January 2016

Cell Type-specific Analysis of Human Interactome and Transcriptome

Shahin Mohammadi
Purdue University

Graduate School Form
30 Updated 12/26/2015

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Shahin Mohammadi

Entitled
CELL TYPE-SPECIFIC ANALYSIS OF HUMAN INTERACTOME AND TRANSCRIPTOME

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

Ananth Grama
Chair

Wojciech Szpankowski

David Gleich

Jennifer Neville

Markus Lill

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy of Integrity in Research" and the use of copyright material.

Approved by Major Professor(s): Ananth Grama

Approved by: William Gorman, Assistant to the Department Head 11/1/2016

Head of the Departmental Graduate Program Date

CELL TYPE-SPECIFIC ANALYSIS OF HUMAN INTERACTOME AND TRANSCRIPTOME

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Shahin Mohammadi

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December 2016

Purdue University

West Lafayette, Indiana

I dedicate this thesis to my mom, whose role in my life I can not even begin to describe. She made my past, present, and future possible. There is no word that can express the sacrifices she made in her life to make sure I will have the best life I can. I am, and will always be, in debt to her.

ACKNOWLEDGMENTS

This journey would have not been possible if not for the great help from my major advisor, Ananth Grama, who not only mentored me, but also believed in me, trusted me, and encouraged me to explore my interests and curiosity. I am also grateful to my co-advisor, Wojciech Szpankowski, who continuously supported me and provided valuable guidance throughout my PhD. During these years, I had the honor and the pleasure of collaborating with great colleagues and made wonderful connections which contributed greatly to my success, my growth, and my achievements. I would like to give special thanks to Shankar Subramaniam who have provided great support throughout a few of our works, all of which resulted in wonderful contributions. I am also in debt to David Gleich who have always been there when I need someone to help me understand the problem, formulate it, or otherwise provide invaluable help. Working with him has been a pleasure and I am hoping to continue this relationship throughout my academic career. Last but not least I would like to give special thanks to Andrea Goldsmith and her postdoc Neta Zuckerman who have helped me form a big part of my current direction of research and provided insight on our joint work. I would also like to thank the members of my thesis committee, Markus Lill and Jennifer Neville, for reading this dissertation and providing helpful comments and feedback.

TABLE OF CONTENTS

Page

LIST OF TABLES viii

LIST OF FIGURES ix

ABSTRACT xii

1 INTRODUCTION 1

2 BIOLOGICAL BACKGROUND 5
2.1 Genes, RNAs, Proteins – And How We Measure Them in Bulk Tissues 5
2.2 From Bulk Tissue Measurements to Single Cell RNA-Seq (scRNA-Seq) 6
2.3 Cell Type Specificity – Ubiquitously Expressed Versus Highly Selective Genes 7
2.4 Adding Cell Type-specificity to Biological Networks 8

3 A BIOLOGICALLY-INSPIRED KERNEL TO MEASURE SIMILARITY OF CELLS 9
3.1 Background 9
3.2 Materials and Methods 11
3.2.1 Datasets 11
3.2.2 Identifying the Shared Subspace among a Group of Tissues . 12
3.2.3 Adjusting Transcriptional Signatures to Control for the Effect of Shared Subspace 15
3.2.4 Computing Signal-to-Noise Ratio (SNR) 16
3.2.5 Assessing the Significance of Marker Detection Methods . . 17
3.2.6 Combining Individual p-values to Compute a Meta p-value . 18
3.3 Results and Discussion 18
3.3.1 Adjusting for the Effect of Housekeeping Genes Enhances Signal-to-Noise Ratio (SNR) for the Known Marker Genes 18
3.3.2 Iterative Application of Adjustment Process Identifies Markers That are Comparable or Better Than the t-test 19
3.3.3 Adjusted Cell Type Signatures Identify Groups of Similar Tissues/Cell Types 22
3.3.4 Putting the Pieces Together: Automated Identification of Cell Types and Their Characteristic Markers 25
3.3.5 Adjusted Signatures Predict Tissue-specific Transcriptional Regulatory Networks 26

Page

4 DE NOVO IDENTIFICATION OF CELL TYPES FROM SINGLE-CELL TRANSCRIPTOME . . . 32
4.1 Background . . . 32
4.2 Materials and Methods . . . 34
4.2.1 Datasets . . . 34
4.2.2 Overview of Prior Methods for Cell-type Identification . . . 35
4.2.3 Overview and Justification for *ACTION*'s Components . . . 36
4.2.4 Step 1: A Biologically-inspired Metric for Similarity of Cells 37
4.2.5 Steps 2 and 3: A Geometric Approach to Identify Principal Functions (Representing Pure Cell Types) . . . 39
4.2.6 Estimating the Total Number of Archetypes Needed to Represent All Cell Types . . . 42
4.2.7 Steps 4 and 5: Constructing the Transcriptional Regulatory Network Corresponding to Each Archetype . . . 44
4.3 Results and Discussion . . . 46
4.3.1 Component 1: Measuring Cell-to-cell Similarity . . . 47
4.3.2 Component 2: A Geometric View to Identify Discrete Cell Types 50
4.3.3 Component 3: Constructing Subclass-specific Transcription Regulatory Network of MITF-associated Patients . . . 53

5 SEPARATING CELL TYPES AND THEIR RELATIVE PERCENTAGES FROM COMPLEX TISSUES . . . 57
5.1 Background . . . 57
5.2 Materials and Methods . . . 60
5.2.1 Datasets . . . 60
5.2.2 Deconvolution: Formal Definition . . . 61
5.2.3 Overview of Prior In Silico Deconvolution Methods . . . 73
5.2.4 Evaluation Measures . . . 76
5.2.5 Implementation . . . 77
5.3 Results and Discussion . . . 78
5.3.1 Effect of Loss Function and Constraint Enforcement on Deconvolution Performance . . . 78
5.3.2 Agreement of Gene Expressions With Sum-to-One (STO) Constraint . . . 83
5.3.3 Range Filtering – Finding an Optimal Threshold . . . 86
5.3.4 Selection of Marker Genes – The Good, Bad, and Ugly . . . 95
5.3.5 To Regularize or Not to Regularize . . . 100
5.3.6 Putting it All Together . . . 101
5.3.7 Summary . . . 104

6 CONSTRUCTING CELL TYPE SPECIFIC INTERACTOMES . . . 107
6.1 Background . . . 107
6.2 Materials and Methods . . . 109

Page
6.2.1 Datasets 109
6.2.2 Constructing Human Tissue-specific Interactome 111
6.2.3 Implementation Details 114
6.3 Results and Discussion 114
6.3.1 Transcriptional Activity Scores Predict Tissue-specificity of Genes 114
6.3.2 Constructing Tissue-specific Interactomes 116
6.3.3 Qualitative Characterization of Tissue-specific Interactomes 117
6.3.4 Tissue-specific Interactome Predicts Context-sensitive Interactions in Known Functional Pathways 118
6.3.5 Tissue-specific Interactions are Enriched among Proteins with Shared Tissue-specific Annotations 120
6.3.6 Tissue-specific Interactions Densely Connect Genes Corresponding to Tissue-specific Disorders 120
6.3.7 Tissue-specific Interactome Identifies Novel Disease-related Pathways – Case Study in Neurodegenerative Disorders 122

7 CONSERVATION OF CELL TYPE-SPECIFIC NETWORKS ACROSS SPECIES 132
7.1 Background 132
7.2 Materials and Methods 136
7.2.1 Datasets 136
7.2.2 Sparse Network Alignment Using Belief Propagation 138
7.2.3 Tissue-specific Random Model (TRAM) for Generating Pseudo-random Tissues 139
7.2.4 Significance of Network Alignments 141
7.2.5 Differential Expression of Genes with Respect to a Group of Tissues 142
7.2.6 Conservation of Genesets Based on the Majority Voting Rule 143
7.3 Results and Discussion 144
7.3.1 Aligning Yeast Interactome with Human Tissue-specific Networks 146
7.3.2 Investigating Roles of Housekeeping Genes and their Conservation across Species 148
7.3.3 Quantifying Similarity of Human Tissues with Yeast 152
7.3.4 Identifying Groups of Coherent Tissues 153
7.3.5 Dissecting Tissue-selective Genes with Respect to Their Conservation 157
7.3.6 Elucidating Functional Roles of the Brain and Blood Selective Genes 161
7.3.7 Assessing the Significance of Tissue-specific Pathologies among Conserved and Human-specific Tissue-selective Genes 164

8 CONCLUSION AND FUTURE DIRECTIONS 167

Page

LIST OF REFERENCES 168
VITA 185

LIST OF TABLES

Table Page

3.1 Number of markers for different cell types and tissues used for validating cell similarity kernel 13

3.2 Significance of matching known functional groups to the clusters identified using the new cell similarity kernel 24

5.1 Summary statistics of each dataset for deconvolution 62

5.2 Best combination of choices for feature selection/regularization for different datasets 76

5.3 Summary of adaptive ranges for each dataset for deconvolution 94

5.4 Best combination of choices for feature selection/regularization for different datasets 106

6.1 Compactness of tissue specific disease genes in their tissue-specific interactome 123

7.1 Tissues with the most significant similarity to the yeast interactome . . 154

7.2 Tissues with the least significant similarity to the yeast interactome . . 155

7.3 Summary of tissue-selective gene partitioning 160

7.4 Enriched disease classes of tissue-selective genes 165

7.5 Comparative analysis of brain-specific pathologies 166

LIST OF FIGURES

Figure Page

3.1 SNR enhancement for marker genes after the adjustment process for housekeeping genes 20

3.2 Significance of marker predictions using two-step adjustment process compared to the standard t-test. 21

3.3 Heatmap of tissue/cell type similarities after the adjustment process for housekeeping genes 23

3.4 Performance of different methods for *de novo* identification of cell types using Label Propagation 25

3.5 Enrichment of top-ranked genes in the top three clusters in cancer cell lines dataset using new similarity measure for cells 27

3.4 Tissue-specific transcriptional regulatory network (tsTRN) of top 3 clusters identified in the cancer cell lines dataset 31

4.1 Five main components of *ACTION* 34

4.2 Example of running PCHA algorithm 42

4.3 Illustration of identification of total number of functions for the *Pollen* dataset 44

4.4 Evaluation of *ACTION* Similarity Metric 47

4.5 Performance of *ACTION* in identifying cell types 50

4.6 A continuous view of cell types in the Melanoma dataset identifies subclasses of immune cells and highlights a MITF-related "axis" 53

4.7 The transcriptional regulatory network (TRN) for MITF-associated Melanoma patients highlights a number of genes that have not previously been associated with Melanoma – along with some known markers 54

5.1 Comparison of different loss functions 68

5.2 Average computational time for each loss function in different datasets 79

5.3 Agreement among different evaluation measures across different datasets 80

Figure Page

5.4 Overall performance of different loss/constraints combinations over all datasets 81

5.5 Overall performance of different loss function/constraints combinations over all datasets (lower the better) 82

5.6 Sample-based error of the Retina dataset, based on $\boldsymbol{\mathcal{L}}_2$ with explicit *NN* and *STO* 83

5.7 Percent of features in each dataset that violate the STO constraint . . 85

5.8 Performance of deconvolution methods after removing violating features 86

5.9 Distribution of gene expression values for mixtures and references . . . 88

5.10 Percent of covered features during range filtering 89

5.11 Performance of PERT datasets during range filtering 90

5.12 Average performance of range filtering over all datasets 91

5.13 Dataset-specific changes in the performance of deconvolution methods after filtering expression ranges to fit within $[2^3 - 2^{12}]$ 92

5.14 Sorted log_2-transformed gene expressions in different datasets 93

5.15 Example of adaptive filtering over the CellLines dataset 93

5.16 Dataset-specific changes in the performance of deconvolution methods after adaptive range filtering 94

5.17 Individual performance plots for range filtering in datasets which range filtering exhibits negative effect on the deconvolution 95

5.18 Effect of marker selection on the performance of deconvolution methods 97

5.19 Effect of marker selection, after range filtering, on the performance of deconvolution methods 98

5.20 High-level functional classification of genes 99

5.21 Effect of L2 regularization on the performance of deconvolution methods 102

5.22 Optimal value of λ for each dataset/configuration pair 102

5.23 Performance of deconvolution before/after applying combined feature selection/regularization. Gray shade is mAD of the original deconvolutions (smaller the better). 105

6.1 Summary of GTEx sample numbers per tissue 110

6.2 Distribution of UPC normalized gene expression values 115

Figure Page

6.3 Evaluation of tissue-specific markers using a threshold value of 0.75 to define expressed genes. 126

6.4 Size of the largest connected component in node removal (NR) method as a function of expression threshold. A rapid disaggregation phase can be spotted around 0.75 127

6.5 Qualitative characteristics of tissue-specific interactomes constructed using different methods. 127

6.6 Decomposition of global interactome into brain-specific network using ERW and ActPro ($\alpha = 0.5$) methods 128

6.7 Gain of Area Under the Curve (AUC) of known context-specific pathway edges among tissue-specific interactions 128

6.8 Performance of ActPro with $\alpha = 0.15$ over different tissues 129

6.9 Tissues with the highest gain of AUC for predicting tissue-specific pathway edges 130

6.10 Mean gain of Area under the curve (AUC) for predicting proteins co-annotated with tissue-specific functions 130

6.11 Tissue-specific pathways in human neurodegenerative disorders. Nodes are colored according to their tissue-specific expressions, with novel identified genes marked in red, accordingly. The thickness of edges represent their confidence with tree edges marked as blue. 131

7.1 Main components of the analysis framework for comparing yeast with human tissue-specific interactome 145

7.2 A high-level classification of human genes 146

7.3 Alignment graph of core human genes 149

7.4 Projection of alignment p-values on the network of tissue-tissue similarities 156

7.5 Membership distribution of non-housekeeping genes in human tissues . 157

7.6 Distribution of tissue-selectivity p-values in different tissue groups . . . 158

7.7 Summary of gene classifications. Housekeeping and tissue-selective genes, in four main groups of human tissues, which are classified into three main classes based on their conservation in yeast 160

7.8 Enrichment map of unique blood-selective functions 162

7.9 Enrichment map of unique brain-selective functions 162

ABSTRACT

Mohammadi, Shahin Ph.D., Purdue University, December 2016. Cell Type-specific Analysis of Human Interactome and Transcriptome. Major Professor: Ananth Grama.

Cells are the fundamental building block of complex tissues in higher-order organisms. These cells take different forms and shapes to perform a broad range of functions. What makes a cell uniquely eligible to perform a task, however, is not well-understood; neither is the defining characteristic that groups similar cells together to constitute a cell type. Even for known cell types, underlying pathways that mediate cell type-specific functionality are not readily available. These functions, in turn, contribute to cell type-specific susceptibility in various disorders.

In this dissertation, I propose a novel measure of similarity between cells and utilize it to identify *de novo* cell types. I show that my method allows us to uncover novel cancer subtypes. Furthermore, by constructing underlying pathways that drive progression of these subtypes, I show that we can pinpoint diagnostic biomarkers and potential therapeutic targets. Then, I develop a method to dissect the cell type composition of complex tissues. Using this snapshot of what tissues/cell types look like, I create a framework for constructing tissue/cell type-specific interactomes to shed light on the systems-level understanding of cellular functions. I use these networks to uncover brain-specific pathways that are involved in Alzheimer's and Parkinson's diseases. Finally, I provide evidence for the conservation of these interactomes across distant species, even down to unicellular organisms, such as yeast.

1 INTRODUCTION

Human cells, while inheriting a similar genetic code, exhibit distinct morphological and functional characteristics and group together uniquely to form complex tissues. Uncovering biochemical processes that drive the transformation of a totipotent cell into various cell types and ultimately tissues is essential to our understanding of living systems. Understanding this complex machinery determines how tissues differ in terms of their anatomy, physiology, morphology, and, more importantly, how various cellular control mechanisms contribute to the observed similarities/ differences.

A fundamental challenge in understanding cellular biology is to classify cells according to their common functions. This allows us to study cell types as a whole, and to extend our understanding to the behavior of individual cells. Traditionally, cells are classified into a few hundred different cell types according to their morphological properties and cell cell surface markers. However, emerging knowledge suggests that seemingly identical cell types may exhibit varying transcriptional characteristics, leading to vastly different functions. This, in turn, motivates the development of new approaches for defining refined groupings of cells.

With the availability of single cell transcriptomic data, there is an unprecedented opportunity to analyze and model cellular processes at a resolution that was not possible before. These technologies have the potential to radically redefine our view of cell type identity. However, there are a number of challenges to realize the potential of these datasets. The first major challenge is to define what makes a pair of cells similar. This similarity is at the very core of any algorithm that aims to find groups of similar tissue, and by extension, coherent cell types. To this end, my first attempt was to develop a method for uncovering true similarity between cells, while accounting for biological and technical noise. In this framework, I aim to remove the common part of the identity of cells to boost their distinguishing (informative) signals. To this end, I

project cellular signatures to a suitably regularized orthogonal subspace, which allows better identification of cells, as well as their similarities. I show that this reduction step enhances the signal-to-noise ratio (SNR) for known markers. Moreover, I show that repeated application of subspace reduction within groups of cell types allows us to identify highly specific markers.

Armed with a measure of cell-to-cell similarities, the next logical objective is to group coherent cells to identify *de novo* cell types. To this end, we need to account for a few important considerations. First, there are extremely important but very rare cells, such as circulating tumor cells, the identity of which is of considerable significance. Traditional clustering algorithms typically fail to capture such trends, and there is a need to develop specific tools that are robust to the sampling density of cells. Another key challenge is that even after identifying a cell type, it is unclear what distinguishes a cell type from other cell types. Transcriptional regulatory networks (TRNs) are at the heart of this differentiating process. As such, understanding cell type-specific TRNs has the potential to unlock cell type identity. To address these issues, I developed a method that uses our measure of cell-to-cell similarity as a kernel, and constructs a geometric representation of the functional space of cells. This representation characterizes principal functions that are performed by cells. I then couple this geometric approach with a new statistical framework to reconstruct the underlying transcriptional regulatory networks that mediate characteristic behavior of each cell type.

Once we have characterized a cell-type, the next step is to analyze how various cell types come together to form complex tissues. A fundamental question then is to deconvolve complex tissues to identify their constituting cell types and their relative fractions. This has applications in the removal of contaminants (e.g., surrounding cells) from tumor biopsies, as well as in monitoring changes in the cell population in response to treatment or infection. This problem is known as source separation in the signal processing community and has attracted considerable attention. In this dissertation, I focused on one specific problem in this area: knowing individual cell

types profiles, can we identify the composition of these cell types in a given tissue? To answer this question, I performed a comprehensive study to investigate the effect of different loss functions, constraints on the solution, preprocessing and data filtering, feature selection, and regularization on deconvolution quality. I developed prescriptive recipes that yield the best performance and showed how these recipes could be used in practice.

The next question I address is how various cell types and tissues come together at a system level? How do various gene products interact? What are the emergent properties of these complex interactions? To answer these questions, I developed a method for constructing accurate tissue/cell type-specific networks. This approach formulates network inference as a suitably regularized convex optimization problem. The objective function of the optimization problem has two terms – the first term corresponds to a diffusion kernel that propagates activity of genes through interactions (network links). The second term is a regularizer that penalizes differences between transcriptional and functional activity scores. I use these functional activity scores to compute tissue-specificity of each edge in the global interactome, which I show, through a number of validation tests, are significantly better than state-of-the-art methods in the field. Finally, I couple these networks with Prize-Collecting Steiner Tree (PCST) method to identify novel disease/tissue-specific pathways that drive neurodegenerative disorders. This method is platform-independent and can be applied directly to single-channel, double-channel, and RNA-Seq expression datasets.

Finally, I addressed the problem of transferrability of molecular mechanisms across organisms. Budding yeast, S. cerevisiae, has been used extensively as a model organism for studying cellular processes in evolutionarily distant species, including humans. However, The extent to which a unicellular organism, such as yeast, can be used to model tissue-specific processes has never been assessed. To answer this question, I developed a novel framework to systematically quantify the suitability of yeast as a model organism for different human tissues. To this end, I first used network alignment to map human tissue-specific networks to the yeast interactome. Then, I

devised a statistical model to assign an empirical p-value to each alignment, assessing the overall suitability of yeast to model the systems biology of each tissue. My framework not only helps in classifying human tissues/cell types as either compatible with the yeast model or not, but also provides missing functional elements in yeast for each tissue. These functional constructs can further be utilized to engineer humanized yeast models that mimic the biology of specific human tissues, and can be used as high-throughput, tissue-specific model to study different diseases.

In summary, my dissertation extends our understanding of the identity of human cell types, their functional pathways, their composition in complex tissues, and their conservation across evolution. In the what follows I will address each of these questions in sequential order in each chapter of this dissertation.

2 BIOLOGICAL BACKGROUND

2.1 Genes, RNAs, Proteins – And How We Measure Them in Bulk Tissues

The *central dogma* of biology describes the flow of genetic information within cells – the genetic code, represented in DNA molecules, is first transcribed to an intermediate construct, called *messenger RNA (mRNA)*, which in turn translates into *proteins*. These proteins are the functional workhorses of the cell. Genes, defined as the minimal coding sections of the DNA, contain the recipe for making proteins. These instructions are utilized dynamically by the cell to adapt to different conditions.

The amounts of various proteins in a cell can be measured at a time point. This corresponds to the level of *protein expression*. This process is limited by the availability of high-quality antibodies that can specifically target each protein. The amount of active *mRNA* in a cell, however, can be measured at the genome scale using high-throughput technologies such as *microarrays* and *RNASeq*. The former is an older technology that relies on the binding affinity of complementary base pairs (alphabets used in the DNA/RNA molecules), while the latter is a newer technique, using *next generation sequencing (NGS)*. This technique estimates gene expression based on the overlap of mRNA fragments with known genomic features. Since microarrays have been used for years, extensive databases from different studies are publicly available. RNASeq datasets, in comparison, are relatively smaller but growing rapidly in scale and coverage. Both of these technologies provide reliable proxies for the amount of proteins in cells, with RNASeq being more sensitive, especially for lowly expressed genes.

The expression level of genes is tightly regulated in different stages of cellular development, as well as in response to environmental changes. In addition to these *biological variations* due to cellular state, intermediate steps in each technology in-

troduce *technical variations* in repeated measurement of gene expression in the same cell-type. To enhance reproducibility of measurements, one normally includes multiple instances of the same cell-type in each experiment, known as *technical* replicates. The expression profiles from these experiments provide a snapshot of the cell under different conditions. In addition to the biological variation of genes within the same cell type, there is an additional level of variation when we look across different cell types.

2.2 From Bulk Tissue Measurements to Single Cell RNA-Seq (scRNA-Seq)

Traditionally, biological samples are disaggregated and measured in bulk. However, different tissues typically consist of a heterogeneous population of cells of different type and at different fractions. An important example is the tumor microenvironment. In this case, not only we have immune, stromal, and cancerous cells all mixed in the tumor biopsy, but also the tumor samples themselves consist of different subtypes.

Many different groups have focused their energy to develop technologies capable of measuring RNA quantities at the single cell level. A major challenge to achieve this goal is the limited RNA quantity at single cell level. The total amount of RNA in a single cell is in order of picograms, whereas most recent RNA-seq technologies need at least tens of nanograms to be able to measure RNA levels. To compensate for this gap, amplification techniques before performing RNA-Seq are mandatory. Another challenge is to isolate single cells from bulk tissue without perturbing their expression profile. Furthermore, different techniques may have biases towards different ends of RNA or according to the specific sequence of it. Finally, the maximum length of RNAs that each method can measure can vary across methods. Methods such as Smart-Seq [151], and its successor Smart-Seq2 [144], aim to sequence the whole length of genes. On the other hand, more recently developed methods, such as InDrop [92] and DropSeq [114], sacrifice full-length coverage to gain of significantly higher throughput.

These methods are under constant development, with microfluidics and emulsion-based techniques being at the leading front of this development.

2.3 Cell Type Specificity – Ubiquitously Expressed Versus Highly Selective Genes

Proteins are basic workhorses of living cells. Their overall quantity is tightly regulated across different tissues and cell-types to manifest tissue-specific biology and pathobiology. These regulatory controls orchestrate cellular machinery at different levels of resolution, including, but not limited to, gene regulation [62, 120], epigenetic modification [24,121], alternative splicing [19,46], and post-translational modifications [77, 194]. Transcriptional regulation is a fundamental component of this hierarchical regulation, which has been widely used to study context-specific phenotypes. In the context of human tissues/ cell types, genes can exhibit varying levels of specificity in their expression. They can be broadly classified as (i) tissue-specific (unique to one cell-type); (ii) tissue-selective (shared among coherent groups of cell-types); and (iii) housekeeping (utilized in all cell-types). Housekeeping genes comprise a subset of human genes that are universally expressed across all tissues and are responsible for maintaining core cellular functions needed by all tissues, including translation, RNA processing, intracellular transport, and energy metabolism [23, 39, 172]. These genes are under stronger selective pressure, compared to tissue-specific genes, and evolve more slowly [221]. In contrast, certain genes are specifically or preferentially expressed in one, or a set of biologically relevant tissue types [22, 23, 181, 206]. These *marker genes* are critical for distinguishing various cells. In fact, *cell surface markers* have long been used to sort different subsets of immune cells. Tissue-specific/selective genes have significant applications in drug discovery, as they have been shown to be more likely drug targets [40]. Tissue-specific transcription factors (tsTFs) are significantly implicated in human diseases [123, 150], including cancers [197].

2.4 Adding Cell Type-specificity to Biological Networks

The majority of human proteins do not work in isolation but take part in pathways, complexes, and other functional modules. These complex interactions are typically represented as a graph. This graph can be undirected, in the case of protein-protein interaction networks (PINs), or directed, as in transcriptional regulatory networks (TRNs). In PINs, each node represents a protein and each edge indicates a physical interaction between a pair of proteins. These interactions are measured in vitro by technologies such as yeast two-hybrid (Y2H) or affinity purificationmass spectrometry (AP/MS). The basic premise of these technologies is to assess if two proteins *can* interact. That is conditioned on if they are both expressed at high enough levels, co-localized, and post-translationally modified (if needed). As such, PINs provide a *superset* of all possible interactions that can happen in the cell. Of course, this set also contains a lot of false negatives since not all pairs of proteins are systematically measured for interactions, and even if they have been, there is a high false positive rate associated with these technologies as well (especially with AP/MS). In TRNs, each node can be either a transcription factor (TF) (a protein) or a target gene (TG). Interactions are regulatory interactions and are directed. If a protein is linked to a gene if means that it regulates the expression of that genes. This regulation can be either positive, or activation, or negative, or inhibition.

In context-specific networks, we add a spatial and/or temporal context to these networks. What this information provides is a realistic snapshot of what is going on inside a specific cell type at a given moment (that the data was captured). Perturbations that impact interacting interfaces of proteins are significantly enriched among tissue-specific, disease-causing variants [155, 157, 208]. Additionally, disease-related protein complexes tend to be over-expressed in tissues in which defects cause pathology [103]. In terms of topology, tissue-specific genes typically reside in the periphery of the interactome, are enriched among signaling and cell surface receptors, and highly associated with the onset of tissue-specific disorders [216].

3 A BIOLOGICALLY-INSPIRED KERNEL TO MEASURE SIMILARITY OF CELLS

3.1 Background

An embryonic stem cell encapsulates all of the genetic information needed to develop an individual; it differentiates into various cell types, which group together to shape tissues, combine to constitute organs, and assemble into organ systems. Various differentiated tissues/ cell types, while inheriting a similar genetic code, exhibit unique anatomical and physiological features. Traditionally, these cell types/ tissues have been classified using their high-level phenotypic characterizations, such as location and morphology. However, more recently, single-cell technologies have revealed an unprecedented heterogeneity among what were, until recently, believed to be identical cell types [162]. This heterogeneity is achieved through systematic control of cellular machinery at different levels, including transcriptional, translational, and post-translational regulations, to orchestrate tissue-specific functions and dynamic responses to environmental stimuli.

Transcriptional regulation is among the best-studied aspects of this control. It is manifested in the observed differences in expression levels of genes across tissues. *Housekeeping genes* constitute the subset of the transcriptome that is universally expressed in human tissues. These genes are responsible for core cellular functions [23, 39, 172], and their corresponding pathways, are essential to all cells for their normal activity. However, they are not informative, with respect to the identity of cells, nor do they provide any power to classify cells into coherent groups of cell types. In contrast, certain genes are specifically or preferentially expressed in one, or a set of biologically relevant tissue types [22, 23, 181, 206]. These *marker genes* are critical for distinguishing various cells. In fact, *cell surface markers* have long been used to sort

different subsets of immune cells. These genes play a crucial role in the physiology and the pathophysiology of human tissues. Many of the known disease genes are tissue-specific and are under/ over-expressed in the specific tissue(s) where the gene defect causes pathology [60, 103].

The use of transcriptomic profile as a genome-scale phenotype to identify unique cell types has attracted considerable attention [186]. However, identifying transcriptionally related cell types and their key marker genes remains a challenging task. One of the complicating factors in this paradigm is the hierarchical relationships among cell types. At the highest level, all cells are highly similar due to the expression of housekeeping genes. These genes are typically expressed at high levels and strongly impact the computed cell-cell distances (using any of the existing distance measures). After peeling this common layer, cell types split into groups with common functionality, which can be represented using the *community affiliation graph* model [215]. Here, we can model common functionalities such as "affiliations", which are used to annotate cell types. However, these affiliations are not known *a priori*. Furthermore, cell types are not uniformly spaced and form a hierarchical structure linking them together. As we move deeper into this hierarchy, shared functionalities become more detailed, and distances among cell types reduce – necessitating use of rigorous statistical models and methods to assess the "proximity" of cell types.

In this chapter, I propose an iterative, multi-step process to simultaneously identify groups of similar cell types as well as their characteristic marker genes that are specifically expressed within each group of cell types. The two main operators in my framework are *subspace reduction*, in which we identify the unique signature of a given *expression domain*, and *clustering*, in which we group similar tissues/ cell types in the *reduced space* to define new *expression domains*. In this framework, I make an implicit assumption that genes do not work alone, but rather, as part of functional pathways. These pathways can be viewed as *barcodes* that uniquely identify their corresponding cell types/ tissues.

Motivated by these considerations, I develop a novel algorithm for *de novo* identification of cell types and their corresponding markers. I show that this subspace reduction step significantly enhances the signal-to-noise ratio (SNR) for markers, and that repeated application of reduction step within known groups of cells can identify their markers. Next, I show that, in the absence of known groupings, my method can automatically identify similar cell types using a clustering algorithm. I use this as a hierarchical prior for characterizing the expression domain of genes. Finally, I show that my method is able to reconstruct highly accurate models of tissue-specific transcriptional regulatory networks (tsTRN). my framework is particularly well-suited for applications in single-cell analysis, in which the true identity of cell types, as well as their corresponding markers, is critical.

3.2 Materials and Methods

3.2.1 Datasets

Gene Expression Profiles

In my experiments, I used two separate datasets derived from different technologies. The first dataset, which I will refer to as *immune cell types*, is the expression profile of 38 distinct subpopulations of hematopoietic cells measured using Affymetrix GeneChip microarray [130]. This dataset consists of the gene expression of $12,074$ genes in a total of 211 samples. The second dataset contains a comprehensive compendium of 675 cancer cell lines [93]. The origin of these cell lines can be classified into 17 different tissues. I will focus on these 17 distinct groups, but collectively refer to this dataset as the *cancer cell lines* dataset.

Gold Standard Marker Genes

To evaluate identified markers and the impact of adjustment, I collected marker genes from two independent studies. For immune cell types, I adopted the LM22

dataset from Newman *et al.* [129]. First, for each cell type in LM22, I identified genes that are highly expressed. Then, I computed the mean expression of these markers in each of the immune cell types in my dataset. I constructed a weighted bipartite graph between cell types in these two datasets and identified matches using a maximum-weight bipartite matching algorithm [102], followed by manual assessment. Table 3.1a shows the final results for the immune marker set.

For the cancer cell lines dataset, I downloaded the gold standard tissue-specific markers from the Human Protein Atlas (HPA) [193]. I manually matched ten different tissues of origin to the markers in HPA, and limited my focus on the markers that have both transcriptomic and proteomic evidence. Among these ten, *pancreas* markers were not significantly expressed in the pancreas-originated cell lines, and thus I removed this from my set. The final set consists of nine tissues, shown in Table 3.1b.

Transcriptional Regulatory Network (TRN)

I collected transcription factor (TF) – target gene (TG) interactions from the RegNetwork database [110], which aggregates data from 25 different databases. This dataset contains a total of $151,214$ regulatory interactions between $1,408$ TFs and $20,230$ TGs.

3.2.2 Identifying the Shared Subspace among a Group of Tissues

A given set of tissues/ cell-types typically share a common set of genes/ pathways, while specializing through preferential genes that control and regulate this core shared set. I represent the *raw transcriptional signature* of these tissues using a matrix $T \in \mathbb{R}^{n_g \times n_t}$, in which rows correspond to genes and columns correspond to various tissues. We are interested in finding the subspace of common genes, and to use it to adjust the transcriptional signatures. When the given set includes all, or majority of, human cell-types, the shared subspace represents the signature of housekeeping genes.

Table 3.1: Number of markers for different cell types and tissues used for validating cell similarity kernel

(a) Immune cell types

LM22 cell type	mapped cell type	number of markers
B cells naive	Naive B-cells	118
B cells memory	Mature B-cell class able to switch	106
Plasma cells	Mature B-cells	109
T cells CD8	CD8+ Effector Memory	142
T cells CD4 naive	Naive CD4+ T-cell	121
T cells CD4 memory activated	CD4+ Effector Memory	107
NK cells activated	Mature NK cell_CD56+ CD16+ CD3-	109
Monocytes	Monocyte	104
Dendritic cells activated	Myeloid Dendritic Cell	121
Eosinophils	Eosinophill	159
Neutrophils	Granulocyte (Neutrophilic Metamyelocyte)	140

(b) Cancer cell lines

cell line origin	number of markers
brain	336
colo-rectal	72
kidney	158
liver	185
lung	56
ovary	35
stomach	77
urinary bladder	27
skin	133

There are a number of methods for approximating this common signature in T, the simplest of which would be to compute the mean of its columns. An alternate

approach involves decomposing T into sum of rank-one matrices, using methods such as *singular value decomposition (SVD)* or *non-negative matrix under-approximation (NMU)*. The general goal of these methods is to represent T as a sum of outer products of vectors. More formally, I write T as follows:

$$T = U_r \Sigma_r V_r = \sum_{i=1}^{r} \sigma_i u_i {v_i}^T, \tag{3.1}$$

where $r \leq \min(n_g, n_t)$ is the rank of the approximation. In the SVD formulation, u_i and v_i vectors are called left and right singular vectors, respectively. These vectors constitute an orthonormal basis, that is, both $u_i u_j^T = \delta_{ij}$ and $v_i v_j^T = \delta_{ij}$ for all i and j. Additionally, for any r, an SVD is the optimal rank-r approximation of T. When all entries of T are positive, Perron-Frobenius theorem ensures that all entries of the both left and right singular vectors are positive. However, the first residual matrix, $R_1 = M - \sigma_1 u_1 {v_1}^T$, can, and typically does, contain negative elements to ensure orthonormality. On the other hand, the *NMU* formulation does not ensure orthonormality, but, rather enforces an additional constraint on the optimization problem, which is that R_k should consist of only positive elements. Unfortunately, while SVD has an optimal solution, the additional non-negativity constraint of NMU makes its computation non-convex, though heuristics exist to approximate the solution.

Here, I use a rank-one approximation of matrix T, that is $r = 1$, to identify a unique signature that closely represents the common signature in T. I use the first singular vector of matrix T, after z-score normalization, as a proxy for the housekeeping signature throughout my study.

3.2.3 Adjusting Transcriptional Signatures to Control for the Effect of Shared Subspace

Let us denote the transcriptional profile of the ith tissue by $\mathbf{T}_i$. In order to compute the *raw transcriptional similarity* between each given pair of tissues, I compute the Pearson's correlation as follows:

$$r_{\mathbf{T}_i\mathbf{T}_j} = \frac{\sum_{k=1}^{n}(t_{ki} - \bar{t_i})(t_{kj} - \bar{t_j})}{\sqrt{\sum_{k=1}^{n}(t_{ki} - \bar{t_i})^2}\sqrt{\sum_{k=1}^{n}(t_{kj} - \bar{t_j})^2}} \tag{3.2}$$

where, n represents the total number of genes, t_{ki} and t_{kj} are the expression levels of the kth gene in the ith and j^{th} tissues, respectively. Similarly, $\bar{t_i}$ and $\bar{t_j}$ represent the average expression levels of genes in the corresponding tissues. Let $\mathbf{Z}_i$ denote the Z-score normalized version of $\mathbf{T}_i$, defined as $\frac{\mathbf{T}_i - \mu(\mathbf{T}_i)}{\sigma(\mathbf{T}_i)}$. I refer to $\mathbf{Z}_i$ as the *raw transcriptional signature* of tissue i. Using this formulation, I can simplify the raw transcriptional similarity as the normalized dot product of raw transcriptional signatures:

$$r_{\mathbf{T}_i\mathbf{T}_j} = \frac{\mathbf{Z}_i\mathbf{Z}_j}{n} \tag{3.3}$$

The raw transcriptional similarity of tissues is artificially inflated due to the ubiquitous expression of housekeeping genes across all tissues. To control for this effect, I first define *housekeeping transcriptional signature*, denoted by vector $\mathbf{S}$, as the left singular vector of matrix $\mathbf{Z}$. Using this notation, I revise my similarity scores by computing the partial Pearson's correlation between $\mathbf{T_i}$ and $\mathbf{T_j}$, after controlling for the effect of $\mathbf{S}$ as follows:

$$r_{\mathbf{T}_i\mathbf{T}_j\bullet\mathbf{S}} = \frac{r_{\mathbf{T}_i\mathbf{T}_j} - r_{\mathbf{T}_i\mathbf{S}}r_{\mathbf{T}_j\mathbf{S}}}{\sqrt{1 - r^2_{\mathbf{T}_i\mathbf{S}}}\sqrt{1 - r^2_{\mathbf{T}_j\mathbf{S}}}} \tag{3.4}$$

As before, we can rewrite this using the Z-score formulation. Let us denote the adjusted transcriptional profile of tissue i as $\mathbf{Y}_i = \mathbf{Z}_i - r_{\mathbf{T}_i\mathbf{S}}\mathbf{S}$. I define the *adjusted transcriptional signature* of tissue i as $\hat{Z}_i = \frac{\mathbf{Y}_i - \mu(\mathbf{Y}_i)}{\sigma(\mathbf{Y}_i)}$. Finally, we can rewrite the adjusted transcriptional similarity as:

$$r_{\mathbf{T}_i\mathbf{T}_j\bullet\mathbf{S}} = \frac{\hat{\mathbf{Z}}_i\hat{\mathbf{Z}}_j}{n} \tag{3.5}$$

I use this approach to remove the shared subspace of a given set of expression profiles, and to construct the corresponding adjusted transcriptional signatures. Significantly positive transcriptional similarities in this framework are indicators of shared tissue-specific pathways. I use these adjusted transcriptional signatures in my study to identify marker genes. However, when applying methods that rely on the positivity of the input expression matrix, one can use the sigmoid transform of these scores as follows:

$$\hat{p}_{ki} = \frac{1}{1 + e^{-\hat{z}_{ki}}} \tag{3.6}$$

Please note that this transformation, when applied to the raw transcriptional signatures, is equivalent to the previously known softmax normalization:

$$\begin{aligned} p_{ki} &= \frac{1}{1 + e^{-z_{ki}}} \\ &= \frac{1}{1 + e^{-(\frac{t_{ki} - \mu(\mathbf{T}_i)}{\sigma(\mathbf{T}_i)})}} \end{aligned} \tag{3.7}$$

This normalization is known to remove the effect of outliers, while preserving a linear relationship for mid-range values.

3.2.4 Computing Signal-to-Noise Ratio (SNR)

Signal-to-Noise Ratio (SNR) is a commonly used measure for evaluating the quality of a desired *signal* by comparing the power of the signal to the power of (undesired) *noise*. I define the desired signal as the expression of marker genes in their corresponding tissue/ cell type of origin. Similarly, I define noise as the expression of the rest of the genes in that cell type. Let us assume there are k replicas of a given tissue/ cell type, and a total of n genes, represented in a matrix $\mathbf{A} \in \mathbb{R}^{n \times k}$. We are also given a subset $\mathcal{S}$ of rows that are designated as markers. I compute the power of signal as: $P_{signal} = \frac{\|A(\mathcal{S},:)\|_F^2}{|\mathcal{S}|}$. The numerator can be also expressed as $\| \mathbf{vec}(A(\mathcal{S},:)) \|_2^2$, where the **vec** operator vectorizes a matrix by stacking up its columns. Similarly, we can

compute the power of noise as: $P_{signal} = \frac{\|A(\mathcal{S}',:)\|_F^2}{|\mathcal{S}'|}$, where $\mathcal{S}' = \{1..n\} \setminus \mathcal{S}$. Then, we can compute SNR as:

$$SNR = 10log_{10}(\frac{P_{signal}}{P_{noise}}), \tag{3.8}$$

which is in unit of decibels (dB).

3.2.5 Assessing the Significance of Marker Detection Methods

I use the *hypergeometric* p-*value* as a statistical measure of the overlap among sets. A typical use case for this formulation is in over-representation analysis (ORA). The classical approach to this problem is to select a predefined cutoff l to identify top-ranked genes, and then to compute the enrichment p-value using the hypergeometric distribution. Let us denote the total number of gene products by N. Given a set of known gene annotations (true positives) of size A, I encode these annotations using a binary vector $\lambda = \lambda_1, \lambda_2, ...\lambda_N \in \{0,1\}^N$. Let the random variable T denote the number of positive genes in the target set, if we distribute genes randomly. In this formulation, the hypergeometric p-value is defined as:

$$\begin{aligned} p\text{-value}(T = b_l(\lambda)) &= Prob(b_l(\lambda) \leq T) \\ &= HGT(b_l(\lambda)|N, A, l) \\ &= \sum_{t=b_l(\lambda)}^{min(A,l)} \frac{C(A,t)C(N-A, l-t)}{C(N,l)} \end{aligned} \tag{3.9}$$

where HGT is the tail of hypergeometric distribution and $b_l(\lambda) = \sum_{i=1}^{l} \lambda_i$ counts the total number of true positives in top-l observations. The drawback of this approach is that we need a predefined cutoff value, l. To remedy this, Eden *et al.* [43] propose a two-step process for computing the exact enrichment p-value, called *mHG p-value*, without the need for a predefined cutoff value of l. First, an optimal cutoff value is chosen among all possible values of $1 \leq l \leq N$. The computed value for this optimal cutoff is called the *minimum hypergeometric (mHG) score*, and is defined as:

$$mHG(\lambda) = min_{1 \leq l \leq N} HGT(b_l(\lambda)|N, A, l) \tag{3.10}$$

Next, a dynamic programming (DP) method is used to compute the exact p-value of the observed mHG score, in the state space of all possible λ vectors of size N having exactly A ones (please refer to Eden *et al.* [43] for algorithmic details, and Eden [42] for an efficient implementation).

3.2.6 Combining Individual p-values to Compute a Meta p-value

When we compute individual p-values for each tissue/ cell type, we then need to combine them in order to define a meta p-value that can be used to assess each selection. To combine a set of computed p-values, I use the Fisher's method [49]. This method computes a statistic $\mathcal{S} = -2\sum_{i=1}^{k} ln(p_i)$ for a set of k given *p-values* p_i. Then, I can use χ^2 test with $2k$ degrees of freedom to assess the significance of the meta-analysis, assuming that p_is are independent.

3.3 Results and Discussion

In this section, I validate the following hypotheses: (i) a single level of adjustment (reducing the effect of housekeeping genes) enhances the signal-to-noise ratio (SNR); (ii) repeated application of the reduction process over groups of cell types allows us to recover cell type-specific markers; (iii) automatic identification of *putative cell types* using label propagation based clustering yields reliable grouping of cell types; and (iv) cluster-specific, adjusted signatures yield highly accurate models of cell type/ tissue-specific transcriptional regulatory circuits. All of these hypothesis are validated using known cell type groupings and markers.

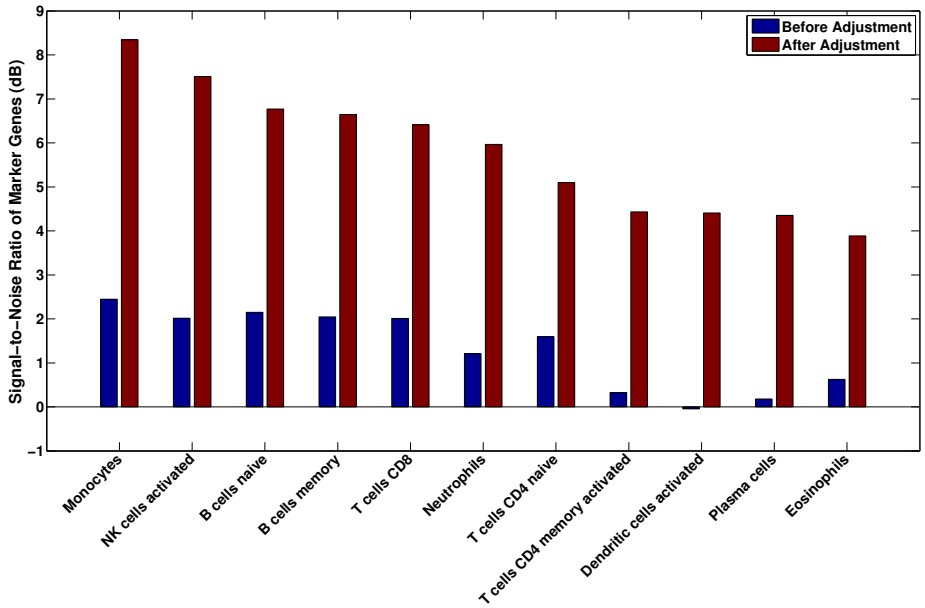
3.3.1 Adjusting for the Effect of Housekeeping Genes Enhances Signal-to-Noise Ratio (SNR) for the Known Marker Genes

I hypothesize that the global expression of housekeeping genes, which are universally expressed genes that perform core cellular functions, masks the true signal from

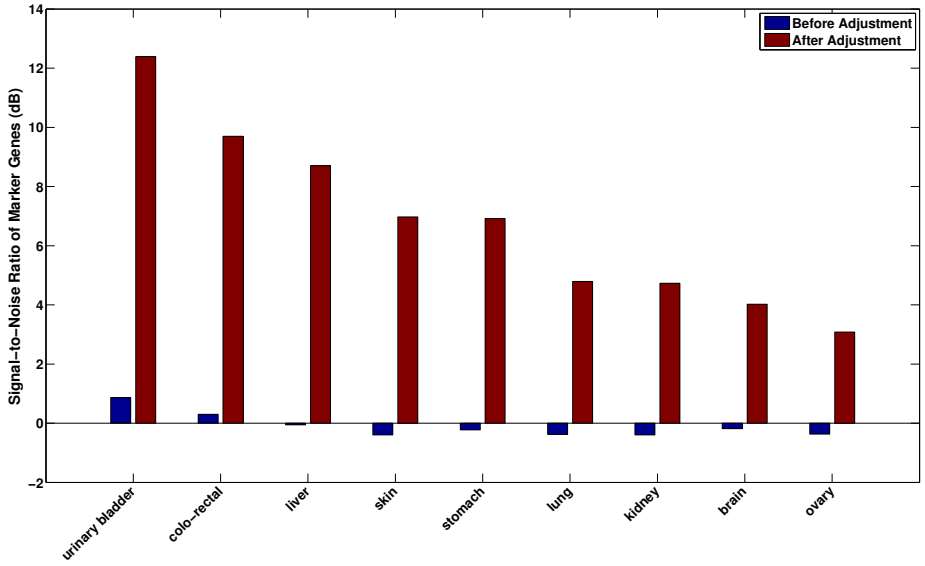
tissue-specific markers. Thus, adjusting for this common signature should enhance the signal-to-noise ratio (SNR) of marker detection methods. To systematically evaluate my hypothesis, I compute SNR using Equation 3.8, for immune cells and cancer cell line markers, respectively. The results of this adjustment are presented in Figure 3.1. In the two cases shown, we observe a significant improvement over the raw expressions. However, we note that in the cancer cell line dataset, there are cases in which the power of non-marker genes is stronger than the power of marker genes, thus the negative dB values. This effect is remedied in all cases after adjustment, which suggests that the proposed adjustment process deflates housekeeping gene expression effectively, but does not negatively influence the expression of marker genes.

3.3.2 Iterative Application of Adjustment Process Identifies Markers That are Comparable or Better Than the t-test

In this experiment, I quantify the extent to which highly expressed genes in the adjusted profile can be used to identify tissue/ cell type-specific markers. I apply the same adjustment process to each group of cells/tissues, after adjusting for housekeeping effect. The result is a single shared signature for each group of cell types/ tissues. I rank genes according to their expression level in this signature and assess the over-representation of known markers among the higher ranked elements in this list. I use mHG p-values, introduced in Section 3.2.5, to assess the significance of each case. Similarly, I compute the mHG p-values for results of one-sided and two-sided t-tests, which correspond to the most commonly used methods for identifying differentially expressed genes. my final results are presented in Figure 3.2. For the immune cell dataset, the iterative adjustment process yields superior results in every single case. However, in the cancer cell line datasets, the results are more varied. In this case, I removed *ovary* from my study, since none of the methods had significant p-values. To systematically evaluate different methods, I use Fisher's method [49] to combine individual p-value into a *meta* p-value, the details of which are presented in



(a) Immune cell types

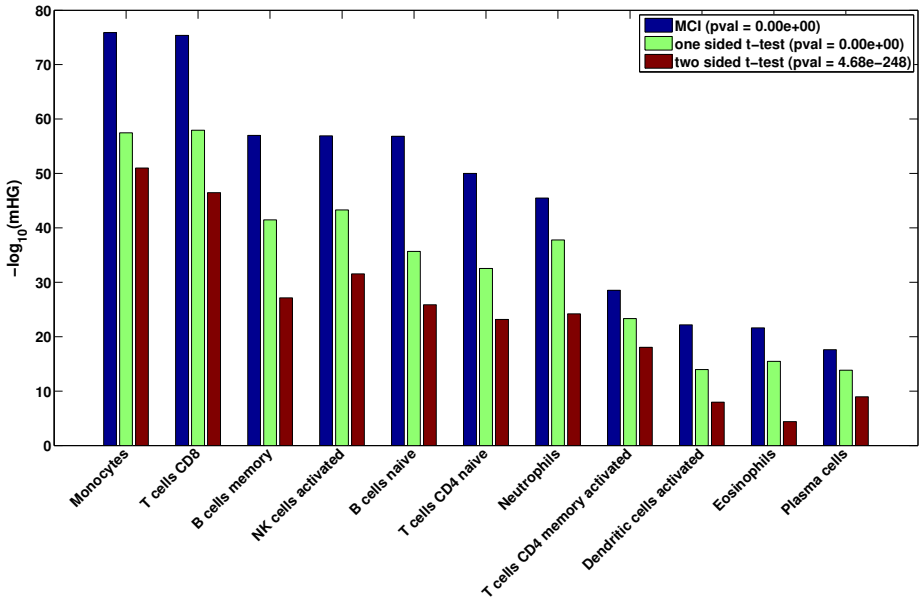


(b) Cancer cell lines

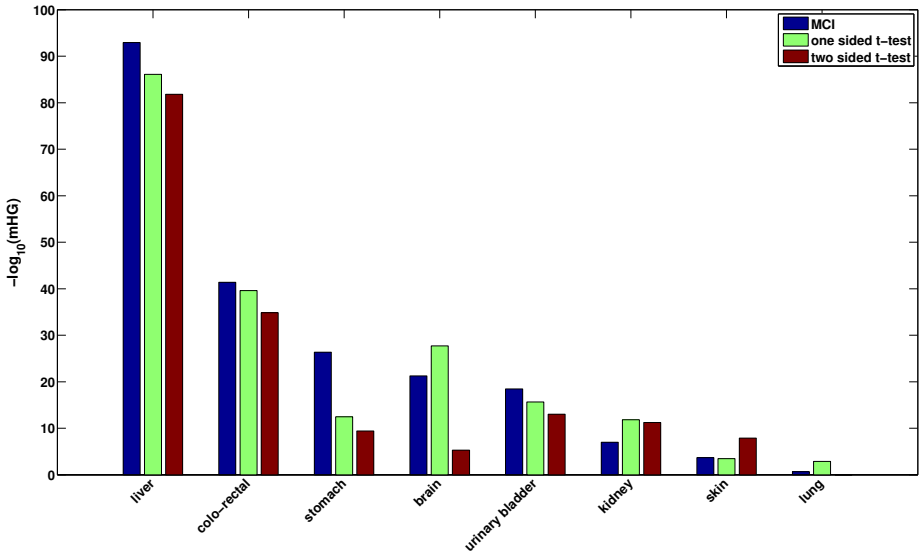
Figure 3.1.: SNR enhancement for marker genes after the adjustment process for housekeeping genes

Section 3.2.6. This results in the combined p-values of 2.5×10^{-197}, 2×10^{-185}, and 8.1×10^{-150}, for my method, one-sided t-test, and two-sided t-test, respectively.

In summary, in both cases my method significantly outperforms the standard t-test but, *more importantly*, as I show in the next section, it does not depend on a predefined grouping and can automatically identify relevant expression domains.



(a) Immune cell types



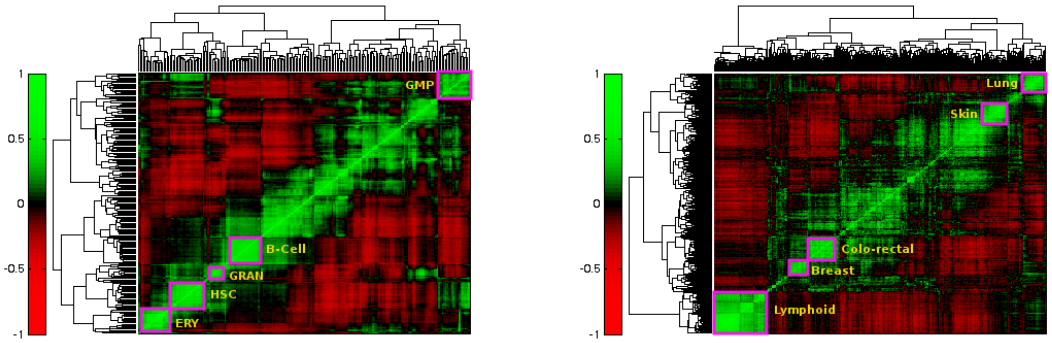
(b) Cancer cell lines

Figure 3.2.: Significance of marker predictions using two-step adjustment process compared to the standard t-test.

3.3.3 Adjusted Cell Type Signatures Identify Groups of Similar Tissues/Cell Types

Having established that iterative application of the adjustment process can identify marker genes within given groups of cells, I now study whether these groups can be identified from the data directly. Given a compendium of cells, this would allow us to automatically identify major subgroups corresponding to cell types, as well as key marker genes associated with each group. In order to evaluate if such structure exists in the adjusted data, I perform bi-clustering on the similarity matrix between tissues/ cell types. I compute similarities using Pearson's correlation, after adjusting expression profiles for the effect of housekeeping genes. Figure 3.3 shows the clustered heat-map of samples in each of my datasets. Each coherent group of samples is marked according to the majority of cell types/ tissues in the group. For the immune cell dataset, *B-cell (mature)* and *hematopoietic stem cell (HSC)* are two of the largest coherent groups, followed by *erythrocyte (ERY), granulocyte/monocyte progenitor (GMP)*, and *granulocyte (GRAN)*. In the cancer cell line dataset, *lymphoid* tissues comprise the largest coherent group, followed by *lung, skin, colo-rectal*, and *breast* groups. These groups are the major *separable* clusters at the first level of hierarchy. The size of each group is related to the total number of samples for that tissue/ cell type, whereas consistency within the group is related to the homogeneity of cell types. For example, *lymphoid* tissues exhibit three separate subtypes in the heat-map, which correspond to *bone marrow, lymph node* and *blood*.

I use a recent method proposed by Gaiteri [52] to automatically identify these *separable* groups. This algorithm is a modification of the label propagation clustering that corrects for the global frequency of labels, which in turn allows it to identify more refined clusters. I compute similarity matrices before and after adjustment, and remove all negative entries after computing the correlation scores. Next, I match each identified cluster to known groups in each dataset. I first construct a weighted bipartite graph between clusters on one side and known groups of cell types on the other, by assigning a hypergeometric p-value to the size of their overlap. I then use



(a) Immune cell types

(b) Cancer cell lines

Figure 3.3.: Heatmap of tissue/cell type similarities after the adjustment process for housekeeping genes

a maximum-weight bipartite matching algorithm [102] to compute the best match for each cluster. I rank each tissue/ cell type according to the best matched cluster, i.e., how well identified clusters capture each group. Table 3.2 summarizes the set of tissues/ cell types in each dataset best matched to the identified set of clusters. Interestingly, all major separable groups in the cancer cell line dataset are captured by at least one cluster. In addition, brain and pancreatic tissues both have a corresponding cluster, even though in the heat-map, they were not distinguishable from the rest of tissues. On the other hand, for immune cell types, clusters cover a majority of separable cell types, with the exception of *GMP*, which is a heterogeneous group by itself consisting of a group of progenitor cells in the myeloid branch. Memory T-cells are also strongly connected in the heat-map, but are split into different groups, with the groups themselves being fairly homogeneous. We note that, in general, known tissues in the cancer cell lines dataset are better represented by their clusters than the immune cell types, in terms of their overlap p-value. I hypothesize that this phenomena is due to higher underlying similarity among immune cell types that is not separable using only one level of clustering.

Table 3.2: Significance of matching known functional groups to the clusters identified using the new cell similarity kernel

(a) Immune cell types

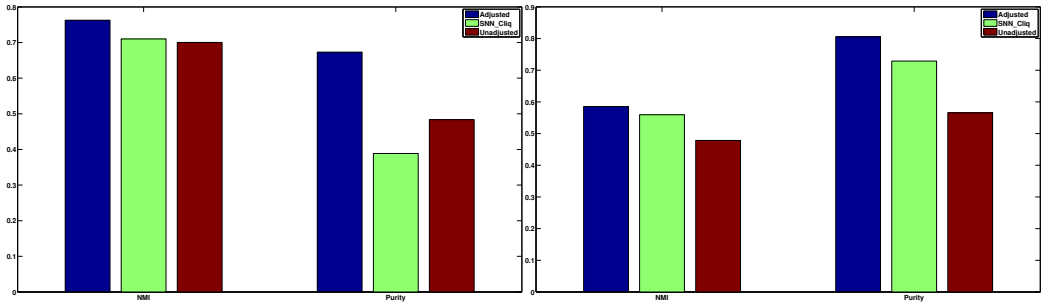
Celltype	$-log_{10}(p-val)$
Hematopoietic stem cell	8.36
Erythroid	7.63
Mature B-cell class able to switch	4.90
CD4+ Central Memory	4.34
Granulocyte (Neutrophil)	3.88
Basophils	3.60

(b) Cancer cell lines

Cellline origin	$-log_{10}(p-val)$
lymphoid	120.02
skin	51.45
breast	38.86
colo-rectal	38.43
brain	13.17
lung	11.85
pancreas	11.75

In order to assess the performance of my method, I applied the same procedure on each of the clusters, representing major cell types, to identify more refined clusters, where each cluster represents a sub-cell type. In addition to the cell type hierarchy identified using adjusted/unadjusted transcriptional signatures, I also identified cell types using *SNN_Cliq* method [213], which is shown to outperform both *k-means* and *DBSCAN* methods in identifying cell types. Figure 3.4 compares the final clustering results, using *NMI* and *Purity* measures, which are two of the most well-used ex-

trinsic measures to evaluate clustering results. It can be seen that in all case, label propagation clustering using *adjusted* signature outperforms the other two methods.



(a) Immune cell types (b) Cancer cell lines

Figure 3.4.: Performance of different methods for *de novo* identification of cell types using Label Propagation

In summary, label propagation applied to the similarity scores after adjustment for housekeeping genes can automatically identify groups of cell types/ tissues with coherent functions/ expression. These groups can be used as a hierarchical prior to define the expression domain of tissue/ cell type-specific genes and their corresponding pathways, as I demonstrate in the next section.

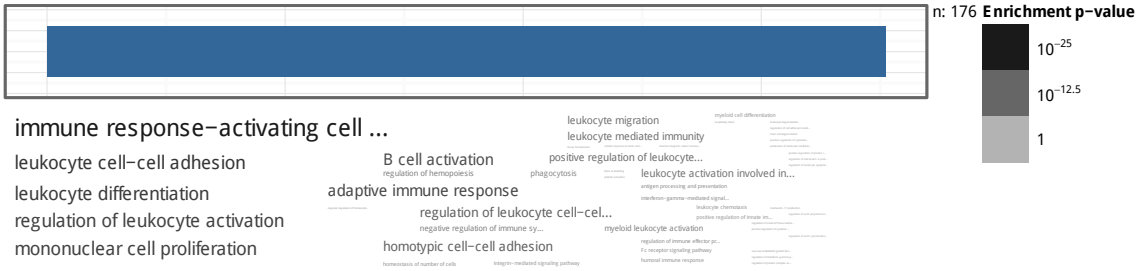
3.3.4 Putting the Pieces Together: Automated Identification of Cell Types and Their Characteristic Markers

I have, thus far, shown that adjusted transcriptional signatures are capable of identifying highly accurate cell-type markers. Furthermore, being accurate representations of cell type/ tissue-specific functionality, these signatures are better suited to quantifying cell type-cell type and tissue-tissue similarities. These similarities, in turn, can be used to identify coherent groups of cell types/ tissues. Here, I show that highly expressed genes within identified clusters are enriched with tissue/ cell type-specific pathways. I select the top three clusters that correspond to top three best-covered tissues of origin in the cell lines dataset as my test cases. First, I ap-

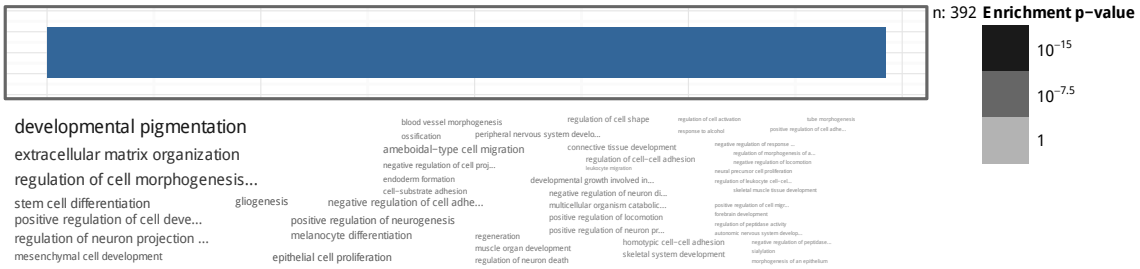
ply the adjustment process over each cluster, instead of known groups. I then filter each cluster signature vector to select anything above z-score threshold of 1.96. I use these three genesets and performed GO enrichment analysis over each one of them using the *GOsummaries* package in R/Bioconductor [95]. This package relies on the g:Profiler [153] package to identify and summarize enriched terms using their hierarchical relationships, but also generates a word cloud of the final, simplified results. Figure 3.5 shows the enrichment for the top three clusters in the cancer cell line dataset, which are *lymphoid, skin,* and *breast*, respectively. I note that the annotations of each cluster are consistent with the matched pair of known groups. Furthermore, each cluster is highly enriched with respect to related tissue-specific functions. This validates the fact that the grouping/ marker detection process is able to automatically identify cell types/ tissues, and to identify highly specific markers.

3.3.5 Adjusted Signatures Predict Tissue-specific Transcriptional Regulatory Networks

Tissue/ cell type-specific transcription factors (tsTFs) are significantly implicated in various human disorders [123, 150], including cancers [197]. Having established that adjusted signatures can be used to identify marker genes from among identified clusters, I now construct core regulatory networks responsible, in each tissue, for defining its identity. I focus on the same set of tissues as in Section 3.3.4. For each tissue, I first identify the set of transcription factors that are *highly expressed*, specifically in that tissue. I then assign a p-value to each of these TFs by looking at their target genes. I identify how many total targets each TF has, how many of them are expressed (above z-score of 1.96), and how many total genes are expressed in the adjusted signature. Using these statistics, I compute the p-value of tissue-specificity for each selected TF using the tail of hypergeometric distribution. A TF is deemed signficant in a given tissue if it is specifically expressed highly in that tissue, after the iterative adjustment process, and has a significantly large number of targets that are



(a) Lymphoid



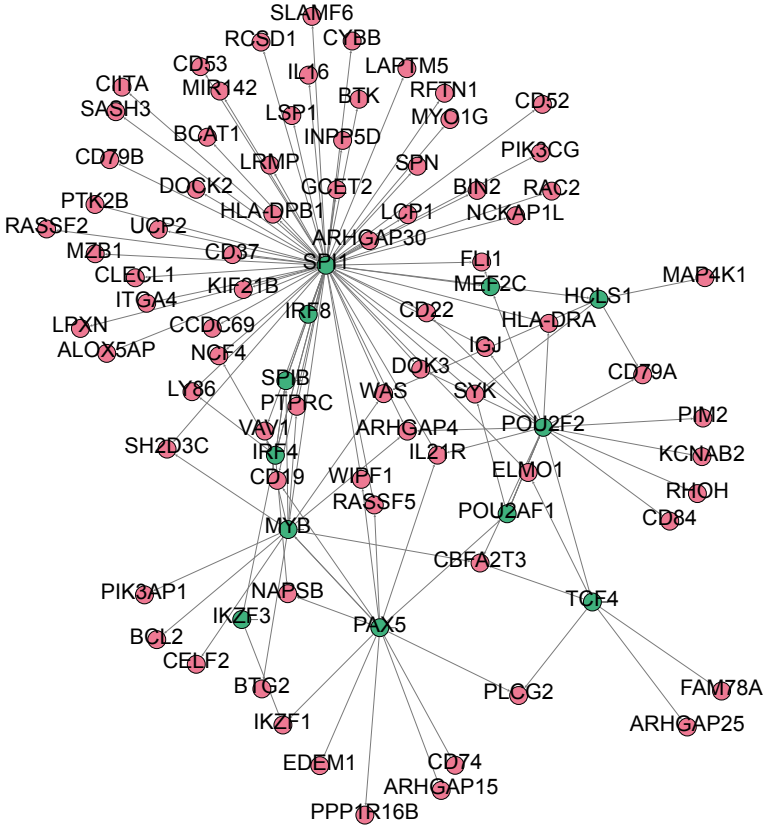
(b) Skin

(c) Breast

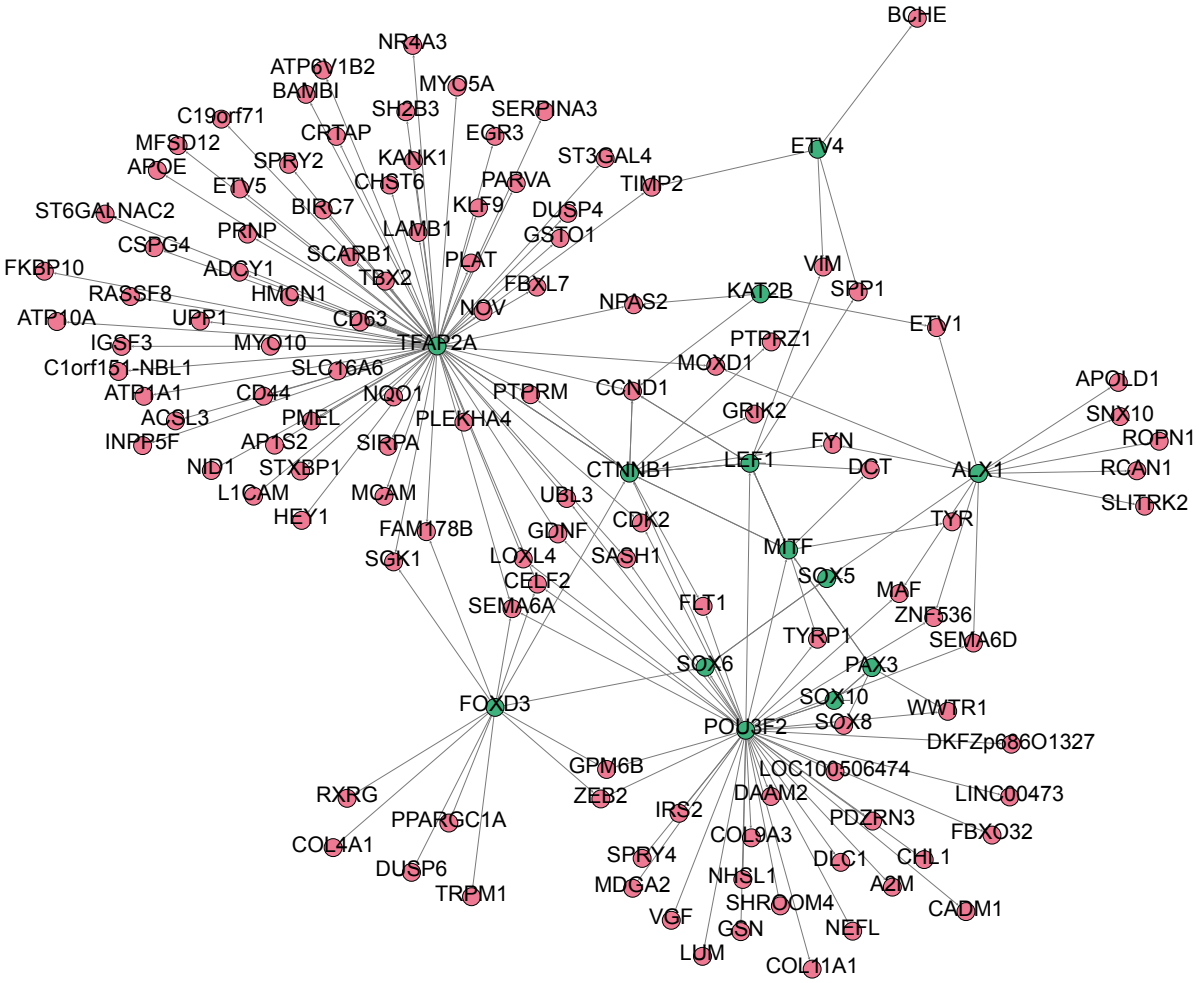
Figure 3.5.: Enrichment of top-ranked genes in the top three clusters in cancer cell lines dataset using new similarity measure for cells

also highly expressed. I identify a minimal set of 12, 14 and 8 TFs for *lymphoid, skin,* and *breast* tissues, respectively. I then construct the tissue-specific transcriptional regulatory network (tsTRN) as the bipartite graph consisting of the selected TFs, together with their highly expressed gene targets. For breast, *GRHL1* just has a self-loop, whereas, in skin network, *NDN* TF is only connected to *NGFR*. I exclude these two TFs from further study. Figure 4.7 shows three networks corresponding to the tsTRN of these tissues.

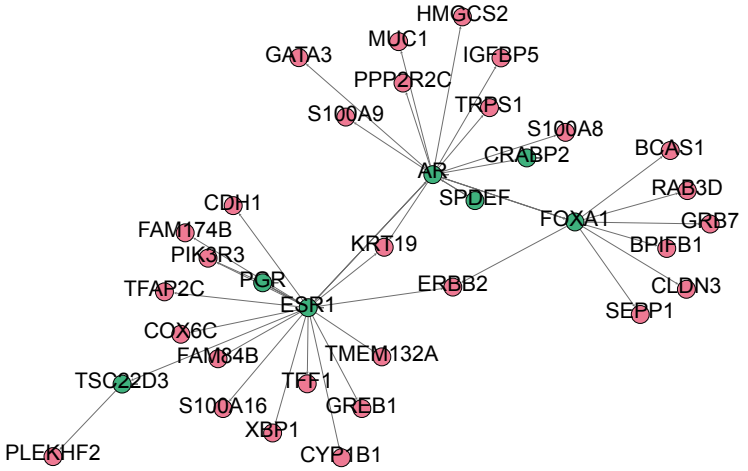
Functional enrichment analysis of identified TFs shows a very significant and very relevant set of functions. *Myb* is a known proto-oncogene and its over-expression plays a key role in development of chronic B-lymphocytic leukemia (B-CLL) [198]. On the other hand, *POU2F2, SPI1, MEF2C, MYB, IRF4, IRF8, IKZF3*, and *HCLS1* are all involved in the *hematopoiesis* (GO:0030097 p-val $= 3.6 \times 10^{-9}$). Among these genes, *SPI1* has the highest connectivity in the constructed lymphoid-specific TRN (Figure 3.6a). This TF regulates gene expression during myeloid and B-lymphoid cell development. In skin-specific TRN (Figure 3.5b), *TFAP2A* has the highest connectivity, but *CTNNB1* has a higher centrality. Interestingly, a subset of TFs in this network, *LEF1,CTNNB1*, and *ALX1*, are known to be involved in the *positive regulation of epithelial to mesenchymal transition* (p-val $= 3.2 \times 10^{-6}$). This suggests that the skin-specific network can be used to identify new targets for trans-differentiation. Finally, breast-specific TRN is centered around *Estrogen Receptor 1 (ESR1)*, *Androgen Receptor (AR)*, and *Forkhead Box A1 (FOXA1)* TFs. These TFs, together with *progesterone receptor (PGR)*, constitute the core of the *steroid hormone mediated signaling pathway* (p-val $= 9.4 \times 10^{-7}$), and essential for sexual development and reproductive function. In summary, these tsTRNs, identified automatically from the given cell type/ tissue-specific transcriptome, capture highly relevant functionalities that are fundamental to the core identity of each cell type. I conclude that, my framework can identify hypothesized groups of related cells, identify their common markers, and construct the underlying circuits that regulate the context-specific machinery.


SLAMF6
RCSD1
CYBB
CD53
IL16
LAPTM5
CIITA
MIR142
BTK
RFTN1
SASH3
LSP1
MYO1G
CD52
BCAT1
INPP5D
PIK3CG
CD79B
LRMP
SPN
DOCK2
GCET2
BIN2
RAC2
PTK2B
HLA-DPB1
LCP1
NCKAP1L
RASSF2
UCP2
ARHGAP30
MZB1
CD37
SPI1
FLI1
CLECL1
MEF2C
MAP4K1
KIF21B
ITGA4
HCLS1
IRF8
CD22
LPXN
CCDC69
HLA-DRA
ALOX5AP
NCF4
IGJ
DOK3
CD79A
SPIB
LY86
WAS
SYK
PTPRC
PIM2
VAV1
ARHGAP4
POU2F2
SH2D3C
IRF4
IL21R
KCNAB2
CD19
WIPF1
ELMO1
RHOH
RASSF5
POU2AF1
CD84
MYB
CBFA2T3
PIK3AP1
NAPSB
TCF4
IKZF3
PAX5
BCL2
CELF2
BTG2
PLCG2
FAM78A
IKZF1
ARHGAP25
EDEM1
CD74
ARHGAP15
PPP1R16B

(a) Lymphoid


BCHE
NR4A3
ATP6V1B2
BAMBI
MYO5A
SH2B3
SERPINA3
C19orf71
CRTAP
EGR3
MFSD12
SPRY2
KANK1
ST3GAL4
ETV4
APOE
ETV5
CHST6
PARVA
TIMP2
ST6GALNAC2
BIRC7
KLF9
DUSP4
PRNP
LAMB1
GSTO1
CSPG4
SCARB1
PLAT
VIM
FKBP10
ADCY1
TBX2
FBXL7
SPP1
RASSF8
HMCN1
NOV
NPAS2
KAT2B
ATP10A
UPP1
CD63
ETV1
IGSF3
MYO10
TFAP2A
PTPRZ1
C1orf151-NBL1
SLC16A6
MOXD1
APOLD1
ATP1A1
CD44
NQO1
PTPRM
CCND1
SNX10
AQSL3
PMEL
PLEKHA4
GRIK2
ROPN1
INPP5F
AP1S2
SIRPA
FYN
NID1
STXBP1
CTNNB1
LEF1
DCT
ALX1
RCAN1
L1CAM
MCAM
UBL3
SLITRK2
HEY1
CDK2
TYR
FAM178B
GDNF
MITF
SGK1
LOXL4
SASH1
SOX5
CELF2
MAF
SEMA6A
FLT1
ZNF536
TYRP1
SEMA6D
SOX6
PAX3
SOX10
WWTR1
FOXD3
POU3F2
SOX8
DKFZp686O1327
GPM6B
LOC100506474
RXRG
ZEB2
DAAM2
LINC00473
IRS2
PDZRN3
PPARGC1A
COL9A3
FBXO32
COL4A1
SPRY4
DLC1
CHL1
DUSP6
NHSL1
A2M
MDGA2
TRPM1
SHROOM4
CADM1
VGF
GSN
NEFL
LUM
COL11A1

(b) Skin



(c) Breast

Figure 3.4.: Tissue-specific transcriptional regulatory network (tsTRN) of top 3 clusters identified in the cancer cell lines dataset

4 DE NOVO IDENTIFICATION OF CELL TYPES FROM SINGLE-CELL TRANSCRIPTOME

4.1 Background

Complex tissues consist of heterogeneous populations of interacting cells that are specialized to perform different functions. With rapid growth in single cell transcriptomic technologies, the observed diversity of known cell types has greatly expanded. What were once believed to be homogeneous groups of cells can now viewed as ecosystems of varying cell types [186]. In tumor microenvironments, for example, immune, stromal, and cancerous cells coexist, cooperate, and compete for resources. The exact composition of these cells, as well as their molecular makeup, have significant impact on diagnosis, prognosis, and treatment of cancer patients [129]. Single cell technologies have already been proven useful for dissecting this complex microenvironment [156]. Using the rapidly growing datasets of single cell gene expression profiles, a key challenge is to identify *de novo* cell types directly from genome-wide transcriptomic phenotypes [176]. An important problem in cell type identification is the existence of rare but key cell types, such as circulating tumor cells [145]. Beyond identifying cell types, it is also import to identify factors that distinguish them from other cell types.

I propose a new method, called *Archetypal-analysis for cell type identificaTION (ACTION)*, to identify cell types from single cell expression datasets. My method is robust to biological noise, identifies a wide range of cell types with varying relative populations, and provides a novel mechanism for constructing transcriptional regulatory networks (TRN) that mediate characteristic behaviors of each cell type. At the core of my method is a biologically-inspired metric for similarity of cells, as characterized by their transcriptional profiles. This metric accounts for specificity of marker

genes and defines a signature for each cell that is robust to noise. At the same time, it is sensitive enough to capture weak cell type-specific signals. This metric helps us construct a geometric representation for the space of principal functions, which are groups of distinguishing functions that are uniquely performed by specialized cells. In this space, assigning cells to their closest principal function accurately identifies cell types. Finally, I develop a statistical framework to identify key marker genes, as well as transcription factors that are responsible for mediating the observed expression of these markers. I use these regulatory elements to construct cell type-specific transcriptional regulatory networks.

My method provides a flexible approach for directly mapping characteristic transcriptional regulatory networks of cells from the raw transcriptomic data. I apply my method to the problem of subtyping Melanoma patients and identify a coherent subclass, which closely resembles noninvasive tumors [201]. For this subclass, I characterized key marker genes, as well as their underlying pathways. This analysis highlights a MITF-associated regulatory network and suggests a potential mechanism for distinguishing invasive and proliferative types of melanoma.

Significance. A few methods have been proposed for the problem of cell type identification [65, 70, 82, 96, 117, 213, 219]. A common theme underlying these methods is to cluster coherent cells as putative cell types [176]. At the core of these clustering methods is a similarity measure that defines relationships among cells. A majority of prior methods rely on classical measures such as correlation or Euclidean distance to define such relationships. However, this approach is confounded by ubiquitously and highly expressed levels of housekeeping genes. Cell type-specific markers, on the other hand, have a weaker signal in comparison. This, in turn, causes a majority of traditional techniques to be driven by biological noise contributed by housekeeping genes [125]. To overcome this, methods – such as *ACTION* – that are robust to biological noise but are sensitive enough to identify cell type-specific signals are critically needed. Once the identity of a cell has been established, it is unclear what

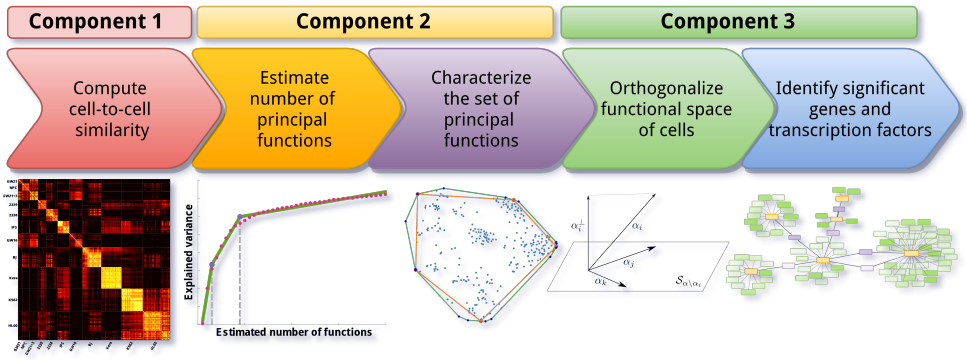


Figure 4.1.: Five main components of *ACTION*

distinguishes it from other cell types. Transcriptional regulatory networks (TRNs) are important aspects of this differentiation process. Understanding cell type-specific TRNs has the potential to explain distinguishing mechanisms underlying observed transcriptional phenotypes. *ACTION* is among the first set of methods to directly infer cell type-specific networks from single cell expression datasets.

4.2 Materials and Methods

4.2.1 Datasets

Single cell gene expression datasets For all my studies, I rely on the following datasets collected from publicly available sources:

***Immune* (from Supplementary Material)** : Comprehensive qPCR based assay of 1522 immune cells. This dataset spans 30 different types of stem, progenitor, and fully differentiated cells [67].

***Melanoma* (GEO: GSE72056)** : This dataset measures the expression profile of 4,645 malignant, immune, and stromal cells isolated from 19 freshly procured human melanoma tumors. These cells are classified into 7 major types [183].

***MouseBrain* (GEO: GSE60361)** : This dataset contains the expression profile of 3005 cells from the mouse cortex and hippocampus. These cells classify into 7 major types, including *astrocytes-ependymal, endothelial-mural, interneurons, microglia, oligodendrocytes, pyramidal CA1*, and *pyramidal SS* [219].

***Pollen* (SRA: SRP041736)** : This is a small, but commonly used dataset that contains different cell types in developing cerebral cortex. It consists of 301 cells that classify into 11 distinct cell types [146].

Immune subtype markers I collected immune cell markers for 22 subclasses from a recent paper [129]. This dataset contains a total of 547 markers, spanning 7 different T-cell subtypes, B-cells, NK cells, and myeloid derived subclasses. This dataset is collected and heavily curated from publicly available databases.

Transcriptional Regulatory Network (TRN) I collect transcription factor (TF) – target gene (TG) interactions from the RegNetwork database [110], which aggregates data from 25 different databases. This dataset contains a total of $151,214$ regulatory interactions between $1,408$ TFs and $20,230$ TGs.

4.2.2 Overview of Prior Methods for Cell-type Identification

Various methods have been developed to tackle the problem of cell type identification. **SNN-Cliq** [213] computes a similarity graph among cells, referred to as *shared nearest neighbor (SNN)*. It then uses a graph-based clustering algorithm to identify dense subgraphs. **TSCAN** [82] starts by grouping genes with similar expression patterns into "modules" and represents all cells in this reduced space. It then performs principal component analysis (PCA) over the module space to further reduce dimensions. Finally, cells are clustered by fitting a mixture of multivariate normal distributions to the data, with the number of components estimated using the Bayesian Information Criterion (BIC). **SCUBA** [117] first uses k-means with gap statistic to cluster data along an initial binary tree by analyzing bifurcation events

for time-course data. Then,it refines the tree using a maximum likelihood scheme. BackSPIN [219] is based on SPIN algorithm, which permutes correlation matrix of cell types to extract its underlying structure. BackSPIN then couples it with a divisive splitting procedure to identify clusters from the ordered similarity matrix. Two methods are specifically designed to identify rare cell types. RaceID [65] uses k-means to first cluster cells, with the number of clusters identified using gap statistic. Then, it identifies rare cell types as outliers that are not explained by an appropriate noise model, accounting for both biological and technical variations. GiniClust [83] aims to identify marker genes that are specific to rare cell types using the concept of Gini index. Then, it computes distances between cell types in this reduced subspace and uses DBSCAN clustering algorithm to identify cell types. In addition to these methods, there are approaches that visualize cell types on a continuous spectrum in a given space. Haghverdi *et al.* [68] proposed to use diffusion maps to model the continuous spectrum of cells. On the other hand, Korem *et al.* [96], adopted a previously developed method, called **Pareto task inference (ParTI)** method [70], and applied it to single cell datasets.

4.2.3 Overview and Justification for *ACTION*'s Components

In the following sections, I describe various components of *ACTION*, as shown in Figure 4.1. I first explain exactly how the metric, illustrated in Figure 4.4a, is computed from a matrix of raw cell expression profile data (Step 1 in the overview). Next, I explain how *ACTION* identifies the principal functions of a set of cells, assuming it knows the number of principal functions (Step 3 in the overview). I use an elbow method based on the quality of the principal functions to choose the actual number of principal functions (Step 2 in the overview). Finally, I explain how to estimate the transcriptional regulatory network for a specific principal function (Step 5) by orthogonalizing the functional space of cells (Step 4).

4.2.4 Step 1: A Biologically-inspired Metric for Similarity of Cells

Justification The transcriptome of each cell consists of genes that are expressed at different levels and have different specificity with respect to the underlying cell types. *Housekeeping genes* are the subset of genes responsible for mediating core cellular functions, such as translation, transcription, and DNA repair. These functions are needed by all cells to function properly, which result in ubiquitous expression of these genes across all cell types [45]. While fundamental to cellular function, these genes are not informative with respect to the identity of cells. That is, the fact that a housekeeping gene is expressed in a cell does not provide any information regarding its cell type. On the other hand, cell type-specific genes are preferentially expressed in one or a few selected group of cell types to perform cell type-specific functions. Unlike housekeeping genes, cell type-specific genes are highly relevant for grouping cells according to their common functions. My goal here is to define a similarity measure between cells that suppresses the noise contributed by housekeeping genes and enhances the signal contained in cell type-specific genes.

Suppressing housekeeping genes To suppress the ubiquitously high expression of housekeeping genes, I adopt a method that I developed recently for bulk tissue measurements and extend it to single cell analysis [125]. The core of this method is to project a standardized representation of expression profiles of cells onto the orthogonal subspace of housekeeping genes. Let us denote given expression profiles of cells using matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$, where each row corresponds to a gene and each column represents a cell. I use the shorthand $\boldsymbol{x}_i$ to denote the expression profile of ith cell. In addition, let us denote the signature vector of housekeeping genes by $\boldsymbol{v}$. As a first order estimate, housekeeping signature is computed by taking the average expression over all cells: $\boldsymbol{v} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$. This choice is optimal in a least-square sense when the chance of observing a gene is uniform across all cells. Then, I z-score normalize the profile of each cell: $\boldsymbol{z}_i = \frac{\boldsymbol{x}_i - \mu_i}{\sigma_i}$, where μ_i and σ_i are the mean and sample standard deviation of the entries in the ith cell profile. Similarly, I z-score

normalize the signature vector of housekeeping genes, $\boldsymbol{v}$, to create a new vector $\boldsymbol{z}_v$. Finally, I project out the impact of the housekeeping gene expressions on each cell's profile as follows:

$$\boldsymbol{z}_i^{\perp} = \left(\mathbf{I} - \frac{\boldsymbol{z}_v \boldsymbol{z}_v^T}{\|\boldsymbol{z}_v\|_2^2}\right)\boldsymbol{z}_i. \tag{4.1}$$

This operation projects z_i to the orthogonal complement of the space spanned by the housekeeping genes. I then concatenate the column vectors $\boldsymbol{z}_i^{\perp}$ to create a matrix $\mathcal{Z}^{\perp}$.

Enhancing signal from cell type-specific genes Next, to enhance the signal contributed by preferentially expressed genes, I propose an information theoretic approach, which in essence is similar to the one used previously for marker detection [160]. The idea is to use Shannon's entropy to measure the informativeness of gene expressions. If a gene is uniformly expressed across cells, it contains less information as opposed to the case in which it is selectively expressed in a few cells. To this end, I first shift all entries of $\mathcal{Z}^{\perp}$ by its minimum value to ensure positivity. Then, I normalize this shifted matrix to construct a new matrix $\mathbf{P}$, in which every row has sum one. Let $\boldsymbol{p}_j$ be the row vector associated with the jth gene. Then, I compute the entropy of $\boldsymbol{p}_j$ as: $H(j) = -\sum_j p_{ji} \log(p_{ji})$, where p_{ji} is an entry in the matrix $\mathbf{P}$. Finally, I use these entropy values as a basis to boost contributions from the most informative genes. To this end, I compute a scaling factor for each gene as follows. First, I partition genes as either informative or noninformative by finding the location of the most rapid shift in uniformity values, which resembles a L-shaped curve. Let us denote the entropy of the gene on the edge of this partition by H^*. Then for each gene j, I define a scaling factor as $s_j = H^*/H(j)$. Finally, I compute the kernel matrix as follows:

$$\mathbf{K} = (\boldsymbol{Z}^{\perp})^T \mathbf{diag}(\boldsymbol{w}) \boldsymbol{Z}^{\perp} \tag{4.2}$$

where function **diag**() creates a diagonal matrix from elements of a given vector, and each entry $w_i = s_j^2$. In this formulation, if I denote $\mathbf{Q} = \mathbf{diag}(\boldsymbol{c})\boldsymbol{Z}^{\perp}$, then $\mathbf{K} = \mathbf{Q}^T\mathbf{Q}$ defines a dot-product kernel.

4.2.5 Steps 2 and 3: A Geometric Approach to Identify Principal Functions (Representing Pure Cell Types)

Transcriptional profiles of cells that perform multiple functions can be represented using a limited repertoire of principal functions. The functional space of cells, thus, can be represented by a low-dimensional geometric construct.

The convex hull of a given set of points is the minimum volume polytope that encloses all points. This can be envisioned as a rubber band fitting to the outermost points. The functional space of cells that perform multiple functions can be represented using a limited repertoire of principal functions, which has recently been shown to be embedded within a reduced convex hull [70]. The corners, or archetypes, of this space represent principal functions, associated with specialized groups of cells. Identifying the enclosing convex hull in high-dimensional space is computationally expensive and susceptible to noise and overfitting. As an alternative, I seek a limited number of points on the convex hull that enclose as many points as possible, while being resilient to noise and outliers. To this end, I first use the *successive projection algorithm (SPA)* to identify k transcriptional profiles as initial corners for the covering convex hull, each of which corresponds to a pure cell that is specialized to perform a set of unique principal functions. Then, I use principal convex hull algorithm (PCHA) combined with my distance kernel to adjust these corners by allowing others cells to contribute to the identity of each archetype/corner. This is combined with a standard model selection technique to estimate the number of principal functions.

A quick sketch of my procedure is as follows. I expand on this description in subsequent sections. For each $k = 1, \ldots, K_{\max}$, (i) ***identify potential "pure" cells***: use SPA on the raw expression data $\mathbf{X}$ to find k pure cells that are near

extreme points of the functional space; and (ii) ***adjust the corners***: initialize PCHA using the profiles of those k cells and iterate using the kernel $\mathcal{K}$. Then let $V(k)$ be the PCHA objective function with k archetypes. Finally after all models have been adjusted, (iii) ***estimate the number of cell types*** from $V(k)$ such that it balances the number of cells and the total explained variance.

Estimating "pure" Cells As Extreme Corners of the Functional Subspace of Cells

Given a raw expression matrix $\mathbf{X}$, I aim to identify an "optimal" set $\mathcal{S}$ of k "pure cells." These cells can be viewed as extreme corners of the convex hull of the functional space of cells, and all other samples can be written as convex combinations of these basis vectors. Under a strict assumption, known as *separability*, I seek to identify k columns such that $\mathbf{X} = \mathbf{X}(:, \mathcal{S})\mathbf{H}$, where $\mathcal{S}$ is the selected column subspace of matrix $\mathbf{X}$ and $\mathbf{H}$ is non-negative. This means that every column of $\mathbf{X}$ is a non-negative linear combination of a subset $\mathcal{S}$ of all columns. In terms of cells, this means that every cell's expression profile is a combination of a few cells. However, this is a very strong assumption that rarely holds in real data. A relaxation of this assumption, referred to as *near-separability*, seeks to estimate $\mathbf{X} \approx \mathbf{X}(:, \mathcal{S})\mathbf{H} + \mathbf{N}$, where the noise is bounded: $\|\mathbf{N}(:, j)\|_2 \leq \varepsilon$. This decomposition is known as *near-separable Nonnegative Matrix Factorization (NMF)*. The *Successive Projection Algorithm (SPA)* is an efficient algorithm for solving near-separable NMF with provable performance guarantees [57]. If ε satisfies the technical condition $\varepsilon \leq \mathcal{O}\left(\frac{\sigma_{min}(\mathbf{W})}{\sqrt{k}\kappa^2(\mathbf{W})}\right)$, then:

$$\min_{0 \leq \mathbf{H}} \|\mathbf{X} - \mathbf{X}(:, \mathcal{S})\mathbf{H}\| \leq \mathcal{O}\Big(\epsilon\kappa^2(\mathbf{W})\Big) \tag{4.3}$$

More recently, other techniques have been developed to enhance the robustness of *SPA* to noise [58]. These methods are based on the fact that premultiplying matrix $\mathbf{X}$ by an orthogonal matrix $\mathbf{Q}$ preserves its separability. Thus, by carefully choosing matrix $\mathbf{Q}$, I can enhance the conditioning of the problem. Here, I use the *prewhitening technique*, which uses SVD decomposition of matrix $\mathbf{X}$ to estimate a

noise-reduced approximation matrix. Algorithm 1 presents the SPA algorithm combined with prewhitening technique that I use to estimate a set of k cells.

Input: $\mathbf{X} \in \mathbb{R}^{m \times n}$: expression profile of cells

Output: $\mathcal{S}$: selected subset of columns in matrix $\mathbf{X}$

1: $[\mathbf{U}_k, \mathbf{\Sigma}_k, \mathbf{V}_k] = \mathbf{SVD}(\mathbf{X}, k)$

2: $\widetilde{\mathbf{X}} = \underbrace{\mathbf{\Sigma}_k^{-1}\mathbf{U}_k^T}_{\mathbf{Q}}\mathbf{X} = \mathbf{V}_k^T$ {Prewhitening}

3: $\mathcal{S} = \{\}, \mathbf{R} = \widetilde{\mathbf{X}} \Rightarrow$ *Initialize*

4: **for** $i = \{1, \cdots, k\}$ **do**

5: $\quad \alpha = \mathbf{argmax}_j \|\boldsymbol{r}_j\|_2$ {$\boldsymbol{r}_j$ is the jth column}

6: $\quad \boldsymbol{\beta} = \mathbf{R}(:, \alpha)$

7: $\quad \mathbf{R} \leftarrow (\mathbf{I} - \frac{\boldsymbol{\beta}\boldsymbol{\beta}^T}{\boldsymbol{\beta}^T\boldsymbol{\beta}})\mathbf{R}$ {Orthogonal Projection}

8: $\quad \mathcal{S} \leftarrow \mathcal{S} \cup \{\boldsymbol{\beta}\}$

9: **end for**

Algorithm 1: SPA algorithm with prewhitening

Adjusting Selected Corners to Allow Contributions From All Cells

Archetypal-analysis (AA) [35] can be viewed as a generalization of near-separable NMF. While in near-separable NMF all columns are represented using k columns in $\mathbf{X}$, in AA this constraint is relaxed to be a convex combination of all columns in $\mathbf{X}$. Formally, I can formulate AA as follows:

$$\begin{aligned} \underset{\mathbf{C},\mathbf{H},\boldsymbol{\alpha}}{\text{minimize}} \quad & \|\mathbf{X} - \mathbf{XCH}\| \\ \text{subject to} \quad & \|\mathbf{C}(:, i)\|_1 = 1. \\ & \|\mathbf{H}(:, i)\|_1 = 1. \\ & 0 \leq \mathbf{C}, 0 \leq \mathbf{H} \end{aligned} \tag{4.4}$$

Near-separable NMF is a special case of AA in which $\mathbf{C}$ has exactly k nonzeros and none of the columns have more than one element. The matrix $\mathbf{W} = \mathbf{XC}$ here stores

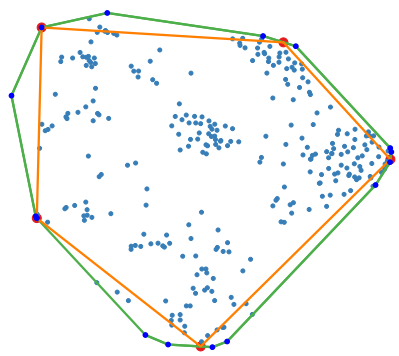


Figure 4.2.: Example of running PCHA algorithm

the *archetypes*. Joint column stochasticity of $\mathbf{C}$ and $\mathbf{H}$ indicates that archetypes are convex combinations of data points, and each data point can be represented as convex combination of archetypes.

There is an algorithm, called *Principal Convex Hull Analysis (PCHA)*, to solve the above problem. The intuition behind PCHA is to fit a polytope to the data points, which approximates the optimal polytope containing as many data points as possible. Figure 4.2 illustrates this phenomena.

I use a kernelized version of PCHA algorithm that minimizes the objective:

$$\text{trace}(-\mathbf{X}^T\mathbf{X}\mathbf{C}\mathbf{H} - \mathbf{H}^T\mathbf{C}^T\mathbf{X}^T\mathbf{X} + \mathbf{H}^T\mathbf{C}^T\mathbf{X}^T\mathbf{X}\mathbf{C}\mathbf{H}) \tag{4.5}$$

in which I directly provide the *ACTION* kernel $\mathcal{K}$ as $\mathbf{X}^T\mathbf{X}$ and initialize $\mathbf{C}$ based on the solution to SPA.

4.2.6 Estimating the Total Number of Archetypes Needed to Represent All Cell Types

A key challenge in all parametric methods is to identify the optimal configuration for associated parameters. In my formulation, the total number of archetypes (corner points) must be provided by the user or directly estimated from the data. To automatically identify this number, one can use various measures of *"goodness"* to

assess overall performance as I increase the number of archetypes. A balance between the number of archetypes and the goodness of solution provides an optimal compromise. I use variance explained by the fit as a measure to find the optimal number of archetypes. For each archetype count (up to a max value), I fit a convex hull to the data and compute explained variance.

The explained variance has an elbow-shape, meaning that it starts increasing rapidly, then it plateaus. The corner of this L-curve is an optimal choice for the number of archetypes. To find this point automatically, I fit a piecewise linear model to the data with two split points. This allows us to distinguish both rapid and more gradual shift patterns in the L-curve. Formally:

$$f(c) = \begin{cases} m_1 c + b_1, & \text{for } 0 \leq c < c_i \\ m_2 c + b_2, & \text{for } c_i \leq c < c_j \\ m_3 c + b_3, & \text{for } c < c_j \leq c_{max} \end{cases} \tag{4.6}$$

where c is the archetype count and c_i and c_j are two free parameters. I evaluate every pair of $(c_i, c_j); 1 \leq c_i < c_j \leq c_{max}$ and fit a minimum least squares fit to each piece. The configuration with minimum overall error is selected as c_i^{best} and c_j^{best}. For this specific configuration, let m_2 and m_3 represent the slope of the second and the third linear fits. Then, if $\frac{m_2}{m_3}$ is less that or equal to a user-defined parameter $\text{threshold}_{\text{min}}$, then I select the first split point (c_i^{best}). Otherwise, I have a rapidly shifting curve and the slopes of second and third segments are very close. Thus, I select the second split point as the choice of k. Figure 4.3 illustrates an example of fitting process. The pink dots represent the explained variance for archetypal fits with increasing number of archetypes. Green lines show the piecewise linear fit to the data. The optimal number of archetypes is selected according to $best_j$ in this case, which is nine.

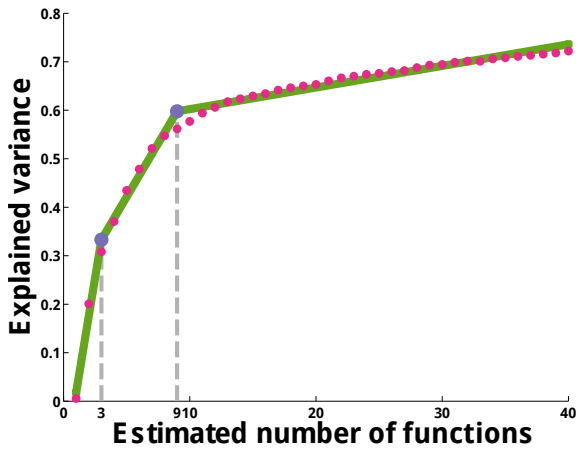


Figure 4.3.: Illustration of identification of total number of functions for the *Pollen* dataset

4.2.7 Steps 4 and 5: Constructing the Transcriptional Regulatory Network Corresponding to Each Archetype

Each archetype represents a principal function performed by a group of cells. However, what makes these functions unique and the functional specializations they represent is not clear from the archetype signatures. To identify marker genes in each archetype, and to shed light on the underlying network regulating the observed transcriptional phenotype, I developed a novel approach based on orthogonalizing the space of principal functions.

Archetype orthogonalization to Identify Cell Type-specific Markers

A key factor in analyzing principal functions represented by each archetype is to identify what distinguishes one archetype from others. To identify shared and unique aspects represented by each archetype, I present a new method, called *arechetype orthogonalization*. The idea is to remove effects that are shared with any other archetypes before analyzing a given archetype.

Recall the result of PCHA is $\mathbf{C}$ and $\mathbf{H}$. The result $\mathbf{XC}$ represents the archetypes in the space of gene expression profiles. Let us denote the vector representation of archetype i by $\boldsymbol{a}_i$ and let $\mathbf{A}$ be the matrix of all archetypes. Let $\mathbf{A}_{-i}$ denote the matrix of archetypes without the ith column. Then our goal is to project $\boldsymbol{a}_i$ into the subspace orthogonal to the columns spanned by $\mathbf{A}_{-i}$. This can be computed as:

$$\boldsymbol{a}_i^{\perp} = \Big(\mathbf{I} - \mathbf{A}_{-i}(\mathbf{A}_{-i}^T\mathbf{A}_{-i})^{-1}\mathbf{A}_{-i}^T\Big)\boldsymbol{a}_i \tag{4.7}$$

For each archetype, I can sort all genes according to their *"residual expression"* after orthogonalization.

Identifying Cell Type-specific Transcriptional Regulatory Network (TRN)

Given *residual expression* vectors for each archetype, I can identify key regulatory circuits responsible for the observed transcriptional phenotype. I construct induced subgraphs of the global transcriptional regulatory network (TRN), which drive characteristic behavior of each cell type. First, I order all genes according to their residual expression for a given archetype. Then, for each transcription factor (TF), I identify the over-representation of its target genes (TGs) among top-ranked genes with respect to that archetype. To this end, I use minimum hypergeometric (mHG) p-value. This method is nonparametric, in the sense that I do not need to predefine a fixed cut. Let us represent the total number of genes by m. Given a set of target genes, of size T, I construct a binary vector of true positives (targets) as $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, ...\lambda_m] \in \{0,1\}^m$. Let the random variable Z denote the number of target genes among a fixed number of l top-ranked genes, if I distribute genes randomly. In this formulation, I can express the p-value in terms of the hypergeometric distribution:

$$\begin{aligned} p\text{-value}(Z = b_l(\lambda)) &= \text{Prob}(b_l(\lambda) \leq Z) \\ &= \text{HGT}(b_l(\lambda)|m, T, l) \\ &= \sum_{x=b_l(\lambda)}^{min(T,l)} \frac{\binom{T}{x}\binom{m-T}{l-x}}{\binom{m}{l}} \end{aligned} \tag{4.8}$$

where HGT is the tail of hypergeometric distribution and $b_l(\lambda) = \sum_{i=1}^{l} \lambda_i$ counts the total number of true positives in top-l observations. The drawback of this approach is that I still need a predefined cutoff value, l. To remedy this, Eden *et al.* [43] proposed a two-step process for computing the exact enrichment p-value, called *mHG p-value*, without the need for a predefined cutoff value of l. First, an optimal cutoff value is chosen among all possible values of $1 \leq l \leq N$. The computed value for this optimal cutoff is called the *minimum hypergeometric (mHG) score*, and is defined as:

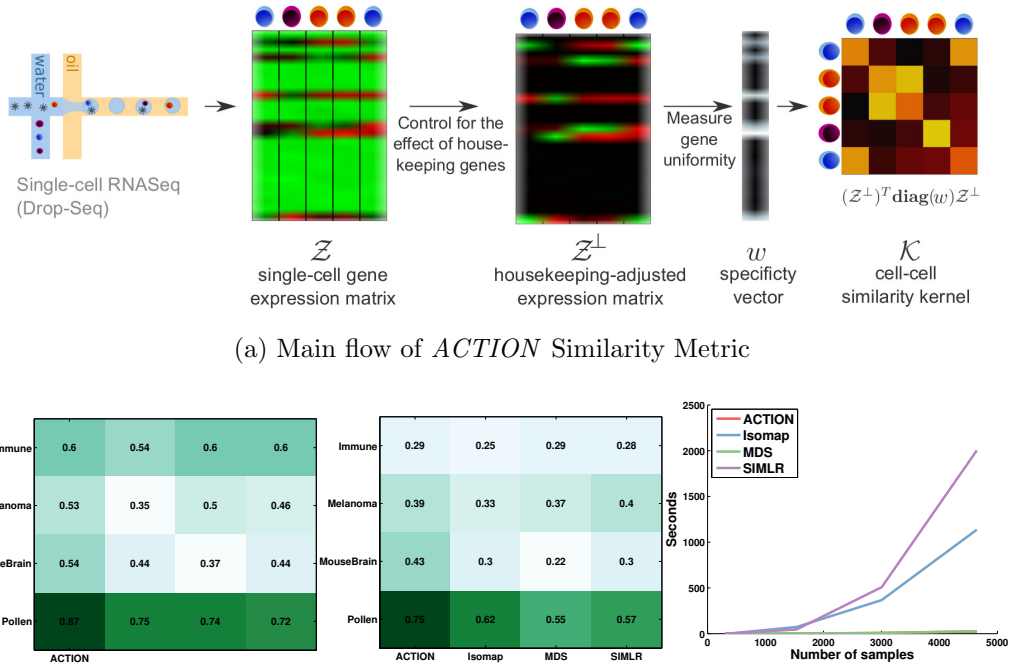
$$\text{mHG}(\lambda) \quad = \quad \min_{1 \leq l \leq m} p\text{-value}(Z = b_l(\lambda)) \tag{4.9}$$

Next, a dynamic programming (DP) method is used to compute the exact p-value of the observed mHG score, in the state space of all possible λ vectors of size m having exactly T ones.

I use this formulation to identify significant transcription factors based on the number of target genes (TGs) with high residual expression. This, in turn, splits TGs of each TF into top vs bottom-ranked genes. I then select all significant TFs, together with their top-ranked target genes and construct a node-weighted induced subgraph of the global TRN, which represents the cell type-specific TRN.

4.3 Results and Discussion

The *ACTION* framework consists of three major components, shown in Figure 4.1: (i) A robust measure of cell-to-cell similarity, (ii) A geometric approach for identification of principal functions, and (iii) a statistical framework for constructing cell-type specific transcriptional regulatory networks (TRNs). My cell-to-cell similarity metric is rooted in the notion that functional roles of a cell form an embedded hierarchy, with successively refined set of tissue-specific functions. When used with a classic clustering algorithm such as k-means, *ACTION* metric surpasses all other measures of cell similarity in identifying cell types. The next component of my method is a geometric approach for identifying principal functions of cells, each represented by an archetype (corner) of the convex hull in the functional space of cells. Finally,



(a) Main flow of *ACTION* Similarity Metric

(b) Normalized Mutual Information

(c) Adjusted Rand Index

(d) Timing (s)

Figure 4.4.: Evaluation of *ACTION* Similarity Metric

ACTION uses a novel method that utilizes the geometric view of cell functions to construct the transcriptional regulatory network (TRN) that mediates characteristic behavior of each cell type. In what follows, I describe, validate, and discuss each component in detail.

4.3.1 Component 1: Measuring Cell-to-cell Similarity

An essential component of any method for identifying cell types is the ability to quantify similarity between individual cells. Most prior methods rely on traditional measures, such as Euclidean distance, that are not specifically targeted towards transcriptomic profiles. In contrast, I define a similarity metric, or formally a kernel, specifically designed for measuring similarity between cells [125]. My approach is based on the observation that housekeeping genes, while not informative of cell type

identity, significantly impact traditional measures of cell similarity due to their ubiquitous and high expression levels. Suppressing these genes significantly enhances the signal-to-noise ratio (SNR) in expression profiles, allowing us to extract a stronger cell type-specific signal.

Novel methodology my method starts by projecting transcriptional signatures to the orthogonal subspace spanned by housekeeping genes. I then boost the contribution of cell type-specific genes using an information theoretic approach. Finally, I combine these two measures to define a robust measure of cell-to-cell similarity. This approach is illustrated in Figure 4.4. The mathematical models underlying the metric are described in the Methods section.

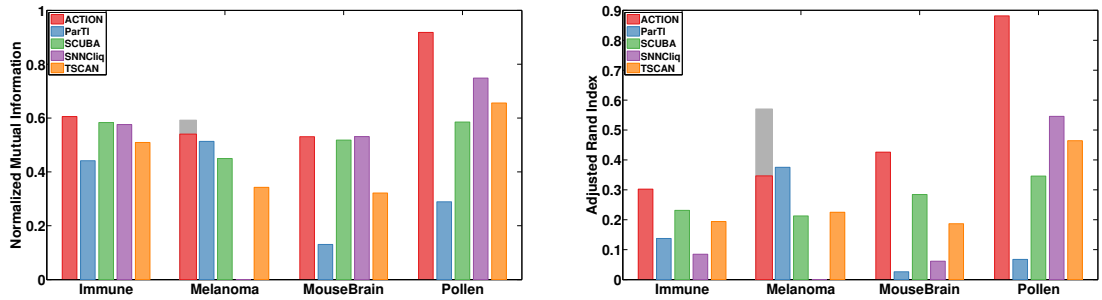
Validation To establish the superiority of my metric, I compare it against one measure specifically designed for single cell analysis, *SIMLR*, and two general measures: *multidimensional scaling (MDS)*, and Isomap. SIMLR [205], combines a number of distance metrics to learn a joint similarity score that maximizes the block diagonal structure of the resulting matrix. Both *MultiDimensional Scaling (MDS)* and *Isomap* are nonlinear dimension reduction techniques. The former method projects points into a low-dimensional space, such that distances between samples are preserved to the extent possible. The latter method first computes the nearest neighborhood graph of data points. It then uses shortest path between vertices as a measure of distance between them. Finally it uses MDS to embed these distances in a low-dimensional space. After projecting the data to a lower dimension space in either MDS or Isomap, one can use linear correlation in the transformed subspace to measure similarity between cells. While *ACTION* is a non-parametric method, other methods need additional input. For SIMLR, I need to provide the true number of cell types. In order to give the other methods the best chance at competing with *ACTION*, I evaluate them using ten different values for dimension of projected subspace (from 5 to 50 with increments of 5) and report the best results obtained over all configurations.

To assess the quality of computed similarities between cells, I use each of the four measures to cluster cells and identify cell types. Each cluster is assumed to represent a unique cell type, and the clusters are determined using the commonly used kernel k-means algorithm. I compare the computed cell types with the true (known) cell types in terms of *Normalized Mutual Information (NMI)* and *Adjusted Rand Index (ARI)*. Normalized Mutual Information is an information theoretic measure that is zero for random clustering (when the identified clustering contains no information about true cell types), and one for a clustering that perfectly matches a given gold standard. The ARI measure is also between zero and one; however, it evaluates the cases in which a given pair of cells are either co-clustered in both true and identified, or classified separately in both.

In each case, I perform 100 independent clusterings with random initialization and report the average of NMI and ARI scores as quality measures (relative ordering of results is robust with respect to other aggregating functions, such as median or max). These experiments are independently performed for each dataset. Figures 4.4b-d present the performance of the cell type identification technique operating with different similarity measures, both in terms of their clustering quality (NMI and ARI) and total running time.

Discussion of results on similarity metric To evaluate performance of each similarity metric, I analyzed four different datasets, which are listed in Section 5.2.1. These datasets have different number of cells, ranging from hundreds to thousands, span a wide range of normal and cancerous cells, and are measured using different single cell technologies.

For both *MouseBrain* and *Pollen* datasets, *ACTION* metric significantly outperforms other metrics in terms of both NMI and ARI measures. For the *Melanoma* dataset, *ACTION* has significantly better NMI, but there is a tie between *ACTION*, MDS, and SIMLR with respect to the ARI measure. Finally, for the *Immune* dataset, there is a tie between *ACTION*, MDS, and SIMLR for both measures. In all studies,



(a) NMI (b) ARI

Figure 4.5.: Performance of *ACTION* in identifying cell types

t-test with p-val $\leq 10^{-2}$ has been used to assess significance of difference between observed NMI/ARI values. In summary, my results demonstrate that in all cases *ACTION* metric is either significantly better or at least as good as any other methods. Thus establishes the *ACTION* metric as a *fast*, *nonparametric*, and *accurate* method for computing similarity among single cells. I use this measure throughout the rest of my study. I note however, that my overall framework is flexible with respect to choice of other similarity metrics.

4.3.2 Component 2: A Geometric View to Identify Discrete Cell Types

Novel methodology Using the *ACTION* metric as a measure of similarity between cells, I develop a new method for identifying *de novo* cell types in a given experiment. My method is based on a geometric interpretation of cellular functions. Each cell is a data-point in a high-dimensional space. My method identifies *"extreme"* corners in this space, and each cell is characterized by its distance to every corner. The corners identified by *ACTION* represent "pure" cells that are specialized to perform a principal function. This is in contrast to methods such as unsupervised clustering (e.g., k-medoids) that identify *the most common* centers. My focus on identifying the extreme points (and thus, principal functions), allows us to better identify rare cell types.

Validation. Each corner or archetype represents a principal function. I first validate these by considering each archetype as a *characteristic cell type*. I then identify the type of each cell by determining the closest archetype and assigning this type. I compare my method to four recently proposed methods: SCUBA [117], SNNCliq [213], single-cell ParTI [70, 96], and TSCAN [82]. Details of these methods are given in the methods section. Clique size and density of quasi cliques of *SNN_Cliq* are left as default parameters ($k = 3$ and $r = 0.7$). Increasing clique size k did not improve performance, but significantly increased the running time. With these parameters, *SNNCliq* did not terminate in *72h* for the largest dataset (Melanoma), after which I stopped the experiment. I present a comprehensive analysis of the results for all other combinations of datasets/methods.

Discussion of results on cell-type identification. Figure 4.5 shows comparative performance of different methods in predicting cell types in various datasets. In all cases, except ARI for the *Melanoma* dataset, *ACTION* yields superior results compared to the state-of-the-art methods for cell-type identification. In general, NMI measure exhibits lower range of variation across methods, whereas ARI has a higher range of variability. To further investigate the difference between *ParTI* and *ACTION* on the Melanoma dataset, I manually evaluated each archetype identified in these methods. My results indicate that the source of difference is that *ACTION* identifies more refined subtypes of T-cells and subclasses of tumor cells, whereas *ParTI* combines these subtypes/classes. These subgroup details are missing from the annotations provided for the dataset by authors. Combining cell types that are classified as different subtypes of T-cells or subclasses of tumor cells significantly enhances the computed performance measures of *ACTION* in this dataset. This is shown using gray boxes in the corresponding figure.

Analysis and validation of the principal functions. While cells can be classified based on their closest archetype, they can also be viewed on a continuum [70]. To illustrate this *continuous view*, I use the distance from each archetype as a low-

dimensional embedding of the cells. I use the Fielder embedding, followed by adjustment using Stochastic Neighbor Embedding (SNE) method to visualize this low-dimensional embedding in Figure 4.6. Each archetype is marked with a text labeled (A1, . . . , A11) point and assigned a unique color. Each point corresponds to a cell. I interpolate its color using its distance to all archetypes to highlight the continuous nature of the data. The labels for the groups are based on three sources. First, I perform enrichment analysis on the cells assigned to each archetype. Then, I use markers provided in the original datasets to identify the cell type-specific expression in each archetype. Finally, I use markers from *LM22* dataset [129] to classify subtypes of immune cells.

Figure 4.6 illustrates the ability of my method to identify both isolated cell-types with specialized principal functions, as well as cells with a combination of functions. As an example, different subclasses of T-cell constitute a spectrum with the corners (or archetypes) representing specialized functions that are performed by a pure T-cell subtype. In addition to given cell types, I also find an additional archetype, *A6*, which links between T-cells and B-cells and I hypothesize to be a lymphocyte progenitor.

In terms of tumor cells, many of the patients form their own archetypes. The two exceptions to this rule, A5 and A10, define a "*MITF axis*", which is shown in the subfigure (MITF is one of the transcription factors known be related to various types of Melanoma [183,201]). Archetype *A5* is enriched in five patients with varying degrees of expression for MITF from mid to high. I collectively refer to patients in Archetype A5 as *MITF-associated* patients. Archetype *A10*, on the other hand, contains patients 81 and 82, both of who have low levels of MITF. In what follows, I construct the transcriptional regulatory network responsible for mediating observed phenotype of MITF-associated patients in A5.

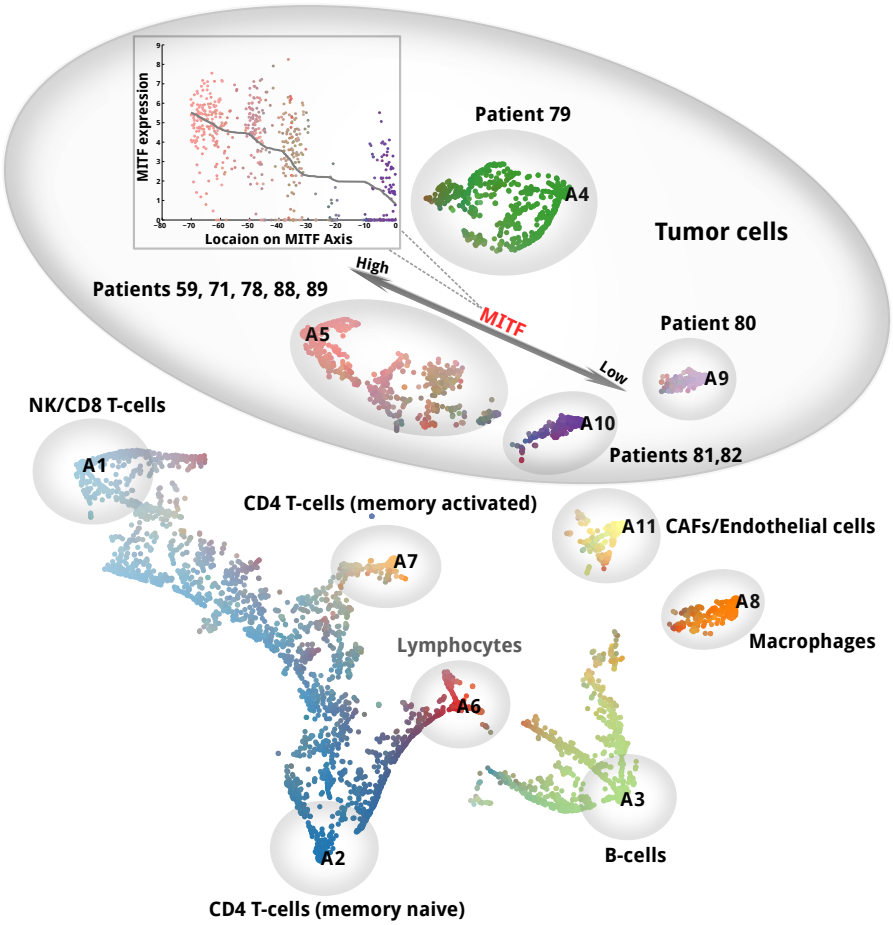


Figure 4.6.: A continuous view of cell types in the Melanoma dataset identifies subclasses of immune cells and highlights a MITF-related "axis"

4.3.3 Component 3: Constructing Subclass-specific Transcription Regulatory Network of MITF-associated Patients

Novel Methodology I propose a new method to construct regulatory pathways responsible for mediating phenotypes associated with each archetype. To this end, I first perform an *archetype orthogonalization* (details described in Section 4.2.7), to compute residual expression and identify marker genes that are unique to the archetype. Then, I rank all genes according to their *residual expression*. Finally, I project these scores to the transcriptional regulatory network (TRN) to find key transcription fac-

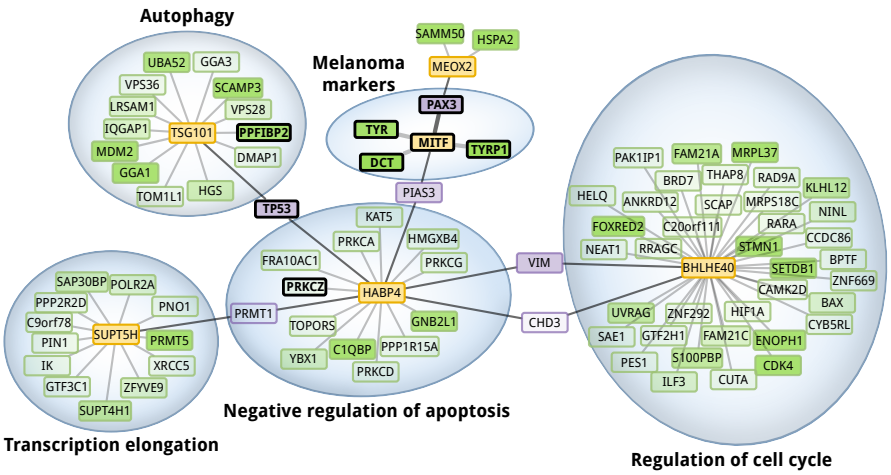


Figure 4.7.: The transcriptional regulatory network (TRN) for MITF-associated Melanoma patients highlights a number of genes that have not previously been associated with Melanoma – along with some known markers

tors (TFs) responsible for mediating the observed transcriptional phenotype. For each TF, I assess the over-representation of its targets among top-ranked genes (according to the residual expression score). I use a dynamic programming algorithm [43] to assign exact p-values to each TF. For each TF, its "top ranked" target genes, according to the cut that yields the minimum hypergeometric score, are also selected as part of the regulatory network.

I apply this technique to identify regulatory pathways of MITF-associated samples. A p-value threshold of 0.05 is used to identify significant TFs. The final constructed network is presented in Figure 4.7. This network consists of six key transcription factors (in yellow), 85 target genes (in green/purple). Purple nodes are target genes that are jointly regulated by two TFs. I marked enriched functions of each group in the figure, accordingly, and highlighted elements that are already known to be associated with Melanoma.

Validation *MITF* is one of the best-characterized markers for Melanoma, and is also used in the original paper to classify patients [183]. It is notable here that my

method identified MITF directly using data from the activity of its targets. Furthermore, since these transcription factors are identified based on the activity of their target, they are "related" to the subclasses, however, the mechanism of their control can be diverse.

Among other factors, *BHLHE40* has the highest number of activated targets. This factor, among other functions, regulates *M-MITF*, a melanocyte-restricted isoform of *MITF*, and potently reduces expression of *MITF* under hypotoxic conditions [48]. Angiogenesis, or growth of blood vessels, is a hallmark of cancer. *MEOX2* plays multiple roles in this process. At low levels, it activates nuclear factor-κB (NK-κB), a proangiogenic signaling pathway, whereas in high doses, it has an inhibitory role [27]. Similarly, *TSG101* plays different roles depending on the context. In fibroblasts, it acts as a tumor suppressor gene, whereas it has a tumor-enhancing role in some epithelial tumors. This bidirectional regulation is postulated to be through expression of *MMP-9*in different cell types [159]. The role of other factors is less-studied.

Experimental evidence To further validate my results, I use the transcriptome of 10 patients with *invasive* and *proliferative* melanoma subtypes from Verfaillie *et al.* [201]. *Proliferative* subtype is characterized by high levels of *MITF*, as well as *SOX10* and *PAX3*. In contract, *invasive* subtype is known to have low levels of *MITF* and high levels of epithelial-to-mesenchymal (EMT) transcription factor ZEB1, and is associated with metastatic dissemination. Nodes in my *MITF*-associatied TRN resemble the *proliferative* subtype. Thus, I use marker genes for this class to validate my results. There are a total of 770 marker genes for the *proliferative* subtype and among 91 total genes in my network, 8 genes coincide with them (p-value = 0.01). These genes include *DCT, MITF, PAX3, PPFIBP2, PRKCZ, TP53, TYR,* and *TYRP1*, all of which have high residual expression compared to all other nodes. Beside the *MITF* subnetwork, *TP53, PRKCZ,* and *PPFIBP2* are also enriched in this set. Interestingly, a key factor involved in the *invasive* subtype, *MEOX2*, is also identified as a node in my network. As mentioned earlier, depending on the level

of its expression, this gene can play different roles for proliferative versus invasive subclasses.

Collectively, these results illustrate the effectiveness of the *ACTION* in identifying novel cancer subtypes, their underlying regulatory network, and characteristic markers. This, in turn, presents new avenues for diagnosis and prognosis of melanoma patients, as well as new therapeutic targets for further investigation.

5 SEPARATING CELL TYPES AND THEIR RELATIVE PERCENTAGES FROM COMPLEX TISSUES

5.1 Background

Source separation, or deconvolution, is the problem of estimating individual signal components from their mixtures. This problem arises when source signals are transmitted through a mixing channel and the mixed sensor readings are observed. Source separation has applications in a variety of fields. One of its early applications is in processing audio signals [136, 173, 204, 218]. Here, mixtures of different sound sources, such as speech or music, are recorded simultaneously using several microphones. Various frequencies are convolved by the impulse response of the room and the goal is to separate one or several sources from this mixture. This has direct applications in speech enhancement, voice removal, and noise cancellation in recordings from populated areas. In hyperspectral imaging, the spectral signature of each pixel is observed. This signal is the combination of pure spectral signatures of constitutive elements mixed according to their relative abundance. In satellite imaging, each pixel represents sensor readings for different patches of land at multiple wavelengths. Individual sources correspond to reflectances of materials at different wavelengths that are mixed according to the material composition of each pixel [57, 113, 131, 134, 203].

Beyond these domains, deconvolution has applications in removing noise from biomedical sensors. Tracing electrical current in the brain is widely used as a proxy for spatiotemporal patterns of brain activity. These patterns have significant clinical applications in diagnosis and prediction of epileptic seizures, as well as characterizing different stages of sleep in patients with sleep disorders. Electroencephalography (EEG) and magnetoencephalography (MEG) are two of the most commonly used techniques for cerebral imaging. These techniques measure voltage fluctuations and

changes in the electromagnetic fields, respectively. Superconducting QUantum Interference Device (SQUID) sensors used in the latter technology are susceptible to magnetic coupling due to geometry and must be shielded carefully against magnetic noise. Deconvolution techniques are used to separate different noise sources and to ameliorate the effect of electrical and magnetic coupling in these devices [72, 179, 199, 220].

At a high level, mixing channels can be classified as: (i) linear or nonlinear, (ii) instantaneous, delayed, or convolutive, and (iii) over/under determined. When neither the sources nor the mixing process is available, the problem is known as blind source separation (BSS). This problem is highly under-determined in general, and additional constraints; such as independence among sources, sparsity, or non-negativity; are typically enforced on the sources in practical applications. A new class of methods has been developed recently, known as semi or guided BSS [72, 136, 173, 204]. In these methods, additional information is available a priori on the approximate behavior of either sources or the mixing process. In this chapter, I focus on the class of over-determined, linear instantaneous (LI) mixing processes, for which a deformed prior on sources is available. In this case, the parameters of the linear mixer, as well as the true identity of the original sources are to be determined.

In the context of molecular biology, deconvolution methods have been used to identify constituent cell-types in a tissue, along with their relative proportions. The inherent heterogeneity of tissue samples makes it difficult to identify separated, cell-type specific signatures, i.e., the precise gene expression levels for each cell-type. Relative changes in cell proportions, combined with variations attributed to the changes in the biological conditions, such as disease state, complicate identification of true biological signals from mere technical variations. Changes in tissue composition are often indicative of disease progression or drug response. For example, coupled depletion of specific neuronal cells with the gradual increase in the glial cell population is indicative of neurodegenerative disorders. An increasing proportion of malignant cells, as well as a growing fraction of tumor infiltrating lymphocytes (TIL) compared to surrounding cells, directly influence tumor growth, metastasis, and clinical outcomes

for patients [100,129]. Deconvolving tissue biopsies allows further investigation of the interaction between tumor and micro-environmental cells, along with its role in the progression of cancer.

The expression level of genes, which is a proxy for the number of present copies of each gene product, is one of the most common source factors used for separating cell-types and tissues. In the linear mixing model, the expression of each gene in a complex mixture is estimated as a linear combination of the expression of the same gene in the constitutive cell-types. *In silico* deconvolution methods for separating complex tissues can be coarsely classified as either *full deconvolution*, in which both cell-type specific expressions and the percentages of each cell-type are estimated, or *partial deconvolution* methods, where one of these data sources is used to compute the other [164]. These two classes loosely relate to BSS and guided-BSS problems. Note that in cases where relative abundances are used to estimate cell-type-specific expressions, the problem is highly under-determined. In the complementary case of computing percentages from noisy expressions of purified cells, the problem is highly over-determined. In the former case, I typically have only a handful of known cell types with known percentages, and I need to estimate unknown expression values for thousands of genes. In the latter case, I use known expression values of all genes (or a selected subset that is typically much larger than the number of cell types) to compute percentages of a small population of constituting cells. Thus, in the case of an over-determined system, the key is to select the most reliable features that satisfy the linearity assumption. I provide an in-depth review of recent deconvolution methods in Section 5.2.3.

In contrast to computational methods, a variety of experimental cell separation techniques have been proposed to enrich samples for cell-types of interest. However, these experimental methods not only involve significant time, effort, and expense, but may also result in insufficient RNA abundance for further quantification of gene expression. In this case, amplification steps may introduce technical artifacts into the gene expression data. Furthermore, sorting of cell-types must be embedded in

the experiment design for the desired subset of cells, and any subsequent separation is infeasible. Computational methods, on the other hand, are capable of sorting mixtures at different levels of resolution and for arbitrary cell-type subsets of interest.

The organization of the remainder of the chapter is as follows: The formal definition of the deconvolution problem and its relationship to linear regression is defined in Section 5.2.2. Sections 5.2.2 and 5.2.2 review different choices and examples of the objective function used in regression. An overview of computational methods for biological deconvolution is provided in Section 5.2.3. Datasets and evaluation measures used in this study are described in Sections 5.2.1 and 5.2.4, respectively. The effect of the loss function, constraint enforcement, range filtering, and feature selection choices on the performance of deconvolution methods is evaluated systematically in Sections 5.3.1-5.3.5.

5.2 Materials and Methods

5.2.1 Datasets

In Vivo Mixtures With Known Percentages

I use a total of five datasets with known mixtures. I use CellMix to download and normalize these datasets [54], which uses the *soft* format data available from Gene Expression Omnibus (GEO).

- **BreatBlood** [61] (GEO ID: *GSE29830*): Breast and blood from human specimens are mixed in three different proportions and each of the mixtures is measured three times, with a total of nine samples.

- **CellLines** [2] (GEO ID: *GSE11058*): Mixture of human cell lines Jurkat (T cell leukemia), THP-1 (acute monocytic leukemia), IM-9 (B lymphoblastoid multiple myeloma) and Raji (Burkitt B-cell lymphoma) in four different concentrations, each of which is repeated three times, resulting in a total of 12 samples.

- **LiverBrainLung** [165] (GEO ID: *GSE19830*): This dataset contains three different rat tissues, namely brain, liver, and lung tissues, which are mixed in 11 different concentrations with each mixture having three technical replicates, for a total of 33 samples.

- **RatBrain** [101] (GEO ID: *GSE19380*): This contains four different cell-types, namely rat's neuronal, astrocytic, oligodendrocytic and microglial cultures, and two replicates of five different mixing proportions, for a total of 10 samples.

- **Retina** [166] (GEO ID: *GSE33076*): This dataset pools together retinas from two different mouse lines and mixed them in eight different combinations and three replicates for each mixture, resulting in a total of 24 samples.

Mixtures With Available Cell-sorting Data Through Flow-cytometry

For this experiment, I use two datasets available from Qiao *et al.* [148]. I directly download these datasets from the supplementary material of the paper. These datasets are post-processed by the supervised normalization of microarrays (SNM) method to correct for batch effects. Raw expression profiles are also available for download under GEO ID *GSE40830*. This dataset contains two sub-datasets:

- **PERT_Uncultured**: This dataset contains uncultured human cord blood mono nucleated and lineage depleted (Lin-) cells on the first day.

- **PERT_Cultured**: This dataset contains culture-derived lineage-depleted human blood cells after four days of cultivation.

Table 5.1 summarizes overall statistics related to each of these datasets.

5.2.2 Deconvolution: Formal Definition

I introduce formalisms and notation used in discussing different aspects of *in silico* deconvolution of biological signals. I focus on models that assume *linearity*, that is,

Table 5.1: Summary statistics of each dataset for deconvolution

Dataset	# features	# samples	# references
BreastBlood	54675	9	2
CellLines	54675	12	4
LiverBrainLung	31099	33	3
PERT_Cultured	22215	2	11
PERT_Uncultured	22215	4	11
RatBrain	31099	10	4
Retina	22347	24	2

the expression signature of the mixture is a weighted sum of the expression profile for its constitutive cell-types. In this case, sources are cell-type specific references and the mixing process is determined by the relative fraction of cell-types in the mixture.

I first introduce the mathematical constructs used:

- $\mathbf{M} \in \mathbb{R}^{n \times p}$: Mixture matrix, where each entry $\mathbf{M}(i, j)$ represents the raw expression of gene i, $1 \leq i \leq n$, in heterogeneous sample j, $1 \leq j \leq p$. Each sample, represented by $\boldsymbol{m}$, is a column of the matrix $\mathbf{M}$, and is a combination of gene expression profiles from constituting cell types in the mixture.

- $\mathbf{H} \in \mathbb{R}^{n \times r}$: Reference signature matrix for the expression of primary cell types, with multiple biological/technical replicates for each cell-type. In this matrix, rows correspond to the same set of genes as in $\mathbf{M}$, columns represent replicates and there is an underlying grouping among columns that collects profiles corresponding to the same cell-type.

- $\mathbf{G} \in \mathbb{R}^{n \times q}$: Reference expression profile, where the expression of similar cell-types in matrix $\mathbf{H}$ is represented by the average value.

- $\mathbf{C} \in \mathbb{R}^{q \times p}$: Relative proportions of each cell-type in the mixture sample. Here, rows correspond to cell-types and columns represent samples in mixture matrix $\mathbf{M}$.

Using this notation, I can formally define deconvolution as an optimization problem that seeks to identify "optimal" estimates for matrices $\mathbf{G}$ and $\mathbf{C}$, denoted by $\hat{\mathbf{G}}$ and $\hat{\mathbf{C}}$, respectively. Since $\mathbf{G}$ and/or $\mathbf{C}$ are not known a priori, I use an approximation that is based on the linearity assumption. In this case, I aim to find $\hat{\mathbf{G}}$ and $\hat{\mathbf{C}}$ such that their product is close to the mixture matrix, $\mathbf{M}$. Specifically, given a function $\boldsymbol{\delta}$ that measures the distance between the true and approximated solutions, also referred to as the loss function, I aim to solve:

$$\min_{0 \leq \hat{\mathbf{G}}, \hat{\mathbf{C}}} \boldsymbol{\delta}(\hat{\mathbf{G}}\hat{\mathbf{C}}, \mathbf{M}) \tag{5.1}$$

In partial deconvolution, either $\mathbf{C}$ or $\mathbf{G}$, or their noisy representation, is known a priori and the goal is to find the other unknown matrix. When matrix $\mathbf{G}$, referred to as the *reference profile*, is known, the problem is over-determined and I seek to distinguish features (genes) that closely conform to the linearity assumption, from the rest of the (variable) genes. In this case, I can solve the problem individually for each mixture sample. Let us denote by $\boldsymbol{m}$ and $\hat{\boldsymbol{c}}$ the expression profile and estimated cell-type proportion of a mixture sample, respectively. Then, I can rewrite Equation 5.1 as:

$$\min_{0 \leq \hat{\boldsymbol{c}}} \boldsymbol{\delta}(\mathbf{G}\hat{\boldsymbol{c}}, \boldsymbol{m}) \tag{5.2}$$

This formulation is essentially a linear regression problem, with an arbitrary loss function. On the other hand, in the case of full deconvolution, I can still estimate $\mathbf{C}$ in a column-by-column fashion. However, estimating $\mathbf{G}$ is highly under-determined and I must use additional sources to restrict the search space. One such source of information is the variation across samples in $\mathbf{M}$, depending on the cell-type concentrations in the latest estimated value of $\mathbf{C}$. In general, most regression-based methods for full deconvolution use an iterative scheme that starts from either noisy estimates of $\mathbf{G}$ and $\mathbf{C}$, or a random sample that satisfies given constraints on these matrices, and

successively improves over this initial approximation. This iterative process can be formalized as follows:

$$\begin{aligned} \hat{\mathbf{C}} &\leftarrow \underset{0 \leq \hat{\mathbf{C}}}{\mathbf{argmin}}(\boldsymbol{\delta}(\hat{\mathbf{G}}\hat{\mathbf{C}} - \mathbf{M})) \\ \hat{\mathbf{G}} &\leftarrow (\underset{0 \leq \hat{\mathbf{G}}}{\mathbf{argmin}}(\boldsymbol{\delta}(\hat{\mathbf{C}}^T\hat{\mathbf{G}}^T - \mathbf{M})^T))^T \end{aligned} \tag{5.3}$$

Please note that the updating $\hat{\mathbf{G}}$ is typically row-wise (for each gene), whereas updating $\hat{\mathbf{C}}$ is column-wise (for each sample). Non-negative matrix factorization (NMF) is a dimension reduction technique that aims to factor each column of the given input matrix as a nonnegative weighted sum of non-negative basis vectors, with the number of basis vectors being equal or less than the number of columns in the original matrix. The *alternating non-negative least squares formulation (ANLS)* for solving NMF can be formulated using the framework introduced in Equation 5.3. There are additional techniques for solving NMF, including the *multiplicative updating rule* and the *hierarchical alternating least squares (HALS)* methods, all of which are special cases of block-coordinate descent [91]. Two of the most common loss functions used in NMF are the Frobenius and Kullback-Leibler (KL) divergence [91].

In addition to *non-negativity (NN)*, an additional *sum-to-one (STO)* constraint is typically applied over columns of the matrix $\hat{\mathbf{C}}$, or the sample-specific vector $\hat{\boldsymbol{c}}$. This constraint restricts the search space, which can potentially enhance the accuracy of the results. It also simplifies the interpretation of values in $\hat{\boldsymbol{c}}$ as relative percentages. Finally, another fundamental assumption that is mostly neglected in prior work is the **similar cell quantity (SCQ)** constraint. The similar cell quantity assumption states that all reference profiles and corresponding mixtures must be normalized to ensure that they represent the expression level of the "same number of cells." If this constraint is not satisfied, differences in the cell-type counts directly affect concentrations by rescaling the estimated coefficients to adjust for the difference.

In this chapter, I focus on different loss functions ($\boldsymbol{\delta}$ functions), as well as the role of constraint enforcement strategies, in estimating $\hat{\boldsymbol{c}}$. These constitute the key building blocks of both partial and full deconvolution methods.

Choice of Objective Function

In linear regression, often a slightly different notation is used, which I describe here. I subsequently relate it to the deconvolution problem. Given a set of samples, $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$, where $\boldsymbol{x}_i \in \mathbb{R}^k$ and $y_i \in \mathbb{R}$, the regression problem seeks to find a function $f(\boldsymbol{x})$ that minimizes the aggregate error over all samples. Let us denote the fitting error by $r_i = y_i - f(\boldsymbol{x}_i)$. Using this notation, I can write the regression problem as:

$$\underset{f \in \mathcal{F}}{\textbf{argmin}} \sum_{i=1}^{m} \boldsymbol{\mathcal{L}}(r_i) \tag{5.4}$$

where the *loss function* $\boldsymbol{\mathcal{L}}$ measures the cost of estimation error. I focus on the class of linear functions, that is $f_{\boldsymbol{w}}(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x}$, for which I have $r_i = y_i - \boldsymbol{w}^T\boldsymbol{x}_i$. In this formulation, y_i corresponds to the expression level of a gene in the mixture, vector $\boldsymbol{x}_i$ is the expression level of the same gene in the reference cell types, and $\boldsymbol{w}$ is the fraction of each cell-type in the mixture. I can represent $\{\boldsymbol{x}_i\}_{i=1}^m$ in a compact form by matrix $\mathbf{X}$, in which row i corresponds to $\boldsymbol{x}_i$.

In cases where the number of parameters is greater than the number of samples, minimizing Equation 5.4 directly can result in *over-fitting*. Furthermore, when features (columns of $\mathbf{X}$) are highly correlated, the solution may change drastically in response to small changes in the samples, particularly among the correlated features. This condition, known as *multicollinearity*, can result in inaccurate estimates, in which coefficients of similar features are vastly different. To remedy these problems, I can add a *regularization term* that incorporates additional constraints (such as sparsity

or flatness) to enhance the stability of results. I re-write the problem with the added regularizer as:

$$\underset{\boldsymbol{w}\in\mathbb{R}^k}{\textbf{argmin}}\{\underbrace{\sum_{i=1}^{m}\boldsymbol{\mathcal{L}}(y_i-\boldsymbol{w}^T\boldsymbol{x}_i)}_{\text{Overall loss}}+\underbrace{\lambda\boldsymbol{\mathcal{R}}(\boldsymbol{w})}_{\text{Regularizer}}\} \tag{5.5}$$

where the λ parameter controls the relative importance of estimation error versus regularization. There are different choices and combinations for the loss function $\boldsymbol{\mathcal{L}}$ and regularizer function $\boldsymbol{\mathcal{R}}$, which I describe in the following sections.

Choice of Loss Functions

There are a variety of options for suitable loss functions. Some of these functions are known to be asymptotically optimal for a given noise density, whereas others may yield better performance in practice when assumptions underlying the noise model are violated. I summarize the most commonly used set of loss functions:

- If I assume that the underlying model is perturbed by Gaussian white noise, the squared or quadratic loss, denoted by $\boldsymbol{\mathcal{L}}_2$, is known to be asymptotically optimal. This loss function is used in classical least squares regression and is defined as:

$$\boldsymbol{\mathcal{L}}_2(r_i)=r_i^2=(y_i-\boldsymbol{w}^T\boldsymbol{x}_i)^2$$

- Absolute deviation loss, denoted by $\boldsymbol{\mathcal{L}}_1$, is the optimal choice if noise follows a Laplacian distribution. Formally, it is defined as:

$$\boldsymbol{\mathcal{L}}_1(r_i)=|r_i|=|y_i-\boldsymbol{w}^T\boldsymbol{x}_i|$$

  Compared to $\boldsymbol{\mathcal{L}}_2$, the choice of $\boldsymbol{\mathcal{L}}_1$ is preferred in the presence of outliers, as it is less sensitive to extreme values

- Huber's loss function, denoted by $\boldsymbol{\mathcal{L}}_{huber}^{(M)}$, is a parametrized combination of $\boldsymbol{\mathcal{L}}_1$ and $\boldsymbol{\mathcal{L}}_2$. The main idea is that $\boldsymbol{\mathcal{L}}_2$ loss is more susceptible to outliers, while it is more sensitive to small estimation errors. To combine the best of these

two functions, I can define a half-length parameter M, which I use to transition from $\boldsymbol{\mathcal{L}}_2$ to $\boldsymbol{\mathcal{L}}_1$. More formally:

$$\boldsymbol{\mathcal{L}}_{Huber}^{(M)}(r_i) = \begin{cases} r_i^2, & \text{if } |r_i| \leq M \\ M(2|r_i| - M), & \text{otherwise} \end{cases}$$

- The loss function used in *support vector regression (SVR)* is the ϵ-insensitive loss, denoted by $\boldsymbol{\mathcal{L}}_{\epsilon}^{(\epsilon)}$. Similar to Huber loss, there is a transition phase between small and large estimation errors. However, ϵ-insensitive loss does not penalize the errors that are smaller than a threshold. Formally, I define ϵ-insensitive loss as:

$$\begin{aligned} \boldsymbol{\mathcal{L}}_{\epsilon}^{(\epsilon)}(r_i) &= \mathbf{max}(0, |r_i| - \epsilon) \\ &= \begin{cases} 0, & \text{if } |r_i| \leq \epsilon \\ |r_i| - \epsilon, & \text{otherwise} \end{cases} \end{aligned}$$

Figure 5.1 provides a visual representation of these loss functions, in which I use $M = 1$ and $\epsilon = \frac{1}{2}$ for the Huber and ϵ-insensitive loss functions, respectively. Note that for small residual values, $|r_i| \leq M = 1$, Huber and square loss are equivalent. However, outside this region Huber loss becomes linear.

Choice of Regularizers

When the reference profile contains many cell-types that may not exist in mixtures, or in cases where constitutive cell-types are highly correlated, regularizing the objective function can sparsify the solution or enhance the conditioning of the problem. I describe two commonly used regularizers here:

- The norm-2 regularizer is used to shrink the regression coefficient vector $\boldsymbol{w}$ to ensure that it is as flat as possible. A common use of this regularizer is in

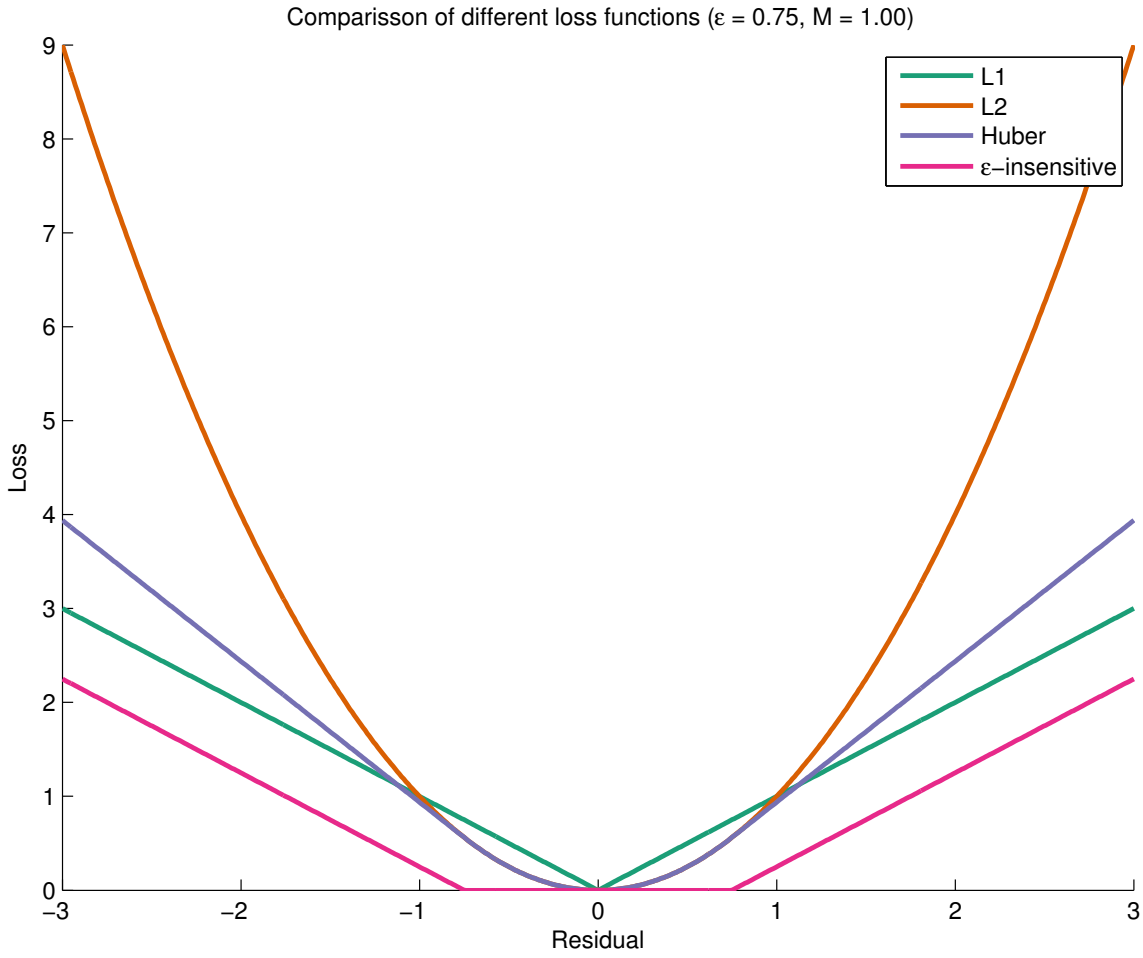


Figure 5.1.: Comparison of different loss functions

conjunction with $\boldsymbol{\mathcal{L}}_2$ loss to remedy the multicollinearity problem in classical least squares regression. This regularizer is formally defined as:

$$\boldsymbol{\mathcal{R}}_2(\boldsymbol{w}) = \| \boldsymbol{w} \|_2^2 = \sum_{i=1}^{k} w_i^2. \tag{5.6}$$

- Another common regularizer is the norm-1 regularizer, which is used to enforce sparsity over $\boldsymbol{w}$. Formally, it can be defined as:

$$\boldsymbol{\mathcal{R}}_1(\boldsymbol{w}) = \| \boldsymbol{w} \|_1 = \sum_{i=1}^{k} |w_i|. \tag{5.7}$$

In addition to these two regularizers, their combinations have also been introduced in the literature. One such example is *elastic net*, which uses a convex combination of the two, that is $\boldsymbol{\mathcal{R}}_{elastic}(\boldsymbol{w}) = \alpha\boldsymbol{\mathcal{R}}_1(\boldsymbol{w}) + (1-\alpha)\boldsymbol{\mathcal{R}}_2(\boldsymbol{w})$. Another example is *group Lasso*, which, given a grouping G among cell-types, enforces flatness among members of the group, while enhancing the sparsity pattern across groups. This regularizer function can be written as $\boldsymbol{\mathcal{R}}_{group} = \sum_{G_i} \boldsymbol{\mathcal{L}}_2(\boldsymbol{w}(G_i))$, where $\boldsymbol{w}(G_i)$ is the weight of cell-types in the ith group.

Examples of Objective Functions Used In Practice

Ordinary Least Squares (OLS) The formulation of OLS is based on squared loss, $\boldsymbol{\mathcal{L}}_2$. Formally, I have:

$$\begin{aligned} \mathbf{min}_{\boldsymbol{w}}\{\sum_{i=1}^{m} \boldsymbol{\mathcal{L}}_2(r_i)\} &= \mathbf{min}_{\boldsymbol{w}}\{\sum_{i=1}^{m} (y_i - \boldsymbol{w}^T \boldsymbol{x}_i)^2\} \\ &= \mathbf{min}_{\boldsymbol{w}} \| y - \mathbf{X}\boldsymbol{w} \|_2^2 \end{aligned}$$

where row i of the matrix $\mathbf{X}$, also known as the *design matrix*, corresponds to $\boldsymbol{x}_i$. This formulation has a closed form solution given by:

$$\hat{\boldsymbol{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}$$

In this formulation, I can observe that norm-2 regularization is especially useful in cases where the matrix $\mathbf{X}$ is ill-conditioned and near-singular, that is,

columns are dependent on each other. By shifting $\mathbf{X}^T\mathbf{X}$ towards the identity matrix, I ensure that the eigenvalues are farther from zero, which enhances the conditioning of the resulting combination.

Ridge Regression One of the main issues with the OLS formulation is that the design matrix, $\mathbf{X}$, should have full column rank k. Otherwise, if I have highly correlated variables, the solution suffers from the *multicollinearity* problem. This condition can be remedied by incorporating a norm-2 regularizer. The resulting formulation, known as *ridge regression*, is as follows:

$$\begin{aligned}&\mathbf{min}_{\boldsymbol{w}}\{\sum_{i=1}^{m}\boldsymbol{\mathcal{L}}_2(r_i)+\lambda\boldsymbol{\mathcal{R}}_2(\boldsymbol{w})\}\\&=\mathbf{min}_{\boldsymbol{w}}\parallel y-\mathbf{X}\boldsymbol{w}\parallel_2^2+\lambda\parallel\boldsymbol{w}\parallel_2^2\end{aligned}$$

Similar to OLS, I can differentiate w.r.t. $\boldsymbol{w}$ to find the close form solution for Ridge regression given by:

$$\hat{\boldsymbol{w}}=(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^T\boldsymbol{y}$$

Least Absolute Selection and Shrinkage Operator (LASSO) Combining the OLS with a norm 1 regularizer, we have the LASSO formulation:

$$\begin{aligned}&\mathbf{min}_{\boldsymbol{w}}\{\sum_{i=1}^{m}\boldsymbol{\mathcal{L}}_2(r_i)+\lambda\boldsymbol{\mathcal{R}}_1(\boldsymbol{w})\}\\&=\mathbf{min}_{\boldsymbol{w}}\parallel y-\mathbf{X}\boldsymbol{w}\parallel_2^2+\lambda\parallel\boldsymbol{w}\parallel_1\end{aligned}$$

This formulation is especially useful for producing sparse solutions by introducing zero elements in vector $\boldsymbol{w}$. However, while being convex, it does not have a closed form solution.

Robust Regression It is known that $\boldsymbol{\mathcal{L}}_2(\boldsymbol{r})$ is dominated by the largest elements of the residual vector $\boldsymbol{r}$, which makes it sensitive to outliers. To remedy this problem, different robust regression formulations have been proposed that use

alternative loss functions. Two of the best-known formulations are based on the $\boldsymbol{\mathcal{L}}_1$ and $\boldsymbol{\mathcal{L}}_{huber}$ loss functions. The $\boldsymbol{\mathcal{L}}_1$ formulation can be written as:

$$\begin{aligned}\mathbf{min}_{\boldsymbol{w}}\{\sum_{i=1}^{m}\boldsymbol{\mathcal{L}}_1(r_i)\} &= \mathbf{min}_{\boldsymbol{w}}\{\sum_{i=1}^{m}|y_i-\boldsymbol{w}^T\boldsymbol{x}_i|\}\\ &= \mathbf{min}_{\boldsymbol{w}} \parallel y-\mathbf{X}\boldsymbol{w} \parallel_1\end{aligned}$$

However, for the Huber loss function, while it can be defined similarly, it is usually formulated as an alternative convex Quadratic Program (QP):

$$\begin{gathered}\mathbf{min}_{\boldsymbol{x},\boldsymbol{z},\boldsymbol{t}}\{\frac{1}{2}\parallel \boldsymbol{z} \parallel_2^2 + M\mathbf{1}^T\boldsymbol{t}\}\\ \text{Subject to: } -\boldsymbol{t} \leq \mathbf{X}\boldsymbol{w}-\boldsymbol{y}-\boldsymbol{z} \leq \boldsymbol{t}\end{gathered} \tag{5.8}$$

which can be solved more efficiently using the following equivalent QP variant [116]:

$$\begin{gathered}\mathbf{min}_{\boldsymbol{x},\boldsymbol{z},\boldsymbol{r},\boldsymbol{s}}\{\frac{1}{2}\parallel \boldsymbol{z} \parallel_2^2 + M\mathbf{1}^T(\boldsymbol{r}+\boldsymbol{s})\}\\ \text{Subject to: } \begin{cases}\mathbf{X}\boldsymbol{w}-\boldsymbol{y}-\boldsymbol{z}=\boldsymbol{r}-\boldsymbol{s}\\ 0 \leq \boldsymbol{r},\boldsymbol{s}\end{cases}\end{gathered} \tag{5.9}$$

In both of these formulations, the scalar M corresponds to half-length parameter of the Huber's loss function.

Support Vector Regression In machine learning, Support Vector Regression (SVR) is a commonly used technique that aims to find a regression by maximizing the margins around the estimated separator hyperplane from the closest data points on each side of it. This margin provides the region in which estimation errors are ignored. SVR has been recently used to deconvolve biological mixtures, where it has been shown to outperform other methods [129]. One of the variants of

SVR is ϵ-SVR, in which parameter ϵ defines the margin, or the ϵ-tube. The primal formulation of ϵ-SVR with linear kernel can be written as [196]:

$$\mathbf{min}_{\boldsymbol{w},\boldsymbol{\xi}_i^+,\boldsymbol{\xi}_i^-}\{\frac{1}{2}\parallel \boldsymbol{w} \parallel_2^2 + C\sum_{i=1}^{m}(\boldsymbol{\xi}_i^+ + \boldsymbol{\xi}_i^-)\}$$

$$\text{Subject to:} \begin{cases} y_i - \boldsymbol{w}\cdot\boldsymbol{x}_i \leq \epsilon + \boldsymbol{\xi}_i^+ \\ -(\epsilon + \boldsymbol{\xi}_i^-) \leq y_i - \boldsymbol{w}\cdot\boldsymbol{x}_i \\ 0 \leq \boldsymbol{\xi}_i^+, \boldsymbol{\xi}_i^- \end{cases} \tag{5.10}$$

in which, given the *unit norm assumption* introduced in Section 5.2.2, I assume that $b = 0$. The dual problem for the primal in Equation 5.10 can be written in matrix form as:

$$\mathbf{max}_{\boldsymbol{\alpha}^+,\boldsymbol{\alpha}^-} \quad \Big\{\mathbf{1}^T\big((\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-)\odot \boldsymbol{y}\big) - \epsilon\mathbf{1}^T(\boldsymbol{\alpha}^+ + \boldsymbol{\alpha}^-) - (\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-)^T\mathbf{K}(\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-)\Big\}$$

$$\text{Subject to:} \begin{cases} \mathbf{1}^T(\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-) = 0 \\ 0 \leq \boldsymbol{\alpha}^+, \boldsymbol{\alpha}^- \leq C \end{cases} \tag{5.11}$$

In this formulation, $\mathbf{1}$ is a vector of all ones, $\odot$ is the element-wise product, and $\mathbf{K}$ is the kernel matrix defined as $\mathbf{K} = \mathbf{X}\mathbf{X}^T$. The dual formulation is often used to solve ϵ-SVR, because it can be easily extended to use different kernel functions to map $\boldsymbol{x}_i$ to a d-dimensional non-linear feature space. Additionally, when $m \ll k$, such as the case of high-dimensional feature spaces, it provides a better way to solve the SVR problem. However, the primal problem provides a more straightforward interpretation. In addition, in the case where $k \ll m$, it provides superior performance. To show the similarity with Equation 5.5, I can rewrite Equation 5.10 using the ϵ-insensitive loss function as follows:

$$\mathbf{min}_{\boldsymbol{w}}\{\sum_{i=1}^{m}\mathcal{L}_\epsilon(y_i - \boldsymbol{w}^T\boldsymbol{x}_i) + \lambda\boldsymbol{\mathcal{R}}_2(\boldsymbol{w})\} \tag{5.12}$$

where $\lambda = \frac{1}{2C}$ [170].

5.2.3 Overview of Prior In Silico Deconvolution Methods

A majority of existing deconvolution methods fall into two groups – they either use a regression-based framework to compute $\mathbf{G}$, $\mathbf{C}$, or both; or perform statistical inference over a probabilistic model. Abbas *et al.* [2] present one of the early regression-based methods for estimating $\mathbf{C}$. This method is designed to identify cell-type concentrations from a known reference profile of immune cells. Their method is based on Ordinary Least Squares (OLS) regression and does not consider either non-negativity or sum-to-one constraints explicitly, but rather it enforces these constraints implicitly after the optimization procedure. An extension of this approach is proposed by Qiao *et al.* [148], which uses non-negative least squares (NNLS) to explicitly enforce non-negativity as part of the optimization. Gong *et al.* [61] present a quadratic programming (QP) framework to explicitly encode both constraints in the optimization problem formulation. They also propose an extension to this method, called DeconRNASeq, which applies the same QP framework to RNASeq datasets. More recently Newman *et al.* [129] propose robust linear regression (RLR) and ν-SVR regression instead of $\boldsymbol{\mathcal{L}}_2$ based regression, which is highly susceptible to noise. Digital cell quantification (DCQ) [5] is another approach designed for monitoring the immune system during infection. Compared to prior methods, DCQ forces sparsity by combining $\boldsymbol{\mathcal{R}}_2$ and $\boldsymbol{\mathcal{R}}_1$ regularization into an *elastic net*. This regularization is essential for successfully identifying the subset of active cells at each stage, given the larger number of cell-types included in their panel (213 immune cell sub-populations). In contrast to these techniques, Shen-Orr *et al.* [165] propose a method, call *csSAM*, which is specifically designed to identify genes that are differentially expressed among purified cell-types. The core of this method is regression over matrix $\mathbf{C}$ to estimate matrix $\mathbf{G}$.

Full regression-based methods correspond to unsupervised approaches in the sense that they do not rely on either $\mathbf{G}$ or $\mathbf{C}$. They are either fully *ab initio*, or they use variations of block-coordinate descent to successively identify better estimates for

both **C** and **G** [91]. Venet *et al.* [200] present one of the early methods in this class, which uses an NMF-like method coupled with a heuristic to decorrelate columns of **G** in each iteration. Repsilber *et al.* [154] propose an algorithm called *deconf*, which uses alternating non-negative least squares (ANLS) for solving NMF, without the decorrelation step of Vennet *et al.*, while implicitly applying constraints on **C** and **G** at each iteration. Inspired by the work of Pauca *et al.* on hyperspectral image deconvolution [134], Zuckerman *et al.* [225] propose an NMF method based on the Frobenius norm for gene expression deconvolution. They use gradient descent to solve for **C** and **G** at each step, which converges to a local optimum of the objective function. Given that the expression domain of cell-type specific markers is restricted to unique cells in the reference profile, Gaujoux *et al.* [53] present a semi-supervised NMF (ss-NMF) method that explicitly enforces an orthogonality constraint at each iteration over the subset of markers in the reference profile. This constraint both enhances the convergence of the NMF algorithm, and simplifies the matching of columns in the estimated cell-type expression to the columns of the reference panel, **G**. The Digital Sorting Algorithm (DSA) [222] works as follows: if concentration matrix **C** is known a priori, it directly uses quadratic programming (QP) with added constraints on the lower/upper bound of gene expressions to estimate matrix **G**. Otherwise, if fractions are also unknown, it uses the average expression of given marker genes that are only expressed in one cell-type, combined with the *STO* constraint, to estimate concentrations matrix **C** first. Population-specific expression analysis (PSEA) [101] performs a linear least squares regression to estimate quantitative measures of cell-type-specific expression levels, in a similar fashion as the update equation for estimating $\hat{\mathbf{G}}$ in Equation 5.3. In cases where the matrix **C** is not known a priori, *PSEA* exploits the average expression of marker genes that are exclusively expressed in one of the reference profiles as *reference signals* to track the variation of cell-type fractions across multiple mixture samples.

More recently, a new class of methods, collectively referred to as *convex analysis of mixtures (CAM)*, have been proposed to directly infer marker genes from mixture

profiles [25, 187, 207]. The *CAM* family of methods aim to use a geometric approach to identify corners of the scatter simplex for mixed expression profiles. The key to the success of these methods is a recently proven bijection between the scatter simplex of mixed profiles and a transformed (rotated and compressed) version of the scatter simplex for constituent cell types [207]. Following this, "marker genes" are concentrated around the corners of this simplex. After identifying these semi-orthogonal markers, one can recover cell type percentages using any of the marker-based methods mentioned above, such as PSEA [101] or DSA [222]. Similar techniques have been also proposed earlier to infer tumor phylogeny using microarray measurements of tumor populations [161].

In addition to regression-based methods, a large class of methods is based on probabilistic modeling of gene expression. Erikkila *et al.* [47] introduce a method, called *DSection*, which formulates the deconvolution problem using a Bayesian model. It incorporates a Bayesian prior over the noisy observation of given concentration parameters to account for their uncertainty, and employs a MCMC sampling scheme to estimate the posterior distribution of the parameters/latent variables, including $\mathbf{G}$ and a refined version of $\mathbf{C}$. The in-silico NanoDissection method [85] uses a classification algorithm based on linear SVM coupled with an iterative adjustment process to refine a set of provided, positive and negative, marker genes and infer a ranked list of genome-scale predictions for cell-type-specific markers. Quon *et al.* [149] propose a probabilistic deconvolution method, called *PERT*, which estimates a global, multiplicative perturbation vector to correct for the differences between provided reference profiles and the true cell-types in the mixture. *PERT* formulates the deconvolution problem in a similar framework as Latent Dirichlet Allocation (LDA), and uses the conjugate gradient descent method to cyclically optimize the joint likelihood function with respect to each latent variable/parameter. Finally, microarray microdissection with analysis of differences (MMAD) [107] incorporates the concept of the *effective RNA fraction* to account for source and sample-specific bias in the cell-type fractions for each gene. They propose different strategies depending on the availability of

Table 5.2: Best combination of choices for feature selection/regularization for different datasets

Reference	Name	Method	Loss	Non-negativity	Sum-to-one	Regularizer
Abbas *et al.* (2009)	LS	Ordinary Least Squares (OLS)	$\mathcal{L}_2$	Imp	Imp	-
Gong *et al.* (2011)	QP	Quadratic Programming	$\mathcal{L}_2$	Exp	Exp	-
Qiao *et al.* (2012)	NNLS	Non-negative Least Squares (NNLS)	$\mathcal{L}_2$	Exp	Imp	-
Altboum *et al.* (2014)	DCQ	Elastic Net	$\mathcal{L}_2$	Imp	Imp	$\mathcal{L}_1/\mathcal{L}_2$
Newman *et al.* (2015)	RLR	Robust Linear Regression (RLR)	Huber	Imp	Imp	-
Newman *et al.* (2015)	CIBERSORT	ν-SVR	ϵ-insensitive	Imp	Imp	$\mathcal{L}_2$

x

additional data sources. In cases where no additional information is available, they identify genes with the highest variation in mixtures as markers and assign them to different reference cell-types using k-means clustering, and finally use these *de novo* markers to compute cell-type fractions. *MMAD* uses a MLE approach over the residual sum of squares to estimate unknown parameters in their formulation.

In this chapter I focus on partial deconvolution methods for recovering matrix $\mathbf{C}$ using given reference profiles for constituent tissues/cell types. Table 5.2 summarizes different combinations proposed in literature thus far. I cover all these configurations here, as well as missing combinations that have not been studied in current literature.

5.2.4 Evaluation Measures

Let us denote the actual and estimated coefficient matrices by $\mathbf{C}$ and $\hat{\mathbf{C}}$, respectively. I first normalize these measures to ensure each column sums to one. Then, I define the corresponding percentages as $\mathbf{P} = 100 \times \mathbf{C}_{norm}$ and $\hat{\mathbf{P}} = 100 \times \hat{\mathbf{C}}_{norm}$. Finally, let $r_{jk} = p_{jk} - \hat{p}_{jk}$ be the residual estimation error of cell-type k in sample j. Using this notation, I can define three commonly used measures of estimation error as follows:

1. **Mean absolute difference (mAD)**: This is among the easiest measures to interpret. It is defined as the average of all differences for different cell-type percentages in different mixture samples. More specifically:

$$mAD = \frac{1}{p \times q} \sum_{j=1}^{p} \sum_{k=1}^{q} |r_{jk}|$$

2. **Root mean squared distance (RMSD)**: This measure is one of the most commonly used distance functions in the literature. It is formally defined as:

$$mAD = \sqrt{\frac{1}{p \times q} \sum_{j=1}^{p} \sum_{k=1}^{q} r_{jk}^2}$$

3. **Pearson's correlation distance**: Pearson's correlation measures the linear dependence between estimated and actual percentages. Let us vectorize percentage matrices as $\boldsymbol{p} = \mathbf{vec}(\mathbf{P})$ and $\hat{\boldsymbol{p}} = \mathbf{vec}(\hat{\mathbf{P}})$. Using this notation, the correlation between these two vectors is defined as:

$$\rho_{\boldsymbol{p},\hat{\boldsymbol{p}}} = \frac{\mathbf{cov}(\boldsymbol{p}, \hat{\boldsymbol{p}})}{\sigma(\boldsymbol{p})\sigma(\hat{\boldsymbol{p}})} \tag{5.13}$$

where **cov** and σ correspond to covariance and standard variation of vectors, respectively. Finally, I define the correlation distance measure as $R^2D = 1 - \rho_{\boldsymbol{p},\hat{\boldsymbol{p}}}$.

5.2.5 Implementation

All codes and experiments have been implemented in Matlab. To implement different formulations of the deconvolution problem, I used CVX, a package for specifying and solving convex programs [1,63]. I used Mosek together with CVX, which is a high-performance solver for large-scale linear and quadratic programs [126]. All codes and datasets are freely available at `github.com/shmohammadi86/DeconvolutionReview`.

5.3 Results and Discussion

I now present a comprehensive evaluation of various formulations for solving deconvolution problems. Some of these algorithimic combinations have been proposed in literature, while others represent new algorithmic choices. I systematically assess the impact of these algorithmic choices on the performance of in-silico deconvolution.

5.3.1 Effect of Loss Function and Constraint Enforcement on Deconvolution Performance

I perform a systematic evaluation of the four different loss functions introduced in Section 5.2.2, as well as implicit and explicit enforcement of *non-negativity (NN)* and *sum-to-one (STO)* constraints over the concentration matrix ($\hat{\mathbf{C}}$), on the overall performance of deconvolution methods for each dataset. There are 16 configurations of loss functions/constraints for each test case. Additionally, for Huber and Hinge loss functions, where M and ϵ are unknown, I perform a grid search with 15 values in multiples of 10 spanning the range $\{10^{-7}, \cdots, 10^{7}\}$ to find the best values for these parameters. In order to evaluate an upper bound on the "potential" performance of these two loss functions, I use the true concentrations in each sample, $\mathbf{c}$, to evaluate each parameter choice. In practical applications, the RMSD of residual error between $\boldsymbol{m}$ and $\mathbf{G}\hat{\boldsymbol{c}}$ is often used to select the optimal parameter. This is not always in agreement with the choice made based on known $\boldsymbol{c}$.

For each test dataset, I compute the three evaluation measures defined in Section 5.2.4. Additionally, for each of these measures, I compute an empirical p-value by sampling random concentrations from a Uniform distribution and enforcing *NN* and *STO* constraints on the resulting random sample. In my study, I sampled $10,000$ concentrations for each dataset/measure, which results in a lower bound of 10^{-4} on the estimated p-values. Figure 5.2 presents the time each loss function takes to compute per sample, averaged over all constraint combinations. The actual times taken for Huber and Hinge losses are roughly 15 times those reported here, which is the

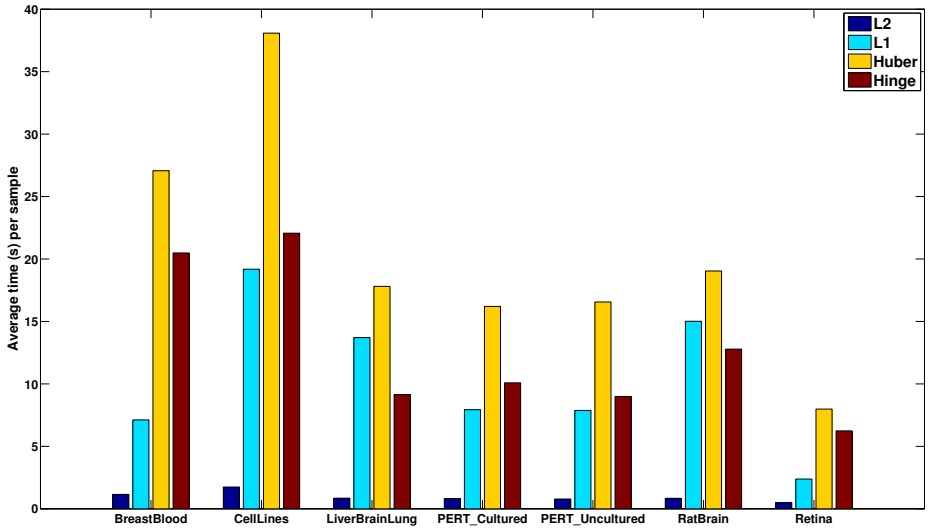


Figure 5.2.: Average computational time for each loss function in different datasets

number of experiments performed to find the optimal parameters for these loss functions. From these results, $\boldsymbol{\mathcal{L}}_2$ can be observed to have the fastest computation time, whereas $\boldsymbol{\mathcal{L}}_{Huber}$ is the slowest. Measures $\boldsymbol{\mathcal{L}}_1$ and $\boldsymbol{\mathcal{L}}_{Hinge}$ fit in between these two extremes, with $\boldsymbol{\mathcal{L}}_1$ being faster the majority of times. I can directly compare these computation times, because I formulate all methods within the same framework; thus, differences in implementations do not impact direct comparisons.

Computation time, while important, is not the critical measure in my evaluation. The true performance of a configuration (selection of loss function and constraints) is measured by its estimation error. In order to rank different configurations, I first assess the agreement among different measures. To this end, I evaluate each dataset as follows: for each experiment, I compute $\mathbf{mAD}, \mathbf{RMSD}$, and $\mathbf{R^2D}$ independently. Then, I use *Kendall* rank correlation, a non-parametric hypothesis test for statistical dependence between two random variables, between each pair of measures and compute a log-transformed p-value for each correlation. Figure 5.3 shows the agreement among these measures across different datasets. Overall, $RMSD$ and mAD measures show higher consistency, compared to R^2D measure. However, the mAD

Figure 5.3.: Agreement among different evaluation measures across different datasets

measure is easier to interpret as a measure of percentage loss for each configuration. Consequently, I choose this measure for my evaluation in this study.

Using mAD as the measure of performance, I evaluate each configuration over each dataset and sort the results. Figure 5.4 shows various combinations for each dataset. The **RatBrain**, **LiverBrainLung**, **BreastBlood**, and **CellLines** datasets achieve high performance. Among these datasets, **RatBrain**, **LiverBrainLung**, and **BreastBlood** had the $\mathcal{L}_2$ loss function as the best configuration, with the **CellLines** dataset being less sensitive to the choice of the loss function. Another surprising observation is that for the majority of configurations, enforcing the *sum-to-one* constraint worsens the results. I investigate this issue in greater depth in Section 5.3.2.

For **Retina**, as well as both **PERT** datasets, the overall performance is worse than the other datasets. In the case of **PERT**, this is expected, since the flow-sorted proportions are used as an estimate of cell-type proportions. Furthermore, the reference profiles come from a different study and therefore have greater difference with the true cell-types in the mixture. However, the **Retina** dataset exhibits unusually low performance, which may be attributed to multiple factors. As an initial investigation, I performed a quality control (QC) over different samples to see if errors are similarly distributed across samples. Figure 5.6 presents per-sample error, measured

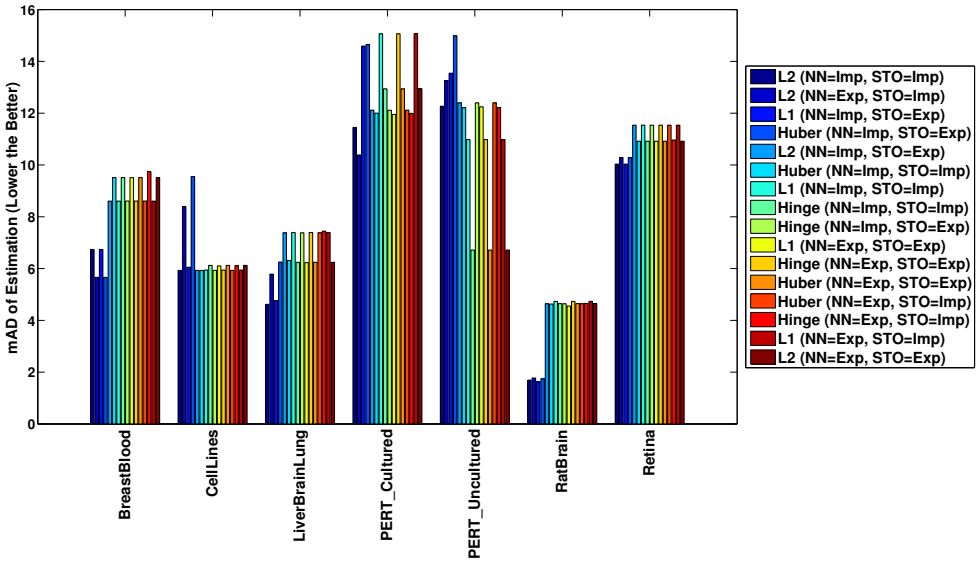
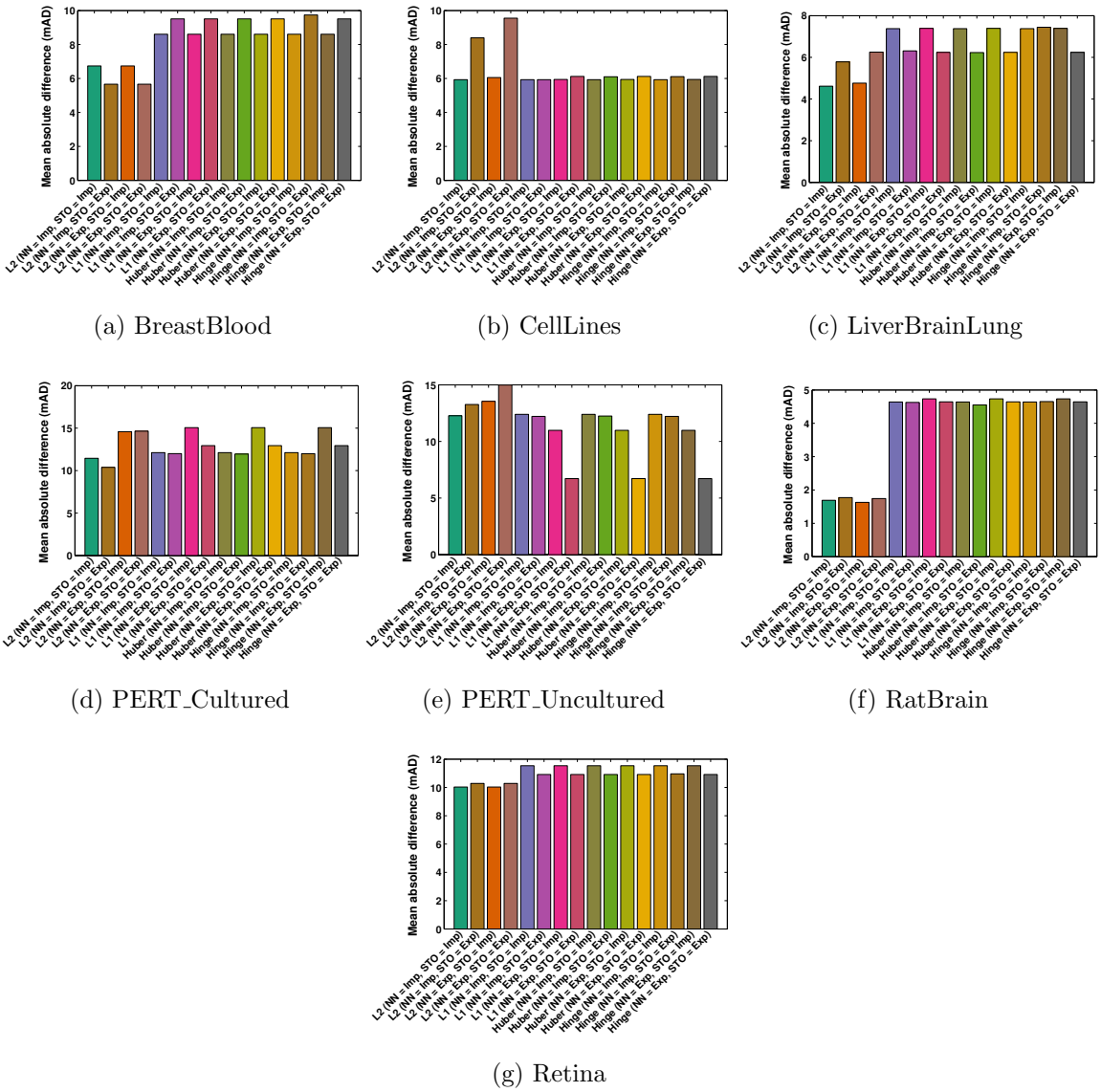


Figure 5.4.: Overall performance of different loss/constraints combinations over all datasets



(a) BreastBlood (b) CellLines (c) LiverBrainLung

(d) PERT_Cultured (e) PERT_Uncultured (f) RatBrain

(g) Retina

Figure 5.5.: Overall performance of different loss function/constraints combinations over all datasets (lower the better)

by mAD, with median and median absolute deviation (MAD) marked accordingly. Interestingly, for the $4^{th}, 6^{th}$, and 8th mixtures, the third replicate has much higher error than the rest. In the expression matrix, I observed a lower correlation between these replicates and the other two replicates in the batch. Additionally, for the 7th mixture, all three replicates show high error rates. I expand on these results in

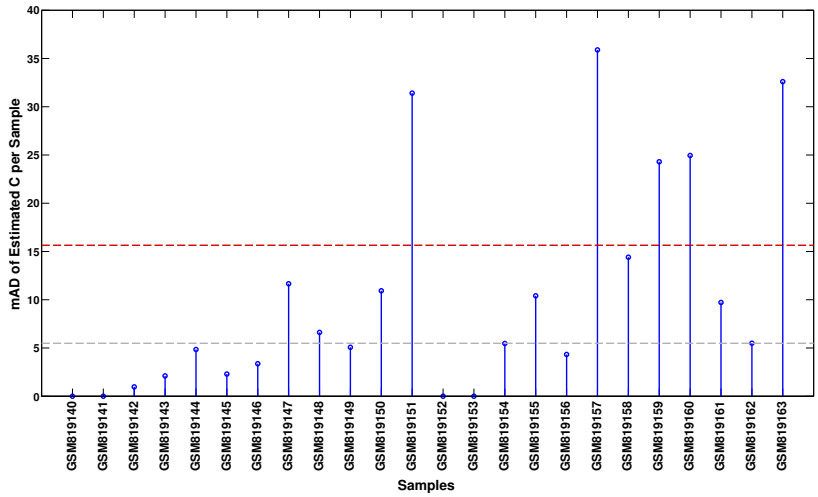


Figure 5.6.: Sample-based error of the Retina dataset, based on $\boldsymbol{\mathcal{L}}_2$ with explicit *NN* and *STO*

later sections to identify additional reasons that contribute to the low deconvolution performance of the **Retina** dataset.

Finally, I note that in all test cases the performance of $\boldsymbol{\mathcal{L}}_1, \boldsymbol{\mathcal{L}}_{Huber}$, and $\boldsymbol{\mathcal{L}}_{Hinge}$ are comparable, while $\boldsymbol{\mathcal{L}}_{Huber}$ and $\boldsymbol{\mathcal{L}}_{Hinge}$ needed an additional step of parameter tuning. Consequently, I only consider $\boldsymbol{\mathcal{L}}_1$ as a representative of this "robust" group of loss functions in the rest of my study.

5.3.2 Agreement of Gene Expressions With Sum-to-One (STO) Constraint

Considering the lower performance of configurations that explicitly enforce STO constraints, I aim to investigate whether features (genes) in each dataset respect this constraint. Under the *STO* and *NN* constraints, I use simple bounds for identifying violating features, for which there is no combination of concentration values that can satisfy both *STO* and *NN*. Let $\boldsymbol{m}(i)$ be the expression value of the ith gene in the given mixture, and $\mathbf{G}(i,1), \cdots, \mathbf{G}(i,q)$ be the corresponding expressions in different reference cell-types. Let $\mathbf{G}_{min}(i) = \mathbf{min}\{\mathbf{G}(i,1), \cdots, \mathbf{G}(i,q)\}$ and $\mathbf{G}_{max}(i) = \mathbf{max}\{\mathbf{G}(i,1), \cdots, \mathbf{G}(i,q)\}$. Given that all concentrations are bound be-

tween $0 \leq \boldsymbol{c}(k) \leq 1; \forall 1 \leq k \leq k$, the minimum and maximum values that an estimated mixture value for the ith gene can attain are $\mathbf{G}_{min}(i)$ and $\mathbf{G}_{max}(i)$, respectively (by setting $\boldsymbol{c}(k) = 1$ for min/max value, and 0 everywhere else). Using this notation, I can identify features that violate *STO* as follows:

$$\boldsymbol{m}(i) \leq \mathbf{G}_{min}(i) \quad \forall 1 \leq i \leq n \quad \{\text{Violating reference}\}$$
$$\mathbf{G}_{max}(i) \leq \boldsymbol{m}(i) \quad \forall 1 \leq i \leq n \quad \{\text{Violating mixture}\}$$

The first condition holds because expression values in reference profiles are so large that I need the sum of concentrations to be lower than one to be able to match the corresponding gene expression in the mixture. The second condition holds in cases where the expression of a gene in the mixture is so high that I need the sum of concentrations to be greater than one to be able to match it. In other words, for feature i, these constraints identify extreme expression values in reference profiles and mixture samples, respectively. Using these conditions, I compute the total number of features violating *STO* condition in each dataset.

Figure 5.7 presents violating features in mixtures and reference profiles, averaged over all mixture samples in each dataset. I normalize and report the percent of features to account for differences in the total number of features in each dataset. We first observe that for the majority of datasets, except **Retina** and **BreastBlood**, the percent of violating features is much smaller than violating features in reference profiles. These two datasets also have the highest number of violating features in their reference profiles, summing to a total of approximately 60% of all features. This observation is likely due to the normalization used in preprocessing microarray profiles. Specifically, one must not only normalize **M** and **G** independently, but also with respect to each other. I suggest using control genes that are expressed in all cell-types with low variation to normalize expression profiles. A recent study aimed to identify subsets of housekeeping genes in human tissues that respect these conditions [45]. Another choice is using ribosomal proteins, the basic building blocks

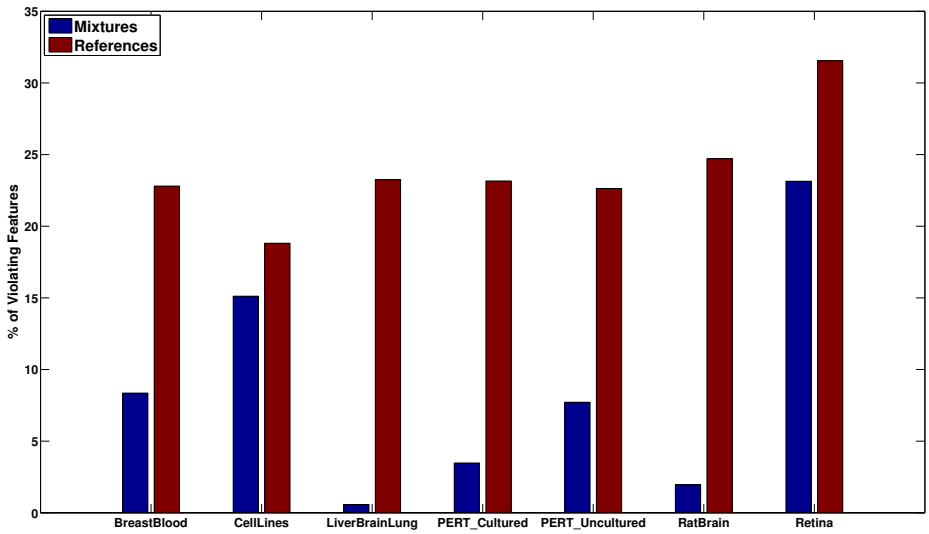


Figure 5.7.: Percent of features in each dataset that violate the STO constraint

of the cellular translation machinery, which are expressed in a wide range of species. The Remove Unwanted Variation (RUV) [51] method is developed to remove batch effects from microarray and RNASeq expression profiles, but also to normalize them using control genes. A simple extension of this method can be adopted to solve the normalization difference between mixtures and references.

Next, I evaluate how filtering these features affects deconvolution performance of each dataset. For each case, I run deconvolution using all configurations and report the change (delta mAD) independently. Figure 5.8 presents changes in the mAD estimation error after removing violating features in both $\boldsymbol{m}$ and $\mathbf{G}$ before performing deconvolution. Similar to previous experiments, the **Retina** dataset exhibits widely different behavior than the rest of the datasets. Removing this dataset from further consideration, I find that the overall performance over all datasets improves, with the exception of the **RatBrain** dataset. In the case of the **RatBrain** dataset, I hypothesize that the initially superior performance can be attributed to highly expressed features. These outliers, that happens to agree with the true solution, result in *overfitting*. Finally, I note a correlation between observed enhancements and the level of

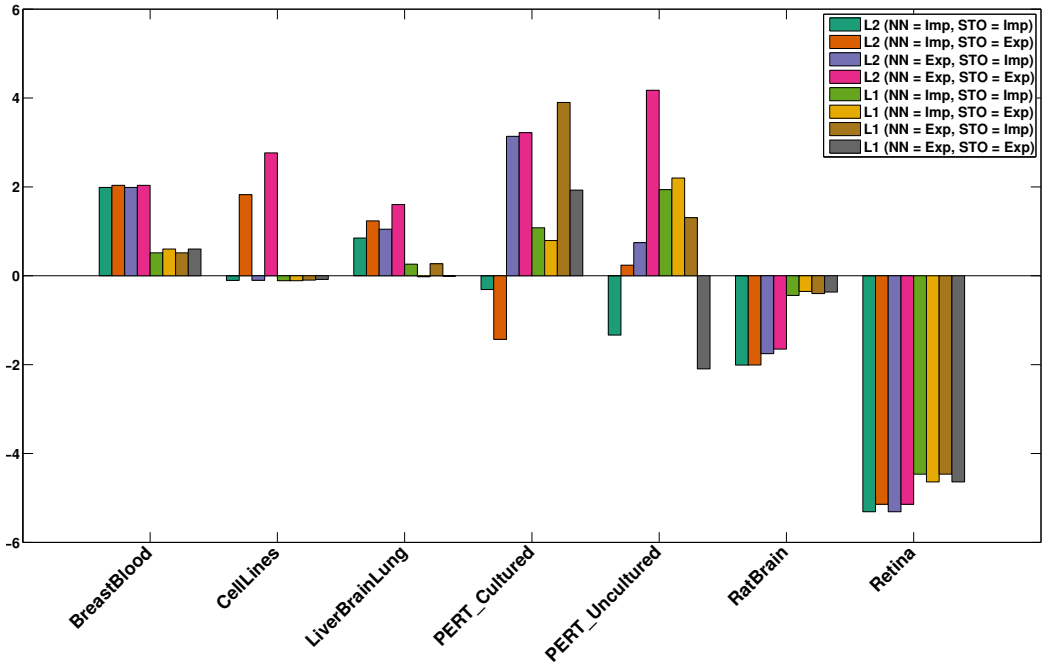


Figure 5.8.: Performance of deconvolution methods after removing violating features

violation of features in $\boldsymbol{m}$. Consistent with this observation, I obtain similar results when I only filter violating features from mixtures, but not reference profiles.

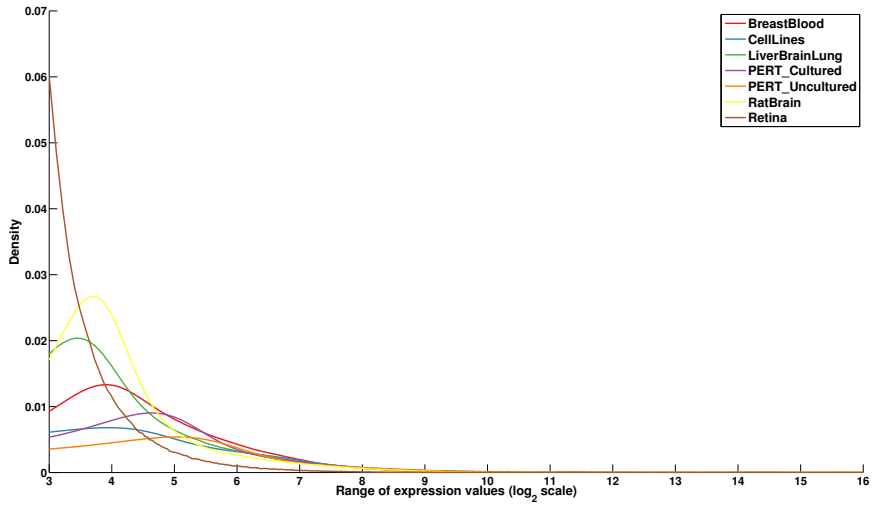
5.3.3 Range Filtering – Finding an Optimal Threshold

Different upper/lower bounds have been proposed in the literature to prefilter expression values prior to deconvolution. For example, Gong *et al.* [61] suggest an effective range of $[0.5, 5000]$, whereas Ahn *et al.* [4] observe an optimal range of $[2^4 - 2^{14}]$. To facilitate the choice of expression bounds, I seek a systematic way to identify an optimal range for different datasets. Kawaji *et al.* [88] recently report on an experiment to assess whether gene expression is quantified linearly in mixtures. To this end, they mix two cell-types (THP-1 and HeLa cell-lines) and see if experimentally measured expressions match with the computationally simulated datasets. They observe that expression values for microarray measurements are skewed for the lowly expressed genes (approximately < 10). This allows us to choose the lower bound

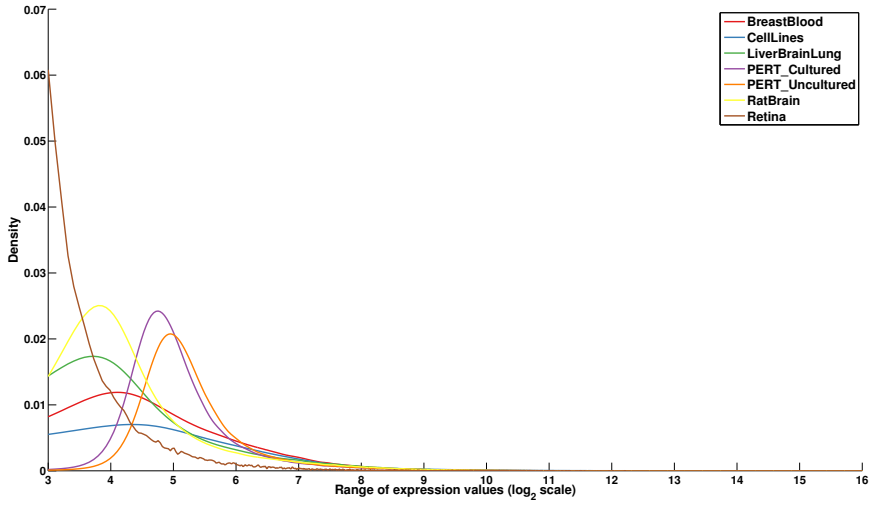
based on experimental evidence. In my study, I search for the optimal bounds over a log_2-linear space; thus, I set a threshold of 2^3 on the minimum expression values, which is closest to the bound proposed by Kawaji *et al.* [88].

Choosing an upper bound on the expression values is a harder problem, since it relates to enhancing the performance of deconvolution methods by removing outliers. Additionally, there is a known relationship between the mean expression value and its variance [188], which makes these outliers noisier than the rest of the features. This becomes even more important when dealing with purified cell-types that come from different labs, since highly expressed time/micro-environment dependent genes would be significantly different than the ones in the mixture [148]. A simple argument is to filter genes that the range of expression values in *Affymetrix* microarray technology is bounded by 2^{16} (due to initial normalization and image processing steps). Measurements close to this bound are not reliable as they might be saturated and inaccurate. However, practical bounds used in previous studies are far from these extreme values. In order to examine the overall distribution of expression values, I analyze different datasets independently. For each dataset, I separately analyze mixture samples and reference profiles, encoded by matrices $\mathbf{M}$ and $\mathbf{G}$, respectively. For each of these matrices, I vectorize the expression values and perform kernel smoothing using the Gaussian kernel to estimate the probability density function.

Figure 5.9a and Figure 5.9b show the distribution of log-transformed expression values for mixtures and reference profiles, respectively. These expression values are greater than my lower bound of 2^3. In agreement with my previous results, I observe an unusually skewed distribution for the **Retina** dataset, which in turn contributes to its lower performance compared to other ideal mixtures. Additionally, I observe that approximately 80% of the features in this dataset are smaller than 2^3, which are filtered and not shown in the distribution plot. For the rest of the datasets, in both mixtures and references, I observe a bell-shaped distribution with most of the features captured up to an upper bound of $2^8 - 2^{10}$. Another exception to this pattern



(a) Mixtures



(b) Reference Profiles

Figure 5.9.: Distribution of gene expression values for mixtures and references

is the **CellLines** dataset, which has a heavier tail than other datasets, especially in its reference profile.

Next, I systematically evaluate the effect of range filtering by analyzing upper bounds increasing in factors of 10 in the range $\{2^5, \cdots, 2^{16}\}$. In each case, I remove all features that at least one of the reference profiles or mixture samples has a value exceeding this upper bound. Figure 5.10 illustrates the percent of features that are

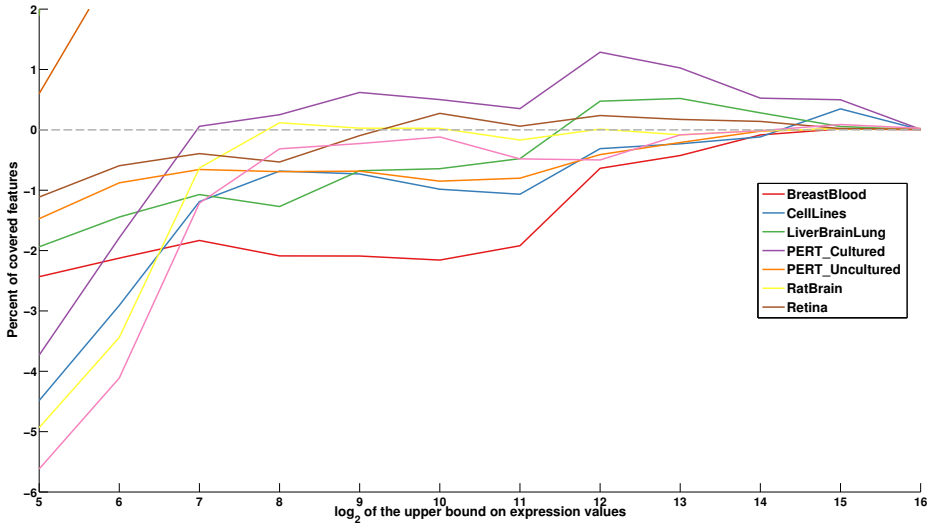
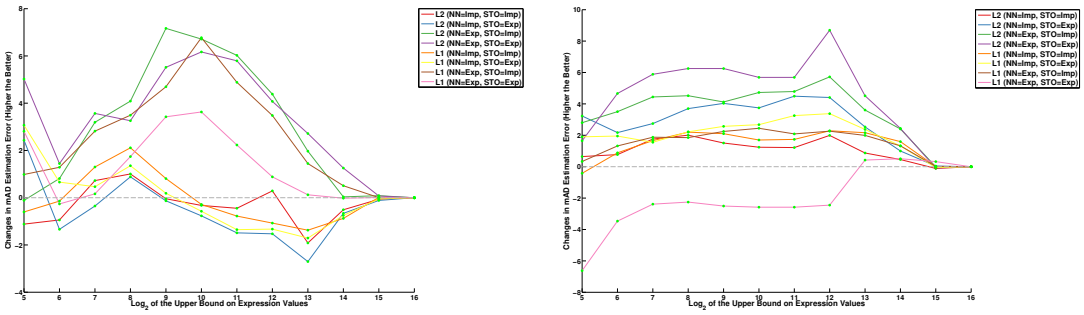


Figure 5.10.: Percent of covered features during range filtering

retained, as I increase the upper bound. As mentioned earlier, approximately 80% of the features in the **Retina** dataset are lower than 2^3, which is evident from the maximum percent of features left to be bounded by 20% in this figure. Additionally, consistent with my previous observation over expression densities, more that 80% of the features are covered between $2^8 - 2^{10}$, except for the **CellLine** dataset.

Finally, I perform deconvolution using the remaining features given each upper bound. The results are mixed, but a common trend is that removing highly expressed genes decreases performance of ideal mixtures with known concentrations, while enhancing the performance of **PERT** datasets. Figure 5.11a and Figure 5.11b show the changes in mAD error, compared to unfiltered deconvolution, for the **PERT** dataset. In each case, I observe improvements up to 7 and 8 percent, respectively. The red and green points on the diagram show the significance of deconvolution. Interestingly, while both methods show similar improvements, all data points for cultured PERT seem to be insignificant, whereas uncultured PERT shows significance for the majority of data-points. This is due to the weakness of my random model, which is dependent on the number of samples and is not comparable across datasets. Uncultured PERT has twice as many samples as cultured PERT, which makes it less likely to have any



(a) Cultured (b) Uncultured

Figure 5.11.: Performance of PERT datasets during range filtering

random samples achieving as good an mAD as the observed estimation error. This dependency on the number of samples can be addressed by defining sample-based p-values. Another observation is that for the uncultured dataset, all measures have been improved, except $\boldsymbol{\mathcal{L}}_1$ with explicit NN and STO constraints. On the other hand, for the cultured dataset, both $\boldsymbol{\mathcal{L}}_1$ and $\boldsymbol{\mathcal{L}}_2$ with the explicit NN constraint perform well, whereas implicitly enforcing NN deteriorates their performance. Cultured and uncultured datasets have their peak at 2^{10} and 2^{12}, respectively.

For the rest of the datasets, range filtering decreased performance in a majority of cases, except the **Retina** dataset, which had an improved performance at 2^6 with the best result achieved with $\boldsymbol{\mathcal{L}}_1$ with both explicit NN and STO enforcement. This changed the best observed performance of this datasest, measured as mAD, to be close to 7. These mixed results make it harder to choose a threshold for the upper bound, so I average results over all datasets to find a balance between improvements in PERT and overall deterioration in other datasets. Figure 5.12 presents the averaged mAD difference across all datasets. This suggests a "general" upper bound filter of 2^{12} to be optimal across all datasets.

I use this threshold to filter all datasets and perform deconvolution on them. Figure 5.13 presents the dataset-specific performance of range filtering with fixed bounds, measured by changes in the mAD value compared to the original deconvolution. As

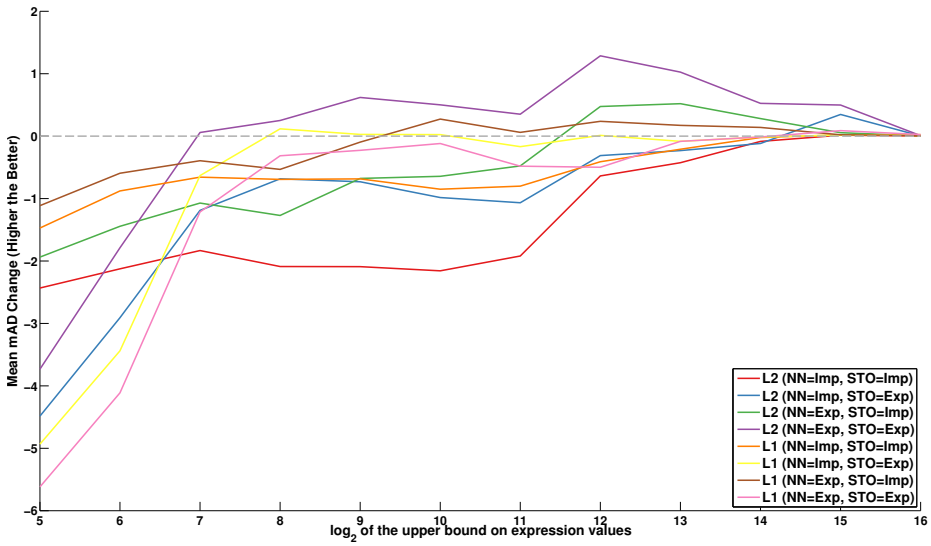


Figure 5.12.: Average performance of range filtering over all datasets

observed from individual performance plots, range filtering is most effective in cases where the reference profiles differ significantly from the true cell-types in the mixture, such as the case with the **PERT** datasets. In ideal mixtures, since cell-types are measured and mixed at the same time/laboratory, this distinction is negligible. In these cases, highly expressed genes in mixtures and references coincide with each other and provide additional clues for the regression. Consequently, removing these highly expressed genes often degrades the performance of deconvolution methods. This generalization of the upper bound threshold, however, should be adopted with care, since each dataset exhibits different behavior in response to range filtering. Ideally, one must filter each dataset individually based on the distribution of expression values. Furthermore, in practical applications, gold standards are not available to aid in the choice of cutoff threshold.

I now introduce a new method that adaptively identifies an effective range for each dataset. Figure 5.14 illustrates the log_2 normalized value of maximal expression for each gene in matrices **M** and **G**, sorted in ascending order. In all cases, intermediate values exhibit a gradual increase, whereas the top and bottom elements in the sorted list show a steep change in their expression. I aim to identify the critical points

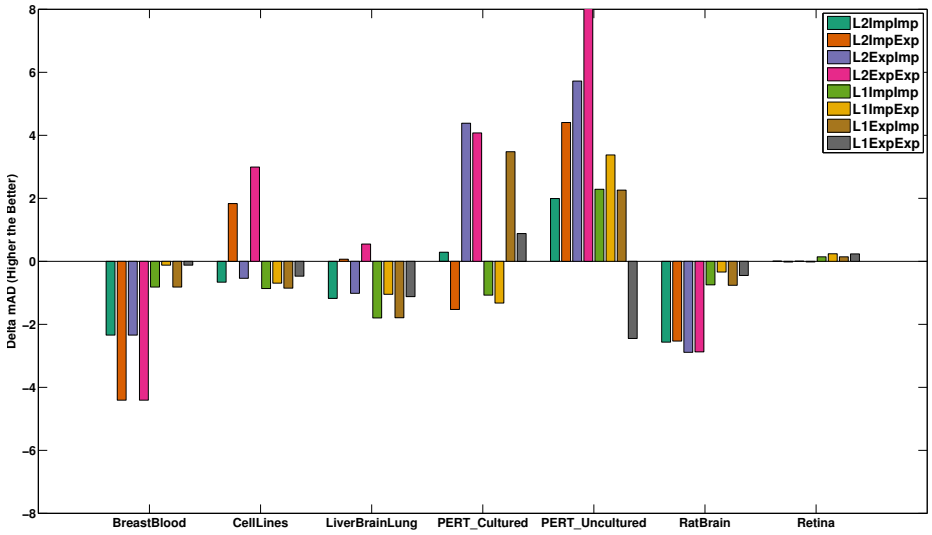


Figure 5.13.: Dataset-specific changes in the performance of deconvolution methods after filtering expression ranges to fit within $[2^3 - 2^{12}]$

corresponding to these sudden changes in the expression values for each dataset. To this end, I select the middle point as a point of reference and analyze the upper and lower half, independently. For each half, I find the point on the curve that has the longest distance from the line connecting the first (last) element to the middle element. Application of this process over the **CellTypes** dataset is visualized in Figure 5.15. Green points in this figure correspond to the critical points, which are used to define the lower and upper bound for the expression values of this dataset.

I use this technique to identify adaptive ranges for each dataset prior to deconvolution. Table 5.3 summarizes the identified critical points for each dataset. Figure 5.16 presents the dataset-specific performance of each method after adaptive range filtering. While in most cases the results for fixed and adaptive range filtering are compatible, and in some cases adaptive filtering gives better results, the most notable difference is the degraded performance of **LiverBrainLung**, and, to some extent, **RatBrain** datasets. To further investigate this observation, I examine individual experiments for these datasets for fixed thresholds. Figure 5.17 illustrates individual plots for each dataset. The common trend here is that in both cases range filtering,

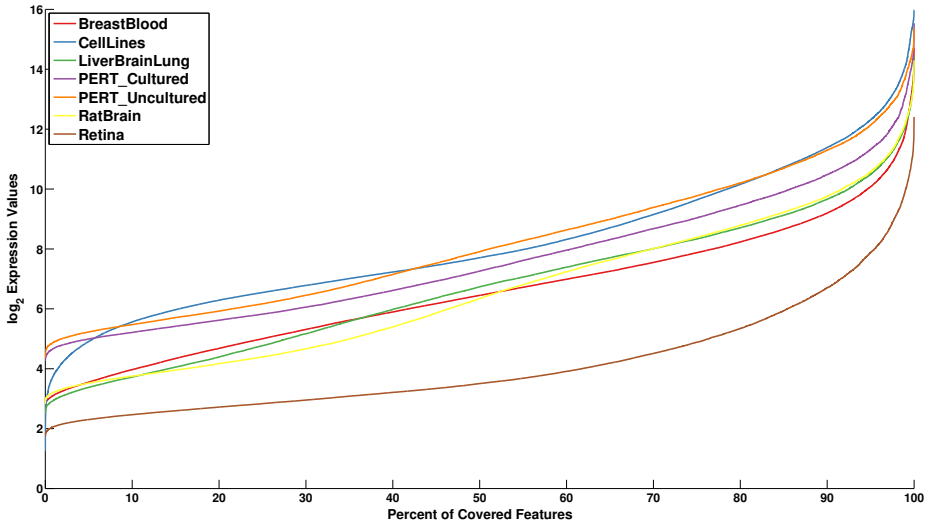


Figure 5.14.: Sorted log_2-transformed gene expressions in different datasets

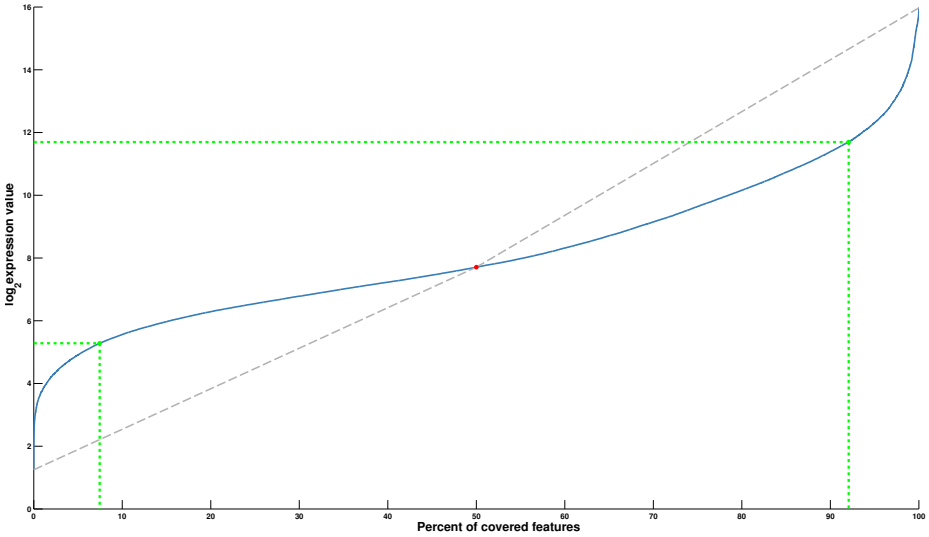


Figure 5.15.: Example of adaptive filtering over the CellLines dataset

in general, degrades the performance of deconvolution methods for all configurations. In other words, extreme values in these datasets are actually helpful in guiding the regression, and any filtering negatively impacts performance. This suggests that range filtering, in general, is not always helpful in enhancing the deconvolution performance,

Table 5.3: Summary of adaptive ranges for each dataset for deconvolution

	LowerBound	UpperBound
BreastBlood	4.2842	9.4314
CellLines	5.2814	11.6942
LiverBrainLung	3.3245	9.9324
PERT_Cultured	4.9416	10.9224
PERT_Uncultured	5.1674	11.5042
RatBrain	3.3726	9.9698
Retina	2.4063	6.7499

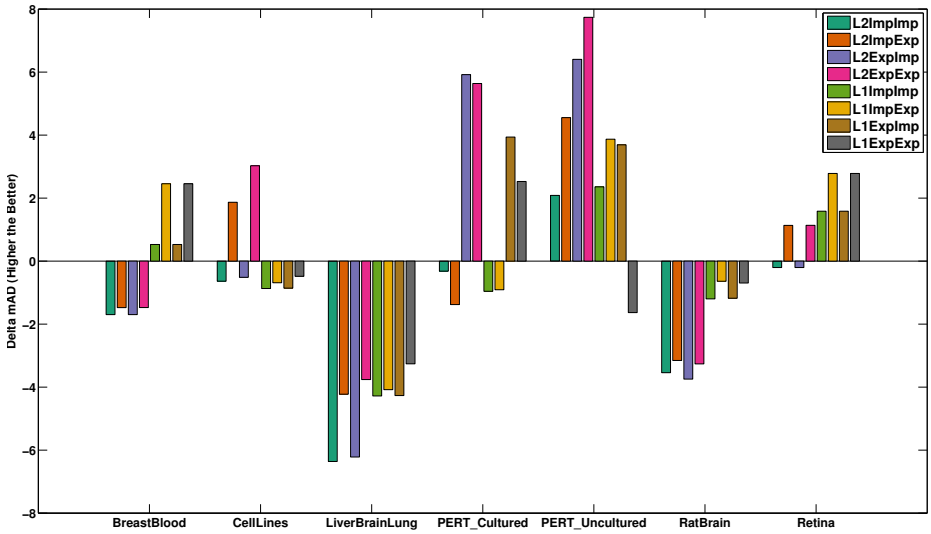
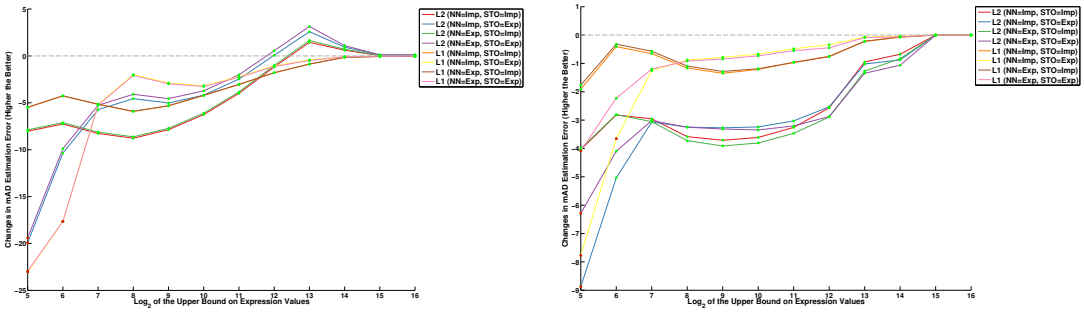


Figure 5.16.: Dataset-specific changes in the performance of deconvolution methods after adaptive range filtering

and in fact in some cases; for example the ideal datasets such as **LiverBrainLung**, **RatBrain**, and **BreastBlood**; it can be counterproductive.



(a) LiverBrainLung (b) RatBrain

Figure 5.17.: Individual performance plots for range filtering in datasets which range filtering exhibits negative effect on the deconvolution

5.3.4 Selection of Marker Genes – The Good, Bad, and Ugly

Selecting marker genes that uniquely identify a certain tissue or cell-type, prior to deconvolution, can help in improving the conditioning of matrix $\mathbf{G}$, thus improving its discriminating power and stability of results, as well as decreasing the overall computation time. A key challenge in identifying "marker" genes is the choice of method that is used to assess selectivity of genes. Various parametric and nonparametric methods have been proposed in literature to identify differentially expressed genes between two groups [32, 80] or between a group and other groups [195]. Furthermore, different methods have been developed in parallel to identify *tissue-specific* and *tissue-selective* genes that are unique markers with high specificity to their host tissue/cell type [14, 22, 86, 124]. While choosing/developing accurate methods for identifying reliable markers is an important challenge, an in-depth discussion of the matter is beyond the scope of this article. Instead, I adopt two methods used in the literature. Abbas *et al.* [2] present a framework for choosing genes based on their overall differential expression. For each gene, they use a t-test to compare the cell-type with the highest expression with the second and third highest expressing cell-type. Then, they sort all genes and construct a sequence of basis matrices with increasing sizes. Finally, they use condition number to identify an "optimal" cut among top-ranked

genes that minimizes the condition number. Newman *et al.* [129] propose a modification to the method of Abbas *et al.*, in which genes are not sorted based on their overall differential expression, but according to their tissue-specific expression when compared to all other cell-types. After prefiltering differentially expressed genes, they sort genes based on their expression fold ratio and use a similar cutoff that optimizes the condition number. Note that the former method increases the size of the basis matrix by one at each step, while the latter method increases it by q (number of cell-types). The method of Newman *et al.* has the benefit that it chooses a similar number of markers per cell-type, which is useful in cases where one of the references has a significantly higher number of markers.

I implement both methods and assess their performance over the datasets. I observe slightly better performance with the second method and use it for the rest of my experiments. Due to unexpected behavior of the **Retina** dataset, as well as a low number of significant markers in all my trials, I eliminate this dataset from further study. In identifying differentially expressed genes, a key parameter is the q-value cutoff to report significant features. The distribution of corrected p-values exhibits high similarity among ideal mixtures, while differing significantly in **CellLines** mixtures and both **PERT** datasets. I find the range of $10^{-3} - 10^{-5}$ to be an optimal balance between these two cases and perform experiments to test different cutoff values. Figure 5.18 shows changes in the mAD measure after applying marker detection, using a q-value cutoff of 10^{-3}, which resulted in the best overall performance in my study. I observe that the **PERT_Uncultured** and **LiverBrainLung** datasets have the highest gain across the majority of configurations, while **BreastBlood** and **RatBrain** exhibit an improvement in experiments with $\boldsymbol{\mathcal{L}}_1$ while their $\boldsymbol{\mathcal{L}}_2$ performance is greatly decreased. Finally, for the **PERT_Cultured** and **CellLines** datasets, I observe an overall decrease in performance in almost all configurations.

Next, I note that the internal sorting based on fold-ratio intrinsically prioritizes highly expressed genes and is susceptible to noisy outliers. To test this hypothesis, I perform a range selection using a global upper bound of 10^{12} prior to the marker

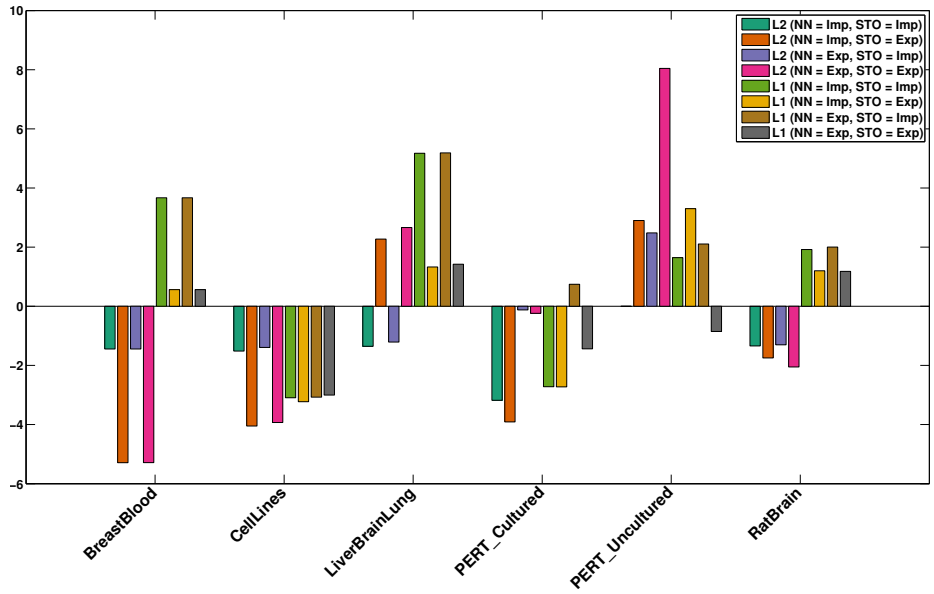


Figure 5.18.: Effect of marker selection on the performance of deconvolution methods

selection method and examine if this combination can enhance my previous results. I find the q-value threshold of 10^{-5} to be the better choice in this case. Figure 5.19 shows changes in performance of different methods when I prefilter expression ranges prior to marker selection. The most notable change is that both the **PERT_Cultured** and the **CellLines**, which were among the datasets with the lowest performance in the previous experiment, are now among the best-performing datasets, in terms of overall mAD enhancement. I still observe a higher negative impact on $\boldsymbol{\mathcal{L}}_2$ in this case, but the overall amplitude of the effect has been dampened in both **BreastBlood** and **RatBrain** datasets.

I note that there is no prior knowledge as to the "proper" choice of the marker selection method in the literature and that their effect on the deconvolution performance is unclear. An in-depth comparison of marker detection methods can benefit future developments in this field. An ideal marker should serve two purpose: (i) be highly informative of the cell-type in which it is expressed, (ii) shows low variance due to spatiotemporal changes in the environment (changes in time or microenvironment). Figure 5.20 shows a high-level classification of genes. An ideal marker is an invari-

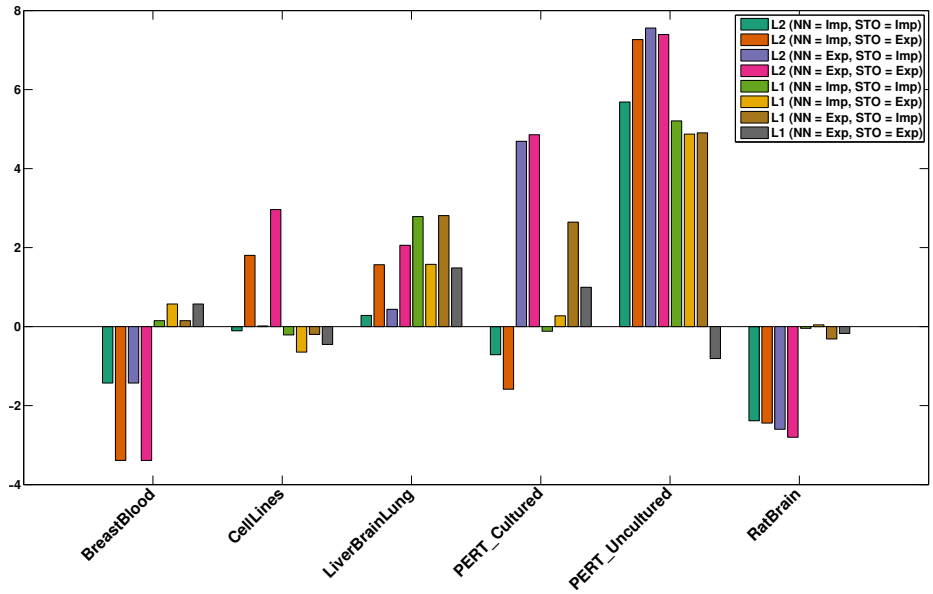


Figure 5.19.: Effect of marker selection, after range filtering, on the performance of deconvolution methods

ant, cell-type specific gene, marked with green in the diagram. On the other hand, variant genes, both universally expressed and tissue-specific, are not good candidates, especially in cases where references are adopted from a different study. These genes, however, comprise ideal subsets of genes that should be updated in full deconvolution while updating matrix **G**, since their expression in the reference profile may differ significantly from the true cell-types in the mixture. It is worth mentioning that the proper ordering to identify best markers is to first identify tissue-specific genes and then prune them based on their variability. Otherwise, when selecting invariant genes, I may select many housekeeping genes, since their expression is known to be more uniform compared to tissue-specific genes.

Another observation relates to the case in which groups of profiles of cell-types have high similarity within the group, but are significantly distant from the rest. This makes identifying marker genes more challenging for these groups of cell-types. An instance of this problem is when I consider markers in the **PERT** datasets. In this case, erythrocytes have a much larger number of distinguishing markers compared to

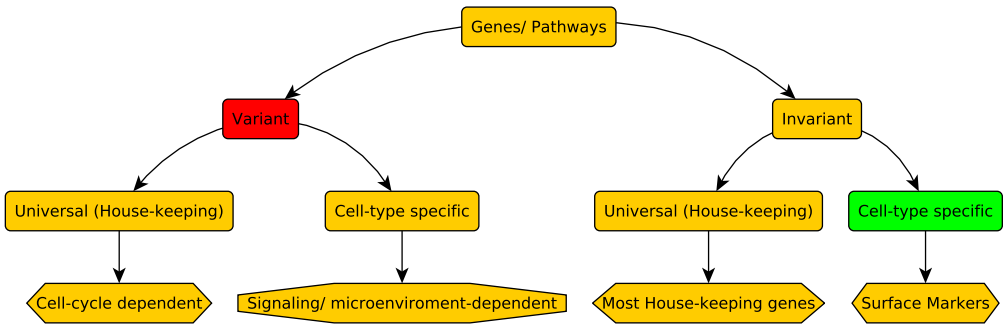


Figure 5.20.: High-level functional classification of genes

other references. This phenomenon is primarily attributed to the underlying similarity between undifferentiated cell-types in the **PERT** datasets, and their distance from the fully differentiated red blood cells. In these cases, it is beneficial to summarize each group of similar tissues using a "representative profile" for the whole group, and to use a hierarchical structure to recursively identify markers at different levels of resolution [124].

Finally, I examine the common choice of condition number as the optimal choice to select the number of markers. First, unlike the "U" shape plot reported in previous studies, in which condition number initially decreases to an optimal point and then starts increasing, I observe variable behavior in the condition number plot, both for Newman *et al.* and Abbas *et al.* methods. This makes the generalization of condition number as a measure applicable to all datasets infeasible. Additionally, I note that the lowest condition number is achieved if $\mathbf{G}$ is fully orthogonal, that is $\mathbf{G}^T\mathbf{G} = \kappa\mathbf{I}$ for any constant κ. By selecting tissue-selective markers, I can ensure that the product of columns in the resulting matrix is close to zero. However, the norm-2 of each column can still be different. I developed a method that specifically grows the basis matrix by accounting for the norm equality across columns. I find that in all cases my basis matrix has a lower condition number than both the Newman *et al.* and Abbas *et al.* methods, but it did not always improve the overall performance of deconvolution

methods using different loss functions. Further study on the optimal choice of the number of markers is another key question that needs further investigation

5.3.5 To Regularize or Not to Regularize

I now evaluate the impact of regularization on the performance of different deconvolution methods. To isolate the effect of the regularizer from prior filtering/feature selection steps, I apply regularization on the original datasets. The $\boldsymbol{\mathcal{R}}_1$ regularizer is typically applied in cases where the solution space is large, that is, the total number of available reference cell-types is a superset of the true cell-types in the mixture. This type of regularization acts as a "selector" to choose the most relevant cell-types and zero-out the coefficients for the rest of the cell-types. This has the effect of enforcing a sparsity pattern. Datasets used in this study are all controlled benchmarks in which references are hand-picked to match the ones in the mixture; thus, sparsifying the solution does not add value to the deconvolution process. On the other hand, an $\boldsymbol{\mathcal{R}}_2$ regularizer, also known as Tikhonov regularization, is most commonly used when the problem is ill-posed. This is the case, for example, when the underlying cell-types are highly correlated with each other, which introduces dependency among columns of the basis matrix. In order to quantify the impact of this type of regularization on the performance of deconvolution methods, I perform an experiment similar to the one in Section 5.3.1 with an added $\boldsymbol{\mathcal{R}}_2$ regularizer. In this experiment, I use $\boldsymbol{\mathcal{L}}_1$ and $\boldsymbol{\mathcal{L}}_2$ loss functions, as I previously showed that the performance of the other two loss functions is similar to $\boldsymbol{\mathcal{L}}_1$. Instead of using Ridge regression, I implement an equivalent formulation, $\| \boldsymbol{m} - \mathbf{G}\boldsymbol{c} \|_2 + \lambda \| \boldsymbol{c} \|_1$, which traces the same path but has higher numerical accuracy. To identify the optimal value of the λ parameter that balances the relative importance of solution fit versus regularization, I search over the range of $\{10^{-7}, \cdots, 10^{7}\}$. It is notable here that when λ is close to zero, the solution is identical to the one without regularization, whereas when $\lambda \to \infty$ the deconvolu-

tion process is only guided by the solution size. Similar to the range filtering step in Section 5.3.3, I use the minimum mAD error to choose the optimal value of λ.

Figure 5.21 presents changes in mAD error, compared to original errors, after regularizing loss functions with the $\boldsymbol{\mathcal{R}}_2$ regularizer. From these observations, it appears that **PERT_Cultured** has the most gain due to regularization, whereas for **PERT_Uncultured**, the changes are smaller. A detailed investigation, however, suggests that in the majority of cases for **PERT_Cultured**, the performance gain is due to over shrinkage of vector $\boldsymbol{c}$ to the case of being almost uniform. Interestingly, the choice of uniform $\boldsymbol{c}$ has lower mAD error for this dataset compared to most other results. Overall, both of the **PERT** datasets show significant improvements compared to the original solution, which can be attributed to the underlying similarity among hematopoietic cells. On the other hand, an unexpected observation is the performance gain over $\boldsymbol{\mathcal{L}}_1$ configurations for the **BreastBlood** dataset. This is primarily explained by the limited number of cell-types (only two), combined with the similar concentrations used in all samples (only combinations of 67% and 33%).

To gain additional insight into the parameters used in each case during deconvolution, I plot the optimal λ values for each configuration in each dataset. Figure 5.22 summarizes the optimal values of the λ parameter. Large values indicate a beneficial effect for regularization, whereas small values are suggestive of negative impact. In all cases where the overall mAD score has been improved, their corresponding λ parameter was large. However, large values of λ do not necessarily indicate a significant impact on the final solution, as is evident in the **CellLines** and **LiverBrainLung** datasets. Finally, we observe that cases where the value of λ is close to zero are primarily associated with the $\boldsymbol{\mathcal{L}}_2$ loss function.

5.3.6 Putting it All Together

Having analyzed each individual aspect that impacts the performance of the deconvolution process, in this section I put all of the pieces together and evaluate the overall

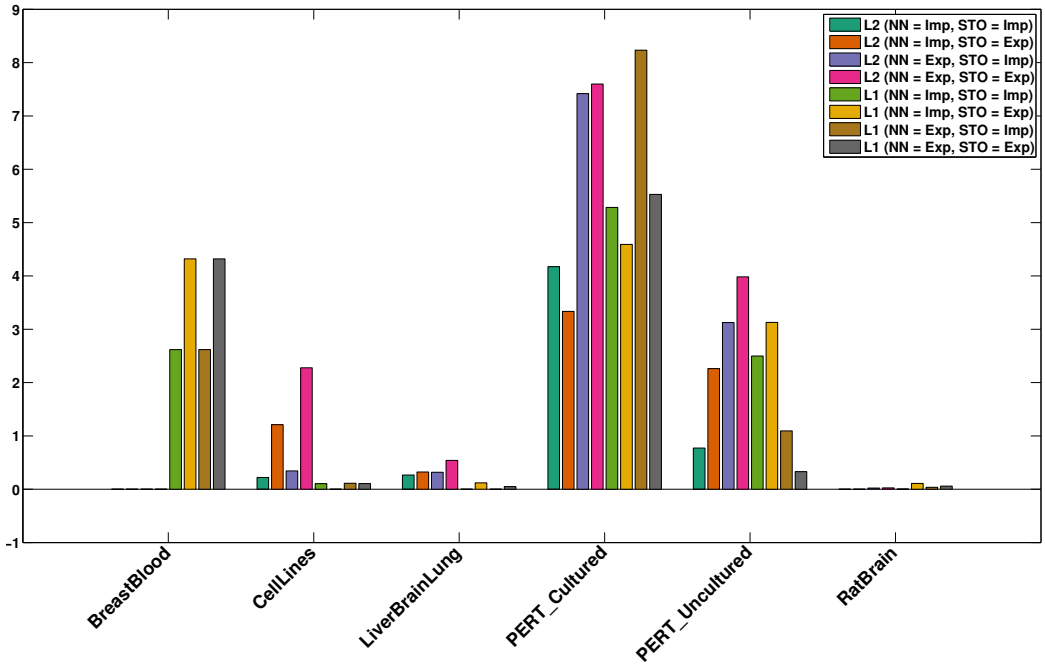


Figure 5.21.: Effect of L2 regularization on the performance of deconvolution methods

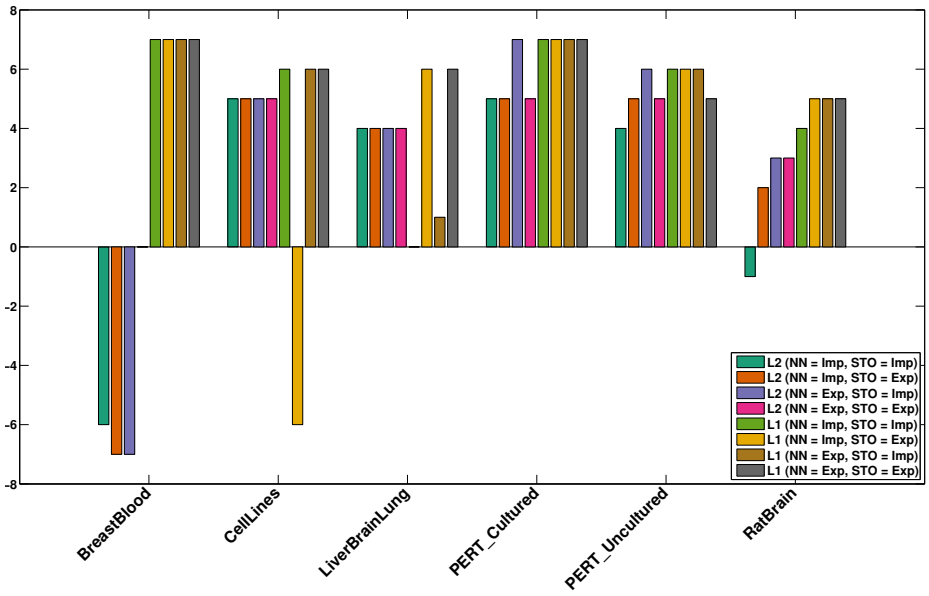


Figure 5.22.: Optimal value of λ for each dataset/configuration pair

performance over each dataset. I remove the **Retina** dataset from this study due to the observed discrepancies. For the remaining six datasets, I assess performance of both $\boldsymbol{\mathcal{L}}_1$ and $\boldsymbol{\mathcal{L}}_2$ objectives with different combinations of NN/STO enforcement, for a total of eight configurations. For each configuration, I use my previous results to determine proper feature selection, i.e. whether or not to remove violating features and/ or to select marker genes. Finally, I note that my results in Section 5.3.5, while instructive, are not directly applicable here due to differences in the selected subset of genes. Thus, for each configuration, I rerun the experiment without regularization, as well as with regularization with $\lambda \in \{10^{-7}, \ldots, 10^{7}\}$.

Table 5.4 summarizes the settings used to solve each instance. There are some general patterns to note here. First, for the **RatBrain** dataset, I are at the lowest attainable mAD error, and neither removing violating features, nor selecting markers can boost this. For this dataset, mAD is significantly lower than the rest of datasets. This is likely due to the existence of highly expressed genes that just happen to align, with low variation, between reference profiles and mixtures. This, however, has the potential downfall of overfitting, in which case, the best configurations identified in this dataset are not generalizable to other datasets.

Next, I observe that in a majority of cases, filtering violating features, on average, either decreases mAD error or at least it does not increase it, with the previously mentioned exception of **RatBrain**. Similarly, selecting marker genes, combined with range filtering, in most cases improves deconvolution results, except in **RatBrain** and **BreastBlood**. For the **BreastBlood** dataset, I argue that the quality of selected markers might be affected because I only have two cell-types, but this needs further validation. Finally, I note either marker selection or range filtering alone performs much worse than combining them together.

The final results of my experiments, before and after feature selection/ regularization, are shown in Figure 5.23. Shaded bars correspond to the original performance for each configuration, and colored bars are the final mAD errors computed. Interestingly, after suitable feature selection, the results for most of the datasets are within

similar error ranges, approximately within the range of $[5-7]$ mAD. Furthermore, $\boldsymbol{\mathcal{L}}_2$ seems to perform equally as good as $\boldsymbol{\mathcal{L}}_1$, given the proper subspace of genes to perform deconvolution in, if not better. This is consistent with my understanding, since $\boldsymbol{\mathcal{L}}_2$ has higher sensitivity compared to $\boldsymbol{\mathcal{L}}_1$. Even in both **PERT** datasets, I observed only minor differences between these two objectives. Furthermore, $Loss_2$ has much more efficient solvers compared to $\boldsymbol{\mathcal{L}}_1$. Thus, in general I recommend using $\boldsymbol{\mathcal{L}}_2$, and to only resort to $\boldsymbol{\mathcal{L}}_1$ if $\boldsymbol{\mathcal{L}}_2$ does not perform well. However, I can not generalize this claim to all datasets. In cases with high level of noise and/ or imperfect marker selection, $Loss_1$ still makes a better choice. Finally, contrary to traditional wisdom, I observe a dataset-dependency for the effect of constraint enforcement. That is, explicitly encoding all constraints in the objective function does not always enhance the quality of final results. A deeper understanding of this observation can guide one to choosing the right formulation for the problem at hand.

5.3.7 Summary

Based on my observations, I propose the following guidelines for the deconvolution of expression datasets:

1. Preprocess reference profiles and mixtures using invariant, universally expressed (housekeeping) genes to ensure that the *similar cell quantity (SCQ)* constraint is satisfied.

2. Filter violating features that cannot satisfy the *sum-to-one (STO)* constraint.

3. Filter lower and upper bounds of gene expressions using adaptive range filtering.

4. Select invariant (among references and between references and samples) cell-type-specific markers to enhance the discriminating power of the basis matrix.

5. Solve the regression using the $\boldsymbol{\mathcal{L}}_2$ loss function, together with an $\boldsymbol{\mathcal{R}}_2$ regularizer, or group LASSO if sparsity is desired among groups of tissues/cell-types.

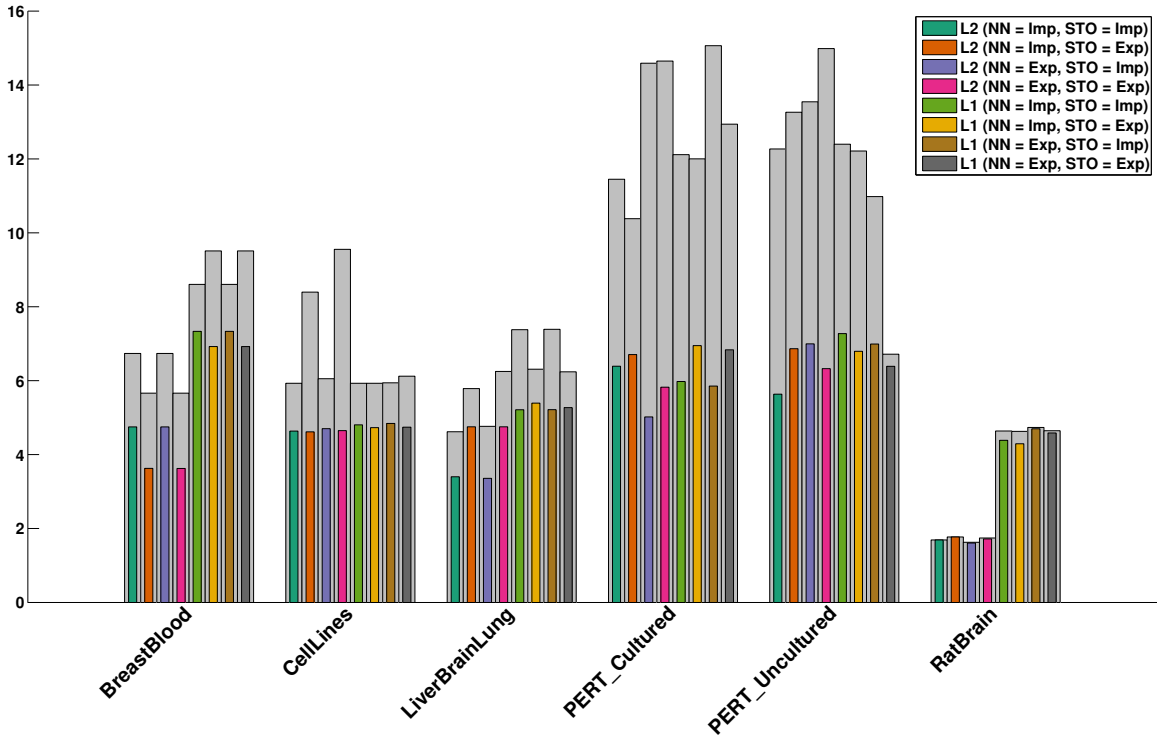


Figure 5.23.: Performance of deconvolution before/after applying combined feature selection/regularization. Gray shade is mAD of the original deconvolutions (smaller the better).

Table 5.4: Best combination of choices for feature selection/regularization for different datasets

Loss Function		$\mathcal{L}_2$				$\mathcal{L}_1$			
Non-negativity		+		-		+		-	
Sum-to-One		-	+	-	+	-	+	-	+
BreastBlood	Remove violating features	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	Filter markers/range	No	No	No	No	No	Yes	No	Yes
	Best lambda	0	0.001	1E-06	1	10000000	100000	10000000	100000
CellLines	Remove violating features	No	Yes	No	Yes	No	No	No	No
	Filter markers/range	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	Best lambda	0.0001	10	100	10	10	10000	0.001	10000
LiverBrainLung	Remove violating features	Yes	Yes	Yes	Yes	Yes	No	Yes	No
	Filter markers/range	No	Yes	No	Yes	Yes	Yes	Yes	Yes
	Best lambda	10000	10000	10000	10000	1000	1000	1000	0.0001
PERT_Cultured	Remove violating features	No	No	Yes	Yes	Yes	Yes	Yes	Yes
	Filter markers/range	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	Best lambda	10000	10000	1000	1000	10000	100000	10000	10000
PERT_Uncultured	Remove violating features	No	Yes	Yes	Yes	Yes	Yes	Yes	No
	Filter markers/range	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
	Best lambda	1000	1000	1000	1000	10000	10000	1000	100000
RatBrain	Remove violating features	No	No	No	No	No	No	No	No
	Filter markers/range	No	No	No	No	Yes	Yes	No	No
	Best lambda	0.1	100	1000	1000	1E-06	0.001	100000	100000

6. Use the L-curve method to identify the optimal balance between the regression fit and the regularization penalty.

6 CONSTRUCTING CELL TYPE SPECIFIC INTERACTOMES

6.1 Background

Proteins are basic workhorses of living cells. Their overall quantity is tightly regulated across different tissues and cell-types to manifest tissue-specific biology and pathobiology. These regulatory controls orchestrate cellular machinery at different levels of resolution, including, but not limited to, gene regulation [62, 120], epigenetic modification [24, 121], alternative splicing [19, 46], and post-translational modifications [77, 194]. Transcriptional regulation is a key component of this hierarchical regulation, which has been widely used to study context-specific phenotypes. In the context of human tissues/ cell-types, genes can exhibit varying levels of specificity in their expression. They can be broadly classified as: (i) tissue-specific (unique to one cell-type); (ii) tissue-selective (shared among coherent groups of cell-types); and (iii) housekeeping (utilized in all cell-types). Tissue-specific/selective genes have significant applications in drug discovery, since they have been shown to be more likely drug targets [40]. Tissue-specific transcription factors (tsTFs) are significantly implicated in human diseases [123, 150], including cancers [197]. Finally, disease genes and protein complexes tend to be over-expressed in tissues in which defects cause pathology [103].

The majority of human proteins do not work in isolation but take part in pathways, complexes, and other functional modules. Tissue-specific proteins are known to follow a similar trend. Perturbations that impact interacting interfaces of proteins are significantly enriched among tissue-specific, disease-causing variants [155,157,208]. This emphasizes the importance of constructing tissue-specific interactomes and their constitutive pathways for understanding mechanisms that differentiate cell-types and make them uniquely susceptible to tissue-specific disorders. Prior attempts at re-

constructing human tissue-specific interactomes rely on a set of *"expressed genes"* in each tissue, and use this set as the baseline of transcriptional activity. The node removal (NR) method [16] constructs tissue-specific interactomes by identifying the induced subgraph of the expressed genes. Magger *et al.* [115] propose a method called *"Edge ReWeighting (ERW)"*, which extends the NR method to weighted graphs. This method penalizes an edge once, if one of its end-points is not expressed, and twice, if both end-points are missing from the expressed gene-set.

While these methods have been used to study tissue-specific interactions, their underlying construction relies only on the immediate end-points of each interaction to infer tissue-specificity. Furthermore, they threshold expression values, often using ad-hoc choices of thresholds to classify genes as either expressed or not. Finally, it is hard to integrate expression datasets from multiple platforms, or from multiple labs, into a single framework. These constraints are primarily dictated by limitations of high-throughput technologies for assaying gene expression. In these technologies, one can easily compare expression of the same gene across different samples to perform differential analysis; however, expression of different genes in the same sample are not directly comparable due to technical biases, differences in baseline expression, and GC content of genes. A recently proposed method, *Universal exPression Code (UPC)* [143], addresses many of these issues by removing platform-specific biases and converting raw expressions to a unified transcriptional activity score. These scores are properly normalized and can be compared across different genes and platforms.

Leveraging the UPC method, I propose a novel approach that uses the topological context of an interaction to infer its specificity score. my approach formulates the inference problem as a suitably regularized convex optimization problem. The objective function of the optimization problem has two terms – the first term corresponds to a *diffusion kernel* that propagates activity of genes through interactions (network links). The second term is a *regularizer* that penalizes differences between *transcriptional* and *functional* activity scores. I use these functional activity scores to compute tissue-specificity for each edge in the global interactome, which I show, through a

number of validation tests, are significantly better than prior methods. my method is widely applicable and can be applied directly to single-channel, double-channel, and RNA-Seq expression datasets processed using UPC/SCAN. Furthermore, it can be easily adapted to cases where expression profiles are only available in preprocessed form.

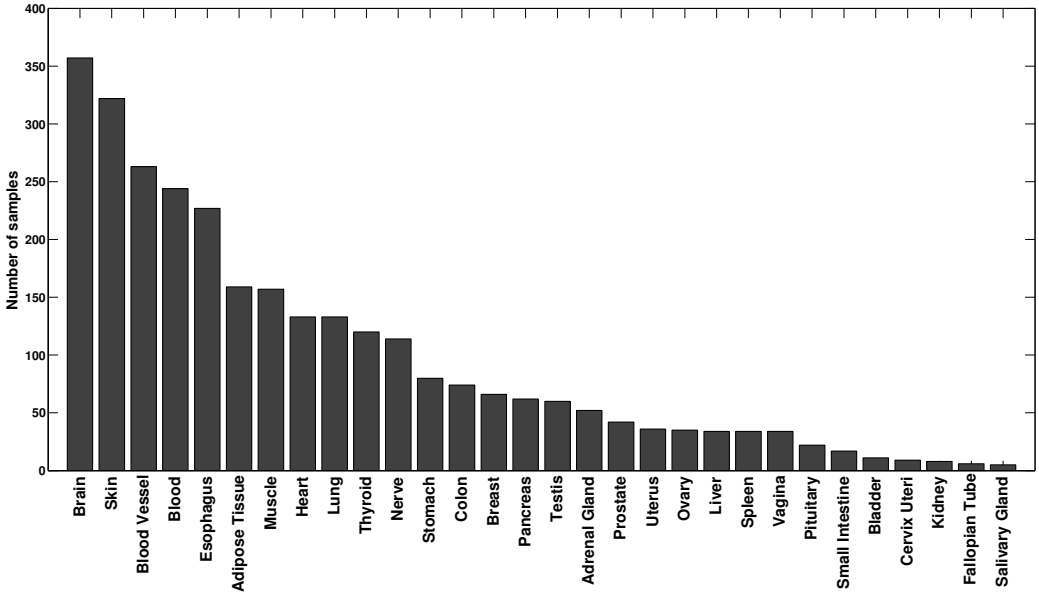
The rest of this chapter is organized as follows: In Section 6.2.1 I provide details of the datasets used in my study. Next, I introduce a new method method, called *Activity Propagation (ActPro)*, and provide a consistent notation to formalize previous methods. I evaluate the effectiveness of UPC transcriptional activity scores to predict tissue-specific genes in Section 6.3.1. Details of procedure for constructing tissue-specific networks and their parameter choices are discussed in Section 6.3.2. Section 6.3.3 provides qualitative assessment of my tissue-specific networks, whereas Sections 6.3.4-6.3.6 present validation studies for tissue-specific interactions using known pathway edges, co-annotation of proteins, and GWAS disease genes. Finally, in Section 6.3.7, I use the brain-specific interactome constructed using my method to identify novel disease-related pathways and use them to identify candidate targets for neurodegenerative disorders.

6.2 Materials and Methods

6.2.1 Datasets

I downloaded the RNASeq dataset version 4.0 (*dbGaP accession phs000424.v4.p1*) from the The Genotype-Tissue Expression (GTEx) project [6, 120]. This dataset contains 2,916 samples from 30 different tissues/cell types, the summary of which is presented in Figure 6.1. I processed each sample using the UPC method [143], a novel platform-independent normalization technique that corrects for platform-specific technical variations and estimates the probability of transcriptional activity for each gene in a given sample. The benefit of this method is that activation probability scores are highly consistent across different technologies, and more im-

Figure 6.1.: Summary of GTEx sample numbers per tissue



portantly, they are comparable across different genes in a given sample. For each gene, I recorded the transcript with the highest activation probability in the sample. Finally, I averaged replicate samples within each group to construct a unique transcription signature vector for each tissue/ cell type. The final dataset contains the expression value of 23,243 genes across 30 different tissues/ cell types.

In addition, I extracted human protein-protein interactions from the iRefIndex database [152], which consolidates protein interactions from different databases. Edges in this dataset are weighted using an MI (MINT-Inspired) score, which measures the confidence of each interaction based on three different evidence types, namely (i) the interaction types (binary/complex) and experimental method used for detection, (ii) the total number of unique PubMed publications reporting the interaction, and (iii) the cumulative evidence of interlogous interactions from other species. Finally, I map transcription data to the human interactome by converting all gene IDs to Entrez Gene IDs and only retaining genes that both have a corresponding node in the interactome and have been profiled by the GTEx project. This yields a global in-

teractome with 147,444 edges, corresponding to protein-protein interactions, between 14,658 nodes, representing gene products.

6.2.2 Constructing Human Tissue-specific Interactome

The global human interactome is a superset of all *possible* physical interactions that can take place in the cell. It does not provide any information as to which interactions actually occur in a given context. There are a variety of factors, including co-expression of genes corresponding to a pair of proteins, their co-localization, and post-translational modification, that mediate protein interactions at the right time and place. Quantifiable expression of both proteins involved in an interaction is one of the most important factors that determine the existence of an interaction. Different methods have been proposed in literature to utilize this source of information to construct human tissue-specific interactomes. Here, I briefly review existing methods, their drawbacks, and propose a new method, called *Activity Propagation (ActPro)*, which addresses noted shortcomings.

Previous Methods

Let us denote the adjacency matrix of the global interatome by $\mathbf{A}$, where element $\boldsymbol{a}_{ij}$ is the weight (confidence) of the edge connecting vertices v_i and v_j. Let $\boldsymbol{z}$ encode expression of genes in a tissue and $\underline{\boldsymbol{z}}$ be the binarized version of $\boldsymbol{z}$ for a fixed threshold. Finally, let $\boldsymbol{diag}$ operator applied to a given vector be the diagonal matrix with the vector on the main diagonal. my aim is to compute a matrix $\hat{\mathbf{A}}$, which is the adjacency matrix of the tissue-specific interactome for a given expression profile. Using this notation, I can summarize prior methods for constructing tissue-specific interactomes as follows.

- **Node Removal (NR)** This method computes the induced subgraph of the "expressed" gene products [16].

$$\hat{\mathbf{A}} = \boldsymbol{diag}(\underline{\boldsymbol{z}}) * \mathbf{A} * \boldsymbol{diag}(\underline{\boldsymbol{z}}) \tag{6.1}$$

- **Edge Re-Weighting (ERW)** This method penalizes edges according to the expression state (active/ inactive) of its end points [115]. Given a penalty parameter $0 \leq rw \leq 1$, ERW penalizes each edge by rw once, if only one of its end-points is active, and twice, if both incident vertices are inactive. Formally:

$$\hat{\mathbf{A}} = \boldsymbol{diag}(rw^{(\boldsymbol{e}-\underline{\boldsymbol{z}})}) * \mathbf{A} * \boldsymbol{diag}(rw^{(\boldsymbol{e}-\underline{\boldsymbol{z}})}) \tag{6.2}$$

  where $\boldsymbol{e}$ is the vector of all ones.

Proposed Method

The main assumption of *ERW* and *NR* methods is that *transcriptional activity* of a gene is a reliable proxy for its *functional activity*. While this holds in a majority of cases, there are situations in which these scores differ significantly. First, the basis for *transcriptional activity* estimation is that genes with higher expression levels have higher chance of being functionally active in a given context. While this is generally true, there are genes that only need a low expression level to perform their function; i.e., their functionally active concentration is much lower than the rest of genes. Second, there is noise associated with measurement of gene expression, and converting measured expression values to UPC scores can over/ under-estimate *transcriptional activity*. Finally, we note that there are genes whose *down-regulation* corresponds to their functional activity (as opposed to the other way around).

Based on these observations, I propose a novel framework, called *Activity Propagation (ActPro)* to identify the most functionally active subnetwork of a given interactome. my method incorporates global network topology to propagate activity scores, while simultaneously minimizing the number of changes to the gene activity

scores. To this end, I first define a smoothed functional activity score defined by the following optimization problem:

$$\boldsymbol{x}^* = \underset{\boldsymbol{x}}{\mathbf{argmin}}\left\{ \frac{\alpha}{|E|}\boldsymbol{x}\mathbf{L}\boldsymbol{x} + \frac{(1-\alpha)}{|V|} \parallel \boldsymbol{x} - \boldsymbol{z} \parallel_1 \right\}$$
$$\text{Subject to: } \begin{cases} \mathbf{1}^T\boldsymbol{x} = 1 \\ 0 \leq \boldsymbol{x} \end{cases} \tag{6.3}$$

In this problem, $\mathbf{L}$ is the *Laplacian* matrix, defined as $\mathbf{A} - \mathbf{\Delta}$, where element δ_{ii} of $\mathbf{\Delta}$ is the weighted degree of ith vertex in the global interactome. The Laplacian operator $\mathbf{L}$ acts on a given function defined over vertices of a graph, such as $\boldsymbol{x}$, and computes the smoothness of $\boldsymbol{x}$ over adjacent vertices. More specifically, we can expand the first term in Equation 6.3 as $\sum_{i,j} w_{i,j}(x_i - x_j)^2$, which is the accumulated difference of values between adjacent nodes scaled by the weight of the edge connecting them. This term defines a *diffusion kernel* that propagates activity of genes through network links. The second term is a *regularizer*, which penalizes changes by enforcing sparsity over the vector of differences between *transcriptional* and *functional* activities. This minimizes deviation from original transcriptome. It should be noted here that use of norm-1 is critical, since norm-2 regularization blends the transcriptional activity scores and significantly reduces their discriminating power. This negative aspect of norm-2 minimization is confirmed by my experiments. Finally, constraint $\mathbf{1}^T\boldsymbol{x} = 1$ is known as the fixed budget. It ensures that vector $\boldsymbol{x}$ is normalized and bounded. Parameter α determines the relative importance of regularization versus loss. We can equivalently define a penalization parameter $\lambda = \frac{1-\alpha}{\alpha}$, which is the standard notation in optimization framework. This problem is a classical convex optimization problem and we can solve it using efficient solvers to identify its global optimum.

After solving Equation 6.3, I first scale $\boldsymbol{x}^*$ by $|V|$. These scores are centered around 1, which allows us to perform minimal changes to the weight of interactions in the global interactome. Using these smoothed activity scores, I can re-weight the global human interactome as follows:

$$\hat{\mathbf{A}} = \boldsymbol{diag}(\boldsymbol{x}^*) * \mathbf{A} * \boldsymbol{diag}(\boldsymbol{x}^*) \tag{6.4}$$

We can also derive an alternative formulation for *ActPro* which, instead of using *transcriptional activity* scores computed by UPC, uses expression values computed through more common methods such as RMA or MAS5.0 [108]. I call this method *penalty propagation*, or *PenPro* for short. In this framework, computed expression values are not directly comparable and I need to threshold them to classify genes as either expressed or not. Using the same notation defined previously, I can define *functional activity* scores by solving the following problem:

$$\boldsymbol{x}^* = \underset{\boldsymbol{x}}{\mathbf{argmin}} \left\{ \frac{\alpha}{|E|} \boldsymbol{x}\mathbf{L}\boldsymbol{x} + \frac{(1-\alpha)}{|V|} \parallel \boldsymbol{x} - \underline{\boldsymbol{z}} \parallel_1 \right\}$$
$$\text{Subject to: } \begin{cases} \mathbf{1}^T \boldsymbol{x} = 1 \\ 0 \leq \boldsymbol{x} \end{cases} \tag{6.5}$$

The only difference here is that, instead of *transcriptional activity* vector $\boldsymbol{z}$, I use the binarized expression vector $\underline{\boldsymbol{z}}$. We observe similar performance for *ActPro* and *PenPro*, with *ActPro* being marginally superior in all cases, and thus I will only present results for *ActPro*.

6.2.3 Implementation Details

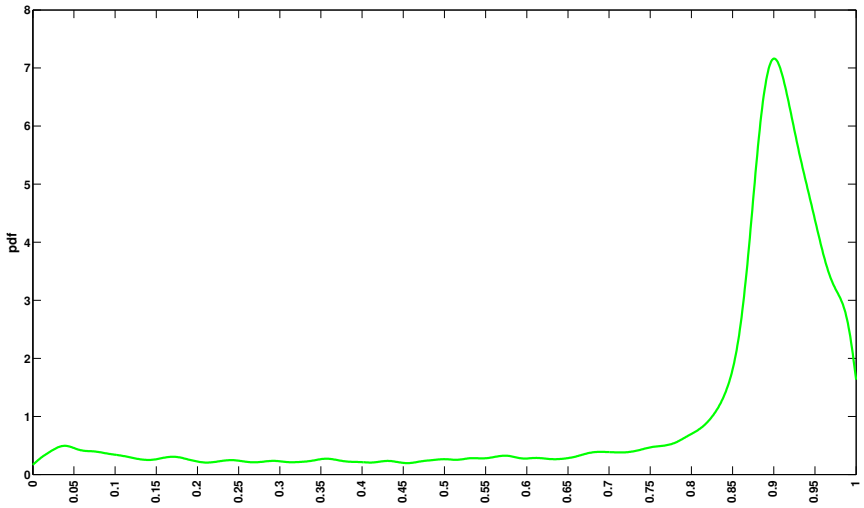
All codes used in my experiments have been implemented in Matlab. To solve the convex problem in Equation 6.3, I used CVX, a package for specifying and solving convex programs [63]. I used Mosek together with CVX, which is a high-performance solver for large-scale linear and quadratic programs [126].

6.3 Results and Discussion

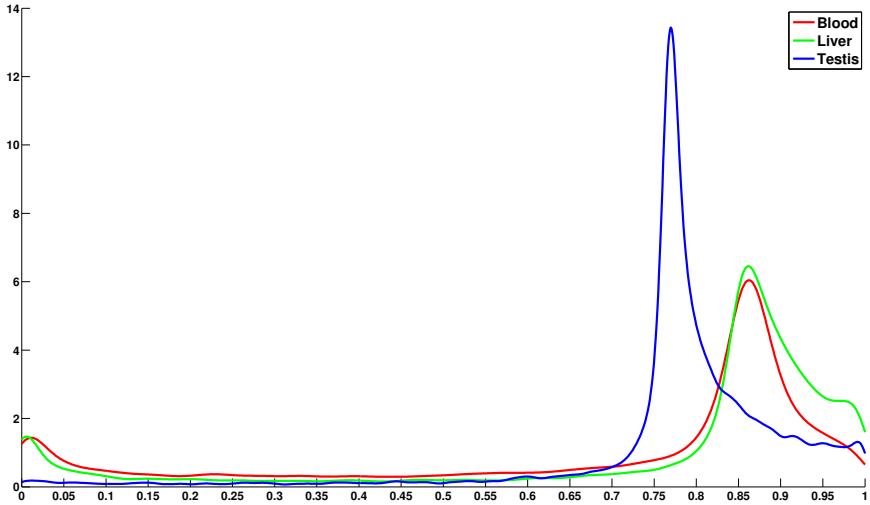
6.3.1 Transcriptional Activity Scores Predict Tissue-specificity of Genes

To validate the quality of UPC normalized expression values, I first analyze the distribution of gene expressions across all tissues. Figure 6.2a shows the distribution of transcriptional activities, averaged over all samples. The overall distribution exhibits

a bimodal characteristic that has a clear separation point that distinguishes expressed genes from others. I set a global threshold of 0.75 for identifying genes that are expressed in each tissue. These genes are used in evaluating NR and ERW methods. It should be noted that the distribution of UPC values vary across cell-types, as shown in Figure 6.2b; however, the separation point is robust.



(a) Average expression across all samples



(b) Three tissues with low, medium, and high number of expressed genes

Figure 6.2.: Distribution of UPC normalized gene expression values

Expression value of genes across tissues can be classified as *specific, selective*, or *housekeeping*. Housekeeping (HK) genes are ubiquitously expressed across all tissues to perform core cellular functions. On the other hand, tissue specific/selective genes are uniquely expressed in a given tissue context to perform tissue-specific functions. These genes typically reside in the periphery of the network, are enriched among signaling and cell surface receptors, and highly associated with the onset of tissue-specific disorders [216]. Figure 6.3a shows the total number of genes identified in each tissue as preferentially expressed (either specific or selective). Testis tissue exhibits the largest number of preferentially expressed genes (I refer to these as markers), with more that $1,400$ genes, while blood samples have the fewest markers with only ~ 250 marker genes. In order to assess whether the sets of preferentially expressed genes can predict tissue-specific functions, I performed GO enrichment analysis over different sets of tissue-specific markers using *GOsummaries* package in R/Bioconductor [95]. This package uses g:Profiler [153] as backend for enrichment analysis and provides a simple visualization of the results as a word cloud. The coverage of available annotations for different tissues is not uniform; that is, some tissues are better annotated for specific terms than the others. I chose six well-annotated tissues with high, mid, and low number of identified markers for further study. I limited terms to the ones with at least 20 and at most 500 genes to avoid overly generic/specific terms. Finally, I used a strong hierarchical filtering to remove duplicate GO terms and thresholded terms at p-value of 0.05. Figure 6.3b shows the enrichment word-cloud for each tissue. It can be seen that all terms identified here are highly tissue-specific and representative of main functions for each tissue, which supports the validity of computed transcriptional activity scores from UPC.

6.3.2 Constructing Tissue-specific Interactomes

Node Removal (NR) and Edge ReWeighting (ERW) methods need a predefined set of expressed genes in each tissue to construct tissue-specific interactomes (or a

given lower bound to threshold expression values). I use the set of all genes with transcriptional activity greater than or equal to 0.75 as the set of expressed genes for these methods. I chose this threshold based on the averaged distribution of gene expressions, as well as further manual curation of genes at different thresholds.

Node Removal (NR) method is known to disintegrate the network with stringent expression values [115]. To evaluate the performance of NR over different expression thresholds and assess its sensitivity to the choice of threshold, I computed the size of largest connected components, while varying the value of expression threshold. Figure 6.4 shows stable behavior up to threshold value of 0.75, after which the size of largest component exhibit a rapid shift and the network starts to disintegrate. This suggests that the expression value of 0.75 is also the optimal topological choice for NR method.

For the ActPro algorithm, I evaluated the results over three different values of α in set $\{0.15, 0.5, 0.85\}$ and reported the result for each case.

6.3.3 Qualitative Characterization of Tissue-specific Interactomes

A key feature of tissue specific networks is their ability to discriminate positive edges that manifest in each tissue from the entire set of potential interactions in the global interactome. In case of Node Removal (NR) and Edge ReWeighting (ERW) methods, it is easy to distinguish positive and negative edges: every edge for which at least one of the endpoints is not expressed can be classified as a negative edge. The latter method updates edge weights, to account for expression of their end-points, whereas the former method sets a hard threshold to either include an edge or not. In the case of *ActPro*, I first notice that the distribution of edge weights is very different between ActPro and previous methods. Whereas NR and ERW methods never increase the weight of an edge, in ActPro edge weights can increase or decrease. This behavior, however, is biased towards the positive end. To decompose each network into its HK, positive, and negative subspaces, I use the following strategy:

for each tissue-specific network constructed by a given method, I first compute the relative weight change between the global interactome and the tissue-specific network. I then normalize these changes using Z-score normalization and define positive and negative subspaces according to the sign of normalized relative changes. I further define and separate HK edges as the subset of positive edges that are positive in at least half of the tissues. Figure 6.5 summarizes the average statistics for constructed networks using different methods. As a general observation, *ActPro* classifies fewer interactions as housekeeping and provides more specific positive and negative edges. Furthermore, as I increase the α parameter, representing the diffusion depth, we observe that these edges are more evenly distributed across vertices. To give a concrete example, I constructed the brain-specific network using *ERW* and *ActPro* methods. Figure 6.6 illustrates the final statistics of the constructed networks. Consistent with the average statistics, I observe much smaller positive/negative nodes/edges in *ERW*.

6.3.4 Tissue-specific Interactome Predicts Context-sensitive Interactions in Known Functional Pathways

To evaluate the power of tissue-specific interactions in capturing context-sensitive physical interactions in known pathways, I first use Edge Set Enrichment Analysis (ESEA) to rank pathway edges according to their gain/loss of mutual information in each tissue context [69]. ESEA aggregates pathways from seven different sources (KEGG; Reactome; Biocarta, NCI/Nature Pathway Interaction Database; SPIKE; HumanCyc; and Panther) and represents them as a graph with edges corresponding to biological relationships, resulting in over 2,300 pathways spanning 130,926 aggregated edges. It then uses an information-theoretic measure to quantify dependencies between genes based on gene expression data and ranks edges, accordingly. Formally, for each pathway edge, ESEA computes the differential correlation score (EdgeScore) as follows:

$$EdgeScore = \boldsymbol{MI}_{all}(i,j) - \boldsymbol{MI}_{control}(i,j) \tag{6.6}$$

where $\boldsymbol{MI}_{all}$ is the mutual information of the gene expression profiles for genes i and j across all cell-types. Here, $\boldsymbol{MI}_{control}$ measures the mutual information only in the given tissue context. Each edge can be classified as either a gain of correlation (GoC), loss of correlation (LoC), or no change (NC) depending on the value of $EdgeScore$. I use GoC edges, that is, a pair of genes with positive gain of mutual information in the tissue context, as true positive edges in each tissue. Similarly, I use all positive edges in all tissues but the tissue of interest as true negatives.

To assess agreement between ESEA scores over known pathway edges and computed tissue-specific interactions, I rank all edges according to the difference of their weights in the human tissue-specific interactome compared to the global interactome and evaluate the enrichment of true positive pathway edges among top-ranked edges. I compute the receiver operating characteristic (ROC) curve for each tissue and average the area under the curve (AUC) gain, compared to random baseline, over all tissues. Figure 6.7 presents the relative performance of each method. All three configurations of the *ActPro* algorithm are ranked at the top of the list – demonstrating the superior performance of my proposed method.

To further investigate tissue-specific details for the top-ranked method, *ActPro* with $\alpha = 0.15$, I sorted AUC gain for each tissue, shown in Figure 6.8. This plot exhibits high level of heterogeneity, and surprisingly, four of the tissues had worse than random performance. This was consistent across all of the methods. To further understand this, I investigated the ranked list of edges and identified a high enrichment of edges with LoC among top-ranked edges. I performed enrichment analysis over these negative edges and identified significant tissue-specific functions among them, which suggests that the poor observed performance for these tissues is attributed to their misclassification as negative edges.

At the other end of the spectrum, *Fallopian Tube*, *Vagina*, and *Cervix Uteri* had consistently high AUC gain across different methods. Figure 6.9 shows the ROC curve for these tissues.

6.3.5 Tissue-specific Interactions are Enriched among Proteins with Shared Tissue-specific Annotations

I hypothesize that tissue-specific edges are enriched with proteins that participate in similar tissue-specific functions. To evaluate my hypothesis, I collected a set of manually curated tissue-specific Gene Ontology (GO) annotations from a recent study [64]. I mapped tissues to GTEx tissues and identified tissue-specific GO annotations for genes in each tissue-specific interactome. I excluded tissues with less that 100 edges with known annotations. This resulted in 10 tissues, *Adipose Tissue, Blood Vessel, Blood, Brain, Breast, Heart, Kidney, Lung*, and *Muscle*, for which I had enough annotations. I use the same strategy employed in previous section to identify the mean gain of AUC for each method, which is illustrated in Figure 6.10. It should be noted that the gain of AUC is much smaller here than the case with ESEA edges, which can be attributed to the sparsity of tissue-specific GO annotations. Unlike ESEA, *ActPro* with $\alpha = 0.5$ outperforms the case with $\alpha = 0.15$.

Among the ten tissues, *Adipose* and *Muscle* tissues performed marginally better than the others with AUC of 0.59 and 0.58, respectively. On the other hand, *Lung* tissue had the worst performance with lower than random AUC of 0.47.

6.3.6 Tissue-specific Interactions Densely Connect Genes Corresponding to Tissue-specific Disorders

Disease genes are densely connected to each other in the interactome, which provides the basis for a number of methods for network-based disease gene prioritization [94]. Tissue-specific interactomes have been shown to have higher accuracy in predicting disease-related genes using the random-walk method [115]. More recently, Cornish *et al.* [33] used the concept of "geneset compactness", and showed that the average distance among nodes corresponding to a given disorder is significantly smaller in tissue-specific networks, compared to an ensemble of random graphs.

Here, I adopt this concept to measure how closely tissue-specific genes related to human disorders are positioned in networks constructed using different methods. First, I use a symmetric diffusion process instead of Random-Walk with Restart (RWR), which is a better measure of distance. Second, I use an alternative random model in which I hypothesize that genes corresponding to tissue-specific disorders are strongly connected to each other, compared to random genesets of the same size.

To validate my hypothesis, I gather genes corresponding to tissue-specific disorders from a recent study [73]. These genes are extracted from the GWAS Catalog by mapping known associations to disease-specific loci. Among a total of 99 disorders, I focused on the gold standard set of 29 diseases with at least 10 high-quality primary targets. I successfully mapped 27 of these diseases to GTEx tissues, which are used for the rest of my study. Consistent with previous studies [115], I observed a small subset of disease genes not to be expressed in the tissue in which they cause pathology. Among all disease genes, I only retained genes that are connected in the global interactome and are expressed above 0.1 UPC score.

For a given tissue-specific interactome represented by its adjacency matrix, $\mathbf{A}_T$, I define a stochastic matrix $\mathbf{S} = \mathbf{\Delta}^{-\frac{1}{2}}\mathbf{A}_T\mathbf{\Delta}^{-\frac{1}{2}}$, where $\mathbf{\Delta}$ is the diagonal matrix, with entries δ_{ii} being the degree of node i in the human tissue-specific interactome. Using this matrix, I can compute degree-weighted random-walk scores among gene pairs as:

$$\mathbf{P} = (1-\alpha)(\mathbf{I} - \alpha\mathbf{S})^{-1} \tag{6.7}$$

I define the random-walk distance as $d_{ij} = -log_{10}(p_{ij})$, after replacing zero elements of $\mathbf{P}$ with $\epsilon = 2^{-52}$. Given a disease geneset Γ, I measure its compactness as the normalized average of distances for all pairs of nodes in the geneset:

$$\kappa_\Gamma = \frac{\sum_{i \neq j \in \Gamma} d_{ij}}{\binom{|\Gamma|}{2}} \tag{6.8}$$

Finally, I sampled without replacement, 100K vertex samples of size $|\Gamma|$ from the tissue-specific interactome and computed the compactness for each of the samples, individually. I defined an empirical p-value as the fraction of random instances with

higher compactness (lower κ) compared to Γ. I removed disorders for which none of the methods yield significant p-value given a threshold of 0.05. The final dataset consists of 15 diseases with significantly compact interactions. To combine the p-values for different disorders, I use the Edgington method [44]. This method gathers a statistic $\mathcal{S} = \sum_{i=1}^{k} p_i$ for a set of k given *p-values*, and computes the meta p-value by assigning significance to $\mathcal{S}$ as:

$$\sum_{j=0}^{\lfloor \mathcal{S} \rfloor} -1^{j} \binom{k}{j} \frac{(\mathcal{S}-j)^{k}}{k!} \tag{6.9}$$

The list of all individual and combined p-values is shown in Table 6.1. In these experiments, *ActPro* ($\alpha = 0.85$) had the most significant results, closely followed by *ActPro* ($\alpha = 0.5$). This suggests that propagating information using diffusion kernel in ActPro enhances its prediction power for tissue-specific pathologies. Furthermore, there are four diseases for which the global interactome had more significant predictions compared to tissue-specific networks, among which *primary biliary cirrhosis* and *psoriasis* had the highest difference. This difference may be attributed to misclassification of disease/ tissue in Himmelstein *et al.* [73], or existence of cross tissue mechanisms of action for the disease.

6.3.7 Tissue-specific Interactome Identifies Novel Disease-related Pathways – Case Study in Neurodegenerative Disorders

I now investigate whether tissue-specific interactomes can help in predicting novel pathways that are involved in the progression of neurodegenerative disorders. I perform a case study of *Alzheimer's* and *Parkinson's* diseases, both of which were among disorders with high compactness in brain tissue. I use Prize-Collecting Steiner Tree (PCST) algorithm to identify the underlying pathway among disease-genes identified by GWAS studies. Formally, PCST problem can be formulated as:

$$\underset{<v,e>\in T}{\textbf{argmin}} \Big\{ \sum_{e} c_e - \lambda \sum_{v} b_v \Big\}, \tag{6.10}$$

Table 6.1: Compactness of tissue specific disease genes in their tissue-specific interactome

	global	**ActPro_0.15**	**ActPro_0.50**	**ActPro_0.85**	**ERW**	**NR**
Alzheimer's disease	**4.12E-3**	6.96E-3	5.98E-3	5.44E-3	5.32E-3	9.60E-2
breast carcinoma	1.83E-3	1.11E-3	8.40E-4	**8.30E-4**	4.09E-3	8.15E-2
chronic lymphocytic leukemia	8.20E-4	7.40E-4	**4.80E-4**	5.10E-4	8.50E-4	2.94E-2
coronary artery disease	3.95E-1	1.58E-1	1.09E-1	1.03E-1	1.33E-1	**1.93E-2**
Crohn's disease	2.56E-2	1.93E-2	1.50E-2	**1.44E-2**	8.54E-2	4.14E-1
metabolic syndrome X	1.11E-2	1.09E-2	**1.07E-2**	1.12E-2	1.02E-1	7.39E-1
Parkinson's disease	1.59E-2	1.25E-2	9.89E-3	**9.50E-3**	1.34E-2	9.62E-2
primary biliary cirrhosis	**7.20E-4**	1.32E-3	3.16E-3	3.40E-3	2.80E-2	6.86E-1
psoriasis	**2.10E-4**	1.10E-3	1.16E-3	9.50E-4	4.67E-3	3.24E-1
rheumatoid arthritis	1.70E-2	**9.28E-3**	1.06E-2	1.10E-2	6.39E-2	3.61E-1
systemic lupus erythematosus	4.98E-2	1.19E-2	7.56E-3	7.22E-3	2.55E-3	**1.60E-4**
type 1 diabetes mellitus	2.64E-2	3.01E-2	**2.38E-2**	2.40E-2	2.64E-1	9.39E-1
type 2 diabetes mellitus	1.57E-3	2.90E-4	2.40E-4	**1.80E-4**	5.60E-4	7.90E-3
vitiligo	**1.17E-3**	2.13E-3	3.04E-3	3.54E-3	1.84E-2	5.69E-1
schizophrenia	3.47E-1	2.13E-1	1.93E-1	1.84E-1	1.40E-1	**4.10E-2**
combined	1.53E-13	1.24E-17	6.62E-19	**3.70E-19**	9.03E-14	2.43E-03

where T is an induced tree of the given graph, v and e are the set of vertices and edges in T, respectively, c_e is the cost of choosing edge e, and b_v is the reward/prize of collecting node v. Similar methods have been proposed previously to connect upstream signaling elements to downstream transcriptional effector genes [190, 191].

To identify disease-related pathways, I first prune non-specific interactions in the network by removing vertices that have more than 500 interactions. I transform edge confidence values (conductances) to edge penalties (resistances) by inverting each edge weight. Node prizes are defined as the ratio of their incident edges that fall within disease-related genes to the total degree of a node. I assigned a node prize of 1,000 to disease genes to ensure that they are selected as terminal nodes. Finally, I use a recent message passing algorithm [8] to identify PCST rooted at each disease-related gene and choose the best tree as the backbone of the disease-related pathway. Over

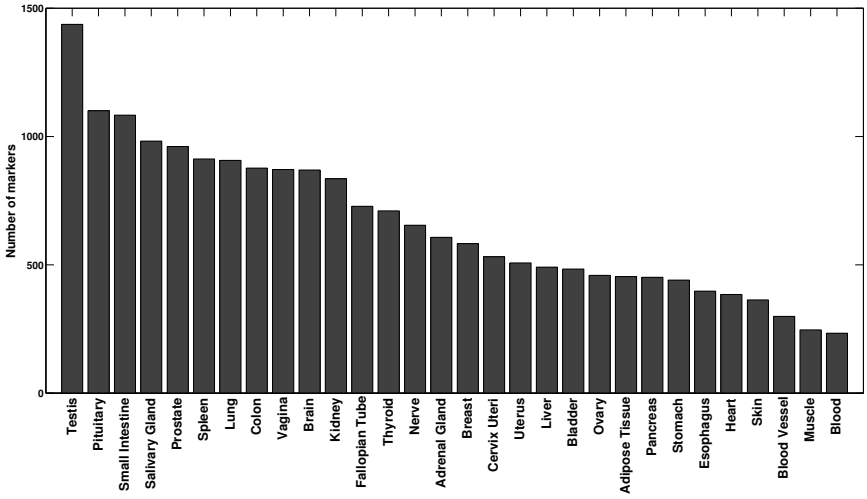
each node, I use a maximum depth of 4 and $\lambda = 1$ as parameters to the message passing algorithm. Figure 6.11 shows final tissue-specific pathways for Alzheimer's and Parkinson's diseases.

Alzheimer's disease (AD) network contains two distinct subnetworks, one centered around *CLTC* and the other centered around *ABL1*. *PICALM, CLU, APOE*, and *SORL1* are all known genes involved in AD, which are also involved *negative regulation of amyloid precursor protein catabolic process*. All four of these genes converge on *CLCT* gene, but through different paths. *PICALM* gene is known to play a central role in clathrin-related endocytosis. This protein directly binds to CLTC and recruits clathrin and adaptor protein 2 (AP-2) to the plasma membrane [21]. On the other hand, *CLU, APOE*, and *SORL1* are linked to the *CLTC* through novel linker genes *XRCC6, MAPT/BIN1* and *GG2A/HGS*, respectively. Gamma-adaptin gene, GGA2, binds to clathrins and regulates protein traffic between the Golgi network and the lysosome. This network is postulated to be an important player in AD [21]. *HGS* gene is a risk factor age-related macular degeneration (AMD) and has been hypothesized to be a shared factor for AD [111]. Interestingly, *MAPT*, a novel marker identified in this study, is a risk factor for Parkinson's disease and very recently shown to also be linked to AD [38]. A second component in AD network is centered around *ABL1* gene, which, together with *CBL, INPPL1, CD2AP*, and *MAPT* share the *SH3 domain binding* function. *INPPL1* gene, a metabolic syndrome risk factor, has been hypothesized to link AD with the recently posed term *"type 3 diabetes"* [3]. Finally, we note that *MAPT* gene is one of the central genes that links these two main components, the role of which warrants further investigation.

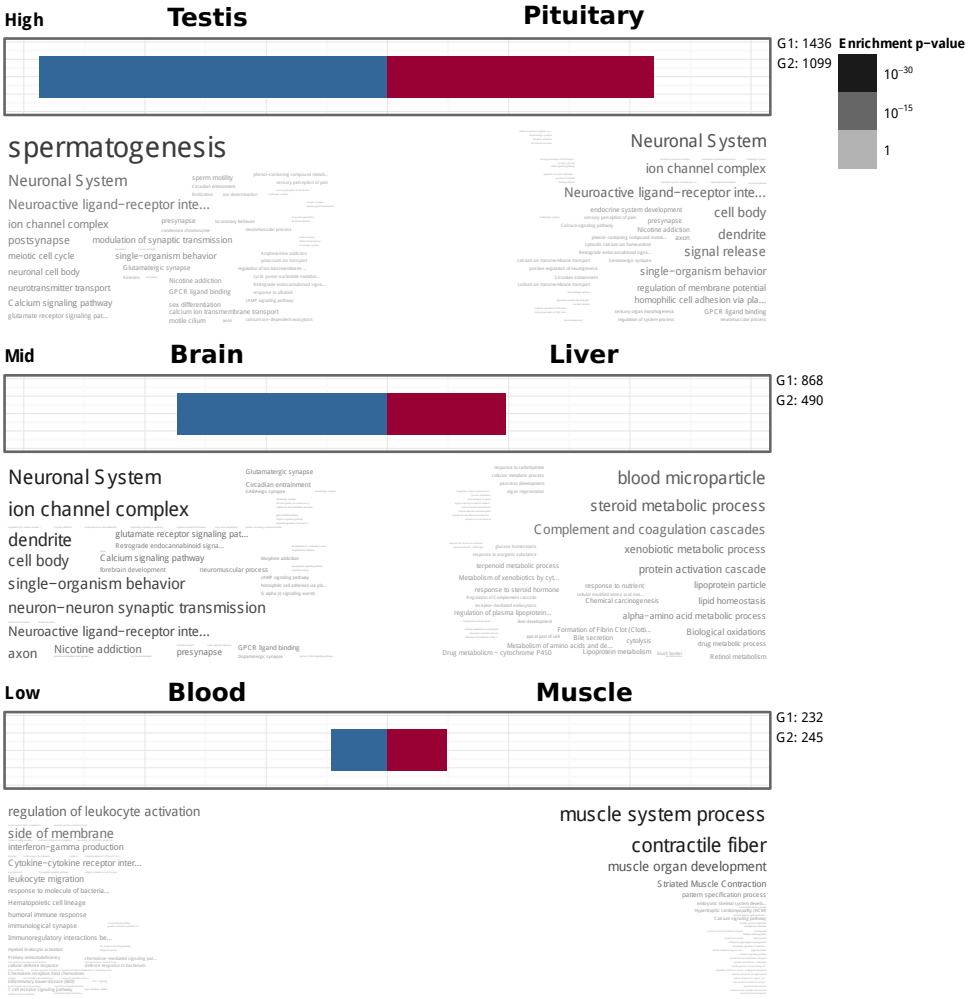
Parkinson's disease (PD) network, on the other hand, contains one densely connected core centered around *MAPT* gene. There are two main branches converging on *MAPT*. On the left, *WNT3, FZD1*, and *GSK3B* constitute upstream elements of the WNT signaling pathway, which is known to play an important role in PD neurodegeneration [12]. *GSK3* gene product is postulated to directly interact with *MAPT* (τ) and *LRKK2*, while implicitly regulating *SNCA* (α-Syn) in a β-cat depen-

dent manner. However, we observed direct interaction between *GSK3B* and **SNCA**, and parallel pathways connecting it to *LRRK2* via *SNCA* and *MAPT*. Both *SNCA* and *MAPT* also take part in the right branch, together with *CAV1* and *RHOA*, which is enriched in *reactive oxygen species metabolic process*. Accumulation of ROS contributes to mitochondrial dysfunction and protein misfolding, both of which are linked to progression of PD. *RIT2* enzyme is identified independently and confirmed as PD susceptibility factor [133]. Pankratz *et al.* also suggested *CALM1* as the bridge linking *RIT2* with *MAPT* and *SNCA*, which confirms my findings. Cyclin G associated kinase (*GAK*) is a known risk factor for PD. I identified *HSPA8* as a key link between *GAK*, WNT signaling pathway, and *CSNK1E* with central PD genes, *MAPT, SNCA,* and *LRRK2*. *HSPA8* gene has been proposed as a biomarker for diagnosis of PD [105]. Finally, myelin basic protein (*MBP*) interacts closely with *CALM1* and *LRRK2*. This gene has been previously shown to be differentially expressed in PD and proposed as potential biomarker for PD [90].

In summary, I show that the brain-specific interactome derived from my method helps in uncovering tissue-specific pathways that are involved in neurodegenerative diseases. Similar analysis of other human tissues can potentially contribute to identification of new therapeutic targets for other human disorders.



(a) Number of selected markers per tissue/cell type



(b) GO Enrichment of tissue-specific markers

Figure 6.3.: Evaluation of tissue-specific markers using a threshold value of 0.75 to define expressed genes.

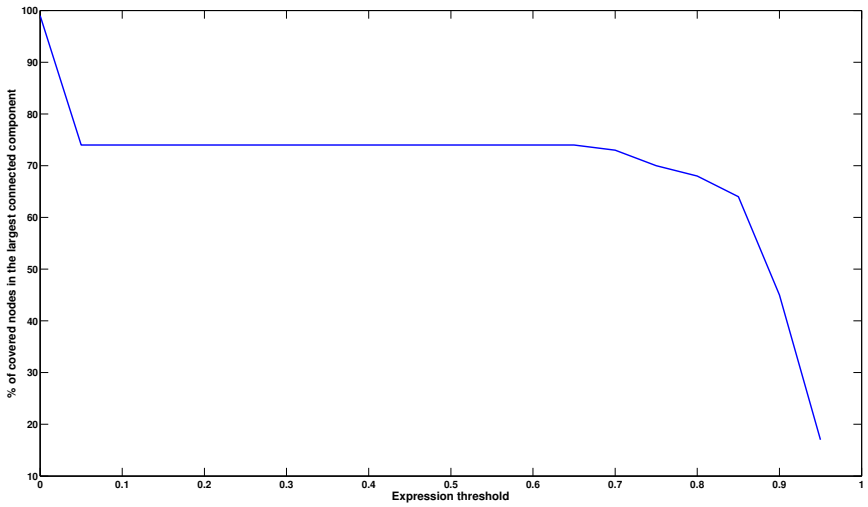
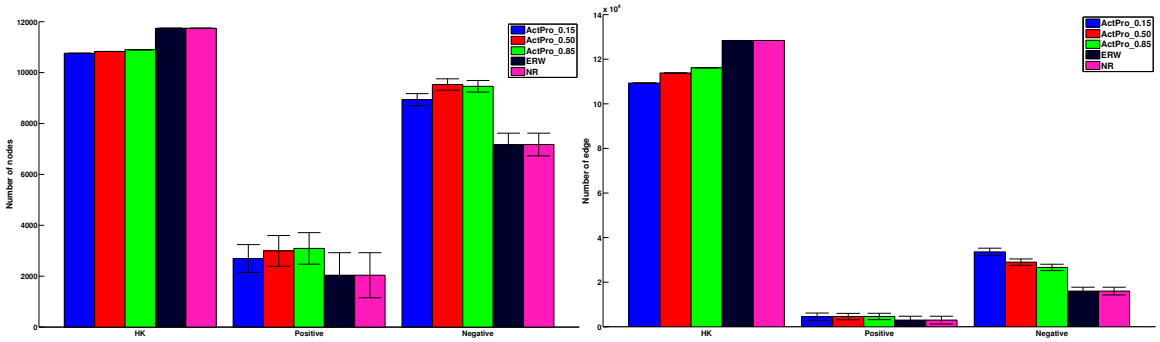


Figure 6.4.: Size of the largest connected component in node removal (NR) method as a function of expression threshold. A rapid disaggregation phase can be spotted around 0.75



(a) Number of nodes (b) Number of edges

Figure 6.5.: Qualitative characteristics of tissue-specific interactomes constructed using different methods.

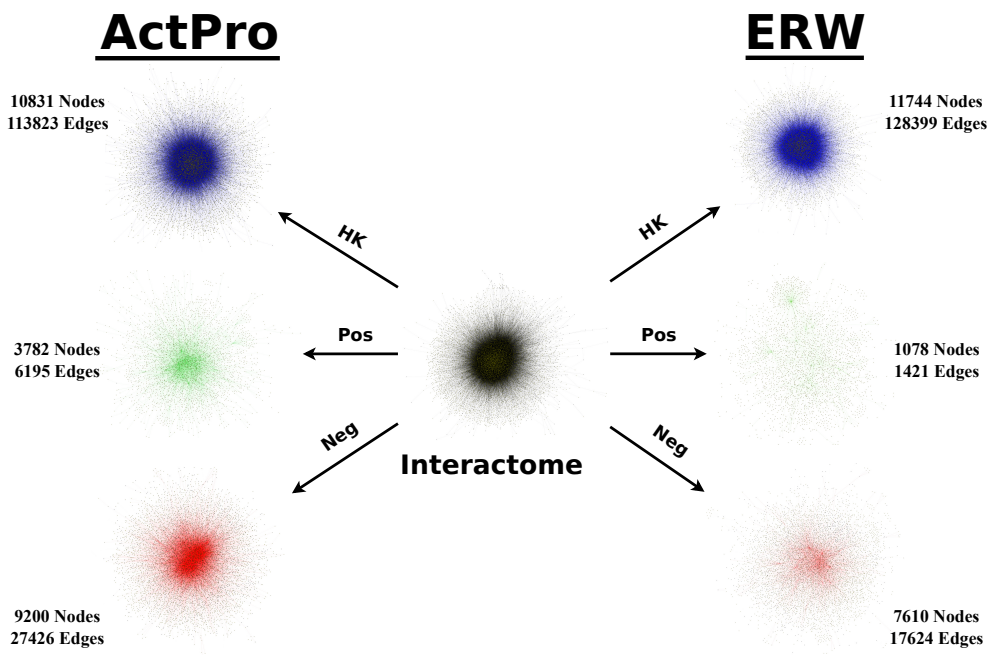


Figure 6.6.: Decomposition of global interactome into brain-specific network using ERW and ActPro ($\alpha = 0.5$) methods

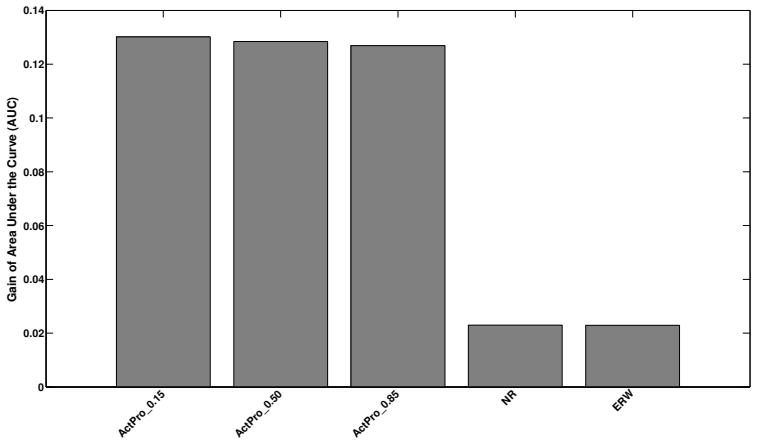


Figure 6.7.: Gain of Area Under the Curve (AUC) of known context-specific pathway edges among tissue-specific interactions

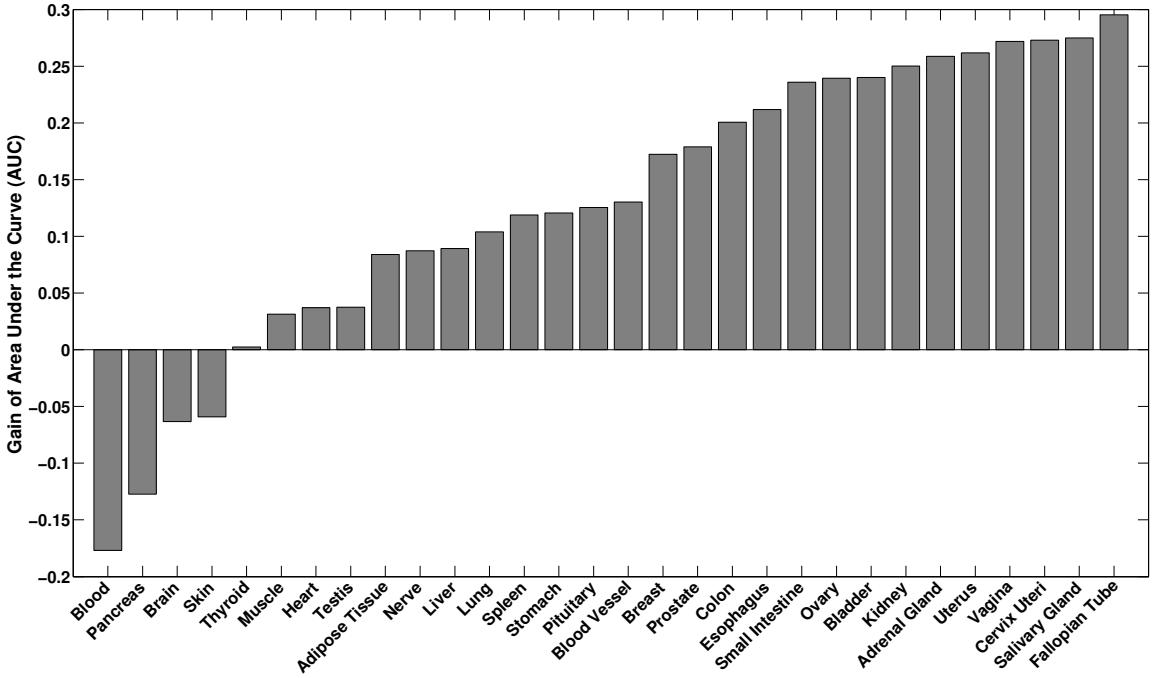
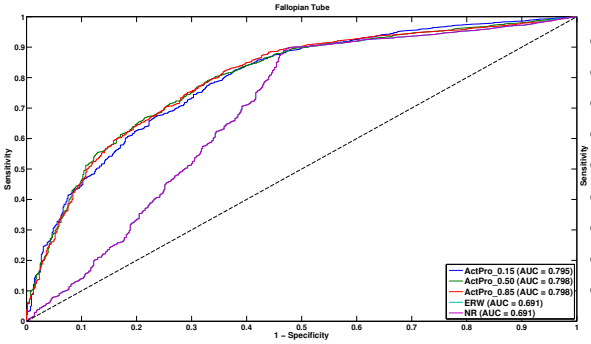
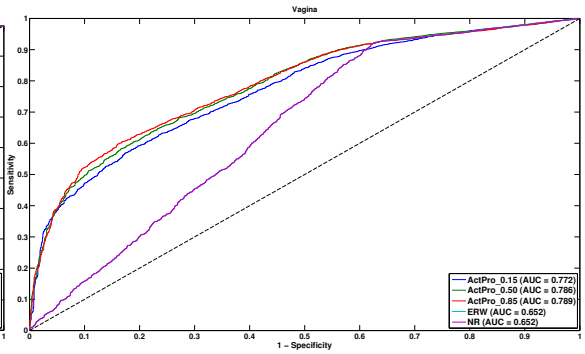


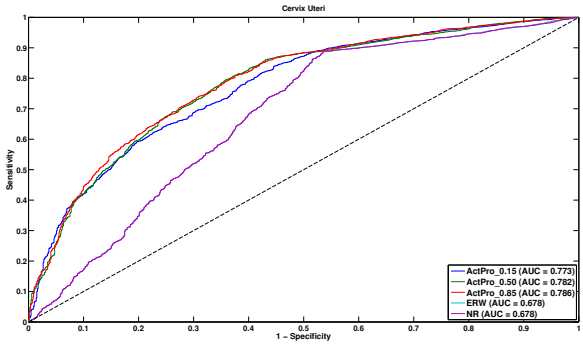
Figure 6.8.: Performance of ActPro with $\alpha = 0.15$ over different tissues



(a) Fallopian Tube



(b) Vagina



(c) Cervix Uteri

Figure 6.9.: Tissues with the highest gain of AUC for predicting tissue-specific pathway edges

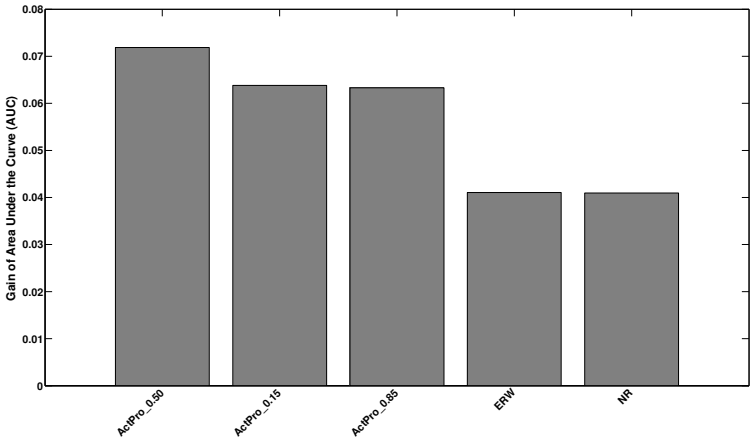
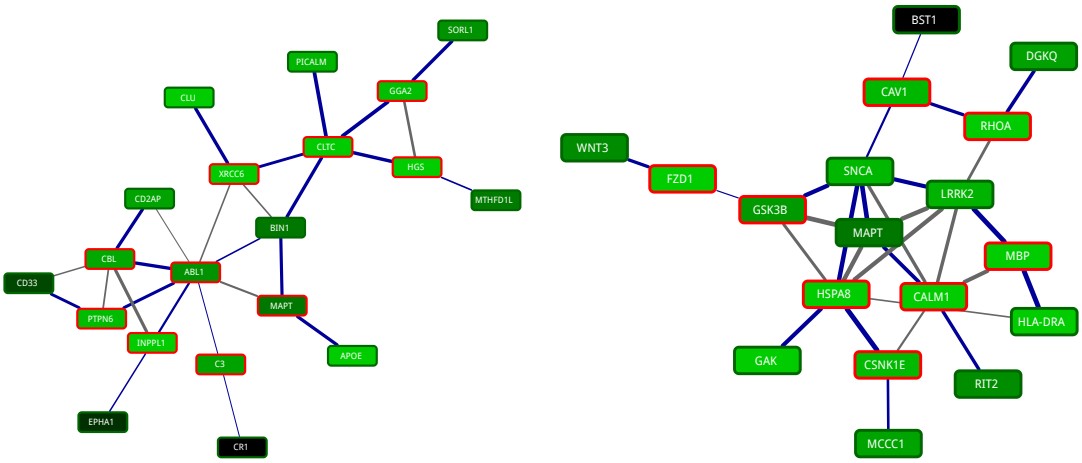


Figure 6.10.: Mean gain of Area under the curve (AUC) for predicting proteins co-annotated with tissue-specific functions



(a) Alzheimer's Disease (b) Parkinson's Disease

Figure 6.11.: Tissue-specific pathways in human neurodegenerative disorders. Nodes are colored according to their tissue-specific expressions, with novel identified genes marked in red, accordingly. The thickness of edges represent their confidence with tree edges marked as blue.

7 CONSERVATION OF CELL TYPE-SPECIFIC NETWORKS ACROSS SPECIES

7.1 Background

Budding yeast, *S. cerevisiae*, is widely used as an experimental system, due to its ease of manipulation in both haploid and diploid forms, and rapid growth compared to animal models. Coupled with the continuous development of new experimental methodologies for manipulating various aspects of its cellular machinery, it has served as the primary model organism for molecular and systems biology [17]. Motivated by the availability of its full genome in 1996 as the first eukaryotic organism to be sequenced [59], an array of functional genomics tools emerged, including a comprehensive collection of yeast deletion mutants [56, 211], genome-wide over-expression libraries [84], and green fluorescent protein (GFP)-tagged yeast strains [55, 76]. The maturity of yeast's genetic and molecular toolbox has, in turn, positioned it as the primary platform for development of many high-throughput technologies, including transcriptome [29, 37, 104], proteome [224], and metabolome [81, 202] screens. These *-omic* datasets, all originally developed in yeast, aim to capture dynamic snapshots of the state of biomolecules during cellular activities. With the advent of "systems modeling", a diverse set of methods have been devised to assay the interactions, both physical and functional, among different active entities in the cell, including protein-protein [78, 98, 192], protein-DNA [79, 106], and genetic [34, 184, 185] interactions. These interactions, also referred to as the *interactome*, embody a complex network of functional pathways that closely work together to modulate the cellular machinery. Comparative analysis of these pathways relies on network alignment methods, much the same way as sequence matching and alignments are used for individual genes and proteins. Network alignments use both the homology of genes, as

well as their underlying interactions, to project functional pathways across different species [97, 99, 163, 167]. These methods have been previously applied to detection of ortholog proteins, projection of functional pathways, and construction of phylogenetic trees.

Yeast and humans share a significant fraction of their functional pathways that control key aspects of eukaryotic cell biology, including the cell cycle [71], metabolism [142], programmed cell death [20, 127], protein folding, quality control and degradation [18], vesicular transport [15], and many key signaling pathways, such as mitogen-activated protein kinase (MAPK) [26, 209], target of rapamycin (TOR) [36], and insulin/IGF-I [9] signaling pathways. In the majority of cases, yeast has been the model organism in which these pathways were originally identified and studied. These conserved biochemical pathways drive cellular growth, division, trafficking, stress-response, and secretion, among others, all of which are known to be associated with various human pathologies. This explains the significant role for yeast as a model organism for human disorders [140, 141, 168]. Yeast has contributed to my understanding of cancers [66, 138, 139] and neurodegenerative disorders [89, 137, 182]. Having both chronological aging (amount of time cells survive in post-mitotic state) and replicative aging (number of times a cell can divide before senescence occurs), yeast is also used extensively as a model organism in aging research. It has contributed to the identification of, arguably, more human aging genes than any other model organism [112].

Depending on the conservation of the underlying pathways, there are two main approaches to studying them in yeast. It has been estimated that, out of 2,271 known disease-associated genes, 526 genes ($\sim 23\%$) have a close ortholog in the yeast genome, spanning 1 out of every 10 yeast genes [50]. For these orthologous pairs of disease-associated genes, I can directly increase the gene dosage of the endogenous yeast protein by using overexpression plasmids, or decrease it, through either gene knockout or knockdown experiments, in order to study gain- or loss-of-function phenotypes, respectively. A key challenge in phenotypic screens is that disrupting genes,

even when they have close molecular functions, can result in characteristically different organism-level phenotypes. *Phenologs*, defined as phenotypes that are related by the orthology of their associated genes, have been proposed to address this specific problem [119]. A recent example of such an approach is the successful identification of a highly conserved regulatory complex implicated in human leukemia [178]. This complex, named COMPASS (Complex of Proteins Associated with Set1), was originally identified by studying protein interactions of the yeast Set1 protein, which is the ortholog of the human mixed-lineage leukemia (MLL) gene, and years later was shown to be conserved from yeast to fruit flies to humans. On the other hand, if the disease-associated gene(s) in humans does not have close orthologs in yeast, heterologous expression of the human disease-gene in yeast, also referred to as *"humanized yeast"*, can be used to uncover conserved protein interactions and their context, to shed light on the molecular mechanisms of disease development and progression. For the majority of disease-genes with known yeast orthologs, heterologous expression of the mammalian gene is functional in yeast and can compensate for the loss-of-function phenotype in yeast deletion strains [17]. This approach has already been used to construct humanized yeast model cells to study cancers [66], apoptosis-related diseases [31], mitochondrial disorders [147], and neurodegenerative diseases [137]. Perhaps one of the more encouraging examples is the very recent discovery of a new compound, N-aryl benzimidazole (NAB), which strongly protects cells from α-synuclein toxicity in the humanized yeast model of Parkinson's disease [180]. In a follow-up study, they tested an analog of the NAB compound in the induced pluripotent stem (iPS) cells generated from the neuron samples of Parkinson's patients with α-synuclein mutations. They observed that the same compound can reverse the toxic effects of α-synuclein aggregation in neuron cells [30]. Using this combined phenotypic screening, instead of the traditional target-based approach, they were not only able to discover a key compound targeting similar conserved pathways in yeast and humans, but also uncover the molecular network that alleviates the toxic effects of α-synuclein. These humanized yeast models have also been used to study human genetic variations [41].

Various successful instances of target identification, drug discovery, and disease network reconstruction using humanized yeast models have established its role as a model system for studying human disorders. When coupled with more physiologically relevant model organisms to cross-validate predictions, yeast can provide a simple yet powerful first-line tool for large-scale genetic and chemical screening [137, 139]. However, as a unicellular model organism, yeast fails to capture organism-level phenotypes that emerge from inter-cellular interactions. Perhaps, more importantly, it is unclear how effectively it can capture tissue-specific elements that make a tissue uniquely susceptible to disease. All human tissues inherit the same genetic code, but they exhibit unique functional and anatomical characteristics. Similar sets of molecular perturbations can cause different tissue-specific pathologies given the network context in which the perturbation takes place. For example, disruption of energy metabolism can contribute to the development of neurodegenerative disorders, such as Alzheimer's, in the nervous system, while causing cardiomyopathies in muscle tissues [13]. These context-dependent phenotypes are driven by genes that are specifically or preferentially expressed in one or a set of biologically relevant tissue types, also known as *tissue-specific* and *tissue-selective* genes, respectively. Disease genes, and their corresponding protein complexes, have significant tendencies to selectively express in tissues where defects cause pathology [60, 103]. How tissue-selective pathways drive tissue-specific physiology and pathophysiology is not completely understood; neither is the extent to which I can use yeast as an effective model organism to study these pathways.

I propose a quantitative framework to assess the scope and limitations of yeast as a model organism for studying human tissue-specific pathways. This framework is grounded in a novel statistical model for effectively assessing the similarity of each tissue with yeast, considering both expressed genes and their underlying physical interactions as a part of functional pathways. To understand the organization of human tissues, I present a computational approach for partitioning the functional space of human proteins and their interactions based on their conservation both across species

and among different tissues. Using this methodology, I identify a set of *core genes*, defined as the subset of the most conserved housekeeping genes between humans and yeast. These core genes are not only responsible for many of the fundamental cellular processes, including translation, protein targeting, ribosome biogenesis, and mRNA degradation, but also show significant enrichment in terms of viral infectious pathways. On the other hand, human-specific housekeeping genes are primarily involved in cell-to-cell communication and anatomical structure development, with the exception of mitochondrial complex I, which is also human-specific. Next, I identify comprehensive sets of tissue-selective functions that contribute the most to the computed overall similarity of each tissue with yeast. These conserved, tissue-selective pathways provide a comprehensive catalog for which yeast can be used as an effective model organism. Conversely, human-specific, tissue-selective genes show the highest correlation with tissue-specific pathologies and their functional enrichment resembles highly specific pathways that drive normal physiology of tissues.

Comparative analysis of yeast and human tissues to construct conserved and non-conserved functional tissue-specific networks can be used to elucidate molecular/ functional mechanisms underlying dysfunction. Moreover, it sheds light on the suitability of the yeast model for the specific tissue/ pathology. In cases where suitability of yeast can be established, through conservation of tissue-specific pathways in yeast, it can serve as an experimental model for further investigations of new biomarkers, as well as pharmacological and genetic interventions.

7.2 Materials and Methods

7.2.1 Datasets

Protein-protein Interaction (PPI) Networks

I adopted human tissue-specific networks from *Bossi et al.* [16]. They integrated protein-protein interactions from 21 different databases to create the whole human

interactome consisting of 80,922 interactions among 10,229 proteins. Then, they extracted the set of expressed genes in each tissue from GNF Gene Atlas and used it to construct the tissue-specific networks, defined as the vertex-induced subgraphs of the entire interactome with respect to the nodes corresponding to the expressed genes in each tissue.

Additionally, I obtained the yeast interactome from the BioGRID [175] database, update 2011 [174], version 3.1.94, by extracting all physical interactions, excluding interspecies and self interactions. This resulted in a total of 130,483 (76,282 non-redundant) physical interactions among 5,799 functional elements in yeast (both RNA and protein). Next, I downloaded the list of annotated CDS entries from the Saccharomyces Genome Database (SGD) [28] and restricted interactions to the set of pairs where both endpoints represent a protein-coding sequence, i.e., protein-protein interactions. The final network consists of 71,905 interactions between 5,326 proteins in yeast.

Protein Sequence Similarities Between Yeast and Humans

I downloaded the protein sequences for yeast and humans in FASTA format from Ensembl database, release 69, on Oct 2012. These datasets are based on the GRCh37 and EF4 reference genomes, each of which contain 101,075 and 6,692 protein sequences for *H. Sapiens* and *S. Cerevisiae*, respectively. Each human gene in this dataset has, on average, 4.49 gene products (proteins). I identified and masked low-complexity regions in protein sequences using *pseg* program [212]. The *ssearch36* tool, from *FASTA* [135] version 36, was then used to compute the local sequence alignment of the protein pairs using the Smith-Waterman algorithm [169]. I used this tool with the BLOSUM50 scoring matrix to compute sequence similarity of protein pairs in humans and yeast. All sequences with E-values less than or equal to 10 are recorded as possible matches, which results in a total of 664,769 hits between yeast and human proteins. For genes with multiple protein isoforms, coming from alternatively spliced

variants of the same gene, I only record the most significant hit. The final dataset contains 162,981 pairs of similar protein-coding genes.

7.2.2 Sparse Network Alignment Using Belief Propagation

Analogous to the sequence alignment problem, which aims to discover conserved genomic regions across different species, network alignment is motivated by the need for extracting shared functional pathways that govern cellular machinery in different organisms. The network alignment problem in its abstract form can be formulated as an optimization problem with the goal of identifying an optimal mapping between the nodes of the input networks, which maximizes both sequence similarity of aligned proteins and conservation of their underlying interactions. At the core of every alignment method are two key components: i) a scoring function and ii) an efficient search strategy to find the optimal alignment. The scoring function is usually designed to favor the alignment of similar nodes, while simultaneously accounting for the number of conserved interactions between the pair of aligned nodes. Biologically speaking, this translates to identifying functional *orthologs* and *interologs*, respectively.

Given a pair of biological networks, $\boldsymbol{G} = (\mathcal{V}_{\boldsymbol{G}}, \mathcal{E}_{\boldsymbol{G}})$ and $\boldsymbol{H} = (\mathcal{V}_{\boldsymbol{H}}, \mathcal{E}_{\boldsymbol{H}})$, with $n_{\boldsymbol{G}} = |V_G|$ and $n_{\boldsymbol{H}} = |V_H|$ vertices, respectively, we can represent the similarity of vertex pairs between these two networks using a weighted bipartite graph $\boldsymbol{L} = (\mathcal{V}_{\boldsymbol{G}} * \mathcal{V}_{\boldsymbol{H}}, \mathcal{E}_{\boldsymbol{L}}, \boldsymbol{w})$, where $\boldsymbol{w} : \mathcal{E}_{\boldsymbol{L}} \rightarrow \mathcal{R}$ is a weight function defined over edges of $\boldsymbol{L}$. I will denote mapping between vertices $v_i \in \mathcal{V}_{\boldsymbol{G}}$ and $v_{i'} \in \mathcal{V}_{\boldsymbol{H}}$ with (i, i') and ii', interchangeably. Let us encode the edge conservations using matrix $\mathbf{S}$, where $\mathbf{S}(ii', jj') = 1$, iff alignment of $v_i \rightarrow v_{i'}$ together with $v_j \rightarrow v_{j'}$ will conserve an edge between graphs $\boldsymbol{G}$ and $\boldsymbol{H}$, and $\mathbf{S}(ii', jj') = 0$, otherwise. Then, the network

alignment problem can be formally represented using the following integer quadratic program:

$$\max_{\boldsymbol{x}} \quad (\alpha \boldsymbol{w}^T \boldsymbol{x} + \frac{\beta}{2} \boldsymbol{x}^{\mathbf{T}} \mathbf{S} \boldsymbol{x}) \tag{7.1}$$

$$\text{Subject to:} \quad \begin{cases} \mathbf{C}\boldsymbol{x} \leq \mathbf{1}_{n_{\boldsymbol{G}} * n_{\boldsymbol{H}}} & \text{Matching constraints;} \\ x_{ii'} \in \{0, 1\}, & \text{Integer constraint.} \end{cases}$$

Here, $\mathbf{C}$ and $\boldsymbol{w}$ are the incidence matrix and edge weights of the graph $\boldsymbol{L}$, respectively, whereas $\boldsymbol{x}$ is the matching indicator vector. Vector $\boldsymbol{w}$, which encodes the *prior* knowledge of node-to-node similarity between the input pair of networks, defines the search space of *potential orthologs* and can be computed using sequence, structural, or functional similarity of the proteins corresponding to node pairs. In this study, I chose sequence similarity of aligned protein sequences to assign edge weights in the bipartite graph defined by $\boldsymbol{L}$. When $\boldsymbol{L}$ is a complete bipartite graph, i.e. each pair of vertices between $\boldsymbol{G}$ and $\boldsymbol{H}$ represents a viable ortholog candidate, I will have $\mathbf{S} = \boldsymbol{G} \otimes \boldsymbol{H}$. However, *Bayati et al.* [10] recently proposed an efficient method, based on the message passing algorithm, for cases where $\boldsymbol{L}$ is sparse, i.e., $|\mathcal{E}_{\boldsymbol{L}}| << n_{\boldsymbol{G}} * n_{\boldsymbol{H}}$, by restricting the search space to the subset of promising candidates that are provided by $\mathcal{E}_{\boldsymbol{L}}$. I will use this algorithm throughout this chapter for solving the network alignment problem.

7.2.3 Tissue-specific Random Model (TRAM) for Generating Pseudo-random Tissues

Let us denote the global human interactome by $\boldsymbol{G} = (\mathcal{V}_{\boldsymbol{G}}, \mathcal{E}_{\boldsymbol{G}})$, and each tissue-specific network by $\boldsymbol{T} = (\mathcal{V}_{\boldsymbol{T}}, \mathcal{E}_{\boldsymbol{T}})$, respectively. Using this notation, we have $n_{\boldsymbol{T}} = |\mathcal{V}_{\boldsymbol{T}}|$, $\mathcal{V}_{\boldsymbol{T}} \subset \mathcal{V}_{\boldsymbol{G}}$, and $\mathcal{E}_{\boldsymbol{T}} \subset \mathcal{E}_{\boldsymbol{G}}$ is the subset of all edges from $\boldsymbol{G}$ that connect vertices in $\mathcal{V}_{\boldsymbol{T}}$, i.e., $\boldsymbol{T}$ is the vertex-induced subgraph of $\boldsymbol{G}$ under $\mathcal{V}_{\boldsymbol{T}}$. This is the formal description of the model used by Bossi *et al.* [16] to construct human tissue-specific networks. Using this construction model, we note that every tissue-specific network inherits a shared core of interactions among housekeeping genes that are universally

expressed to maintain basic cellular functions. Let us denote this subset of genes by $\mathcal{V}_{\boldsymbol{U}} \subset \mathcal{V}_{\boldsymbol{T}}$, having $n_{\boldsymbol{U}} = |\mathcal{V}_{\boldsymbol{U}}|$ members, and the corresponding induced core sub-graph using $\boldsymbol{U} = (\mathcal{V}_{\boldsymbol{U}}, \mathcal{E}_{\boldsymbol{U}})$.

In this setting, I propose a new random model to explicitly mimic the topology of tissue-specific networks. Formally, given each human tissue-specific network, I seed an ensemble of *pseudo-random tissues* denoted by $\mathbb{R}_{\boldsymbol{T}} = \boldsymbol{G}(\mathcal{V}_{\mathbb{R}}, \mathcal{E}_{\mathbb{R}})$, in which every instance shares two main characteristics from the original network: (i) the total number of vertices, (ii) the shared core of housekeeping interactions. To summarize, this random graph sampling scheme is as follows: first, I initialize the vertex set $\mathcal{V}_{\mathbb{R}}$ using $\mathcal{V}_{\boldsymbol{U}}$, which includes $n_{\boldsymbol{U}}$ housekeeping genes. Next, to ensure that the newly generated random instance has the same number of vertices as the seed network, we sample $n_{\boldsymbol{T}} - n_{\boldsymbol{U}}$ vertices without replacement from the remaining vertices, $\mathcal{V}_{\boldsymbol{G}} \setminus \mathcal{V}_{\boldsymbol{U}}$. Finally, we construct the random graph as the vertex induced sub-graph of the global human interactome imposed by $\mathcal{V}_{\mathbb{R}}$.

It can be noted that my random model not only provides a pseudo-random network seeded on each tissue-specific network, but also provides a node-to-node similarity score between the newly generated graph and the yeast interactome. This is a critical component of my framework, which distinguishes it from other *random graph models*, such as Erdos-Renyi, network growth, or preferential attachment. The only other effort to combine topology with the node-to-node similarity score is proposed by Sahraeian *et al.* [158], which fits a gamma distribution over the the known pairs of ortholog/ non-orthologs proteins in three species (according to their KEGG pathways), and uses the fitted distribution to sample new sequence similarity scores. However, this model does not benefit from the structural knowledge of the tissue-specific networks. Moreover, its sequence similarity generation model loosely fits the observed data and does not provide a fine-tuned model to assess the significance of tissue-specific alignments. My model, one the other hand, is grounded in the same construction model as the original tissue-specific networks, and provides enough se-

lectivity to distinguish similarity/ dissimilarity of aligned networks with yeast and to assign an empirical p-value to each alignment.

7.2.4 Significance of Network Alignments

For each optimal alignment of a human tissue-specific network with yeast, given by its indicator variable $\boldsymbol{x}$, I quantify the overall sequence similarity of aligned proteins with the matching score of the alignment, $\hat{w} = \boldsymbol{w}^T\boldsymbol{x}$, and the total number of conserved edges by the alignment overlap, $\hat{o} = \frac{1}{2}\boldsymbol{x}^T\mathbf{S}\boldsymbol{x}$. These measures can be used to rank different network alignments. However, without a proper reference to compare with, it is almost impossible to interpret these values in a statistical sense. To address this issue, I sample an ensemble of $k_{\mathbb{R}}$ random networks from the *tissue-specific random model (TRAM)*, independently align each instance to the yeast interactome, and empirically compute a *topological*, a *homological* (sequence-based), and a *mixed* alignment p-value for each alignment using Monte-Carlo simulation.

Let $\hat{w}_{\mathbb{R}}$ and $\hat{o}_{\mathbb{R}}$ be the random vectors representing the weight and overlap of aligning random tissues with yeast, respectively. First, I define individual p-values for the conservation of network topology and sequence homology. Let us denote by $k_P^{(\hat{w})}$ and $k_P^{(\hat{o})}$ the number of random samples that have weight and overlap greater than or equal to the original alignment, respectively. Then, we can define the following p-values:

$$p - val_{\mathbf{homolgy}} = \frac{k_P^{(\hat{w})}}{k_{\mathbb{R}}} \tag{7.2}$$

$$p - val_{\mathbf{topology}} = \frac{k_P^{(\hat{o})}}{k_{\mathbb{R}}} \tag{7.3}$$

Before I define the *mixed* p-value, I define upper and lower bounds on the p-value. These bounds are independent of the mixing parameter. For cases where both $\hat{o} \leq \hat{o}_{\mathbb{R}}(i)$ and $\hat{w} \leq \hat{w}_{\mathbb{R}}(i)$, for $1 \leq i \leq k_{\mathbb{R}}$, I can report that the random alignment is at least as good as the original alignment. Conversely, if both $\hat{o}_{\mathbb{R}}(i) < \hat{o}$ and $\hat{w}_{\mathbb{R}}(i) < \hat{w}$, we can assert that the original alignment outperforms the random alignment. Let us

denote the number of such cases by k_P and k_N, respectively. Using this formulation, I can compute the following bounds on the mixed p-value of the alignment:

$$\delta_{\mathbb{R}} = \frac{k_P}{k_{\mathbb{R}}} \leq \text{alignment p-value} \leq 1 - \frac{k_N}{k_{\mathbb{R}}} = \Delta_{\mathbb{R}} \tag{7.4}$$

Please note that Δ_R and δ_R are not p-values themselves, rather, they represent α-independent bounds on the mixed p-values. I can use these bounds to estimate the similarity of each tissue-specific network to the yeast interactome. Tissues for which the upper-bounds on the alignment p-value are smaller than a given threshold α_u are considered similar to yeast, while tissues with lower-bounds larger than α_l are considered dissimilar. For cases where the following conditions hold: $\hat{o}_{\mathbb{R}}(i) < \hat{o}$ and $\hat{w} < \hat{w}_{\mathbb{R}}(i)$, or $\hat{o} < \hat{o}_{\mathbb{R}}(i)$ and $\hat{w}_{\mathbb{R}}(i) < \hat{w}$, the p-values are α-dependent. To quantify this ambiguity, I define the *reliability* of a p-value as $\frac{k_N+k_P}{k_{\mathbb{R}}}$. When there is no ambiguity, that is, both the homological and topological p-values of each case are either concurrently significant or concurrently insignificant, the reliability score is one. Otherwise, in cases where one of them is significant while the other is not, the reliability score decreases, accordingly. Finally, I define an unadjusted mixed p-value similar to the convex combination used in network alignment. Let us define a new random variable $o\hat{w}_{\mathbb{R}} = \alpha * \hat{o}_{\mathbb{R}} + \beta * \hat{w}_{\mathbb{R}}$. Using this notation, I define the mixed p-value as:

$$p - value = Prob(\alpha * \hat{o} + \beta * \hat{w} \leq o\hat{w}_{\mathbb{R}}) \tag{7.5}$$

7.2.5 Differential Expression of Genes with Respect to a Group of Tissues

Given a homogenous group of human tissues/cell types, I first identify all *expressed genes* in the group, i.e., all non-housekeeping genes that are expressed in *at least* one of the tissue members. Next, in order to identify the subset of expressed genes that are *selectively* expressed, I use a *hypergeometric* random model. A gene is identified as selectively expressed if it is expressed in significantly higher number of tissues in the given group than randomly selected tissue subsets of the same size. Let N and n denote the total number of tissues in this study and the subset of tissues in the given

group, respectively. Moreover, let us represent by c_N the number of all tissues in which a given gene is expressed, whereas c_n similarly represents the number of tissues in the given group that the gene is expressed. Finally, let the random variable X be the number of tissues in which the gene is expressed, if we randomly select subsets of tissues of similar size. Using this formulation, we can define the *tissue-selectivity* p-*value* of each expressed gene in the given group as follows:

$$\begin{aligned} p\text{-value}(X = c_n) &= Prob(c_n \leq X) \\ &= HGT(c_n|N, n, c_N) \\ &= \sum_{x=c_n}^{min(c_N,n)} \frac{C(c_N, x)C(N - c_N, n - x)}{C(N, n)} \end{aligned} \tag{7.6}$$

In order to partition genes into *selective* and *ubiquitous* genesets, I derive the tissue-selectivity p-value distribution of all expressed non-housekeeping genes in the given group. I use the Gaussian kernel to smooth this distribution and then find the critical points of the smoothed density function to threshold for tissue-selective genes. The motivation behind my choice is that these points provide shifts in the underlying distribution, from tissue-selective to ubiquitous genes. Given the bi-modal characteristic of the distribution, it has three expected critical points. I use the first of these points as my cutoff point. This provides highest precision for declared tissue-selective genes, but lower recall than the other two choices.

7.2.6 Conservation of Genesets Based on the Majority Voting Rule

Given a set of genes that are selectively expressed in a homogenous group of tissues/cell types, I am interested in tri-partitioning them into either *conserved*, *human-specific*, or *unclassified* genes. *Conserved genes* are the subset of tissue-selective genes that are consistently aligned in majority of aligned tissues in the given group. Conversely, *human-specific genes* are the subset of tissue-selective genes that are consistently unaligned in majority of tissues in the given group. Finally, *unclassified genes*

are the subset of tissue-selective genes for which we do not have enough evidence to classify them as either conserved or human-specific.

The key data-structure I use to tri-partition genesets is the *alignment consistency table*. Let C be a group of homogenous tissues with $n = |C|$. Furthermore, let $g_C^{\mathbf{TS}}$ represent the set of tissue-selective genes with respect to C, such that $k_C^{\mathbf{TS}} = |g_C^{\mathbf{TS}}|$. The alignment consistency table is a table of size $k_C^{\mathbf{TS}} \times n$, represented by $\mathcal{T}_C^{\mathbf{TS}}$, in which $\mathcal{T}_C^{\mathbf{TS}}(i,j)$ is the aligned yeast partner of ith tissue selective gene under the network alignment of jth tissue in C, or $'-'$ (gap), if it is unaligned. I find the most common partner for each tissue-selective gene and use a *consensus rate*, represented by τ, to summarize each rows of the alignment consistency table. If a gene is consistently aligned to the same yeast partner in at least $\tau * n$ tissues in C, I declare it as conserved. Similarly, if it is unaligned in at least $\tau * n$ tissues in C, I classify it as human-specific. If neither one of these conditions hold, I report it as unclassified.

7.3 Results and Discussion

In this section, I present a comparative framework for investigating the scope and limitations of yeast as a model organism for studying tissue-specific biology in humans. Figure 7.1 illustrates the high-level summary of my study design. We start by aligning each of the human tissue-specific networks with the yeast interactome. I couple the alignment module with a novel statistical model to assess the significance of each alignment and use it to infer the respective similarity/ dissimilarity of human tissue-specific networks with their corresponding counterparts in yeast. Using a network of tissue-tissue similarities computed using their transcriptional profile, I show that my network alignment p-values are consistent with groupings derived from transcriptional signatures. I use this network of tissue similarities to identify four major groups of tissues/ cell-types. These groups; representing brain tissues, blood cells, ganglion tissues, and testis-related tissues; are further used to identify tissue-selective genes that are active within each group compared to the rest of tissues.

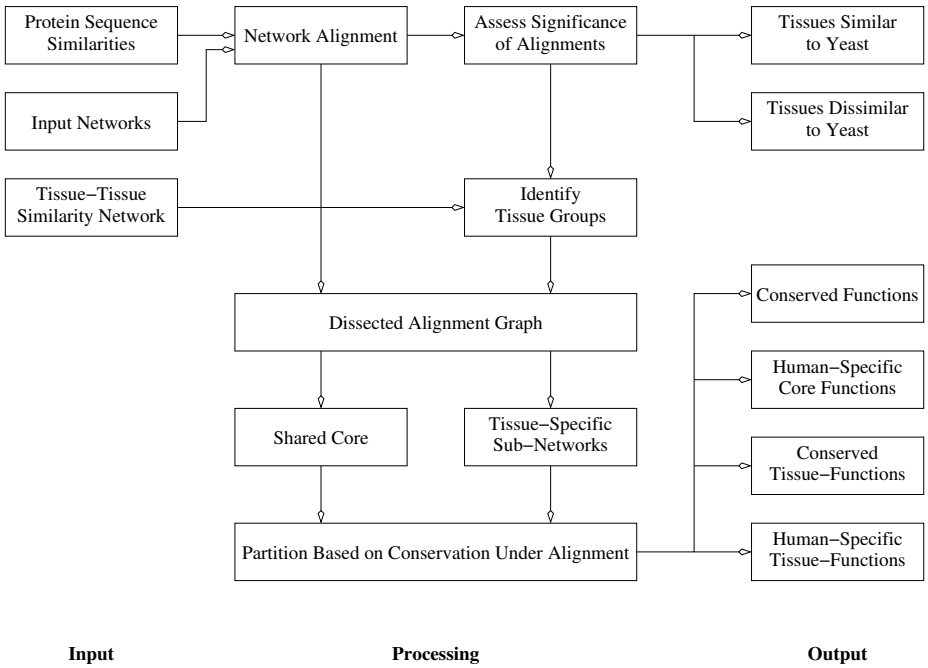


Figure 7.1.: Main components of the analysis framework for comparing yeast with human tissue-specific interactome

I partition both housekeeping and tissue-selective subsets of human genes separately into the conserved and human-specific subsections. I provide extensive validation for the selective genes with respect to blood cells and brain tissues. Figure 7.2 illustrates the overall partitioning of the genes and their relative subsets. I provide an in-depth analysis of each of these subsets, and show that while conserved subsets provide the *safe zone* for which yeast can be used as an ideal model organism, the human-specific subset can shed light on the *shadowed subspace* of the human interactome in yeast. This subset can provide future directions for constructing humanized yeast models.

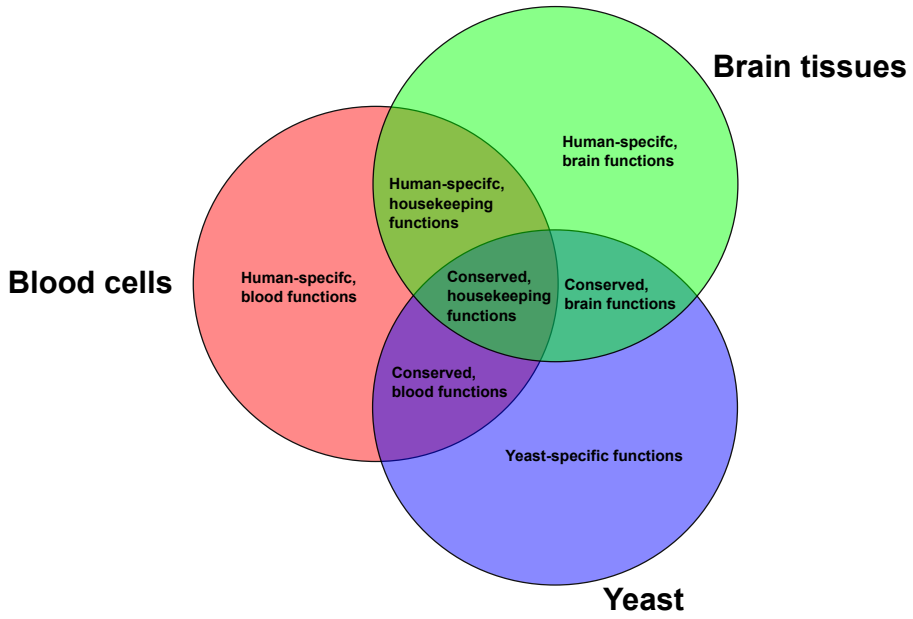


Figure 7.2.: A high-level classification of human genes

7.3.1 Aligning Yeast Interactome with Human Tissue-specific Networks

The *global* human interactome represents a static snapshot of potential physical interactions that *can* occur between pairs of proteins. However, it does not provide any information regarding the spatiotemporal characteristics of the actual protein interactions. These interactions have to be complemented with a dynamic *context*, such as expression measurements, to help interpret cellular rewiring under different conditions.

[16] overlaid the mRNA expression level of each transcript (transcriptome) in different human tissues [177] on top of the *global* human interactome, integrated from 21 PPI databases, and constructed a set of 79 reference tissue-specific networks. I adopt these networks and align each one of them separately to the yeast interactome that I constructed from the BioGRID database.

In order to compare these human tissue-specific networks with the yeast interactome, considering both the sequence similarity of proteins and the topology of their

interactions, I employ a recently proposed sparse network alignment method, based on the Belief Propagation (BP) approach. This method is described in the Materials and methods section [10].

Genes, and their corresponding proteins, do not function in isolation; they form a complex network of interactions among coupled biochemical pathways in order to perform their role(s) in modulating cellular machinery. Moreover, each protein may be involved in multiple pathways to perform a diverse set of functions. Using a network alignment approach to project these pathways across species allows us to not only consider their first-order dynamics, through co-expression of homologous protein pairs, but also the context in which they are expressed.

To construct the state space of potential homologous pairs, we align all protein sequences in human and yeast and pre-filter hits with sequence similarity E-values greater than 10. For genes with multiple protein isoforms I only store the most significant hit. Using these sequence-level homologies, I construct a matrix L that encodes pairwise sequence similarities between yeast and human proteins. Entries in matrix L can be viewed as edge weights for a bipartite graph connecting human genes on one side, and the yeast genes, on the other side. I use this matrix to restrict the search space of the BP network alignment method (please see Supplementary Methods for details on E-value normalization and Materials and Methods section for BP alignment method).

Parameters α and $\beta(= 1 - \alpha)$ control the relative weight of sequence similarity (scaled by α) as compared to topological conservation (scaled by β) in the BP network alignment. Using a set of preliminary simulations aligning the global human interactome with its tissue-specific sub-networks, for which we have the *true* alignment, with various choices of α in the range of 0.1 to 0.9, I identify the choices of $\alpha = \frac{1}{6}$ and $\beta = \frac{5}{6}$ to perform the best in my experiments. I use the same set of parameters to align each tissue-specific network with the yeast interactome, as it provides a balanced contribution from sequence similarities and the number of conserved edges.

7.3.2 Investigating Roles of Housekeeping Genes and their Conservation across Species

Housekeeping genes comprise a subset of human genes that are universally expressed across all tissues and are responsible for maintaining core cellular functions needed by all tissues, including translation, RNA processing, intracellular transport, and energy metabolism [23,39,172]. These genes are under stronger selective pressure, compared to tissue-specific genes, and evolve more slowly [221]. As such, we expect to see a higher level of conservation among human housekeeping genes compared with yeast genes. I refer to the most conserved subset of housekeeping genes between humans and yeast, computed using network alignment of tissues-specific networks with the yeast network, as the *core genes*.

I identify a gene as housekeeping if it is expressed in *all* 79 tissues. I identify a total of 1,540 genes that constitute the shared section of human tissue-specific networks. These genes, while having similar set of interactions among each other, are connected differently to the set of tissue-selective genes.

Using the alignment partners of all housekeeping genes in the yeast interactome, I construct an alignment consistency table of size $1,540 \times 79$, which summarizes the network alignments over the shared subsection of tissue-specific networks. Then, I use the majority voting method to classify housekeeping genes as *core*, which are conserved in yeast, *human-specific*, which are consistently unaligned across human tissues, and *unclassified*, for which we do not have enough evidence to classify it as either one of the former cases.

Network alignments are noisy and contain both false-positive (defined as aligned pairs that are not functionally related), as well as false-negatives (pairs of functional orthologs that are missed in the alignment). These errors can come from different sources, including gene expression data (node errors), interactome (edge errors), or the alignment procedure (mapping errors). I propose a method based on majority voting across different alignments to (partially) account for these errors. Given a set of network alignments, I consider a pair of entities consistently aligned (either

matched or unmatched) if they are consistent in at least $100 * \tau\%$ of alignments in the set. The parameter τ, called the *consensus rate*, determines the level of accepted disagreement among different alignments. A higher value of consensus rate increases the precision of the method at the cost of decreased sensitivity. In order to select the optimal consensus rate parameter, I tried values in range $[0.5 - 1.0]$ with increments of $\frac{1}{2}$. I identified the parameter choice of $\tau = 0.9$, equivalent to 90% agreement among aligned tissues, to perform the best in classifying human-specific and conserved genes, while keeping the sets well-separated. Using this approach, I was able to tri-partition 1,540 housekeeping genes into 595 conserved, 441 human-specific, and 504 unclassified genes, respectively.

In order to investigate the conserved sub-network of core genes, I construct their alignment graph as the Kronecker product of the subgraph induced by core genes in the human interactome and its corresponding aligned subgraph in yeast. Conserved edges in this network correspond to interologs, i.e., orthologous pairs of interacting proteins between yeast and human [217].

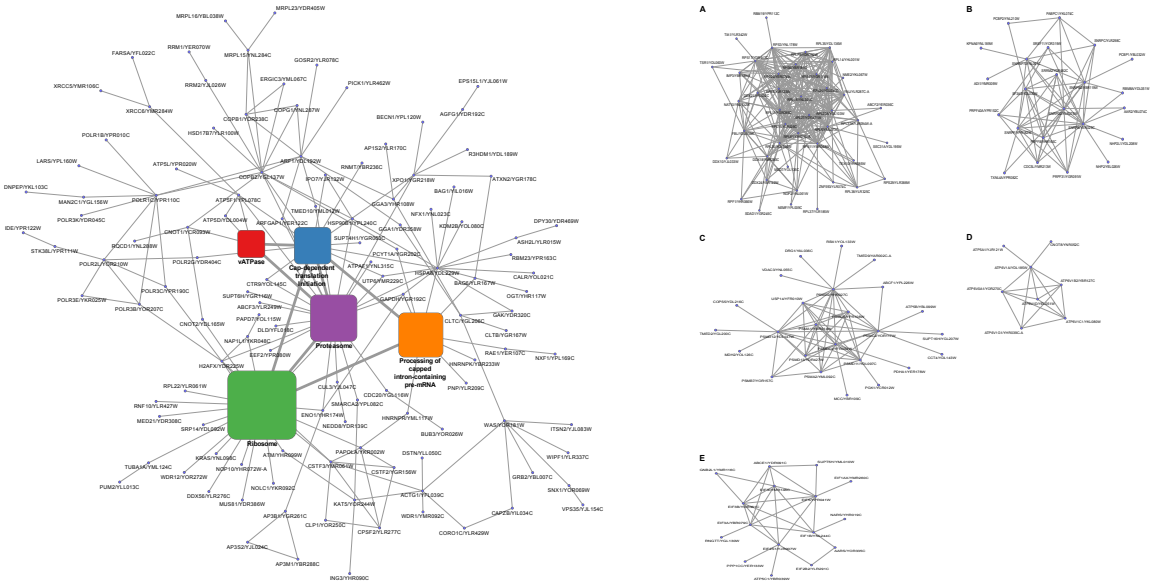


Figure 7.3.: Alignment graph of core human genes

Figure 7.3 shows the largest connected component of this constructed alignment graph. I applied the MCODE [7] network clustering algorithm on this graph to identify highly interconnected regions corresponding to putative protein complexes. I identified five main clusters, which are color-coded on the alignment graph, and are shown separately on the adjacent panels. Ribosome is the largest, central cluster identified in the alignment graph of core genes, and together with proteasome and spliceosome, constitutes the three most conserved complexes in the alignment graph. This complex is heavily interconnected to the eIFs, to modulate eukaryotic translation initiation, as well as proteasome, which controls protein degradation. Collectively, these complexes regulate protein turnover and maintain a balance between synthesis, maturation, and degradation of cellular proteins.

In order to further analyze the functional roles of these housekeeping genes, I use the g:Profiler [153] R package to identify highly over-represented terms. Among functional classes, I focus on the gene ontology (GO) biological processes, excluding electronic annotations, KEGG pathways, and CORUM protein complexes to provide a diverse set of functional roles. I use the Benjamini-Hochberg procedure to control for false-discovery rate (FDR), with p-value threshold of $\alpha = 0.05$, and eliminate all enriched terms with more than 500 genes to prune overly generic terms. Using this procedure, I identify enriched functional terms for both core and human-specific subsets of housekeeping genes.

I manually group the most significant terms (p-value $\leq 10^{-10}$) in core genes, which results in five main functional classes, namely ribosome biogenesis, translation, protein targeting, RNA splicing, and mRNA surveillance. First, we observe a one-to-one mapping between enriched terms and identified putative complexes corresponding to translation initiation (p-value $= 7.1 * 10^{-17}$) and ribosome (p-value $= 5.97 * 10^{-11}$). In addition, translation termination and elongation are also enriched with decreasing levels of significance. Moreover, these processes are tightly linked to SRP-dependent co-translational protein targeting (p-value $= 2.7 * 10^{-15}$). This, in turn, suggests protein synthesis as one of the most conserved aspects of eukaryotic cells. Next, we

note that both mRNA splicing (p-value $= 7.04 * 10^{-10}$) and nonsense-mediated decay (p-value $= 4.66 * 10^{-16}$) are also enriched among the most significant functional terms, which supports my earlier hypothesis related to the role of splicesome in the alignment graph of core genes. Finally, I find that the most significant functional term, as well as a few other related terms, are involved in viral infection, which suggests that (a subset of the) core genes provides a *viral gateway* to mammalian cells. This can be explained in light of two facts: i) viral organisms rely on the host machinery for their key cellular functions, and ii) housekeeping genes are more ancient compared to tissue-selective genes, and core genes provide the most conserved subset of these housekeeping genes. As such, these genes may contain more conserved protein interaction domains and be structurally more "familiar" as interacting partners for the viral proteins and provide ideal candidates for predicting host-pathogen protein interactions.

Next, I perform a similar procedure for the human-specific housekeeping genes. This subset, unlike core genes, is mostly enriched with terms related to anatomical structure development and proximal cell-to-cell communication (paracrine signaling), with the exception of complex I of the electron transport chain, which is the strongest identified term. This NADH-quinone oxidoreductase is the largest of the five enzyme complexes in the respiratory chain of mammalian cells. However, this complex is not present in yeast cells and has been replaced with a single subunit NADH dehydrogenase encoded by gene NDI1. Impairment of complex I has been associated with various human disorders, including Parkinson's and Huntington's disease. Transfecting complex I-defective cells with yeast NDI1 as a therapeutic agent has been proposed as a successful approach to rescue complex I defects [118, 214]. This technique, also known as *NDI1 therapy*, opens up whole new ways in which yeast can contribute to the research and development on human diseases: not only yeast can be used as a model organism, but also can provide candidates that can be used for gene therapy in mammalian cells.

A key observation here is that the human-specific subset of housekeeping genes is not only associated with fewer functional terms, but is also less significantly associated

with these terms. This effect can be attributed to two factors. First, we note that some of the genes predicted to be human-specific might be an artifact of the method. For example, the belief propagation (BP) method enforces sequence similarity as a necessary, but not sufficient, condition for a pair of genes to be aligned, which means that any human gene with no sequence similarity to yeast genes will not be aligned, resulting in genes being artificially classified as human-specific. Second, and more importantly, a majority of functional annotations for human genes are initially attributed in other species, specially yeast, and transferred across ortholog groups. Based on my construction, human-specific genes are defined as the subset of housekeeping genes with no orthology with yeast. As such, it can be expected that these genes span the *shadowed subspace* of the functional space of human genes that is under-annotated.

7.3.3 Quantifying Similarity of Human Tissues with Yeast

Housekeeping genes are shared across all human tissues and cell types. They provide a conserved set of functions that are fundamental to cellular homeostasis. However, these genes do not provide direct insight into how different tissues utilize these key functions to exhibit their dynamic, tissue-specific characteristics. To assess the similarity of each tissue with yeast, I propose a novel statistical model, called *tissue-specific random model (TRAM)*, which takes into account the ubiquitous nature of housekeeping genes and mimics the topological structure of tissue-specific networks (please see Materials and Methods section for the details of the random model).

I use the alignment score of each tissue-yeast pair as the objective function. To asses the significance of each alignment score, I use a Monte Carlo simulation method to sample from the underlying probability distribution of alignment scores.

For each tissue-specific network, I sample $k_{\mathbb{R}} = 10,000$ pseudo-random tissues of the same size from TRAM, separately align them with the yeast interactome, and compute the number of conserved edges and sequence similarity of aligned protein

pairs as alignment statistics, in order to compute the empirical p-values. For each network alignment, I compute a *topological*, a *homological* (sequence-based), and a *mixed* (alignment score) p-value. Additionally, I use cases in which alignment quality is significantly better in the original tissue alignment, both in terms of sequence and topology, to quantify an *upper bound* on the alignment p-values. Conversely, cases in which both of these measures are improved in the random samples can be used to define a *lower bound* on the alignment p-value.

First, we note that all tissues with significant *mixed* p-values also have both significant topological and homological (sequence-based) p-values. For a majority of tissues with insignificant *mixed* p-values, we still observe significant homological, but insignificant topological p-values. I summarize the most and the least similar tissues to yeast by applying a threshold of $\alpha_l = \alpha_u = 10^{-2}$ to the p-value upper and lower bounds, respectively. Using the p-value upper bound ($\Delta_{\mathbb{R}}$) of 10^{-2}, I identify a total of 23 out of 79 tissues with high similarity to yeast. These are listed in Table 7.1. Among them, blood cells consistently show high significance, without even a single instance from $10,000$ samples having either the alignment weight or the edge overlap of the random sample exceeding the original alignment. Similarly, immune cell lines and male reproductive tissues also show significant alignment p-values, but with lower reliability scores. Conversely, there are 19 out of 79 tissues that have $\delta_{\mathbb{R}} > 10^{-2}$. These are least similar to yeast. Among these tissues, listed in Table 7.2, ganglion tissues consistently show the least similarity to yeast. An interesting observation is that tissues and cell types at either end of the table (either the most or the least similar) usually have very high reliability scores, that is both their topology and homology p-values are consistent.

7.3.4 Identifying Groups of Coherent Tissues

Next, I investigate the correlation between the similarity of human tissues among each other and how it relates to their corresponding alignment p-values with yeast

Table 7.1: Tissues with the most significant similarity to the yeast interactome

Name	pval lower bound	overall pval	pval upper bound	reliability
Myeloid Cells	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
Monocytes	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
Dentritic Cells	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
NK Cells	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
T-Helper Cells	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
Cytotoxic T-Cells	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
B-Cells	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
Endothelial	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
Hematopoietic Stem Cells	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
MOLT-4	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
B Lymphoblasts	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
HL-60	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
K-562	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
Early Erythroid	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
Bronchial Epithelial Cells	< 1.00e-04	< 1.00e-04	0.0002	0.9998
Colorectal Adenocarcinoma	< 1.00e-04	< 1.00e-04	0.0004	0.9996
Daudi	< 1.00e-04	< 1.00e-04	0.0009	0.9991
Testis Seminiferous Tubule	< 1.00e-04	< 1.00e-04	0.0012	0.9988
Smooth Muscle	< 1.00e-04	< 1.00e-04	0.0016	0.9984
Blood (Whole)	< 1.00e-04	< 1.00e-04	0.0053	0.9947
Thymus	< 1.00e-04	0.0001	0.0062	0.9938
Testis Interstitial	< 1.00e-04	0.0004	0.0086	0.9914

in order to better understand the transitivity of this relationship. I expect that similar tissues should exhibit consistent alignment p-values, resulting in groups of homogenous tissues with coherent alignments scores.

To this end, I first construct a network of tissue-tissue similarities (TTSN) using the global transcriptome of human tissues from the GNF Gene Atlas, including 44,775 human transcripts covering both known, as well as predicted and poorly characterized

Table 7.2: Tissues with the least significant similarity to the yeast interactome

Name	pval lower bound	overall pval	pval upper bound	reliability
Trigeminal Ganglion	0.9947	0.9994	1	0.9947
Superior Cervical Ganglion	0.9847	0.9991	1	0.9847
Ciliary Ganglion	0.9407	0.9813	0.9964	0.9443
Atrioventricular Node	0.8746	0.9792	0.9921	0.8825
Skin	0.8355	0.9297	0.9809	0.8546
Heart	0.7934	0.9585	0.9815	0.8119
Appendix	0.7596	0.9371	0.973	0.7866
Dorsal Root Ganglion	0.7065	0.933	0.9717	0.7348
Skeletal Muscle	0.3994	0.5902	0.7866	0.6128
Uterus Corpus	0.233	0.7736	0.8769	0.3561
Lung	0.0771	0.3853	0.5544	0.5227
Pons	0.0674	0.5201	0.6983	0.3691
Salivary Gland	0.0639	0.3449	0.5173	0.5466
Liver	0.0600	0.6857	0.8519	0.2081
Ovary	0.0388	0.2735	0.4481	0.5907
Trachea	0.0259	0.2376	0.4146	0.6113
Globus Pallidus	0.0206	0.2471	0.4336	0.587
Cerebellum	0.0127	0.1950	0.3783	0.6344

genes. For each pair of tissues/ cell types, I compute a similarity score using the Pearson correlation of their transcriptional signatures and use the 90th percentile of similarity scores to select the most similar pairs. I annotate each node in the TTSN with its corresponding alignment p-value as a measure of similarity with the yeast interactome. This meta-analysis allows us to investigate how linear measurements of gene/protein activity project to the space of protein interactions, in order to re-wire the underlying interactome in each human tissue.

Figure 7.4 presents the final network. In this network, each node represents a human tissue/cell type and each weighted edge illustrates the extent of overall tran-

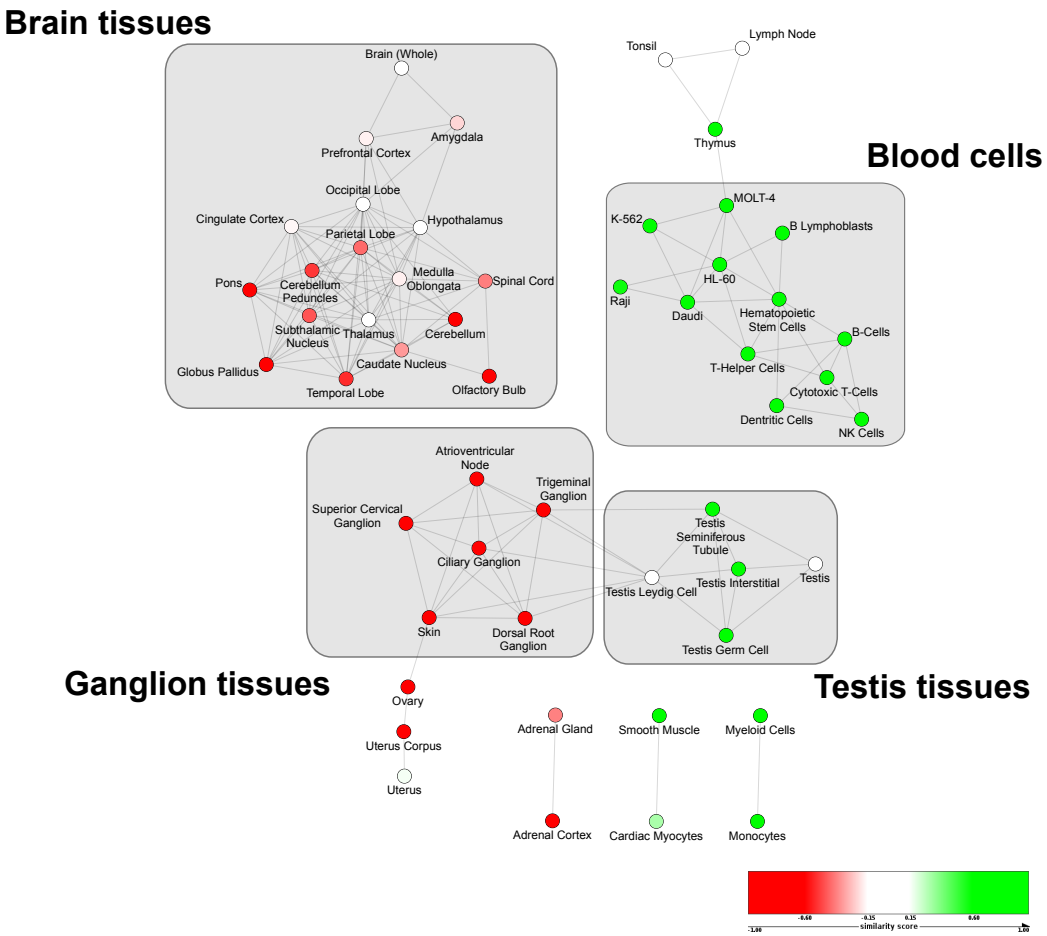


Figure 7.4.: Projection of alignment p-values on the network of tissue-tissue similarities

scriptional similarity between pairs of tissues. This network is filtered to include only tissue pairs with the highest overlap with each other. In order to assign color to each node, I use z-score normalization on the log-transformed alignment mixed p-values. Green and red nodes correspond to the highly positive and highly negative range of z-scores, which represent similar and dissimilar tissues to yeast, respectively.

Preliminary analysis of this network indicates that the alignment p-value of tissues highly correlates with their overall transcriptional overlap. Furthermore, these high-level interactions coincide with each other and fall within distinct *groups* with

consistent patterns. I manually identified four such groups and separately annotated them in the network. These groups correspond to brain tissue, blood cells, ganglion tissues, and testis tissues. Among these groups, blood cells and testis tissues exhibit consistent similarity with yeast, whereas brain and ganglion tissues bear consistent dissimilarity.

The existence of homogenous group of tissues with consistent similarity with yeast suggests an underlying conserved machinery in these clusters. This raises the question of what is consistently aligned within each tissue group and how it relates to the computed alignment p-values? I address this question, and relate it to the onset of tissue-specific pathologies in the remaining subsections.

7.3.5 Dissecting Tissue-selective Genes with Respect to Their Conservation

In this subsection, I investigate the subset of non-housekeeping genes in each homogenous group of human tissues and partition them into sets of genes, and their corresponding pathways that are either conserved in yeast or are human-specific. Next, I analyze how these pathways contribute to the overall similarity/ dissimilarity of human tissues with yeast.

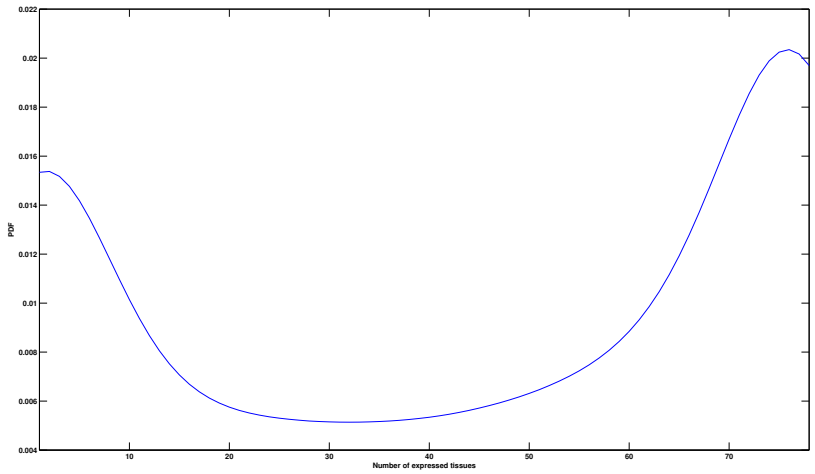


Figure 7.5.: Membership distribution of non-housekeeping genes in human tissues

Figure 7.5 presents the probability density function for the membership distribution of non-housekeeping genes in different human tissues. The observed bi-modal distribution suggests that most non-housekeeping genes are either expressed in a very few selected tissues or in the majority of human tissues. I use this to partition the set of expressed non-housekeeping genes, with the goal of identifying genes that are selectively expressed in each group of human tissues.

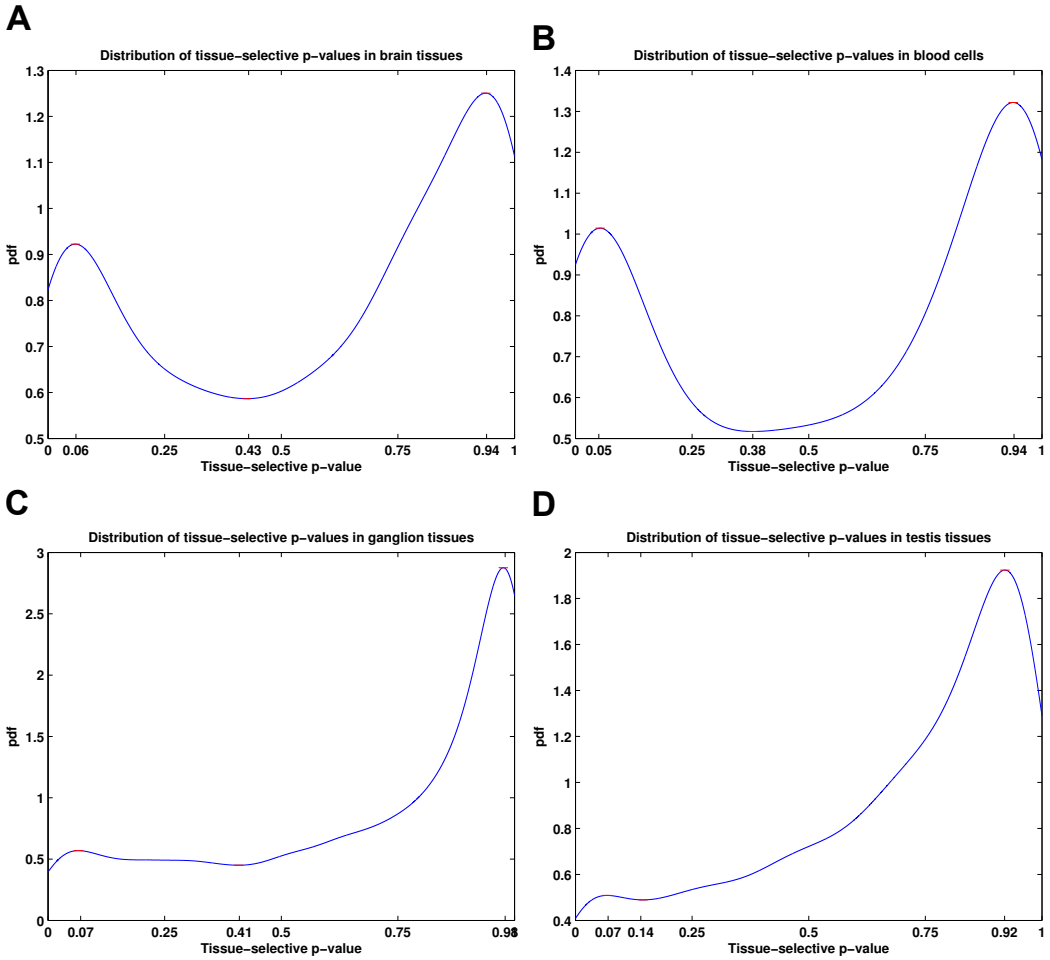


Figure 7.6.: Distribution of tissue-selectivity p-values in different tissue groups

I start with all *expressed non-housekeeping genes* in each tissue group, i.e., genes that are expressed in *at least* one of the tissue members. Next, in order to identify the subset of expressed genes that are *selectively* expressed in each group, I use the *tissue-*

selectivity p-*value* of each gene. In this formulation, a gene is identified as selectively expressed if it is expressed in a significantly higher number of tissues in the given group than randomly selected tissue subsets of the same size (see Materials and Methods section for details). Figure 7.6 illustrates the distribution of tissue-selectivity p-values of expressed genes with respect to four major groups in Figure 7.4. Each of these plots exhibit a bi-modal characteristic similar to the membership distribution function in Figure 7.5. This can be explained by the fact that membership distribution is a mixture distribution, with the underlying components being the same distribution for the subset of genes that are expressed in different tissue groups. I use critical points of the p-value distributions to threshold for tissue-selective genes. The motivation behind this choice is that these points provide shifts in the underlying distribution, from tissue-selective to ubiquitous genes. Given the bi-modal characteristics of these distributions, they all have three critical points, the first of which I use as the cutoff point. This provides highest precision for declared tissue-selective genes, but lower recall than the other two choices.

Having identified the subset of tissue-selective genes with respect to each tissue group, I use the majority voting scheme to tri-partition these sets based on their alignment consistency with yeast. Similar to the procedure I used to tri-partition housekeeping genes, I tried different choices of consensus rate parameter from 50% to 100% with increments of 5%. The percent of unclassified genes decreases linearly with the consensus rate, while relative portions of human-specific/ conserved genes remain the same. I chose 90% for my final results, as it leaves the least number of genes as unclassified, while keeping human-specific and conserved genes well-separated.

Table 7.3 presents the number of expressed genes, selectively expressed genes, and the percent of tissue-selective genes that are conserved, human-specific, or unclassified within each group of tissues. There is a similar relationship between the ratio of conserved/human-specific genes within each group of tissues and their alignment p-values, suggesting that alignment p-values are highly correlated with the conservation

Table 7.3: Summary of tissue-selective gene partitioning

Cluster name	# expressed genes	# TS genes	# CG (%)	# HS (%)	# unclassified (%)
Brain Tissues	5936	891	273 (30.64 %)	401 (45.01 %)	217 (24.35 %)
Blood Cells	6092	1093	460 (42.09 %)	385 (35.22 %)	248 (22.69 %)
Testis Tissues	5358	328	119 (36.28 %)	126 (38.41 %)	83 (25.30 %)
Ganglion Tissues	5278	274	76 (27.74 %)	136 (49.64 %)	62 (22.63 %)

of tissue-selective genes and their corresponding pathways. Figure 7.7 illustrates the relative sizes of each subset of genes identified in this study.

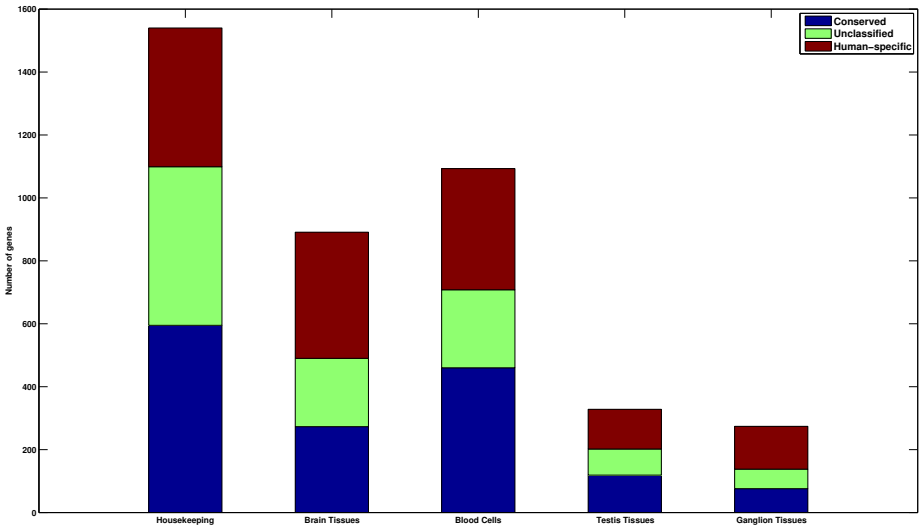


Figure 7.7.: Summary of gene classifications. Housekeeping and tissue-selective genes, in four main groups of human tissues, which are classified into three main classes based on their conservation in yeast

Conserved genes and their corresponding pathways comprise the functional subspace in which we can use yeast as a suitable model organism to study tissue-specific physiology and pathophysiology. On the other hand, human-specific genes provide a complementary set that can be used to construct *tissue-engineered* humanized yeast models. They also provide promising candidates for tissue-specific gene therapies in a

similar fashion to NDI1 therapy, in cases where an alternative functional mechanism can be found in yeast. To further investigate these subsets, I focus on blood cells and brain tissues, which illustrate the clearest separation between their tissue-selective and conserved genes in their TSS distribution, and subject them to more in depth functional analysis in next subsections.

7.3.6 Elucidating Functional Roles of the Brain and Blood Selective Genes

I use g:ProfileR on both human-specific and conserved genes to identify their enriched functions. These two subsets share many common terms, due to the underlying prior of both being subsets of tissue-selective genes. To comparatively analyze these functions and rank them based on their human-specificity, I use the log of p-value ratios between human-specific and conserved genes to filter terms that are at least within 2-fold enrichment. I focus on GO biological processes, KEGG pathways, and CORUM protein complexes and remove all genesets with more than 500 genes to filter for overly generic terms. Finally, to group these terms together and provide a visual representation of the functional space of genes, I use EnrichmentMap (EM) [122], a recent Cytoscape [171] plug-in, to construct a network (map) of the enriched terms. I use the log ratio of p-values to color each node in the graph. Figures 7.8 and 7.9 illustrate the final enrichment map of unique human-specific and conserved blood-selective and brain-selective functions, respectively.

Conserved blood-selective functions, shown in Figure 7.8 (A), are primarily enriched with terms related to DNA replication, cellular growth, and preparing cell for cell-cycle. Among these terms, DNA replication-is tightly linked to both DNA repair and telomere maintenance related terms. Telomere maintenance, specially via telomerase enzyme, is one of the cellular functions that is known to be conserved in yeast, but only active in a selected subset of differentiated human tissues and cell types, including hematopoietic stem cells and male reproductive tissues [109]. Functional terms involved in DNA conformation changes, including condensin complex, as well

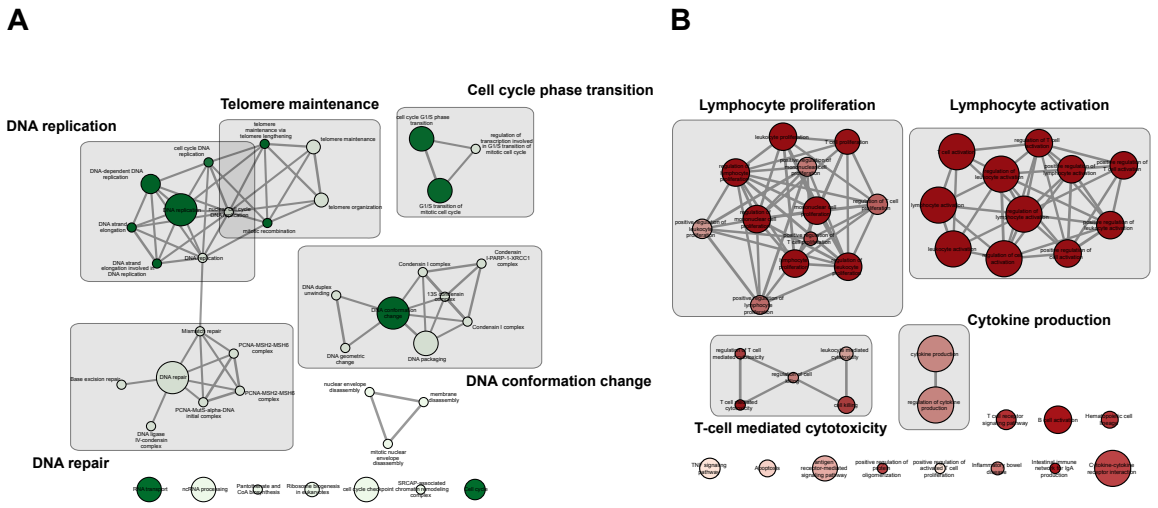


Figure 7.8.: Enrichment map of unique blood-selective functions

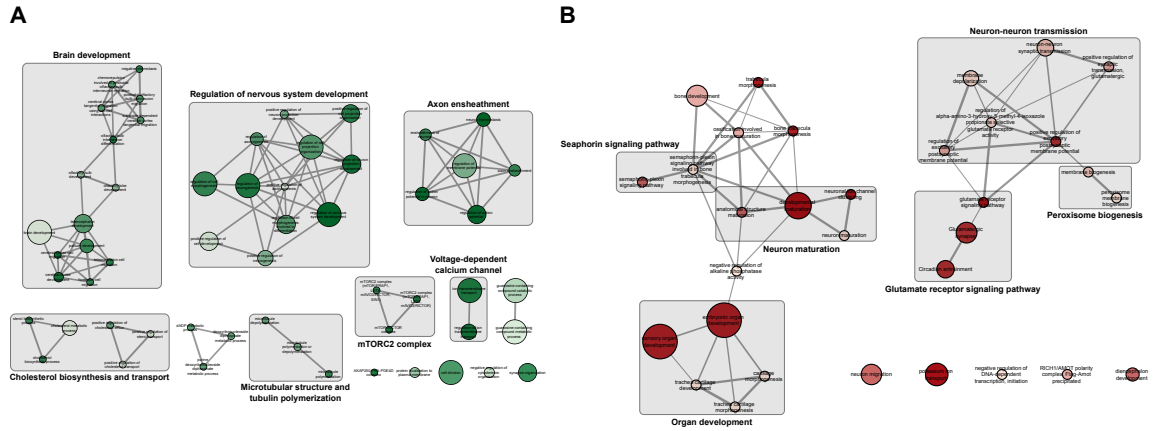


Figure 7.9.: Enrichment map of unique brain-selective functions

as cell cycle phase transition, specifically from G1 to S phases, are two other groups of conserved functional terms that are highly conserved from yeast to human. On the other hand, human-specific blood-selective functions, shown in Figure 7.8 (B), are mainly involved in lymphocyte proliferation and activation. Terms in these two groups are also tightly related to each other and form a larger cluster together. In addition, cytokine production and T-cell mediated cytotoxicity also exhibit human-

specific, blood-selective characteristics. This is partially expected, as these functions are highly specialized immune-cell functions that are evolved particularly in humans to ensure his survival in less-favorable conditions.

Figure 7.9 (A) shows the functional space of conserved brain-selective functions. Many of these terms correspond to various aspects of brain development, including olfactory bulb, telencephalon, pallium, and cerebral cortex development, as well as the regulatory circuit that controls nervous system development. Considering the unicellular nature of yeast, the exact mechanisms in which orthologs of these pathways modulate yeast cellular machinery is less studied. An in-depth analysis to identify matching phenologs can help us use yeast to study various disorders related to brain development. Another functional aspect that exhibits high conservation is the mTOR complex 2. The target of rapamycin (TOR) signaling is a highly conserved pathway, which forms two structurally distinct protein complexes, mTORC1 and mTORC2. The former complex has a central role in nutrient-sensing and cell growth, and as such, has been used extensively to study calorie restriction (CR) mediated lifespan extension. On the other hand, mTORC2 has been recently proposed to modulate consolidation of long-term memory [75]. Cholesterol biosynthesis and transport is another conserved functional aspect that differs significantly from other human tissues. As the most cholesterol-rich organ in the body, expression of genes corresponding to lipoprotein receptors and apolipoproteins is tightly regulated among different brain cells and plays an important role in normal brain development. Dysregulation of these metabolic pathways is implicated in various neurological disorders, such as Alzheimer's disease [132].Finally, microtubular structure and tubulin polymerization also shows significant conservation and is known to play a key role in brain development [189]. These cytoskeletal proteins have recently been associated with brain-specific pathologies, including epilepsy [87].

Finally, we study human-specific brain functions, which are shown in Figure 7.9 (B). One of the key functional aspects in this group is the semaphorin-plexin signaling pathway. This pathway was initially characterized based on its role in the anatomical

structure maturation of the brain, specifically via the repulsive axon guidance, but later was found to be essential for morphogenesis of a wide range of organ systems, including sensory organs and bone development [223]. Another human-specific signaling pathway identified in brain is the glutamate receptor signaling pathway, which also cross-talks with circadian entrainment, as well as neuron-neuron transmission. This pathway plays a critical role in neural plasticity, neural development and neurodegeneration [128]. It has also been associated with both chronic brain diseases, such as schizophrenia, as well as neurodegenerative disorders, such as Alzheimer's disease [210].

Both conserved and human-specific genes play important roles in tissue-specific pathologies. In addition, these genes, which are enriched with regulatory and signaling functions, cross-talk with housekeeping genes to control cellular response to various factors. As such, a complete picture of disease onset, development, and progression can only be achieved from a systems point of view. From this perspective, we study not only the genes (or their states) that are frequently altered in disease, but also the underlying tissue-specific and housekeeping pathways in which they interact to exhibit the observed phenotype(s). In the next subsection, I further investigate this hypothesis. I study the potential of different subsets of the identified tissue-selective genes for predicting tissue-specific pathologies.

7.3.7 Assessing the Significance of Tissue-specific Pathologies among Conserved and Human-specific Tissue-selective Genes

To further study the predictive power of tissue-selective genes for human pathologies, I use the *genetic association database (GAD)* disease annotations as my gold standard [11]. This database collects gene-disease associations from genetic association studies. Additionally, each disease has been assigned to one of the 19 different disease classes in GAD database. I use DAVID functional annotation tool for disease enrichment analysis of tissue-selective genes [74].

Table 7.4: Enriched disease classes of tissue-selective genes

	Conserved genes		Human-specific genes	
	Disease class	p-value	Disease class	p-value
Blood cells	Cancer	$9.3 * 10^{-4}$	Immune	$1.2 * 10^{-5}$
Brain tissues	Psych	$3.6 * 10^{-4}$	Psych	$5.7 * 10^{-8}$
	Chemdependency	$2.6 * 10^{-3}$	Neurological	$3.0 * 10^{-2}$
	Pharmacogenomic	$9.7 * 10^{-2}$		

First, I seek to identify which disease classes are significantly enriched among each set of tissue-selective genes. Table 7.4 shows the disease classes enriched in each group of brain and blood selective genes. Conserved blood-selective genes are predominantly enriched with cancers, whereas human-specific blood-selective genes are mainly associated with immune disorders. This can be linked to the previous results indicating that conserved subset is mainly involved in regulating growth, DNA replication, and cell cycle, whereas human-specific genes are primarily involved in lymphocyte proliferation and activation. Conversely, brain-selective genes show higher similarities in terms of disease classes that they can predict. Both conserved and human-specific brain-selective genes can predict psychiatric disorders, but human-specific subset seems to be a more accurate predictor. On the other hand, neurological disorders are only enriched in human-specific subset of brain-selective genes, whereas disorders classified as pharmacogenomic and chemdependency show higher enrichment in conserved genes.

To summarize the specific disorders that are enriched in each subset of brain-selective genes, I integrate all identified diseases and rank them based on their enrichment p-value, if it is only enriched in one set, or their most significant p-value, if it is enriched in both sets. Table 7.5 shows the top 10 disease terms enriched in either human-specific or conserved brain-selective genes. In majority of cases, human-specific genes are more significantly associated with brain-specific pathologies

Table 7.5: Comparative analysis of brain-specific pathologies

Disorder	Conserved genes	Human-specific genes
schizophrenia	0.008573	8.4905E-06
autism	0.048288	0.00077448
dementia	0.0014356	-
schizophrenia; schizoaffective disorder; bipolar disorder	-	0.0021433
myocardial infarct; cholesterol, HDL; triglycerides; atherosclerosis, coronary; macular degeneration; colorectal cancer	0.0051617	-
epilepsy	0.071562	0.0064716
seizures	-	0.020381
bipolar disorder	0.048288	0.022016
attention deficit disorder conduct disorder oppositional defiant disorder	0.032444	0.023865

than conserved genes. In addition, there are unique disorders, such as schizophrenia, bi-polar disorder, and seizures, that are only enriched among human-specific genes.

In conclusion, both conserved and human-specific subsets of tissue-selective genes are significantly associated with different human disorders. However, the human-specific subset shows higher association with tissue-specific pathologies. To this end, they guide us to appropriate molecular constructs (gene insertions) in yeast to explore molecular/functional mechanisms that cause tissue-specific dysfunction. Such mechanisms can be tested in humans, and if validated, yeast can serve as an experimental model for further investigations of biomarkers and pharmacological and genetic interventions.

8 CONCLUSION AND FUTURE DIRECTIONS

In this dissertation, I developed computational methods coupled with statistical models to analyze human transcriptomics and interactomics datasets from single cell level up to complex tissues. These methods lay the foundation to study cell type-specific biology and pathobiology at a scale that has not been possible before. To this end, I have proposed a novel algorithms to (i) measure similarity of cells, (ii) identify cell types from single cell datasets, (iii) separate cell types from complex tissues, (iv) reconstruct tissue-specific interactomes, and (v) assess conservation of these tissue-specific pathways.

One major direction for extending this work is to combine gene expression deconvolution with single cell analysis. Single cell transcriptomics can provide a rough sketch of what each cell type should look like. This cell type-specific expression panel then can be used to perform supervised deconvolution. On the other hand, one major challenge in single cell analysis is the lack of ability to estimate underlying fractions in complex mixtures, such as tumor microenvironment. Deconvolution techniques provide these cellular decompositions, which can be additionally incorporated into single cell analysis to correct for sampling biases, among other confounding factors.

LIST OF REFERENCES

LIST OF REFERENCES

[1] CVX, Matlab Software for Disciplined Convex Programming, Version 2.1. `http://cvxr.com/cvx`, March 2014.

[2] Alexander R Abbas, Kristen Wolslegel, Dhaya Seshasayee, Zora Modrusan, and Hilary F Clark. Deconvolution of Blood Microarray Data Identifies Cellular Activation Patterns in Systemic Lupus Erythematosus. *PLOS ONE*, 4(7):e6098, January 2009.

[3] Giulia Accardi, Calogero Caruso, Giuseppina Colonna-Romano, et al. Can Alzheimer Disease Be a form of Type 3 Diabetes? *Rejuvenation Research*, 15(2):217–221, 2012.

[4] Jaeil Ahn, Ying Yuan, Giovanni Parmigiani, et al. DeMix: Deconvolution for Mixed Cancer Transcriptomes Using Raw Measured Data. *Bioinformatics (Oxford, England)*, 29(15):1865–71, August 2013.

[5] Z. Altboum, Y. Steuerman, E. David, et al. Digital Cell Quantification Identifies Global Immune Cell Dynamics during Influenza Infection. *Molecular Systems Biology*, 10(2):720–720, 2014.

[6] K. G. Ardlie, D. S. Deluca, A. V. Segre, et al. The Genotype-Tissue Expression (GTEx) Pilot Analysis: Multitissue Gene Regulation in Humans. *Science*, 348(6235):648–660, May 2015.

[7] Gary D Bader and Christopher W V Hogue. An Automated Method for Finding Molecular Complexes in Large Protein Interaction Networks. *BMC Bioinformatics*, 4:2, January 2003.

[8] M. Bailly-Bechet, C. Borgs, A. Braunstein, et al. Finding Undetected Protein Associations in Cell Signaling by Belief Propagation. *Proceedings of the National Academy of Sciences*, 108(2):882–887, January 2011.

[9] Michelangela Barbieri, Massimiliano Bonafè, Claudio Franceschi, and Giuseppe Paolisso. Insulin/IGF-I-signaling Pathway: An Evolutionarily Conserved Mechanism of Longevity from Yeast to Humans. *American Journal of Physiology, Endocrinology, and Metabolism*, 285(5):E1064–71, 2003.

[10] Mohsen Bayati, David F Gleich, Amin Saberi, and Ying Wang. Message-Passing Algorithms for Sparse Network Alignment. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 7(1):3:1—-3:31, 2013.

[11] Kevin G Becker, Kathleen C Barnes, Tiffani J Bright, and S Alex Wang. The Genetic Association Database. *Nature Genetics*, 36(5):431–2, May 2004.

[12] Daniel C Berwick and Kirsten Harvey. The Importance of Wnt Signalling for Neurodegeneration in Parkinson's Disease. *Biochemical Society Transactions*, 40(5):1123–8, 2012.

[13] Ethan Bier and William Mcginnis. Model Organisms in the Study of Development And Disease. In *Inborn Errors of Development: The Molecular Basis of Clinical Disorders of Morphogenesis*. 2008.

[14] Kenneth D Birnbaum and Edo Kussell. Measuring Cell Identity in Noisy Biological Systems. *Nucleic Acids Research*, 39(21):9093–107, 2011.

[15] Juan S Bonifacino and Benjamin S Glick. The Mechanisms of Vesicle Budding and Fusion. *Cell*, 116(2):153–66, January 2004.

[16] Alice Bossi and Ben Lehner. Tissue Specificity and the Human Protein Interaction Network. *Molecular Systems Biology*, 5:260, January 2009.

[17] David Botstein and Gerald R Fink. Yeast: An Experimental organism for 21st Century Biology. *Genetics*, 189(3):695–704, 2011.

[18] Jeffrey L Brodsky and William R Skach. Protein Folding and Quality Control in the Endoplasmic Reticulum: Recent Lessons from Yeast and Mammalian Cell Systems. *Current Opinion in Cell Biology*, 23(4):464–75, 2011.

[19] Marija Buljan, Guilhem Chalancon, Sebastian Eustermann, et al. Tissue-Specific Splicing of Disordered Segments That Embed Binding Motifs Rewires Protein Interaction Networks. *Molecular Cell*, 46(6):871–883, 2012.

[20] D Carmona-Gutierrez, C Ruckenstuhl, M Bauer, et al. Cell Death in Yeast: Growing Applications of a Dying Buddy. *Cell Death and Differentiation*, 17(5):733–4, May 2010.

[21] Chris Carter. Alzheimer's Disease: APP, Gamma Secretase, APOE, CLU, CR1, PICALM, ABCA7, BIN1, CD2AP, CD33, EPHA1, and MS4A2, and their Relationships with Herpes Simplex, C. Pneumoniae, Other Suspect Pathogens, and the Immune System. *International Journal of Alzheimer's Disease*, 2011:1–34, 2011.

[22] Florence M G Cavalli, Richard Bourgon, Wolfgang Huber, Juan M Vaquerizas, and Nicholas M Luscombe. SpeCond: A Method to Detect Condition-specific Gene Expression. *Genome Biology*, 12(10):R101, January 2011.

[23] Cheng-Wei Chang, Wei-Chung Cheng, Chaang-Ray Chen, et al. Identification of Human Housekeeping Genes and Tissue-selective Genes by Microarray Meta-analysis. *PLOS ONE*, 6(7):e22859, January 2011.

[24] Raghunath Chatterjee and Charles Vinson. CpG Methylation Recruits Sequence Specific Transcription Factors Essential for Tissue Specific Gene Expression. *Biochimica et Biophysica Acta*, 1819(7):763–70, 2012.

[25] Li Chen, Tsung-Han Chan, Peter L Choyke, et al. CAM-CM: A Signal Deconvolution tool for in Vivo Dynamic Contrast-enhanced Imaging of Complex Tissues. *Bioinformatics (Oxford, England)*, 27(18):2607–9, September 2011.

[26] Raymond E Chen and Jeremy Thorner. Function and Regulation in MAPK Signaling Pathways: Lessons Learned from the Yeast Saccharomyces Cerevisiae. *Biochimica et Biophysica Acta*, 1773(8):1311–40, 2007.

[27] Yun Chen, Arnold B. Rabson, and David H. Gorski. MEOX2 Regulates Nuclear Factor-B Activity in Vascular Endothelial Cells Through Interactions with P65 and IB. *Cardiovascular Research*, 87(4):723–731, September 2010.

[28] J Michael Cherry, Eurie L Hong, Craig Amundsen, et al. Saccharomyces Genome Database: The Genomics Resource of Budding Yeast. *Nucleic Acids Research*, 40(Database issue):D700–5, January 2012.

[29] R J Cho, M J Campbell, E a Winzeler, et al. A Genome-wide Transcriptional Analysis of the Mitotic Cell Cycle. *Molecular Cell*, 2(1):65–73, July 1998.

[30] Chee Yeun Chung, Vikram Khurana, Pavan K Auluck, et al. Identification and Rescue of α-synuclein toxicity in Parkinson Patient-derived Neurons. *Science*, 342(6161):983–7, 2013.

[31] Caitlin Clapp, Liam Portt, Chamel Khoury, et al. Untangling the Roles of Anti-Apoptosis in Regulating Programmed Cell Death Using Humanized Yeast Cells. *Frontiers in Oncology*, 2:59, January 2012.

[32] Neil R Clark, Kevin S Hu, Axel S Feldmann, et al. The Characteristic Direction: A Geometrical Approach to Identify Differentially Expressed Genes. *BMC Bioinformatics*, 15(1):79, 2014.

[33] Alex J. Cornish, Ioannis Filippis, Alessia David, and Michael J.E Sternberg. Exploring the Cellular Basis of Human Disease Through a Large-scale Mapping of Deleterious Genes to Cell Types. *Genome Medicine*, 7(1):95, 2015.

[34] Michael Costanzo, Anastasia Baryshnikova, Jeremy Bellay, et al. The Genetic Landscape of a Cell. *Science*, 327(5964):425–31, January 2010.

[35] Adele Cutler and Leo Breiman. Archetypal Analysis. *Technometrics*, 36(4):338, November 1994.

[36] Claudio De Virgilio and Robbie Loewith. The tOR Signalling Network from Yeast to Man. *The International Journal of Biochemistry & Cell Biology*, 38(9):1476–81, January 2006.

[37] J L DeRisi, V R Iyer, and P O Brown. Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale. *Science*, 278(5338):680–6, 1997.

[38] R S Desikan, a J Schork, Y Wang, et al. Genetic Overlap Between Alzheimers Disease and Parkinsons Disease At the MAPT Locus. *Molecular Psychiatry*, 20(12):1588–1595, 2015.

[39] Zoltán Dezso, Yuri Nikolsky, Evgeny Sviridov, et al. A Comprehensive Functional Analysis of Tissue Specificity of Human Gene Expression. *BMC Biology*, 6:49, January 2008.

[40] Zoltán Dezso, Yuri Nikolsky, Evgeny Sviridov, et al. A Comprehensive Functional Analysis of Tissue Specificity of Human Gene Expression. *BMC Biology*, 6:49, 2008.

[41] Maitreya J Dunham and Douglas M Fowler. Contemporary, Yeast-based Approaches to Understanding Human Genetic Variation. *Current Opinion in Genetics & Development*, 23(6):658–64, 2013.

[42] Eran Eden. *Discovering Motifs in Ranked Lists of DNA Sequences*. PhD thesis, Technion - Israel Institute of Technology, 2007.

[43] Eran Eden, Doron Lipson, Sivan Yogev, and Zohar Yakhini. Discovering Motifs in Ranked Lists of DNA Sequences. *PLOS Computational Biology*, 3(3):e39, 2007.

[44] Eugene S Edgington. An Additive Method for Combining Probability Values from Independent Experiments. *The Journal of Psychology*, 80(2):351–363, 1972.

[45] Eli Eisenberg and Erez Y Levanon. Human housekeeping genes, revisited, 2013.

[46] Jonathan D. Ellis, Miriam Barrios-Rodiles, Recep Çolak, et al. Tissue-specific Alternative Splicing Remodels Protein-protein Interaction Networks. *Molecular Cell*, 46(6):884–92, 2012.

[47] Timo Erkkilä, Saara Lehmusvaara, Pekka Ruusuvuori, et al. Probabilistic Analysis of Gene Expression Measurements from Heterogeneous Tissues. *Bioinformatics (Oxford, England)*, 26(20):2571–7, October 2010.

[48] Erez Feige, Satoru Yokoyama, Carmit Levy, et al. Hypoxia-induced Transcriptional Repression of the Melanoma-associated Oncogene MITF. *Proceedings of the National Academy of Sciences of the United States of America*, 108(43):E924–33, October 2011.

[49] R. A Fisher. *Statistical Methods for Research Workers*. Cosmo study guides. Cosmo Publications, 1925.

[50] Kristoffer Forslund, Fabian Schreiber, Nattaphon Thanintorn, and Erik L L Sonnhammer. OrthoDisease: Tracking Disease Gene Orthologs Across 100 Species. *Briefings in Bioinformatics*, 12(5):463–73, 2011.

[51] Johann a Gagnon-Bartsch and Terence P Speed. Using Control Genes to Correct for Unwanted Variation in Microarray Data. *Biostatistics (Oxford, England)*, 13(3):539–52, July 2012.

[52] Chris Gaiteri, Mingming Chen, Boleslaw Szymanski, et al. Identifying Robust Communities and Multi-community Nodes by Combining top-down and Bottom-up Approaches to Clustering. *Scientific Reports*, 5:16361, November 2015.

[53] Renaud Gaujoux and Cathal Seoighe. Semi-supervised Nonnegative Matrix Factorization for Gene Expression Deconvolution: A Case Study. *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, 12(5):913–21, July 2012.

[54] Renaud Gaujoux and Cathal Seoighe. CellMix: A Comprehensive Toolbox for Gene Expression Deconvolution. *Bioinformatics (Oxford, England)*, 29(17):2211–2, September 2013.

[55] Sina Ghaemmaghami, Won-Ki Huh, Kiowa Bower, et al. Global Analysis of Protein Expression in Yeast. *Nature*, 425(6959):737–41, 2003.

[56] Guri Giaever, Angela M Chu, Li Ni, et al. Functional Profiling of the Saccharomyces Cerevisiae Genome. *Nature*, 418(6896):387–91, July 2002.

[57] Nicolas Gillis. Successive Nonnegative Projection Algorithm for Robust Nonnegative Blind Source Separation. *SIAM Journal on Imaging Sciences*, 7(2):1420–1450, January 2014.

[58] Nicolas Gillis and Stephen A. Vavasis. Semidefinite Programming Based Preconditioning for More Robust Near-Separable Nonnegative Matrix Factorization. *SIAM Journal on Optimization*, 25(1):677–698, January 2015.

[59] A Goffeau, B G Barrell, H Bussey, et al. Life with 6000 Genes. *Science*, 274(5287):546, 563–7, 1996.

[60] Kwang-Il Goh, Michael E Cusick, David Valle, et al. The Human Disease Network. *Proceedings of the National Academy of Sciences of the United States of America*, 104(21):8685–90, May 2007.

[61] Ting Gong, Nicole Hartmann, Isaac S Kohane, et al. Optimal Deconvolution of Transcriptional Profiling Data Using Quadratic Programming with Application to Complex Clinical Blood Samples. *PLOS ONE*, 6(11):e27156, January 2011.

[62] Harald Göring. Tissue Specificity of Genetic Regulation of Gene Expression. *Nature Genetics*, 44(10):1077–1078, 2012.

[63] Michael Grant and Stephen Boyd. Graph Implementations for Nonsmooth Convex Programs. In V. Blondel, S. Boyd, and H. Kimura, editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited, 2008. `http://stanford.edu/~boyd/graph_dcp.html`.

[64] Casey S Greene, Arjun Krishnan, Aaron K Wong, et al. Understanding Multicellular Function and Disease With Human Tissue-specific Networks. *Nature Genetics*, 32(4):453–465, 2015.

[65] Dominic Grün, Anna Lyubimova, Lennart Kester, et al. Single-cell Messenger RNA Sequencing Reveals Rare Intestinal Cell Types. *Nature*, 525(7568):251–5, September 2015.

[66] Nicoletta Guaragnella, Vanessa Palermo, Alvaro Galli, et al. The Expanding Role of Yeast in Cancer Research and Diagnosis: Insights Into the Function of the Oncosuppressors P53 and BRCA1/2. *FEMS Yeast Research*, 2013.

[67] Guoji Guo, Sidinh Luc, Eugenio Marco, et al. Mapping Cellular Hierarchy by Single-Cell Analysis of the Cell Surface Repertoire. *Cell Stem Cell*, 13(4):492–505, October 2013.

[68] Laleh Haghverdi, Florian Buettner, and Fabian J. Theis. Diffusion Maps for High-dimensional Single-cell Analysis of Differentiation Data. *Bioinformatics*, 31(18):2989–2998, September 2015.

[69] Junwei Han, Xinrui Shi, Yunpeng Zhang, et al. ESEA: Discovering the Dysregulated Pathways Based On Edge Set Enrichment Analysis. *Scientific Reports*, 5:13044, 2015.

[70] Yuval Hart, Hila Sheftel, Jean Hausser, et al. Inferring Biological Tasks Using Pareto Analysis of High-dimensional Data. *Nature Methods*, 12(3):233–235, January 2015.

[71] Leland H Hartwell. Nobel Lecture. Yeast and Cancer. *Bioscience Reports*, 22(3-4):373–94, 2002.

[72] C.W. Hesse and C.J James. On Semi-Blind Source Separation Using Spatial Constraints with Applications in EEG Analysis. *Biomedical Engineering, IEEE Transactions on*, 53(12):2525–2534, December 2006.

[73] Daniel S. Himmelstein and Sergio E Baranzini. Heterogeneous Network Edge Prediction: A Data Integration Approach to Prioritize Disease-Associated Genes. *PLOS Computational Biology*, 11(7):e1004259, July 2015.

[74] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources. *Nature Protocols*, 4(1):44–57, January 2009.

[75] Wei Huang, Ping June Zhu, Shixing Zhang, et al. mTORC2 Controls Actin Polymerization Required for Consolidation of Long-term Memory. *Nature Neuroscience*, 16(4):441–8, 2013.

[76] Won-Ki Huh, James V Falvo, Luke C Gerke, et al. Global Analysis of Protein Localization in Budding Yeast. *Nature*, 425(6959):686–91, 2003.

[77] Keisuke Ikegami, Xiao-Hui Liao, Yuta Hoshino, et al. Tissue-Specific Posttranslational Modification Allows Functional Targeting of Thyrotropin. *Cell Reports*, 9(3):801–809, 2014.

[78] T Ito, T Chiba, R Ozawa, et al. A Comprehensive Two-hybrid Analysis to Explore the Yeast Protein Interactome. *Proceedings of the National Academy of Sciences of the United States of America*, 98(8):4569–74, 2001.

[79] V R Iyer, C E Horak, C S Scafe, et al. Genomic Binding Sites of the Yeast Cell-cycle Transcription Factors SBF and MBF. *Nature*, 409(6819):533–8, January 2001.

[80] Marine Jeanmougin, Aurelien de Reynies, Laetitia Marisa, et al. Should We Abandon the T-test in the Analysis of Gene Expression Microarray Data: A Comparison of Variance Modeling Strategies. *PLOS ONE*, 5(9):e12336, January 2010.

[81] Michael C Jewett, Gerald Hofmann, and Jens Nielsen. Fungal Metabolite Analysis in Genomics and Phenomics. *Current Opinion in Biotechnology*, 17(2):191–7, 2006.

[82] Zhicheng Ji and Hongkai Ji. TSCAN: Pseudo-time Reconstruction and Evaluation In Single-cell RNA-seq Analysis. *Nucleic Acids Research*, 44(13):e117–e117, July 2016.

[83] Lan Jiang, Huidong Chen, Luca Pinello, and Guo-Cheng Yuan. GiniClust: Detecting Rare Cell Types From Single-cell Gene Expression Data with Gini Index. *Genome Biology*, 17(1):144, December 2016.

[84] Grace Marie Jones, Jim Stalker, Sean Humphray, et al. A Systematic Library for Comprehensive Overexpression Screens in Saccharomyces Cerevisiae. *Nature Methods*, 5(3):239–41, 2008.

[85] Wenjun Ju, Casey S Greene, Felix Eichinger, et al. Defining Cell-type Specificity At the Transcriptional Level in Human Disease. *Genome Research*, 23(11):1862–73, November 2013.

[86] Koji Kadota, Jiazhen Ye, Yuji Nakai, Tohru Terada, and Kentaro Shimizu. ROKU: A Novel Method for Identification of Tissue-specific Genes. *BMC Bioinformatics*, 7:294, January 2006.

[87] Ludmyla Kandratavicius, Mariana Raquel Monteiro, Jaime Eduardo Hallak, et al. Microtubule-associated Proteins in Mesial Temporal Lobe Epilepsy with and Without Psychiatric Comorbidities and their Relation with Granular Cell Layer Dispersion. *BioMed Research International*, 2013:960126, January 2013.

[88] H. Kawaji, M. Lizio, M. Itoh, et al. Comparison of CAGE and RNA-seq Transcriptome Profiling Using Clonally Amplified and Single-molecule Next-generation Sequencing. *Genome Research*, 24(4):708–717, 2014.

[89] Vikram Khurana and Susan Lindquist. Modelling Neurodegeneration in Saccharomyces Cerevisiae: Why Cook with Baker's Yeast? *Nature Reviews. Neuroscience*, 11(6):436–49, 2010.

[90] Jeong-Min Kim, Kyu-Hwa Lee, Yeo-Jin Jeon, et al. Identification of Genes Related to Parkinson's Disease Using Expressed Sequence Tags. *DNA research: An International Journal for Rapid Publication of Reports on Genes and Genomes*, 13(6):275–86, 2006.

[91] Jingu Kim, Yunlong He, and Haesun Park. Algorithms for Nonnegative Matrix and Tensor Factorizations: A Unified View Based on Block Coordinate Descent Framework. *Journal of Global Optimization*, 58(2):285–319, March 2013.

[92] Allon M Klein, Linas Mazutis, Ilke Akartuna, et al. Droplet Barcoding for Single-cell Transcriptomics Applied to Embryonic Stem Cells. *Cell*, 161(5):1187–201, May 2015.

[93] Christiaan Klijn, Steffen Durinck, Eric W Stawiski, et al. A Comprehensive Transcriptional Portrait of Human Cancer Cell Lines. *Nature Biotechnology*, 33(3):306–312, December 2014.

[94] Sebastian Köhler, Sebastian Bauer, Denise Horn, and Peter N Robinson. Walking the Interactome for Prioritization of Candidate Disease Genes. *American Journal of Human Genetics*, 82(4):949–58, 2008.

[95] Raivo Kolde. *GOsummaries: Word Cloud Summaries of GO Enrichment Analysis*, 2014. R package version 2.0.0.

[96] Yael Korem, Pablo Szekely, Yuval Hart, et al. Geometry of the Gene Expression Space of Individual Cells. *PLOS Computational Biology*, 11(7):e1004224, July 2015.

[97] Mehmet Koyutürk, Yohan Kim, Umut Topkara, et al. Pairwise Alignment of Protein Interaction Networks. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, 13(2):182–99, 2006.

[98] Nevan J Krogan, Gerard Cagney, Haiyuan Yu, et al. Global Landscape of Protein Complexes in the Yeast Saccharomyces Cerevisiae. *Nature*, 440(7084):637–43, 2006.

[99] Oleksii Kuchaiev and Natasa Przulj. Integrative Network Alignment Reveals Large Regions of Global Network Similarity in Yeast and Human. *Bioinformatics (Oxford, England)*, 27(10):1390–6, May 2011.

[100] Alexandre Kuhn, Azad Kumar, Alexandra Beilina, et al. Cell Population-specific Expression Analysis of Human Cerebellum. *BMC Genomics*, 13:610, January 2012.

[101] Alexandre Kuhn, Doris Thu, Henry J Waldvogel, Richard L M Faull, and Ruth Luthi-Carter. Population-specific Expression Analysis (PSEA) Reveals Molecular Changes in Diseased Brain. *Nature Methods*, 8(11):945–7, November 2011.

[102] H. W. Kuhn and Bryn Yaw. The Hungarian Method for the Assignment Problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955.

[103] Kasper Lage, Niclas Tue Hansen, E Olof Karlberg, et al. A Large-scale Analysis of Tissue-specific Pathology and Gene Expression of Human Disease Genes and Complexes. *Proceedings of the National Academy of Sciences of the United States of America*, 105(52):20870–5, 2008.

[104] D a Lashkari, J L DeRisi, J H McCusker, et al. Yeast Microarrays for Genome Wide Parallel Genetic and Gene Expression Analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 94(24):13057–62, 1997.

[105] Edward C Lauterbach. Psychotropics Regulate Skp1a, Aldh1a1, and Hspa8 Transcription Potential to Delay Parkinson's Disease. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 40:236–239, January 2013.

[106] J D Lieb, X Liu, D Botstein, and P O Brown. Promoter-specific Binding of Rap1 Revealed By Genome-wide Maps of Protein-DNA Association. *Nature Genetics*, 28(4):327–34, 2001.

[107] David a Liebner, Kun Huang, and Jeffrey D Parvin. MMAD: Microarray Microdissection with Analysis of Differences Is a Computational tool for Deconvoluting Cell Type-specific Contributions from Tissue Samples. *Bioinformatics (Oxford, England)*, 30(5):682–9, March 2014.

[108] W. K. Lim, K. Wang, C. Lefebvre, and A Califano. Comparative Analysis of Microarray Normalization Procedures: Effects on Reverse Engineering Gene Networks. *Bioinformatics*, 23(13):i282–i288, July 2007.

[109] Jue Lin, Elissa Epel, Joshua Cheon, et al. Analyses and Comparisons of Telomerase Activity and Telomere Length in Human T and B Cells: Insights for Epidemiology of Telomere Maintenance. *Journal of Immunological Methods*, 352(1-2):71–80, January 2010.

[110] Zhi-Ping Liu, Canglin Wu, Hongyu Miao, and Hulin Wu. RegNetwork: An Integrated Database of Transcriptional and Post-transcriptional Regulatory Networks in Human and Mouse. *Database*, 2015:bav095, September 2015.

[111] Mark W. Logue, Matthew Schu, Badri N. Vardarajan, et al. A Search for Age-related Macular Degeneration Risk Variants in Alzheimer Disease Genes and Pathways. *Neurobiology of Aging*, 35(6):1510.e7–1510.e18, 2014.

[112] Valter D Longo, Gerald S Shadel, Matt Kaeberlein, and Brian Kennedy. Replicative and Chronological Aging in Saccharomyces Cerevisiae. *Cell Metabolism*, 16(1):18–31, July 2012.

[113] Wing-Kin Ma, Jose M. Bioucas-Dias, Tsung-Han Chan, et al. A Signal Processing Perspective on Hyperspectral Unmixing: Insights from Remote Sensing. *IEEE Signal Processing Magazine*, 31(1):67–81, January 2014.

[114] Evan Z Macosko, Anindita Basu, Rahul Satija, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161(5):1202–14, May 2015.

[115] Oded Magger, Yedael Y Waldman, Eytan Ruppin, and Roded Sharan. Enhancing the Prioritization of Disease-causing Genes Through Tissue Specific Protein Interaction Networks. *PLOS Computational Biology*, 8(9):e1002690, 2012.

[116] Olvi L. Mangasarian and David R. Musicant. Robust Linear and Support Vector Regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(9):950–955, 2000.

[117] Eugenio Marco, Robert L. Karp, Guoji Guo, et al. Bifurcation Analysis of Single-cell Gene Expression Data Reveals Epigenetic Landscape. *Proceedings of the National Academy of Sciences*, 111(52):E5643–E5650, December 2014.

[118] Mathieu Marella, Byoung Boo Seo, Takao Yagi, and Akemi Matsuno-Yagi. Parkinson's Disease and Mitochondrial Complex I: A Perspective on the Ndi1 therapy. *Journal of Bioenergetics and Biomembranes*, 41(6):493–7, 2009.

[119] Kriston L McGary, Tae Joo Park, John O Woods, et al. Systematic Discovery of Nonobvious Human Disease Models Through orthologous Phenotypes. *Proceedings of the National Academy of Sciences of the United States of America*, 107(14):6544–9, 2010.

[120] M. Mele, P. G. Ferreira, F. Reverter, et al. The Human Transcriptome Across Tissues and Individuals. *Science*, 348(6235):660–665, May 2015.

[121] I Mendizabal, T E Keller, J Zeng, and Soojin V Yi. Epigenetics and Evolution. *Integrative and Comparative Biology*, 54(1):31–42, 2014.

[122] Daniele Merico, Ruth Isserlin, Oliver Stueker, Andrew Emili, and Gary D Bader. Enrichment Map: A Network-based Method for Gene-set Enrichment Visualization and Interpretation. *PLOS ONE*, 5(11):e13984, January 2010.

[123] David N Messina, Jarret Glasscock, Warren Gish, and Michael Lovett. An oRFeome-based Analysis of Human Transcription Factor Genes and the Construction of a Microarray to Interrogate their Expression. *Genome Research*, 14(10B):2041–7, 2004.

[124] Shahin Mohammadi and Ananth Grama. A Novel Method to Enhance the Sensitivity of Marker Detection Using a Refined Hierarchical Prior of Tissue Similarities. Technical report, bioRxiv, 2015.

[125] Shahin Mohammadi and Ananth Grama. De novo identification of cell type hierarchy with application to compound marker detection. In *ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM BCB)*, 2016.

[126] MOSEK-ApS. *The MOSEK Optimization toolbox for MATLAB Manual. Version 7.1 (Revision 28)*, 2015.

[127] Ana Joyce Munoz, Kwanjeera Wanichthanarak, Eugenio Meza, and Dina Petranovic. Systems Biology of Yeast Cell Death. *FEMS Yeast Research*, 12(2):249–65, 2012.

[128] S Nakanishi, Y Nakajima, M Masu, et al. Glutamate Receptors: Brain Function and Signal Transduction. *Brain Research. Brain Research Reviews*, 26(2-3):230–5, May 1998.

[129] Aaron M Newman, Chih Long Liu, Michael R Green, et al. Robust Enumeration of Cell Subsets from Tissue Expression Profiles. *Nature Methods*, (2014):1–10, 2015.

[130] Noa Novershtern, Aravind Subramanian, Lee N Lawton, et al. Densely Interconnected Transcriptional Circuits Control Cell States in Human Hematopoiesis. *Cell*, 144(2):296–309, January 2011.

[131] D. Nuzillard and A Bijaoui. Blind Source Separation and Analysis of Multispectral Astronomical Images. *Astronomy & Astrophysics Supplement Series*, 147:129–138, 2000.

[132] Matthias Orth and Stefano Bellosta. Cholesterol: Its Regulation and Role in Central Nervous System Disorders. *Cholesterol*, 2012:292598, January 2012.

[133] Nathan Pankratz, Gary W. Beecham, Anita L. DeStefano, et al. Meta-analysis of Parkinson's Disease: Identification of a Novel Locus, RIT2. *Annals of Neurology*, 71(3):370–384, 2012.

[134] V. Paul Pauca, J. Piper, and Robert J Plemmons. Nonnegative Matrix Factorization for Spectral Data Analysis. *Linear Algebra and its Applications*, 416(1):29–47, July 2006.

[135] W R Pearson and D J Lipman. Improved tools for Biological Sequence Analysis. *Proceedings of the National Academy of Sciences*, 85:2444–2448, 1988.

[136] M.S. Pedersen, U. Kjems, K.B. Rasmussen, and L.K Hansen. Semi-blind source separation using head-related transfer functions [speech signal separation]. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 5, pages V–713–16 vol.5, May 2004.

[137] Clara Pereira, Cláudia Bessa, Joana Soares, Mariana Leão, and Lucília Saraiva. Contribution of Yeast Models to Neurodegeneration Research. *Journal of Biomedicine & Biotechnology*, 2012:941232, January 2012.

[138] Clara Pereira, Isabel Coutinho, Joana Soares, et al. New Insights Into Cancer-related Proteins Provided By the Yeast Model. *The FEBS Journal*, 279(5):697–712, 2012.

[139] Clara Pereira, Isabel Coutinho, Joana Soares, et al. New Insights Into Cancer-related Proteins Provided By the Yeast Model. *The FEBS Journal*, 279(5):697–712, 2012.

[140] Fabiana Perocchi, Eugenio Mancera, and Lars M Steinmetz. Systematic Screens for Human Disease Genes, From Yeast to Human and Back. *Molecular Biosystems*, 4(1):18–29, January 2008.

[141] Dina Petranovic and Jens Nielsen. Can Yeast Systems Biology Contribute to the Understanding of Human Disease? *Trends in Biotechnology*, 26(11):584–90, 2008.

[142] Dina Petranovic, Keith Tyo, Goutham N Vemuri, and Jens Nielsen. Prospects of Yeast Systems Biology for Human Health: Integrating Lipid, Protein and Energy Metabolism. *FEMS Yeast Research*, 10(8):1046–59, 2010.

[143] Stephen R Piccolo, Michelle R Withers, Owen E Francis, Andrea H Bild, and W Evan Johnson. Multiplatform Single-sample Estimates of Transcriptional Activation. *Proceedings of the National Academy of Sciences of the United States of America*, 110(44):17778–83, 2013.

[144] Simone Picelli, Omid R Faridani, Asa K Björklund, et al. Full-length RNA-seq from Single Cells Using Smart-seq2. *Nature Protocols*, 9(1):171–81, January 2014.

[145] V. Plaks, C. D. Koopman, and Z. Werb. Circulating Tumor Cells. *Science*, 341(6151):1186–1188, September 2013.

[146] Alex a Pollen, Tomasz J Nowakowski, Joe Shuga, et al. Low-coverage Single-cell MRNA Sequencing Reveals Cellular Heterogeneity and Activated Signaling Pathways in Developing Cerebral Cortex. *Nature Biotechnology*, 32(10):1053–1058, August 2014.

[147] Yufeng Qian, Aashiq H Kachroo, Christopher M Yellman, Edward M Marcotte, and Kenneth A Johnson. Yeast Cells Expressing the Human Mitochondrial DNA Polymerase Reveal Correlations Between Polymerase Fidelity and Human Disease Progression. *The Journal of Biological Chemistry*, 289(9):5970–85, 2014.

[148] Wenlian Qiao, Gerald Quon, Elizabeth Csaszar, et al. PERT: A Method for Expression Deconvolution of Human Blood Samples from Varied Microenvironmental and Developmental Conditions. *PLOS Computational Biology*, 8(12):e1002838, January 2012.

[149] Gerald Quon, Syed Haider, Amit G Deshwar, et al. Computational Purification of Individual Tumor Gene Expression Profiles Leads to Significant Improvements in Prognostic Prediction. *Genome Medicine*, 5(3):29, January 2013.

[150] Towfique Raj, Katie Rothamel, Sara Mostafavi, et al. Polarization of the Effects of Autoimmune and Neurodegenerative Risk Alleles in Leukocytes. *Science*, 344(6183):519–23, 2014.

[151] Daniel Ramsköld, Shujun Luo, Yu-Chieh Wang, et al. Full-length MRNA-Seq from Single-cell Levels of RNA and Individual Circulating Tumor Cells. *Nature Biotechnology*, 30(8):777–782, July 2012.

[152] Sabry Razick, George Magklaras, and Ian M Donaldson. iRefIndex: A Consolidated Protein Interaction Database with Provenance. *BMC Bioinformatics*, 9(1):405, 2008.

[153] J. Reimand, T. Arak, and J Vilo. g:Profiler–a Web Server for Functional Interpretation of Gene Lists (2011 Update). *Nucleic Acids Research*, 39(suppl):W307–W315, July 2011.

[154] Dirk Repsilber, Sabine Kern, Anna Telaar, et al. Biomarker Discovery in Heterogeneous Tissue Samples - Taking the In-silico Deconfounding Approach. *BMC Bioinformatics*, 11:27, January 2010.

[155] Thomas Rolland, Murat Taan, Benoit Charloteaux, et al. A Proteome-Scale Map of the Human Interactome Network. *Cell*, 159(5):1212–1226, 2014.

[156] Assieh Saadatpour, Shujing Lai, Guoji Guo, and Guo-Cheng Yuan. Single-Cell Analysis in Cancer Genomics. *Trends in Genetics*, 31(10):576–86, October 2015.

[157] Nidhi Sahni, Song Yi, Mikko Taipale, et al. Widespread Macromolecular Interaction Perturbations in Human Genetic Disorders. *Cell*, 161(3):647–660, 2015.

[158] Sayed Mohammad Ebrahim Sahraeian and Byung-June Yoon. A Network Synthesis Model for Generating Protein Interaction Network Families. *PLOS ONE*, 7:e41474, 2012.

[159] Xu Bin Sai, Tomohiko Makiyama, Hiroshi Sakane, et al. TSG101, a Tumor Susceptibility Gene, Bidirectionally Modulates Cell Invasion Through Regulating MMP-9 MRNA Expression. *BMC Cancer*, 15(1):933, December 2015.

[160] Jonathan Schug, Winfried-Paul Schuller, Claudia Kappen, et al. Promoter Features Related to Tissue Specificity As Measured by Shannon Entropy. *Genome Biology*, 6(4):R33, 2005.

[161] Russell Schwartz and Stanley E Shackney. Applying Unmixing to Gene Expression Data for Tumor Phylogeny Inference. *BMC Bioinformatics*, 11:42, 2010.

[162] Ehud Shapiro, Tamir Biezuner, and Sten Linnarsson. Single-cell Sequencing-based Technologies Will Revolutionize Whole-organism Science. *Nature Reviews. Genetics*, 14(9):618–630, July 2013.

[163] Roded Sharan, Silpa Suthram, Ryan M Kelley, et al. Conserved Patterns of Protein Interaction in Multiple Species. *Proceedings of the National Academy of Sciences of the United States of America*, 102(6):1974–9, 2005.

[164] Shai S Shen-Orr and Renaud Gaujoux. Computational Deconvolution: Extracting Cell Type-specific Information from Heterogeneous Samples. *Current Opinion in Immunology*, 25(5):571–8, 2013.

[165] Shai S Shen-Orr, Robert Tibshirani, Purvesh Khatri, et al. Cell Type-specific Gene Expression Differences In Complex Tissues. *Nature Methods*, 7(4):287–9, April 2010.

[166] Sandra Siegert, Erik Cabuy, Brigitte Gross Scherf, et al. Transcriptional Code and Disease Map for Adult Retinal Cell Types. *Nature Neuroscience*, 15(3):487–495, 2012.

[167] Rohit Singh, Jinbo Xu, and Bonnie Berger. Global Alignment of Multiple Protein Interaction Networks with Application to Functional Orthology Detection. *Proceedings of the National Academy of Sciences of the United States of America*, 105(35):12763–8, 2008.

[168] Michael G Smith and Michael Snyder. Yeast As a Model for Human Disease. *Current Protocols in Human Genetics*, Chapter 15:Unit 15.6, 2006.

[169] T F Smith and M S Waterman. Identification of Common Molecular Subsequences. *Journal of Molecular Biology*, 147(1):195–7, 1981.

[170] Alex J. Smola and Bernhard Schölkopf. A Tutorial on Support Vector Regression. *Statistics and Computing*, 14(3):199–222, August 2004.

[171] Michael E Smoot, Keiichiro Ono, Johannes Ruscheinski, Peng-Liang Wang, and Trey Ideker. Cytoscape 2.8: New Features for Data Integration and Network Visualization. *Bioinformatics (Oxford, England)*, 27(3):431–2, 2011.

[172] Ouissem Souiai, Emmanuelle Becker, Carlos Prieto, et al. Functional Integrative Levels in the Human Interactome Recapitulate organ organization. *PLOS ONE*, 6(7):e22051, January 2011.

[173] Nathan Souviraà-Labastie, Anaik Olivero, Emmanuel Vincent, and Frédéric Bimbot. Multi-channel Audio Source Separation Using Multiple Deformed References. *IEEE Transactions on Audio, Speech and Language Processing*, 23(11):1775–1787, 2015.

[174] Chris Stark, Bobby-Joe Breitkreutz, Andrew Chatr-Aryamontri, et al. The BioGRID Interaction Database: 2011 Update. *Nucleic Acids Research*, 39(Database issue):D698–704, January 2011.

[175] Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, et al. BioGRID: A General Repository for Interaction Datasets. *Nucleic Acids Research*, 34(Database issue):D535–9, January 2006.

[176] Oliver Stegle, Sarah A. Teichmann, and John C. Marioni. Computational and Analytical Challenges in Single-Cell Transcriptomics. *Nature Review Genetics*, 16(3):133–145, 03 2015.

[177] Andrew I Su, Tim Wiltshire, Serge Batalov, et al. A Gene Atlas of the Mouse and Human Protein-encoding Transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(16):6062–7, 2004.

[178] Yoh-hei Takahashi, Gerwin H Westfield, Austin N Oleskie, et al. Structural Analysis of the Core COMPASS Family of Histone H3K4 Methylases from Yeast to Human. *Proceedings of the National Academy of Sciences of the United States of America*, 108(51):20526–31, 2011.

[179] Akaysha C Tang, Barak a Pearlmutter, Michael Zibulevsky, and Scott a Carter. Blind Source Separation of Multichannel Neuromagnetic Responses. *Neurocomputing*, 3233:1115 – 1120, 2000.

[180] Daniel F Tardiff, Nathan T Jui, Vikram Khurana, et al. Yeast Reveal a "Druggable" Rsp5/Nedd4 Network That Ameliorates α-synuclein Toxicity in Neurons. *Science*, 342(6161):979–83, 2013.

[181] Shaolei Teng, Jack Y Yang, and Liangjiang Wang. Genome-wide Prediction and Analysis of Human Tissue-selective Genes Using Microarray Expression Data. *BMC Medical Genomics*, 6 Suppl 1:S10, January 2013.

[182] Sandra Tenreiro, Matthias C Munder, Simon Alberti, and Tiago F Outeiro. Harnessing the Power of Yeast to Unravel the Molecular Basis of Neurodegeneration. *Journal of Neurochemistry*, 127(4):438–52, 2013.

[183] I. Tirosh, B. Izar, S. M. Prakadan, et al. Dissecting the Multicellular Ecosystem of Metastatic Melanoma by Single-cell RNA-seq. *Science*, 352(6282):189–196, April 2016.

[184] a H Tong, M Evangelista, a B Parsons, et al. Systematic Genetic Analysis with ordered Arrays of Yeast Deletion Mutants. *Science*, 294(5550):2364–8, 2001.

[185] Amy Hin Yan Tong, Guillaume Lesage, Gary D Bader, et al. Global Mapping of the Yeast Genetic Interaction Network. *Science*, 303(5659):808–13, 2004.

[186] Cole Trapnell. Defining Cell Types and States with Single-cell Genomics. *Genome Research*, 25(10):1491–1498, October 2015.

[187] Tsung-Han Chan, Wing-Kin Ma, Chong-Yung Chi, and Yue Wang. A Convex Analysis Framework for Blind Separation of Non-Negative Sources. *IEEE Transactions on Signal Processing*, 56(10):5120–5134, October 2008.

[188] Y Tu, G Stolovitzky, and U Klein. Quantitative Noise Analysis for Gene Expression Microarray Experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 99(22):14031–14036, 2002.

[189] R P Tucker. The Roles of Microtubule-associated Proteins in Brain Morphogenesis: A Review. *Brain Research. Brain Research Reviews*, 15(2):101–20, 1990.

[190] Nurcan Tuncbag, Alfredo Braunstein, Andrea Pagnani, et al. Simultaneous Reconstruction of Multiple Signaling Pathways Via the Prize-collecting Steiner forest Problem. *Journal of Computational Biology: A Journal of Computational Molecular Cell biology*, 20(2):124–36, 2013.

[191] Nurcan Tuncbag, Scott McCallum, Shao Shan Carol Huang, and Ernest Fraenkel. SteinerNet: A Web Server for Integrating 'omic' Data to Discover Hidden Components of Response Pathways. *Nucleic Acids Research*, 40:1–5, 2012.

[192] P Uetz, L Giot, G Cagney, et al. A Comprehensive Analysis of Protein-protein Interactions in Saccharomyces Cerevisiae. *Nature*, 403(6770):623–7, 2000.

[193] M. Uhlen, L. Fagerberg, B. M. Hallstrom, et al. Tissue-based Map of the Human Proteome. *Science*, 347(6220):1260419–1260419, January 2015.

[194] Krithika Vaidyanathan and Lance Wells. Multiple Tissue-specific Roles for the O -GlcNAc Post-translational Modification in the Induction of and Complications Arising from Type II Diabetes. *Journal of Biological Chemistry*, 289(50):34466–34471, 2014.

[195] K Van Deun, H Hoijtink, L Thorrez, et al. Testing the Hypothesis of Tissue Selectivity: The Intersection-Union Test and a Bayesian Approach. *Bioinformatics (Oxford, England)*, 25(19):2588–94, 2009.

[196] Vladimir Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[197] Juan M Vaquerizas, Sarah K Kummerfeld, Sarah Teichmann, and Nicholas Luscombe. A Census of Human Transcription Factors: Function, Expression and Evolution. *Nature Reviews. Genetics*, 10(4):252–263, 2009.

[198] Karin Vargova, Nikola Curik, Pavel Burda, et al. MYB Transcriptionally Regulates the MiR-155 Host Gene in Chronic Lymphocytic Leukemia. *Blood*, 117(14):3816–25, April 2011.

[199] Carlos Vaya, José J. Rieta, César Sanchez, and David Moratal. Convolutive Blind Source Separation Algorithms Applied to the Electrocardiogram of Atrial Fibrillation: Study of Performance. *IEEE Transactions on Biomedical Engineering*, 54(8):1530–1533, 2007.

[200] D. Venet, F. Pecasse, C. Maenhaut, and H Bersini. Separation of Samples Into their Constituents Using Gene Expression Data. *Bioinformatics*, 17(Suppl 1):S279–S287, June 2001.

[201] Annelien Verfaillie, Hana Imrichova, Zeynep Kalender Atak, et al. Decoding the Regulatory Landscape of Melanoma Reveals TEADS As Regulators of the Invasive Cell State. *Nature Communications*, 6:6683, 2015.

[202] Silas G Villas-Bôas, Joel F Moxley, Mats Akesson, Gregory Stephanopoulos, and Jens Nielsen. High-throughput Metabolic State Analysis: The Missing Link in Integrated Functional Genomics of Yeasts. *The Biochemical Journal*, 388(Pt 2):669–77, 2005.

[203] E. Villeneuve and H Carfantan. Hyperspectral Data Deconvolution for Galaxy Kinematics with MCMC. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pages 2477–2481, August 2012.

[204] Emmanuel Vincent, Nancy Bertin, Rémi Gribonval, and Frédéric Bimbot. From Blind to Guided Audio Source Separation: How Models and Side Information Can Improve the Separation of Sound. *IEEE Signal Processing Magazine*, 31(3):107–115, May 2014.

[205] Bo Wang, Junjie Zhu, Emma Pierson, Daniele Ramazzotti, and Serafim Batzoglou. Visualization and Analysis of Single-cell RNA-seq Data by Kernel-based Similarity Learning. Technical report, bioRxiv, May 2016.

[206] Liangjiang Wang, Anand K Srivastava, and Charles E Schwartz. Microarray Data Integration for Genome-wide Analysis of Human Tissue-selective Gene Expression. *BMC Genomics*, 11 Suppl 2:S15, January 2010.

[207] Niya Wang, Eric P. Hoffman, Lulu Chen, et al. Mathematical Modelling of Transcriptional Heterogeneity Identifies Novel Markers and Subpopulations in Complex Tissues. *Scientific Reports*, 6:18909, January 2016.

[208] Xiujuan Wang, Xiaomu Wei, Bram Thijssen, et al. Three-dimensional Reconstruction of Protein Networks Provides Insight Into Human Genetic Disease. *Nature Biotechnology*, 30(2):159–164, 2012.

[209] C Widmann, S Gibson, M B Jarpe, and G L Johnson. Mitogen-activated Protein Kinase: Conservation of A Three-kinase Module from Yeast to Human. *Physiological Reviews*, 79(1):143–80, January 1999.

[210] Stacey S Willard and Shahriar Koochekpour. Glutamate, Glutamate Receptors, and Downstream Signaling Pathways. *International Journal of Biological Sciences*, 9(9):948–59, January 2013.

[211] E Winzeler, D Shoemaker, A Astromoff, et al. Functional Characterization of the S. Cerevisiae Genome by Gene Deletion and Parallel Analysis. *Science*, 285(5429):901–6, 1999.

[212] John C Wootton and Scott Federhen. Statistics of Local Complexity in Amino Acid Sequences and Sequence Databases. *Computers & Chemistry*, 17(2):149–163, 1993.

[213] Chen Xu and Zhengchang Su. Identification of Cell Types from Single-cell Transcriptomes Using a Novel Clustering Method. *Bioinformatics (Oxford, England)*, 31(12):1974–80, June 2015.

[214] Takao Yagi, Byoung Boo Seo, Eiko Nakamaru-Ogiso, et al. Can a Single Subunit Yeast NADH Dehydrogenase (Ndi1) Remedy Diseases Caused by Respiratory Complex I Defects? *Rejuvenation Research*, 9(2):191–7, January 2006.

[215] Jaewon Yang and Jure Leskovec. Community-Affiliation Graph Model for Overlapping Network Community Detection. In *12th IEEE International Conference on Data Mining, ICDM 2012, Brussels, Belgium, December 10-13, 2012*, pages 1170–1175, 2012.

[216] Esti Yeger-Lotem and Roded Sharan. Human Protein Interaction Networks Across Tissues and Diseases. *Frontiers in Genetics*, 6(August):1–5, 2015.

[217] Haiyuan Yu, Nicholas M Luscombe, Hao Xin Lu, et al. Annotation Transfer Between Genomes: Protein-protein Interologs and Protein-DNA Regulogs. *Genome Research*, 14(6):1107–18, 2004.

[218] Meng Yu, Wenye Ma, Jack Xin, and Stanley Osher. Multi-Channel L1 Regularized Convex Speech Enhancement Model and Fast Computation by the Split Bregman Method. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(2):661–675, 2012.

[219] A. Zeisel, A. B. M. Manchado, S. Codeluppi, et al. Cell Types in the Mouse Cortex and Hippocampus Revealed by Single-cell RNA-seq. *Science*, 347(6226):1138–42, March 2015.

[220] Kun Zhang and Aapo Hyvärinen. Source Separation and Higher-Order Causal Analysis of MEG,and EEG. *CoRR*, abs/1203.3533, 2012.

[221] Liqing Zhang and Wen-Hsiung Li. Mammalian Housekeeping Genes Evolve More Slowly Than Tissue-specific Genes. *Molecular Biology and Evolution*, 21(2):236–9, 2004.

[222] Yi Zhong, Ying-Wooi Wan, Kaifang Pang, Lionel Chow, and Zhandong Liu. Digital Sorting of Complex Tissues for Cell Type-specific Gene Expression Profiles. *BMC Bioinformatics*, 14:89, January 2013.

[223] Yeping Zhou, Rou-Afza F Gunput, and R Jeroen Pasterkamp. Semaphorin Signaling: Progress Made and Promises Ahead. *Trends in Biochemical Sciences*, 33(4):161–70, 2008.

[224] H Zhu, M Bilgin, R Bangham, et al. Global Analysis of Protein Activities Using Proteome Chips. *Science*, 293(5537):2101–5, 2001.

[225] Neta S Zuckerman, Yair Noam, Andrea J Goldsmith, and Peter P Lee. A Self-directed Method for Cell-type Identification and Separation of Gene Expression Microarrays. *PLOS Computational Biology*, 9(8):e1003189, January 2013.

VITA

VITA

Shahin Mohammadi was a computer science student at Purdue from 2010 to 2016. He got his M.S. and Ph.D. degrees in Dec 2012 and 2016, respectively. His research interests include computational biology, machine learning, and parallel computing. His work spans different areas of bioinformatics/systems biology and aims to develop computational methods coupled with statistical models for data-intensive problems, with application in mining the cell type-specific transcriptome and interactome.