 44, pp. 1084–1089, oct 2012.

[26] R. S. N. Fehrmann, R. C. Jansen, J. H. Veldink, H. J. Westra, D. Arends, M. J. Bonder, J. Fu, P. Deelen, H. J. M. Groen, A. Smolonska, R. K. Weersma, R. M. W. Hofstra, W. A. Buurman, S. Rensen, M. G. M. Wolfs, M. Platteel, A. Zhernakova, C. C. Elbers, E. M. Festen, G. Trynka, M. H. Hofker, C. G. J. Saris, R. A. Ophoff, L. H. van den Berg, D. A. van Heel, C. Wijmenga, G. J. Meerman, and L. Franke, “Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA,” *PLoS Genetics*, vol. 7, p. e1002197, aug 2011.

[27] ENCODE Project Consortium, E. Birney, J. A. Stamatoyannopoulos, A. Dutta, R. Guigó, T. R. Gingeras, E. H. Margulies, Z. Weng, M. Snyder, E. T. Dermitzakis, R. E. Thurman, M. S. Kuehn, C. M. Taylor, S. Neph, C. M. Koch, S. Asthana, A. Malhotra, I. Adzhubei, J. A. Greenbaum, R. M. Andrews,

P. Flicek, P. J. Boyle, H. Cao, N. P. Carter, G. K. Clelland, S. Davis, N. Day, P. Dhami, S. C. Dillon, M. O. Dorschner, H. Fiegler, P. G. Giresi, J. Goldy, M. Hawrylycz, A. Haydock, R. Humbert, K. D. James, B. E. Johnson, E. M. Johnson, T. T. Frum, E. R. Rosenzweig, N. Karnani, K. Lee, G. C. Lefebvre, P. A. Navas, F. Neri, S. C. J. Parker, P. J. Sabo, R. Sandstrom, A. Shafer, D. Vetrie, M. Weaver, S. Wilcox, M. Yu, F. S. Collins, J. Dekker, J. D. Lieb, T. D. Tullius, G. E. Crawford, S. Sunyaev, W. S. Noble, I. Dunham, F. Denoeud, A. Reymond, P. Kapranov, J. Rozowsky, D. Zheng, R. Castelo, A. Frankish, J. Harrow, S. Ghosh, A. Sandelin, I. L. Hofacker, R. Baertsch, D. Keefe, S. Dike, J. Cheng, H. A. Hirsch, E. A. Sekinger, J. Lagarde, J. F. Abril, A. Shahab, C. Flamm, C. Fried, J. Hackermüller, J. Hertel, M. Lindemeyer, K. Missal, A. Tanzer, S. Washietl, J. Korbel, O. Emanuelsson, J. S. Pedersen, N. Holroyd, R. Taylor, D. Swarbreck, N. Matthews, M. C. Dickson, D. J. Thomas, M. T. Weirauch, J. Gilbert, J. Drenkow, I. Bell, X. Zhao, K. G. Srinivasan, W.-K. Sung, H. S. Ooi, K. P. Chiu, S. Foissac, T. Alioto, M. Brent, L. Pachter, M. L. Tress, A. Valencia, S. W. Choo, C. Y. Choo, C. Ucla, C. Manzano, C. Wyss, E. Cheung, T. G. Clark, J. B. Brown, M. Ganesh, S. Patel, H. Tammana, J. Chrast, C. N. Henrichsen, C. Kai, J. Kawai, U. Nagalakshmi, J. Wu, Z. Lian, J. Lian, P. Newburger, X. Zhang, P. Bickel, J. S. Mattick, P. Carninci, Y. Hayashizaki, S. Weissman, T. Hubbard, R. M. Myers, J. Rogers, P. F. Stadler, T. M. Lowe, C.-L. Wei, Y. Ruan, K. Struhl, M. Gerstein, S. E. Antonarakis, Y. Fu, E. D. Green, U. s. Karaöz, A. Siepel, J. Taylor, L. A. Liefer, K. A. Wetterstrand, P. J. Good, E. A. Feingold, M. S. Guyer, G. M. Cooper, G. Asimenos, C. N. Dewey, M. Hou, S. Nikolaev, J. I. Montoya-Burgos, A. Löytynoja, S. Whelan, F. Pardi, T. Massingham, H. Huang, N. R. Zhang, I. Holmes, J. C. Mullikin, A. Ureta-Vidal, B. Paten, M. Seringhaus, D. Church, K. Rosenbloom, W. J. Kent, E. A. Stone, NISC Comparative Sequencing Program, Baylor College of

Medicine Human Genome Sequencing Center, Washington University Genome Sequencing Center, Broad Institute, Children's Hospital Oakland Research Institute, S. Batzoglou, N. Goldman, R. C. Hardison, D. Haussler, W. Miller, A. Sidow, N. D. Trinklein, Z. D. Zhang, L. Barrera, R. Stuart, D. C. King, A. Ameur, S. Enroth, M. C. Bieda, J. Kim, A. A. Bhinge, N. Jiang, J. Liu, F. Yao, V. B. Vega, C. W. H. Lee, P. Ng, A. Yang, Z. Moqtaderi, Z. Zhu, X. Xu, S. Squazzo, M. J. Oberley, D. Inman, M. A. Singer, T. A. Richmond, K. J. Munn, A. Rada-Iglesias, O. Wallerman, J. Komorowski, J. C. Fowler, P. Couttet, A. W. Bruce, O. M. Dovey, P. D. Ellis, C. F. Langford, D. A. Nix, G. Euskirchen, S. Hartman, A. E. Urban, and Kra..., "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project," *Nature*, vol. 447, pp. 799–816, jun 2007.

[28] A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, Others, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y.-C. Wu, A. R. Pfenning, X. Wang, M. ClaussnitzerYaping Liu, C. Coarfa, R. Alan Harris, N. Shoresh, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. David Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. Scott Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K.-H. Farh, S. Feizi, R. Karlic, A. A.-R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, N. Abdennur, M. Adli, M. Akerman, L. Barrera, J. Antosiewicz-Bourget, T. Ballinger, M. J. Barnes, D. Bates, R. J. A. Bell, D. A. Bennett, K. Bianco, C. Bock, P. Boyle, J. Brinchmann, P. Caballero-Campo, R. Camahort, M. J. Carrasco-Alfonso, T. Charnecki, H. Chen, Z. Chen, J. B. Cheng, S. Cho, A. Chu, W.-Y. Chung,

C. Cowan, Q. Athena Deng, V. Deshpande, M. Diegel, B. Ding, T. Durham, L. Echipare, L. Edsall, D. Flowers, O. Genbacev-Krtolica, C. Gifford, S. Gillespie, E. Giste, I. A. Glass, A. Gnirke, M. Gormley, H. Gu, J. Gu, D. A. Hafler, M. J. Hangauer, M. Hariharan, M. Hatan, E. Haugen, Y. He, S. Heimfeld, S. Herlofsen, Z. Hou, R. Humbert, R. Issner, A. R. Jackson, H. Jia, P. Jiang, A. K. Johnson, T. Kadlecek, B. Kamoh, M. Kapidzic, J. Kent, A. A.-R. Kim, M. Kleinewietfeld, S. Klugman, J. Krishnan, S. Kuan, T. Kutyavin, A.-Y. Lee, K. Lee, J. Li, N. Li, Y. Li, K. L. Ligon, S. Lin, Y. Lin, J. Liu, Y. Y. Liu, C. J. Luckey, Y. P. Ma, C. Maire, A. Marson, J. S. Mattick, M. Mayo, M. McMaster, H. Metsky, T. Mikkelsen, D. Miller, M. Miri, E. Mukame, R. P. Nagarajan, F. Neri, J. Nery, T. Nguyen, H. O'Geen, S. Paithankar, T. Papayannopoulou, M. Pelizzola, P. Plettner, N. E. Propson, S. Raghuraman, B. J. Raney, A. Raubitschek, A. P. Reynolds, H. Richards, K. Riehle, P. Rinaudo, J. F. Robinson, N. B. Rockweiler, E. Rosen, E. Rynes, J. Schein, R. Sears, T. Sejnowski, A. Shafer, L. Shen, R. Shoemaker, M. Sigaroudinia, I. Slukvin, S. Stehling-Sun, R. Stewart, S. L. Subramanian, K. Suknuntha, S. Swanson, S. Tian, H. Tilden, L.-H. L. Tsai, M. Urich, I. Vaughn, J. Vierstra, S. Vong, U. Wagner, H. Wang, T. T. Wang, Y. Wang, A. Weiss, H. Whitton, A. Wildberg, H. Witt, K.-J. Won, M. Xie, X. Xing, I. Xu, Z. Xuan, Z. Ye, C.-a. Yen, P. Yu, X. X. Zhang, X. X. Zhang, J. Zhao, Y. Zhou, J. Zhu, Y. Zhu, S. Ziegler, A. E. Beaudet, L. A. Boyer, P. L. De Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. M. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L.-H. L. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. T. Wang, M. Kellis, A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward,

A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y.-C. Wu, A. R. Pfenning, X. Wang, M. Claussnitzer, Y. Y. Liu, C. Coarfa, R. A. Harris, N. Shoresh, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K.-H. Farh, S. Feizi, R. Karlic, A. A.-R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, A. E. Beaudet, L. A. Boyer, P. L. De Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. M. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L.-H. L. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. T. Wang, M. Kellis, Roadmap Epigenomics Consortium, A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y.-C. Wu, A. R. Pfenning, X. Wang, M. Claussnitzer, Y. Y. Liu, C. Coarfa, R. A. Harris, N. Shoresh, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K.-H. Farh, S. Feizi, R. Karlic, A. A.-R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, A. E. Beaudet, L. A. Boyer, P. L. De Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. M. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L.-H. L. Tsai,

W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. T. Wang, and M. Kellis, "Integrative analysis of 111 reference human epigenomes," *Nature*, vol. 518, pp. 317–330, feb 2015.

[29] H. G. Stunnenberg, M. Hirst, S. Abrignani, D. Adams, M. de Almeida, L. Altucci, V. Amin, I. Amit, S. E. Antonarakis, S. Aparicio, T. Arima, L. Arrigoni, R. Arts, V. Asnafi, M. Esteller, J. B. Bae, K. Bassler, S. Beck, B. Berkman, B. E. Bernstein, M. Bilenky, A. Bird, C. Bock, B. Boehm, G. Bourque, C. E. Breeze, B. Brors, D. Bujold, O. Burren, M. J. Bussemakers, A. Butterworth, E. Campo, E. Carrillo-de Santa-Pau, L. Chadwick, K. M. Chan, W. Chen, T. H. Cheung, L. Chiapperino, N. H. Choi, H. R. Chung, L. Clarke, J. M. Connors, P. Cronet, J. Danesh, M. Dermitzakis, G. Drewes, P. Durek, S. Dyke, T. Dylag, C. J. Eaves, P. Ebert, R. Eils, J. Eils, C. A. Ennis, T. Enver, E. A. Feingold, B. Felder, A. Ferguson-Smith, J. Fitzgibbon, P. Flicek, R. S. Foo, P. Fraser, M. Frontini, E. Furlong, S. Gakkhar, N. Gasparoni, G. Gasparoni, D. H. Geschwind, P. Gla?ar, T. Graf, F. Grosveld, X. Y. Guan, R. Guigo, I. G. Gut, A. Hamann, B. G. Han, R. A. Harris, S. Heath, K. Helin, J. G. Hengstler, A. Heravi-Moussavi, K. Herrup, S. Hill, J. A. Hilton, B. C. Hitz, B. Horsthemke, M. Hu, J. Y. Hwang, N. Y. Ip, T. Ito, B. M. Javierre, S. Jenko, T. Jenuwein, Y. Joly, S. J. Jones, Y. Kanai, H. G. Kang, A. Karsan, A. K. Kiemer, S. C. Kim, B. J. Kim, H. H. Kim, H. Kimura, S. Kinkley, F. Klironomos, I. U. Koh, M. Kostadima, C. Kressler, R. Kreuzhuber, A. Kundaje, R. Kuppers, C. Larabell, P. Lasko, M. Lathrop, D. H. Lee, S. Lee, H. Lehrach, E. Leitao, T. Lengauer, A. Lernmark, R. D. Leslie, G. K. Leung, D. Leung, M. Loeffler, Y. Ma, A. Mai, T. Manke, E. R. Marcotte, M. A. Marra, J. H. Martens, J. I. Martin-Subero, K. Maschke, C. Merten, A. Milosavljevic, S. Minucci, T. Mitsuyama, R. A. Moore, F. Muller, A. J. Mungall, M. G. Netea, K. Nordstrom, I. Norstedt, H. Okae, V. Onuchic,

F. Ouellette, W. Ouwehand, M. Pagani, V. Pancaldi, T. Pap, T. Pastinen, R. Patel, D. S. Paul, M. J. Pazin, P. G. Pelicci, A. G. Phillips, J. Polansky, B. Porse, J. A. Pospisilik, S. Prabhakar, D. C. Procaccini, A. Radbruch, N. Rajewsky, V. Rakyan, W. Reik, B. Ren, D. Richardson, A. Richter, D. Rico, D. J. Roberts, P. Rosenstiel, M. Rothstein, A. Salhab, H. Sasaki, J. S. Satterlee, S. Sauer, C. Schacht, F. Schmidt, G. Schmitz, S. Schreiber, C. Schroder, D. Schubeler, J. L. Schultze, R. P. Schulyer, M. Schulz, M. Seifert, K. Shirahige, R. Siebert, T. Sierocinski, L. Siminoff, A. Sinha, N. Soranzo, S. Spicuglia, M. Spivakov, C. Steidl, J. S. Strattan, M. Stratton, P. Sudbeck, H. Sun, N. Suzuki, Y. Suzuki, A. Tanay, D. Torrents, F. L. Tyson, T. Ulas, S. Ullrich, T. Ushijima, A. Valencia, E. Vellenga, M. Vingron, C. Wallace, S. Wallner, J. Walter, H. Wang, S. Weber, N. Weiler, A. Weller, A. Weng, S. Wilder, S. M. Wiseman, A. R. Wu, Z. Wu, J. Xiong, Y. Yamashita, X. Yang, D. Y. Yap, K. Y. Yip, S. Yip, J. I. Yoo, D. Zerbino, and G. Zipprich, “The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery,” *Cell*, vol. 167, pp. 1145–1149, Nov 2016.

[30] F. W. Albert and L. Kruglyak, “The role of regulatory variation in complex traits and disease,” *Nature Reviews Genetics*, vol. 16, pp. 197–212, apr 2015.

[31] F. A. Wright, P. F. Sullivan, A. I. Brooks, F. Zou, W. Sun, K. Xia, V. Madar, R. Jansen, W. Chung, Y.-H. Zhou, A. Abdellaoui, S. Batista, C. Butler, G. Chen, T.-H. Chen, D. D’Ambrosio, P. Gallins, M. J. Ha, J.-J. Hottenga, S. Huang, M. Kattenberg, J. Kochar, C. M. Middeldorp, A. Qu, A. Shabalin, J. Tischfield, L. Todd, J.-Y. Tzeng, G. van Grootheest, J. M. Vink, Q. Wang, W. Wang, W. Wang, G. Willemsen, J. H. Smit, E. J. de Geus, Z. Yin, B. W. J. H. Penninx, and D. I. Boomsma, “Heritability and genomics of gene expression in peripheral blood,” *Nature Genetics*, vol. 46, pp. 430–437, may 2014.

[32] K. G. Ardlie, D. S. Deluca, A. V. Segre, T. J. Sullivan, T. R. Young, E. T.

Gelfand, C. A. Trowbridge, J. B. Maller, T. Tukiainen, M. Lek, L. D. Ward, P. Kheradpour, B. Iriarte, Y. Meng, C. D. Palmer, T. Esko, W. Winckler, J. N. Hirschhorn, M. Kellis, D. G. MacArthur, G. Getz, A. A. Shabalin, G. Li, Y.-H. Zhou, A. B. Nobel, I. Rusyn, F. A. Wright, T. Lappalainen, P. G. Ferreira, H. Ongen, M. A. Rivas, A. Battle, S. Mostafavi, J. Monlong, M. Sammeth, M. Mele, F. Reverter, J. M. Goldmann, D. Koller, R. Guigo, M. I. McCarthy, E. T. Dermitzakis, E. R. Gamazon, H. K. Im, A. Konkashbaev, D. L. Nicolae, N. J. Cox, T. Flutre, X. Wen, M. Stephens, J. K. Pritchard, Z. Tu, B. Zhang, T. Huang, Q. Long, L. Lin, J. Yang, J. Zhu, J. Liu, A. Brown, B. Mestichelli, D. Tidwell, E. Lo, M. Salvatore, S. Shad, J. A. Thomas, J. T. Lonsdale, M. T. Moser, B. M. Gillard, E. Karasik, K. Ramsey, C. Choi, B. A. Foster, J. Syron, J. Fleming, H. Magazine, R. Hasz, G. D. Walters, J. P. Bridge, M. Miklos, S. Sullivan, L. K. Barker, H. M. Traino, M. Mosavel, L. A. Siminoff, D. R. Valley, D. C. Rohrer, S. D. Jewell, P. A. Branton, L. H. Sobin, M. Barcus, L. Qi, J. McLean, P. Hariharan, K. S. Um, S. Wu, D. Tabor, C. Shive, A. M. Smith, S. A. Buia, A. H. Undale, K. L. Robinson, N. Roche, K. M. Valentino, A. Britton, R. Burges, D. Bradbury, K. W. Hambright, J. Seleski, G. E. Korzeniewski, K. Erickson, Y. Marcus, J. Tejada, M. Taherian, C. Lu, M. Basile, D. C. Mash, S. Volpi, J. P. Struewing, G. F. Temple, J. Boyer, D. Colantuoni, R. Little, S. Koester, L. J. Carithers, H. M. Moore, P. Guan, C. Compton, S. J. Sawyer, J. P. Demchok, J. B. Vaught, C. A. Rabiner, N. C. Lockhart, K. G. Ardlie, G. Getz, F. A. Wright, M. Kellis, S. Volpi, E. T. Dermitzakis, and GTEx Consortium, "The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans," *Science*, vol. 348, pp. 648–660, may 2015.

[33] 1000 Genomes Project Consortium, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, and G. R. Abecasis, "A global reference for human genetic variation,"

Nature, vol. 526, pp. 68–74, oct 2015.

[34] L. J. Carithers, K. Ardlie, M. Barcus, P. A. Branton, A. Britton, S. A. Buia, C. C. Compton, D. S. DeLuca, J. Peter-Demchok, E. T. Gelfand, P. Guan, G. E. Korzeniewski, N. C. Lockhart, C. A. Rabiner, A. K. Rao, K. L. Robinson, N. V. Roche, S. J. Sawyer, A. V. Segrè, C. E. Shive, A. M. Smith, L. H. Sobin, A. H. Undale, K. M. Valentino, J. Vaught, T. R. Young, H. M. Moore, and GTEx Consortium, "A novel approach to high-quality postmortem tissue procurement: The GTEx project," *Biopreservation and Biobanking*, vol. 13, pp. 311–319, oct 2015.

[35] J. O'Connell, D. Gurdasani, O. Delaneau, N. Pirastu, S. Ulivi, M. Cocca, M. Traglia, J. Huang, J. E. Huffman, I. Rudan, R. McQuillan, R. M. Fraser, H. Campbell, O. Polasek, G. Asiki, K. Ekoru, C. Hayward, A. F. Wright, V. Vitart, P. Navarro, J.-F. Zagury, J. F. Wilson, D. Toniolo, P. Gasparini, N. Soranzo, M. S. Sandhu, and J. Marchini, "A general approach for haplotype phasing across the full spectrum of relatedness," *PLoS Genetics*, vol. 10, p. e1004234, apr 2014.

[36] B. Howie, J. Marchini, and M. Stephens, "Genotype imputation with thousands of genomes," *G3*, vol. 1, pp. 457–470, nov 2011.

[37] 1000 Genomes Project Consortium, G. R. Abecasis, A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin, R. E. Handsaker, H. M. Kang, G. T. Marth, G. A. McVean, D. M. Altshuler, R. M. Durbin, G. R. Abecasis, D. R. Bentley, A. Chakravarti, A. G. Clark, P. Donnelly, E. E. Eichler, P. Flicek, S. B. Gabriel, R. A. Gibbs, E. D. Green, M. E. Hurles, B. M. Knoppers, J. O. Korbel, E. S. Lander, C. Lee, H. Lehrach, E. R. Mardis, G. T. Marth, G. A. McVean, D. A. Nickerson, J. P. Schmidt, S. T. Sherry, J. Wang, R. K. Wilson, R. A. Gibbs, H. Dinh, C. Kovar, S. Lee, L. Lewis, D. Muzny, J. Reid, M. Wang, J. Wang,

X. Fang, X. Guo, M. Jian, H. Jiang, X. Jin, G. Li, J. Li, Y. Li, Z. Li, X. Liu, Y. Lu, X. Ma, Z. Su, S. Tai, M. Tang, B. Wang, G. Wang, H. Wu, R. Wu, Y. Yin, W. Zhang, J. Zhao, M. Zhao, X. Zheng, Y. Zhou, E. S. Lander, D. M. Altshuler, S. B. Gabriel, N. Gupta, P. Flicek, L. Clarke, R. Leinonen, R. E. Smith, X. Zheng-Bradley, D. R. Bentley, R. Grocock, S. Humphray, T. James, Z. Kingsbury, H. Lehrach, R. Sudbrak, M. W. Albrecht, V. S. Amstislavskiy, T. A. Borodina, M. Lienhard, F. Mertes, M. Sultan, B. Timmermann, M. L. Yaspo, S. T. Sherry, G. A. McVean, E. R. Mardis, R. K. Wilson, L. Fulton, R. Fulton, G. M. Weinstock, R. M. Durbin, S. Balasubramaniam, J. Burton, P. Danecek, T. M. Keane, A. Kolb-Kokocinski, S. McCarthy, J. Stalker, M. Quail, J. P. Schmidt, C. J. Davies, J. Gollub, T. Webster, B. Wong, Y. Zhan, A. Auton, R. A. Gibbs, F. Yu, M. Bainbridge, D. Challis, U. S. Evani, J. Lu, D. Muzny, U. Nagaswamy, J. Reid, A. Sabo, Y. Wang, J. Yu, J. Wang, L. J. Coin, L. Fang, X. Guo, X. Jin, G. Li, Q. Li, Y. Li, Z. Li, H. Lin, B. Liu, R. Luo, N. Qin, H. Shao, B. Wang, Y. Xie, C. Ye, C. Yu, F. Zhang, H. Zheng, H. Zhu, G. T. Marth, E. P. Garrison, D. Kural, W. P. Lee, W. F. Leong, A. N. Ward, J. Wu, M. Zhang, C. Lee, L. Griffin, C. H. Hsieh, R. E. Mills, X. Shi, M. von Grotthuss, C. Zhang, M. J. Daly, M. A. DePristo, D. M. Altshuler, E. Banks, G. Bhatia, M. O. Carneiro, G. del Angel, S. B. Gabriel, G. Genovese, N. Gupta, R. E. Handsaker, C. Hartl, E. S. Lander, S. A. McCarroll, J. C. Nemesh, R. E. Poplin, S. F. Schaffner, K. Shakir, S. C. Yoon, J. Lihm, V. Makarov, H. Jin, W. Kim, K. C. Kim, J. O. Korbel, T. Rausch, P. Flicek, K. Beal, L. Clarke, F. Cunningham, J. Herrero, W. M. McLaren, G. R. Ritchie, R. E. Smith, X. Zheng-Bradley, A. G. Clark, S. Gottipati, A. Keinan, J. L. Rodriguez-Flores, P. C. Sabeti, S. R. Grossman, S. Tabrizi, R. Tariyal, D. N. Cooper, E. V. Ball, P. D. Stenson, D. R. Bentley, B. Barnes, M. Bauer, R. Cheetham, T. Cox, M. Eberle, S. Humphray, S. Kahn, L. Murray, J. Peden, R. Shaw, K. Ye, M. A. Batzer, M. K. Konkel,

J. A. Walker, D. G. MacArthur, M. Lek, R. Sudbrak, V. S. Amstislavskiy, R. Herwig, M. D. Shriver, C. D. Bustamante, J. K. Byrnes, F. M. De La Vega, S. Gravel, E. E. Kenny, J. M. Kidd, P. Lacroute, B. K. Maples, A. Moreno-Estrada, F. Zakharia, E. Halperin, Y. Baran, D. W. Craig, A. Christoforides, N. Homer, T. Izatt, A. A. Kurdoglu, S. A. Sinari, K. Squire, S. T. Sherry, C. Xiao, J. Sebat, V. Bafna, K. Ye, E. G. Burchard, R. D. Hernandez, C. R. Gignoux, D. Haussler, S. J. Katzman, W. J. Kent, B. Howie, A. Ruiz-Linares, E. T. Dermitzakis, T. Lappalainen, S. E. Devine, X. Liu, A. Maroo, L. J. Tallon, J. A. Rosenfeld, L. P. Michelson, G. R. Abecasis, H. M. Kang, P. Anderson, A. Angius, A. Bigham, T. Blackwell, F. Busonero, F. Cucca, C. Fuchsberger, C. Jones, G. Jun, Y. Li, R. Lyons, A. Maschio, E. Porcu, F. Reinier, S. Sanna, D. Schlessinger, C. Sidore, A. Tan, M. K. Trost, P. Awadalla, A. Hodgkinson, G. Lunter, G. A. McVean, J. L. Marchini, S. Myers, C. Churchhouse, O. Delaneau, A. Gupta-Hinch, Z. Iqbal, I. Mathieson, A. Rimmer, D. K. Xifara, T. K. Oleksyk, Y. Fu, X. Liu, M. Xiong, L. Jorde, D. Witherspoon, J. Xing, E. E. Eichler, B. L. Browning, C. Alkan, I. Hajirasouliha, F. Hormozdiari, A. Ko, P. H. Sudmant, E. R. Mardis, K. Chen, A. Chinwalla, L. Ding, D. Dooling, D. C. Koboldt, M. D. McLellan, J. W. Wallis, M. C. Wendl, Q. Zhang, R. M. Durbin, M. E. Hurles, C. Tyler-Smith, C. A. Albers, Q. Ayub, S. Balasubramaniam, Y. Chen, A. J. Coffey, V. Colonna, P. Danecek, N. Huang, L. Jostins, T. M. Keane, H. Li, S. McCarthy, A. Scally, J. Stalker, K. Walter, Y. Xue, Y. Zhang, M. B. Gerstein, A. Abyzov, S. Balasubramanian, J. Chen, D. Clarke, Y. Fu, L. Habegger, A. O. Harmanci, M. Jin, E. Khurana, X. J. Mu, C. Sisu, Y. Li, R. Luo, H. Zhu, C. Lee, L. Griffin, C. H. Hsieh, R. E. Mills, X. Shi, M. von Grotthuss, C. Zhang, G. T. Marth, E. P. Garrison, D. Kural, W. P. Lee, A. N. Ward, J. Wu, M. Zhang, S. A. McCarroll, D. M. Altshuler, E. Banks, G. del Angel, G. Genovese, R. E. Handsaker, C. Hartl, J. C. Nemesh, K. Shakir,

S. C. Yoon, J. Lihm, V. Makarov, J. Degenhardt, P. Flicek, L. Clarke, R. E. Smith, X. Zheng-Bradley, J. O. Korbel, T. Rausch, A. M. Stutz, D. R. Bentley, B. Barnes, R. Cheetham, M. Eberle, S. Humphray, S. Kahn, L. Murray, R. Shaw, K. Ye, M. A. Batzer, M. K. Konkel, J. A. Walker, P. Lacroute, D. W. Craig, N. Homer, D. Church, C. Xiao, J. Sebat, V. Bafna, J. J. Michaelson, K. Ye, S. E. Devine, X. Liu, A. Maroo, L. J. Tallon, G. Lunter, Z. Iqbal, D. Witherspoon, J. Xing, E. E. Eichler, C. Alkan, I. Hajirasouliha, F. Hormozdiari, A. Ko, P. H. Sudmant, K. Chen, A. Chinwalla, L. Ding, M. D. McLellan, J. W. Wallis, M. E. Hurles, B. Blackburne, H. Li, S. J. Lindsay, Z. Ning, A. Scally, K. Walter, Y. Zhang, M. B. Gerstein, A. Abyzov, J. Chen, D. Clarke, E. Khurana, X. J. Mu, C. Sisu, R. A. Gibbs, F. Yu, M. Bainbridge, D. Challis, U. S. Evani, C. Kovar, L. Lewis, J. Lu, D. Muzny, U. Nagaswamy, J. Reid, A. Sabo, J. Yu, X. Guo, Y. Li, R. Wu, G. T. Marth, E. P. Garrison, W. F. Leong, A. N. Ward, G. del Angel, M. A. DePristo, S. B. Gabriel, N. Gupta, C. Hartl, R. E. Poplin, A. G. Clark, J. L. Rodriguez-Flores, P. Flicek, L. Clarke, R. E. Smith, X. Zheng-Bradley, D. G. MacArthur, C. D. Bustamante, S. Gravel, D. W. Craig, A. Christoforides, N. Homer, T. Izatt, S. T. Sherry, C. Xiao, E. T. Dermitzakis, G. R. Abecasis, H. M. Kang, G. A. McVean, E. R. Mardis, D. Dooling, L. Fulton, R. Fulton, D. C. Koboldt, R. M. Durbin, S. Balasubramaniam, T. M. Keane, S. McCarthy, J. Stalker, M. B. Gerstein, S. Balasubramanian, L. Habegger, E. P. Garrison, R. A. Gibbs, M. Bainbridge, D. Muzny, F. Yu, J. Yu, G. del Angel, R. E. Handsaker, V. Makarov, J. L. Rodriguez-Flores, H. Jin, W. Kim, K. C. Kim, P. Flicek, K. Beal, L. Clarke, F. Cunningham, J. Herrero, W. M. McLaren, G. R. Ritchie, X. Zheng-Bradley, S. Tabrizi, D. G. MacArthur, M. Lek, C. D. Bustamante, F. M. De La Vega, D. W. Craig, A. A. Kurdoglu, T. Lappalainen, J. A. Rosenfeld, L. P. Michelson, P. Awadalla, A. Hodgkinson, G. A. McVean, K. Chen, C. Tyler-Smith, Y. Chen, V. Colonna, A. Frankish, J. Harrow, Y. Xue,

M. B. Gerstein, A. Abyzov, S. Balasubramanian, J. Chen, D. Clarke, Y. Fu, A. O. Harmanci, M. Jin, E. Khurana, X. J. Mu, C. Sisu, R. A. Gibbs, G. Fowler, W. Hale, D. Kalra, C. Kovar, D. Muzny, J. Reid, J. Wang, X. Guo, G. Li, Y. Li, X. Zheng, D. M. Altshuler, P. Flicek, L. Clarke, J. Barker, G. Kelman, E. Kulesha, R. Leinonen, W. M. McLaren, R. Radhakrishnan, A. Roa, D. Smirnov, R. E. Smith, I. Streeter, I. Toneva, B. Vaughan, X. Zheng-Bradley, D. R. Bentley, T. Cox, S. Humphray, S. Kahn, R. Sudbrak, M. W. Albrecht, M. Lienhard, D. W. Craig, T. Izatt, A. A. Kurdoglu, S. T. Sherry, V. Ananiev, Z. Belaia, D. Beloslyudtsev, N. Bouk, C. Chen, D. Church, R. Cohen, C. Cook, J. Garner, T. Hefferon, M. Kimelman, C. Liu, J. Lopez, P. Meric, C. O'Sullivan, Y. Ostapchuk, L. Phan, S. Ponomarov, V. Schneider, E. Shekhtman, K. Sirotkin, D. Slotta, C. Xiao, H. Zhang, D. Haussler, G. R. Abecasis, G. A. McVean, C. Alkan, A. Ko, D. Dooling, R. M. Durbin, S. Balasubramaniam, T. M. Keane, S. McCarthy, J. Stalker, A. Chakravarti, B. M. Knoppers, G. R. Abecasis, K. C. Barnes, C. Beiswanger, E. G. Burchard, C. D. Bustamante, H. Cai, H. Cao, R. M. Durbin, N. Gharani, R. A. Gibbs, C. R. Gignoux, S. Gravel, B. Henn, D. Jones, L. Jorde, J. S. Kaye, A. Keinan, A. Kent, A. Kerasidou, Y. Li, R. Mathias, G. A. McVean, A. Moreno-Estrada, P. N. Ossorio, M. Parker, D. Reich, C. N. Rotimi, C. D. Royal, K. Sandoval, Y. Su, R. Sudbrak, Z. Tian, B. Timmermann, S. Tishkoff, L. H. Toji, C. Tyler-Smith, M. Via, Y. Wang, H. Yang, L. Yang, J. Zhu, W. Bodmer, G. Bedoya, A. Ruiz-Linares, C. Z. Ming, G. Yang, C. J. You, L. Peltonen, A. Garcia-Montero, A. Orfao, J. Dutil, J. C. Martinez-Cruzado, T. K. Oleksyk, L. D. Brooks, A. L. Felsenfeld, J. E. McEwen, N. C. Clemm, A. Duncanson, M. Dunn, E. D. Green, M. S. Guyer, and J. L. Peterson, "An integrated map of genetic variation from 1,092 human genomes," *Nature*, vol. 491, pp. 56–65, nov 2012.

[38] A. A. Shabalin, "Matrix eQTL: Ultra fast eQTL analysis via large matrix

operations," *Bioinformatics*, vol. 28, pp. 1353–1358, may 2012.

[39] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B*, vol. 57, no. 1, pp. 289–300, 1995.

[40] J. H. Sul, B. Han, C. Ye, T. Choi, and E. Eskin, "Effectively identifying eQTLs from multiple tissues by combining mixed model and meta-analytic approaches," *PLOS Genetics*, vol. 9, p. e1003491, jun 2013.

[41] B. Han and E. Eskin, "Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies," *American Journal of Human Genetics*, vol. 88, pp. 586–598, may 2011.

[42] M. Bogomolov, C. B. Peterson, Y. Benjamini, and C. Sabatti, "Testing hypotheses on a tree: new error rates and controlling strategies," *arXiv preprint arXiv:1705.07529*, May 2017.

[43] C. Giambartolomei, D. Vukcevic, E. E. Schadt, L. Franke, A. D. Hingorani, C. Wallace, and V. Plagnol, "Bayesian test for colocalisation between pairs of genetic association studies using summary statistics," *PLoS Genetics*, vol. 10, p. e1004383, may 2014.

[44] A. Saha and A. Battle, "False positives in trans-eqtl and co-expression analyses arising from rna-sequencing alignment errors," *F1000Research*, vol. 7, 2018.

[45] O. Stegle, L. Parts, R. Durbin, and J. Winn, "A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eqtl studies," *PLoS Comput Biol*, vol. 6, 2010.

[46] C. B. Peterson, M. Bogomolov, Y. Benjamini, and C. Sabatti, "TreeQTL: Hierarchical error control for eQTL findings," *Bioinformatics*, vol. 32, pp. 2556–2558, aug 2016.

[47] A. Buil, A. Viñuela, A. Brown, M. Davies, I. Padioleau, D. Bielser, L. Romano,

D. Glass, P. Di Meglio, K. Small, T. Spector, and E. T. Dermitzakis, "Quantifying the degree of sharing of genetic and non-genetic causes of gene expression variability across four tissues," *bioRxiv*, p. 53355, may 2016.

[48] V. Hore, A. Viñuela, A. Buil, J. Knight, M. I. McCarthy, K. Small, and J. Marchini, "Tensor decomposition for multiple-tissue gene expression experiments," *Nature Genetics*, vol. 48, pp. 1094–1100, sep 2016.

[49] S. B. Montgomery and E. T. Dermitzakis, "From expression QTLs to personalized transcriptomics," *Nature Reviews Genetics*, vol. 12, pp. 277–282, apr 2011.

[50] J. Bryois, A. Buil, D. M. Evans, J. P. Kemp, S. B. Montgomery, D. F. Conrad, K. M. Ho, S. Ring, M. Hurles, P. Deloukas, Others, G. Davey Smith, and E. T. Dermitzakis, "Cis and trans effects of human genomic variants on gene expression," *PLoS Genetics*, vol. 10, p. e1004461, jul 2014.

[51] T. Huan, T. Esko, M. J. Peters, L. C. Pilling, K. Schramm, C. Schurmann, B. H. Chen, C. Liu, R. Joehanes, A. D. Johnson, C. Yao, S.-x. Ying, P. Courchesne, L. Milani, N. Raghavachari, R. Wang, P. Liu, E. Reinmaa, A. Dehghan, A. Hofman, A. G. Uitterlinden, D. G. Hernandez, S. Bandinelli, A. Singleton, D. Melzer, A. Metspalu, M. Carstensen, H. Grallert, C. Herder, T. Meitinger, A. Peters, M. Roden, M. Waldenberger, M. Dörr, S. B. Felix, T. Zeller, I. C. f. B. P. G. (icbp), R. Vasan, C. J. O'Donnell, P. J. Munson, X. Yang, H. Prokisch, U. Völker, J. B. J. van Meurs, L. Ferrucci, D. Levy, J. B. J. van Meurs, L. Ferrucci, D. Levy, and Others, "A meta-analysis of gene expression signatures of blood pressure and hypertension," *PLOS Genetics*, vol. 11, p. e1005035, mar 2015.

[52] D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorff, and H. Parkinson, "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations," *Nucleic Acids Research*,

vol. 42, pp. D1001–D1006, jan 2014.

[53] A. C. Lidral, H. Liu, S. A. Bullard, G. Bonde, J. Machida, A. Visel, L. M. Uribe, X. Li, B. Amendt, and R. A. Cornell, "A single nucleotide polymorphism associated with isolated cleft lip and palate, thyroid cancer and hypothyroidism alters the activity of an oral epithelium and thyroid enhancer near FOXE1," *Human Molecular Genetics*, vol. 24, pp. 3895–3907, jul 2015.

[54] N. Eriksson, J. Y. Tung, A. K. Kiefer, D. A. Hinds, U. Francke, J. L. Mountain, and C. B. Do, "Novel associations for hypothyroidism include known autoimmune risk loci," *PLoS ONE*, vol. 7, no. 4, p. e34442, 2012.

[55] J. C. Denny, D. C. Crawford, M. D. Ritchie, S. J. Bielinski, M. A. Basford, Y. Bradford, H. S. Chai, L. Bastarache, R. Zuvich, P. Peissig, D. Carrell, A. H. Ramirez, J. Pathak, R. A. Wilke, L. Rasmussen, X. Wang, J. A. Pacheco, A. N. Kho, M. G. Hayes, N. Weston, M. Matsumoto, P. A. Kopp, K. M. Newton, G. P. Jarvik, R. Li, T. A. Manolio, I. J. Kullo, C. G. Chute, R. L. Chisholm, E. B. Larson, C. A. McCarty, D. R. Masys, D. M. Roden, and M. de Andrade, "Variants near FOXE1 are associated with hypothyroidism and other thyroid conditions: using electronic medical records for genome- and phenome-wide studies," *American Journal of Human Genetics*, vol. 89, pp. 529–542, oct 2011.

[56] M. De Felice, C. Ovitt, E. Biffali, a. Rodriguez-Mallon, C. Arra, K. Anastassiadis, P. E. Macchia, M. G. Mattei, a. Mariano, H. Schöler, V. Macchia, and R. Di Lauro, "A mouse model for hereditary thyroid dysgenesis and cleft palate," *Nature Genetics*, vol. 19, pp. 395–398, aug 1998.

[57] N. Agrawal, R. Akbani, B. A. Aksoy, A. Ally, H. Arachchi, S. L. Asa, J. T. Auman, M. Balasundaram, S. Balu, S. B. Baylin, M. Behera, B. Bernard, R. Beroukhim, J. A. Bishop, A. D. Black, T. Bodenheimer, L. Boice, M. S. Bootwalla, J. Bowen, R. Bowlby, C. A. Bristow, R. Brookens, D. Brooks, R. Bryant, E. Buda, Y. S. N.

Butterfield, T. Carling, R. Carlsen, S. L. Carter, S. E. Carty, T. A. Chan, A. Y. Chen, A. D. Cherniack, D. Cheung, L. Chin, J. Cho, A. Chu, E. Chuah, K. Cibulskis, G. Ciriello, A. Clarke, G. L. Clayman, L. Cope, J. A. Copland, K. Covington, L. Danilova, T. Davidsen, J. A. Demchok, D. DiCara, N. Dhalla, R. Dhir, S. S. Dookran, G. Dresdner, J. Eldridge, G. Eley, A. K. El-Naggar, S. Eng, J. A. Fagin, T. Fennell, R. L. Ferris, S. Fisher, S. Frazer, J. Frick, S. B. Gabriel, I. Ganly, J. Gao, L. A. Garraway, J. M. Gastier-Foster, G. Getz, N. Gehlenborg, R. Ghossein, R. A. Gibbs, T. J. Giordano, K. Gomez-Hernandez, J. Grimsby, B. Gross, R. Guin, A. Hadjipanayis, H. A. Harper, D. N. Hayes, D. I. Heiman, J. G. Herman, K. A. Hoadley, M. Hofree, R. A. Holt, A. P. Hoyle, F. W. Huang, M. Huang, C. M. Hutter, T. Ideker, L. Iype, A. Jacobsen, S. R. Jefferys, C. D. Jones, S. J. M. Jones, K. Kasaian, E. Kebebew, F. R. Khuri, J. Kim, R. Kramer, R. Kreisberg, R. Kucherlapati, D. J. Kwiatkowski, M. Ladanyi, P. H. Lai, P. W. Laird, E. Lander, M. S. Lawrence, D. Lee, E. Lee, S. Lee, W. Lee, K. M. Leraas, T. M. Lichtenberg, L. Lichtenstein, P. Lin, S. Ling, J. Liu, W. Liu, Y. Liu, V. A. LiVolsi, Y. Lu, Y. Ma, H. S. Mahadeshwar, M. A. Marra, M. Mayo, D. G. McFadden, S. Meng, M. Meyerson, P. A. Mieczkowski, M. Miller, G. Mills, R. A. Moore, L. E. Mose, A. J. Mungall, B. A. Murray, Y. E. Nikiforov, M. S. Noble, A. I. Ojesina, T. K. Owonikoko, B. A. Ozenberger, A. Pantazi, M. Parfenov, P. J. Park, J. S. Parker, E. O. Paull, C. S. Pedamallu, C. M. Perou, J. F. Prins, A. Protopopov, S. S. Ramalingam, N. C. Ramirez, R. Ramirez, B. J. Raphael, W. K. Rathmell, X. Ren, S. M. Reynolds, E. Rheinbay, M. D. Ringel, M. Rivera, J. Roach, A. G. Robertson, M. W. Rosenberg, M. Rosenthal, S. Sadeghi, G. Saksena, C. Sander, N. Santoso, J. E. Schein, N. Schultz, S. E. Schumacher, R. R. Seethala, J. Seidman, Y. Senbabaoglu, S. Seth, S. Sharpe, K. R. M. Shaw, J. P. Shen, R. Shen, S. Sherman, M. Sheth, Y. Shi, I. Shmulevich, G. L. Sica, J. V. Simons, R. Sinha, P. Sipahimalani, R. C. Smallridge, H. J. Sofia,

M. G. Soloway, X. Song, C. Sougnez, C. Stewart, P. Stojanov, J. M. Stuart, S. O. Sumer, Y. Sun, B. Tabak, A. Tam, D. Tan, J. Tang, R. Tarnuzzer, B. S. Taylor, N. Thiessen, L. Thorne, V. Thorsson, R. M. Tuttle, C. B. Umbricht, D. J. Van Den Berg, F. Vandin, U. Veluvolu, R. G. W. Verhaak, M. Vinco, D. Voet, V. Walter, Z. Wang, S. Waring, P. M. Weinberger, N. Weinhold, J. N. Weinstein, D. J. Weisenberger, D. Wheeler, M. D. Wilkerson, J. Wilson, M. Williams, D. A. Winer, L. Wise, J. Wu, L. Xi, A. W. Xu, L. L. Yang, L. L. Yang, T. I. Zack, M. A. Zeiger, D. Zeng, J. C. Zenklusen, N. Zhao, H. Zhang, J. J. Zhang, J. J. Zhang, W. Zhang, E. Zmuda, L. Zou, C. G. A. R. Network, and Others, "Integrated genomic characterization of papillary thyroid carcinoma," *Cell*, vol. 159, pp. 676–690, oct 2014.

[58] T. Taniguchi, K. Ogasawara, A. Takaoka, and N. Tanaka, "IRF family of transcription factors as regulators of host defense," *Annual Review of Immunology*, vol. 19, pp. 623–655, 2001.

[59] L. C. White, K. L. Wright, N. J. Felix, H. Ruffner, L. F. L. Reis, R. Pine, and J. P. Y. Ting, "Regulation of LMP2 and TAP1 genes by IRF-1 explains the paucity of CD8+ T cells in IRF-1(-/-) mice," *Immunity*, vol. 5, pp. 365–376, oct 1996.

[60] J. M. Penninger, C. Sirard, H. W. Mittrücker, A. Chidgey, I. Kozieradzki, M. Nghiem, A. Hakem, T. Kimura, E. Timms, R. Boyd, T. Taniguchi, T. Matsuyama, and T. W. Mak, "The interferon regulatory transcription factor IRF-1 controls positive and negative selection of CD8+ thymocytes," *Immunity*, vol. 7, pp. 243–254, aug 1997.

[61] P. K. M. Kim, M. Armstrong, Y. Liu, P. Yan, B. Bucher, B. S. Zuckerbraun, A. Gambotto, T. R. Billiar, and J. H. Yim, "IRF-1 expression induces apoptosis and inhibits tumor growth in mouse mammary cancer cells in vitro and in vivo,"

Oncogene, vol. 23, pp. 1125–1135, feb 2004.

[62] A. Dehghan, Q. Yang, A. Peters, S. Basu, J. C. Bis, A. R. Rudnicka, M. Kavousi, M.-H. Chen, J. Baumert, G. D. O. Lowe, and Others, "Association of novel genetic loci with circulating fibrinogen levels," *Circulation: Cardiovascular Genetics*, vol. 2, no. 2, pp. 125–133, 2009.

[63] D. Davalos and K. Akassoglou, "Fibrinogen as a key regulator of inflammation in disease," *Seminars in Immunopathology*, vol. 34, pp. 43–62, jan 2012.

[64] C. J. Mann, E. Perdiguero, Y. Kharraz, S. Aguilar, P. Pessina, A. L. Serrano, and P. Muñoz-Cánoves, "Aberrant repair and fibrosis development in skeletal muscle," *Skeletal Muscle*, vol. 1, no. 1, p. 21, 2011.

[65] M. Suelves, B. Vidal, A. L. Serrano, M. Tjwa, J. Roma, R. López-Alemany, A. Luttun, M. M. De Lagrán, M. À. Díaz, M. Jardí, M. Roig, M. Dierssen, M. Dewerchin, P. Carmeliet, and P. Muñoz-Cánoves, "uPA deficiency exacerbates muscular dystrophy in MDX mice," *Journal of Cell Biology*, vol. 178, pp. 1039–1051, sep 2007.

[66] M. Suelves, R. López-Alemany, F. Lluís, G. Aniorte, E. Serrano, M. Parra, P. Carmeliet, and P. Muñoz-Cánoves, "Plasmin activity is required for myogenesis in vitro and skeletal muscle regeneration in vivo," *Blood*, vol. 99, pp. 2835–2844, apr 2002.

[67] A. Liberzon, C. Birger, H. Thorvaldsd??ttir, M. Ghandi, J. P. Mesirov, and P. Tamayo, "The Molecular Signatures Database Hallmark Gene Set Collection," *Cell Systems*, vol. 1, pp. 417–425, dec 2015.

[68] O. Stegle, C. Lippert, J. M. Mooij, N. D. Lawrence, and K. Borgwardt, "Efficient inference in matrix-variate gaussian models with\iid observation noise," in *Advances in neural information processing systems*, pp. 630–638, 2011.

[69] C. Gao, I. C. McDowell, S. Zhao, C. D. Brown, and B. E. Engelhardt, "Context

specific and differential gene co-expression networks via bayesian biclustering," *PLoS computational biology*, vol. 12, no. 7, 2016.

[70] B. Zhang and S. Horvath, "A general framework for weighted gene co-expression network analysis," *Statistical applications in genetics and molecular biology*, vol. 4, no. 1, 2005.

[71] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.

[72] S. Freytag, J. Gagnon-Bartsch, T. P. Speed, and M. Bahlo, "Systematic noise degrades gene co-expression signals but can be corrected," *BMC bioinformatics*, vol. 16, no. 1, p. 309, 2015.

[73] J. M. Akey, S. Biswas, J. T. Leek, and J. D. Storey, "On the design and analysis of gene expression studies in human populations," *Nature genetics*, vol. 39, no. 7, pp. 807–808, 2007.

[74] V. Van Noort, B. Snel, and M. A. Huynen, "The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model," *EMBO reports*, vol. 5, no. 3, pp. 280–284, 2004.

[75] Princy Parsana, C. Ruberman, A. E. Jaffe, M. C. Schatz, A. Battle, and J. T. Leek, "Addressing confounding artifacts in reconstruction of gene co-expression networks," 2019.

[76] A. Buja and N. Eyuboglu, "Remarks on parallel analysis," *Multivariate behavioral research*, vol. 27, no. 4, pp. 509–540, 1992.

[77] G. Consortium, L. analysts:, D. A. . C. C. L. Laboratory, N. program management:, B. collection:, Pathology:, eQTL manuscript working group:, A. Battle, C. D. Brown, B. E. Engelhardt, and S. B. Montgomery, "Genetic effects on gene expression across human tissues," *Nature*, vol. 550, pp. 204–213, 10 2017.

[78] P. Langfelder and S. Horvath, "Wgcna: an r package for weighted correlation

network analysis," *BMC bioinformatics*, vol. 9, no. 1, p. 1, 2008.

[79] E. Y. Chen, C. M. Tan, Y. Kou, Q. Duan, Z. Wang, G. V. Meirelles, N. R. Clark, and A. Ma'ayan, "Enrichr: interactive and collaborative html5 gene list enrichment analysis tool," *BMC bioinformatics*, vol. 14, no. 1, p. 128, 2013.

[80] M. V. Kuleshov, M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann, *et al.*, "Enrichr: a comprehensive gene set enrichment analysis web server 2016 update," *Nucleic acids research*, vol. 44, no. W1, pp. W90–W97, 2016.

[81] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, and J. P. Mesirov, "Molecular signatures database (msigdb) 3.0," *Bioinformatics*, vol. 27, pp. 1739–1740, 06 2011.

[82] A.-L. Barabási, R. Albert, and H. Jeong, "Scale-free characteristics of random networks: the topology of the world-wide web," *Physica A: Statistical Mechanics and its Applications*, vol. 281, no. 1, pp. 69–77, 2000.

[83] M. R. Carlson, B. Zhang, Z. Fang, P. S. Mischel, S. Horvath, and S. F. Nelson, "Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks," *BMC genomics*, vol. 7, no. 1, p. 1, 2006.

[84] J. T. Leek and J. D. Storey, "Capturing heterogeneity in gene expression studies by surrogate variable analysis," *PLoS Genet*, vol. 3, no. 9, pp. 1724–1735, 2007.

[85] J. K. Pickrell, J. C. Marioni, A. A. Pai, J. F. Degner, B. E. Engelhardt, E. Nkadori, J.-B. Veyrieras, M. Stephens, Y. Gilad, and J. K. Pritchard, "Understanding mechanisms underlying human gene expression variation with RNA sequencing," *Nature*, vol. 464, pp. 768–772, apr 2010.

[86] T. i. m. o. t. h. é. e. Flutre, X. i. a. o. q. u. a. n. Wen, J. o. n. a. t. h. a. n. Pritchard, and M. a. t. t. h. e. w. Stephens, "A statistical framework for joint eqtl analysis in multiple tissues," *PLoS Genetics*, vol. 9, 2013.

[87] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies," *Nature Genetics*, vol. 38, pp. 904–909, aug 2006.

[88] C.-J. Hsieh, M. A. Sustik, I. S. Dhillon, and P. Ravikumar, "Quic: quadratic approximation for sparse inverse covariance estimation.," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2911–2947, 2014.

[89] V. Copois, F. Bibeau, C. Bascoul-Mollevi, N. Salvetat, P. Chalbos, and C. Bareil, "Impact of rna degradation on gene expression profiles: assessment of different methods to reliably determine rna quality," *J Biotechnol.*, vol. 127, 2007.

[90] M. I. Love, J. B. Hogenesch, and R. A. Irizarry, "Modeling of rna-seq fragment sequence bias reduces systematic errors in transcript abundance estimation," *Nat Biotechnol*, vol. 34, 2016.

[91] A. E. Jaffe, R. Tao, A. L. Norris, M. Kealhofer, A. Nellore, and J. H. Shin, "qsva framework for rna quality correction in differential expression analysis," *Proc Natl Acad Sci U S A*, vol. 114, 2017.

[92] K. D. Hansen, R. A. Irizarry, and Z. Wu, "Removing technical variability in rna-seq data using conditional quantile normalization," *Biostatistics*, vol. 13, no. 2, pp. 204–216, 2012.

[93] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford, "Salmon provides fast and bias-aware quantification of transcript expression," *Nat Methods*, vol. 14, 2017.

[94] I. Gallego Romero, A. A. Pai, J. Tung, and Y. Gilad, "Rna-seq: impact of rna degradation on transcript quantification," *BMC Biol*, vol. 12, 2014.

[95] S. Liebhaber, "mrna stability and the control of gene expression.," no. 36, pp. 29–32, 1996.

[96] P. Mehlen and A. Puisieux, “Metastasis: a question of life or death,” *Nat Rev Cancer*, vol. 6, 2006.

[97] J. H. Tsai and J. Yang, “Epithelial-mesenchymal plasticity in carcinoma metastasis,” *Genes Dev*, vol. 27, 2013.

[98] C. L. Chaffer and R. A. Weinberg, “A perspective on cancer cell metastasis,” *science*, vol. 331, no. 6024, pp. 1559–1564, 2011.

[99] J. P. Thiery, H. Acloque, R. Y. J. Huang, and M. A. Nieto, “Epithelial-mesenchymal transitions in development and disease,” *Cell*, vol. 139, 2009.

[100] R. Kalluri and R. A. Weinberg, “The basics of epithelial-mesenchymal transition,” *J Clin Invest*, vol. 119, 2009.

[101] B. De Craene and G. Berx, “Regulatory networks defining emt during cancer initiation and progression,” *Nature Reviews Cancer*, vol. 13, no. 2, pp. 97–110, 2013.

[102] A. Bergamaschi, Y. H. Kim, K. A. Kwei, Y. Choi, M. Bocanegra, and A. Langerød, “Camk1d amplification implicated in epithelial-mesenchymal transition in basal-like breast cancer,” *Mol Oncol*, vol. 2, 2008.

[103] Y.-L. Choi, M. Bocanegra, M. J. Kwon, Y. K. Shin, S. J. Nam, J.-H. Yang, J. Kao, A. K. Godwin, and J. R. Pollack, “Lyn is a mediator of epithelial-mesenchymal transition and a target of dasatinib in breast cancer,” *Cancer research*, vol. 70, no. 6, pp. 2296–2306, 2010.

[104] P. Papageorgis, A. W. Lambert, S. Ozturk, F. Gao, H. Pan, and U. Manne, “Smad signaling is required to maintain epigenetic silencing during breast cancer progression,” *Cancer Res*, vol. 70, 2010.

[105] A. Deshiere, E. Duchemin-Pelletier, E. Spreux, D. Ciais, F. Combes, and Y. Vandenbrouck, “Unbalanced expression of ck2 kinase subunits is sufficient to drive epithelial-to-mesenchymal transition by snail1 induction,” *Oncogene*,

vol. 32, 2013.

[106] J. Cai, H. Guan, L. Fang, Y. Yang, X. Zhu, and J. Yuan, "Microrna-374a activates wnt/β-catenin signaling to promote breast cancer metastasis," *J Clin Invest*, vol. 123, 2013.

[107] K. L. Andarawewa, A. C. Erickson, W. S. Chou, S. V. Costes, P. Gascard, and J. D. Mott, "Ionizing radiation predisposes nonmalignant human mammary epithelial cells to undergo transforming growth factor ??-induced epithelial to mesenchymal transition," *Cancer Res*, vol. 67, 2007.

[108] T. Joyce, D. Cantarella, C. Isella, E. Medico, and A. Pintzas, "A molecular signature for epithelial to mesenchymal transition in a human colon cancer cell system is revealed by large-scale microarray analysis," *Clin Exp Metastasis*, vol. 26, 2009.

[109] W.-L. Hwang, M.-H. Yang, M.-L. Tsai, H.-Y. Lan, S.-H. Su, S.-C. Chang, H.-W. Teng, S.-H. Yang, Y.-T. Lan, S.-H. Chiou, *et al.*, "Snail regulates interleukin-8 expression, stem cell–like activity, and tumorigenicity of human colorectal carcinoma cells," *Gastroenterology*, vol. 141, no. 1, pp. 279–291, 2011.

[110] S. Ohashi, M. Natsuizaka, S. Naganuma, S. Kagawa, S. Kimura, and H. Itoh, "A notch3-mediated squamous cell differentiation program limits expansion of emt-competent cells that express the zeb transcription factors," *Cancer Res*, vol. 71, 2011.

[111] F. Zijl, S. Mall, G. Machat, C. Pirker, R. Zeillinger, and A. Weinhaeusel, "A human model of epithelial to mesenchymal transition to monitor drug efficacy in hepatocellular carcinoma progression," *Mol Cancer Ther*, vol. 10, 2011.

[112] J. M. Drake, G. Strohbehn, T. B. Bair, J. G. Moreland, and M. D. Henry, "Zeb1 enhances transendothelial migration and represses the epithelial phenotype of prostate cancer cells," *Mol Biol Cell*, vol. 20, 2009.

[113] O. Leshem, S. Madar, I. Kogan-Sakin, I. Kamer, I. Goldstein, and R. Brosh, "Tmprss2/erg promotes epithelial to mesenchymal transition through the zeb1/zeb2 axis in a prostate cancer model," *PLoS One*, vol. 6, 2011.

[114] D. Kong, S. Banerjee, A. Ahmad, Y. Li, Z. Wang, and S. Sethi, "Epithelial to mesenchymal transition is mechanistically linked with stem cell signatures in prostate cancer cells," *PLoS One*, vol. 5, 2010.

[115] H. Roca, J. Hernandez, S. Weidner, R. C. McEachin, D. Fuller, and S. Sud, "Transcription factors ovol1 and ovol2 induce the mesenchymal to epithelial transition in human cancer," *PLoS One*, vol. 8, 2013.

[116] E. Takahashi, O. Nagano, T. Ishimoto, T. Yae, Y. Suzuki, and T. Shinoda, "Tumor necrosis factor-α regulates transforming growth factor-β-dependent epithelial-mesenchymal transition by promoting hyaluronan-cd44-moesin interaction," *J Biol Chem*, vol. 285, 2010.

[117] R. Edgar, M. Domrachev, and A. E. Lash, "Gene expression omnibus: Ncbi gene expression and hybridization array data repository," *Nucleic Acids Res*, vol. 30, 2002.

[118] J. Rung and A. Brazma, "Reuse of public genome-wide gene expression data," *Nat Rev Genet*, vol. 14, 2012.

[119] M. Pierre, B. DeHertogh, A. Gaigneaux, B. DeMeulder, F. Berger, and E. Bareke, "Meta-analysis of archived dna microarrays identifies genes regulated by hypoxia and involved in a metastatic phenotype in cancer cells," *BMC Cancer*, vol. 10, 2010.

[120] H. M. J. Sontrop, W. F. J. Verhaegh, M. J. T. Reinders, and P. D. Moerland, "An evaluation protocol for subtype-specific breast cancer event prediction," *PLoS One*, vol. 6, 2011.

[121] M. Chen, K. Wang, L. Zhang, C. Li, and Y. Yang, "The discovery of putative

urine markers for the specific detection of prostate tumor by integrative mining of public genomic profiles," *PLoS One*, vol. 6, 2011.

[122] C. J. Gröger, M. Grubinger, T. Waldhör, K. Vierlinger, and W. Mikulits, "Meta-analysis of gene expression signatures defining the epithelial to mesenchymal transition during cancer progression," *PLoS One*, vol. 7, 2012.

[123] K. S. Button, J. P. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. Robinson, and M. R. Munafò, "Power failure: why small sample size undermines the reliability of neuroscience," *Nature Reviews Neuroscience*, vol. 14, no. 5, pp. 365–376, 2013.

[124] S. Zöllner and J. K. Pritchard, "Overcoming the winner's curse: estimating penetrance parameters from case-control data," *Am J Hum Genet*, vol. 80, 2007.

[125] J. K. Choi, U. Yu, S. Kim, and O. J. Yoo, "Combining multiple microarray studies and modeling interstudy variation," *Bioinformatics*, vol. 19, 2003.

[126] A. C. Eklund and Z. Szallasi, "Correction of technical bias in clinical microarray data improves concordance with known biological information," *Genome Biol*, vol. 9, 2008.

[127] B. M. Bolstad, R. A. Irizarry, M. Åstrand, and T. P. Speed, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics*, vol. 19, 2003.

[128] A. A. Shabalin, H. Tjelmeland, C. Fan, C. M. Perou, and A. B. Nobel, "Merging two gene-expression studies via cross-platform normalization," *Bioinformatics*, vol. 24, 2008.

[129] F. B. Baker and L. J. Hubert, "Measuring the power of hierarchical cluster analysis," *J Am Stat Assoc*, vol. 70, 1975.

[130] D. R. Rhodes, J. Yu, K. Shanker, N. Deshpande, R. Varambally, and D. Ghosh, "Oncomine: a cancer microarray database and integrated data-mining platform,"

Neoplasia, vol. 6, 2004.

[131] J. J. Smith, N. G. Deane, F. Wu, N. B. Merchant, B. Zhang, and A. Jiang, "Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer," *Gastroenterology*, vol. 138, 2010.

[132] M. Schmidt, D. Böhm, C. Törne, E. Steiner, A. Puhl, and H. Pilch, "The humoral immune system has a key prognostic impact in node-negative breast cancer," *Cancer Res*, vol. 68, 2008.

[133] C. Hatzis, L. Pusztai, V. Valero, D. J. Booser, L. Esserman, A. Lluch, T. Vidaurre, F. Holmes, E. Souchon, H. Wang, *et al.*, "A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer," *Jama*, vol. 305, no. 18, pp. 1873–1881, 2011.

[134] S. Glück, J. S. Ross, M. Royce, E. F. McKenna, C. M. Perou, and E. Avisar, "Tp53 genomics predict higher clinical and pathologic tumor response in operable early-stage breast cancer treated with docetaxel-capecitabine ??," *Trastuzumab Breast Cancer Res Treat*, vol. 132, 2012.

[135] C. Desmedt, F. Piette, S. Loi, Y. Wang, F. Lallemand, and B. Haibe-Kains, "Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the transbig multicenter independent validation series," *Clin Cancer Res*, vol. 13, 2007.

[136] K. Shedden, J. Taylor, S. Enkemann, M. Tsao, T. Yeatman, W. Gerald, S. Eschrich, I. Jurisica, T. Giordano, D. Misek, *et al.*, "Director's challenge consortium for the molecular classification of lung adenocarcinoma," *Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. Nat Med*, vol. 14, no. 8, pp. 822–827, 2008.

[137] H. Okayama, T. Kohno, Y. Ishii, Y. Shimada, K. Shiraishi, and R. Iwakawa, "Identification of genes upregulated in alk-positive and egfr/kras/alk-negative

lung adenocarcinomas," *Cancer Res*, vol. 72, 2012.

[138] B. S. Taylor, N. Schultz, H. Hieronymus, A. Gopalan, Y. Xiao, and B. S. Carver, "Integrative genomic profiling of human prostate cancer," *Cancer Cell*, vol. 18, 2010.

[139] D. Sean and P. S. Meltzer, "Geoquery: a bridge between the gene expression omnibus (geo) and bioconductor," *Bioinformatics*, vol. 23, 2007.

[140] G. R. Grimmett, "On the number of clusters in the percolation model," *Journal of the London Mathematical Society*, vol. 2, no. 2, pp. 346–350, 1976.

[141] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences*, vol. 102, pp. 15545–15550, oct 2005.

[142] S. Yokoyama and H. Asahara, "The myogenic transcriptional network," *Cell Mol Life Sci*, vol. 68, 2011.

[143] D. K. Reaves, K. D. Fagan-Solis, K. Dunphy, S. D. Oliver, D. W. Scott, and J. M. Fleming, "The role of lipolysis stimulated lipoprotein receptor in breast cancer and directing breast cancer cell behavior," *PloS one*, vol. 9, no. 3, 2014.

[144] X. Wang, J. Yang, J. Qian, Z. Liu, H. Chen, and Z. Cui, "S100a14, a mediator of epithelial-mesenchymal transition, regulates proliferation, migration and invasion of human cervical cancer cells," *Am J Cancer Res*, vol. 5, 2015.

[145] X. Xu, B. Su, C. Xie, S. Wei, Y. Zhou, and H. Liu, "Sonic hedgehog-gli1 signaling pathway regulates the epithelial mesenchymal transition (emt) by mediating a new target gene, s100a4, in pancreatic cancer cells," *PLoS One*, vol. 9, 2014.

[146] T. Kawahara, N. Hotta, Y. Ozawa, S. Kato, K. Kano, and Y. Yokoyama, "Quantitative proteomic profiling identifies dpysl3 as pancreatic ductal adenocarcinoma-

associated molecule that regulates cell adhesion and migration by stabilization of focal adhesion complex," *PLoS One*, vol. 8, 2013.

[147] M. Kanda, S. Nomoto, H. Oya, D. Shimizu, H. Takami, and S. Hibino, "Dihydropyrimidinase-like 3 facilitates malignant behavior of gastric cancer," *J Exp Clin cancer Res CR*, vol. 33, 2014.

[148] Y. Li, Y. Zeng, S. M. Mooney, B. Yin, A. Mizokami, and M. Namiki, "Resistance to paclitaxel increases the sensitivity to other microenvironmental stresses in prostate cancer cells," *J Cell Biochem*, vol. 112, 2011.

[149] K. Steketee, "Ziel-van der made acj, van der korput hagm, houstmuller ab," *Trapman J A bioinformatics-based functional analysis shows that the specifically androgen-regulated gene SARG contains an active direct repeat androgen response element in the first intron J Mol Endocrinol*, vol. 33, 2004.

[150] J. Milara, T. Peiro, A. Serrano, and J. Cortijo, "Epithelial to mesenchymal transition is increased in patients with copd and induced by cigarette smoke," *Thorax BMJ Publishing Group Ltd*, vol. 68, 2013.

[151] N. S. Nagathihalli, P. P. Massion, A. L. Gonzalez, P. Lu, and P. K. Datta, "Smoking induces epithelial-to-mesenchymal transition in non-small cell lung cancer through hdac-mediated downregulation of e-cadherin," *Mol Cancer Ther*, vol. 11, 2012.

[152] K. R. Fischer, A. Durrans, S. Lee, J. Sheng, F. Li, and S. T. C. Wong, "Epithelial-to-mesenchymal transition is not required for lung metastasis but contributes to chemoresistance," *Nature Nature Research*, vol. 527, 2015.

[153] X. Zheng, J. L. Carstens, J. Kim, M. Scheible, J. Kaye, and H. Sugimoto, "Epithelial-to-mesenchymal transition is dispensable for metastasis but induces chemoresistance in pancreatic cancer," *Nature Nature Publishing Group*, vol. 527, 2015.

[154] E. McKiernan, E. W. McDermott, D. Evoy, J. Crown, and M. J. Duffy, "The role of s100 genes in breast cancer progression," *Tumor Biol*, vol. 32, 2011.

[155] P. Mak, I. Leav, B. Pursell, D. Bae, X. Yang, and C. A. Taglienti, "Erbeta impedes prostate cancer emt by destabilizing hif-1alpha and inhibiting vegf-mediated snail nuclear localization: implications for gleason grading," *Cancer Cell NIH Public Access*, vol. 17, 2010.

[156] L. I. Furlong, "Human diseases through the lens of network biology," *Trends Genet*, vol. 29, 2013.

[157] Z. Guan, G. Parmigiani, and P. Patil, "Merging versus ensembling in multi-study machine learning: Theoretical insight from random effects," *arXiv preprint arXiv:1905.07382*, 2019.

[158] P. Patil and G. Parmigiani, "Training replicable predictors in multiple studies," *Proceedings of the National Academy of Sciences*, vol. 115, no. 11, pp. 2578–2583, 2018.

[159] S. S. Kim, C. Dai, F. Hormozdiari, B. van de Geijn, S. Gazal, Y. Park, L. O'Connor, T. Amariuta, P.-R. Loh, H. Finucane, *et al.*, "Genes with high network connectivity are enriched for disease heritability," *The American Journal of Human Genetics*, vol. 104, no. 5, pp. 896–913, 2019.

[160] H. K. Finucane, B. Bulik-Sullivan, A. Gusev, G. Trynka, Y. Reshef, P.-R. Loh, V. Anttila, H. Xu, C. Zang, K. Farh, *et al.*, "Partitioning heritability by functional annotation using genome-wide association summary statistics," *Nature genetics*, vol. 47, no. 11, p. 1228, 2015.

[161] S. Gazal, H. K. Finucane, N. A. Furlotte, P.-R. Loh, P. F. Palamara, X. Liu, A. Schoech, B. Bulik-Sullivan, B. M. Neale, A. Gusev, *et al.*, "Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection," *Nature genetics*, vol. 49, no. 10, p. 1421, 2017.

[162] J. Zhu, “Sradb,” 2017.

[163] V. Lagani, A. D. Karozou, D. Gomez-Cabrero, G. Silberberg, and I. Tsamardinos, “A comparative evaluation of data-merging and meta-analysis methods for reconstructing gene-gene interactions,” *BMC bioinformatics*, vol. 17, no. 5, p. S194, 2016.

[164] B. Zhang and S. Horvath, “A general framework for weighted gene co-expression network analysis,” *Stat Appl Genet Mol Biol*, vol. 4, 2005.

[165] D. M. Bravata and I. Olkin, “Simple pooling versus combining in meta-analysis,” *Evaluation & the health professions*, vol. 24, no. 2, pp. 218–230, 2001.

[166] C. A. Cassa, D. Weghorn, D. J. Balick, D. M. Jordan, D. Nusinow, K. E. Samocha, A. O’Donnell-Luria, D. G. MacArthur, M. J. Daly, D. R. Beier, *et al.*, “Estimating the selective effects of heterozygous protein-truncating variants from human exome data,” *Nature genetics*, vol. 49, no. 5, pp. 806–810, 2017.

[167] G. Csardi and T. Nepusz, “The igraph software package for complex network research,” *InterJournal*, vol. Complex Systems, p. 1695, 2006.

[168] P. Danaher, P. Wang, and D. M. Witten, “The joint graphical lasso for inverse covariance estimation across multiple classes,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 76, no. 2, pp. 373–397, 2014.

[169] J. T. Leek, “Asymptotic conditional singular value decomposition for high-dimensional genomic data,” *Biometrics*, vol. 67, no. 2, pp. 344–352, 2011.

[170] S. K. Kim, J. Lund, M. Kiraly, K. Duke, M. Jiang, J. M. Stuart, A. Eizinger, B. N. Wylie, and G. S. Davidson, “A gene expression map for caenorhabditis elegans,” *Science*, vol. 293, no. 5537, pp. 2087–2092, 2001.

[171] A. Rzhetsky and S. M. Gomez, “Birth of scale-free molecular networks and the number of distinct dna and protein domains per genome,” *Bioinformatics*, vol. 17, no. 10, pp. 988–996, 2001.

[172] A. Bhan, D. J. Galas, and T. G. Dewey, "A duplication growth model of gene expression networks," *Bioinformatics*, vol. 18, no. 11, pp. 1486–1493, 2002.

[173] I. K. Jordan, L. Mariño-Ramírez, Y. I. Wolf, and E. V. Koonin, "Conservation and coevolution in the scale-free human gene coexpression network," *Molecular biology and evolution*, vol. 21, no. 11, pp. 2058–2070, 2004.

[174] M. E. Newman, "The structure and function of complex networks," *SIAM review*, vol. 45, no. 2, pp. 167–256, 2003.

[175] B. Zhang, C. Gaiteri, L.-G. Bodea, Z. Wang, J. McElwee, A. A. Podtelezhnikov, C. Zhang, T. Xie, L. Tran, R. Dobrin, *et al.*, "Integrated systems approach identifies genetic nodes and networks in late-onset alzheimer's disease," *Cell*, vol. 153, no. 3, pp. 707–720, 2013.

[176] L. J. Kogelman, S. Cirera, D. V. Zhernakova, M. Fredholm, L. Franke, and H. N. Kadarmideen, "Identification of co-expression gene networks, regulatory genes and pathways for obesity based on adipose tissue rna sequencing in a porcine model," *BMC medical genomics*, vol. 7, no. 1, p. 57, 2014.

[177] J. Xue, S. V. Schmidt, J. Sander, A. Draffehn, W. Krebs, I. Quester, D. De Nardo, T. D. Gohel, M. Emde, L. Schmidleithner, *et al.*, "Transcriptome-based network analysis reveals a spectrum model of human macrophage activation," *Immunity*, vol. 40, no. 2, pp. 274–288, 2014.

[178] J. A. Miller, S.-L. Ding, S. M. Sunkin, K. A. Smith, L. Ng, A. Szafer, A. Ebbert, Z. L. Riley, J. J. Royall, K. Aiona, *et al.*, "Transcriptional landscape of the prenatal human brain," *Nature*, vol. 508, no. 7495, p. 199, 2014.

[179] M. Hawrylycz, J. A. Miller, V. Menon, D. Feng, T. Dolbeare, A. L. Guillozet-Bongaarts, A. G. Jegga, B. J. Aronow, C.-K. Lee, A. Bernard, *et al.*, "Canonical genetic signatures of the adult human brain," *Nature neuroscience*, vol. 18, no. 12, p. 1832, 2015.

[180] M. S. Breen, A. X. Maihofer, S. J. Glatt, D. S. Tylee, S. D. Chandler, M. T. Tsuang, V. B. Risbrough, D. G. Baker, D. T. O'Connor, C. M. Nievergelt, *et al.*, "Gene networks specific for innate immunity define post-traumatic stress disorder," *Molecular psychiatry*, vol. 20, no. 12, p. 1538, 2015.

[181] P. Bailey, D. K. Chang, K. Nones, A. L. Johns, A.-M. Patch, M.-C. Gingras, D. K. Miller, A. N. Christ, T. J. Bruxner, M. C. Quinn, *et al.*, "Genomic analyses identify molecular subtypes of pancreatic cancer," *Nature*, vol. 531, no. 7592, p. 47, 2016.

[182] C. Gao, I. C. McDowell, S. Zhao, C. D. Brown, and B. E. Engelhardt, "Context specific and differential gene co-expression networks via bayesian biclustering," *PLOS Computational Biology*, vol. 12, pp. 1–39, 07 2016.

[183] M. Fromer, P. Roussos, S. K. Sieberts, J. S. Johnson, D. H. Kavanagh, T. M. Perumal, D. M. Ruderfer, E. C. Oh, A. Topol, H. R. Shah, *et al.*, "Gene expression elucidates functional impact of polygenic risk for schizophrenia," *Nature neuroscience*, vol. 19, no. 11, p. 1442, 2016.

[184] K. A. Hoadley, C. Yau, T. Hinoue, D. M. Wolf, A. J. Lazar, E. Drill, R. Shen, A. M. Taylor, A. D. Cherniack, V. Thorsson, *et al.*, "Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer," *Cell*, vol. 173, no. 2, pp. 291–304, 2018.

[185] M. V. Lombardo, H. M. Moon, J. Su, T. D. Palmer, E. Courchesne, and T. Pramparo, "Maternal immune activation dysregulation of the fetal brain transcriptome and relevance to the pathophysiology of autism spectrum disorder," *Molecular psychiatry*, vol. 23, no. 4, p. 1001, 2018.

Appendix A

Addressing confounding artifacts in reconstruction of gene coexpression networks

A.1 Conditions for and Proof of Convergence of Principal Components

Lemma: *Let* $\mathbf{X}$ *be a high-dimensional matrix of expression data with signal both due to artifacts* $\mathbf{A}$*, and due to a genuine network of linear expression relationships. Then under the conditions below and provided that the node degree distribution of the network follows a power-law, the principal components of* X *consistently estimate a linear space spanning the artifacts* A *and not the network structure.*

Proof:

Decompose a gene expression matrix with n samples and m genes $\mathbf{X}_{m\times n} = (\mathbf{x}_1, ..., \mathbf{x}_m)^T$ as follows:

$$\mathbf{X} = \boldsymbol{\mu} \times \mathbf{1} + \boldsymbol{\Gamma}_{\mathbf{A}}\mathbf{A} + \boldsymbol{\Gamma}_{\mathbf{N}}\mathbf{N} + \mathbf{U}$$

where,

- $\boldsymbol{\mu} = (\mu_1, ..., \mu_m)^T$ is an m dimensional column vector with $\mu_i := E\left[\mathbf{x}_i\right]$, $i =$

$1, ..., m$ and $\mathbf{1}$ is an n dimensional row vector of 1's.

- There are L artifacts or confounders ($L < n$), forming an $L \times n$ matrix A with an associated coefficient matrix Γ_A.
- N is an $m \times n$ matrix of expression data without any network structure, with associated $m \times m$ coefficient vector Γ_N. Features i and k are share an edge if γ_{ik}^N or γ_{ki}^N are nonzero. This represents a linear relationship between the expression levels of genes. To avoid circularity, the diagonal entries of Γ_N are set to zero.
- $\mathbf{U}$ is an $m \times n$ matrix of pairwise independent mean zero random noise

Based on our previous work [169], given a high-dimensional matrix with the number of features much larger than the number of samples ($m >> n$) we make the following additional assumptions about the behavior of the data in the experiment.

1. The number of non-zero entries in the network Γ_N follows a power-law distribution with an exponential coefficient $2 < \alpha < 3$ [82]. As we point out in the main text power-law degree distributions have been observed in gene expression networks, for example yeast co-expression networks [74, 83] and *Caenorhabditis elegans* [170], and the preferential attachment model characteristic of scale-free networks has been explained by gene duplication [171–173]. Further, network inference algorithms such as WGCNA also employ this assumption.
2. The entries in the artifact and network coefficient, pre-network expression data, and independent noise matrices have bounded fourth moment:

$$
\begin{aligned}
0 < E\left[\left(\gamma_{A_{i,j}}\right)^4\right] &\leq B_{\gamma_A} \\
0 < E\left[\left(\gamma_{N_{i,j}}\right)^4\right] &\leq B_{\gamma_N} \\
0 < E\left[\left(N_{i,j}\right)^4\right] &\leq B_N \\
0 < E\left[\left(u_{i,j}\right)^4\right] &\leq B_U.
\end{aligned}
$$

Therefore, by Lyapunov's inequality, there exist (finite) bounds B'_{γ_A}, B'_{γ_N}, B'_N, and B'_U, on the variances:

$$\begin{aligned}
0 < \text{Var}\left(\gamma_{A_{i,j}}\right) &= E\left[(\gamma_{A_{i,j}})^2\right] \leq B'_{\gamma_A} \\
0 < \text{Var}\left(\gamma_{N_{i,j}}\right) &= E\left[(\gamma_{N_{i,j}})^2\right] \leq B'_{\gamma_N} \\
0 < \text{Var}\left(N_{i,j}\right) &= E\left[(N_{i,j})^2\right] \leq B'_N \\
0 < \text{Var}\left(u_{i,j}\right) &= E\left[(u_{i,j})^2\right] \leq B'_U.
\end{aligned}$$

This is true for most common distributions used to model gene expression data or a suitably transformed version.

3. There exists a positive definite matrix Δ for which the following hold:

   (a) $\lim\limits_{m\to\infty} \|\frac{1}{m}A^T\Gamma_A^T\Gamma_A A - A^T\Delta A\|_F = 0$

   (b) $A^T\Delta A$ has eigenvalues $\lambda_1 > .... > \lambda_L > \lambda_{L+1} = ... = \lambda_n = 0$

   This assumption means that the batch effects and other artifacts are sufficiently widespread as to affect a fixed and non-negligable percentage of the genes in the data set.

Additionally, we assume without loss of generality, that expression levels of each gene in $\mathbf{X}$ is centered.

4. $\boldsymbol{\mu} = \vec{0}$.

5. The expression data in the absence of any network structure, N, has mean $E[N] = \vec{0}$ where $\vec{0}$ is an m-dimensional column vector. Further, in the absence of network structure, the genes are pairwise independent. Therefore, by Assumption 2 the entries of N converge almost surely to zero.

Based on this model, we show that the principal components of the matrix $\mathbf{X}$ (with a fixed n - sample size) estimate the artifacts and are not corrupted by the signal from the network terms.

The eigen-vectors of the matrix $\frac{1}{m}\mathbf{X}^T\mathbf{X}$ are equal to the right singular vectors of the matrix $\mathbf{X}$. Given observed data $\mathbf{X}$, the empirical variance-covariance matrix of the data $\hat{\Sigma}$ takes the form:

$$
\begin{aligned}
\hat{\Sigma} &= \frac{1}{m}\mathbf{X}^T\mathbf{X} \\
&= \frac{1}{m}\left(\Gamma_A A + \Gamma_N N + \mathbf{U}\right)^T\left(\Gamma_A A + \Gamma_N N + \mathbf{U}\right) \\
&= \frac{1}{m}\left(A^T\Gamma_A^T + N^T\Gamma_N^T + \mathbf{U}^T\right)\left(\Gamma_A A + \Gamma_N N + \mathbf{U}\right) \\
&= \frac{1}{m}(A^T\Gamma_A^T\Gamma_A A + A^T\Gamma_A^T\Gamma_N N + A^T\Gamma_A^T\mathbf{U} + N^T\Gamma_N^T\Gamma_A A + N^T\Gamma_N^T\Gamma_N N + N^T\Gamma_N^T\mathbf{U}+ \\
&\qquad \mathbf{U}^T\Gamma_A A + \mathbf{U}^T\Gamma_N N + \mathbf{U}^T\mathbf{U}) \\
&= \frac{1}{m}\left(A^T\Gamma_A^T\Gamma_A A + A^T\Gamma_A^T\Gamma_N N + N^T\Gamma_N^T\Gamma_A A + N^T\Gamma_N^T\Gamma_N N\right) + \\
&\quad \frac{1}{m}\left(A^T\Gamma_A^T\mathbf{U} + N^T\Gamma_N^T\mathbf{U} + \mathbf{U}^T\Gamma_A A + \mathbf{U}^T\Gamma_N N + \mathbf{U}^T\mathbf{U}\right) \\
&= \frac{1}{m}A^T\Gamma_A^T\Gamma_A A + \frac{1}{m}A^T\Gamma_A^T\Gamma_N N + \frac{1}{m}N^T\Gamma_N^T\Gamma_A A + \frac{1}{m}N^T\Gamma_N^T\Gamma_N N + \\
&\quad \frac{1}{m}A^T\Gamma_A^T\mathbf{U} + \frac{1}{m}N^T\Gamma_N^T\mathbf{U} + \frac{1}{m}\mathbf{U}^T\Gamma_A A + \frac{1}{m}\mathbf{U}^T\Gamma_N N + \frac{1}{m}\mathbf{U}^T\mathbf{U}
\end{aligned}
$$

We will show that as the number of features (i.e. genes) grows, the empirical variance-covariance matrix, after centering by an estimate of the background variation, converges to the same thing as if there were no network structure:

$$\tilde{\mathbf{X}}_{\mathbf{unstr}} := \mathbf{\Gamma_A A} + \mathbf{U}.$$

Then we can show that the principal components of the confounded matrix are consistent estimators of the confounding variables.

Therefore, we will show that, holding the number of observations n fixed, there exits an $n \times n$ matrix $\mathcal{L}$ so that:

$$
\begin{aligned}
\lim_{m\to\infty}\frac{1}{m}\left(\tilde{\mathbf{X}}^{\text{unstr}}\right)^T\tilde{\mathbf{X}}^{\text{unstr}} - \hat{\sigma}^{\mathbf{2}}_{\mathbf{ave}}\mathbf{I} = \mathcal{L} \\
\lim_{m\to\infty}\frac{1}{m}\mathbf{X}^T\mathbf{X} - \hat{\sigma}^2_{ave}\mathbf{I} = \mathcal{L}
\end{aligned}
$$

where, borrowing the notation from Leek 2011, we let $V_L(\mathbf{X}) = \{v_1(\mathbf{X}), ..., v_L(\mathbf{X})\}$ be a matrix of the first L right singular vectors of $\mathbf{X}$ and $\hat{\Gamma}_L$ the least squares estimates from regressing $\mathbf{X}$ on $V_L(\mathbf{X})$. Then, we define:

$$\sigma^2_{ave} := \frac{1}{m(n-L)} \|\mathbf{X} - \hat{\Gamma}_L V_L(\mathbf{X})\|_F,$$

where we estimate L using a permutation approach through the 'num.sv' function in the *sva* package.

To determine $\mathcal{L}$, we write:

$$\begin{aligned}
\frac{1}{m}\left(\tilde{\mathbf{X}}^{\text{unstr}}\right)^T \tilde{\mathbf{X}}^{\text{unstr}} - \hat{\sigma}^{\mathbf{2}}_{\mathbf{ave}}\mathbf{I} &= \frac{1}{m}\left(\Gamma_A A + U\right)^T\left(\Gamma_A A + U\right) - \hat{\sigma}^2_{ave}\mathbf{I} \\
&= \frac{1}{m}\left(A^T\Gamma_A^T + U^T\right)\left(\Gamma_A A + U\right) - \hat{\sigma}^2_{ave}\mathbf{I} \\
&= \frac{1}{m}A^T\Gamma_A^T\Gamma_A A + \frac{1}{m}A^T\Gamma_A^T U + \frac{1}{m}U^T\Gamma_A A + \frac{1}{m}^T U - \hat{\sigma}^2_{ave}\mathbf{I}
\end{aligned}$$

Letting m (number of genes) grow,

$$\begin{aligned}
&\lim_{m\to\infty}\frac{1}{m}\left(\tilde{\mathbf{X}}^{\text{unstr}}\right)^T \tilde{\mathbf{X}}^{\text{unstr}} - \hat{\sigma}^{\mathbf{2}}_{\mathbf{ave}}\mathbf{I} \\
&\quad = \lim_{m\to\infty}\frac{1}{m}\Gamma_A^T\Gamma_A A + \lim_{m\to\infty}\frac{1}{m}A^T\Gamma_A^T U + \lim_{m\to\infty}\frac{1}{m}U^T\Gamma_A A + \lim_{m\to\infty}\frac{1}{m}U^T U - \hat{\sigma}^2_{ave}\mathbf{I} \\
&\quad = A^T\Delta A + \lim_{m\to\infty}\frac{1}{m}A^T\Gamma_A^T U + \lim_{m\to\infty}\frac{1}{m}U^T\Gamma_A A + \lim_{m\to\infty}\frac{1}{m}U^T U - \hat{\sigma}^2_{ave}\mathbf{I}
\end{aligned}$$

Leek 2011 shows that the terms $\lim_{m\to\infty}\frac{1}{m}A^T\Gamma_A^T U + \lim_{m\to\infty}\frac{1}{m}U^T\Gamma_A A$ both converge almost surely to zero by the Kolmogorov Strong Law of Large Numbers (KSLLN). Further, Leek 2011 uses KSLLN to show that the off diagonal elements of $\frac{1}{m}U^T U$ converge almost surely to zero, while the diagonals converge almost surely to $\hat{\sigma}^2_{ave}$. Therefore,

$$\lim_{m\to\infty}\frac{1}{m}\left(\tilde{\mathbf{X}}^{\text{unstr}}\right)^T \tilde{\mathbf{X}}^{\text{unstr}} - \hat{\sigma}^{\mathbf{2}}_{\mathbf{ave}}\mathbf{I} = \mathbf{A^T\Delta A},$$

and

$$\mathcal{L} = A^T\Delta A.$$

The limit of the empirical variance-covariance matrix is as follows:

$$\begin{aligned}
&\lim_{m\to\infty}\frac{1}{m}\mathbf{X}^T\mathbf{X}-\hat{\sigma}^2_{ave}\mathbf{I}\\
&=\lim_{m\to\infty}\frac{1}{m}A^T\Gamma_A^T\Gamma_A A+\lim_{m\to\infty}\frac{1}{m}\Gamma_A A^T\Gamma_N N+\lim_{m\to\infty}\frac{1}{m}N^T\Gamma_N^T\Gamma_A A+\lim_{m\to\infty}\frac{1}{m}N^T\Gamma_N^T\Gamma_N N+-\hat{\sigma}^2_{ave}\mathbf{I}\\
&\quad\lim_{m\to\infty}\frac{1}{m}A^T\Gamma_A^T\mathbf{U}+\lim_{m\to\infty}\frac{1}{m}N^T\Gamma_N^T\mathbf{U}+\lim_{m\to\infty}\frac{1}{m}\mathbf{U}^T\Gamma_A A+\lim_{m\to\infty}\frac{1}{m}\mathbf{U}^T\Gamma_N N+\lim_{m\to\infty}\frac{1}{m}\mathbf{U}^T\mathbf{U}-\hat{\sigma}^2_{ave}\\
&=\lim_{m\to\infty}\frac{1}{m}A^T\Gamma_A^T\Gamma_A A+\lim_{m\to\infty}\frac{1}{m}\Gamma_A A^T U+\lim_{m\to\infty}\frac{1}{m}U^T\Gamma_A A+\lim_{m\to\infty}\frac{1}{m}U^T U-\hat{\sigma}^2_{ave}\mathbf{I}+\\
&\quad\lim_{m\to\infty}\frac{1}{m}A^T\Gamma_A^T\Gamma_N N+\lim_{m\to\infty}\frac{1}{m}N^T\Gamma_N^T\Gamma_A A+\lim_{m\to\infty}\frac{1}{m}N^T\Gamma_N^T\Gamma_N N+\lim_{m\to\infty}\frac{1}{m}\mathbf{U}^T\Gamma_N N+\lim_{m\to\infty}\frac{1}{m}N^T\\
&=A^T\Delta A+\lim_{m\to\infty}\frac{1}{m}U^T U-\hat{\sigma}^2_{ave}\mathbf{I}+\lim_{m\to\infty}\frac{1}{m}A^T\Gamma_A^T\Gamma_N N+\lim_{m\to\infty}\frac{1}{m}N^T\Gamma_N^T\Gamma_A A+\\
&\quad\lim_{m\to\infty}\frac{1}{m}N^T\Gamma_N^T\Gamma_N N+\lim_{m\to\infty}\frac{1}{m}\mathbf{U}^T\Gamma_N N+\lim_{m\to\infty}\frac{1}{m}N^T\Gamma_N^T\mathbf{U}\\
&=A^T\Delta A+\underbrace{\lim_{m\to\infty}\frac{1}{m}A^T\Gamma_A^T\Gamma_N N}_{(1)}+\underbrace{\lim_{m\to\infty}\frac{1}{m}N^T\Gamma_N^T\Gamma_A A}_{(2)}+\underbrace{\lim_{m\to\infty}\frac{1}{m}N^T\Gamma_N^T\Gamma_N N}_{(3)}+\\
&\quad\underbrace{\lim_{m\to\infty}\frac{1}{m}\mathbf{U}^T\Gamma_N N}_{(4)}+\underbrace{\lim_{m\to\infty}\frac{1}{m}N^T\Gamma_N^T\mathbf{U}}_{(5)}
\end{aligned}$$

We consider the convergence of (1) through (5) separately:

1.

$$\lim_{m\to\infty}\frac{1}{m}A^T\Gamma_A^T\Gamma_N N=\lim_{m\to\infty}A^T\frac{1}{m}\Gamma_A^T\Gamma_N N$$

We first consider $Q := \frac{1}{m}\Gamma_A^T\Gamma_N$, an $L \times m$ matrix with entries indexed by

$l \in \{1, ..., L\}, k \in \{1, ..., m\}$:

$$\begin{aligned} q_{lk} &= Q_{l,k} \\ &= \frac{1}{m} \sum_{j=1}^{m} \Gamma_{A_{j,l}} \Gamma_{N_{j,k}} \\ &= \frac{1}{m} \sum_{j=1}^{m} \gamma_{A_{j,l}} \gamma_{N_{j,k}} \\ &= \frac{1}{m} \sum_{\{j : \gamma_{N_{j,k}} \neq 0\}} \gamma_{A_{j,l}} \gamma_{N_{j,k}} + \frac{1}{m} \sum_{\{j : \gamma_{N_{j,k}} = 0\}} \gamma_{A_{j,l}} \gamma_{N_{j,k}} \\ &= \frac{1}{m} \sum_{\{j : \gamma_{N_{j,k}} \neq 0\}} \gamma_{A_{j,l}} \gamma_{N_{j,k}} + \frac{1}{m} \sum_{\{j : \gamma_{N_{j,k}} = 0\}} \gamma_{A_{j,l}} \times 0 \\ &= \frac{1}{m} \sum_{\{j : \gamma_{N_{j,k}} \neq 0\}} \gamma_{A_{j,l}} \gamma_{N_{j,k}} \end{aligned}$$

Suppose that there are $0 \leq d \leq m$ indices j for which $\gamma_{N\,j,k} \neq 0$, so that there are d terms $\gamma_{A_{j,l}} \gamma_{N_{j,k}}$ in the summation contributing to q_{lk}. We can re-index these terms as $\gamma_{A_{j',l}} \gamma_{N_{j',k}}, j' = 1, ..., d$.

For any fixed k, whenever $\gamma_{N_{j,k}}$, necessarily genes k and j share an edge. Therefore, given d non-zero coefficients $\gamma_{N_{j,k}}$, gene k has at least degree d. However, [174] show that for scale free networks following a power-law degree distribution $p_k \sim k^{\alpha-1}$, as assumed in our framework, the maximum degree of a vertex in the network follows $k_{\max} \sim m^{\frac{1}{\alpha-1}}$, and $d \leq m^{\frac{1}{\alpha-1}}$. Therefore, we can write each element as:

$$\begin{aligned}
q_{lk} &= \frac{1}{m} \sum_{\{j:\gamma_{N_{j,k}} \neq 0\}} \gamma_{A_{j,l}} \gamma_{N_{j,k}} \\
&= \frac{1}{m} \sum_{j'=1}^{d} \gamma_{A_{j',l}} \gamma_{N_{j',k}} \\
&= \frac{d}{m} \frac{1}{d} \sum_{j'=1}^{d} \gamma_{A_{j',l}} \gamma_{N_{j',k}} \\
&\leq \frac{m^{\frac{1}{\alpha-1}}}{m} \frac{1}{d} \sum_{j'=1}^{d} \gamma_{A_{j',l}} \gamma_{N_{j',k}} \\
&= m^{-1} m^{\frac{1}{\alpha-1}} \frac{1}{d} \sum_{j'=1}^{d} \gamma_{A_{j',l}} \gamma_{N_{j',k}} \\
&= m^{\frac{2-\alpha}{\alpha-1}} \frac{1}{d} \sum_{j'=1}^{d} \gamma_{A_{j',l}} \gamma_{N_{j',k}} \\
&= m^{\frac{-(\alpha-2)}{\alpha-1}} \frac{1}{d} \sum_{j'=1}^{d} \gamma_{A_{j',l}} \gamma_{N_{j',k}} \\
&= \frac{1}{m^{\frac{\alpha-1}{\alpha-2}}} \frac{1}{d} \sum_{j'=1}^{d} \gamma_{A_{j',l}} \gamma_{N_{j',k}}
\end{aligned}$$

By Assumption 1 ($2 < \alpha < 3$), so that $\frac{\alpha-1}{\alpha-2} > 1$ and $\lim_{m\to\infty} \frac{1}{m^{\frac{\alpha-1}{\alpha-2}}} = 0$.

Now, consider the expectation of the terms inside of the summation. For any j', applying the Cauchy-Schwarz inequality to $|\gamma_{A_{j',l}}|$ and $|\gamma_{N_{j',k}}|$

$$\begin{aligned}
E\left[|\gamma_{A_{j',l}} \gamma_{N_{j',k}}|\right] &\leq \sqrt{E\left[|\gamma_{A_{j',l}}|^2\right] E\left[|\gamma_{N_{j',k}}|^2\right]} \\
&= \sqrt{E\left[(\gamma_{A_{j',l}})^2\right] E\left[(\gamma_{N_{j',k}})^2\right]} \\
&\leq \sqrt{B'_{\gamma_A} \times B'_{\gamma_N}} \quad \text{By Assumption 2} \\
&= B^* \quad \text{where we define the bound } \; B^* := \sqrt{B'_{\gamma_A} \times B'_{\gamma_N}},
\end{aligned}$$

and

$$-\infty < -B^* \leq E\left[\gamma_{A_{j',l}} \gamma_{N_{j',k}}\right] \leq B^* < \infty,$$

and by the Strong Law of Large Numbers,

$$\frac{1}{d} \sum_{j'=1}^{d} \gamma_{A_{j',l}} \gamma_{N_{j',k}} \xrightarrow{a.s.} E\left[\gamma_{A_{j',l}} \gamma_{N_{j',k}}\right],$$

therefore, for each l, k:

$$q_{lk} \leq \frac{1}{m^{\frac{\alpha-1}{\alpha-2}}} \frac{1}{d} \sum_{j'=1}^{d} \gamma_{A_{j',l}} \gamma_{N_{j',k}} \xrightarrow{a.s.} 0$$

s

and

$$Q \xrightarrow{a.s.} 0.$$

Recall, the matrix of artifacts A is $L \times n$ dimensional, so that it is fixed with respect to m, and, as shown in Assumption 5, $N \xrightarrow{a.s.} 0$, so that by Slutsky's Theorem:

$$\lim_{m\to\infty} \frac{1}{m} A^T \Gamma_A^T \Gamma_N N = 0$$

2.

$$\lim_{m\to\infty} \frac{1}{m} N^T \Gamma_N^T \Gamma_A A$$

By symmetry, the same argument as in (1) holds, and

$$\lim_{m\to\infty} \frac{1}{m} N^T \Gamma_N^T \Gamma_A A = 0$$

3.

$$\lim_{m\to\infty} \frac{1}{m} N^T \Gamma_N^T \Gamma_N N = \lim_{m\to\infty} N^T \frac{1}{m} \Gamma_N^T \Gamma_N N$$

We will first consider $P := \frac{1}{m} \Gamma_N^T \Gamma_N$, an $m \times m$ matrix with entries indexed by $l, k \in \{1, ..., m\}$:

$$\begin{aligned} p_{lk} &= P_{l,k} \\ &= \frac{1}{m} \sum_{j=1}^{m} \Gamma_{N_{j,l}} \Gamma_{N_{j,k}} \\ &= \frac{1}{m} \sum_{j=1}^{m} \gamma_{N_{j,l}} \gamma_{N_{j,k}} \end{aligned}$$

We will consider the diagonal and off-diagonal entries of P separately. The diagonal entries ($k = l$) take the form:

$$\begin{aligned}
p_{ll} &= \frac{1}{m}\sum_{j=1}^{m} \gamma_{N_{j,l}}\gamma_{N\,j,l} \\
&= \frac{1}{m}\sum_{j=1}^{m} \gamma_{N_{j,l}}^2 \\
&= \frac{1}{m}\sum_{\{j:\gamma_{N_{j,l}}\neq 0\}} \gamma_{N_{j,l}}^2 + \frac{1}{m}\sum_{\{j:\gamma_{N_{j,l}}=0\}} \gamma_{N_{j,l}}^2 \\
&= \frac{1}{m}\sum_{\{j:\gamma_{N_{j,l}}\neq 0\}} \gamma_{N_{j,l}}^2
\end{aligned}$$

Now, whenever $\gamma_{N_{j,l}} \neq 0$, by definition, genes j and l share an edge, so that d', the number of j such that $\gamma_{N_{j,l}} \neq 0$ is equal to the degree of vertex l. Following the argument from the proof of (1) , $d' \leq m^{\frac{1}{\alpha-1}}, 2 < \alpha < 3$ and:

$$\begin{aligned}
p_{ll} &= \frac{1}{m}\sum_{\{j:\gamma_{N_{j,l}}\neq 0\}} \gamma_{N_{j,l}}^2 \\
&= \frac{1}{m}\sum_{j'=1}^{d'} \gamma_{N_{j',l}}^2 \\
&= \frac{d'}{m}\frac{1}{d'}\sum_{j'=1}^{d'} \gamma_{N_{j',l}}^2 \\
&\leq \frac{1}{m^{\frac{\alpha-1}{\alpha-2}}}\frac{1}{d'}\sum_{j'=1}^{d'} \gamma_{N_{j',l}}^2
\end{aligned}$$

Again, by Assumption 1

$$\lim_{m\to\infty} \frac{1}{m^{\frac{\alpha-1}{\alpha-2}}} = 0.$$

Further, by Assumption 2

$$E\left[\gamma_{N_{j',l}}^4\right] \leq B_{\gamma_N},$$

so that applying the Strong Law of Large Numbers,

$$\frac{1}{d}\sum_{j'=1}^{d'} \gamma_{N_{j',l}}^2 \xrightarrow{a.s.} E\left[\gamma_{N_{j',l}}^2\right] \leq B'_{\gamma_N},$$

and for each l:

$$0 \le p_{ll} \le \frac{1}{m^{\frac{\alpha-1}{\alpha-2}}} \frac{1}{d'} \sum_{j'=1}^{d'} \gamma^2_{N_{j',l}} \xrightarrow{a.s.} 0.$$

We now consider the off-diagonal entries($k \ne l$):

$$\begin{aligned} p_{lk} &= P_{l,k} \\ &= \frac{1}{m} \sum_{j=1}^{m} \Gamma_{N_{j,l}} \Gamma_{N_{j,k}} \\ &= \frac{1}{m} \sum_{j=1}^{m} \gamma_{N_{j,l}} \gamma_{N{j,k}} \\ &= \frac{1}{m} \sum_{\{j: \gamma_{N_{j,l}} \ne 0 \text{ and } \gamma_{\mathrm{N_{j,k}}} \ne 0\}} \gamma_{N_{j,l}} \gamma_{N{j,k}} + \frac{1}{m} \sum_{\{j: \gamma_{N_{j,l}} = 0 \text{ or } \gamma_{\mathrm{N_{j,k}}} = 0\}} \gamma_{N_{j,l}} \gamma_{N{j,k}} \\ &= \frac{1}{m} \sum_{\{j: \gamma_{N_{j,l}} \ne 0 \text{ and } \gamma_{\mathrm{N_{j,k}}} \ne 0\}} \gamma_{N_{j,l}} \gamma_{N{j,k}} \end{aligned}$$

If both $\gamma_{N_{j,l}} \ne 0$ and $\gamma_{N_{j,k}} \ne 0$ then gene j shares and edge with both genes l and k, so that d', the number of j such that $\gamma_{N_{j,l}} \ne 0$ and $\gamma_{N_{j,k}} \ne 0$ will be bounded by the maximum of the degrees of vertices l and k. The same argument as used for the diagonal entries then follows:

$$p_{lk} \le \frac{1}{m^{\frac{\alpha-1}{\alpha-2}}} \frac{1}{d''} \sum_{j'=1}^{d'} \gamma_{N_{j',l}} \gamma_{N_{j',k}},$$

and

$$\lim_{m \to \infty} \frac{1}{m^{\frac{\alpha-1}{\alpha-2}}} = 0.$$

Further, for any j', by Assumption 2 and the Cauchy-Schwarz inequality to $|\gamma_{N_{j',l}}|$ and $|\gamma_{N_{j',k}}|$

$$\begin{aligned} E\left[|\gamma_{N_{j',l}} \gamma_{N_{j',k}}|\right] &\le \sqrt{E\left[|\gamma_{N_{j',l}}|^2\right] E\left[|\gamma_{N_{j',k}}|^2\right]} \\ &= \sqrt{E\left[(\gamma_{N_{j',l}})^2\right] E\left[(\gamma_{N_{j',k}})^2\right]} \\ &\le \sqrt{B'_{\gamma_N} \times B'_{\gamma_N}} \end{aligned}$$

and

$$-\infty < -(B'_{\gamma_N})^2 \le E\left[\gamma_{N_{j',l}} \gamma_{N_{j',k}}\right] (B'_{\gamma_N})^2 < \infty,$$

and by the Strong Law of Large Numbers,

$$\frac{1}{d''}\sum_{j'=1}^{d''}\gamma_{N_{j',l}}\gamma_{N_{j',k}} \xrightarrow{a.s.} E\left[\gamma_{N_{j',l}}\gamma_{N_{j',k}}\right],$$

therefore, for each $l \neq k$:

$$p_{lk} \leq \frac{1}{m^{\frac{\alpha-1}{\alpha-2}}}\frac{1}{d''}\sum_{j'=1}^{d''}\gamma_{N_{j',l}}\gamma_{N_{j',k}} \xrightarrow{a.s.} 0.$$

Therefore, both the diagonal and off-diagonal entries in P converge to zero, and

$$P \xrightarrow{a.s.} 0.$$

As shown in Assumption 5, $N \xrightarrow{a.s.} 0$, so that by Slutsky's Theorem:

$$\lim_{m\to\infty}\frac{1}{m}N^T\Gamma_N^T\Gamma_N N = 0$$

4.

$$\lim_{m\to\infty}\frac{1}{m}\mathbf{U}^T\Gamma_N N$$

This term converges almost surely to zero by the KSLLN since $E\left[U\right] = 0$ and Γ_N and U have bounded fourth moments.

5.

$$\lim_{m\to\infty}\frac{1}{m}N^T\Gamma_N^T\mathbf{U}$$

This term converges almost surely to zero by the KSLLN since $E\left[U\right] = 0$ and Γ_N and U have bounded fourth moments.

Therefore, all of the terms (1)-(5) converge almost surely to zero and the limit of the empirical variance-covariance matrix is

$$\lim_{m\to\infty}\frac{1}{m}\mathbf{X}^T\mathbf{X} - \hat{\sigma}^2_{ave}\mathbf{I} = \mathbf{A^T}\boldsymbol{\Delta}\mathbf{A} + \underbrace{\lim_{m\to\infty}\frac{1}{m}A^T\Gamma_A^T\Gamma_N N}_{(1)} + \underbrace{\lim_{m\to\infty}\frac{1}{m}N^T\Gamma_N^T\Gamma_A A}_{(2)} +$$

$$\underbrace{\lim_{m\to\infty}\frac{1}{m}N^T\Gamma_N^T\Gamma_N N}_{(3)} + \underbrace{\lim_{m\to\infty}\frac{1}{m}\mathbf{U}^T\Gamma_N N}_{(4)} + \underbrace{\lim_{m\to\infty}\frac{1}{m}N^T\Gamma_N^T\mathbf{U}}_{(5)} = A^T\Delta A = \mathcal{L}$$

The principal components of this matrix consistently estimate the space spanned by the confounding artifacts as we have previously demonstrated [169].

Therefore we show that given confounded high-dimensional gene expression data where the number of genes is much larger than the number of samples - top principal components will consistently estimate artifacts, and not network structure.

A.2 Effect of fewer PC correction on reconstruction of co-expression networks with WGCNA and graphical lasso

Since broad trends in co-expression may sometimes reflect distant regulatory relationships between genes,to ensure that we are not removing true long range signals, we also reconstructed networks with data corrected for one quarter and half the number of PCs estimated by our correction method. With WGCNA, we found that using a half of the estimated number of PCs sometimes performed better in lung and skin. For the remaining tissues half-PC correction does reduce false discoveries compared to uncorrected data, however using the complete number of estimated PCs performs better (Supplementary Figure 6).

With graphical lasso networks, correcting data with fewer PCs does improve FDR compared to uncorrected data. However, the networks built with data corrected with complete PCs performed either better or similar to fewer number of PCs(Supplementary Figure 7).

Supplementary Tables

Study	Network reconstruction method	Correction approach
[175]	WGCNA	known technical factors - RIN, pH, PMI, age, batch, preservation, and gender
[15]	WGCNA	none
[176]	WGCNA	none, voom normalization
[177]	WGCNA	none, quantile normalization
[178]	WGCNA	none, quantile normalization
[179]	WGCNA	batch correction
[180]	WGCNA	none prior to network reconstruction. After networks were reconstructed, tested for confounding through module eigengene-trait correlations. however these did not include technical confounders like batch, etc.
[181]	WGCNA	none, tmm normalization
[182]	Bayesian biclustering	network learning method jointly models hidden confounders
[183]	WGCNA	known technical covariates: diagnosis status, Age of death, sex, PMI, pH, RIN, clustered processing batch, and ancestry markers
[14]	Graphical lasso	hidden factor correction
[184]	WGCNA	batch correction
[185]	WGCNA	none, quantile normalization

Table A-I. Few studies account for artifacts during re-construction of co-expression networks.

Tissue	Known covariate
Adipose Subcutaneous	- Code for BSS collection site - RNA integrity number (RIN) - Type of nucleic acid isolation batch - Estimated library size - Mean coefficient of variance - Transcripts detected - Intronic rate - Expression profiling efficiency - # transcripts that have at least one read in their 5' end - % intragenic End 2 reads sequenced in sense direction - gene GC%
Lung	- Autolysis score - Code for BSS collection site - RNA integrity number (RIN) - Type of nucleic acid batch - End 2 mapping rate - 3' 50-base normalization - Transcripts detected - Gap percentage - Intronic rate - % intragenic End 1 reads sequenced in sense direction - % intragenic End 2 reads sequenced in sense direction - Gene GC%
Skeletal Muscle	- Code for BSS collection site - Type of nucleic acid isolation batch - chimeric pairs - 3' 50-base normalization - Library size - Intergenic rate - Transcripts detected - Gap percentage - Intronic rate - Mapped unique rate of total - % intragenic End 1 reads sequenced in sense direction - # transcripts that have at least one read in their 5' end - Duplication rate of mapped - Gene GC%
Thyroid	- Code for BSS collection site - Autolysis score - Type of nucleic acid isolation batch - RNA integrity number - 3' 50 base normalization

	- Library size - Intergenic rate - Reads designated as failed by sequencer - Transcripts detected - Intronic rate - Expression profiling efficiency - # transcripts that have at least one read in their 5' end - Duplication rate of mapped - % intragenic end 2 reads sequenced in sense direction - Gene GC%
Whole Blood	- Mapped read count - Code for BSS collection site - RNA integrity number (RIN) - Time point reference for Start and End times of sample procurement - Chimeric pairs - 5' 50-base normalization - 3' 50-base normalization - mean coverage per base - Library size - Reads designated as failed by sequencer - Mean coefficient of variance - Transcripts detected - Gap percentage - Intronic rate - Alternative alignments - % intragenic end 2 reads sequenced in sense direction - Gene GC%

Table A-II. Known covariates regressed from gene-expression data for multiple covariate based correction. The expression variance explained (adjusted R^2) by these covariates was $>= 0.01$.

Supplementary Figures

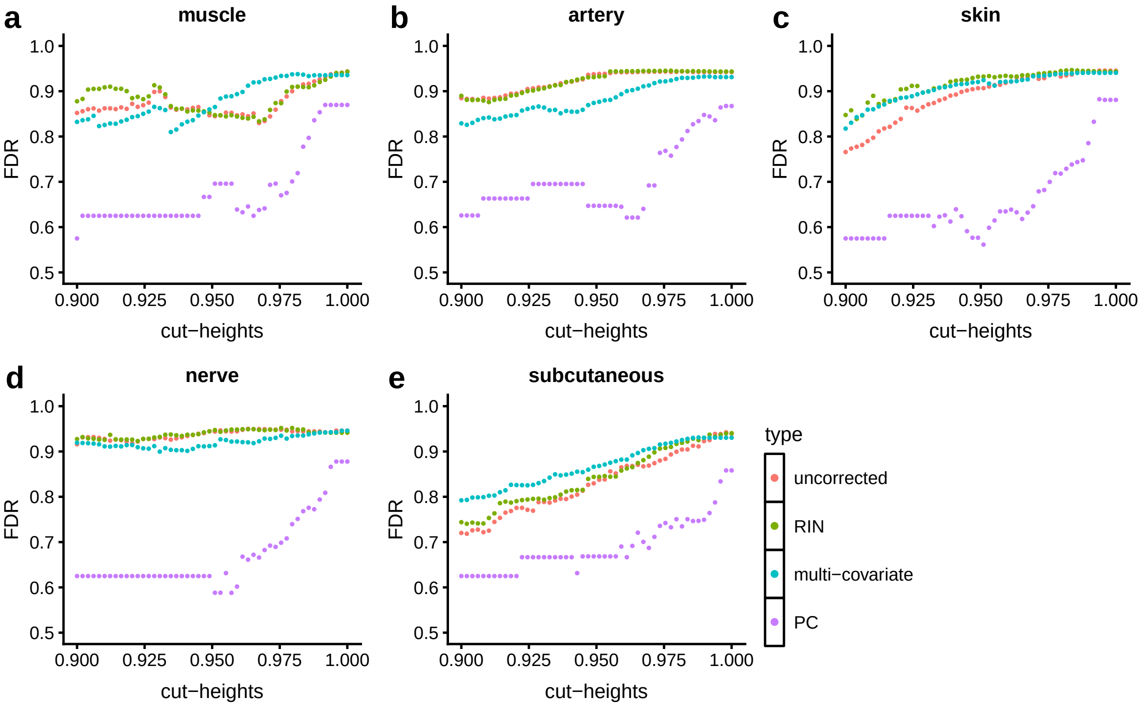


Figure A-1. False discovery rates of WGCNA networks obtained at a varying cut-heights with uncorrected, RIN corrected, multiple covariate corrected and PC corrected data. Most tissues show considerable reduction in false discoveries after PC correction. PC correction shows only moderate improvement on FDR in sun-exposed skin.

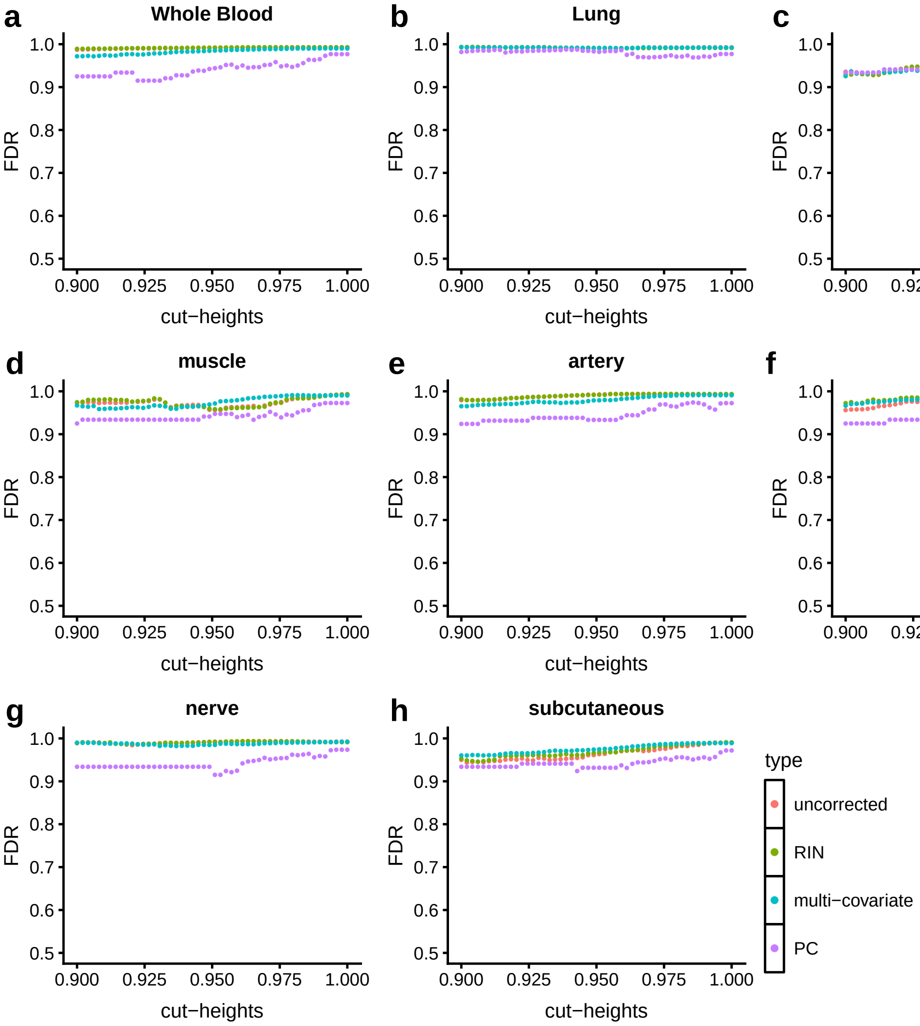


Figure A-2. False discovery rates of WGCNA networks using shared list of true positives obtained from canonical pathway database (gene pairs present in at least two pathway databases). Each color corresponds to the correction approach, and each point in the figure corresponds to FDR of the network at specific cut-height.

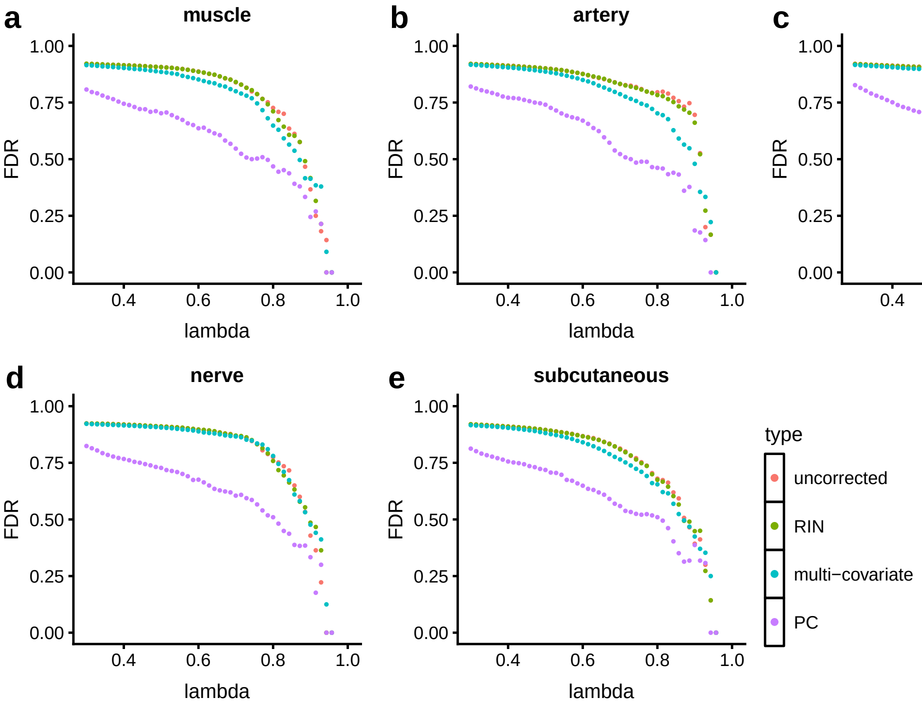


Figure A-3. False discovery rates of graphical lasso networks using canonical pathway databases. Networks were obtained at a varying values of penalty parameter (0.3 - 1.0). Each color corresponds to the correction approach, and each point corresponds to the network obtained at a specific lambda.

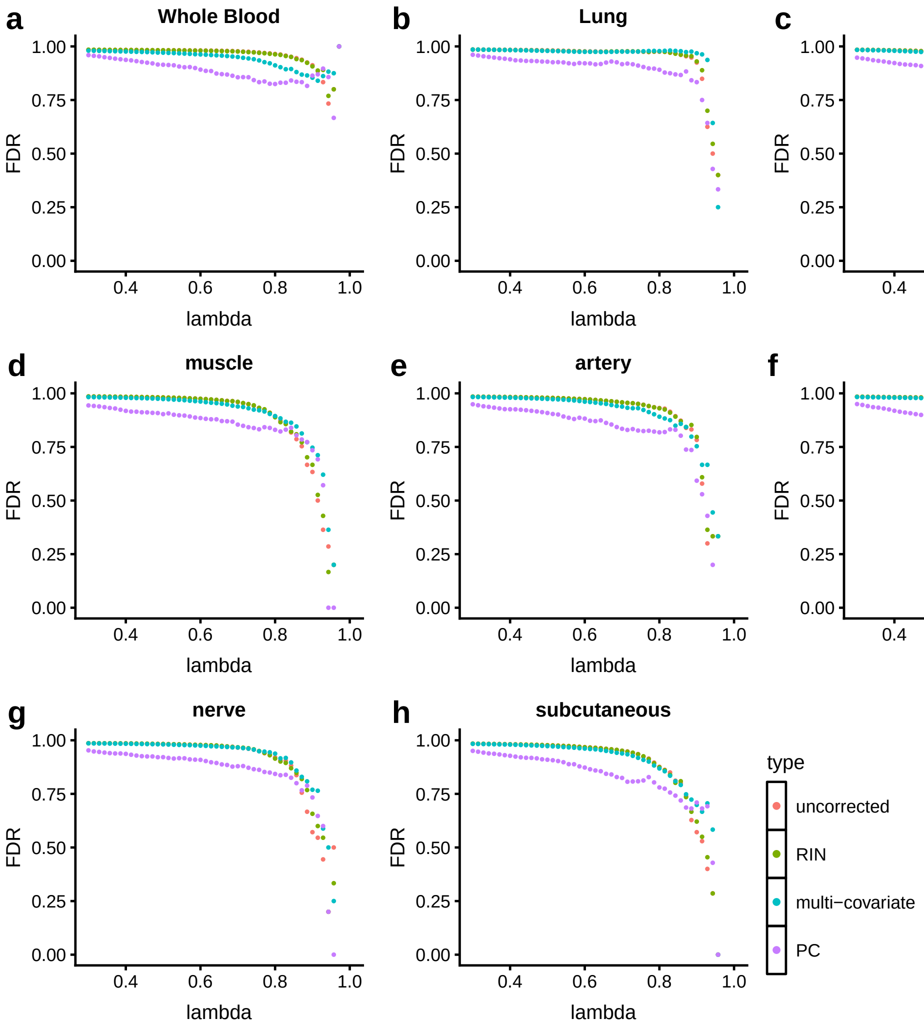
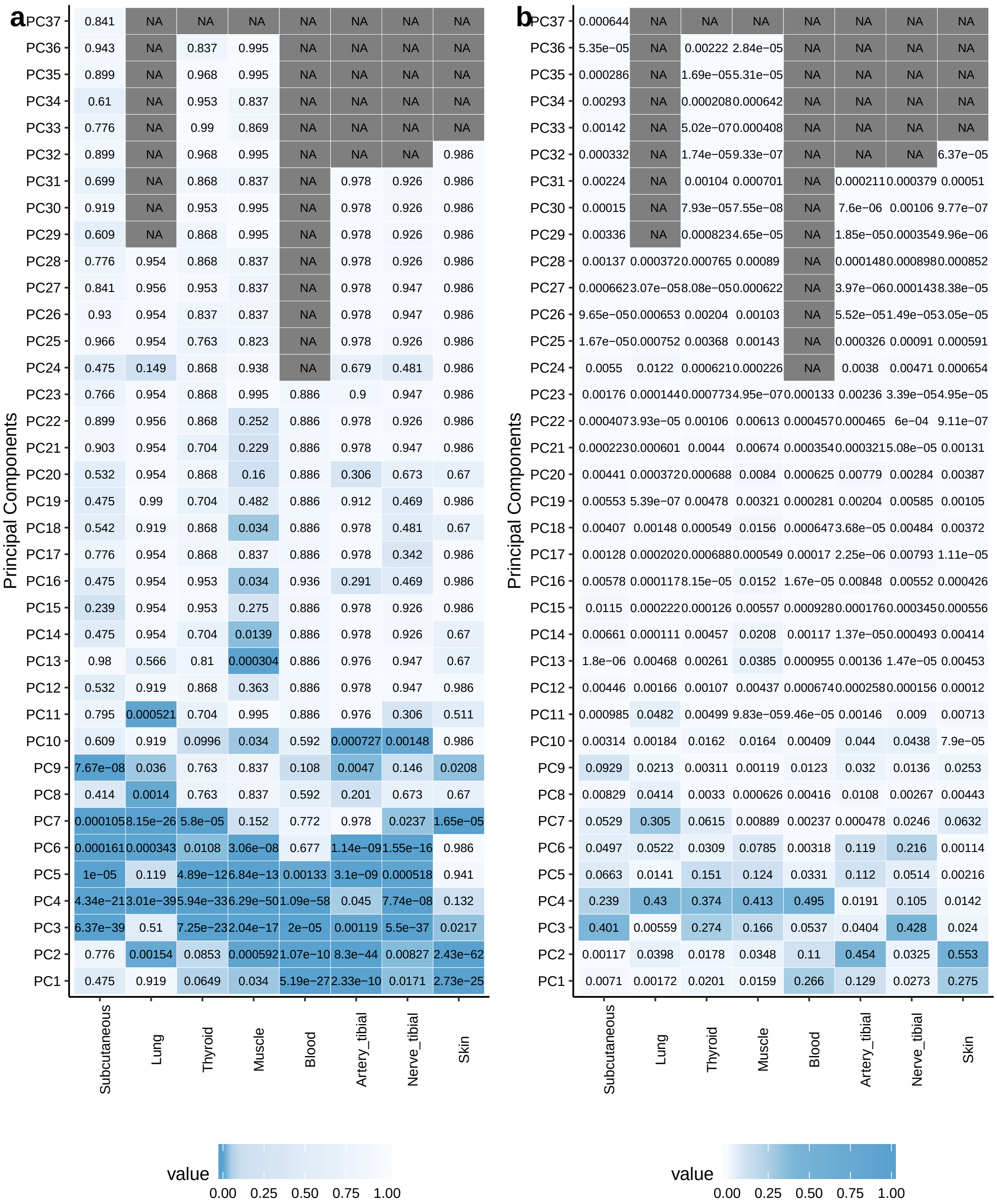


Figure A-4. False discovery rates of graphical lasso networks using shared list of true positives obtained from canonical pathway database (gene pairs present in at least two pathway databases). Each color corresponds to the correction approach, and each point in the figure corresponds to FDR of the network at specific value of penalty parameter value (lambda = [0.3, 1.0]).



a

	Subcutaneous	Lung	Thyroid	Muscle	Blood	Artery_tibial	Nerve_tibial	Skin
PC37	0.841	NA	NA	NA	NA	NA	NA	NA
PC36	0.943	NA	0.837	0.995	NA	NA	NA	NA
PC35	0.899	NA	0.968	0.995	NA	NA	NA	NA
PC34	0.61	NA	0.953	0.837	NA	NA	NA	NA
PC33	0.776	NA	0.99	0.869	NA	NA	NA	NA
PC32	0.899	NA	0.968	0.995	NA	NA	NA	0.986
PC31	0.699	NA	0.868	0.837	NA	0.978	0.926	0.986
PC30	0.919	NA	0.953	0.995	NA	0.978	0.926	0.986
PC29	0.609	NA	0.868	0.995	NA	0.978	0.926	0.986
PC28	0.776	0.954	0.868	0.837	NA	0.978	0.926	0.986
PC27	0.841	0.956	0.953	0.837	NA	0.978	0.947	0.986
PC26	0.93	0.954	0.837	0.837	NA	0.978	0.947	0.986
PC25	0.966	0.954	0.763	0.823	NA	0.978	0.926	0.986
PC24	0.475	0.149	0.868	0.938	NA	0.679	0.481	0.986
PC23	0.766	0.954	0.868	0.995	0.886	0.9	0.947	0.986
PC22	0.899	0.956	0.868	0.252	0.886	0.978	0.926	0.986
PC21	0.903	0.954	0.704	0.229	0.886	0.978	0.947	0.986
PC20	0.532	0.954	0.868	0.16	0.886	0.306	0.673	0.67
PC19	0.475	0.99	0.704	0.482	0.886	0.912	0.469	0.986
PC18	0.542	0.919	0.868	0.034	0.886	0.978	0.481	0.67
PC17	0.776	0.954	0.868	0.837	0.886	0.978	0.342	0.986
PC16	0.475	0.954	0.953	0.034	0.936	0.291	0.469	0.986
PC15	0.239	0.954	0.953	0.275	0.886	0.978	0.926	0.986
PC14	0.475	0.954	0.704	0.0139	0.886	0.978	0.926	0.67
PC13	0.98	0.566	0.81	0.000304	0.886	0.976	0.947	0.67
PC12	0.532	0.919	0.868	0.363	0.886	0.978	0.947	0.986
PC11	0.795	0.000521	0.704	0.995	0.886	0.976	0.306	0.511
PC10	0.609	0.919	0.0996	0.034	0.592	0.000727	0.00148	0.986
PC9	7.67e−08	0.036	0.763	0.837	0.108	0.0047	0.146	0.0208
PC8	0.414	0.0014	0.763	0.837	0.592	0.201	0.673	0.67
PC7	0.000105	8.15e−26	5.8e−05	0.152	0.772	0.978	0.0237	1.65e−05
PC6	0.000161	0.000343	0.0108	3.06e−08	0.677	1.14e−09	1.55e−16	0.986
PC5	1e−05	0.119	4.89e−12	6.84e−13	0.00133	3.1e−09	0.000518	0.941
PC4	4.34e−21	3.01e−39	5.94e−33	6.29e−50	1.09e−58	0.045	7.74e−08	0.132
PC3	6.37e−39	0.51	7.25e−23	2.04e−17	2e−05	0.00119	5.5e−37	0.0217
PC2	0.776	0.00154	0.0853	0.000592	1.07e−10	8.3e−44	0.00827	2.43e−62
PC1	0.475	0.919	0.0649	0.034	5.19e−27	2.33e−10	0.0171	2.73e−25

b

	Subcutaneous	Lung	Thyroid	Muscle	Blood	Artery_tibial	Nerve_tibial	Skin
PC37	0.000644	NA	NA	NA	NA	NA	NA	NA
PC36	5.35e−05	NA	0.00222	2.84e−05	NA	NA	NA	NA
PC35	0.000286	NA	1.69e−05	5.31e−05	NA	NA	NA	NA
PC34	0.00293	NA	0.000208	0.000642	NA	NA	NA	NA
PC33	0.00142	NA	5.02e−07	0.000408	NA	NA	NA	NA
PC32	0.000332	NA	1.74e−05	9.33e−07	NA	NA	NA	6.37e−05
PC31	0.00224	NA	0.00104	0.000701	NA	0.000211	0.000379	0.00051
PC30	0.00015	NA	7.93e−05	7.55e−08	NA	7.6e−06	0.00106	9.77e−07
PC29	0.00336	NA	0.000823	4.65e−05	NA	1.85e−05	0.000354	9.96e−06
PC28	0.00137	0.000372	0.000765	0.00089	NA	0.000148	0.000898	0.000852
PC27	0.000662	3.07e−05	8.08e−05	0.000622	NA	3.97e−06	0.000143	8.38e−05
PC26	9.65e−05	0.000653	0.00204	0.00103	NA	5.52e−05	1.49e−05	3.05e−05
PC25	1.67e−05	0.000752	0.00368	0.00143	NA	0.000326	0.00091	0.000591
PC24	0.0055	0.0122	0.000621	0.000226	NA	0.0038	0.00471	0.000654
PC23	0.00176	0.000144	0.000773	4.95e−07	0.000133	0.00236	3.39e−05	4.95e−05
PC22	0.000407	3.93e−05	0.00106	0.00613	0.000457	0.000465	6e−04	9.11e−07
PC21	0.000223	0.000601	0.0044	0.00674	0.000354	0.000321	5.08e−05	0.00131
PC20	0.00441	0.000372	0.000688	0.0084	0.000625	0.00779	0.00284	0.00387
PC19	0.00553	5.39e−07	0.00478	0.00321	0.000281	0.00204	0.00585	0.00105
PC18	0.00407	0.00148	0.000549	0.0156	0.000647	3.68e−05	0.00484	0.00372
PC17	0.00128	0.000202	0.000688	0.000549	0.00017	2.25e−06	0.00793	1.11e−05
PC16	0.00578	0.000117	8.15e−05	0.0152	1.67e−05	0.00848	0.00552	0.000426
PC15	0.0115	0.000222	0.000126	0.00557	0.000928	0.000176	0.000345	0.000556
PC14	0.00661	0.000111	0.00457	0.0208	0.00117	1.37e−05	0.000493	0.00414
PC13	1.8e−06	0.00468	0.00261	0.0385	0.000955	0.00136	1.47e−05	0.00453
PC12	0.00446	0.00166	0.00107	0.00437	0.000674	0.000258	0.000156	0.00012
PC11	0.000985	0.0482	0.00499	9.83e−05	9.46e−05	0.00146	0.009	0.00713
PC10	0.00314	0.00184	0.0162	0.0164	0.00409	0.044	0.0438	7.9e−05
PC9	0.0929	0.0213	0.00311	0.00119	0.0123	0.032	0.0136	0.0253
PC8	0.00829	0.0414	0.0033	0.000626	0.00416	0.0108	0.00267	0.00443
PC7	0.0529	0.305	0.0615	0.00889	0.00237	0.000478	0.0246	0.0632
PC6	0.0497	0.0522	0.0309	0.0785	0.00318	0.119	0.216	0.00114
PC5	0.0663	0.0141	0.151	0.124	0.0331	0.112	0.0514	0.00216
PC4	0.239	0.43	0.374	0.413	0.495	0.0191	0.105	0.0142
PC3	0.401	0.00559	0.274	0.166	0.0537	0.0404	0.428	0.024
PC2	0.00117	0.0398	0.0178	0.0348	0.11	0.454	0.0325	0.553
PC1	0.0071	0.00172	0.0201	0.0159	0.266	0.129	0.0273	0.275

Figure A-5. Principal component loadings of gene expression are significantly associated with estimates of sample specific GC bias. Association was tested using a linear model. Panel (a) shows BH adjusted p-values and (b) shows R-squared.

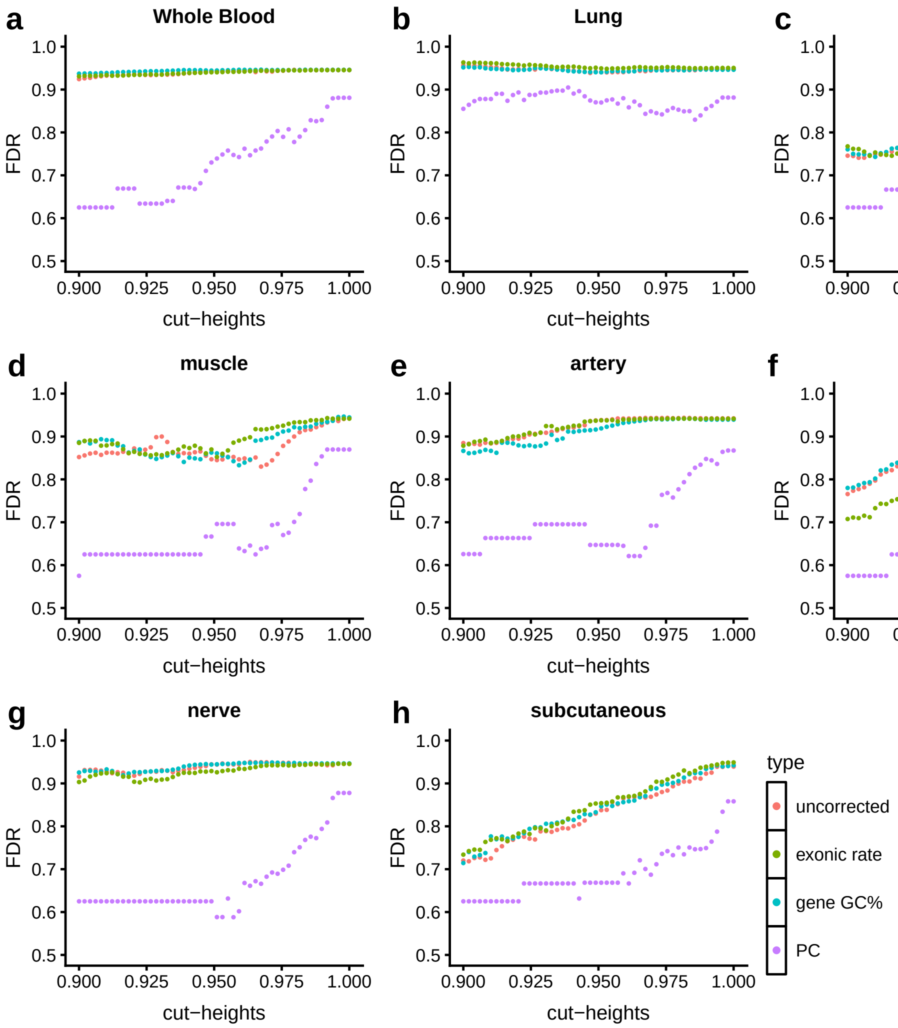


Figure A-6. False discovery rates of WGCNA modules using canonical pathway databases. Each color corresponds to the correction approach, and each point in the figure corresponds to FDR of the network at specific cut-height. Exonic rate and gene GC% are the known confounder used in this figure.

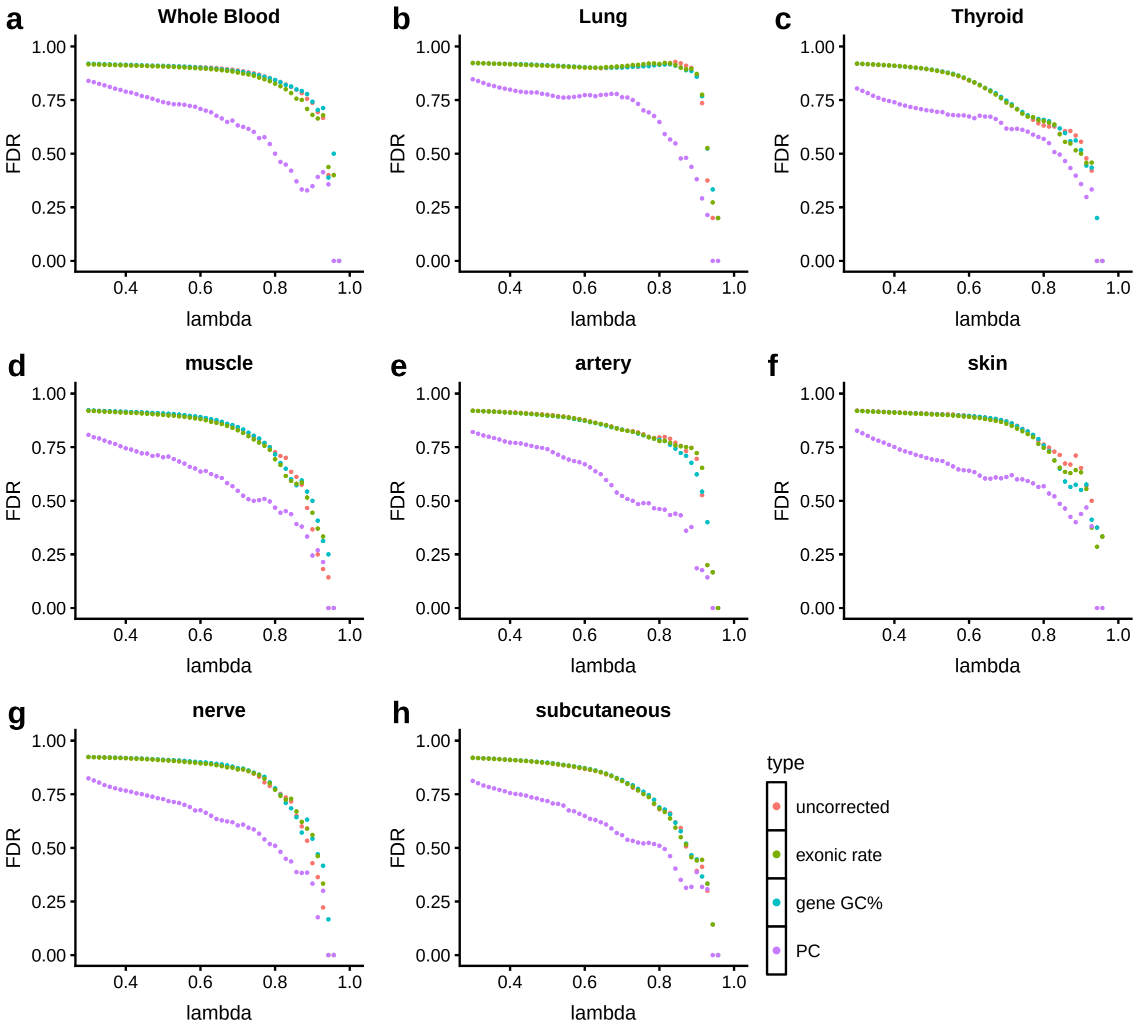


Figure A-7. False discovery rates of graphical lasso networks using canonical pathway databases. Networks were obtained at a varying values of penalty parameter (0.3 - 1.0). Each color corresponds to the correction approach, and each point corresponds to the network obtained at a specific lambda. Exonic rate and gene GC% are the known confounder used in this figure.

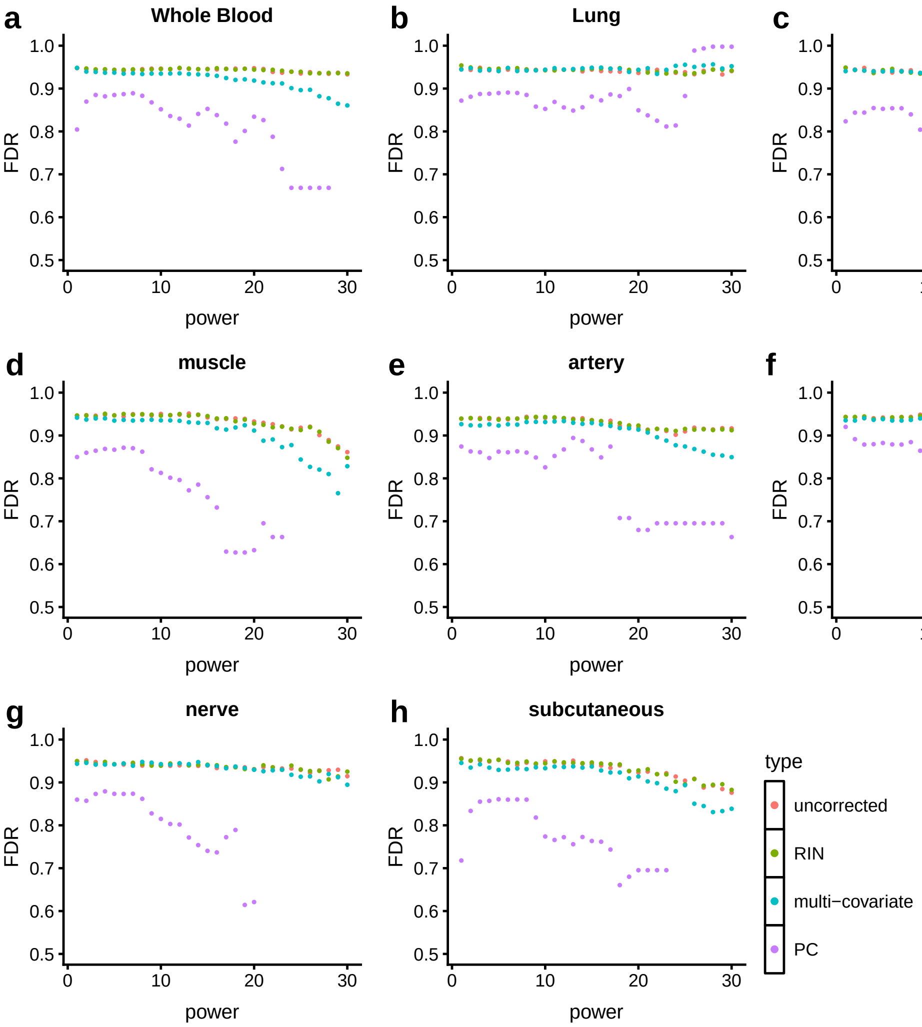


Figure A-8. False discovery rates of networks inferred with signed WGCNA networks using canonical pathways. Each color corresponds to the correction approach, and each point in the figure corresponds to FDR of networks obtained at different values of power transform β, ranging from 1 to 30.

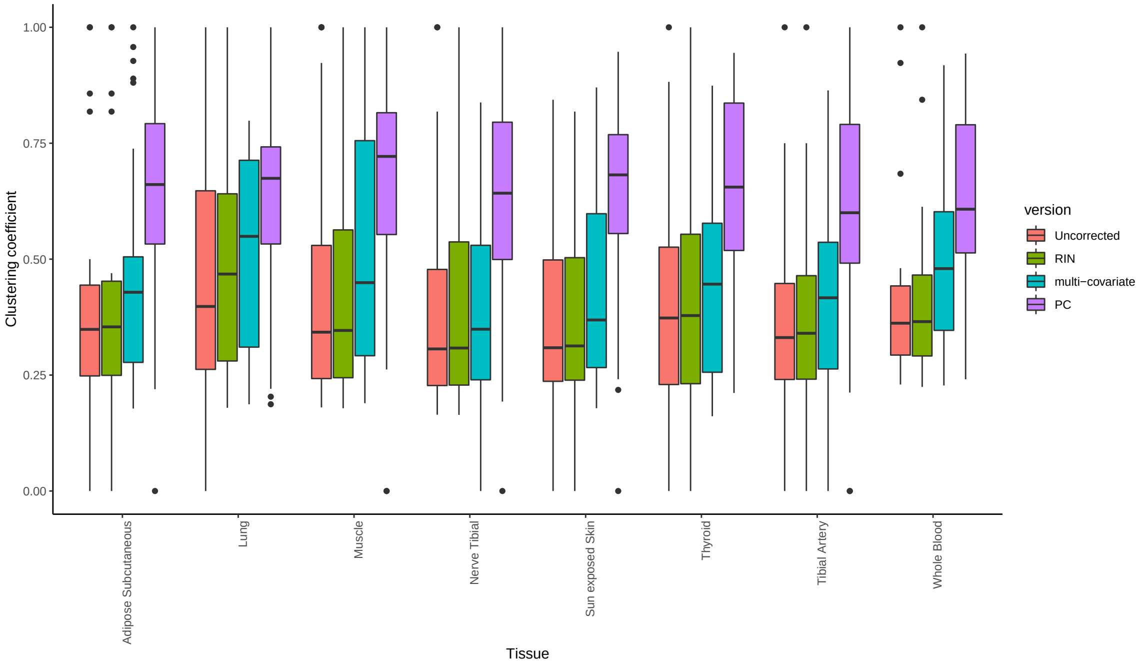


Figure A-9. Graphical lasso networks reconstructed after PC correction of gene expression measurements show higher clustering coefficient compared to uncorrected networks across all tissues. Both scale-free and small-world networks have high clustering coefficient.

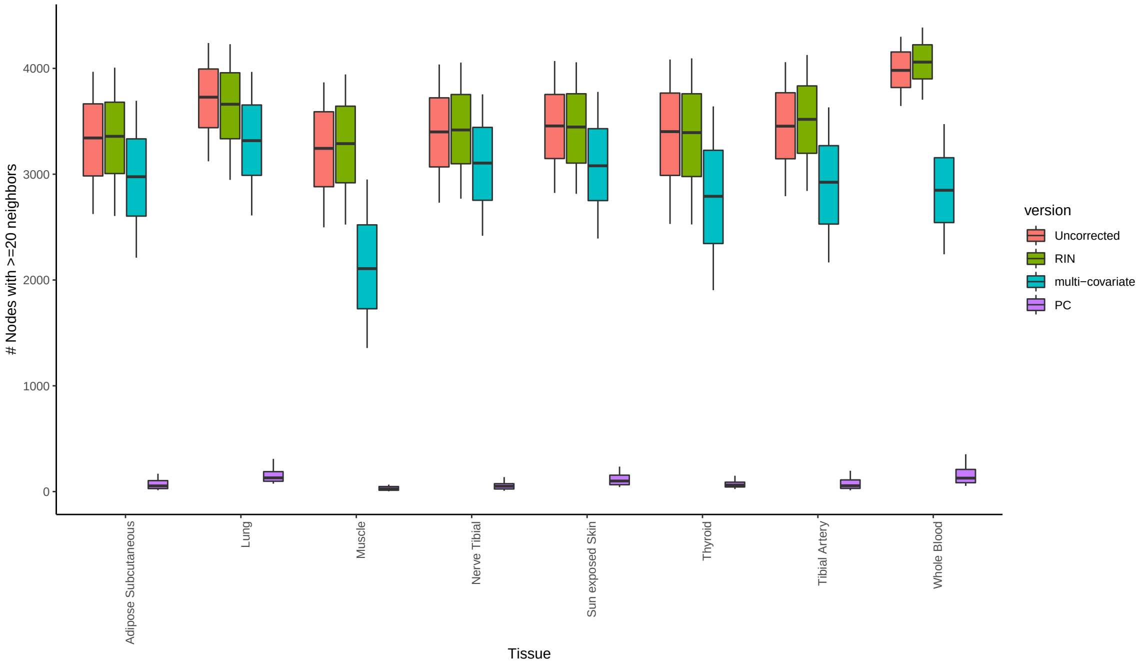


Figure A-10. Graphical lasso networks ($\lambda = [0.3, 0.43]$) reconstructed after PC correction of gene expression measurements show considerably fewer hub nodes compared to uncorrected networks across all tissues. Scale-free networks have few hub nodes.

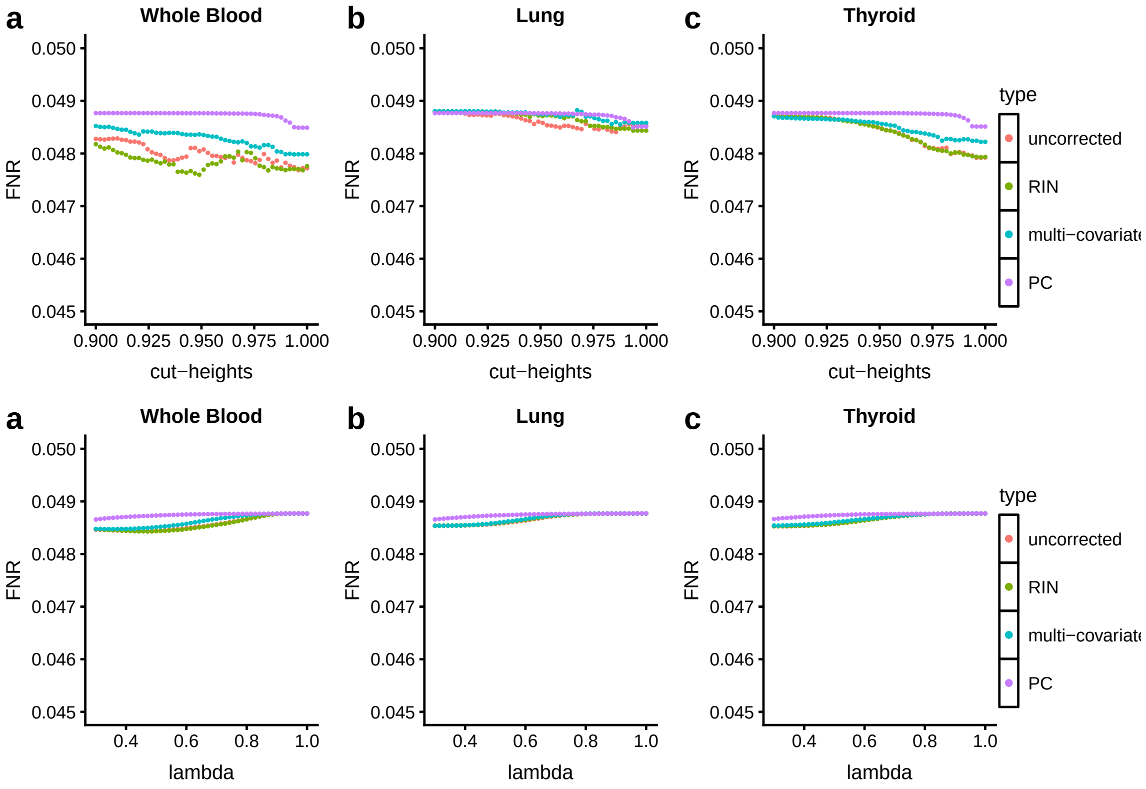


Figure A-11. Graphical lasso networks reconstructed before and after PC correction of gene expression measurements show no improvement on false negative rates.

Appendix B

Multi-study integration to identify global expression pat-terns and key regulators of Epithelial to Mesenchymal transition (EMT) in cancer

B.1 Supplementary Figures

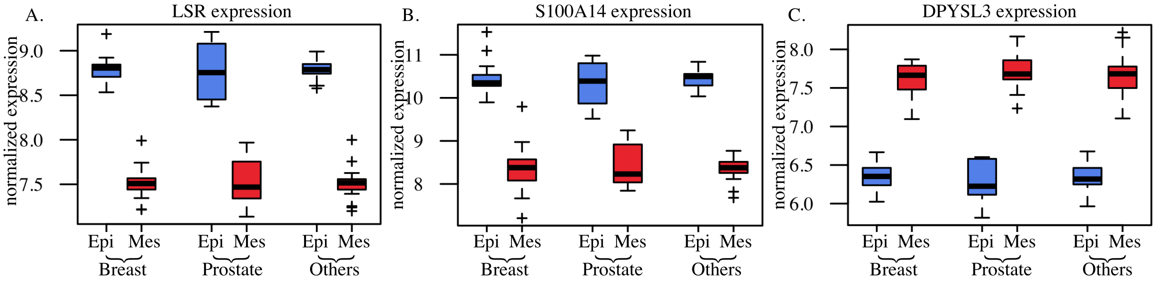


Figure B-1. Expression of EMT genes previously unknown in prostate cancer in integrated cell lines data Expression of LSR (A), S100A14 (B) and DPYSL3 (C) in breast, prostate and others (retinal pigment, liver, colon and esophageal) cancer cell lines from QN + SVA normalized integrated data

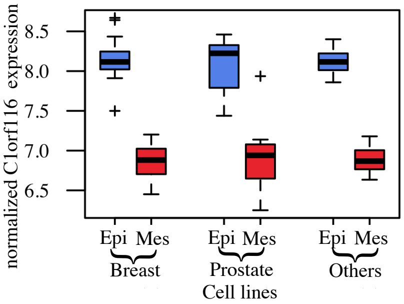


Figure B-2. Expression of C1orf116 in breast, prostate and others (retinal pigment, liver, colon and esophageal) cancer cell lines from integrated data.

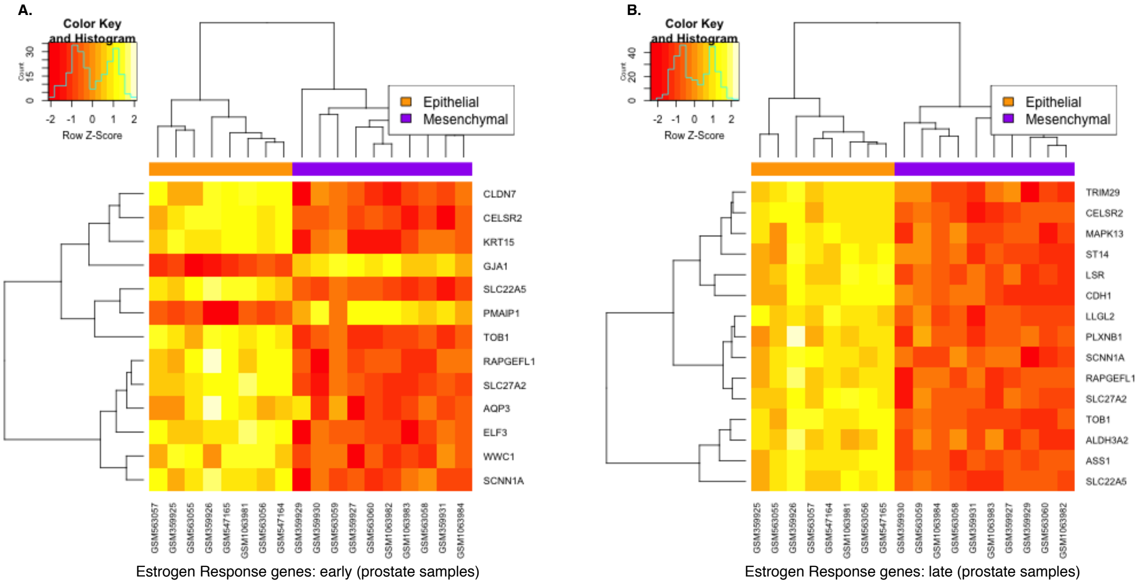


Figure B-3. Expression of Estrogen responsive genes - (A) early and (B) late in prostate cancer cell line samples from integrated data

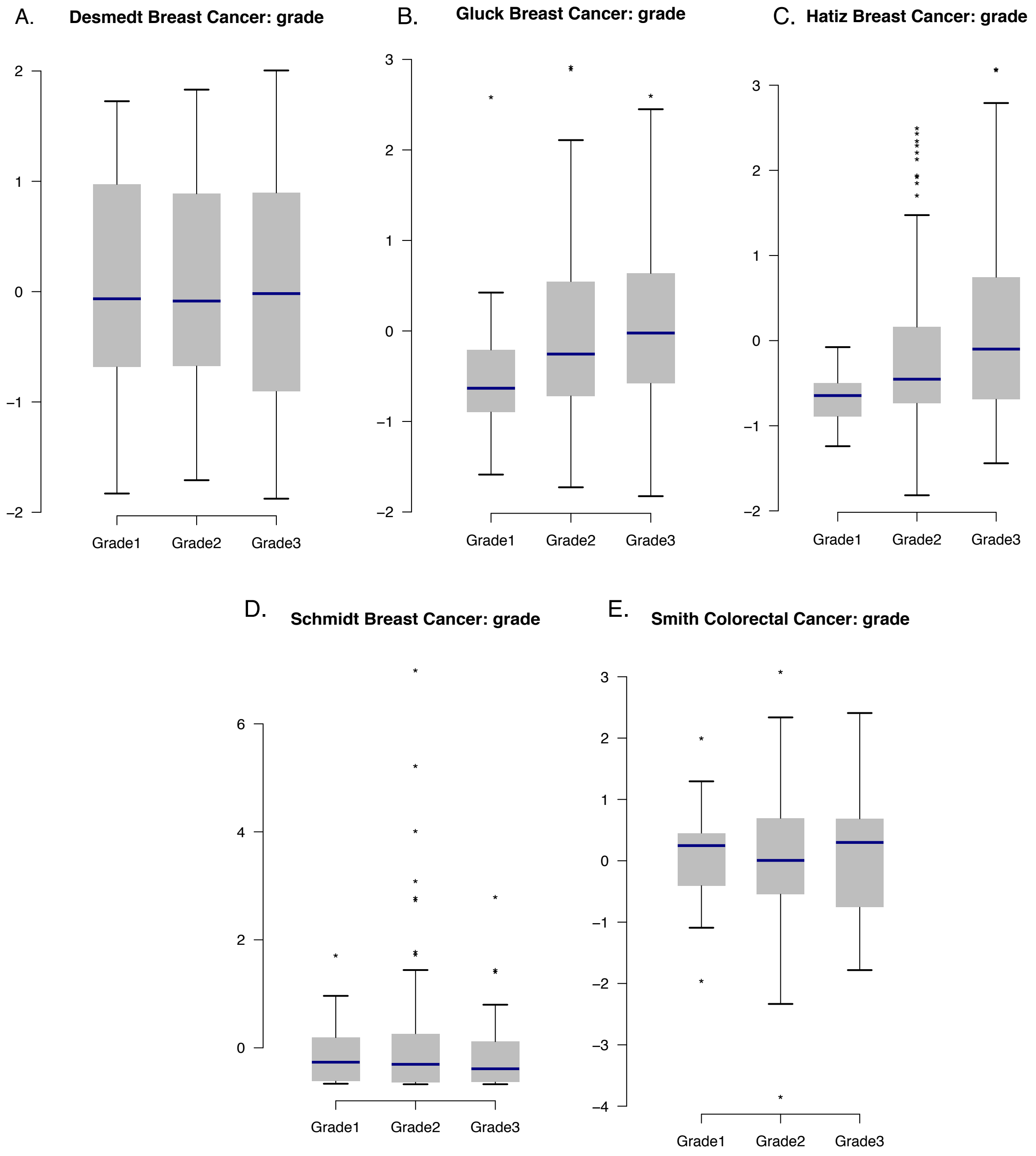


Figure B-4. C1orf116 expression in clinical patient data from breast and colorectal cancer.

B.2 Supplementary Tables

List of Antibodies
Anti-C1orf116 (HPA011888)
Anti-LSR (HPA007270)
Anti-S100A14 (HPA27613)
Anti-DPYSL3 (SAB2103348)
Anti-beta-actin (A5411) 680RD Goat anti-mouse IgG (H+L) (926-6870)

Table B-I. List of antibodies

Gene	Average rank	Gene	Average rank
MAP7	15	HJURP	190.25
FXYD3	17.5	SPINT1	191.75
EMP3	18	RASA3	195
VCAN	19.25	FAAH	198
ACSF2	30.5	HRH1	198.25
GEM	33.75	GREM1	199
EPS8L2	34.875	MTUS1	199.25
MEF2C	35	ITM2C	199.5
MMP28	35	PLAUR	199.5
LSR	35.75	DSP	203
CDH2	36.25	JAG2	203
ZEB1	36.75	CTSD	203.25
SPOCK1	40	CD320	203.5
COL4A2	40.25	WWC1	203.5
FBN1	40.5	STAP2	204
PPL	42.25	SH2D3A	204.25
PTN	44.75	CHN1	205
GNG11	46	CTGF	206.75
GFPT2	46.5	MST1R	206.75
RGL1	46.5	COL1A2	207.5
PMP22	49.25	EML1	208.25
COL5A1	51.25	SLC22A5	209.75
ASS1	52.75	CXADR	212.25
CDH1	58	LIMK2	212.75
EXTL2	60.375	PLOD1	212.75
ERMP1	60.75	HOOK2	213.375
LLGL2	67.25	MMP14	216
KDELR3	67.75	PEA15	218.5
JUP	70.25	LOX	219.25
RAPGEFL1	74.25	EPB41L3	222.25
ELF3	74.75	MPZL2	222.25

CLDN7	76.75	HAS2	222.75
SLIT2	77.25	KIRREL	224.75
MYO5C	80.75	PPFIBP2	224.75
MLLT11	81.75	ETS1	226.75
EDIL3	84.375	RHOQ	228.25
CELSR2	84.75	FJX1	228.375
SLC27A2	84.75	INADL	234.75
AKAP12	85.5	FBXO5	236
TIMP1	86	FOXG1	236.75
GLS2	87	TRIB2	238.75
DPYSL3	89	KCNMA1	239.25
COL6A3	90.25	PTPRF	243
SRPX	90.75	VIM	244
PCOLCE2	99.75	PVRL3	245
MAP1B	102	GJC1	245.25
TUBA1A	102.25	AP1M2	247.5
KRT15	105.5	FERMT2	247.5
EPB41L4B	105.75	SMURF2	249.125
FAM64A	110.25	POSTN	249.75
ST14	111	ORAI2	249.875
SLC22A4	111.5	LRBA	250
AKT3	115.25	RBMS3	252.5
FAP	117.5	CEP170	253.375
PDLIM7	117.5	FBN2	254.5
SNAPC1	119	CD70	254.75
HMOX1	120	BCAR3	255
HEY1	121.75	CHMP7	255.25
CLMN	122	GJA1	256.25
ALDH4A1	124.375	DDR2	258
RECK	125	SERPINE1	261.5
GRB7	126	SERPINE2	261.5
CXCL3	126.5	MOXD1	265.25
TIMP2	127.75	MCAM	265.5
TCF4	129.5	SHCBP1	266
CHST2	131	TMEM158	267.25
TRIM29	132.75	RAB25	267.5
PMAIP1	133.75	DAB2	269
OAS1	135	MBNL3	270.125
C1orf116	136	IL1RN	270.5
TOB1	138.25	COL1A1	272.25
LPAR2	138.5	DAAM1	273.5
CDH11	140	AQP3	274.625
ALDH3A2	144.5	MMP2	275.5
TRIM26	145.75	CLPX	279
ABCA12	146	PSIP1	280.75

S100A14	146.75	DHRS1	281.25
LHFP	150.875	NMNAT2	283
AP1G2	152.25	TWF2	283.25
CDS1	152.875	ZEB2	283.875
TGFB1	154.75	PCOLCE	284.5
HSD17B8	156.75	BCL2A1	284.75
ERBB3	157.375	VAMP8	284.75
SPARC	164.25	SLC2A9	285.75
GLT8D2	166.125	NAV1	286
NCAM1	169	DTX4	288
PKP2	170.5	ENO2	290
SLC35D2	172.5	SLC25A37	290
COL6A1	172.75	ANTXR1	292.75
CYR61	173.25	CASK	294.75
FZD5	173.5	LGALS1	300.5
PLXNB1	173.5	TSPAN5	300.75
LRP8	177	CREG1	305
LRRC1	177.25	FZD2	306.25
WNT5A	179.5	SCNN1A	306.375
MAPK13	180	DDR1	307
JAM3	181	CLN3	308.375
CD59	183	ECH1	310
PRSS8	183.5	SLC27A3	314
SULT1A1	189.25	CEBPA	314.5

Table B-II. List of top 200 ranked differentially expressed genes

Genes common with Groger et. al
CDH1
CDH11
CDH2
CDS1
COL1A1
COL5A1
COL6A1
COL6A3
CTGF
CXADR
ELF3
EML1
EMP3
FBN1

FXYD3
HAS2
JUP
KRT15
LOX
LSR
MAP1B
MAP7
MMP2
MPZL2
MTUS1
PKP2
PLXNB1
PMP22
PPL
PRSS8
RECK
SERPINE1
SERPINE2
SLC22A4
SLC27A2
SPINT1
SPOCK1
TMEM158
TUBA1A
VCAN
VIM
WNT5A
ZEB1

Table B-III. Common genes with Groger et. al. study and 200 DE genes

Vita

Princy Parsana's undergraduate training was in Biotechnology from Padmashree Dr. D. Y. Patil University, Navi Mumbai. She received her Masters in Bioinformatics from Johns Hopkins University (JHU) in 2013. Currently, Princy is a member of Dr. Alexis Battle's lab, and a PhD candidate in Computer Science at JHU. She is also co-advised by Dr. Kenneth Pienta. Her research aims to build machine learning models to discover functional relationships between genes, and identify patterns of sharing and specificity across human tissues and diseases. Princy enjoys mentoring and working with undergraduate, masters, and junior PhD students. She has co-instructed a freshman class on data science in genomics and healthcare; and has given guest lectures for computational genomics courses at Hopkins. Princy has been an advocate for building community among students and cares about diversity, inclusion and equity. She co-led the effort to form CS graduate student council, and continues to serve as Academic Head. Princy was also President of the Indian Graduate Students Association, student-faculty liaison at JHU-CS, and professional development chair of graduate women in CS and ECE. Princy is invested in applying her background in machine learning and genetics to develop methods that will enable discovery of interpretable biomakers for patient stratification, including early diagnosis. After finishing her graduate work, Princy will be joining Guardant Health as a Bioinformatics Scientist.