THE INTERPRETATION OF GENE COEXPRESSISON IN SYSTEMS BIOLOGY

by

Marjan Farahbod

MSc, Chalmers University of Technology, 2013

BSc, University of Tehran, 2010

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Bioinformatics)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

December 2019

© Marjan Farahbod, 2019
The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

The interpretation of gene coexpression in systems biology

submitted by Marjan Farahbod in partial fulfillment of the requirements for the degree of Doctor of philosophy in Bioinformatics

Examinining Committee:

Dr. Paul Pavlidis
Supervisor

Dr. Alexandre Bouchard Côté
Supervisory Committee Member

Dr. Joerg Gsponer
Supervisory Committee Member

Dr. Marco A. Marra
University Examiner

Dr. T. Michael Underhill
University Examiner

Additional Supervisory Committee Members:

Dr. Wyeth Wasserman
Supervisory Committee Member

Supervisory Committee Member
Abstract

One of the key features of transcriptomic data is the similarity of expression patterns among groups of genes, referred to as coexpression. It has been shown that coexpressed genes tend to share similar functions. Based on this, a common assumption is that gene coexpression is a result of transcriptional regulation and therefore, regulatory relationships could be inferred from coexpression. However, success in inferring such relationships has been limited and there are questions about the source and interpretation of coexpression. Here I explore coexpression as an observed signal from the data, examine its source and assess its relevance for inferring regulatory relationships.

In chapter 2 I studied differential coexpression, which refers to the alteration of gene coexpression between biological conditions. It is commonly assumed that differential coexpression can reveal rewiring of transcription regulatory networks, specifically among the genes that maintain their average expression level between the conditions. However, I show that to a large extent and in contrast to this common assumption, differential coexpression is more parsimoniously explained by changes in average expression levels. This finding demonstrates limitations for inference of regulatory rewiring from coexpression and poses questions for the underlying causes of the observed coexpression.

In Chapter 3, I studied cellular composition variation among bulk tissue samples as a source of variance and the observed coexpression. I found that for most genes, differences in expression levels across cell types account for a large fraction of their variance and as a result genes with similar cell-type expression profiles appear to be coexpressed. Finally, I showed that this coexpression dominates the underlying intra-cell-type coexpression and also has the two prominent features of coexpression in the bulk tissue: reproducibility and biological relevance.

Through my studies, I was able to provide an explanation for much of the observed coexpression in the bulk tissue and shed light on its resolution and limitation for inference of regulatory relationships. I also studied coexpression in single-nucleus data and show that some of the observed coexpression in it is likely to be attributed to the transcriptional regulation, which could be a subject for future studies.
Lay Summary

Based on the cell’s state and context, functional units of the cellular machinery are
generated from the genome through a process called gene expression. These units work in
collaborating groups to perform various cellular functions. Coexpression is a process in
which a group of genes are expressed coordinately, likely due to their required
collaboration in a cellular process. Therefore, understanding which genes, and why and
when they are coexpressed, is extremely helpful for untangling the complexity of cellular
mechanisms. Coexpression analysis can roughly tell us which genes are being expressed
simultaneously. While this type of analysis has potential for increasing our understanding
of cellular functions, there are certain issues with the interpretation of the observed signal.
Here I address these issues by examining the signal from a data analytics perspective to
delineate its limits and pitfalls in interpretation, in order to ultimately present alternative
approaches for improvement.
Preface

All the work presented in this dissertation was conducted in the Michael Smith Laboratories at the University of British Columbia. I was solely responsible for conducting the research and the primary author of every chapter and corresponding publications. My supervisor, Paul Pavlidis, contributed ideas, supervision as well as text, and editorial suggestions for all chapters.

A version of chapter 2 has been published [1]. I was the lead investigator, responsible for all major areas of concept formation, data collection and analysis, as well as manuscript composition. Paul Pavlidis was the supervisory author on this project and was involved throughout the project in concept formation and manuscript composition and writing.

At the time of writing, a version of Chapter 3 has been made available online [2] and is currently under review for publication. I was the lead investigator, responsible for all major areas of concept formation, data collection and analysis, as well as manuscript composition. Paul Pavlidis was the supervisory author on this project and was involved throughout the project in concept formation and manuscript composition and writing.


Table of Contents

Abstract........................................................................................................................................... iii
Lay Summary................................................................................................................................. iv
Preface............................................................................................................................................... v
Table of Contents .......................................................................................................................... vi
List of Tables .................................................................................................................................... x
List of Figures.................................................................................................................................... xi
List of Supplementary Material ................................................................................................. xiii
List of Abbreviations .................................................................................................................... xv
Acknowledgements ........................................................................................................................ xvii
Dedication .......................................................................................................................................... xviii

Chapter 1: Introduction ................................................................................................................ 1
  1.1 Codifying gene function ........................................................................................................ 4
  1.2 Regulation of transcription .................................................................................................. 5
  1.3 Transcriptomic data ............................................................................................................. 8
    1.3.1 Platforms and specifics .................................................................................................. 8
    1.3.2 Noise and confounding factors in expression data ..................................................... 9
  1.4 Coexpression analysis ......................................................................................................... 11
    1.4.1 Gene clustering in transcriptomic data ....................................................................... 12
    1.4.2 Coexpression networks ............................................................................................. 14
      1.4.2.1 Similarity measure ............................................................................................... 14
      1.4.2.2 Network characteristics and interpretation ........................................................ 15
    1.4.3 Reproducibility and conservation of coexpression ................................................... 18
2.3.2 Many links are tissue-specific.................................................................51
2.3.3 Expression level changes explain much of the tissue specific
coexpression........................................................................................................52
2.3.4 Reproducibility and validation of the networks.................................54
2.3.5 Characterization of “pure” differential coexpression .......................57
2.4 Discussion........................................................................................................58

Chapter 3: Untangling the Effect of Cellular Composition on Coexpression Analysis ..63
3.1 Introduction.....................................................................................................63
3.2 Methods..........................................................................................................64
  3.2.1 Data ...........................................................................................................64
  3.2.2 Single-nucleus data..................................................................................65
  3.2.3 GTEx datasets and networks.................................................................66
  3.2.4 Identification of marker genes...............................................................67
  3.2.5 Modelling expression level of genes in the GTExBulk dataset.............68
  3.2.6 Enrichment of functional terms .............................................................68
  3.2.7 Synthesized bulk datasets .....................................................................69
3.3 Results..............................................................................................................69
  3.3.1 Gene variance is highly affected by cellular composition.................69
  3.3.2 Compositional variation explains much of the bulk tissue variance ....75
  3.3.3 Cellular composition effects can mask the intra-cell-type coexpression.80
3.4 Discussion.......................................................................................................89

Chapter 4: Conclusion..........................................................................................93

Bibliography .........................................................................................................100

viii
Appendices.................................................................................................................................. 122

Appendix A Supplementary Materials for Chapter 2 .................................................................122

A.1  Disease annotation for the genes..............................................................122
A.2  Tissue specificity score for GTEx data.......................................................122
A.3  Examining topological overlap for partners of pure links .................122
A.4  Contribution of the pure links to tissue-specific enrichment of
functional terms in TSNs ..................................................................................124
A.5  Contribution of expression-induced links to the tissue-specific
enrichment of functional terms in TSNs.........................................................126
A.6  Enrichment of the disease ontology terms in the networks ..........127
A.7  Figures.........................................................................................................128
A.8  Tables .........................................................................................................134
A.9  List of supplementary files and their content.....................................140

Appendix B Supplementary Materials for Chapter 3 .........................................................145

B.1  Modeling and simulation.................................................................145
B.2  Gene variance explained by the variance of marker genes ..............147
B.3  Figures.........................................................................................................149
B.4  Tables .........................................................................................................159
B.5  Details and description of additional files......................................161
List of Tables

Table A- 1 Affymetrix blood datasets .................................................................134
Table A- 2 Affymetrix brain datasets .................................................................135
Table A- 3 Affymetrix liver datasets .................................................................136
Table A- 4 Affymetrix lung datasets .................................................................137
Table A- 5 Affymetrix skeletal muscle datasets .................................................138
Table A- 6 TAN attributes *ND = node degree ..................................................138
Table A- 7 TSN attributes *ND = node degree ....................................................139
Table A- 8 Pure attributes *ND = node degree ....................................................139
Table A- 9 GTEx binary networks attributes *ND = node degree .......................139
Table A- 10 Count of tissue-specific genes, FDR < 0.1, log transformed FC = 0 .......140
Table A- 11 Count of genes exclusively expressed in each of the tissues ...............140
Table A- 12 Count of expression-induced links ....................................................140
Table A- 13 Supplementary file content information ..........................................141
Table A- 14 Supplementary file content information ..........................................141
Table A- 15 Supplementary file content information ..........................................143

Table B- 1 Count of links and genes in each of the networks ...............................159
Table B- 2 Count of genes and links in snuc-RNAseq populations ........................160
List of Figures

Figure 1-1 Gene function, regulation, and coexpression ......................................................... 3
Figure 1-2 Coexpression analysis in forms of datasets ............................................................. 22
Figure 1-3 Guilt By Association framework ............................................................................. 27
Figure 1-4 Analyses and datasets ............................................................................................. 35
Figure 2-1 Different types of tissue-specific links ................................................................. 42
Figure 2-2 Overview of the network construction ................................................................... 46
Figure 2-3 Tissue Specificity Score of the links ...................................................................... 52
Figure 2-4 Identification of pure links ..................................................................................... 54
Figure 2-5 Reproducibility and functional relevance of the links .......................................... 56
Figure 3-1 Schematic of cellular composition effect ............................................................... 71
Figure 3-2 Much of the observed variance is explained by cellular composition ............... 75
Figure 3-3 Much coexpression is explained by cellular composition effects ....................... 79
Figure 3-4 Cellular composition can mask intra-cell-type coexpression ............................ 84
Figure 3-5 Coexpression of NRGN and CALM3 in Excitatory cell types ........................... 87
Figure 3-6 Reproducibility of robust snuc-RNAseq clusters in GTEx networks ............... 89

Figure A-1 Count of samples and genes for each tissue ......................................................... 128
Figure A-2 Proportion of links with low expressed genes .................................................... 129
Figure A-3 Comparison of the TSS and p-values from Wilcoxon-rank-sum test .......... 130
Figure A-4 Reproducibility of the TSN and TAN links in GTEx datasets ....................... 131
Figure A-5 Distribution of TOP for gene pairs ...................................................................... 132
Figure A- 6 Additional examples of pure links ...............................................................133

Figure B- 1 Comparison of two sets of marker genes ..................................................149
Figure B- 2 Distribution of R-squared values for marker genes.....................................150
Figure B- 3 Count of cells in each population of snuc-RNAseq ..................................151
Figure B- 4 Variance of the CT profiles versus the observed variance in bulk tissue......152
Figure B- 5 Mean expression level of genes in each cluster........................................154
Figure B- 6 Link overlap ratio ....................................................................................156
Figure B- 7 Reproducibility of GTExBulk clusters.......................................................157
Figure B- 8 Representation of the links in a snuc-RNAseq population.........................158
List of Supplementary Material

TSlinksTSS_blood_pureLinks_GTEx_exp.txt
TSlinksTSS_brain_cortex_pureLinks_GTEx_exp.txt
TSlinksTSS_liver_pureLinks_GTEx_exp.txt
TSlinksTSS_lung_pureLinks_GTEx_exp.txt
TSlinksTSS_skeletalMuscle_pureLinks_GTEx_exp.txt
blood_TSExclusiveDiseaseTerms_linkRemoval.txt
brain_TSExclusiveDiseaseTerms_linkRemoval.txt
liver_TSExclusiveDiseaseTerms_linkRemoval.txt
lung_TSExclusiveDiseaseTerms_linkRemoval.txt
skeletalMuscle_TSExclusiveDiseaseTerms_linkRemoval.txt
blood_TSExclusiveFunctionTerms_linkRemoval.txt
brain_TSExclusiveFunctionTerms_linkRemoval.txt
liver_TSExclusiveFunctionTerms_linkRemoval.txt
lung_TSExclusiveFunctionTerms_linkRemoval.txt
skeletalMuscle_TSExclusiveFunctionTerms_linkRemoval.txt
AdditionalFileB1.txt
AdditionalFileB2.txt
AdditionalFileB3.txt
AdditionalFileB4.txt
AdditionalFileB5.txt
AdditionalFileB6.txt
AdditionalFileB7.txt
### List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AD</td>
<td>Alzheimer disease</td>
</tr>
<tr>
<td>CPM</td>
<td>counts per million</td>
</tr>
<tr>
<td>CT</td>
<td>cell type</td>
</tr>
<tr>
<td>DAG</td>
<td>directed acyclic graph</td>
</tr>
<tr>
<td>DCA</td>
<td>differential coexpression analysis</td>
</tr>
<tr>
<td>DEA</td>
<td>differential expression analysis</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>DREAM</td>
<td>dialogue for reverse engineering assessments and methods</td>
</tr>
<tr>
<td>eQTL</td>
<td>expression quantitative trait loci</td>
</tr>
<tr>
<td>FDR</td>
<td>false discovery rate</td>
</tr>
<tr>
<td>GBA</td>
<td>guilt by association</td>
</tr>
<tr>
<td>GRN</td>
<td>gene regulatory network</td>
</tr>
<tr>
<td>GO</td>
<td>gene ontology</td>
</tr>
<tr>
<td>KEGG</td>
<td>Kyoto encyclopedia of genes and genomes</td>
</tr>
<tr>
<td>MI</td>
<td>mutual information</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger ribonucleic acid</td>
</tr>
<tr>
<td>PC</td>
<td>principal components</td>
</tr>
<tr>
<td>PCR</td>
<td>principal component regression</td>
</tr>
<tr>
<td>PPI</td>
<td>protein-protein interaction</td>
</tr>
<tr>
<td>PPIN</td>
<td>protein-protein interaction network</td>
</tr>
<tr>
<td>RMA</td>
<td>robust multi-array average</td>
</tr>
<tr>
<td>RNA</td>
<td>ribonucleic acid</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>----------</td>
<td>--------------------------------------</td>
</tr>
<tr>
<td>RNA-seq</td>
<td>RNA sequencing</td>
</tr>
<tr>
<td>snuc-RNAseq</td>
<td>single nucleus RNA-seq</td>
</tr>
<tr>
<td>SOM</td>
<td>self-organizing maps</td>
</tr>
<tr>
<td>SVA</td>
<td>surrogate variable analysis</td>
</tr>
<tr>
<td>SVM</td>
<td>support vector machines</td>
</tr>
<tr>
<td>TAN</td>
<td>tissue aggregated networks</td>
</tr>
<tr>
<td>TOP</td>
<td>topological overlap</td>
</tr>
<tr>
<td>TRN</td>
<td>transcription regulatory networks</td>
</tr>
<tr>
<td>TSN</td>
<td>tissue specific networks</td>
</tr>
<tr>
<td>TSS</td>
<td>tissue specificity score</td>
</tr>
<tr>
<td>WGCNA</td>
<td>weighted gene coexpression network analysis</td>
</tr>
</tbody>
</table>
Acknowledgements

First and foremost, I would like to thank my supervisor Paul Pavlidis whose guidance, encouragement, untiring support and patience was crucial throughout this journey. I would also like to thank my supervisory committee members, Alexandre Bouchard Côté, Joerg Gsponer and Wyeth Wasserman, whose support and advice have guided me throughout my studies.

I would like to thank GTEx consortium and Allen Brain Institute for having made their data available, without which this thesis would not have been possible.

Throughout my research, I have benefited from the support and friendship of many lab members at Pavlidis lab. I also have enjoyed participating in countless fruitful discussion, as well as learning about their various projects and inspiring ambitions. I am thankful to all these wonderful lab members and colleagues.

I would like to extend my gratitude to my family for their love and support. Specifically, my sister, brother, my parents and my grandmother for the colors, passion, laughter, wisdom and beauty they have added to my life. Finally, I would like to thank Soroush, for holding my hand and having walked with me through the various adventures of life and mind during my graduate studies.
To Azra and Nemat
Chapter 1: Introduction

Since the 1970s, the term *coexpression* has been used to refer to the coordinated presence of transcripts or proteins in a cell. A common assumption is that genes involved in the same cellular processes—or genes with similar functions—need to be coexpressed and therefore are likely to be coregulated. An early example of such relationships is the activation of *lac(tose)* operon in bacteria. The lac operon contains three enzymes required for digestion of lactose. In the absence of glucose, the increase of cAMP facilitates binding of the lac operon activator to the DNA, resulting in expression of the enzymes. These enzymes are involved in the same function (lactose metabolism) and are coregulated and coexpressed. Thus, “coexpression” was originally defined as a biological process reflecting common regulation.

With the emergence of high-throughput transcriptomic assays in the 1990s, the term “coexpression” was used in a way subtly separated from its biological basis, to describe an observation from the data, which I refer to as “observed coexpression”. In these studies, coexpression refers to the similarity of RNA expression patterns across samples. It was shown that observed coexpression tends to be associated with functional similarity among the genes. Findings from these studies established two prominent features for observed coexpression: reproducibility and biological relevance. Backed up by these features, coexpression analysis has become prevalent in transcriptomic studies, with the goal to discover functional or

---

1 Earliest Pubmed search results for articles with words co-expression or coexpression in the title or abstract.
regulatory relationships. In other words, coexpression was reframed as an observed signal from high-throughput data that could be used for inference.

The idea behind these high-throughput studies relates to the original coexpression concept, that genes involved in the same cellular process or sharing similar functions are likely to be expressed simultaneously. Subsequently, regulatory relationships are likely to be reflected in genes’ expression patterns. The fact that the observed coexpression is reproducible and represents functional similarity has led to the idea that observed coexpression is a reflection of transcript coexpression, which could be reverse engineered to infer functional or regulatory relationships (see Figure 1-1). However, this scenario has not worked out very well. Despite the abundance of methods and transcriptomic data, prediction of regulatory relationships from coexpression has remained a challenge. In part, this could be explained by the complexity of transcript regulation, the effect of post-transcriptional modifications and also some data driven factors (discussed in Sections 1.2 and 1.6). Yet, even considering such factors, there seems to be a gap in this scenario. On one hand, coexpression is clearly a reproducible and biologically relevant signal, which has led many to assume that it is a readout of regulation. On the other hand, its relevance to regulation (and its power for inferring regulation) should be questioned. My aim in this thesis is to address this gap by focusing on coexpression as an observation from the data and improving our understanding of its meaning and interpretation.
It is commonly assumed that the observed coexpression is a reflection of the simultaneous expression—transcript coexpression—resulted from regulation. Based on this hypothetical layout, the goal in coexpression analysis (dashed line) is to reverse engineer the functional and regulatory relationships among the genes from the observed coexpression.

In this first chapter, I review characteristics, key aspects and applications of observed coexpression in the literature. As explained, coexpression is tangled with *gene function* and *gene regulation*. In the first two sections of this chapter I review these two topics in relevance to coexpression analysis. In Section 1.3 I review the characteristics of transcriptomic data. Coexpression and its application in transcriptomic data analysis is reviewed in Section 1.4. In Sections 1.5 and 1.6, I review the application of coexpression in gene-function prediction and regulatory relationship inference. Section 1.7 has an overview of the research chapters in this thesis (Chapters 2 and 3).

Application of coexpression in transcriptomic data analysis has a history of more than 20 years. In order to reflect this history throughout my literature review articles are mostly cited with respect to the time of publication.
1.1 Codifying gene function

It has been shown that variations in gene coding sequences or their regulatory regions can be associated with disruption in cellular processes and disease development\textsuperscript{15–18}. Understanding the role of gene products in cellular processes is critical in explaining these associations and furthering our knowledge of cell biology. Broadly speaking, the role of gene product(s) in the cell are referred to as gene function. Classically, gene function is determined experimentally and perturbation studies are still relevant\textsuperscript{19,20}. The term \textit{functional genomics} describes the efforts to augment this knowledge by analyzing genomics data\textsuperscript{21,22}. Therefore in bioinformatics, there is a need to formally describe gene function for use in computational analyses, and there are multiple databases that record functional annotations. For example, the Kyoto Encyclopedia for Genes and Genomes (KEGG)\textsuperscript{23} contains information about biological pathways and functional modules. Online Mendelian Inheritance in Man (OMIM)\textsuperscript{24} has gene-disorder annotations. The Gene Ontology (GO)\textsuperscript{25–27} is one of the largest such resources and has two major parts, the ontology itself (GO) and gene-function annotations (GOA). The annotations are commonly used in functional genomics as part of coexpression analysis (detailed in 1.4 and 1.5) and numerous other applications. In Chapters 2 and 3, I use GO/GOA for network based functional enrichment analysis. Here I review GO’s structure and a few of its salient features.

In GO, functional terms are grouped into three categories: \textit{molecular function}, \textit{biological process} and \textit{cellular compartment}. Terms in each category are recorded in a data structure called a directed acyclic graph (DAG). Each node in the DAG represents a
Functional term, and edges are directed from more specific terms (child—closer to the leaf) to the more general ones (parent—closer to the root; see Figure 1-2). The acyclic feature means that there are no loops in this graph, therefore, a node cannot be both descendental and ancestral to any other node. There are different types of edges regarding the relationship they represent. Two of these relationships—*is-a* and *part-of*—indicate an inheritance: if a gene is annotated with a term, then it is also annotated with the term’s parents and ancestors through the *part-of* or *is-a* relationships. Therefore, terms on the nodes toward the leaf are likely to be more specific (associated with only a few genes) and terms on the nodes toward the root are more general (associated with larger group of genes). The DAG structure of the ontology allows for multilevel specification of functional terms. Therefore, genes could share the more general functional terms and also be annotated to their more specific terms.

Databases like GO/GOA are valuable for understanding cellular mechanisms and are employed in various ways for transcriptomic data analysis along with coexpression analysis. However, our knowledge about gene function is far from complete and these databases are constantly evolving and growing\textsuperscript{28–30} based on new biological findings, as well as computational methods for interpretation of transcriptomic data and inference of gene function. Many of these methods use coexpression as will be discussed in Section 1.5.

1.2 Regulation of transcription

Cellular processes are carried out through orchestrated interactions among thousands of gene products. Changes to the presence or abundance of these products could disrupt cellular
functions and result in disease development\textsuperscript{31,32}. It is clear that transcriptional regulation—as a major step of gene expression regulation—plays a critical role in cell’s development and function. A goal in systems biology is to understand this process and as a part of it, to discover the regulatory relationships among the genes. These relationships can help to understand the downstream effects of genetic variants\textsuperscript{33–36} and tissue development\textsuperscript{37,38}. However, identification of these relationships is challenging due to the complexity of transcript regulation.

Initiation of transcription requires the assembly of multiple protein complexes, facilitated by transcription activators and coactivators. This carnival of interacting proteins is specific to each gene and requires certain factors in place. One of the key players in this process are transcription factors (TFs). TFs are proteins with sequence-specific DNA binding coupled to a transcriptional regulatory activity. TFs contribute to the activation or repression of the transcription of their target genes through binding to their regulatory regions on the DNA (the specific binding site is called the cis-regulatory element of the target gene). Therefore, identification of the TF-target relationships is critical for deciphering the regulation of transcription.

A TF’s potential binding site to the DNA can be identified computationally based on its DNA binding \textit{motif}. A motif is a short DNA sequence pattern (often 6-12 bp) with a potentially regulatory role—for TFs, it is their preferred sequence for binding. For example, the human genome encodes about 1,600 TFs and about \( \frac{3}{4} \) of them have described binding motifs\textsuperscript{36}. It is possible to predict binding of a TF to the regulatory region of a gene based on
its motif, but such predictions suffer from extremely low specificity; and, determination of the TF’s regulatory role requires further biological experiments. For example, Cusanovich et al. examined the regulatory function of 29 TFs by monitoring the shift in the expression level of their predicted target genes (based on the binding site) after their knockdown in a cell line. They showed that 46.4–99.1% of the predicted TF bindings did not affect the expression level of the target genes.

One factor contributing to the low performance of these predictions is that TFs work in combination with other regulatory elements. While the potential binding site for each TF can be recognized—to some degree—by its binding affinities to the cis-regulatory elements of its targets, role of the coactivators, epigenetic regulators, as well as other collaborating TFs, are more complicated to account for. As discussed earlier, it is possible that some of the gene-gene regulatory relationships are reflected in coexpression and ideally, the role of coactivators as well as multiple TFs of a target gene could be identified through coexpression analysis. Nonetheless, this has remained an extremely complicated problem and there is little evidence for success in such efforts. Some of the challenges for inferring these relationships using coexpression as well as studies evaluating the methods for such analyses are discussed in Section 1.6.
1.3 Transcriptomic data

A central concept in my thesis is the use of transcriptomic profiling data. In this section I briefly review how such data are generated and prepared for application of coexpression methods (1.3.1), and some of its limitations (1.3.2).

1.3.1 Platforms and specifics

Each sample in a high-throughput transcriptomic dataset provides a snapshot of the transcriptome in the context under study. Transcriptomic datasets commonly used in coexpression analysis can be divided into three major categories regarding their technology: (1) microarray datasets, (2) mRNA-sequencing (mRNA-seq) datasets and (3) Single-cell mRNA-seq datasets. In microarray or mRNA-seq datasets, each sample is a collection of cells from a tissue, cell line or unicellular organism. For single-cell mRNA-seq datasets, each biological sample is an individual cell.

Single-cell data provides opportunity for exploring cell-type heterogeneity \(^{43-45}\) and discovering transcriptional changes during cell differentiation and development \(^{46,47}\). It also has specific challenges in sample preparation and data processing. Cell isolation is a critical step for sample preparation, for which multiple methods are available \(^{48}\). There are also several computational methods available for clustering the cells (samples) into different cell populations \(^{43,45}\) which are generally reproducible \(^{43}\). One issue in single-cell data is drop-outs, which refers to undetected transcripts (read counts of zero in the dataset) as a result of low amount of RNA from a single cell in the sample (only 10-20% of RNA is sequenced \(^{48}\)).
Drop-outs are a challenge for differential expression analysis (DEA) and coexpression analysis\textsuperscript{49,50}. However, it has been suggested that the coexpression observed in single-cell data is generally comparable with bulk tissue coexpression\textsuperscript{51}. In Chapter 3, coexpression networks from different single-cell populations are compared to the coexpression network built from the bulk tissue.

Between microarray and mRNA-seq datasets, the difference in the technologies leads to different pre-processing and quality control processes\textsuperscript{52–54}. Nonetheless, once processed, the resulting matrix of expression measurements for genes in different samples are comparable. The resulting expression matrix can also be treated similarly as inputs for various coexpression analysis. It has been shown that differential expression results are reproducible and coexpression networks are comparable among the two platforms\textsuperscript{55–59}. Results in Section 2.3.4 describe the reproducibility of tissue-specific links between microarray and RNA-seq datasets.

1.3.2 Noise and confounding factors in expression data

As a product of transcriptomic data and the underlying platform limitations, coexpression is affected by noise and technical artifacts such as batch effects\textsuperscript{60,61}. Confounding refers to unwanted or nuisance sources of variation that are correlated with experimental factors of interest. There are multiple methods for removing unwanted variation in expression data. Leek et al.\textsuperscript{62} model the hidden factors by capturing expression heterogeneity among the samples using surrogate variable analysis (SVA). Another method by Johnson et al.\textsuperscript{63} can
remove the known factors (such as batch-effect, where labels for batches are provided), but it performs well with small sample sizes and unbalanced designs.

It has been shown that correcting for extraneous sources of variation improves the statistical power for expression quantitative trait loci (eQTL) studies by increasing the number of significant cis-regulatory elements\textsuperscript{64,65}. The eQTL studies seek to explain expression variation of the genes based on the genetic variants (likely on the cis-regulatory element of the genes). The variation introduced by such factors could mask the underlying variation caused by genetic variants; therefore, removing the variation induced by confounding factors can result in finding more of the target associations. This is commonly done by removing principal components (PCs) from the data. For example, Ng et al.\textsuperscript{65} found that removing 10 PCs of the gene expression data was optimal for their eQTL study, in the sense that removing more PCs would not result in finding more significant associations. Similarly, Mostafavi et al.\textsuperscript{64} remove 30 PCs from the expression data.

Specifically, for coexpression analysis, factors such as variance in cellular composition and batch effects can induce correlations that are not of interest, or can even result in misinterpretations if not considered. The study in Chapter 2 is concerned with the confounding effect of mean expression level on observed coexpression when comparing between groups. In Chapter 3 I report on the effect of cellular composition on observed coexpression in brain tissue. In many coexpression studies, cellular composition acts like a confounding factor, since it affects the observed coexpression in a way that is hard to distinguish from the signal intended to be studied.
1.4 Coexpression analysis

As discussed, coexpression refers to the similarity of expression patterns among genes in high-throughput transcriptomic datasets. Numerous methods have been introduced and compared for measuring coexpression and applying it toward inferring functional or regulatory relationships among genes. The term coexpression analysis refers to any method or analysis that identifies or interprets coexpression.

Application of coexpression analysis has revolved around the basic principle discussed in the beginning of this chapter. Regarding the study type, it can be divided into three major categories:

1. **Analysis of individual transcriptomic data;** coexpression analysis is used to provide extra knowledge about genes under study—reviewed in 1.4.4 and 1.4.5.

2. **General approaches for function prediction;** it is used in the guilt-by-association framework as a form of association between the genes—reviewed in 1.5.

3. **General approaches for discovering regulatory relationships.** Reviewed in 1.6.

In the following sections I will review two fundamental concepts in coexpression analysis: gene clustering and coexpression networks. I then review the state of knowledge on the evolutionary conservation and biological relevance of coexpression (1.4.3) and after that I will discuss the application of coexpression in transcriptomic data analysis.
1.4.1 Gene clustering in transcriptomic data

Clustering methods provide a way to capture and represent groups of genes with similar expression patterns. In the early days of high-throughput transcriptomic data, it was demonstrated through gene clustering analysis that “some genes share their expression patterns, and it is likely that genes with similar expression patterns share function.” In this part I will review some of the key studies that established and reinforced this idea.

The first study of transcript clustering at a reasonably large scale was the 1998 work of Wen et al. 66, who measured the expression levels of 122 genes during nine developmental stages in rat. One of their aims was to determine if there were specific expression patterns that represented functional similarity between the genes. They were able to recognize four major patterns, and some of these patterns were shared among groups of genes with similar functions.

Wen’s work was followed by many studies using clustering and showing that genes with similar expression patterns had shared functions 2–5,67. For example, Eisen et al. 2 identified some well-represented functional groups in yeast time series and human fibroblast datasets. Other studies in yeast documented groups of genes with oscillating expression patterns through the cell cycle 68,69. Coexpression was also quickly applied to inferring regulatory relationships. Spellman et al. 5 and Tavazoie et al. 4 used clustering on yeast time series expression data combined with data from promoter sequences of the genes to identify genes with common regulatory factors relating to the cell cycle. Based on this analysis,
Spellman et al.\textsuperscript{5} showed that some of the genes in the same cluster were likely to be coregulated.

As clustering became a common part of transcriptomic data analysis\textsuperscript{70}, numerous clustering methods, such as k-means, hierarchical, self-organizing maps (SOM) as well as biclustering methods were proposed\textsuperscript{3,71–73}. A necessary step for clustering is to choose a distance or similarity measure for gene pairs (i.e., defining the distance of two genes based on some function of their expression across various samples). As clustering methods became more sophisticated, various measurements were applied, among them are Euclidean distance, Mutual Information and Pearson correlation\textsuperscript{54,71,74,75}.

Through clustering analysis, it became apparent that genes follow different patterns of expression, also that these patterns are likely to represent some underlying biological events, implied by the functional similarity of genes in the same cluster. It is important to notice that grouping of genes based on a clustering algorithm is an observation from the data; therefore, there is no “biological validation” or “ground truth” for how the clusters should form. As a result, choosing a gold standard for the method or similarity measure is dependent on the application—as will be discussed later. Nonetheless, clustering provided context for genes (through groupings), as well as a data structure that illustrated the similarity of their expression patterns and could be compared between studies. Soon afterwards, clustering was often replaced with the concept of a coexpression network, which was deemed to be more convenient or interpretable, as detailed in the next section.
1.4.2 Coexpression networks

Another way of representing the similarity of expression patterns among genes is as a network. Instead of looking at groups of genes with similar expression patterns, coexpression networks represent these similarities at a gene-pair level, where genes are the nodes and the coexpression associations or links are the edges. Networks could be created by simply adding edges based on a threshold on the distance/similarity measure. In this section I will review methods for building coexpression networks and discuss their applications.

1.4.2.1 Similarity measure

Similar to gene clustering, a fundamental step in building coexpression networks is choosing the similarity measures between the genes. There is no single best option for the similarity measure and various methods are applied based on different reasons. For example, Pearson correlation is commonly used\(^{70,76,77}\), but it fails to capture nonlinear relationships\(^1^0\), and for that reason mutual information (MI) is sometimes used\(^70\). However, a comparison between MI and Pearson has shown similar results regarding the obtained links\(^78\). In theory, partial correlation can capture causal relationships and is sometimes proposed for inferring direct regulatory relationships\(^79–81\). It is also common to use hard or soft thresholding for the measurements to obtain binary or weighted networks\(^82–84\). In their popular method for building coexpression networks Weighted Gene Coexpression Network Analysis (WGCNA), Zhang and Horvath\(^82\) use soft thresholding to build the coexpression network and use
topological overlap (TOP) as the similarity measure for network clustering. TOP measures the similarity as the overlap of shared neighbors in the network.

Also similar to gene clustering, a coexpression network is a structure built from the data, therefore evaluation of the similarity measure or the network structure is a challenge due to the fact that there is no ground truth for which gene pairs should be coexpressed. In some cases the similarity measures can be evaluated based on their usefulness in solving a specific problem. For example, Kumari et al.\textsuperscript{11} evaluated the performance of some similarity measures for representation of different functional relationships using the GO and KEGG databases. They reported that some methods are more efficient in predicting pathway genes, whereas some are more efficient in predicting genes for common regulatory factors. It is critical to note that the relevance of a network evaluated or built based on its usability is strictly limited to the intended application.

### 1.4.2.2 Network characteristics and interpretation

The network representation has been used in numerous coexpression analyses. Several studies have examined the biological significance and interpretation of the topological elements in coexpression networks. Some of the most commonly discussed elements in coexpression networks are clusters and hubs. It is also common to observe, or just as commonly, assume, that coexpression networks have a scale-free topology. Here I review some studies relevant to these topics and their implications in coexpression analysis.
It is common to identify gene clusters using network clustering algorithms; clusters are also referred to as subnetworks, subgraphs and coexpression modules\(^{85-89}\). Similar to clustering, it is shown that network modules are likely to be reproducible and enriched with specific functional terms, as in the original concept of clustering expression profiles\(^{90}\). Apart from module identification, some studies focus on comparing modules, coexpression links or other topological elements between the networks to find context-specific characteristics—contexts such as disease, tissues or species\(^{87,88,90-93}\). These studies are known as differential coexpression analysis (DCA), which is discussed in detail in 1.4 and is the main topic of Chapter 2.

Regarding network topology, it has been shown that coexpression networks, like protein-protein interaction networks (PPINs)—a network where proteins are the nodes and they share a link if they interact—tend to have a scale-free topology\(^{94-96}\), meaning that the distribution of node degree follows a power law. In contrast, the WGCNA method enforces scale-freeness as part of the algorithm\(^{82}\). However, there is no evidence suggesting that this assumption results in building a “more accurate” network, since as mentioned before, there is no “true” coexpression network to use as a gold standard.

A popular topological element in coexpression networks is hubs. Hubs refer to genes with a high count of links (high node degree) relative to other genes. Coexpression networks are likely to have a small portion of their nodes identified as hubs, while most of the other nodes have much smaller node degree; that is, the node degree distribution is highly skewed. It is common to examine the hubs for their biological relevance. For example, Carlson et al.\(^{97}\)
show that the number of links for a gene is correlated with its biological essentiality and its evolutionary conservation. As another example, Yang et al.\textsuperscript{85} show that cancer prognostic genes are significantly depleted in hub genes of the network. They also show that this depletion is specific to the cancer, meaning that prognostic genes of one cancer are not significantly depleted in the hub genes of another cancer. While in these studies hubs seem to carry a biologically relevant signal, it is difficult to explain the underlying biological events resulting in such observation. In the absence of a clear interpretation, results from such analyses remain mere observations with no biological or analytical implications.

Representation of coexpression in a network has facilitated the application of many network-based concepts, methods and algorithms. This representation fits with our understanding on how gene products function or interact with each other, like in the PPIN or gene regulatory network (GRN). It also seems more convenient compared to results from a clustering algorithm that provides a mere grouping of the genes with no overlap. In short, the network representation makes it easier for coexpression to be interpreted as a reflection of the PPIs or regulatory relationships. However, for the same reason, this representation is problematic. As mentioned before, coexpression links, unlike the links in GRN or PPIN, have no direct biological interpretation. Coexpression is an observation from the data and although it is more likely for two coexpressed genes to share a protein interaction\textsuperscript{98–101} or a function\textsuperscript{102}, these instances are relatively rare and far from being able to explain the observed coexpression\textsuperscript{103}. Nonetheless, it is common practice in coexpression analysis to report a statistically significant signal with some biological relevance, without providing any
explanation or examination of the underlying biological events likely to have caused the
signal. Unfortunately, without such explanation, these findings remain useless or could even
result in misinterpretation. In the two research projects in Chapter 2 and 3, my aim is to shed
light on the underlying causes of the observed signal in coexpression analysis.

1.4.3 Reproducibility and conservation of coexpression

Coexpression has been explored for both evolutionary conservation \(^{104-106}\) and reproducibility
\(^7,90\). It has been shown that conserved coexpression links and links that are present in multiple
datasets are more likely to represent functional similarities among genes.

Bergmann et al.\(^ {104}\) compared coexpression networks among six different species \(^{ii}\) and
reported the common and specific coexpression modules between the networks. McCarroll et
al.\(^ {106}\) looked for the coexpression of orthologous genes across two species \(^{iii}\) in different
stages of ageing and found highly conserved patterns of expression for fourteen functional
categories. Stuart et al.\(^ {105}\) examined the representation of functional relationships for the
evolutionary conserved coexpression links by building an aggregated coexpression network
among four species \(^{iv}\) and predicted functions for some genes based on the Guilt By
Association principle (discussed in 1.5). Stuart et al.’s work particularly stands out as some of
their predictions were shown to be true based on the subsequent experiments. They had
hypothesized that an aggregated coexpression network from multiple species might be a better

\(^{i}\) *S. cerevisiae, C. elegan, E. coli, A. thaliana, D. melanogaster and H. sapiens*

\(^{ii}\) *Drosophila and C. elegan*

\(^{iii}\) *Human, flies, yeast and worms*
platform for function prediction and that the conserved links might represent functional similarity between the genes for some of the essential functions in the cell. Finally, Lee et al. 7 built an aggregated coexpression network from 60 microarray datasets from human and kept the links that were strongly present in three or more datasets. They reported that the robust coexpression links are more likely to represent functional relevance among the genes. In summary, coexpression of some genes is both reproducible across studies, as well as conserved in evolution. These observations have helped bolster the idea that coexpression is biologically meaningful, but as I discuss, the nature of that biology is not directly readable from these observations.

1.4.4  Coexpression in expression data analysis: overview

In a transcriptomic study, samples are typically collected according to two common experimental designs: time-series and factorial. For a time-series dataset, samples are collected at successive time points (Figure 1-3A). An example of a time-series dataset is the yeast samples collected at periodic time points during the cell cycle 107. In a factorial design, samples are collected in groups, from different conditions, states or treatments (each represent a factor); for example, samples from diseased versus healthy individuals (control) Figure1-3B.

Having either of these experimental designs, the primary goal is to find genes that are of interest relative to the design, and ultimately, identify their role in different cellular processes. What makes a gene interesting in a time-series dataset could be a particular pattern of expression. An example is the oscillation of histone gene expression during the yeast cell
cycle. In factorial designs, it could be a statistically significant shift in the average expression level of the gene between groups of samples. The shift is commonly identified by DEA and might indicate that the gene is causal to or affected by a factor. Once the set of differentially expressed genes is identified, functional enrichment analysis is usually used to summarize processes affected by or involved in the condition.

This basic workflow of DEA followed by enrichment analysis is often enhanced by the application of coexpression analysis. The idea is that coexpression can identify genes with similar expression patterns while other data sources can help to interpret the observed patterns. In time-series datasets, the external information about different time states plays a critical role in explaining the observed coexpression. For example, information about the cell cycle and role of histones in it is critical in explaining oscillation of their expression levels during the cell cycle in yeast.

In factorial datasets, we do not have access to the sequential time points/states and instead we have snapshots of the expression levels from a particular condition. In these studies, it is common to perform a differential coexpression analysis (DCA), in addition to a DEA. As mentioned in 1.4.2, this includes methods for identification of context-specific modules, links or other topological elements between the networks.

It is important to notice the difference between information obtained from DCA and DEA. In theory, DCA can be performed as a complementary tool to DEA and (it is hoped) reveal the underlying regulatory relationships among the genes. However, differential coexpression can happen both in the absence and presence of differential expression. For
example, it has been observed that tissue-specific modules in coexpression networks built from different tissues are enriched with tissue-specific genes $^{90,91}$. That is, much of the observed differential coexpression is between the genes that are not expressed in the other sample groups. In these studies, the expression level acts as a confounding factor. Nonetheless, in most DCAs, the assumption is that the observed differential coexpression is happening in the absence of differential expression and as a result the DCA is implemented and discussed independently of the DEA results. This is problematic for the interpretation of DCA. For the extreme cases where differential coexpression is observed for the genes that are specific to a context, interpretation of the observed differential coexpression as a reflection of the underlying regulatory rewiring is incorrect. For the cases where there is a change in the average expression levels, differential expression acts as a covariate and therefore needs to be considered as such. In Chapter 2, I examine the role of average expression level on observed differential coexpression among five human tissues.
Figure 1-2 Coexpression analysis in forms of datasets
(A) A time series dataset where samples are obtained in sequential time points. (B) Expression patterns for 12 genes in a hypothetical dataset. Genes are clustered based on their expression patterns. (C) Coexpression links for the genes, each dot represents a gene. The top and bottom groups (plots 1 and 4) have distinguishing expression patterns from all other groups. The two middle groups (plots 2 and 3) have similar expression patterns, and therefore the genes are connected between them, with weaker associations (lighter lines). (D) A factorial expression dataset. For each condition several samples are obtained. (E) Each circle represents groups of genes expressed in one condition. The expressed genes have some overlap between the conditions. The overlap is usually the larger proportion of the genes (figure is not accurate). (F) Each figure shows coexpression links in one condition. Each black dot represents a gene. While there are multiple differentially coexpressed links (those that are present in one condition and not the others), there are only two differentially coexpressed links between the genes that are expressed in all the conditions (bright red color for links in C1 and C2).

1.4.5 Coexpression in expression data analysis: applications

Different forms of coexpression-based analysis have been implemented in gene expression studies to gain insight about the genes. Many of these studies have done some form of DCA\textsuperscript{113–116}. Some performed DCA in relevance to the DEA results. In this part I will review examples of such studies.

Torkamani et al.\textsuperscript{64} did a network-based analysis for a group of 101 case and control samples for schizophrenia. Networks were built using MI and WGCNA\textsuperscript{82}. They compared modules in networks from control and schizophrenia cases and reported a module that is not shared between the two networks. They also did a network-based enrichment analysis where
they checked for the modules in the network that were enriched within the differentially expressed genes in schizophrenia versus controls.

As another example, Ray and Zhang \textsuperscript{114}, compared coexpression networks from four brain regions in Alzheimer’s disease (AD). Networks were built using Pearson correlation and thresholding \textsuperscript{76}. They compared network topologies for AD differentially expressed genes and examined the biological differences that the links represented. They reported many AD differentially expressed genes having differences in their links.

Identification of different coexpression modules is one of the most commonly applied methods in DCA. For example, He et al. \textsuperscript{115} compared coexpression networks from hepatitis B and hepatitis C infection to identify different functional modules. Aggarwal et al. \textsuperscript{117} built an aggregated network from gastric cancer tissue from four populations. They report relevant functional modules and coexpression links. McDermott-Roe et al. \textsuperscript{118} studied cardiac hypertrophy and were able to identify EndoG as a determinant gene in the trait based on the change in its coexpression patterns.

It is likely that in most of these studies, modules are enriched with differentially expressed genes that represent the confounding role of differential expression in the observed differential coexpression, as studied in Chapter 2. An example of this is a study by Gu et al. \textsuperscript{119}. They identified the “responsive gene modules”, as modules highly enriched with differentially expressed genes, in inflammation and angiogenesis in a human cell line. Formation of coexpression modules could also be a result of cellular composition variation among the samples, which is the topic of study in Chapter 3. An example for such findings is
the study by Mabbott and Gray. They compared coexpression links in three major B-cell lines in mice and identified modules enriched with cell-type specific genes in each of the networks.

These studies employ different methods for building the networks and comparing them, but the recurrent element in all of them is the identification of coexpression modules. This relates to the fact that coexpression links are hard to interpret by themselves and that identification of robust coexpression links requires multiple datasets. However, as it was mentioned, the identified context-specific modules are likely to represent the confounding role of differential expression, or the effect of cellular composition variation among the samples. Therefore, in spite of being valuable for providing functional context for groups of genes under study, the identified differential coexpression modules are unlikely to represent regulatory rewiring among the genes.

1.5 Coexpression for function prediction

Coexpression has been used in many gene-function prediction studies based on the Guilt By Association (GBA) principle. The GBA principle states that genes with similar attributes (e.g., functions, associated disease terms) are likely to have similar characteristics in different sources of data (e.g., similar expression patterns) and vice versa; that is, the genes are suspected of having an attribute due to their association. The enrichment of functional terms in coexpression clusters is an illustration of this principle, where “expression pattern” is the shared characteristic.
The GBA function-prediction framework takes a set of *query genes* and *training genes* as well as gene associations—this could be a coexpression network or PPIN. Attributes for the query genes are then predicted based on their association to the training genes (see Figure 1-4). This framework is suitable for high-throughput gene-function prediction, but the GBA principle is also the basis for coexpression analysis in many other studies. McDermott-Roe et al.’s study mentioned earlier is a clear example of this. They were able to identify the role of EndoG in hypertrophy based on its association to the mitochondrial genes in the altered coexpression network.

Understanding the application of GBA and its evolution along coexpression analysis helps to illuminate the characteristics, potentials and shortcomings of coexpression. Furthermore, the performance of the high-throughput GBA-based function-prediction could be used as a measurement for the quality of coexpression network.
The inputs are a set of query genes and a set of training genes, as well as the association between them. Based on the association between the two groups of genes, functional terms are assigned to the query gene set.

1.5.1 Methods and frameworks

Solving a GBA function prediction problem has two steps. First is to build the matrix of gene associations and second is to infer functional terms for the query genes (see Figure 1-4). Gene pair coexpression measures as well as PPIs and (in some cases) a combination of different gene similarity matrices are used to build the association matrix.
Methods for inferring functional associations include gene expression clustering \(^{93,121}\) and graph clustering algorithms \(^{82,86,122}\). In this approach, the function of query genes could be inferred from the functions enriched in the cluster they are grouped in. Another approach is label propagation in the network \(^{83,123,124}\). Therefore, association of genes to a specific function is determined by the relative association of the genes in its neighborhood (i.e., direct or indirect links could both be counted) to that function. As a result, each gene obtains a rank for the function under study and is prioritized as a candidate for having an association with the function. It is shown that construction and trimming of the networks in these methods could affect the performance of GBA \(^{83,123}\).

1.5.2 Evaluation and shortcomings

Although the GBA principle is usually referred to in large-scale function prediction studies, the idea was implicit in the majority of coexpression analysis methods from its early days. Perhaps the most distinguishing feature among these studies is evaluation techniques and validation of the findings.

In 1999, Walker et al. \(^{125}\) used coexpression in a GBA framework to identify genes associated with prostate cancer. They predicted eight genes that were novel at the time. However, there was no evaluation of their prediction method. Brown et al. \(^{126}\) used a supervised classifier, support vector machine (SVM), to predict functions for genes and provided statistics for the evaluation of the predictions from coexpression. The novelty of their work was in using supervised learners for GBA implementation and also studying the
performance of the classifier. They also discussed two important shortcomings of GBA, “lack of generality” and “being limited to the at-the-time gene function annotations”. These two shortcomings are critical limitations for the application of GBA and have been largely ignored in later studies.

Wolfe et al. 127 also evaluated the performance of GBA in predicting GO 26 terms. Briefly, they built five gene coexpression networks from mouse, rat, human (two networks from different expression platforms) and a multi-species network by combining the four networks. From each network, they obtained a pairwise value between every gene $g_i$ and every GO term $t_j$, which showed how well $t_j$ is represented by the neighbors of $g_i$ in the network. These values were used to predict functions for $g_i$. Although on average they got good performance (the Area Under the Receiver Operating Curve [AUC ROC] values for some GO categories were around 0.7), for many functions the AUC was below 0.6. Considering that the AUC of 0.5 corresponds to random prediction, we can conclude that GBA is not generally applicable. Nevertheless, the study by Wolfe et al. 127 was influential since it provided a large-scale evaluation of the applicability of GBA.

As it was mentioned in Section 1.4.3, conserved and reproducible coexpression links are more likely to represent functional similarity among the genes. There have also been studies using coexpression from aggregated datasets for GBA for gene function prediction. Guan et al. 128 combined coexpression links from different datasets with tissue-specific expression data to build integrated tissue-specific coexpression networks. They predicted genes associated with male fertility and spermatogenesis phenotypes and one of their top
predictions was also experimentally confirmed. Nayak et al.\textsuperscript{88} built an integrated coexpression network from three human B-cell populations. They predicted functions for more than 50% (36 out of 60) of the poorly characterized genes under study. Some of the predictions were tested in biological experiments later and found to be true.

Coexpression has also been used in a combination of other association matrices to increase performance of function prediction. For example, Guan et al.\textsuperscript{129} combined coexpression networks from different tissues with PPIN and disease association networks to build a network of functional relevance for mouse. Goh et al.\textsuperscript{130} showed that genes associated with the same disease or genes that share protein interactions are more likely to be coexpressed and share similar functions. Some studies showed that the function prediction results improved using aggregated coexpression networks\textsuperscript{123,131–133}. Measuring the wisdom of crowds (both method-wise and data-wise), Verleyen et al.\textsuperscript{132} reported that a system of integrated data sources generates more improved results compared to a system of integrated methods. This might indicate that the bottleneck for GBA studies is not the methods, but rather the quality and limitations of the data.

Apart from the lack of generality and limitations in evaluation, most of the GBA approaches have been shown to suffer from a bias known as multifunctionality\textsuperscript{133}. Gillis and Pavlidis\textsuperscript{133} showed that a single gene list in which genes are ranked by the count of functions they are annotated with can get high average prediction performance for functional terms. Clearly, this is because the genes on top of this list are “true positive” for the many functions they are annotated with. They called these genes “multifunctional genes”. They showed that
the count of functional terms for a gene is correlated with its node degree in biological networks, meaning that multifunctional genes are also likely to have high node degree in the network. As a result, these genes are likely to be ranked higher in the label propagation algorithms in the network for the many functions their neighbors are enriched with. Because the node degree of a gene in many networks is at least partly an artifact of how much the gene is studied, the multifunctionality bias represents lack of specificity in network-based gene-function prediction studies and demonstrates that unlike what is suggested by performance measurements, these methods are not successful in extending our knowledge about gene function.

1.6 Inferring regulatory relationships from transcriptomic data

In 1.3, I reviewed gene regulation and the role of TFs. TF-target relationships are the most popular form of regulatory relationships sought for in coexpression analysis. Here I review some of the challenges for inferring these relationships, as well as a few articles that focus on evaluation of the methods and findings.

1.6.1 Challenges and complexities

Three forms of regulatory relationships can be represented in transcriptomic data and thus, in principle, inferred: TF-target relationships; coactivators and their targets (regulatory proteins that do not directly bind to the cis-regulatory elements of the target); and coregulation (e.g., the set of targets of a TF, also referred to as a regulon). In a coexpression
network, the first two forms might be represented as individual links and regulons could be identified as densely connected clusters. Inferring these relationships from transcriptomic data (via coexpression) entails three major challenges. The first challenge is data-driven limitations such as lack of perturbation experiments and small sample counts compared to the genes\textsuperscript{10,134}. The second is the complexity of gene regulation, specifically transcript regulation and the role of epigenetic\textsuperscript{41} as well as the post-transcriptional regulation, and third is the lack of gold standards that limits the application of methods, specifically the supervised learning methods\textsuperscript{134}.

As mentioned, one issue with the data is the limited size of samples compared to the count of genes. A gene is regulated by multiple TFs, transcription activators and coactivators that work conjointly to activate or repress its transcription. Identification of these regulatory factors among thousands of genes is underspecified due to the small count of samples. This problem is commonly referred to as the curse of dimensionality\textsuperscript{10}, and some studies have tried to bypass this problem by limiting the potential regulators to the TFs that are likely to bind to the promoter region of the targets\textsuperscript{134,135}. The other issue is the dynamic nature of regulatory relationships versus the static nature of transcriptomic samples. It is unlikely that, in the absence of perturbations, variation among samples from the same context represent much of the regulatory dynamics. Also, even with perturbations, samples represent only snapshots of the transcriptome, and it is likely that the inter-state dynamics remain undiscovered. Based on these factors, an ideal transcriptomic dataset for inferring regulatory relationships in an organism is one with thousands of samples from multiple perturbation
experiments, and also from various contexts. Today, studies using coexpression to infer regulatory relationships attempt to do so from data orders of magnitude less rich than this ideal.

The role of post-transcriptional regulation is also important. Research has shown little correlation between protein and RNA levels. This is because the quantity of proteins in a cell is the result of fine tunings in the different steps of gene expression regulation, and RNA is the product of the first step—transcription regulation—in this process. Furthermore, most TFs are primarily regulated post-transcriptionally and also post-translationally, as they need to be activated. Therefore, the ideal form of correlation for inferring a TF-target relationship is the correlation of the TF’s active protein level with the RNA level of its target. Again, this is a critical limitation of the resolution of analyses based purely on transcript-level information. These issues are less likely to affect the identification of regulons and are more likely to affect the identification of direct relationships.

1.6.2 Inference and validation

Among the three regulatory relationships mentioned in the previous section, most studies have focused on inferring TF-target relationships. Multiple review articles have focused on the mathematical properties of the methods for obtaining coexpression links with regard to regulatory applications. I briefly reviewed these methods in 1.4.2. Here, I review a few articles that examined coexpression for inferring regulatory relationships.
The Dialogue for Reverse Engineering Assessments and Methods (DREAM) project evaluated methods for inferring GRN in yeast, bacteria and an *in silico* benchmark. Methods were grouped based on their core strategies for link identification, including regression, MI, correlation and Bayesian networks. Identified links were tested against curated gold standards (for *E. coli*) and high confidence interactions from Chip-seq and binding prediction based on evolutionary conserved motifs (for yeast). The methods had varying performance in each category, while the combined methods had higher performance than the individual methods. In conclusion, performance for yeast was extremely weak (with area under precision-recall curve < 3%), but they had ~40% precision for the newly identified links in *E. coli* (23/53). Previously, Filkov et al. had reported less than 20% sensitivity for predicting yeast regulatory relationships using time series data.

It is commonly believed that causal regulatory relationships are represented in coexpression network, and the major challenge for regulatory inference is to distinguish between false and true positives. To examine this in mammalian data, Djordjevic et al. built coexpression networks from mouse heart and tooth developmental datasets. For their gold standard, they curated 1,518 regulatory relationships for the tooth and 710 for the heart. The correlation values were slightly different for activating and inhibitory regulatory relationships for the heart (higher and lower than the background, respectively), but they were not different for the tooth network. Their conclusion was that due to the lack of precision, coexpression is unlikely to be reliable for inferring regulatory relationships.
1.7 Research contribution

Below I briefly review my objectives and findings from the two research chapters. In the first research project laid out in Chapter 2, I studied differential coexpression between five human tissues. My first goal was to illuminate the resolution of coexpression as an observed signal from the data by quantifying robust differential coexpression between five tissues (blood, brain, liver, lung and skeletal muscle). The second goal was to examine the limits of coexpression for inferring regulatory relationships by identifying robust differential coexpression links among the tissues that were not confounded by differential expression—and therefore could be interpreted as “regulatory rewiring”. This project has two parts, corresponding to these goals.

Figure 1-4 Analyses and datasets

The diagram shows the aims and the main datasets used in the two research chapters of this thesis. Datasets used for each aim are marked.
In the first part, I identified robust tissue-specific coexpression links and showed that they are reproducible in external datasets. Using several datasets, I first built an aggregated coexpression network for each of the five tissues using the microarray datasets. Then, I identified tissue-specific links by comparing coexpression values between the tissues. I showed that the identified tissue-specific links are reproducible in an external dataset (the GTEx datasets).

In the second part of Chapter 2, I modeled the observed coexpression in each of the original datasets with the average expression level of the genes. Based on this model, I was able to show that much of the observed differential coexpression between the five tissues is accompanied by differential expression. This finding demonstrates the limited potential of the observed differential coexpression for representing regulatory rewiring. It also shows that unlike the common belief, differential coexpression rarely happens in the absence of differential expression.

In Chapter 3, I studied the role of cellular composition on coexpression analysis in the brain. My goal was to examine the prevalence of this effect and its implications in the coexpression analysis. This study had three main objectives and is laid out in three parts, each addressing one of the objectives. The first objective was to examine the effect of cellular composition on the observed variance. The second was to examine its role in the observed coexpression, as well as its contribution to its reproducibility and biological relevance. The third objective was to examine a potential method for correcting this signal. In this study, I used two main datasets: a bulk brain dataset (the GTEx dataset, from human brain cortical
tissue) and a high-quality single-nucleus dataset from the brain (Allen institute’s single-nucleus dataset from human brain cortical tissue). Samples (cells) from the single-nucleus dataset were labeled with their identified cell types by the Allen Institute. I was able to estimate genes’ cell-type expression profiles (i.e., the average expression level of genes in each cell type) from the single-nucleus dataset and also build cell-type coexpression networks.

In part one, I modeled gene expression variation in the bulk tissue based on the variation of cell-type marker genes. The results from this model demonstrate that gene variation in the bulk tissue is highly affected by cellular composition. I also showed that for a given gene, its variance in the bulk tissue is correlated with the variance of its expression level between different cell types (obtained from the single-nucleus data). In part two, I investigated the effect of cellular composition on the observed coexpression. Using simulation, I first showed that in a bulk tissue dataset with varying cellular composition among its samples, genes with similar cell-type expression profiles appear to be coexpressed. I then built a coexpression network from the bulk tissue and showed that genes from the same network cluster have similar cell-type expression profiles. I showed that bulk tissue coexpression clusters are highly reproducible between brain coexpression networks and are also enriched with cell-type marker genes and cell-type-specific functional terms. In part three, I addressed two questions: if the cellular composition induced coexpression masks the underlying robust cell-type specific coexpression and if we can correct for it. To answer the first question, I examined the representation of robust coexpression clusters from the single-nucleus data in the bulk tissue. My results demonstrated how the signal is affected in the bulk
tissue. For the second question, I showed that upon using a simple correction method, some of the underlying single-cell coexpression signals can be retrieved.
Chapter 2: Differential Coexpression in Human Tissues

2.1 Introduction

Coexpression networks have been adopted as a convenient representation of the pairwise similarities between gene RNA levels in transcriptomic datasets. A key feature of coexpression is that pairs of coexpressed genes have a tendency to be functionally related. For this reason coexpression networks have been widely used in computational function prediction frameworks, based on the Guilt By Association (GBA) principle and various types of coexpression analysis are prominent features of many transcriptomic studies.

An extension of coexpression analysis is Differential Coexpression Analysis (DCA), referring to changes in coexpression between different conditions. DCA is commonly used as a complementary tool to Differential Expression Analysis (DEA), where it is applied in an effort to reveal changes in regulatory “wiring” between genes which is not otherwise captured through DEA. An issue central to this study is that differential coexpression and differential expression are often confounded. At one extreme, two genes cannot be coexpressed if one (or both) of the genes is not expressed at all and some observed differential coexpression is simply explained by the absence of expression (see Figure 2-1, “expression-induced” links). Also in general the variance of expression is frequently a function of expression level. Examining differential coexpression in the absence of differential expression is also appealing for reasons of parsimony: if two genes are found to be differentially expressed, it is not necessarily clear what new information is gained if one (or
both) of them are also differentially coexpressed. For these reasons, it is of interest to investigate the degree to which differential coexpression is explained by differential expression, and to isolate cases where differential coexpression is not accompanied by changes in mean expression level. The latter “pure differential coexpression” (Figure 2-1) would be good candidates for instances of “rewiring”, though even then the interpretation is not necessarily straightforward.

Here, I study differential coexpression in different human tissues with a specific goal of accounting for the impact of changes in mean expression levels, and to thereby identify “pure” differential coexpression between tissues. Tissue based coexpression networks have been previously used to improve GBA-based prediction for tissue related disease and functions \(^{91,128,149}\). They have also been used in integrated or individual frameworks for studying tissue specific gene regulation \(^{150,151}\). I hypothesized that different tissues should be a relatively rich source of cases of “network rewiring”, due to their wide biological differences, and establishment of the principles would open the door to application in other scenarios such as different disease conditions. To this end, I built coexpression networks for five human tissues by combining coexpression networks from multiple datasets (Tissue Aggregated Networks, TANs). I identified tissue-specific links in the TANs, forming Tissue Specific Networks (TSNs; Figure 2-2). I then modeled how much of the observed differential coexpression between tissues (tissue-specific links) is predictable by the mean expression level of the genes. Based on this, I identified pure links as a subset of TSN links for each tissue. I found that the average expression level of genes is indeed a strong confound for much
tissue-specific coexpression and that pure links are rare. I show that my TAN, TSN and pure links are generally reproducible in external datasets, while pure links are reproduced to a lesser extent. I further searched for the biological significance of the pure links and compared them to the expression-induced links regarding their functional implications in TSNs.
Figure 2-1 Different types of tissue-specific links

(A) Schematic representation of the effect of the average expression levels on differential coexpression, in three extreme cases. In reality links lie on a continuum among these classes. (B) Examples of the link classes from the data, with brain being the target tissue.

Corr: correlation. Exp: Expression levels. Gene symbols are indicated in the legend at right.

Top panel: “pure” differential coexpression between TOMM20 and SLC25A44. The two genes have moderate to high expression levels in all the tissues, but correlation values in
brain are higher. Middle panel: expression-induced 1: AMER2 is only expressed in brain. Bottom panel: expression-induced 2: both AMPH and AP3B2 are only expressed in brain.

2.2 Methods

2.2.1 Datasets

I used two collections of expression datasets to build and examine the reproducibility of the networks, a collection of Affymetrix microarray datasets and RNA-seq data from GTEx. The TANs and TSNs were built from Affymetrix microarray datasets. The pure links are also identified based on the Affymetrix datasets and GTEx data was used for validation of these three groups of links. The Affymetrix datasets consist of 53 datasets that used Affymetrix U133 Plus 2.0 GeneChips, reanalyzed from raw data (see Tables A-1 to A-5, Figure A-1). The final data matrix included 18,494 genes and 3,563 samples. For each tissue, genes were filtered based on their expression level and only genes which were marked as expressed were considered for building the coexpression networks. For GTEx data, I used the gene-level expression data from version 6, downloaded from the GTEx portal for the five tissues blood, brain-cortex, liver, lung and skeletal-muscle.

2.2.1.1 Affymetrix microarray datasets:

I reanalyzed the data from raw CEL files using the RMA algorithm. Outlier samples were detected and removed using the Gemma outlier detection algorithm. Batch effects were removed using the ComBat algorithm in the sva package in R, with the batch information based on the date stamp in each CEL file. For mapping probesets to genes I used annotations
from Gemma. For genes with multiple probesets the mean value was recorded as the expression level; however probesets with very low expression level (lowest tertile) were removed if the gene had at least one other probeset expressed at higher levels. The final data matrix included 18,494 genes and 3,563 samples (Figure A-1).

2.2.1.2 GTEx dataset:
The gene-level expression of GTEx version-6 was downloaded from the GTEx portal \(^{152}\). I used data for blood, brain cortex, liver, lung and skeletal muscle. Genes which had read-counts > 0 in more than 75% of the samples were included in the analysis with no further normalization (Table A-9). In pairwise comparisons between Affymetrix and GTEx datasets, only genes which were marked as expressed in both datasets were considered.

2.2.1.3 Filtering of genes based on expression level for Affymetrix datasets
For each Affymetrix dataset, I filtered the genes based on their expression levels such that genes with mean expression below the first tertile were marked as not expressed, retaining 12,312 genes out of the 18,494 genes on the platform. Genes which were expressed in > 80% of the datasets for a tissue were marked as expressed for that tissue (Tables A6 to A10). For each tissue, only the expressed genes were considered for downstream analysis, including the measurements of network overlaps and densities. This was to avoid spurious correlations due to cross-hybridization (see Figure A-2).
2.2.2 Tissue Aggregated Networks (TAN)

The process of network generation is outlined in Figure 2-2. TANs were built by aggregating binary coexpression networks from Affymetrix datasets for each tissue (Figure 2-2:1), similar to the approach used in Lee et al. 7. Specifically, from each dataset, a raw coexpression matrix was first computed using pairwise Pearson correlations between the gene pairs. Binary networks were built from these raw networks by selecting the gene pairs having coexpression value > 90th percentile within the dataset (Figure 2-2:2). These selected pairs are referred throughout as “links”. For each tissue those links present in at least in n binary networks were marked as present. The value of n was chosen for each tissue to maintain an FDR of 10^-4, using the binomial distribution as the null (Figure 2-2:3). This FDR was chosen to obtain networks of reasonable density, but my key findings are not affected by the exact choice.
Aggregated Networks (TANs) are built by combining binary networks from individual datasets. In each TAN, tissue-specific links are identified based on their correlation values in all the datasets. The tissue-specific links comprise Tissue Specific Networks (TSNs).
2.2.3 Identification of tissue-specific links

Tissue-specific links were identified for each tissue as a subset of the links in the tissue’s TAN. This process has two steps: 1. Measuring a Tissue Specificity Score (TSS) for each link. 2. Setting a threshold TSSs to control false discovery rates.

Measuring the TSS for each of the links in the TANs

My goal was to capture pairs of genes which consistently have relatively high coexpression in one tissue (the target tissue), while having consistently lower coexpression in all the other tissues. To implement this, I first define the correlations for a pair of genes \((g_i, g_j)\) in each of the \(k\) data sets, referred to as \(S_{ij}\):

\[
S_{ij} = \{corr(i,j,k) | i \neq j, 1 \leq k \leq 53\}
\]

Where \(corr(i, j, k)\) is the Pearson correlation value between the genes \(g_i\) and \(g_j\) from dataset \(k\) (\(D_k\)). I modify these raw correlation values for each of the links in two steps. The first step is the normalization between the datasets using a binned rank transformation. The \(corr(i, j, k)\) is replaced with the permille it belongs to between coexpression values from all the expressed genes in \(D_k\) (this binning was used to simplify the implementation). Since I was attempting to capture the presence of coexpression in the target tissue and the absence of it in the other tissues, negative correlations were not of interest, and they could affect the measurements as outliers. Therefore, all negative correlations were assigned to bin 500, corresponding to approximately a zero correlation. More formally:

\[
SB_{ij} = \{\max(bin(r,k),500) | r \in S_{ij}, i \neq j, 1 \leq k \leq 53\}
\]

47
Where \( \text{bin}(r, k) \) gives the permille of the correlation value \( r \) in dataset \( k \). Next, having the tissue \( t_m \) and the link \( l_{ij} \), I define two subsets of \( SB_{ij} \) as follows:

\[
CT^t = \{ r_k \mid r_k \in SB_{ij}, \text{dataset } k \text{ is from tissue } t \}
\]

\[
\overline{CT^t} = SB_{ij} - CT^t
\]

From this, the TSS of a given pair of genes \((g_i, g_j)\) and a given tissue \( t_m \) is calculated as follows:

\[
TSS(g_i, g_j, t_m) = \frac{\sum_{p \in CT^t} \sum_{q \in CT^t} f(r_p - r_q)}{|CT^t| \cdot |\overline{CT^t}|}
\]

\[
f(x) = \begin{cases} x, x > 0 \\ 0, x \leq 0 \end{cases}
\]

I also applied an alternative method using the p-value from a Wilcoxon rank sum test, comparing the ranks of the correlations for the gene pair in one tissue to that in all others. The results of the two approaches are highly correlated (see Figure A-3) and I focused on the TSS as defined above.

**Controlling false discovery rates for TSSs**

To assess the null distribution of TSS, I created random subsets of the 53 datasets pseudo-tissues and computed TSSs for each of them the same way as the real tissues. Each pseudo-tissue was constructed to have all five tissues represented equally. I built 30 sets of pseudo-tissues for each tissue, with the same count of datasets as the tissue. The TSS-FDR for the tissue \( t \) at a given value \( tss \) is calculated as follows:
\[ FDR_t(tss) = \left( \frac{\sum_{pt=1}^{30} |Y^{tss}_{pt}|}{30} \right) \frac{1}{|Y^{tss}_t|} \]

Where \( Y^{tss}_{pt} \) is the set of TSS values for pseudo-tissue \( pt \) which are greater than \( tss \). Similarly, \( Y^{tss}_t \) is defined for tissue \( t \). I selected the TSS threshold for the FDR 0.01 for each tissue.

Selected TS links were further filtered to select those with their TSS <0.4 in all the other tissues.

### 2.2.4 Modeling the effect of expression level on tissue-specific links

I used a linear model to examine the predictive power of expression level of pairs of genes for their coexpression value:

\[
bin(corr(i,j), k) = \beta_0 + \beta_1(e_{ik} + e_{jk}) + \beta_2|e_{ik} - e_{jk}|, i \neq j, 1 \leq k \leq 53
\]

Where \( e_{ik} \) is the mean expression level of gene \( i \) in dataset \( k \). This model captures the predictive value of the total expression of the two genes (\( \beta_1 \)) and their difference (\( \beta_2 \)). The model was fit to all the links in the five TSNs. The \( R^2 \) of the model fit is a continuous measurement of the predictability of each link based on the expressed level of their genes. These \( R^2 \) values were compared to a null \( R^2 \), which was computed by shuffling the mean expression level of the genes across datasets and refitting the models.

### 2.2.5 Coexpression networks built from GTEx datasets for validation

I built binary and raw coexpression networks for the five GTEx datasets. Raw coexpression matrices were computed with the Pearson correlation of the gene expression profiles. To
create binary networks, the threshold was selected so that each resulting network had the same
density to the TAN of the tissue it was compared to.

2.2.6 Gene ontology terms

The Gene Ontology^{26} data was downloaded on 31.03.2015. I used experimentally validated
Gene Ontology terms under the Biological Process sub-group for human. Terms were filtered
to have $\geq 10$ and $\leq 200$ genes assigned to them, resulting in 3,507 GO terms and 8,336
genes with one or more GO terms assigned to them.

2.3 Results

2.3.1 Tissue Aggregated Networks (TANs)

The primary analysis is based on 53 high-quality datasets representing five human tissues
(brain, lung, skeletal muscle, blood and liver), collected on Affymetrix GeneChips (7-15
datasets per tissue, 3,563 samples in total, assaying 18494 genes; see Figure A-1,A). For each
tissue, I identified “expressed” genes based on their mean expression levels (Table A-1, see
Section 2.2.1 for details). I then built Tissue Aggregated Networks (TANs) for each tissue by
aggregating binary coexpression networks constructed from each dataset for that tissue,
retaining links that were observed in multiple data sets, where the threshold minimum number
of dataset networks was set to control the false discovery rate (FDR $< 10^{-4}$, Figure 2-2, see
Section 2.2.2). Each TAN has a different number of links (Table A-6), but this was not
significantly correlated with the number of genes present in the networks or the count of datasets for each tissue. The TANs form the basis for the ensuing analysis.

2.3.2 Many links are tissue-specific

I next identified tissue-specific links in each TAN. Here I define a tissue-specific link as a link which has higher correlation values in one tissue (the “target” tissue) versus all the “other” tissues, and allow for a certain level of noise. To do so I computed a Tissue Specificity Score (TSS) for each TAN link and a tissue, based on the average of the positive difference between normalized correlation values from the target tissue versus the rest of the tissues (see Section 2.2.3). The TSS for a link and a tissue reflects its specificity to that tissue. I generated null distributions for TSS to control the FDR for identification of tissue-specific links. Figure 2-3, A to E shows the distribution of TSS for pseudo-tissues, all the expressed gene pairs and the TAN links. I identified a fraction between 0.03 to 0.32 of the TAN links as tissue-specific (Table A-2, FDR < 0.01).
(A-E) For each tissue, the distribution of Tissue Specificity Scores is plotted for all the links in its TAN (solid line) as compared to ten generated nulls (pseudo-tissues; grey shading) and all the expressed gene pairs in the tissue (dashed line). (B) Summary of links considered tissue-specific or not, selected at FDR .01 compared to the null.

2.3.3 Expression level changes explain much of the tissue specific coexpression

To capture the relation between mean expression level and coexpression, I considered linear models in which the correlation between two genes is modeled as a function of the mean expression levels of the linked genes (see Section 2.2.4). The fitted models were then examined for how well they could explain the correlation of the link in the five tissues.
compared to a null using shuffled expression values. I find that for most of the tissue-specific links, the model $R^2$ is higher compared to the null at FDR $> 0.01$ (Figure 2-4A). This confirms that differences in mean expression can at least partly explain differences in coexpression.

Tissue-specific links that were not substantially associated with mean expression levels (having a low $R^2$ for the model fit) are candidates for pure rewiring, since they are relatively uninfluenced by the confound of overall expression level shifts (as captured by the model). However, failure for the models to predict coexpression is not a sufficient constraint, so I also required that pure links have both genes marked as “expressed” in all the tissues. This results in a smaller number of tissue-specific links, noted with dashed boxes in Figure 2-4B. Thus “pure differential coexpression” links are by definition between genes expressed in all tissues, and variability in mean expression level across tissues does not appear to be associated with tissue specificity of coexpression. I stress that this definition is fairly stringent, but still allows for some variation in mean expression levels. An example of a pure differentially-coexpressed link is shown in Figure 2-1 and additional examples are given in Figure A-6. I conclude that while much of the difference in coexpression among tissues is attributable to changes in mean expression level, there is a small number of links that appear to undergoing “pure” differential coexpression.
2.3.4 Reproducibility and validation of the networks

I examined the reproducibility of the TAN, TSN and pure networks in the five GTEx datasets: blood, brain-cortex, liver, lung and skeletal-muscle. I expected that links coming from a tissue will have generally higher correlation values in the same tissue in GTEx, compared to the other tissues. I examined the distribution of the correlation values in each of these GTEx datasets for any given groups of links (TAN, TSN and pure) from the five tissues. Figure 2-5A shows the distribution of correlation values for different groups of links in GTEx brain-cortex dataset. As expected, TAN, TSN and pure links from brain have generally much higher correlation values than the null in GTEx brain-cortex. TAN links from other tissues have higher correlation values than the null as well (though much less than the brain TAN), and this is expected due to the common links between TANs. On the contrary, TSN and pure links from other tissues have generally much lower correlation values than the null. This is also
expected since the TSN and pure links from other tissues are not expected to have high
correlation values in GTEx brain-cortex. Figure A-4 has similar plot for the rest of the tissues
and Figure 2-5B summarizes the reproducibility of the TAN links in the five GTEx tissues in
terms of likelihood ratio. On average, reproducibility of the pure links is less than TSN links
(52% of tissue-specific links versus 25% of pure links are reproduced in GTEx – see Section
A.4).

One of the known characteristics of coexpression links is that they can capture
functional similarity between genes. In a validation, I examined this for my binary networks
and show that in the TANs, TSNs and GTEx binary networks, genes sharing a coexpression
link are more likely to share some functional terms with each other than that expected by
chance. This is also true for genes with pure links (Figure 2-5C).
(A) Distribution of correlation values for TANs, TSNs and pure links in GTEx brain-cortex dataset. The filled grey violin is the distribution of all the correlation values in GTEx brain-cortex, included as the null reference. Distributions for the brain groups (highlighted) are more dense at higher values compared to other tissues. Generally, links from all the five TANs have higher correlation values compared to the null. For all the TSNs the distribution is close to that from the null, except for the brain TSN. The pure TSN links show similar results to TSN links, although a bit weaker. (B) Summarization of the TAN reproducibility in GTEx binary networks. The dots which represent links from the same tissue as the TAN are circled with black. Arrow points to an example: the count of links from blood TAN in the binary GTEx blood network are >16 times more than expected by chance. (C)
Likelihood ratio of having at least one same functional annotation between the genes that are connected in each of the networks versus the genes that are not connected. The grey dot represents the union of the networks.

2.3.5 Characterization of “pure” differential coexpression

Despite the relatively small number of pure links, it is naturally of interest to understand if they have any particular biological relevance, but this presents a challenge. By definition, the genes participating in pure links are expressed in all of the tissues, and therefore tend not to have any easily identifiable tissue-specific function. Furthermore, 86% of genes expressed in all tissues (those eligible to have pure links) have at least one pure link (6,658/7,745) and 4,267 have pure links in more than one tissue. I was unable to identify any distinguishing functional feature of the genes with pure links compared to the minority that lack them. They also do not stand out in terms of functional similarity: genes associated by a pure link are more likely to be functionally similar, but no more than the other classes of links (Figure 2-5,C). I also found that the neighborhoods of genes paired with pure links show higher topological overlap in the target tissue compared to the other tissues (Figure A-5, see Section A.4). This helps validate pure links as robust differential coexpression, but does not address their biological meaning.

I next turned to analyzing how pure links contribute to the network-level representation of gene function. I defined the enrichment of a term in a network based on the number of links connecting the genes annotated with that term compared to a null, so a function is enriched if genes with that function have a statistically significant number of links
I hypothesized that pure links in a tissue could contribute to tissue-specific enrichment of common cellular processes and tested this by examining what happens if we remove pure links from the network. I first found that 146, 91 and 43 functions are enriched exclusively in the brain, liver and lung TSNs respectively (FDR < 0.01; too few functions were enriched for blood and skeletal muscle for further analysis). When pure links were removed from these TSNs, the count of functions enriched in brain dropped by 20, by 3 in liver and by 4 in lung. The number of such “drop-out” functions for brain and lung is less than those affected by removing equivalent numbers of random links (p < 0.05, see Section A.4). It is also less than removing random sets of expression-induced links (p < 0.01, see Section A.5). Enrichment of five functions in brain were sensitive only to the removal of pure links including “regulation of mitochondrion organization” (GO:0010821) and “phospholipid biosynthetic process” (GO:0008654 – see supplementary tables linkRemoval[...].txt; none of the enriched functions in lung and liver showed such sensitivity). I also performed an analysis of disease associations (rather than functions defined by the Gene Ontology; Section A.6). Brain and lung had 18 and 30 diseases exclusively enriched in their TSN. Of these, 1 and 5 (respectively) were sensitive to the removal of pure links. However, all of these terms were also sensitive to the removal of random sets of links and expression-induced links.

2.4 Discussion

I examined coexpression in five human tissues in an attempt to document the effect of the mean expression level on tissue-specific coexpression. In each of the five tissues that I have
studied, I was able to identify robust coexpression links which are, to a great extent, reproducible in external datasets (TANs) and represent functional similarity between the genes. I was also able to identify tissue-specific links (TSNs). However, I found that, as hypothesized, the majority of differential coexpression between the tissues (tissue-specific links) can be substantially explained by changes in mean expression level of one or both genes and that cases of pure differential coexpression are relatively rare. Many of the pure links are reproduced in GTEx datasets but as a group, pure links contribute to the tissue-specific enrichment of a few functional terms while in comparison, expression-induced links contribute to the tissue-specific enrichment of many functional terms. This study has implications for other studies of differential coexpression, as I now discuss.

I have stressed the importance of controlling for the confounding effects of mean expression level on coexpression. Some previous literature also distinguishes differential expression and differential coexpression and refer to their complementary role. The confounding role of the differential expression has also been noted in the tissue based coexpression networks, where it was observed that the coexpression modules in tissues are enriched with tissue specific genes. Despite this, I am aware of only one attempt to control for expression levels in DCA, in an analysis of a brain development data set. More commonly differential expression is merely reported or commented upon alongside differential coexpression results. But most studies of differential coexpression do not even go that far. I suggest that when differential expression is present, differential coexpression should be interpreted very carefully. This seems especially important given the
biological relevance or meaning of differential coexpression has rarely matched the hopes of revealing “rewiring”. In fact, I am not aware of any case where “regulatory rewiring” has been uncovered in an unbiased differential coexpression analysis, especially in the absence of differential expression.

One of the challenges in identifying differential coexpression is controlling sensitivity and specificity. To help control multiple test penalties that would be incurred by testing every pair of genes, most studies use data reduction approaches such as considering only a subset of genes, or “gene modules” in the networks, comparing their attributes or their preservation. A drawback of the module-based DCA is that it leaves little room to control for the effect of expression level on the individual genes.

Validation also remains a challenge. While some studies report findings based on DCA, to my knowledge they do not present any validation or replication of the differential coexpression. A strength of my work is that I use large quantities of data (as required to obtain robust coexpression results), and I show that some “pure” differential expression is reproduced in the GTEx RNA-seq data set. This was facilitated by the relative ease of finding data sets for tissues compared to more specialized conditions, but even so I was not able to find sufficient data to test more tissues. Similarly, it is unclear how sensitive previous studies have been. Thus most analyses for differential coexpression are likely to be underpowered or overwhelmed by noise, but also have an unclear false positive rate.

At this point I have yet to identify any clear biological significance for any of the pure links individually. This is partly because the nature of the genes makes it difficult to apply
bioinformatic characterization methods to them. First, pure differential coexpression links are between genes expressed in all five tissues. This is in contrast to genes that are tissue-specific, which often have known, or more easily inferred, tissue-specific functions. Thus we cannot ascribe a tissue-specific function of a pure link based on the expression pattern of the genes across tissues. Also, most of the genes expressed in all the tissues have at least one pure link – they are not a “special” small subset of genes. Second, functional annotations of genes are rarely tissue-specific. That is, there are few cases where a gene expressed in multiple tissues has known distinct functions in those tissues (at least, such information is not captured by resources such as the Gene Ontology). Overall, this means there is a dearth of information on tissue-specific gene functions that are not attributable to differences in mean expression level, and thus evaluation of pure links is a challenge. My results from a network-wide functional enrichment analysis showed that pure links in brain contribute to the tissue-specific enrichment of a few functional terms. This does not lead to any clear biological conclusion, but could assist the interpretation of results from other genomic or transcriptomic studies on the brain. Finally, unlike protein-protein interactions, coexpression is not a discrete property of a pair of genes nor an actual physical interaction. Coexpression is derived from the correlation matrix of the entire data set, so any tightly coexpressed pair of genes is essentially guaranteed to be part of a larger pattern (i.e. cluster or module) of genes. Therefore coexpression links do not map cleanly to regulatory relationships nor protein-protein interactions. As a result, the core idea of DCA as reflecting “rewiring” by necessity involves gene neighborhoods, modules or clusters. Taken together, these issues mean that the ability to
extract specific (i.e. biochemical or physical) interactions from differential coexpression is severely limited. This limitation does not necessarily detract from the potential utility of tissue- or other context-specific data to improve the relevance of function predictions from coexpression.
Chapter 3: Untangling the Effect of Cellular Composition on Coexpression Analysis

3.1 Introduction

Coexpression analysis is among the most-used methods in transcriptome data interpretation. The biological underpinnings of coexpression are well-established. Within a cell, genes whose products work together (either directly or indirectly) must be expressed together. This implies some commonality of regulation. Indeed, it is observed that genes with similar functions tend to be coexpressed\(^2,7,90,144\). Based on these observations, coexpression is used in inferential frameworks (often via network-based approaches) to aid prediction of gene function and/or regulation\(^93,141,142,150,164\). In this chapter, I examine the assumptions that underlie such applications of coexpression to “bulk” samples of tissues containing multiple cell types. In particular, I explore the role played by variation in cellular composition.

In bulk brain tissue transcriptome datasets, gene expression clusters (sets of genes which are observed to be coexpressed) are often enriched for cell-type markers\(^165\). Recently it has been proposed that variation in cell type composition between individual samples explains a substantial degree of variation in gene expression in human brain\(^166\). In general, cell-type “deconvolution” methods rely on the idea that cell-type markers can be used to infer cellular composition\(^167,168\). Inferred cellular composition is also used for adjusting statistical models, as in some expression quantitative trait locus (eQTL) analyses\(^169,170\). Thus there is at least implicit awareness that cellular composition is a factor in transcriptome data\(^144,158\). However,
to my knowledge the connection between these observations and the interpretation of coexpression network analysis has not been described in detail.

In this study, I document the effect of cellular composition variability among samples in bulk nervous system tissue, and its downstream effect on network-based functional analysis. Using a combination of bulk tissue and single-cell data analysis, supplemented by simulations, I demonstrate that for a given gene the variance of its expression level in bulk tissue is directly related to its variability across cell types. I then show that this is strongly related to coexpression of genes with each other, such that the dominant signal in bulk tissue is simply due to variation in cellular composition across samples. Because many gene functions are highly associated with specific cell types, my results provide a major reason why clusters enriched for functions are observed in expression data. A further implication is that the utility of bulk tissue coexpression to infer transcriptional regulatory networks beyond uncovering cell-type specific expression patterns is greatly complicated. While this study focuses on expression in the human nervous system, the phenomena I document are likely to play an important role in analyses of other tissues.

3.2 Methods

3.2.1 Data

2. GTEx RNA-seq expression dataset from brain-cortex

3. A set of coexpression networks: binary coexpression networks were built from GTEx RNA-seq blood and liver and a set of Tissue Aggregated Networks (TANs) from blood, brain and liver, from Chapter 2. The TAN networks are built by aggregating several networks from each tissue, built from datasets on Affymetrix platform (see Section 2.2.2). The TSN-brain network is a subset of the TAN-brain network, where the links are identified as specific to the brain, blood or liver. Table B-1 provides counts of genes and links in each of the networks.

3.2.2 Single-nucleus data

The snuc-RNAseq dataset has records from 15,928 nuclei for a total of 50,281 genes, grouped into 75 cell types. I used the read counts from exons only and did not use the intronic reads. I used the labels for the cell-types based on the clustering provided by the Allen Institute. I removed nuclei which had data for less than 2000 genes and nuclei for which the total read count was more than 3 times or less than 1/3 times the median. Genes were filtered for NeuN negative and NeuN positive samples separately. I selected genes expressed in at least 2% of the nuclei or expressed at the highest quartile in the nuclei it is expressed in. The final dataset has data for 16,789 genes and 15,646 nuclei. Figure B-2 shows the count of cells in each group of the 75 cell types. To construct Cell Type (CT) vectors for each gene, I obtained the mean expression level of that gene in each of the 75 cell types. Each of the 16,789 genes yields a CT expression profile vector of 75 elements.
I built coexpression networks for 69 of the 75 cell-types in the snuc-RNAseq data set (six cell-types had too few cells). For each cell type, correlations were computed for each pair of genes using only nuclei in which expression was greater than zero to reduce the impact of drop-outs. Gene pairs with less than 20 usable nuclei were removed. Because of differences in sample size for the correlations (causing different null distributions for the correlation), I omitted the correlation threshold filtering step used for the other data sets, and therefore filtered the one-sided p-values of the correlations to identify the 0.5% of the gene pairs with smallest p-values. Table B-2 has the link count and gene count for the 69 networks.

To construct combined networks, I summed the 64 binary coexpression networks built from Inhibitory and Excitatory neurons. Robust coexpression links were identified as those present in 10 or more of the networks, between genes with more than two such links. The total of 490 genes passed this criterion and were clustered using topological overlap and hierarchical clustering. Sixteen mitochondrial genes and 10 unclustered genes were removed (the presence of mitochondrial genes is likely due to variable mitochondrial contamination of the nuclei). The remaining grey and black clusters have 286 and 178 genes respectively.

3.2.3 GTEx datasets and networks

The read counts per million reads (CPM) values from each of the three GTEx datasets: brain-cortex, liver and blood were filtered to include the genes with CPM > 0 in > 20% of the samples. Expression values were log$_2$ transformed and binary coexpression networks were built using the Pearson correlation, filtered to include the 0.5% of the links with highest
correlation values in each of the three networks. The counts of links and genes included in each network are provided in Table B-1. To cluster the GTExBulk network, I applied hierarchical clustering to the Topological Overlap (TOP)\(^82\). An initial set of 253 clusters were identified for 12,416 of genes, of which 69 clusters had at least 20 genes and were retained for further study. Clustering labels are in Additional file B8. Counts of marker genes for different cell types for each cluster are in Figure B-2.

3.2.4 Identification of marker genes

Marker genes were selected based on two sources. From snuc-RNAseq data, for each of the five major cell types Astrocyte, Oligodendrocyte, Microglia, Endothelial and Pyramidal (labeled as Excitatory cell types), I identified genes with mean count per million \(\geq 2\) fold-change in all other cell-types as marker genes. Additional file B9 has list of markers identified by this manner for each of the five cell types.

As the second source, I used 1,208 marker genes for 18 mouse cerebral cortex cell types identified by Mancarci et al.\(^{171}\). I mapped the marker genes to their human orthologs using the Ensembl database\(^{172}\). Mancarci et al. further refined these markers based on their coexpression in bulk human tissue. To remain consistent with their method, for each cell type I only considered the subset of its marker genes which were highly correlated with each other, using the hierarchical clustering with topological overlap on the marker genes for each cell type. Genes in the cluster with the highest count of links were selected as the markers of the
cell type. The final list includes 256 markers for five major cell types (Additional file B10). Figure B-2 shows overlap of the marker gene sets from the two sources.

3.2.5 Modelling expression level of genes in the GTExBulk dataset

I used linear models to estimate the expression level of genes based on the variation of the marker genes in each of the samples, using the first seven principal components of the whole set of marker genes (the snuc-RNAseq and Mancarcì 171 marker sets separately) in the bulk tissue dataset. Therefore, the expression level of gene \( A \) in sample \( j \) of the GTExBulk dataset is modeled as:

\[
Exp_j(A) = \mu_A + \beta_1 s_1 + \beta_2 s_2 + \cdots + \beta_7 s_7 + \varepsilon_{A,j}
\]

Where \( \mu_A \) is the average expression level of gene \( A \), \( s_i \)'s are the principal component scores for sample \( j \), \( \beta_j \)'s are the parameters of the model and \( \varepsilon_{A,j} \) is residual error.

3.2.6 Enrichment of functional terms

For a given network, each functional term is marked as enriched if the density of the links between the genes associated with the term are significantly higher than the density of the network, where the hypergeometric distribution is used as the null. The FDR was controlled at 0.1 using the method of Benjamini and Hochberg 173. Brain-specific functional terms were identified as the terms enriched in either of the TAN-brain networks or GTExBulk network, but not enriched in TAN-liver and TAN-blood networks. Likewise, housekeeping terms were
identified as terms enriched in either of the TAN-brain or GTExBulk networks, as well as the TAN-blood and TAN-liver networks.

### 3.2.7 Synthesized bulk datasets

Each of the synthesized samples was built using snuc-RNaseq samples (nuclei) from the Allen data described above, as follows. First, nuclei were grouped into five major cell types based on their provided labels. Then, for each synthetic sample, nuclei were randomly sampled with the following baseline proportions: Pyramidal (20%), inhibitory (20%), Oligodendrocyte (43%), Astrocyte (12%) and microglia (5%), based on the estimates of Von Bartheld et al.\textsuperscript{174}, and their gene expression values were added and divided by the total count of nuclei, yielding a final synthetic sample. This was repeated to create multiple synthetic data sets with 100 samples each. To add composition variability, the baseline proportions were randomly varied by drawing each proportion from a normal distribution with variance 33% of its mean.

### 3.3 Results

#### 3.3.1 Gene variance is highly affected by cellular composition

This work builds on two empirically-founded concepts. The first is that many genes are expressed at different levels in different cell types in the brain. The second is that brain tissue samples vary in their precise cellular composition. The latter occurs due to technical (e.g. sampling variability) and biological effects\textsuperscript{174}. The connection between the two in the context
of bulk-tissue transcriptomics can be formalized in the following simple model, schematized in Figure 3-1 (for mathematical details see the Section B.1). For each gene, I define a Cell Type (CT) expression profile, which is a vector of expression levels of the gene in each of \( k \) cell types. In the model, the CT profile is treated as a fixed intrinsic feature of the gene. Second, each bulk tissue sample has a specific cellular composition for those same \( k \) cell types. This forms a cellular composition vector of length \( k \) for each sample, where each element represents the proportion of a cell type in the sample. The observed expression level of a gene in the sample can be modeled as a weighted sum of the values in the CT profile, where the weights are given by the cellular composition vector of the sample. In the toy example shown in Figure 3-1, Gene 1 is only expressed in cell type B and therefore its relative expression in the data precisely tracks the proportion of cell type B present in the samples. This special case is used in many approaches to “cell type deconvolution”, where Gene 1 is considered a “marker gene” for cell type B. In contrast, Gene 5 is expressed equally in all cell types, so it is completely insensitive to differences in cellular composition and its expression level is the same in all the samples (for further mathematical details see Section B.1). The expression pattern becomes more complicated for a case like Gene 4, which is expressed at different levels in each of the two cell types, but because it is expressed at higher levels in cell type A, its pattern in the bulk tissue is positively correlated with the proportion of cell type A. Furthermore, genes that have correlated CT profiles will also be correlated in the bulk tissue (illustrated in Figure 3-1 by genes 1 and 2, and genes 3 and 4).
In general, the model predicts that the more variable a gene’s CT profile, the more its measured expression in bulk tissue will be affected by variability in cellular composition of the samples (see Section B.1 for simulation results demonstrating this). It is important to note that this model ignores all other potential sources of variability including noise or technical artifacts, as well as interactions between genes or cells that can influence expression. My goal is to explore how well this effect explains the observed variance and correlation of genes in bulk tissue data.

**Top: Cell type (CT) profiles for five genes in a hypothetical tissue with two cell types.**

Genes 1 and 2 are marker genes for cell type B. Gene 3 is a marker gene for cell type A. Gene 4 is expressed in both cell types, but at different levels, while Gene 5 is expressed at equal levels. Middle: Hypothetical cellular compositions of five bulk tissue samples. Each sample (alpha) has the same amount of biological material but different proportions of each cell type. **Bottom:** The expected observed expression levels. Gene 1 and Gene 2 are positively correlated and negatively correlated with Gene 3 and 4. Gene 5 is expressed at the same level in all the bulk tissue samples as it is equally expressed in all cell types.
As an initial assessment of whether this model is broadly explanatory, I estimated CT profiles for human cortex from single nucleus RNA-seq data (snuc-RNAseq data, see Section 3.2.2 for the method), yielding expression levels for 16,789 genes in each of 75 different cell types, including all of the major classes of cells expected to be present in bulk cortex. I compared these data to a bulk cortex transcriptome dataset from GTEx (GTExBulk, see Section 3.2.3 for method). As predicted by the model, the variance of a gene’s expression in GTExBulk is correlated with the variance of its CT expression profiles (Spearman’s rho = 0.18; see Figure B-3). Given the many potential sources of error, including noise in the CT profiles as well as the GTExBulk data, the agreement with the naive model is striking.

I next applied an approach related to many deconvolution methods to estimate the amount of variance attributable to cellular composition effects for each gene. As demonstrated in Figure 3-1, expression levels of cell-type marker genes in bulk tissue will reflect the variation of cellular composition among the samples. Therefore, the cellular composition-induced variance of the genes could be modeled by the variation of marker genes in a dataset. Here I used Principal Component Regression (PCR) using the expression of marker genes to predict the variation of the non-marker genes in the GTExBulk dataset (see Section 3.2.5). The amount of variance explained by the model for each gene ($R^2$) is an estimate of the degree to which the gene’s expression pattern is due to variability in cellular composition. In the first analysis, I used sets of marker genes from a high-quality snuc-RNAseq dataset for five major brain cell types (Pyramidal, Microglia, Astrocyte, Oligodendrocyte and Endothelial; see Section 3.2.4). The resulting gene-level values of $R^2$ range up to 0.91 (90th
quantile is 0.79) with a median of 0.46. In contrast, the same models fit to the snuc-RNAseq data, where no effect of cellular composition is expected (barring contamination of individual nuclei), the mean $R^2$ is 0.018, with only 38 genes having values greater than 0.2.

As predicted by the model, $R^2$ values are correlated with the variance of the CT expression profiles ($\rho = 0.28$). To check the robustness of these findings, I tested another set of (largely non-overlapping) marker genes from Mancarci et al. with similar results ($\rho = 0.3$; see Section B.2). I also tested randomly selected sets of non-marker genes instead of markers and found that $R^2$ values are significantly higher for the marker genes when PCs are obtained from marker genes compared to the random selection of genes with similar average expression levels ($p < 0.01$ for average of $R^2$ of marker genes for 100 trials, see Figure B-2). Likewise, the two marker sets also generated higher $R^2$ values for each other than the random gene sets despite their small overlap.

Motivated by reports that coexpression clusters are often associated with tissue-relevant gene functions, I next examined the relationship between gene function and cellular expression patterns. I observed that genes associated with brain-related functional terms (see Section 3.2.6) tend to have higher $R^2$ values, consistent with expected cell-type specific expression patterns in the brain (see Figure 3-2A). That is, genes with a brain-related function tend to have more varying CT profiles -- they are enriched in particular cell types -- which leads to high variation in bulk tissue. For example, genes involved in synaptic transmission are expressed in neurons, while genes involved in myelination are expressed in oligodendrocytes. Examples are genes annotated with “Regulation of synaptic plasticity”
(GO:0048167, mean $R^2 = 0.76$) and genes annotated with “Axon ensheathment”

(GO:0008366, mean $R^2$: 0.68; see Additional file B1). In contrast, terms for housekeeping functions tend to be associated with genes with lower $R^2$ values (Examples: “Histone demethylation”, mean $R^2$ value: 0.61 – GO:0016575; “spliceosomal snRNP assembly”, average $R^2$ value: 0.53 – GO:0000387 – see Additional File B2). In a closer examination, we also see that genes associated with the brain-specific term “Regulation of synaptic plasticity” have significantly higher variance in GTExBulk dataset compared to genes associated with the housekeeping term “Histone demethylation” ($p = 0.005$, t-test). In contrast, in the snuc-RNAseq cell population (a dataset expected to not have cellular composition effects) they have significantly lower variance ($p = 4e-4$; see Figure 3-2B). In summary, these results demonstrate that some of the observed variance of gene expression can be attributed to cell type composition variation, and this is especially true for genes with tissue-specific functions due to their tendency to also have cell-type specific expression patterns.
Figure 3-2 Much of the observed variance is explained by cellular composition

(A) Groups of genes associated with brain-specific functional terms tend to have higher cellular composition-associated variance (indicated by higher average R² values) compared to groups of genes associated with housekeeping terms. The average R² values for two example terms are highlighted with black and blue dots. (B) Distribution of gene-level variance in the GTExBulk and snuc-RNAseq cell population (Exc L2-3 LINC00507 FREM3) for two groups of genes. Genes associated with brain-specific term “Regulation of synaptic plasticity” have higher variance than genes associated with the housekeeping term “Histone demethylation” in the GTExBulk dataset, while they have slightly but significantly lower variance in the snuc-RNAseq cell population.

3.3.2 Compositional variation explains much of the bulk tissue variance

In the previous sections of this chapter I demonstrated that variation in gene expression can be partly accounted for by variation in cellular composition. As illustrated in Figure 3-1, genes which have similar patterns of expression across cell types (as evidenced by correlated CT
profiles) are also expected to have correlated expression in bulk tissue. Importantly, this phenomenon will be observed for any gene which has variability in expression across cell types, not just highly cell-type specific marker genes. For any two genes, in the absence of other factors, as the correlation between their CT expression profiles approaches one (or minus one), their correlation in bulk tissue is expected to approach one (or minus one – see Section B.1, Figure 1 and Figure 3C). I call this “cellular composition-induced coexpression”, to be distinguished from coexpression due to “within-cell” co-regulation. I hypothesized that it is a major source of observed coexpression in bulk tissue.

I first performed clustering of the GTExBulk gene expression profiles, yielding 69 clusters (minimum 20 genes each; see Section 3.2.3). As expected, some clusters are enriched with markers for one cell type (Figure 3-3A). While many of the other genes in these clusters are not markers, they tend to be “quasi-markers” – they are enriched in expression in a cell type (as for Genes 1,2 and 3,4 in the simplified model in Figure 3-1 – see Figure B-5 for expression levels of genes from clusters associated with different cell types, compared to marker genes). Furthermore, clusters that are enriched for markers for the same broad cell types and most of their neighboring clusters have correlated average CT expression profiles (0.993 > rho > 0.42, all p < 5e-4, Figure 3-3A and B). In addition, the average R² values from the regression model are generally high for the marker-enriched clusters, consistent with composition-induced variance in expression (9/10 marker enriched clusters have average R² greater than the median for all clusters (median = 0.65), Figure 3-3A; see Additional file B3 for all values).
The results so far make it apparent that some of the observed coexpression in the bulk brain tissue is explainable by cellular composition variation. Since cell-type specific patterns of expression are likely to be relatively fixed and therefore reproducible, composition-induced coexpression is also likely to be reproducible, and therefore contributing to the reported reproducibility of coexpression clusters among different bulk brain datasets. I examined this by comparing brain bulk tissue coexpression networks with each other and also with coexpression networks from other tissues. In the GTExBulk coexpression network, the intra cluster coexpression links in the clusters enriched with brain marker genes contain 49% of the total links. This is up to 40 times higher than the null expected value for the count of genes in the clusters for the given density of this network (see Figure 3-3D). The same set of genes have a high count of links in TAN networks from Chapter 2, confirming that much of the reproducibility among bulk brain networks can be explained by cellular composition-induced coexpression. Importantly, while most (60-80%) of the genes in these clusters are also expressed in blood and liver, the high degree of observed coexpression among these particular genes is a phenomenon specific to the brain. I also see a large increase of links between the genes in marker enriched clusters in my simulated bulk tissue data upon the introduction of cellular composition variation (Figure 3-3D – see 3.2.7 for methods).

Apart from the marker-enriched clusters, many clusters in the GTEx-derived network are enriched with housekeeping genes and/or functions (see Figure 3-3). Most of these clusters have low mean $R^2$ values (18/28 have mean $R^2$ less than the median of all clusters (0.65) – see Additional files B3, B4), suggesting that their genes have small variability in their
CT expression profiles and their coexpression is less likely to be affected by the cellular composition variation (like gene 5 in Figure 3-1). I hypothesized that some of the coexpression signal among genes from these clusters could have remained obscured due to the prevalence of high correlation values induced by cellular composition variation among other genes. To investigate this, I compared counts of links in different clusters in GTExBulk with counts of links in the GTExBulk residual network (GTEx_residual, a network built from the residuals of the PCR fits). I observed large drops in the count of links in the marker-enriched clusters and an increase in the count of links in clusters with low R² values (Figure 3-3E). The magnitude of these changes highlight how cellular composition-induced coexpression can mask underlying coexpression within cell types. I discuss the use of the GTEx_residual network as a “corrected network” in the next section.
Figure 3-3 Much coexpression is explained by cellular composition effects
(A) Coexpression clusters from GTExBulk dataset network. Clusters are labeled with their IDs. Color indicates if the cluster has markers of a specific cell type, or if it is enriched with housekeeping functional terms or genes. Thickness of the border reflects the mean $R^2$ value for genes in the cluster. (B) Coexpression of mean CT expression profiles for a group of clusters affiliated with Pyramidal cell expression patterns and a group of clusters affiliated with Astrocyte cell expression patterns (affiliation is indicated with presence of markers, high $R^2$ or high inter cluster links) (C) Results from simulated bulk tissue data. Each dot represents data from a pair of genes. Plot shows data for 1000 gene pairs, sampled from a bulk tissue dataset with 100 samples and 10 hypothetical cell types. As demonstrated, for a given pair of genes, their Pearson correlation in the bulk tissue is highly correlated with the Pearson correlation between their CT expression profile. Also the higher the correlation between their CT expression profiles, the more likely their correlation in the bulk tissue is the same as the correlation of their CT expression profiles. (D) Proportion of coexpression involving the set of genes from the clusters enriched with marker genes. (D1) The two brain networks (GTExBulk and the TAN-brain) and the brain-specific network (TSN-brain) have between ~30-50% intra-cluster links in clusters enriched with marker genes. (D2) Portion of links in the same set of clusters in two groups of synthesized bulk tissue networks, modeling the effect of cellular composition variation. (E) Percentage of links increased or decreased for GTExBulk clusters in the residual network (GTEx_residual). Cluster color code as in A. The range for average $R^2$ is shown in the color bar. Many housekeeping clusters (orange) with low average $R^2$ values yield more links in the residual network.

3.3.3 Cellular composition effects can mask the intra-cell-type coexpression

I have shown that a major coexpression signal in bulk tissue comes from cellular composition effects. In my view this presents a shift from the usual interpretation, and raises the question of whether there is substantial coexpression attributable to other sources. This is especially relevant to attempts to infer coregulation. Specifically, the question remains as to whether
coregulatory relationships in the sense typically sought are “visible” in bulk tissue data in the background of cellular composition effects. I do not attempt to fully address this question here, instead concern myself with a simpler one: in the common modes of coexpression analysis of bulk brain tissue, are coexpression patterns present within a cell type detectable? Composition-induced coexpression could in principle mask or amplify the bulk-tissue visibility of intra-cell-type coexpression. In this section, I examine the difference between robust intra-cell-type coexpression as measured in the snuc-RNAseq data and the observed coexpression in GTExBulk, and show that much of the difference can be attributed to the cellular composition effect. I also examine the GTEx_residual network as a form of “corrected network” in retrieving intra-cell-type coexpression.

As a preliminary step, I examined the general agreement of the coexpression networks built from different snuc-RNAseq populations and the GTExBulk dataset, and found that the agreement of network links is up to two times (and in few cases 3-5 times) higher for most of snuc-RNAseq populations than that expected by chance (see Figure B-6). For reference, for the two bulk brain networks TAN-brain and GTExBulk this measure of agreement is 14. Conversely, some of the observed coexpression clusters in the GTExBulk are also reproducible in the larger snuc-RNAseq cell populations (see Figure B-7), including those of many of the housekeeping and some of the brain-specific clusters. This shows that there is some level of agreement between coexpression observed in snuc-RNAseq and bulk data, in agreement with prior work 158.
I then hypothesized that some of the differences between the observed coexpression in snuc-RNAseq data and GTEx bulk could be explained by the cellular composition effect, in a way that is shown schematically in Figure 3-4A. To test this, I compared bulk tissue coexpression with robust intra-cell-type coexpression patterns and found that for the most part, differences between the two are explained by the effect of cellular composition variation. To identify robust intra-cell-type coexpression patterns, I combined 64 snuc-RNAseq coexpression networks built from neuronal cell types (both excitatory and inhibitory) and obtained a consensus (sum) “intra-cell-type” network (see Section 3.2.2 and Additional file B5 for list of links). I focused on the highest-confidence set of links that were present in 10 or more networks, leading to a set of 464 genes with 7,678 highly robust links. Of these links, 32% have correlation values above the 95th quantile in the GTExBulk network but only 63% have correlation values above the median. This sub-network forms two very distinct clusters (Figure 3-4B). The gray cluster is enriched with multiple functional terms associated with neuronal processes and the black cluster is enriched with a few house-keeping functions (see the complete list of functions in Additional file B6). I then looked at the coexpression of these genes in the GTExBulk network (Figure 3-4C). While the genes in the two clusters have partly distinguishable coexpression patterns in GTExBulk, a large number of inter-cluster links are present; that is, the clusters are not as clearly separated. Indeed, the 464 genes appear in multiple clusters in the GTExBulk network (Figure 3-4D). I hypothesized that this is due to similarity of CT expression profiles among some of the genes, causing composition-induced coexpression that “blurs” the underlying cell-type-specific coexpression pattern. In support of
this hypothesis, correlations of the CT profiles are high for some of the genes (Figure 3-4C and D). In particular, this can explain the differentiation between the two clusters ID253, ID168 and Excitatory cell clusters (indicated by green color bar in Figure 3-4D - these are the clusters enriched with Pyramidal markers – see Figure 3-3 for reference). The differentiation is even clearer when CT expression profiles are obtained from neuronal cells only (Figure 3-4E), indicating different expression patterns among neuronal cell types for genes in clusters ID253, ID168, ID64 and Excitatory clusters (Figure 3-4E). My conclusion is that the intra-cell-type coexpression patterns observed in single cell data can be distorted and/or masked in bulk tissue by the effects of cellular composition.
The intra cell type clusters could be divided into separate clusters in the bulk tissue network based on different modes of genes' CT profiles.

Figure 3-4 Cellular composition can mask intra-cell-type coexpression
(A) Schematic showing a coexpression cluster in a specific cell type could be divided into multiple clusters in the bulk tissue dataset, as its genes might have different CT expression profiles. Each circle represents a group of genes. Colors blue, yellow and coral represent different “modes” of CT expression profiles, similar to the mean CT expression profiles for the bulk tissue clusters in Figure 3. (B) The heatmap shows part of the sum network from 64 neuronal snuc-RNAseq datasets where two coexpression clusters are identified. Clusters 1 and 2 (color bar grey and black) are well distinguished from each other. (C) Heat map shows the same set of genes with the same order in the GTExBulk tissue dataset. While the two clusters are somewhat distinguished, a great amount of inter-cluster links is present. (D) The heatmap shows the network for the genes in cluster 1, from the coexpression network built from the correlation of the CT expression profiles obtained from the 75 snuc-RNAseq datasets. Genes are ordered based on their belonging to different GTExBulk clusters, identified by the colorbar and cluster IDs from Figure 3A. Three sub-clusters are mildly distinguished, separating two groups of housekeeping clusters from the Pyramidal clusters (orange versus green). (E) Same plot as D, but the CT expression profiles are obtained from the 64 neuronal cell types only. The distinguished clusters demonstrate the group of genes with different expression levels in the neuronal cell types.

In the previous analysis, I showed that cellular composition effects can mask intra-cell-type coexpression especially when there is a conflict between the correlation of the CT expression profiles and the intra-cell-type coexpression, resulting in loss of the intra-cell-type pattern. In general, there are various scenarios that could occur, and intra-cell-type coexpression patterns might happen to be observed in bulk tissue to varying degrees and for varying reasons. Here I demonstrate this complexity with two genes, CALM3 and NRGN (Figure 3-5). They are robustly correlated in the snRNA-seq excitatory neurons (a link is present in 11 out of 23 of the networks for excitatory neurons), but there is no correlation between them in Inhibitory
neurons, since NRGN is not expressed in inhibitory neurons (Figure 3-5A, B). Accordingly, they have relatively highly correlated CT expression profiles (rho = 0.46), driven by their high expression in excitatory neurons and close to zero expression in the non-neuronal cell types, but moderated by their disjoint expression in inhibitory neurons. This suggests they might be coexpressed in bulk tissue, but for a reason different from that driving their coexpression within excitatory cells. As it happens, their correlation in bulk tissue is relatively high (97th quantile), but not nearly high enough to pass my original 99.5 quantile filter for link selection. Their correlation ranks drop to the 82.3th quantile in the GTEx_residual network. I conclude that the observed coexpression of CALM3 and NRGN in the bulk tissue is primarily caused by correlation of their cell type expression profiles, rather than a reflection of their coexpression in excitatory cells. Also, although their coexpression in the bulk tissue resembles their coexpression in excitatory cells, it is in disagreement with their lack of coexpression in other non-neuronal cell types. In general there is no simple relationship between coexpression within a cell type, and coexpression in a tissue in which that cell type is one of several present.
Figure 3-5 Coexpression of NRGN and CALM3 in Excitatory cell types

(A) From snuc-RNaseq data: NRGN is only expressed in the Excitatory cell types (B) From snuc-RNaseq data: CALM3 is expressed in both Inhibitory and Excitatory cell types (C) The two genes are highly correlated in the Excitatory cell types – based on coexpression networks built from snuc-RNaseq data. They are correlated in GTExBulk dataset but do not meet the threshold for the network (threshold is marked by the red line, it is the 995\textsuperscript{th} quantile).

Given that the effects of cellular composition on coexpression can be viewed as a confound, it is natural to consider whether the data can be corrected. In the previous section we observed that many of the clusters from GTExBulk had a much higher count of links in the residual network. In this framework a natural choice for such a correction are the residuals from my PCR model fits used to obtain the R\textsuperscript{2} estimates. I observe that most of the GTExBulk clusters are significantly reproduced in the GTEx_residual network (see Figure B-6) and many of the
brain-specific and housekeeping terms are enriched in the GTEx_residual network (Additional file B7). However, there is no overall significant improvement in agreement of the links in snuc-RNAseq populations with the GTExBulk_residual compared to the GTExBulk network (see Figure B-7), indicating that correction for composition may not be a panacea. But there is improvement of the precision in recovering snuc-RNAseq coexpression links for some of the GTExBulk-driven clusters. Results for the largest snuc-RNAseq population is presented in Figure B-7. As an example we can consider the neuronal clusters with IDs 239, 242 and 243. These clusters are enriched for pyramidal cell markers, and their genes have high mean $R^2$ values of 0.72, 0.74 and 0.75, respectively, suggesting that correction will have a large effect. In GTEx_residual the link density among the genes in these clusters decreases to approximately $1/2$, $1/6$ and $1/14$ of the density in the GTExBulk network, while the link overlap with the snuc-RNAseq cell populations increases 24 to 175 percent. We also see that, for the most robust snuc-RNAseq clusters (from Figure 3-4B), which were substantially degraded in GTExBulk (Figure 3-4C), their separation is largely restored in the GTexBulk_residual network (Figure 3-6). However, the restoration is mostly due to the removal of inter-cluster links, rather than an increase in the intra-cluster cluster links. These findings suggest that correction for cellular composition effects can be beneficial, but confidence in the results are uncertain without matched single-cell data.
Figure 3-6 Reproducibility of robust snuc-RNAseq clusters in GTEx networks

Inter and intra cluster density values are overlaid in black, and fold-changes in blue. The null density is 0.005 for all the networks. **(A)** This is the Sum network as in Figure 3-4B. For visualization, the data are scaled to be comparable to B and C. Intra cluster density is more than 15 times the inter cluster density for both black and gray clusters. **(B)** As in Figure 3-4C. Intra cluster density is less than the inter cluster density for the black cluster, for the gray one it is less than 2. **(C)** Intra cluster density is more than 10 times higher than the inter cluster density for both black and gray clusters. Notice that the dramatic effect in the density (from GTExBulk to GTEx_residual network) is mostly explained by the relative reduction of inter-cluster links rather than an increase in the intra-cluster links.

### 3.4 Discussion

Coexpression in brain tissue is highly reproducible – that is, there are strong patterns of coexpression that are observed in many independent data sets\(^ {165,175}\). The data I present in this chapter suggest that this is mostly due to the reproducibility of cell-type expression patterns, and the pervasive presence of variability in cellular composition between samples (Figure 3-1 and Figure 3-3). I have shown that when coexpression is observed across cell types or tissues, the dominant patterns are due to cell-type or tissue-specificity of expression, and coexpression
is merely a proxy for differential expression across cell types or tissues. While genes which are expressed specifically in one cell type (for example) can be thought of as having a “shared function”, that function is broad, only reflecting the function of that cell type. There is little expectation that function at the level of individual molecular interactions or pathways would be captured: the distinctness of a cell type cannot be fully described by the activity of a single pathway. Likewise, even for these genes their coregulation may reflect the broad epigenetic state of the cell type \(^{176}\), and finer-grained details of co-regulation are unlikely to be easily captured.

I have also shown that cellular composition-induced coexpression can mask apparently robust cell-type specific coexpression patterns (Figure 3-4). Despite this, a remaining question is whether correction for cellular composition would enable more efficient extraction of coregulation. For this to be the case, underlying patterns due to coregulation would have to be present in the data, and sufficiently separable from cellular proportion effects. For this to be effective, the regulation should ideally not be cell-type specific (otherwise the signal would be that much weaker; Figure 3-6), so the genes involved would have to be expressed in most cells. Since genes which are not cell-type specific tend to have housekeeping functions, it stands to reason that the most apparent coregulatory relationships would be those among housekeeping genes. I note that schemes for correcting bulk tissue data for cell type proportions (either directly or indirectly) are often used in expression QTL studies, and have been shown to increase the number of cis-eQTLs that can be recovered \(^{170}\). This suggests that correcting for cell type proportions and recovering underlying biological
signals is possible, but eQTL studies require large sample sizes (generally at least 100 but often far more, especially for trans-eQTLs). I expect that identification of coregulatory relationships from bulk tissue data will similarly require very large samples sizes and still be most effective at extracting regulation of housekeeping genes rather than cell-type specific genes. Giving these constraints, it would seem preferable to use coexpression data from a single cell type to extract regulatory relations. However, limitations of the most commonly-used single-cell transcriptome methods – discussed in Section 1.3- suggest extracting high-quality regulation information is a challenge. Furthermore, the most commonly used computational method for doing so is designed to simultaneously identify cell types along with building a regulation network, so that the strongest patterns observed are likely dominated by differential expression across cell types, not coexpression within cell types.

This study does have some limitations. First, the analysis of cell-type-level coexpression is based on a single (albeit large and unusually deeply sequenced) data set that used different samples than the bulk tissue. Thus I cannot rule out that the failure to recover some snuc-RNAseq coexpression patterns in bulk tissue might reflect data-specific effects. This might be resolved in the future with additional data sets. Second, I only considered the phenomenon in brain. Intuitively, cellular composition effects should impact any bulk tissue coexpression analysis, but determining whether the inferred effects in other tissues are weaker or stronger than those I observe for brain should be a topic of future research. Finally, the actual cellular composition of the bulk tissue samples I used is not known. While the approach of using cell-type markers to infer composition has been validated many times
I do not claim it is a perfect substitute for accurate direct counts. It remains formally possible that some of the variation I attribute to cellular composition is instead due to complex patterns of gene regulation that mimic compositional effects, but I feel the most parsimonious interpretation of the data is that cellular composition is a major contributor. It is also worth noting that imperfect cell-type-effect measurement could just as easily cause us to underestimate the impact of composition, as the residual would still contain compositional effects.

Beyond the implications for the goal of inferring regulation, my results have important implications for any use of expression data-based gene clustering or module identification in which the patterns are driven by cellular composition effects. First, the representation of the data as a network is potentially misleading, because it is tempting to interpret a network as representing physical relationships. In particular, the idea that “hubs” in coexpression models are especially interesting is highly questionable if that pattern is simply a reflection of the cellular distribution of those transcripts. Second, if cellular composition is of interest, it would be reasonable to analyze composition more directly by inspecting the expression of known markers rather than by using indirect means via clustering and enrichment analysis. This parallels the situation for analysis of differential expression, where changes in measured expression levels can be due to changes in composition. On the other hand, machine learning applications of coexpression to tasks such as gene function prediction are not directly affected by these findings, as success in prediction does not necessarily depend on the biological meaning of the features used.
Chapter 4: Conclusion

The term “coexpression” refers to two tightly linked concepts. The first is defined in the realm of molecular biology as the coordinated transcription of genes by regulatory mechanisms occurring within a cell (i.e., coregulation). The other refers to the observed correlation of RNA levels among the genes in high-throughput transcriptomic datasets, which I refer to as the “observed coexpression”. For decades, coregulation has formed a foundation for understanding genome function: That genes with collaborating products need to be expressed at the same time and are thus coregulated. Meanwhile, coexpression detected in high-throughput datasets (i.e., the observed coexpression) has proven to be a reproducible signal with biological relevance: Genes that are coexpressed have a higher probability of having related functions than those which are not coexpressed. It is often assumed that the observed coexpression is due to the regulatory mechanisms and therefore the latter could be reverse engineered from the former. This assumption is not limited to the coregulatory relationships and in fact is more common for inferring causal regulatory relationships, such as TF-target.

Coexpression analysis has been around for more than two decades and has been implemented in different frameworks on countless transcriptomic datasets, but has rarely resulted in the discovery or recovery of regulatory relationships. There seems to be a gap in this scenario. On one hand, the widely reported signals in coexpression networks with biological relevance, such as hubs and gene clusters, have resulted in the assumption that transcript regulation is behind the observed coexpression. On the other hand, efforts for inferring regulatory relationships from coexpression has remained unsuccessful. In my thesis,
I tried to address this gap from both sides. In the first project (Chapter 2), I examined the potentials and limitations of coexpression in representing regulatory relationships and in the second project (Chapter 3), I studied an alternative explanation for the observed, biologically relevant signal in coexpression. Together, these studies highlighted critical limitations to the interpretation of coexpression as commonly practiced in systems biology.

In Chapter 2, I first examined the resolution of coexpression by studying differential coexpression among five tissues. By collecting several datasets and building an aggregated network for each tissue, I was able to identify robust tissue-specific coexpression links. I further showed that these links are highly reproducible in external datasets. Next, I examined coexpression for its potential in discovering regulatory relationships. The lack of gold standards for regulatory relationships limits such examination. Therefore, I addressed the question by studying “pure” differential coexpression links (i.e., differential coexpression in the absence of differential expression). A common motivation for DCA is to infer regulatory rewiring in the form of pure links. Assuming that coexpression represents regulatory relationships, we should be able to identify a large number of pure links in coexpression networks built from diverse contexts such as tissues. Upon my examination, I found that up to 75% of the observed differential coexpression is substantially explained by the average expression levels of the genes. This finding represents the limited information content of coexpression with regard to regulatory relationships.

Also, reporting differential coexpression without considering the role of differential expression has resulted in misinterpretation. For example, the presence of context-specific
coexpression clusters has fed into the idea that coexpression represents a biological event—that is, context-specific transcript regulation. However, these clusters are commonly enriched with context-specific genes, meaning that much of the observed coexpression is accompanied by differential expression. While context-specific expression is ultimately a result of context-specific regulation, the leap to the conclusion that “differential coexpression links represent regulatory rewiring” is unjustified. Furthermore, the results from Chapter 3 showed that across the tissues, the context-specific coexpression clusters themselves are likely to be a result of the cellular composition effect, which reflects differential expression at the cell-type level.

Results from Chapter 2 showed that the brain datasets had a higher count of tissue-specific links compared to the other four tissues. This is not the result of technical variations, as brain links showed even higher measurements in reproducibility in external datasets. Therefore, it hinted to the presence of a biological signal. Following up on this finding, in Chapter 3, I studied the role of cellular composition on the observed bulk tissue coexpression in the brain, hypothesizing that it could be a source of highly reproducible and biologically relevant coexpression. It is important to note that this effect is likely to be present in other tissues as well, and the reason I studied the brain was due to the available sources and the likely higher prevalence of the effect.

In this study, I demonstrated that genes with similar cell-type expression profiles tend to be coexpressed among the bulk tissue samples. Furthermore, I showed that many clusters in the bulk brain coexpression network are mostly composed of quasi marker genes and as a
result, these clusters are enriched with cell-type-specific functional terms. In short, I found a robust coexpression signal which is generated from a biological source (i.e., cell-type specific expression patterns of the genes) combined by a common experimental variation (i.e., cellular composition variation among the bulk tissue samples). This signal results in biologically relevant findings in coexpression analyses (i.e., enrichment of cell-type specific functional terms in the coexpression clusters). However, the biological event behind this signal is not transcript regulation in the commonly assumed scenario.

Cellular composition variation is likely to be present in transcriptomic datasets from any complex tissue \(^{171,179,180}\). Therefore, an important question to raise is if bulk tissue coexpression is still relevant for transcriptomic studies—more specifically, if there is any coexpression signal in the bulk tissue unaffected by cellular composition and, furthermore, if bulk tissue transcriptomic data could be corrected for the cellular composition effect. I tried to answer these questions in the brain. As demonstrated in Section 3.3.1, cellular composition variation affects genes differently. Clearly, genes with similar expression levels among different cell types have less compositional induced variation—as opposed to genes with varying expression levels among cell types. Therefore, it is possible to observe regulatory induced coexpression among the less-affected genes in the bulk tissue dataset. A well-known (and relatively boring) example of this is the coexpression of ribosome genes. Ribosome genes are known to be coregulated \(^{164}\) and are widely coexpressed. The results from Chapter 3 showed that they are among the genes least affected by cellular composition. Also, as shown in Section 3.3.3, after using a simple correction method on the bulk transcriptomic dataset,
coexpression links showed higher agreement with the robust coexpression links from the snuc-RNAseq data.

Nonetheless, it is valuable to examine the effect of cellular composition in other complex tissues. This could be a subject of future work. The methods implemented in Chapter 3 are not specific to the brain tissue or datasets used in this work, and are applicable to similar studies in other tissues. Ideally, a high-quality single-cell dataset could be employed to identify the cell-type marker genes and obtain both the cell-type expression profiles and the intra-cell-type coexpression networks. However, it is also possible to examine the effect of cellular composition using only the cell-type marker genes. In that case, parameters from the regression model could be substituted for the cell-type expression profiles.

In the two research projects presented in this thesis, I have studied coexpression as a signal from the data in my attempt to understand why coexpression analysis has failed to meet the expectations for discovering or recovering regulatory relationships. I have been able to provide some answers, but what led to this question in the first place reflects a major problem with the analytical approaches toward coexpression. A major problem here is assuming that an observed signal from the data can be explained by specific biological events without detailed examination, and then trying to infer these events from the data without enough evidence that the signal is representing them. As I described in Chapter 1, coexpression was originally used to refer to a biological phenomenon which was observed and studied in low-throughput studies. In high-throughput transcriptomic datasets, coexpression is an observation from the data, and while it is highly reproducible and has some biological relevance, it has
rarely been critically examined for its source. Inference from a signal without a clear understanding of its source is not necessarily problematic as long as it can be shown that the signal carries the information we are interested in. An example of such an approach is the application of coexpression in function prediction studies. It is apparent that coexpression contains information on the functional similarities among genes; therefore, the signal could be harvested for this information, and methods could be adjusted to improve this inference. However, such evidence is lacking for regulatory relationships. This is partially due to the lack of gold standards, but also findings from currently available knowledge are not promising. Without evidence, the causal role of regulatory relationships for the observed coexpression remains a theory, and it continuing to be an assumption is misleading and can result in the misinterpretation of the signal—as has been demonstrated in Chapter 3.

When making a series of reproducible observations, two approaches can be employed: The first is to examine what the signal represents and harvest its relevant content, and the second is to examine the signal’s source so it could be reverse engineered to decipher the underlying events. In the case of coexpression analysis, the results of early-day clustering analyses, which many transcriptomic datasets have revisited and confirmed afterwards, have suggested that coexpression is a robust signal with biological relevance. Studies on functional inference have employed the former approach; they tried to harvest coexpression patterns to predict gene function. However, many other studies on coexpression stop at reporting the existence of a coexpression signal with biological relevance and do not provide any examination or validation for the potential underlying events. For example, the existence of
Coexpression clusters enriched with functional terms has been settled in the field and is not particularly a finding, unless some functional inference is reported or examined. As another example, the status of the hubs in a coexpression network has no biological or analytical interpretation, and therefore, does not demonstrate any validation or contribute to any findings.

Coexpression is interesting in the context of deciphering gene regulatory networks, one of the most complicated problems in systems biology. However, this is only possible through a clear understanding of the data and the information it carries. My effort in this thesis has been to take a step toward this direction and provide a more promising trajectory for the future application of coexpression in regulatory inference analysis. Today, inferring regulatory relationships from coexpression faces two analytical bottlenecks: data limitations and the lack of gold standards – as discussed in Section 1.6. Advances in technology and the development of high-quality single-cell datasets imply a more promising future for such applications. Combined with gold standards obtained through perturbation studies, the application of sophisticated machine learning methods for such inference could be possible.
Bibliography


65. Ng, B. *et al.* An xQTL map integrates the genetic architecture of the human brain’s transcriptome and epigenome. *Nat Neurosci advance online publication*, (2017).


105. Stuart, J. M., Segal, E., Koller, D. & Kim, S. K. A gene-coexpression network for global

106. McCarroll, S. A. *et al.* Comparing genomic expression patterns across species identifies

107. Mariño-Ramírez, L., Hsu, B., Baxevanis, A. D. & Landsman, D. The Histone Database:
A Comprehensive Resource for Histones and Histone Fold-Containing Proteins.

108. Allison, D. B., Cui, X., Page, G. P. & Sabripour, M. Microarray data analysis: from

109. D’haeseleer{"id":1324531, P. *et al.* Tracing genetic information flow from gene
expression to pathways and molecular networks.

110. Mootha, V. K. *et al.* PGC-1α-responsive genes involved in oxidative phosphorylation are


112. Pavlidis, P., Lewis, D. P. & Noble, W. S. Exploring gene expression data with class


120. Mabbott, N. A. & Gray, D. Identification of Co-Expressed Gene Signatures in Mouse B1, Marginal Zone and B2 B-cell Populations. *Immunology* n/a–n/a (2013)


166. Kelley, K. W., Nakao-Inoue, H., Molofsky, A. V. & Oldham, M. C. Variation among
intact tissue samples reveals the core transcriptional features of human CNS cell classes.

167. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression

168. Patrick, E. *et al.* Deconvolving the contributions of cell-type heterogeneity on cortical

e1005223 (2015).

170. Ng, B. *et al.* An xQTL map integrates the genetic architecture of the human brain’s

171. Mancarci, B. O. *et al.* Cross-Laboratory Analysis of Brain Cell Type Transcriptomes


173. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and


Appendices

Appendix A  Supplementary Materials for Chapter 2

A.1  Disease annotation for the genes

The Phenocarta data was extracted on 23.04.2018. The total count of annotations was 26,499, among 6,400 genes.

A.2  Tissue specificity score for GTEx data

To examine the reproducibility of the pure links as tissue-specific links in the GTEx datasets, I define TSS for a given pair of genes \((g_i, g_j)\) and a tissue \(t\) from the five GTEx tissues: blood, brain-cortex, liver, lung and skeletal muscle the same as TSS for the TAN links. The only difference is that for GTEx I have just one coexpression value for each of the tissues. From total count of 506,221 tissue-specific links identified from TANs, I found that 263,302 (52%) have TSS > 0.56 in GTEx. The score 0.56 is the average of the TSS cut-off thresholds from the five tissues previously identified from Affymetrix datasets. From the total count of 80,637 pure links identified from TANs, I found that 20,178 (25%) have TSS > 0.56 in GTEx. Table files in the supplementary data folder TSLinks_TSS have the GTEx TSS for each of the tissue specific links, as well as the expression level of the paired genes in all the tissues.

A.3  Examining topological overlap for partners of pure links

I have identified 80,637 pure links in total across the five tissues (see Table A-8 for a breakdown per tissue). For a given gene and network, I define the neighborhood of the gene in that
network as the set of genes it is connected to. Based on the GBA principle, the neighbourhood of a gene in a coexpression network could suggest its functional role. Therefore, extreme changes in the neighborhood of a gene between the networks could suggest different functionality for the gene in those networks – and thus a different function for the gene in the tissues the networks are built from. Also, for a given pair of genes in a network, the similarity of their neighborhoods can be measured as their Topological overlap (TOP, \(^{182}\)). This similarity (high overlap) could also suggest functional similarity between the gene pair. I applied these ideas to examining if the presence of a pure link in a target TAN suggests higher TOP between the gene pair in that TAN, compared to the other TANs where the pure link is absent. For comparison and as a null, I also examined this for gene pairs with coexpression links present in multiple tissues: for a given pair of genes that have coexpression links in multiple TANs, how much does their TOP change between these TANs, and if the observed change is as extreme as what is observed for the gene pairs with a pure link. The results from this part shows how much we can learn about a gene pair in the absence/presence of a pure differential coexpression link between them.

I used the Jaccard index for measuring the TOP for each gene pair, as the count of shared neighbors (intersect of the neighbors) to the total count of unique neighbors (union). Each gene pair \((g_i, g_j)\) which has a pure link in a target tissue gets five TOP measurements, called TOPJCCs, one from the target tissue TAN and four from the other tissue TANs. For simplicity, for each link I selected the maximum TOPJCC from the other tissues for the
comparisons. Figure A-5 shows the distribution of TOPJCC in the target tissues, compared to those from the maximum in the other tissues.

As explained, for the null I used gene pairs that have links in multiple TANs. I call these multi-tissue-links. I identified 368,411 gene pairs as such, which have coexpression links in two or more TANs. By definition, these gene pairs do not have tissue specific links and therefore the “target” and “other” tissues are not defined for them. To be able to compare their TOPJCC to those from the pure links, for each gene pair I marked the tissue with the highest TOPJCC the “target” and the tissue with the next highest TOPJCC as the “other” tissue. The plot in Figure A-5 shows that between the two genes that are expressed in multiple tissues, the presence of a pure link suggests higher TOPJCC in target tissue and lower TOPJCC in the other tissues (median 0.29 versus 0). This difference is not as extreme for the multi-tissue-links and is a feature unique to the pure link as a differential coexpression link. The difference in TOP overlap could suggest a change between functional similarity of the gene pairs in the target tissue compared to the other tissues for the pure links.

### A.4 Contribution of the pure links to tissue-specific enrichment of functional terms in TSNs

For a given GO term \( f \) and a network \( N \), I mark the term \( f_i \) as enriched in \( N \) if genes annotated with that term have significantly more links between them than the null, where null is the binomial distribution and the density of network \( N \) is the success probability. This was done for all the TSNs and TANs and for all the terms that had \( \geq 20 \) genes annotated with them and
also expressed in that tissue (Table file enrichedTerms_TAN_TSN_table.tsv). I hypothesized that for a given network, some of the functions would lose their enrichment significance if the pure links were removed, and this could further show which cellular processes are affected by the pure rewiring between the networks. I examined this for my TSNs in two steps: first I identified the functions which were exclusively enriched in each of my TSNs. I found that 146, 91 and 43 functions are enriched exclusively in the brain, liver and lung TSNs respectively (the count of enriched functions for blood and skeletal muscle was too low for further analysis). Second, I removed the pure links and measured the significance of the enrichment for the same terms to identify the affected functional terms. When pure links were removed, the count of functions in brain dropped by 20, by 3 in liver and by 4 in lung. As a null comparison for each tissue, I removed sets of randomly selected tissue-specific links (the same count as the pure links) from each TSN, repeated 100 times. The count of drop-out functions for brain and lung (upon removal of the pure links) is less than that generated by removing random sets of links (p < 0.05). However, five functions in brain are specifically sensitive to the removal of pure links and not sensitive to the removal of random sets of links and expression-induced links (See table file brain_TSExclusvieFunctionTerms_linkRemoval.txt, function IDs: 10821, 32388, 6650, 8654, 10506, Column: significantCountIn100_random-PureLinkRemoval, showing the count of times the function remained significant after removal of random sets of links). These functions only lost their significance in < 4 times out of 100 trials of random link removal.
A.5 Contribution of expression-induced links to the tissue-specific enrichment of functional terms in TSNs

In each TSN, links were marked as expression-induced if one or two of their genes were marked as exclusively expressed only in that tissue (see Supplementary Tables 11 and 12 for count of the genes and expression-induced links). The count of expression-induced links is much higher than the pure links in brain, liver and lung and therefore I hypothesized that they have larger contribution to the tissue-specific enrichment of functional terms than that of the pure links. I examined their contribution in two ways: 1. I compared the effect of removing expression-induced links to that from removal of the pure links, 2. I compared the removal of expression-induced links to a null.

For comparing removal of pure links to expression-induced links in brain, liver and lung, I selected random subsets of expression-induced links with the same count as the pure links and removed them from the network. I found that the count of drop-out functions upon the removal of subsets of expression-induced links is significantly higher than that of the pure links (100 trials, p < 0.01). The comparison to the null was similar to what I did for the pure links: I recorded the count of drop-out functions from the removal of the expression-induced links and compared it to the removal of random sets of links to the same count as the expression-induced links. When the expression-induced links were removed from the TSNs, the count of functions dropped by 67 in brain, 45 in lung and 27 in liver. Some of these functions were not affected by removal of random sets of links – their tissue-specific enrichment was primarily due to expression-induced links (see table files.
significantCountIn100_randomEILinkRemoval, showing count of times the function remained significant after removal of random sets of links).

A.6 Enrichment of the disease ontology terms in the networks

I also examined the enrichment of the disease terms in the networks similar to the functional terms. I had 18 and 30 terms exclusively enriched in brain and lung TSNs (the count of terms enriched in other tissues was too small for further analysis and interpretation). I found that only 1 term in brain and 5 in liver are sensitive to the removal of pure links, but the same terms were also sensitive to the removal of random sets of links (see table files [tissue]_TSExclusiveDiseaseTerms_linkRemoval.txt)
A.7 Figures

Figure A-1 Count of samples and genes for each tissue

(A) Distribution of the count of samples for different datasets in each of the tissues. The lighter and darker colors mark the count of samples for different datasets. (B) Overlap of the genes marked as expressed in the five tissues. (C) Overlap of the genes with any edges in the relevant TAN networks. Venn diagram built in

http://bioinformatics.psb.ugent.be/webtools/Venn/
Proportion of coexpression links having one or two of their genes marked as “not expressed” in the datasets. Each box plot represents the portion of the links for each of the 53 datasets at a correlation cut-off threshold.
Comparison of the TSS and p-values from Wilcoxon-rank-sum test. Scatter plots are the unadjusted $-\log_{10}(p$-values$)$ of the Wilcoxon Rank Sum test and the TSS. Red lines are at the cut-off threshold for FDR $< 0.01$. Results from the two methods are highly correlated.
Reproducibility of the TAN, TSN and pure links for GTEx datasets blood, liver, lung and skeletal muscle (plot for GTEx brain-cortex is in Figure 2-3).
Distribution of topological overlap for different groups of gene pairs that are expressed in multiple tissues. Left: distribution of TOP for partners of pure links in the target tissue, versus the maximum TOP from the other tissues. Right: distribution of TOP for gene pairs with coexpression links in multiple tissues. The tissue with highest TOP was considered as “target” and the maximum TOP after that was considered as the “other” tissue. While for the pure differential coexpression links the TOP from “other tissue” is close to zero for almost all the links, for a multi-tissue coexpression link the TOP from the other tissue is above zero for more than half of the links.
Additional examples of pure links (randomly selected). As in Figure 2-1 in the main text, for each link, Pearson correlation values and normalized gene expression values for each of the datasets is plotted. The target tissue is marked with the black box.
### A.8 Tables

#### Table A-1 Affymetrix blood datasets

<table>
<thead>
<tr>
<th>GEO ID</th>
<th>#Samples used/ #Initial Samples</th>
<th>#Samples removed in batch effect/ as outlier</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE13849</td>
<td>120/110</td>
<td>10/0</td>
<td>-</td>
</tr>
<tr>
<td>GSE16028</td>
<td>105/109</td>
<td>1/3</td>
<td>-</td>
</tr>
<tr>
<td>GSE19314</td>
<td>60/65</td>
<td>0/5</td>
<td>-</td>
</tr>
<tr>
<td>GSE25507</td>
<td>143/145</td>
<td>1/1</td>
<td>-</td>
</tr>
<tr>
<td>GSE26050</td>
<td>29/33</td>
<td>1/3</td>
<td>-</td>
</tr>
<tr>
<td>GSE27263</td>
<td>52/92</td>
<td>0/0</td>
<td>40 samples from nasal biopsy removed</td>
</tr>
<tr>
<td>GSE27562</td>
<td>159/162</td>
<td>0/3</td>
<td>-</td>
</tr>
<tr>
<td>GSE33223</td>
<td>29/30</td>
<td>0/1</td>
<td>-</td>
</tr>
<tr>
<td>GSE7753</td>
<td>31/31</td>
<td>0/0</td>
<td>-</td>
</tr>
</tbody>
</table>
Table A- 2 Affymetrix brain datasets

<table>
<thead>
<tr>
<th>GEO ID</th>
<th>#Samples used/ #Initial Samples</th>
<th>#Samples removed in batch effect/ as outlier</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE11882</td>
<td>172/173</td>
<td>0/1</td>
<td>-</td>
</tr>
<tr>
<td>GSE13564</td>
<td>37/44</td>
<td>2/5</td>
<td>-</td>
</tr>
<tr>
<td>GSE17612</td>
<td>45/51</td>
<td>0/6</td>
<td>-</td>
</tr>
<tr>
<td>GSE20146</td>
<td>16/20</td>
<td>0/4</td>
<td>-</td>
</tr>
<tr>
<td>GSE28160</td>
<td>31/38</td>
<td>7/0</td>
<td>-</td>
</tr>
<tr>
<td>GSE3526</td>
<td>145/353</td>
<td>0/2</td>
<td>Samples from different tissues</td>
</tr>
<tr>
<td>GSE35864</td>
<td>68/72</td>
<td>0/4</td>
<td>-</td>
</tr>
<tr>
<td>GSE4036</td>
<td>26/28</td>
<td>0/2</td>
<td>-</td>
</tr>
<tr>
<td>GSE4757</td>
<td>20/20</td>
<td>0/0</td>
<td>-</td>
</tr>
<tr>
<td>GSE5281</td>
<td>160/161</td>
<td>1/0</td>
<td>-</td>
</tr>
<tr>
<td>GSE7621</td>
<td>18/25</td>
<td>7/0</td>
<td>-</td>
</tr>
<tr>
<td>GSE9770</td>
<td>34/34</td>
<td>0/0</td>
<td>-</td>
</tr>
</tbody>
</table>
### Table A- 3 Affymetrix liver datasets

<table>
<thead>
<tr>
<th>GEO ID</th>
<th>#Samples used/ #Initial Samples</th>
<th>#Samples removed in batch effect/ as outlier</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE12720</td>
<td>63/63</td>
<td>0/0</td>
<td>-</td>
</tr>
<tr>
<td>GSE14668</td>
<td>27/28</td>
<td>0/1</td>
<td>-</td>
</tr>
<tr>
<td>GSE15235</td>
<td>43/47</td>
<td>2/2</td>
<td>-</td>
</tr>
<tr>
<td>GSE17183</td>
<td>104/108</td>
<td>2/2</td>
<td>-</td>
</tr>
<tr>
<td>GSE19665</td>
<td>20/20</td>
<td>0/0</td>
<td>-</td>
</tr>
<tr>
<td>GSE26622</td>
<td>20/24</td>
<td>0/4</td>
<td>-</td>
</tr>
<tr>
<td>GSE28619</td>
<td>21/22</td>
<td>0/1</td>
<td>-</td>
</tr>
<tr>
<td>GSE38663</td>
<td>49/52</td>
<td>0/3</td>
<td>-</td>
</tr>
<tr>
<td>GSE40873</td>
<td>28/49</td>
<td>?/3</td>
<td>The rest of samples seem to have dropped at batch effect</td>
</tr>
<tr>
<td>GSE6764</td>
<td>68/75</td>
<td>3/4</td>
<td>-</td>
</tr>
<tr>
<td>GEO ID</td>
<td>#Samples used/ #Initial Samples</td>
<td>#Samples removed in batch effect/ as outlier</td>
<td>Notes</td>
</tr>
<tr>
<td>------------</td>
<td>--------------------------------</td>
<td>---------------------------------------------</td>
<td>--------------------------------------------</td>
</tr>
<tr>
<td>GSE10245</td>
<td>53/58</td>
<td>0/3</td>
<td></td>
</tr>
<tr>
<td>GSE10445</td>
<td>70/72</td>
<td>2/0</td>
<td></td>
</tr>
<tr>
<td>GSE11729</td>
<td>59/60</td>
<td>0/1</td>
<td></td>
</tr>
<tr>
<td>GSE12667</td>
<td>73/75</td>
<td>0/2</td>
<td></td>
</tr>
<tr>
<td>GSE14334</td>
<td>36/38</td>
<td>0/2</td>
<td></td>
</tr>
<tr>
<td>GSE18842-1</td>
<td>43/45</td>
<td>0/2</td>
<td>The normal samples from GSE18842</td>
</tr>
<tr>
<td>GSE18842-2</td>
<td>45/46</td>
<td>0/1</td>
<td>The tumor samples from GSE18842</td>
</tr>
<tr>
<td>GSE21369</td>
<td>25/29</td>
<td>0/4</td>
<td></td>
</tr>
<tr>
<td>GSE27262-1</td>
<td>19/25</td>
<td>3/3</td>
<td>The healthy samples from GSE27262</td>
</tr>
<tr>
<td>GSE27262-2</td>
<td>21/25</td>
<td>3/1</td>
<td>The tumor samples from GSE27262</td>
</tr>
<tr>
<td>GSE27716</td>
<td>34/40</td>
<td>6/0</td>
<td></td>
</tr>
<tr>
<td>GSE28571</td>
<td>89/100</td>
<td>1/11</td>
<td></td>
</tr>
<tr>
<td>GSE30219</td>
<td>293/307</td>
<td>6/0</td>
<td>The 8 normal samples were removed</td>
</tr>
<tr>
<td>GSE31210</td>
<td>245/246</td>
<td>0/1</td>
<td></td>
</tr>
<tr>
<td>GSE37768</td>
<td>35/37</td>
<td>1/2</td>
<td></td>
</tr>
<tr>
<td>GEO ID</td>
<td>#Samples used/ #Initial Samples</td>
<td>#Samples removed in batch effect/ as outlier</td>
<td>Notes</td>
</tr>
<tr>
<td>----------</td>
<td>---------------------------------</td>
<td>-----------------------------------------------</td>
<td>------------------------------------------------</td>
</tr>
<tr>
<td>GSE13070</td>
<td>247/364</td>
<td>5/5</td>
<td>Samples from adipose tissue removed</td>
</tr>
<tr>
<td>GSE17674</td>
<td>17/62</td>
<td>0/4</td>
<td>44 tumor samples were removed</td>
</tr>
<tr>
<td>GSE19420</td>
<td>38/42</td>
<td>0/4</td>
<td>-</td>
</tr>
<tr>
<td>GSE25462</td>
<td>48/50</td>
<td>0/2</td>
<td>-</td>
</tr>
<tr>
<td>GSE35659</td>
<td>22/24</td>
<td>0/2</td>
<td>-</td>
</tr>
<tr>
<td>GSE47881</td>
<td>85/89</td>
<td>0/4</td>
<td>-</td>
</tr>
<tr>
<td>GSE33223</td>
<td>29/30</td>
<td>0/1</td>
<td>-</td>
</tr>
<tr>
<td>GSE7014</td>
<td>29/30</td>
<td>0/0</td>
<td>6 normal samples were removed</td>
</tr>
</tbody>
</table>

Table A- 6 TAN attributes *ND = node degree

<table>
<thead>
<tr>
<th>attribute</th>
<th>blood</th>
<th>brain</th>
<th>liver</th>
<th>lung</th>
<th>skeletal muscle</th>
</tr>
</thead>
<tbody>
<tr>
<td>#expressed genes</td>
<td>10,356</td>
<td>10,689</td>
<td>11,261</td>
<td>11,362</td>
<td>11,014</td>
</tr>
<tr>
<td>#genes with ND &gt; 0</td>
<td>8,747</td>
<td>8,761</td>
<td>10,227</td>
<td>11,181</td>
<td>8,890</td>
</tr>
<tr>
<td>#links</td>
<td>380,875</td>
<td>1,118,791</td>
<td>496,476</td>
<td>752,298</td>
<td>358,029</td>
</tr>
<tr>
<td>Density</td>
<td>0.007</td>
<td>0.020</td>
<td>0.008</td>
<td>0.012</td>
<td>0.006</td>
</tr>
</tbody>
</table>
### Table A- 7 TSN attributes *ND = node degree

<table>
<thead>
<tr>
<th>attribute</th>
<th>blood</th>
<th>brain</th>
<th>liver</th>
<th>lung</th>
<th>skeletal muscle</th>
</tr>
</thead>
<tbody>
<tr>
<td>#genes with ND &gt; 0</td>
<td>5,948</td>
<td>6,422</td>
<td>4,320</td>
<td>8,140</td>
<td>3,712</td>
</tr>
<tr>
<td>#links</td>
<td>33,594</td>
<td>358,531</td>
<td>41,706</td>
<td>62,209</td>
<td>10,181</td>
</tr>
<tr>
<td>Density</td>
<td>4e-4</td>
<td>4e-3</td>
<td>1e-3</td>
<td>5e-4</td>
<td>4e-4</td>
</tr>
</tbody>
</table>

### Table A- 8 Pure attributes *ND = node degree

<table>
<thead>
<tr>
<th>attribute</th>
<th>blood</th>
<th>brain</th>
<th>liver</th>
<th>lung</th>
<th>skeletal muscle</th>
</tr>
</thead>
<tbody>
<tr>
<td>#genes with ND &gt; 0</td>
<td>3,468</td>
<td>3,876</td>
<td>1,452</td>
<td>3,521</td>
<td>998</td>
</tr>
<tr>
<td>#links</td>
<td>8,476</td>
<td>62,179</td>
<td>2,440</td>
<td>6,620</td>
<td>922</td>
</tr>
<tr>
<td>Density</td>
<td>1.6e-4</td>
<td>1.1e-3</td>
<td>3.8e-5</td>
<td>1e-4</td>
<td>1.5e-5</td>
</tr>
</tbody>
</table>

### Table A- 9 GTEx binary networks attributes *ND = node degree

<table>
<thead>
<tr>
<th>attribute</th>
<th>blood</th>
<th>brain_cortex</th>
<th>liver</th>
<th>lung</th>
<th>skeletal muscle</th>
</tr>
</thead>
<tbody>
<tr>
<td>#expressed Genes</td>
<td>14,089</td>
<td>15,996</td>
<td>15,213</td>
<td>15,882</td>
<td>14,810</td>
</tr>
<tr>
<td>#Genes expressed in Affymetrix dataset</td>
<td>9,978</td>
<td>10,389</td>
<td>10,910</td>
<td>11,129</td>
<td>10,301</td>
</tr>
<tr>
<td>#genes with ND &gt; 0</td>
<td>6,989</td>
<td>7,548</td>
<td>8,100</td>
<td>6,434</td>
<td>7,996</td>
</tr>
<tr>
<td>#links</td>
<td>365,775</td>
<td>1,095,477</td>
<td>484,903</td>
<td>735,392</td>
<td>336,358</td>
</tr>
<tr>
<td>Density</td>
<td>0.007</td>
<td>0.020</td>
<td>0.008</td>
<td>0.012</td>
<td>0.006</td>
</tr>
</tbody>
</table>
Table A- 10 Count of tissue-specific genes, FDR < 0.1, log transformed FC = 0

<table>
<thead>
<tr>
<th>tissue</th>
<th>blood</th>
<th>brain</th>
<th>liver</th>
<th>lung</th>
<th>skeletal muscle</th>
</tr>
</thead>
<tbody>
<tr>
<td>skeletal muscle</td>
<td>254</td>
<td>129</td>
<td>250</td>
<td>114</td>
<td>212</td>
</tr>
</tbody>
</table>

Table A- 11 Count of genes exclusively expressed in each of the tissues

<table>
<thead>
<tr>
<th>tissue</th>
<th>blood</th>
<th>brain</th>
<th>liver</th>
<th>lung</th>
<th>skeletal muscle</th>
</tr>
</thead>
<tbody>
<tr>
<td>blood</td>
<td>354</td>
<td>543</td>
<td>365</td>
<td>275</td>
<td>562</td>
</tr>
</tbody>
</table>

Table A- 12 Count of expression-induced links

<table>
<thead>
<tr>
<th>tissue</th>
<th>blood</th>
<th>brain</th>
<th>liver</th>
<th>lung</th>
<th>skeletal muscle</th>
</tr>
</thead>
<tbody>
<tr>
<td>blood</td>
<td>4,714</td>
<td>125,057</td>
<td>16,787</td>
<td>11,953</td>
<td>4,237</td>
</tr>
</tbody>
</table>

A.9  List of supplementary files and their content

The file enrichedTerms_TAN_TSN_table.tsv has the list of functional terms enriched in TANs and TSNs. There are also 15 table files in three groups. Each group has five files, one for each tissue, identified with the tissue name in the file name.
TSlinksTSS\_[tissue]\_pureLinks\_GTE\_exp.csv (5 files, one for each tissue)

These files have 18 columns and each row represents a TSN links, for which its gene symbols, correlation bins in GTEx and the average log2 expression values of the genes are represented.

**Table A-13 Supplementary file content information**

<table>
<thead>
<tr>
<th>Column</th>
<th>Header</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-2</td>
<td>gene_A, gene_B</td>
<td>Gene symbols</td>
</tr>
<tr>
<td>3-7</td>
<td>[tissue]_corr</td>
<td>Correlation bin for the gene pair from GTEx dataset</td>
</tr>
<tr>
<td>8</td>
<td>GTEx_TSS</td>
<td>TSS from the GTEx, calculated as explained in Supplementary Material section 4</td>
</tr>
<tr>
<td>9-18</td>
<td>gene_[A or B][tissue]</td>
<td>The average log2 expression level for the genes in each of the tissues</td>
</tr>
</tbody>
</table>

[tissue]\_TSExclusiveFunctionTerms\_linkRemoval.txt

These files have 15 columns, each row represents a functional term in GO.

**Table A-14 Supplementary file content information**

<table>
<thead>
<tr>
<th>Column</th>
<th>Header</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>functionID</td>
<td>GO ID of the functional term</td>
</tr>
<tr>
<td>2</td>
<td>functionName</td>
<td>Name of the function in GO</td>
</tr>
<tr>
<td>3</td>
<td>geneCount</td>
<td>Count of genes associated with that term in the TSN of the tissue</td>
</tr>
<tr>
<td>4</td>
<td>linkCount(lc)</td>
<td>Count of links between the genes associated with that term in the TSN of the tissue</td>
</tr>
<tr>
<td>Column</td>
<td>Header</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>--------</td>
<td>-------------</td>
</tr>
<tr>
<td>5</td>
<td>lc_minusPureLinks</td>
<td>This is linkCount(lc) minus pure links from the network</td>
</tr>
<tr>
<td>6</td>
<td>lc_minusExpressionInducedLinks</td>
<td>This is linkCount(lc) minus expression-induced links from the network</td>
</tr>
<tr>
<td>7</td>
<td>1-ratioOfPureLinks</td>
<td>Ratio of non-pure linkCount(lc)</td>
</tr>
<tr>
<td>8</td>
<td>1-RatioOfExpressionInducedLinks</td>
<td>Ratio of non-expression-induced linkCount(lc)</td>
</tr>
<tr>
<td>9</td>
<td>SigAfterPureRemoval</td>
<td>If the term is still significant after removing of the pure links</td>
</tr>
<tr>
<td>10</td>
<td>SigAfterExpressionInducedRemoval</td>
<td>If the term is still significant after removing of the expression-induced links</td>
</tr>
<tr>
<td>11</td>
<td>log10Pvalue_plus1-e20</td>
<td>The p-value for the significance of function</td>
</tr>
<tr>
<td>12</td>
<td>log10PvalueAfterPureRemoval_plus1-e20</td>
<td>The p-value for the significance of function after removing the pure links from the network</td>
</tr>
<tr>
<td>13</td>
<td>log10PvalueAfterEIremoval_plus1-e20</td>
<td>The p-value for the significance of function after removing the expression-induced links</td>
</tr>
<tr>
<td>14</td>
<td>significantCountIn100_randomPureLinkRemoval</td>
<td>Count of times the function was significant in 100 trials of removing random sets of links to the count of pure links</td>
</tr>
<tr>
<td>15 or 16</td>
<td>significantCountIn100_randomEILinkRemoval_pureCount</td>
<td>Count of times the function was significant in 100 trials of removing expression-induced links to the count of pure links (this was done only for brain, liver and lung)</td>
</tr>
<tr>
<td>15 or 16</td>
<td>significantCountIn100_randomEILinkRemoval</td>
<td>Count of times the function was significant in 100 trials of removing random sets of links to the count of expression-induced links</td>
</tr>
</tbody>
</table>
These files have 15 columns for blood and skeletal muscle and 16 columns for brain, liver and lung. Each row represents a disease term in GO.

Table A- 15 Supplementary file content information

<table>
<thead>
<tr>
<th>Column</th>
<th>Header</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DOID</td>
<td>Disease-ontology ID of the disease term</td>
</tr>
<tr>
<td>2</td>
<td>Name</td>
<td>Name of the disease in disease ontology</td>
</tr>
<tr>
<td>3</td>
<td>geneCount</td>
<td>Count of genes associated with that term in the TSN of the tissue</td>
</tr>
<tr>
<td>4</td>
<td>linkCount(lc)</td>
<td>Count of links between the genes associated with that term in the TSN of the tissue</td>
</tr>
<tr>
<td>5</td>
<td>lc_minusPureLinks</td>
<td>This is linkCount(lc) minus pure links from the network</td>
</tr>
<tr>
<td>6</td>
<td>lc_minusExpressionInducedLinks</td>
<td>This is linkCount(lc) minus expression-induced links from the network</td>
</tr>
<tr>
<td>7</td>
<td>1-ratioOfPureLinks</td>
<td>Ratio of non-pure linkCount(lc)</td>
</tr>
<tr>
<td>8</td>
<td>1-RatioOfExpressionInducedLinks</td>
<td>Ratio of non-expression-induced linkCount(lc)</td>
</tr>
<tr>
<td>9</td>
<td>SigAfterPureRemoval</td>
<td>If the term is still significant after removing of the pure links</td>
</tr>
<tr>
<td>10</td>
<td>SigAfterExpressionInducedRemoval</td>
<td>If the term is still significant after removing of the expression-induced links</td>
</tr>
<tr>
<td>11</td>
<td>log10Pvalue_plus1-e20</td>
<td>The p-value for the significance of disease</td>
</tr>
<tr>
<td>Column</td>
<td>Header</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>--------</td>
<td>-------------</td>
</tr>
<tr>
<td>12</td>
<td>log10PvalueAfterPureRemoval_plus1-e20</td>
<td>The p-value for the significance of disease after removing the pure links from the network</td>
</tr>
<tr>
<td>13</td>
<td>log10PvalueAfterEIlremoval_plus1-e20</td>
<td>The p-value for the significance of disease after removing the expression-induced links</td>
</tr>
<tr>
<td>14</td>
<td>significantCountIn100_-_randomPureLinkRemoval</td>
<td>Count of times the disease was significant in 100 trials of removing random sets of links to the count of pure links</td>
</tr>
<tr>
<td>15 or 16</td>
<td>significantCountIn100_-_randomEILinkRemoval_pureCount</td>
<td>Count of times the disease was significant in 100 trials of removing expression-induced links to the count of pure links (this was done only for brain, liver and lung)</td>
</tr>
<tr>
<td>15 or 16</td>
<td>significantCountIn100_-_randomEILinkRemoval</td>
<td>Count of times the disease was significant in 100 trials of removing random sets of links to the count of expression-induced links</td>
</tr>
</tbody>
</table>
Appendix B  Supplementary Materials for Chapter 3

B.1  Modeling and simulation

For a given gene $q$ and a tissue with $m$ cell types, I use the vector $c_q$ to represent its Cell Type expression profile (CT profile), which contains the expression level of $q$ in each of the $m$ cell types:

\[
\mathbf{c}_q = [t_{q1}, t_{q2}, \ldots, t_{qm}], \quad t_{q1}, t_{q2}, \ldots, t_{qm} \geq 0, \sum_{i=1}^{m} t_{qi} > 0
\]  

(1)

Where $t_{qi}$ is the mean expression level of $q$ in cell type $i$. I denote the variance of values $t_{qi}$s by $\text{var}(t_q)$, that is, the variance of the expression level of $q$ among different cell types.

For a given dataset with $n$ samples from the tissue, vector $\mathbf{a}_j$ represents the Cell Type Composition (CTC) in sample $j$:

\[
\mathbf{a}_j = [a_{j,1}, a_{j,2}, \ldots, a_{j,m}], \quad a_{j,i} \geq 0, \sum_{i=1}^{m} a_{j,i} = 1
\]  

(2)

Where $m$ is the count of cell types in the tissue as it was in definition 1 and $a_{j,i}$ is the proportion of cell type $i$ in sample $j$. Matrix $\mathbf{A}$ defined as:

\[
\mathbf{A} = [\mathbf{a}_1', \mathbf{a}_2', \ldots, \mathbf{a}_n']
\]  

(3)

contains cellular composition vectors for the $n$ samples of the dataset.
**Simulation 1- variance of a gene in bulk tissue**

Having the above definitions, the expression level of \( q \) in \( n \) samples of the dataset is given as \( e_q \):

\[
e_q = c_q A
\]  \hspace{1cm} (4)

I denote the variance of the expression level of \( q \) among the \( n \) samples of the dataset presented in \( e_q \) by \( \text{var}(q) \). From the above formulation, it is apparent that for the two special cases when \( t_{q1} = t_{q2} = \ldots = t_{qm} \) (i.e. \( \text{var}(t_q) = 0 \)) or when \( \alpha_1 = \ldots = \alpha_n \), I have \( \text{var}(q) = 0 \). In a bulk tissue dataset, these two cases are equivalent to when \( q \) has the same expression level in all cell types and when the cellular composition remains the same among the samples.

For simulation, CT expression profiles were generated for 1000 genes for \( m = 10 \) cell types. For each gene \( q \), \( c_q \) was generated using normal distribution with mean and variance obtained from a CT expression profile of a randomly selected gene from the snuc-RNAseq dataset. Bulk tissue data was generated with \( n = 100 \) samples. Matrix \( A \) was generated using uniform distribution with the criteria that each column has sum one. The results show that among the genes, \( \text{var}(t_q) \) and \( \text{var}(q) \) are highly correlated (see Figure B-3)

**Simulation 2 – correlation of two genes in bulk tissue**
For given two genes $p$ and $q$, with expression vectors $e_p$ and $e_q$ in the bulk tissue dataset, the higher the correlation of their CT profiles $c_p$ and $c_q$ (in any direction), the more likely that $e_p$ and $e_q$ are also correlated in the same direction. Having centered $c_p$ and $c_q$ as $c_p^c = c_p - \bar{c}_p$ and $c_q^c = c_q - \bar{c}_q$ and formulation $4$, since $corr(c_p, c_q) = corr(c_p^c, c_q^c) = \cos(c_p^c, c_q^c)$; correlation between $e_p$ and $e_q$ can be written as:

$$
corr(e_p, e_q) = corr\left( \begin{bmatrix} c_p^c \cdot \alpha'_1, c_p^c \cdot \alpha'_2, ..., c_p^c \cdot \alpha'_n \\ c_q^c \cdot \alpha'_1, c_q^c \cdot \alpha'_2, ..., c_q^c \cdot \alpha'_n \end{bmatrix}, \begin{bmatrix} c_p^c \cdot \alpha_1, c_p^c \cdot \alpha_2, ..., c_p^c \cdot \alpha_n \\ c_q^c \cdot \alpha_1, c_q^c \cdot \alpha_2, ..., c_q^c \cdot \alpha_n \end{bmatrix} \right) = corr\left( \begin{bmatrix} ||\alpha_1||\cos(c_p^c, \alpha_1), ||\alpha_2||\cos(c_p^c, \alpha_2), ..., ||\alpha_n||\cos(c_p^c, \alpha_n) \\ ||\alpha_1||\cos(c_q^c, \alpha_1), ||\alpha_2||\cos(c_q^c, \alpha_2), ..., ||\alpha_n||\cos(c_q^c, \alpha_n) \end{bmatrix} \right)
$$

where $\cos(a,b)$ is the cosine of the angle between the two vectors. As the angle between $c_p^c$ and $c_q^c$ gets smaller (equivalent to their correlation getting higher), the difference between $\cos(c_p^c, \alpha_t)$ and $\cos(c_q^c, \alpha_t)$ gets smaller and therefore correlation between $e_p$ and $e_q$ gets higher. Figure 3-3C shows simulation results for correlation of CT expression profiles versus the observed correlation in the bulk tissue for 1000 gene pairs.

### B.2 Gene variance explained by the variance of marker genes

I examined the results from the Principal Component Regression method from the two sets of marker genes from the same five cell types – Astrocyte, Microglia, Oligodendrocyte, Endothelial and Pyramidal. This was to study the specificity of the model in capturing cellular composition induced variance (reflected specifically in the variation of the marker genes) as
opposed to variance induced by a general confounding factor shared among all the genes. The
sets of marker genes are identified independently and have small overlap (see Figure B-1,C
and Figure B-2).
B.3 Figures

Figure B-1 Comparison of two sets of marker genes

(A) Count of marker genes identified in snuc-RNAseq dataset. (B) Marker genes from Mancarci et al. (2017). (C) Count of marker genes from each of A and B and their overlap. (D) R-squared values obtained from the two marker sets. They are highly correlated.
Comparison of the distribution of R-squared values for marker genes, when PCs are obtained from random selection of genes versus when they are obtained from marker genes.
Count of samples identified by Allen Brain Institute in each of the cell-type populations.
Figure B- 4 Variance of the CT profiles versus the observed variance in bulk tissue

(A) Results from simulation for 1000 genes. Each point is data from one gene. (B) Results from data: CT profiles estimated using snuc-RNAseq data, compared to the observed variance in the GTExBulk tissue dataset.
average expression level of genes from GTExBulk marker-enriched clusters in snuc-RNAseq cells populations

- Astrocyte markers
- Pyramidal markers
- Endothelial markers
- Microglia markers
- Oligodendrocyte markers
- Other markers

Expression level

genes from GTExBulk cluster enriched with Pyramidal markers
genes from GTExBulk cluster enriched with Endothelial markers
genes from GTExBulk cluster enriched with Microglia markers
genes from GTExBulk cluster enriched with Oligodendrocyte markers
genes from GTExBulk cluster enriched with Astrocyte markers
Figure B- 5 Mean expression level of genes in each cluster

Mean expression level of genes from GTExBulk marker-enriched clusters in different cell populations of the snuc-RNAsq data. Each plot shows the expression levels for genes from GTExBulk cluster enriched with markers of specific cell types (from top to bottom: Pyramidal, Endothelial, Microglia, Oligodendrocyte and Astrocyte). Lines show the standard deviation. Generally, genes in marker-enriched clusters have higher expression levels in the relevant cell populations – marked by an arrow – compared to other cell populations.
<table>
<thead>
<tr>
<th>Genes and Ontologies</th>
<th>GTExBulk</th>
<th>GTEx_residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exc L2-3 LINC00507 FREM3</td>
<td>4.304</td>
<td>5.53</td>
</tr>
<tr>
<td>Exc L5-6 THEMIS C1QL3</td>
<td>4.061</td>
<td>4.695</td>
</tr>
<tr>
<td>Exc L3-5 RORB ESPR1</td>
<td>3.144</td>
<td>3.405</td>
</tr>
<tr>
<td>Exc L4-5 RORB FOLH1B</td>
<td>3.572</td>
<td>4.455</td>
</tr>
<tr>
<td>Exc L2 LAMP5 LTK</td>
<td>2.238</td>
<td>2.708</td>
</tr>
<tr>
<td>Exc L4-6 RORB SEMA3E</td>
<td>3.01</td>
<td>3.332</td>
</tr>
<tr>
<td>Inh L2-4 PVALB FPDC2</td>
<td>2.45</td>
<td>2.784</td>
</tr>
<tr>
<td>Exc L5-6 FEZF2 ABO</td>
<td>2.432</td>
<td>2.689</td>
</tr>
<tr>
<td>Inh L1-4 LAMP5 LCP2</td>
<td>2.385</td>
<td>2.594</td>
</tr>
<tr>
<td>Exc L4-6 FEZF2 IL26</td>
<td>2.518</td>
<td>2.656</td>
</tr>
<tr>
<td>Exc L5-6 FEZF2 EFTU1P1</td>
<td>2.626</td>
<td>2.852</td>
</tr>
<tr>
<td>Exc L3-4 RORB CARM1P1</td>
<td>2.581</td>
<td>2.765</td>
</tr>
<tr>
<td>Inh L1-3 SST CALB1</td>
<td>2.451</td>
<td>2.723</td>
</tr>
<tr>
<td>Inh L1 SST NMNR</td>
<td>2.339</td>
<td>2.531</td>
</tr>
<tr>
<td>Inh L2-6 LAMP5 CA1</td>
<td>2.408</td>
<td>2.501</td>
</tr>
<tr>
<td>Astro L1-6 FGFR3 SLC14A1</td>
<td>2.285</td>
<td>2.318</td>
</tr>
<tr>
<td>Oligo L1-6 OPALIN</td>
<td>2.728</td>
<td>2.577</td>
</tr>
<tr>
<td>OPC L1-6 PDGFRA</td>
<td>2.326</td>
<td>2.429</td>
</tr>
<tr>
<td>Inh L4-6 SST B3GAT2</td>
<td>2.51</td>
<td>2.614</td>
</tr>
<tr>
<td>Inh L1-3 VIP CHRM2</td>
<td>2.406</td>
<td>2.404</td>
</tr>
<tr>
<td>Exc L2-4 LINC00507 GLP2R</td>
<td>2.498</td>
<td>2.485</td>
</tr>
<tr>
<td>Exc L4-5 RORB DAPK2</td>
<td>2.485</td>
<td>2.533</td>
</tr>
<tr>
<td>Exc L5-6 RORB TTC12</td>
<td>2.478</td>
<td>2.667</td>
</tr>
<tr>
<td>Exc L3-5 RORB COL22A1</td>
<td>2.57</td>
<td>3.478</td>
</tr>
<tr>
<td>Inh L4-6 PVALB SULF1</td>
<td>2.437</td>
<td>2.555</td>
</tr>
<tr>
<td>Exc L4-6 RORB C1R</td>
<td>2.434</td>
<td>2.517</td>
</tr>
<tr>
<td>Exc L3-5 RORB FILIP1L</td>
<td>2.5</td>
<td>2.655</td>
</tr>
<tr>
<td>Exc L5-6 THEMIS CRABP1</td>
<td>2.608</td>
<td>3.679</td>
</tr>
<tr>
<td>Inh L3-5 SST ADGRG6</td>
<td>2.832</td>
<td>2.459</td>
</tr>
<tr>
<td>Inh L1-2 GAD1 MC4R</td>
<td>2.493</td>
<td>2.454</td>
</tr>
<tr>
<td>Inh L1-2 SST BAGE2</td>
<td>2.557</td>
<td>2.484</td>
</tr>
<tr>
<td>Exc L3-5 RORB TWIST2</td>
<td>2.624</td>
<td>2.561</td>
</tr>
<tr>
<td>Inh L4-5 SST STK32A</td>
<td>2.798</td>
<td>2.598</td>
</tr>
<tr>
<td>Inh L1-2 PAX6 CDH12</td>
<td>2.741</td>
<td>2.651</td>
</tr>
<tr>
<td>Inh L5-6 SST MIR548P2</td>
<td>2.708</td>
<td>2.578</td>
</tr>
<tr>
<td>Exc L5-6 THEMIS FGF10</td>
<td>2.666</td>
<td>2.999</td>
</tr>
<tr>
<td>Inh L5-6 SST NPM1P10</td>
<td>2.902</td>
<td>2.749</td>
</tr>
<tr>
<td>Inh L3-6 VIP HS3ST3A1</td>
<td>2.939</td>
<td>2.637</td>
</tr>
<tr>
<td>Inh L1-3 VIP ADAMTSL1</td>
<td>2.957</td>
<td>2.512</td>
</tr>
<tr>
<td>Inh L1-3 SST GGH</td>
<td>3.12</td>
<td>2.952</td>
</tr>
<tr>
<td>Inh L2-4 VIP CBLN1</td>
<td>3.034</td>
<td>2.517</td>
</tr>
<tr>
<td>Inh L1-3 VIP CCDC184</td>
<td>3</td>
<td>2.597</td>
</tr>
<tr>
<td>Inh L2-4 SST FRZB</td>
<td>3.281</td>
<td>2.878</td>
</tr>
<tr>
<td>Inh L2-5 VIP TYR</td>
<td>3.11</td>
<td>2.749</td>
</tr>
<tr>
<td>Inh L4-5 PVALB MEPE</td>
<td>2.968</td>
<td>2.728</td>
</tr>
<tr>
<td>Inh L5-6 SST KLC2/8A</td>
<td>3.188</td>
<td>2.559</td>
</tr>
<tr>
<td>Inh L3-6 SST HPGD</td>
<td>3.16</td>
<td>2.631</td>
</tr>
<tr>
<td>Inh L1-2 VIP PCDH10</td>
<td>3.219</td>
<td>2.864</td>
</tr>
<tr>
<td>Astro L1-2 FGFR3 GFAP</td>
<td>2.992</td>
<td>3.811</td>
</tr>
<tr>
<td>Exc L5-6 SLC17A7 IL15</td>
<td>3.299</td>
<td>2.858</td>
</tr>
<tr>
<td>Inh L2-5 VIP SERPINF1</td>
<td>3.202</td>
<td>2.65</td>
</tr>
<tr>
<td>Exc L5-6 THEMIS DCSTAMP</td>
<td>2.955</td>
<td>3.177</td>
</tr>
<tr>
<td>Exc L6 FEZF2 SCUBE1</td>
<td>4.006</td>
<td>3.204</td>
</tr>
<tr>
<td>Inh L1-4 VIP OPRM1</td>
<td>3.222</td>
<td>3.025</td>
</tr>
<tr>
<td>Inh L1 SST CHRNA4</td>
<td>3.179</td>
<td>2.684</td>
</tr>
<tr>
<td>Inh L5-6 PVALB LGR5</td>
<td>3.198</td>
<td>2.596</td>
</tr>
<tr>
<td>Inh L1-2 VIP LBH</td>
<td>3.813</td>
<td>3.178</td>
</tr>
<tr>
<td>Inh L2-3 VIP CASC6</td>
<td>3.472</td>
<td>2.896</td>
</tr>
<tr>
<td>Inh L1-2 VIP TSPAN12</td>
<td>3.337</td>
<td>2.878</td>
</tr>
<tr>
<td>Inh L4-6 SST GXYLT2</td>
<td>3.81</td>
<td>2.962</td>
</tr>
<tr>
<td>Inh L2-6 VIP QPCT</td>
<td>3.897</td>
<td>3.441</td>
</tr>
<tr>
<td>Inh L2-4 VIP SPAG17</td>
<td>4.584</td>
<td>3.239</td>
</tr>
<tr>
<td>Inh L2-5 PVALB SCUBE3</td>
<td>4.907</td>
<td>3.189</td>
</tr>
<tr>
<td>Inh L1-3 PAX6 SYT6</td>
<td>4.566</td>
<td>3.074</td>
</tr>
<tr>
<td>Inh L5-6 SST GAD1</td>
<td>5.714</td>
<td>3.736</td>
</tr>
<tr>
<td>Inh L5-6 SST GAD1</td>
<td>5.014</td>
<td>3.306</td>
</tr>
<tr>
<td>Exc L4-5 FEZF2 SCN4B</td>
<td>6.111</td>
<td>3.819</td>
</tr>
<tr>
<td>Inh L1-4 VIP CHRNA6</td>
<td>4.883</td>
<td>2.977</td>
</tr>
<tr>
<td>Inh L1-2 LAMPS DBP</td>
<td>6.31</td>
<td>2.093</td>
</tr>
</tbody>
</table>
Figure B- 6 Link overlap ratio

Ratio of the observed versus the expected link overlap of the links from snuc-RNAseq networks with GTExBulk and GTEx_residual networks. Snuc-RNAseq populations are sorted based on the count of cells, with the top one having the highest count of cells. Generally, the link overlap between GTExBulk and GTEx_residual does not change much and there is some level of agreement between snuc-RNAseq population networks and the two bulk networks.
Each column in the heatmap (and the same column in the bar plot) shows data for one cluster identified in GTExBulk network. The grey color shows that the cluster has significantly high count of links in links (FDR $\leq 0.1$). The top bar plot shows the count of Excitatory and Inhibitory networks built from populations of cell types identified in snuc-RNAseq data. Rows in the heatmap shows the reproducibility of clusters in different networks. The bottom color bar shows if the cluster has markers of specific cell types, enriched by housekeeping functional terms or genes.

Figure B-7 Reproducibility of GTExBulk clusters
Representation of the links in a snuc-RNAseq population (Exc L2-3 LinC00507FREM3). This population was selected because it has the highest count of cells among all the snuc-RNAseq populations and therefore a more complete network regarding the gene count. (A) shows count of links in each of the GTExBulk clusters (clusters were selected to have 10 or more links in the snuc-RNAseq population). (B) shows the ratio of links from GTExBulk and GTEx_residual networks that overlap with links in the snuc-RNAseq network. (C) shows the actual count of links from snuc-RNAseq network retrieved (has overlap) in GTExBulk and GTEx_residual networks. For the Pyramidal clusters in the green box, although the count of links has decreased, precision has almost doubled for the GTEx_residual network. For the housekeeping cluster in the orange box, count of links retrieved is more than double in GTEx_residual network compared to the GTExBulk network and the precision has not changed much.
### Tables

**Table B-1 Count of links and genes in each of the networks**

<table>
<thead>
<tr>
<th>networks</th>
<th>gene count</th>
<th>link count</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTEx brain cortex</td>
<td>14,102</td>
<td>1,774,291</td>
</tr>
<tr>
<td>GTEx blood</td>
<td>11,348</td>
<td>1,312,228</td>
</tr>
<tr>
<td>GTEx liver</td>
<td>16,388</td>
<td>1,481,028</td>
</tr>
<tr>
<td>TAN brain</td>
<td>8,761</td>
<td>1,118,791</td>
</tr>
<tr>
<td>TAN blood</td>
<td>8,747</td>
<td>380,875</td>
</tr>
<tr>
<td>TAN liver</td>
<td>10,227</td>
<td>496,476</td>
</tr>
<tr>
<td>TSN brain</td>
<td>6,422</td>
<td>358,531</td>
</tr>
<tr>
<td>Table B-2 Count of genes and links in snuc-RNAseq populations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Exc L2-3</strong> LINC00507 FREM3</td>
<td>1.061e+04</td>
<td></td>
</tr>
<tr>
<td><strong>Exc L5-6</strong> THEMIS C1OL3</td>
<td>1.077e+04</td>
<td></td>
</tr>
<tr>
<td><strong>Exc L5-3</strong> RORB ESR1</td>
<td>8509</td>
<td></td>
</tr>
<tr>
<td><strong>Exc L4-5</strong> RORB FOLH1B</td>
<td>1.104e+04</td>
<td></td>
</tr>
<tr>
<td><strong>Exc L2</strong> LAMPS LTK</td>
<td>8530</td>
<td></td>
</tr>
<tr>
<td><strong>Exc L4-6</strong> RORB SEMASE</td>
<td>9194</td>
<td></td>
</tr>
<tr>
<td>Inh L4-2 PVALB WFDC2</td>
<td>9107</td>
<td></td>
</tr>
<tr>
<td><strong>Exc L5-6</strong> FEZF2 ABO</td>
<td>1.014e+04</td>
<td></td>
</tr>
<tr>
<td>Inh L1-4 LAMPS LCP2</td>
<td>8169</td>
<td></td>
</tr>
<tr>
<td><strong>Exc L4-6</strong> FEZF2 IL26</td>
<td>9410</td>
<td></td>
</tr>
<tr>
<td><strong>Exc L5-6</strong> FEZF2 EFTUD1P1</td>
<td>1.081e+04</td>
<td></td>
</tr>
<tr>
<td>Inh L3-4 RORB CARM1P1</td>
<td>1.034e+04</td>
<td></td>
</tr>
<tr>
<td>Inh L1-3 SST CALB1</td>
<td>5968</td>
<td></td>
</tr>
<tr>
<td>Inh L1 SST NMBR</td>
<td>7839</td>
<td></td>
</tr>
<tr>
<td>Inh L2-6 LAMPS CA</td>
<td>8656</td>
<td></td>
</tr>
<tr>
<td>Astro L1-6 FGFR3 SLC1A4A</td>
<td>4427</td>
<td></td>
</tr>
<tr>
<td>Oligo L1-6 OPALIN</td>
<td>3286</td>
<td></td>
</tr>
<tr>
<td>OPC L1-6 PDQFRA</td>
<td>3281</td>
<td></td>
</tr>
<tr>
<td>Inh L4-6 SST B3GAT2</td>
<td>7402</td>
<td></td>
</tr>
<tr>
<td>Inh L1-3 VIP CHRM2</td>
<td>7583</td>
<td></td>
</tr>
<tr>
<td>Exc L2-4 LINC00507 GLP2R</td>
<td>8894</td>
<td></td>
</tr>
<tr>
<td>Exc L4-5 RORB DAPK2</td>
<td>8636</td>
<td></td>
</tr>
<tr>
<td><strong>Exc L5-6</strong> RORB TTC12</td>
<td>1.048e+04</td>
<td></td>
</tr>
<tr>
<td>Exc L3-5 RORB COL22A1</td>
<td>8420</td>
<td></td>
</tr>
<tr>
<td>Inh L4-6 PVALB SULF1</td>
<td>6339</td>
<td></td>
</tr>
<tr>
<td>Exc L4-6 RORB C1R</td>
<td>9216</td>
<td></td>
</tr>
<tr>
<td><strong>Exc L5-3</strong> RORB FIP1P1L</td>
<td>9276</td>
<td></td>
</tr>
<tr>
<td><strong>Exc L5-6</strong> THEMIS CRABP1</td>
<td>1.016e+04</td>
<td></td>
</tr>
<tr>
<td>Inh L3-3 SST ADGAR6G6</td>
<td>7035</td>
<td></td>
</tr>
<tr>
<td>Inh L1-2 GAD1 MC4R</td>
<td>6328</td>
<td></td>
</tr>
<tr>
<td>Inh L1-2 SST BAGE2</td>
<td>5777</td>
<td></td>
</tr>
<tr>
<td>Exc L3-6 SST B3GAT2</td>
<td>7885</td>
<td></td>
</tr>
<tr>
<td>Inh L4-5 SST STX32A</td>
<td>5401</td>
<td></td>
</tr>
<tr>
<td>Inh L1-2 PAX6 COH12</td>
<td>7204</td>
<td></td>
</tr>
<tr>
<td>Inh L5-6 SST MIIR548F2</td>
<td>8185</td>
<td></td>
</tr>
<tr>
<td><strong>Exc L5-6</strong> THEMIS FGF10</td>
<td>9303</td>
<td></td>
</tr>
<tr>
<td>Inh L5-6 SST NPM1P10</td>
<td>7986</td>
<td></td>
</tr>
<tr>
<td>Inh L3-6 VIP HSSST3A1</td>
<td>6367</td>
<td></td>
</tr>
<tr>
<td>Inh L1-3 VIP ADAMTLS1</td>
<td>5939</td>
<td></td>
</tr>
<tr>
<td>Inh L1-3 VIP GGH</td>
<td>4470</td>
<td></td>
</tr>
<tr>
<td>Inh L2-4 VIP CBRLN1</td>
<td>3773</td>
<td></td>
</tr>
<tr>
<td>Inh L1-3 VIP CCD184</td>
<td>5904</td>
<td></td>
</tr>
<tr>
<td>Inh L2-4 SST FQZB</td>
<td>4674</td>
<td></td>
</tr>
<tr>
<td>Inh L2-5 VIP TYR</td>
<td>4724</td>
<td></td>
</tr>
<tr>
<td>Inh L4-5 PVALB MEPE</td>
<td>7049</td>
<td></td>
</tr>
<tr>
<td>Inh L5-6 SST KLDIC8A</td>
<td>5764</td>
<td></td>
</tr>
<tr>
<td>Inh L3-6 SST HPD</td>
<td>5986</td>
<td></td>
</tr>
<tr>
<td>Inh L1-2 VIP PCDH20</td>
<td>4353</td>
<td></td>
</tr>
<tr>
<td>Astro L1-2 FGFR3 GAF</td>
<td>917</td>
<td></td>
</tr>
<tr>
<td>Exc L5-6 SLC17AT17 IL15</td>
<td>6668</td>
<td></td>
</tr>
<tr>
<td>Inh L2-5 VIP SERPIN1F1</td>
<td>2972</td>
<td></td>
</tr>
<tr>
<td><strong>Exc L5-6</strong> THEMIS DCSTAMP</td>
<td>7535</td>
<td></td>
</tr>
<tr>
<td>Exc L6 FEZF2 SCUBE1</td>
<td>6841</td>
<td></td>
</tr>
<tr>
<td>Inh L1-4 VIP OPR4M1</td>
<td>2960</td>
<td></td>
</tr>
<tr>
<td>Inh L1 SST CHRNA4</td>
<td>3136</td>
<td></td>
</tr>
<tr>
<td>Inh L5-6 PVALB LGR5</td>
<td>3522</td>
<td></td>
</tr>
<tr>
<td>Inh L2-3 VIP ASC6</td>
<td>4474</td>
<td></td>
</tr>
<tr>
<td>Inh L1-2 VIP TSPAN12</td>
<td>4596</td>
<td></td>
</tr>
<tr>
<td>Inh L4-6 SST OXYLT2</td>
<td>3315</td>
<td></td>
</tr>
<tr>
<td>Inh L2-6 VIP QPCT</td>
<td>1717</td>
<td></td>
</tr>
<tr>
<td>Inh L2-4 VIP SPAG17</td>
<td>1074</td>
<td></td>
</tr>
<tr>
<td>Inh L2-5 PVALB SCUBE3</td>
<td>1516</td>
<td></td>
</tr>
<tr>
<td>Inh L1-3 PAX6 SYT6</td>
<td>2202</td>
<td></td>
</tr>
<tr>
<td>Inh L5-6 SST TH</td>
<td>1450</td>
<td></td>
</tr>
<tr>
<td>Inh L5-6 GAD1 GLP1R</td>
<td>856</td>
<td></td>
</tr>
<tr>
<td>Exc L4-5 FEZF2 SCU4N4</td>
<td>2936</td>
<td></td>
</tr>
<tr>
<td>Inh L1-4 VIP CHRNA6</td>
<td>409</td>
<td></td>
</tr>
<tr>
<td>Inh L1-2 LAMPS DBP</td>
<td>158</td>
<td></td>
</tr>
<tr>
<td><strong>Count of genes in the networks</strong></td>
<td>1.061e+04</td>
<td></td>
</tr>
<tr>
<td><strong>Count of inks in the networks</strong></td>
<td>5.636e+05</td>
<td></td>
</tr>
</tbody>
</table>
B.5  Details and description of additional files

Additional file B1

**Description:** functional terms enriched in brain network

**Content:** The file contains list of Gene Ontology terms that are identified as brain-specific. It has four columns: GOID [Gene Ontology ID], GOTerm [Gene Ontology term], Gene Count [count of genes associated with that term], MeanRsqrd [mean Rsquared value of the genes associated with the term]

Additional file B2

**Description:** functional terms enriched in brain and at least one other tissue (common terms)

**Content:** The file contains list of Gene Ontology terms that are identified as brain-specific. It has four columns: GOID [Gene Ontology ID], GOTerm [Gene Ontology term], Gene Count [count of genes associated with that term], MeanRsqrd [mean Rsquared value of the genes associated with the term]

Additional file B3

**Description:** summary information for GTExBulk clusters

**Content:** The file has three columns: ClusterID [ID of the cluster], memberCount [count of genes in the cluster], averageR2 [average of Rsquared value for genes in the cluster]

Additional file B4

**Description:** functional terms enriched in GTExBulk clusters

**Content:** For each cluster ID, enriched functional terms are written. Only terms with FDR $\geq 0.1$ were included. The file has three columns: GOID [Gene Ontology ID for the enriched term], GOTerm [the enriched term], pvalue [pvalue of the enrichment].
Additional file B5

**Description:** robust snuc-RNAseq links

**Content:** Each row has information about one of the links and the file has nine columns: gene_1 [gene symbol of one of the genes], gene_2 [gene symbol of the other gene], totalRep [total count of snuc-RNAseq networks that the link is present in], inhRep [count of Inhibitory neuron networks that the link is present in], excRep [count of Excitatory neuron networks that the link is present in], GTExCorrRank [rank of the link in GTExBulk network, 1000 is the highest correlation values and 0 is lowest], TAN_presence [the link is present in TAN-brain network or not], gene1_r2 [R squared value for gene_1], gene2_r2 [R squared value for gene_2]

Additional file B6

**Description:** functional terms enriched in robust snuc-RNAseq cluster – the gray cluster

**Content:** The file has three columns: GOID [Gene Ontology ID], GOTerm [Gene Ontology term], FDR [FDR from the enrichment]

Additional file B7

**Description:** functional enrichment results for GTExBulk and GTEx_residual networks

**Content:** The file has four columns: GOID [Gene Ontology ID], GOTerm [Gene Ontology term], GTExBulk_p [corrected p value for GTExBulk network], GTEx_residual_p [corrected p value for GTEx_residual network]

Additional file B8

**Description:** Clustering label and R2 values for genes in GTExBulk

**Content:** The file has three columns: GeneSymbol, clusterID, R2
Additional file B9

**Description:** List of cell type markers from snuc-RNAseq

**Content:** Each column is labeled with a cell type and contains list of marker genes obtained from snuc-RNAseq data

Additional file B10

**Description:** final list of marker genes from Mancarci et al.

**Content:** Each column is labeled with a cell type and contains the final list of markers obtained from Mancarci et al.