RETICULA

A PROJECT TO IMPROVE REACTOME TISSUE SPECIFICITY VIA MACHINE LEARNING APPROACHES USING RNA-SEQ DATA

By

Joshua Garrison Burkhart, M.Sc.

A DISSERTATION

Presented to
the Department of Medical Informatics & Clinical Epidemiology
&
the Oregon Health & Science University School of Medicine
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

June 2021

School of Medicine
Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the Ph.D. dissertation of

Joshua Garrison Burkhart

has been approved

Shannon McWeeney, Ph.D., Mentor

Melissa Wong, Ph.D., Chair

Guanming Wu, Ph.D., Member

Xubo Song, Ph.D., Member

Francesco Raimondi, Ph.D., Member

TABLE OF CONTENTS

Abstract

The functional heterogeneity of human tissues—despite their descendance from a single progenitor cell—poses a substantial complication to interpretation of tissue developmental molecular studies and tissue-specific diseases, such as cancer. This complication significantly impedes the development of precision therapies for individual patients. To address this shortcoming, I have reviewed the major RNA-seq tissue expression data repositories, transformed mRNA transcript counts into coordinates within biochemical reactions in which their protein products participate and generated a human tissue-specific data resource containing 9,115 human tissue samples across 10,726 reactions. This data transformation recapitulates the sample structure found within mRNA transcript abundance values while positioning human tissue samples within the human Reactome preceding/following reaction network and Reactome pathway hierarchy. This additional contextualization enables novel insights using existing reaction and pathway annotations as well as network analytical techniques at the reaction level such as network-based classifiers. I demonstrate the advantage of this approach by executing a novel pathway enrichment method and generating geometric deep learning architectures representative of biological networks, training them to accurately classify human tissue samples in a way that may be described by biochemistry. By investing my trained models, I generate new human tissue hierarchies and find the Reactome preceding/following reaction network exhibits non-random topology, suggesting a novel signature of evolution is embedded within biochemical reaction relationships themselves.

Chapter 1: Introduction and Background

Introduction

While it is known that human tissues harbor characteristic gene expression patterns, gene products exhibit complex relationships with cellular behaviors which hampers straightforward extrapolation from gene expression values to the biochemical reactions carrying out tissues' functions. Unfortunately, even with the considerable advances of contemporary omics-based biological studies, this complexity remains largely hidden from us. Achieving high confidence for which reactions are likely to occur within particular tissue contexts is paramount in order to understand the biochemical mechanisms of tissue development, tissue-specific functions and the etiology of tissue-specific disease.

Within a tissue, proteins, small molecules and other types of molecules participate in interactions and react with each other to form a network. Different tissues are likely characterized by their own discrete reaction networks. Prior decades have seen many large-scale, tissue-specific gene expression datasets generated, processed and stored in multiple data repositories. I hypothesize I can infer tissue-specific reaction networks by using large-scale gene expression data, thereby contributing to the understanding of tissues' functional heterogeneity. To test this, I will leverage the human-annotated biochemical reactions in Reactome, the most comprehensive open source pathway knowledgebase, and apply classical and novel statistical methods and machine learning strategies to connect biochemical reactions with particular tissues.

Background

My research sought to expand our understanding of healthy human tissue function by identifying components of biological networks that exhibit differences among them. If models of human tissue function are able to mechanistically explain specific variations observed across the human population, subtle deviations from healthy function may become detectable at the individual patient level, enabling precise medical intervention.

Precision medicine is a healthcare model that has shown promise in treating cardiovascular disease (Dainis and Ashley, 2018), cystic fibrosis (Manfredi et al., 2019), obesity (Frühbeck et al., 2018), and cancer (Arnedos et al., 2015, Schwaederie et al., 2015, Friedman et al., 2015, Vargas et al., 2016, Dienstmann et al., 2017 and Galle et al., 2018) by considering individual patients' disease characteristics (Ashley, 2015 and Ashley, 2016). A classic example of this approach is patient stratification by ABO blood group when considering blood transfusion while a more contemporary example is classification of the most common type of liver cancer, hepatocellular carcinoma (HCC). HCC has been classified into stages according to the Barcelona-Clinic Liver Cancer staging system whereby patients are classified into one of stage 0, A, B, C or D based on established prognostic criteria (Galle et al., 2018) and each treated according to their stage. Patients diagnosed with stages 0 or A respond to resection, liver transplantation and local ablation while those diagnosed with stage D are instead treated with nutritional, psychological and pain management support (Llovet et al., 2018). Considered together, clinical implementations of precision medicine effect dramatic change in patients' lives where this strategy—across many cancers—has demonstrated increased median response rate, progression free survival and overall survival rates (Schwaederie et al., 2015). This model holds increasing promise as our detection techniques, data collection, integration and analytical methods improve and I better understand the similarities and differences both among and within our bodies. The human population exhibits phenotypic variation as a result of both environmental and genetic differences. Human tissues and even individual cells display phenotypic variation owing largely to their lineage and microenvironment. A key to maximizing the potential benefits of precision medicine is addressing the question, "Given the degree of expected variation, how do I best categorize various disease presentations?".

Considering the introduction of high-throughput cellular and molecular experimental techniques such as flow cytometry and single-cell sequencing, one answer to this question is to achieve a superior understanding of both the etiology and teleology of cellular behavior by developing high-fidelity models of biological networks. Though entire biological networks may never be exactly duplicated in nature, some parts are typified by general patterns and amenable to inference; functional components, after identification and annotation, may be repeatedly observed and eventually mechanistically understood.

Cell behavior is dictated by the state of a cell's biochemical reactions, currently unobservable to us, though a ready measurement I may collect from a cell is its pattern of gene expression. Using RNA-sequencing (Mortazavi et al., 2008), experimentalists are able to approximate the abundance of mRNA present in a cell or group of cells. Though many biochemical reactions involve protein-protein interactions and not interactions of mRNA, specific mRNA synthesis is required for specific protein synthesis and thus mRNA patterns are associated with protein-regulated phenotypes and cell states. Total mRNA abundance is clearly correlated with protein abundance ($R^2 = 0.41$ on log-log scale and $R^2 = 0.44$ following nonlinear transformation, Schwanhäusser et al, 2011) and, when approximate steady state conditions are met, protein abundance has been shown to be primarily determined by mRNA abundance (Liu et al., 2016) with between 56% and 84% of variation in protein abundance shown to be explained by mRNA abundance (Li et al., 2014). However, there may not be direct correspondence between a protein's abundance and its availability to participate in a biochemical reaction, for example, due to competitive occupancy among binding partners or post translational modifications (PTMs). Despite these barriers, I may be able to infer which biochemical reactions are characteristic to particular tissues by identifying repeated patterns among the mRNA transcripts coding for proteins that participate in biochemical reactions, such as collections of transcripts whose products participate the same reaction tending to be expressed at similar levels for particular tissues. I find it reasonable to suggest such patterns of mRNA that associate with particular tissues imply patterns of protein abundance and biochemical reaction states characteristic of those tissues.

Cell phenotypes are defined from myriad perspectives and, at more elementary units of analysis may be considered cell lines, cell types, cell populations or, more abstractly, clusters of similar cells determined by some (here undefined) cell attributes; at more composite units of analysis, cell phenotypes may be thought of as tissue types or disease states.

Though both more elementary and more composite perspectives are of staggering importance to translational medical research, three considerations suggest the latter as currently more amenable to our intention of conducting systematic analysis. First, observations may be captured via so-called *bulk RNA-sequencing*, a technology that's been around longer (Mortazavi et al., 2008) than those required for observations of more elementary units of analysis such as *scRNA-sequencing* (see Wang et al., 2009; Metzker, 2010; Ozsolak et al., 2011; Shapiro et al., 2013; Eberwine et al., 2014; Wu et al., 2014; Chen et al., 2018 for apropos review). Assuming a longer period of market availability translates to more thorough debugging, relying on older technology decreases the likelihood of errors entering our analyses from unknown sources. For example, bulk RNA-sequencing datasets are often scale-normalized when drawn from different experiments to account for potential batch effects (Robinson et al., 2010) that may dampen true signals. Second, not unrelated to the period of market availability, many samples annotated at more composite units of analysis have already been systematically collected, processed and made available to the community by both individual investigators and consortia alike. Though a large volume of data has been gathered through scRNA-sequencing and other experimental techniques relating to more elementary units of analysis, to our knowledge, experimental results generated from systematic processing of tissue-specific samples using these techniques with sample sizes approaching those of existing bulk RNA-sequencing repositories are not currently available. Third, cancer has received a tremendous amount of attention in recent years and the realization that its genetic and phenotypic presentation varies among tissues (Blokzijl et al., 2016) has resulted in available datasets containing thousands of bulk RNA-sequencing results from both healthy tissues and tumors, potentiating near-term analytical findings relevant to human health. The choice to rely upon bulk RNA-sequencing is not without its drawbacks. Healthy human tissues themselves are host to a diversity of cell types, each capable of performing the functions arrived at by evolution. I expect this within-tissue cell type heterogeneity will dilute subtle signals but that the large sample sizes available will empower us to identify patterns of the most prominent cell types within human tissues.

Though particular mutations, genes or gene products may appear insignificantly rare when independently considered, the so-called "pathway principle" (Krogan et al., 2015) holds that the inherent organization of molecular biology is such that their individual effects will coalesce into more robust signals, revealing inherent features undetectable from the narrow perspective of individual genes or proteins. This perspective supports the concept that without the contextualization of individual mRNA expression values offered by the structure of biological networks, meaningful signals are scattered about and the hope for consolidating our insights into a systematic comprehension remains futile. For example, within "hallmark" cancer pathways (Hanahan and Weinberg, 2011), mutations tend to be mutually exclusive, with typically no more than a single mutation per tumor in a particular hallmark pathway, which is taken as evidence for diminishing selective advantage of within-pathway mutations (Ciriello et al., 2012). This supports the pathway principle because it shows mutations are not uniformly random; rather, they are more likely to appear in certain places dictated by pathway architecture. The pathway and network perspective can be leveraged to investigate healthy tissue biology because the same pathway architecture responsible for observed mutational biases similarly governs interactions among unmutated proteins and despite their—more or less—identical DNA, tissues remain able to perform specific functions through characteristic pathway activation states (Huttlin, E., et al., 2010). I argue this observed tissue-specific variation in activation states supports the notion that tissues' pathway architectures vary in characteristic ways and set out to describe this variation through pathway architecture-constrained mRNA expression analysis.

The concept of tissue-specific pathway activation states is well-supported by the literature. For example, tissue-specific pathway activation, measured in ratios of total protein copy numbers relative to mRNA transcripts in mammalian cells, vary across tissues: e.g. murine embryonic NH3T3 cell line protein-mRNA abundance ratio estimated at 2,800 (Schwanhausser et al., 2011) and the protein-mRNA abundance ratio of murine liver cells estimated at 10,000 (Azimifar et al., 2014). These differences suggest striking differential regulation mechanisms. Further, there exists differential chemical and structural microenvironments that enable various immune system interactions and act as nidi for wound healing processes but are also implicated in tissue-specific inflammatory responses and tumorigenesis (Coussens and Werb, 2002). Finally, cancer type-specific mutational frequencies, for example, in the Wnt signaling pathway, which has been shown to be mutated in over 90% of core gastrointestinal cancers and under 10% of genitourinary cancers (Sanchez-Vega et al., 2018), further exemplifies its tissue-specific activity.

Tissue-Specific Networks

Recent years have seen a variety of approaches developed to infer tissue-specific networks. For dynamic system models that describe spatiotemporal changes, approaches to resolve biochemical networks commonly consider network refinement a filtering problem, relying upon models incorporating ordinary differential equations (ODEs) (see Klipp et al., 2006; Aldridge et al., 2006; Tyson et al., 2018 and Loskot et al., 2019 for apropos reviews). These models must be parameterized by accurate estimates of initial chemical concentrations and reaction rates using spatiotemporally resolved data, such as time-series mRNA or protein expression samples. The requirement of accurate parameter estimation for ODE strategies introduces their primary drawback as sufficiently accurate estimates of these parameters remain computationally intractable for complex models despite attention from the field (Nakatsui et al., 2010). For example, dynamic models of the ERK cascade (Rubinfeld et al., 2005) consist of around 100 reactions and 100 protein species (Orton et al., 2005).

Static system models lack spatiotemporal resolution but require fewer parameters to construct and thus scale more efficiently to large systems and require lower-complexity experimental procedures (e.g. time-series gene expression assays are not required). For example, Saha et al. (2017) generated tissue-specific co-expression networks consisting of around 15,000 nodes representing mRNA expression levels and isoform ratios and around 60,000 edges. Static biological networks have been studied using a variety of approaches typically relying upon machine learning/statistical inference. In addition to Saha et al., contemporary examples of studies investigating static biological networks include the construction of a within-tissue gene expression correlation network integrated with existent transcription regulatory networks (Sonawane et al., 2017), Bayesian network generation using Monte Carlo sampling of ensemble simulations (Gendelman et al., 2017), and Bayesian inference relying upon manually curated human proteins integrating with manually curated tissue ontologies (Greene et al., 2015).

Sonawane et al. (2017) constructed tissue-specific networks by combining transcription regulatory (Weirauch et al., 2014) and the protein-protein interaction (Szklarczyk et al., 2015) networks using PANDA, an acronym for Passing Attributes between Networks for Data Assimilation (Glass et al. 2013) that was designed to predict regulatory relationships by integrating multiple biological data sources. Tissue specificity was called based on comparing Jaccard indices of each gene's expression distributions with its connected partners within and among tissues. Inferences regarding tissue function were drawn both from network centrality measures and gene set enrichment analysis. Among their conclusions was the suggestion that tissue-specific transcription regulation occurs at an "intermediate scale" whereby tissue-specific genes tend to be more centrally located in regulatory pathways, rather than by exerting control on dense subnetworks or network hubs. I argue this finding suggests focusing analysis on particular genes or gene products—even highly connected ones—will never yield a satisfactory understanding of differential tissue function and that this intermediate level of abstraction is captured only by carefully expanding focus and constraining analysis with a network perspective.

An investigation of oncogenic breast tissue-specific signaling networks (Gendelman et al., 2017) was conducted using a strategy that calculated Bayesian scores for small subnetworks, constructed ensembles of these small subnetworks, and used a Monte Carlo method to select those consistent with data gleaned from breast cancer cell lines using microarrays. Scoring small subnetworks was conducted by considering all possible combinations of two or three variables considering 149 significant and 20 known associated genes. Functional interpretation of their results was based on gene set enrichment analysis and concordance with prior literature. The unconstrained search used in this approach does not scale well, limiting the size of small subnetworks searched and the number of genes considered. Furthermore, the generative nature of this approach disallows contextualization within a broader network, ignoring potential insights granted by previously reported biological network structures and annotations.

Greene et al. generated tissue-specific networks by first integrating several interaction datasets and refining it using gene expression data from GEO along with annotations from the BRENDA Tissue Ontology (BTO) (Gremse et al., 2010) and other sources, such as the Human Protein Reference

Database (HPRD) (Keshava Prasad et al., 2008; Greene et al., 2015). Their initial interaction network was placed within the BTO tissue/cell-type hierarchy which allowed for interaction network nodes representing genes to be separated into three classes. A class they labeled T' consisted of genes annotated to tissues judged to be "unrelated" based on BTO category. A class labeled U consisted of genes found to be either "widely expressed" in a prior tissue-specific gene expression study, genes for proteins found to be expressed in >75% of tissues assayed in the human protein atlas (Uhlen et al., 2017) or curated "ubiquitous genes" from HPRD. The final class, labeled T consisted of genes annotated to a tissue in the BTO and in neither T' nor U. The interaction network edges were then stratified into four classes, inferring functional roles only within the class where edges connected either an element in T to another element in T or connected an element in T to an element in U. Tissue-specific weights were learned by assessing coexpression within each class using a Bayesian (tissue) classifier. Though this strategy of constraining search space to within each of the four edge classes is protective against multiple hypothesis testing, it does not facilitate discovery of more complex relationships involving dissimilar tissues or more than two genes. To interpret their results, the authors considered results with marginal statistical significance from GWAS studies; this is a novel methodology utilised for functional inference validation in their project. However, neither their interaction network nor the BTO correspond directly to biochemical reactions or annotated pathways, limiting mechanistic insights.

Machine Learning

The aforementioned attempts to develop and refine tissue-specific biological networks use generative or *bottom-up* approaches; however, recent work has shown promise in discriminative or *top-down* network inference (Ma et al. 2018). Such analysis strategies consider the set theoretical union of connections across all comparable biological constructs, relying upon statistical relationships between molecular data and observed phenotype, restricting degrees of freedom using relationships specified by said network. I briefly review machine learning that I may express this concept more thoroughly.

Machine learning may be broadly divided into two categories: unsupervised learning, which applies labels to unlabeled data based on comparing and clustering element attributes and supervised learning, which operates on labeled data to identify patterns of features characteristic to particular labels.

Early on, unsupervised clustering of gene expression data was used to identify temporal mRNA patterns among human and yeast cells (Eisen et al., 1998) and shortly thereafter was used for comparing human tumors with normal colon tissue (Alon et al., 1999). The same year, supervised learning was applied to leukemia subtype identification, separating acute myeloid leukemia (AML) from acute lymphoblastic leukemia (ALL) samples (Golub et al., 1999). Clinically distinct subgroups of large B cell lymphoma and hepatitis B virus-positive metastatic hepatocellular carcinoma specimens were discovered using supervised learning on gene expression data (Wright et al., 2003 and Ye et al., 2003), representing early contributions of gene expression analysis and machine learning to medicine.

Machine learning has become an increasingly prominent analytical technique in biology and medicine as data collection, storage and processing methods improve and society further advances into the information age (see Kourou et al., 2015; Mamoshina et al., 2016; Camacho et al., 2018 and Zitnik et al., 2019 for apropos reviews). An ongoing streak of successful development in the field has imparted general strategies such as incorporating subject matter expertise into feature selection procedures (Costello et al., 2014 and Bilal et al., 2013), sharing features among predicted variables (Costello et al., 2014 and Eduati et al., 2015), testing for both linear and nonlinear relationships between features and predicted variables (Costello et al., 2014 and Eduati et al., 2015) and combining models into ensembles (Bilal et al., 2013).

A machine learning strategy performing discrimination on biological networks is the DCell Visible Neural Network (VNN) generated by Ma et al. (2018), whose architecture was constructed to represent both the Gene Ontology (GO) (Gene Ontology Consortium, 2016) and Clique-extracted Ontology (CliXO) (Kramer et al., 2014). This work resulted in networks containing 97,181 and 22,167 artificial neurons, respectively, representing 2,526 cellular subsystems. This project acts as a touchstone for performing nonlinear inference on biological networks noting the opacity of fully-connected artificial neural network (ANN) models and the fervent desire for model interpretability. With this system, several million yeast single- and double-KO samples from TheCellMap.org (Van Leeuwen et al., 2016; Costanzo et al., 2016 and Usaj et al., 2017), a data repository for yeast genetic interactions, were used to infer the biochemical phenomena underlying both growth rate and resistance to ultraviolet radiation phenotypes and were able to assess whether such phenotypes were primarily the outcomes of a only few cell subsystems or many, finding—for these phenotypes—that ~20% of the subsystems confer ~80% of the importance.

Reactome

For the purpose of modeling human tissue biochemistry, the preeminent network is provided by the Reactome preceding/following reaction network. A hybrid portmanteau of Latin *react-* "done again" and Greek *-some* "body" partial cognate *-ome*, *reactome* may be defined at face value as "repertoire of chemical reactions" but is typically used to refer to the Reactome Pathway Knowledgebase (Fabregat et al., 2017 and Jassal et al., 2020), currently, the most comprehensive and popular open access, manually curated biological pathway knowledgebase. Reactome is freely available online at https://www.reactome.org; exemplary visualizations from the Reactome webpage are shown in Figure 1. The most recent version at the time of this writing (release 77, June 2021) contains 11,084 proteins, 13,662 complexes, 13,827 reactions and 2,536 pathways covering over 50% of total human protein-coding genes and supported by 30,000 published papers. Human pathways in Reactome are manually curated by experienced curators and peer-reviewed by expert reviewers ensuring consistent high quality. Reactome includes annotations that map genes, gene products and other reactants to biochemical reactions and pathways as well as annotations connecting biochemical reactions to each other, forming a reaction network structure. In addition to relationships among reactions, Reactome provides pathway annotations, organized in a hierarchical way based on *containing/contained by* relationships. These pathway annotations and reaction networks are provided for multiple species, including human; however, it does not contain tissue-specific annotations.

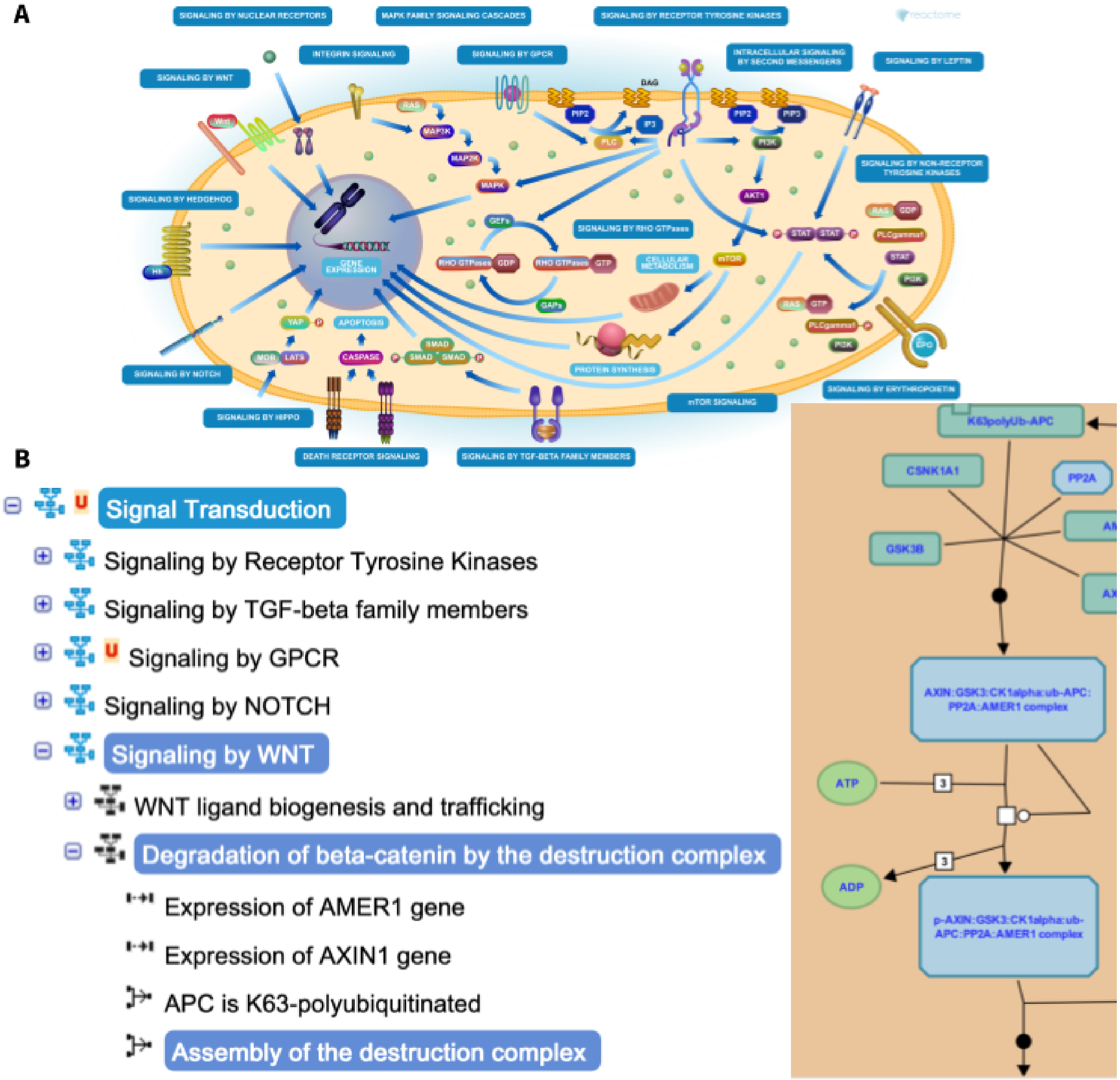

A
SIGNALING BY NUCLEAR RECEPTORS
MAPK FAMILY SIGNALING CASCADES
SIGNALING BY RECEPTOR TYROSINE KINASES
reactome
INTEGRIN SIGNALING
SIGNALING BY GPCR
INTRACELLULAR SIGNALING BY SECOND MESSENGERS
SIGNALING BY LEPTIN
SIGNALING BY WNT
SIGNALING BY NON-RECEPTOR TYROSINE KINASES
SIGNALING BY HEDGEHOG
SIGNALING BY RHO GTPases
GENE EXPRESSION
CELLULAR METABOLISM
PROTEIN SYNTHESIS
APOPTOSIS
SIGNALING BY NOTCH
SIGNALING BY HIPPO
DEATH RECEPTOR SIGNALING
SIGNALING BY TGF-BETA FAMILY MEMBERS
mTOR SIGNALING
SIGNALING BY ERYTHROPOIETIN
Wnt
RAS
MAP3K
MAP2K
MAPK
PIP2
IP3
PLC
DAG
PIP3
PI3K
AKT1
mTOR
STAT
PLCgamma1
GEFs
GAPs
RHO GTPases
GDP
GTP
YAP
MOB
LATS
CASPASE
SMAD
Hh
EPO
B
Signal Transduction
Signaling by Receptor Tyrosine Kinases
Signaling by TGF-beta family members
Signaling by GPCR
Signaling by NOTCH
Signaling by WNT
WNT ligand biogenesis and trafficking
Degradation of beta-catenin by the destruction complex
Expression of AMER1 gene
Expression of AXIN1 gene
APC is K63-polyubiquitinated
Assembly of the destruction complex
K63polyUb-APC
CSNK1A1
PP2A
GSK3B
AXIN:GSK3:CK1alpha:ub-APC:PP2A:AMER1 complex
ATP
ADP
3
3
p-AXIN:GSK3:CK1alpha:ub-APC:PP2A:AMER1 complex

Figure 1: The Reactome knowledgebase is an academically-developed open source database of manually-curated biological pathways that describe biochemical reactions with mechanistic detail. **a**) A cell diagram from Reactome where pathways are labeled with blue boxes about the perimeter of the cell and their constituents and relationships are shown within. **b**) The curated hierarchy considers biochemical reactions and their organization into pathways. Biochemical reactions have relationships with lower-level pathways, which have relationships with higher-level pathways. **c**) The reactant-product network considers the

reactants, catalysts, inhibitors, activators and products of biochemical reactions and connects biochemical reactions when these are shared.

Current biological network analysis strategies leveraging the Reactome tend to use the entire network, including some biochemical relationships thus far experimentally confirmed only in model organisms, such as yeast. Limited by current cost-intensive and low-throughput experimental techniques, human tissues' biochemical reaction networks have not been specifically described; rather, the Reactome contains a single manually-curated human reaction network that describes the superset of reactions thought to occur with high confidence in human cells.

Hypothesis Development

We have abundant evidence about which proteins and small molecules are able to participate in which lower-level reactions and which of those combine to form higher-levels, such as pathways; however, I have incomplete comprehension of two important aspects of this network. First, it is unclear how the microenvironment of reaction participants affect the reaction output. For example, a reaction whose stoichiometry requires an equal number of two proteins to occur may not occur as expected despite equal concentrations of those proteins in the same cellular compartment if those proteins' structural stability, folding or mutual interaction are affected by and subject to differential temperatures or salt concentrations (Baldwin, 1986; Formaneck et al., 2006; Pegram et al., 2010; Zhang, 2012; Reinhard et al., 2015). Further, it is not clear which proteins, small molecules and lower-level reactions are shared among particular tissues or cell types.

Phosphoproteomics research has discovered robust patterns of cell signaling and pathway activation—as judged by protein phosphorylation—associated with tissue-development early in human embryonic stem cell differentiation (Van Hoof et al., 2009) and across human, mouse and rat tissues (Changelian and Fearon, 1986; Huttlin et al., 2010; Karabulut and Frishman, 2016). This suggests network patterns would likely emerge from comprehensive human tissue-specific protein expression assays. Unfortunately, current experimental techniques do not allow us to efficiently observe proteome-wide abundance directly, which requires us to find strategies with which to interrogate tissue-specific biochemical reaction networks by measuring other biological entities. Currently, mRNA profiling appears the most promising avenue into approximating system-wide biochemical reaction network states. Consider, for example, it has been shown that, when cellular steady-state conditions are met or may be sufficiently approximated, assuming neither differentiation nor stress response, protein abundance levels have been shown to be primarily determined by mRNA abundance levels (Liu et al., 2016). Moreover, RNA-sequencing of like tissue has already been shown to reveal distinctive patterns (Pierson et al., 2015; Saha et al., 2017; Sonawane et al., 2017; Söllner et al., 2017, Hanson et al., 2017). Furthermore, the prevalence of gene expression signatures predicting pathway activities in the literature (Bild et al., 2006; Phillips et al., 2006; Lenz et al., 2008; Eroles et al., 2012; Ceccarelli et al., 2016 and Ayers et al., 2017) suggests pathway activity outcomes, such as signal transduction or cell cycle phase transition, are detectable via antecedent gene expression patterns alone, despite largely hidden effects of metabolite and other small molecule gradients, requisite traversal mechanisms of cellular compartmental boundaries, post-translational modifications, RNA and protein degradation rates et cetera. Considering this, our hypothesis is that gene expression signals coalesce across biochemical reactions to form distinct biological network patterns that stratify human tissues and reveal tissue-specific mechanistic details underlying their functional differences. Extrapolation from mRNA profiling data—already shown to separate tissues—strikes us as promising and worthwhile investigating from our proposed reactionwise perspective.

In summary, I find our proposed analysis strategy follows naturally from our premises:

1) Tissues exhibit characteristic protein expression patterns.
2) Protein expression patterns predict pathway activation.
3) Tissues exhibit characteristic mRNA expression patterns.
4) mRNA expression patterns predict pathway activation.
5) Protein abundance at steady state conditions is primarily determined by mRNA abundance.
6) Cadaver tissue is at a steady state.

Aims Overview

In order to determine to which degree gene expression profiles communicate biochemical reaction network states that grant insight into human tissue functions, I tested our hypothesis by pursuing two specific aims. In our first aim, I performed an analytical review of relevant publicly available data sources along with exploratory data analysis, quality filtering, normalization and other preprocessing steps common to RNA-sequencing analysis pipelines and necessary for generalizing findings to new datasets. In this aim, I conducted gene-based clustering analysis in order to review our batch-effect corrections and test for biochemical reaction and network-free associations between human gene expression and human tissue identity. In our second aim, I developed a machine learning/graph analytical strategy with which to construct tissue-specific networks and infer the relative importance of particular biochemical reactions and pathways among human tissues.

Publicly available human tissue-specific RNA-sequencing datasets typically apply pipeline-specific processing, normalization and batch-effect reduction strategies that lead to challenges for their collective integration. After assessing batch effects and other discrepancies across several data sources, I focused on healthy human tissue samples from GTEx alone. Briefly, I arrived at our dataset by downloading reprocessed samples from Recount2 and normalizing them using the DESeq2 library, combining and storing results in a single large data matrix (See Chapter 2 for additional details). In order to maximize community impact, I provide all intermediate data results and a final resource with identifiers matching Reactome, enabling straightforward integration of tissue-specific gene expression datasets with Reactome's web and Cytoscape applications. This aim resulted in a single data repository containing a 9,115 healthy human tissue-specific RNA-sequencing samples across 51 tissues along with software to reproduce its derivation and our analysis. I then used this data resource to group gene expression values according to the reactions in which their protein products participate in order that representative reactions may then be used as features in downstream machine learning techniques in order to explain particular tissues from a biological networks perspective.

The reactions characterized in the first aim are connected separately within both the Reactome pathway hierarchy and the Reactome preceding/following reaction network such that parsimonious tissue-specific networks were discovered and—to within a reasonable degree—their elements contextualized. Quantitative assessment of machine learning algorithm performance with held-out test error are provided and contribute to final result validation. For further validation of our final results, across tissue-specific networks, I performed hierarchical clustering based on input data and network similarity to construct several tissue clusterings and compared them among each other and with prior literature.

In order to consider gene expression values from a reactionwise perspective, I used the data resource generated by our first aim to exploit the many to many gene-reaction relationships manually annotated in Reactome. This work resulted in a set of tissue-specific reaction networks, which may enhance our understanding of healthy human tissue, development and cancer. Additionally, this work demonstrates a computational approach that may be extended to infer cell-type specific reaction networks and pathways based on single cell gene expression data, further improving our understanding of life and disease.

Chapter 2: Harmonize and Integrate Human Tissue-specific RNA-seq Data

The application of experimental results to a clinical setting is stymied without tissue-specific context for biological pathway and network modeling. To address this, I have processed and reviewed publicly available RNA-seq data from several sources, ultimately opting to use data from The Genotype-Tissue Expression (GTEx) project, the most comprehensive human tissue-specific gene expression data repository. I contextualized this data with annotations from Reactome, the most comprehensive open access, manually curated pathway knowledgebase. I transformed mRNA transcript counts into coordinates within spaces defined by biochemical reactions in which their protein products participate and generated a human tissue-specific data resource containing a total of 9,115 human tissue samples across 10,726 reactions. This data transformation preserves the tissue sample structure observed across mRNA transcript abundance values while positioning samples within the Reactome preceding/following reaction network. This additional contextualization enables novel insights using extant Reactome reaction and pathway annotations as well as network analytical techniques at the reaction level such as graph-based machine learning strategies. Additionally, I validate our transformation with significant convergence of pathway enrichment results and follow with a demonstrative analysis of tissue proliferation at the pathway, reaction and gene transcript level.

A prime motivation for developing large repositories of tissue-specific gene expression profiles is to improve our understanding of tissue-specific behavior by associating variation of gene expression with that of cellular function. One of the considerations in performing such association studies is the notion that gene products often act in concert when participating in biochemical reactions and that these gene product relationships themselves vary among cell and tissue types. Thus, by investigating the relationship between tissue-specific gene expression and the states of biochemical reactions, I may be able to improve our understanding of cell and tissue type variation in both transcriptomic and functional regards.

To illustrate this point with a more concrete motivational narrative, I imagine a case where treatment and control groups have been shown to have significant differential gene expression for several genes and moderate or low differential gene expression for others. I would be interested to know which biochemical reactions differ between treatment and control groups but find none of the reactions in which the significantly differentially expressed genes participate exhibit significant differences. In fact, the reactions which differ most significantly, in this case, involve only gene products from genes which exhibit moderate or low expression differences. These reactions could not possibly be recovered via enrichment analysis of significantly differentially expressed genes; indeed, to reveal these reactions, samples' gene expression values must be transformed and considered in a reactionwise manner. As this narrative is merely theoretical; to investigate systems-level gene expression and reaction state information asymmetry, I looked to prominent human tissue-specific RNA-seq data sources for this study and investigated batch effects and other undesired variation in an effort to remove issues that may mislead or otherwise preclude reactionwise insights.

We review data sources maintained by the National Center for Biotechnology Information (NCBI) that catalogue information from disparate experiments. Next I describe large-scale studies of *ex vivo* human tissue samples organized by international consortia. And finally, I discuss results of aggregated reprocessed data sources using combinations of the above.

NCBI Data Sources

The Gene Expression Omnibus (GEO) (Edgar et al., 2002; Barrett et al., 2012) and Sequence Read Archive (SRA) (Leinonen et al., 2010) are two popular resources freely available for academic use without constraints. They act as primary repositories for functional genomics and biological sequence data and each hold tens of thousands of human RNA-sequencing samples with tissue annotations. Because samples are neither collected nor processed uniformly, combined cross-study analyses using raw values from these sources are subject to biases introduced by batch effects that must be carefully considered and adjusted for (Gibson, 2008; Leek et al., 2010; Lazar et al., 2012; Tung et al., 2017). In spite of this limitation, they are relied upon by the following secondary repositories, subsets of which I attempted to harmonize in order to accomplish our first aim.

Large-Scale Human Tissue Studies

The Genotype-Tissue Expression (GTEx) Project and The Cancer Genome Atlas (TCGA) are two of the largest-scale human tissue-specific RNA-sequencing studies conducted over the preceding decade; they have been made publicly available and are commonly subject to reanalysis. GTEx, available online at https://commonfund.nih.gov/gtex, provides access to human whole-genome DNA sequencing and non-diseased tissue-specific RNA-sequencing data from several hundred human cadaver samples across 53 tissues (Carithers et al., 2015). To our knowledge, this is the single largest source of human non-diseased tissue-specific RNA-sequencing data currently available. TCGA (Wang et al., 2018) is a project sponsored by both the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) that collected the clinical and genetic information of over 20,000 human tumors across 33 cancer types including RNA-sequencing data for 29 healthy normal human tissue types intended for use as comparative references.

Aggregated Reprocessed Data Sources

ARCHS4 makes RNA-sequencing data readily publicly available at the gene and transcript levels and went about uniformly processing over eighty thousand human samples from GEO and SRA spanning 53 tissues and 67 cell lines (Lachmann et al., 2018). This project noted the ease of use in performing subsequent analyses on GTEx and TCGA and sought to represent GEO data in a similarly accessible way, though this project did not address batch effect discovery or removal, it does provide raw read counts along with GEO metadata. This project performed uniform transcript quantification using Kallisto (Bray et al., 2016), with which they report to have achieved processing costs of < $0.01 per sample (Lachmann et al., 2018). Uniform processing of multiple study results does reduce researcher degrees of freedom, though, without batch effect adjustment, systematic analysis remains statistically dubious as study-specific biases may occur upstream of the data processing procedure itself.

Recount2 is a project aiming to uniformly process approximately seventy thousand GTEx, TCGA and SRA samples (Collado-Torres et al., 2017) with a single pipeline and claims to aid in reducing the number of researcher degrees of freedom by preselecting pipeline parameters. This project quantified transcripts using Rail-RNA, a cloud-based aligner with reported sample processing costs ranging from $0.72 to $0.91 per sample (Lackmann et al., 2018, Nellore et al., 2017). In a similar fashion to ARCHS4, Recount2 did not apply batch effect adjustment, impeding comparison of results from multiple studies. However, the decision to process large datasets such as GTEx and TCGA uniformly seems a boon to comprehension of post hoc analyses as study-specific effects of smaller studies may be better contextualized.

RNASeqDB unifies GTEx and TCGA data (Wang et al., 2018) and conducted uniform processing and gene and transcript quantification using the STAR aligner (Dobin et al., 2013). This project also applied batch effect adjustment using ComBat and SVAseq (Johnson et al., 2007; Leek 2014) in an effort to remove undesired variation due to study, a known issue in performing comparisons between such datasets.

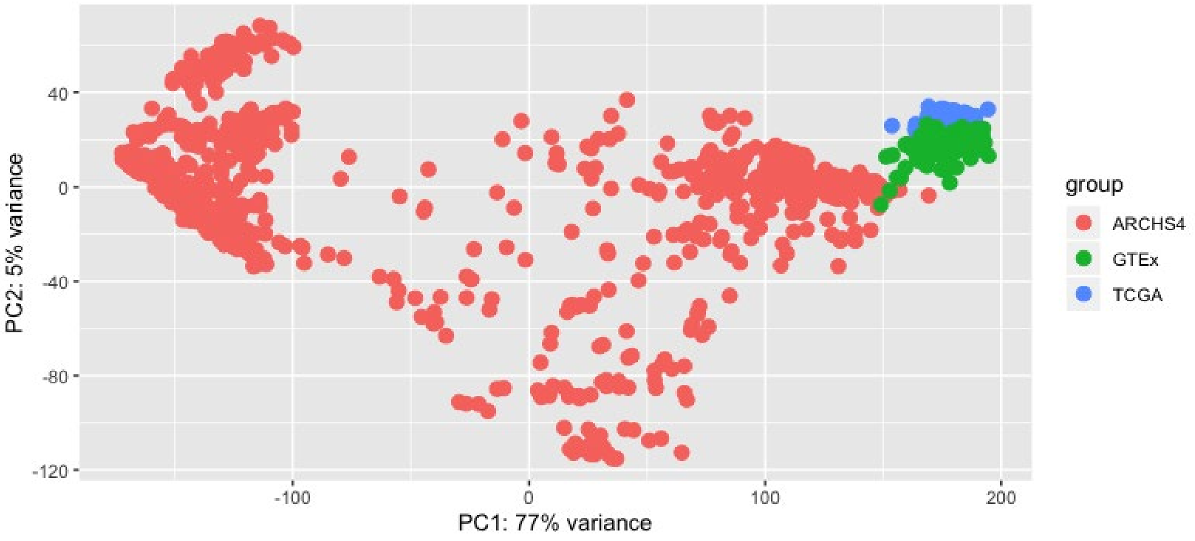
To evaluate these data sources I focused both on the evaluation of the aggregated datasets from the above outlined reprocessing pipelines individually and the assessment of their potential for integration in our methods development. I demonstrated the importance of pre-processing, normalization and exploratory data analysis of mRNA expression values resulting from multiple experiments in our previous research (Burkhart, 2016; Fourati et al., 2018), concepts already reinforced throughout the literature (Gibson, 2008; Leek et al., 2010; Lazar et al., 2012; Tung et al., 2017). Early on, Gibson suggested—as they are abiotic sources of variation—technical effects could be considered sources of environmental contribution to trait variation, further noting batch effects may impede comparative gene expression studies (Gibson, 2008). Indeed, in our prior studies, adjusting mRNA expression levels for technical variation improved the accuracy of machine learning models in predicting viral challenge outcomes on independent validation data held out from an initial training dataset. Our intention is to analytically explore our proposed data sources in order to better understand them and how compromises at this stage may affect downstream research efforts.

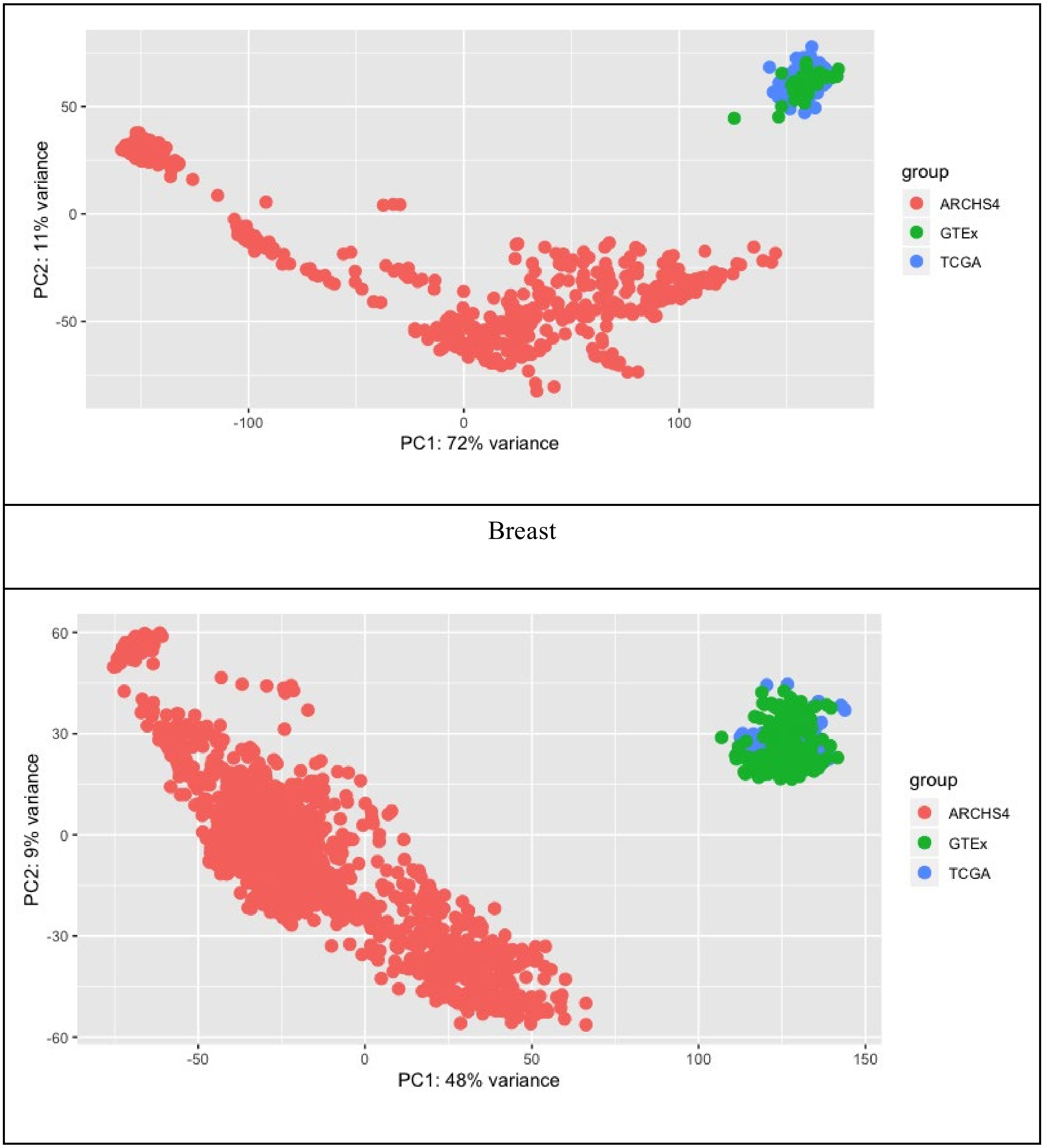
Exploratory Data Analysis

Initial quality assessment consisted of automated comparison of sample identifiers across local datasets and checks for missing values. I used the DESeq2::vst() function in order to adjust for library size and transform raw RNA-sequencing counts to log2 scale values. In order to explore sources of variation within tissue, I compared sample distributions using principal component analysis (PCA), a classical analytical technique used for representing systems via lower-dimensional planes (Pearson, 1901); this showed differences across both data source and tissue type. After recognizing prominent data source-specific batch effects were present in our data, I applied batch effect adjustment using the limma::removeBatchEffect() function, which fits a linear model to batch and removes it from the data, yielding transformed values. However, performing

PCA once again on the transformed values showed data source-specific batches were merely mean-centered and not convincingly adjusted for. In order to compare variation within and among data sources this way, I compared sample distributions using gene-based hierarchical clustering across tissues for several biochemical reactions represented within our dataset.

We began by tabulating the healthy human tissues available in ARCHS4, GTEx and TCGA and, to best comprehend data source-specific biases, sought to include tissue samples from those tissues present in each data source. ARCHS4 bulk tissue download R scripts were generated and downloaded using the https://amp.pharm.mssm.edu/archs4/data.html web interface. GTEx and TCGA tissue samples were downloaded using the Recount2 web interface at https://jhubiostatistics.shinyapps.io/recount/. Transcripts with missing or zero values across all samples were removed from each data source. I applied pathway overrepresentation analysis to test for pathway-specific biases among those missing or zero value transcripts finding no significant enrichment using a significance threshold of FDR <= 0.05. Remaining transcript identifiers were converted to Ensembl (Howe et al., 2021) gene identifier using the biomaRt::getBM() function. Transcripts without Ensembl gene identifiers were removed and tested for pathway-specific biases (as above) without finding significant enrichment. The remaining data from each source was combined into a single R dataframe using the dplyr::bind_cols() function. Datasource and tissue annotations are stored separately as character vectors. Because some samples have some instances of undetected transcripts (zero-values) and some samples have some instances of more than the maximum integer value (machine integer maximum) for our computing environment, the R dataframe is scaled to the *machine integer maximum* - 1 and then shifted by adding 1 to each value. This is necessary preparation for log-transformation, which I then apply using the DESeq2::vst() function. I then created a design matrix using the aforementioned tissue vector and remove batch effects by parameterizing our call to limma::removeBatchEffects() with our datasource vector specified as *batch* and our tissue vector specified as *design* arguments.

Liver

Kidney

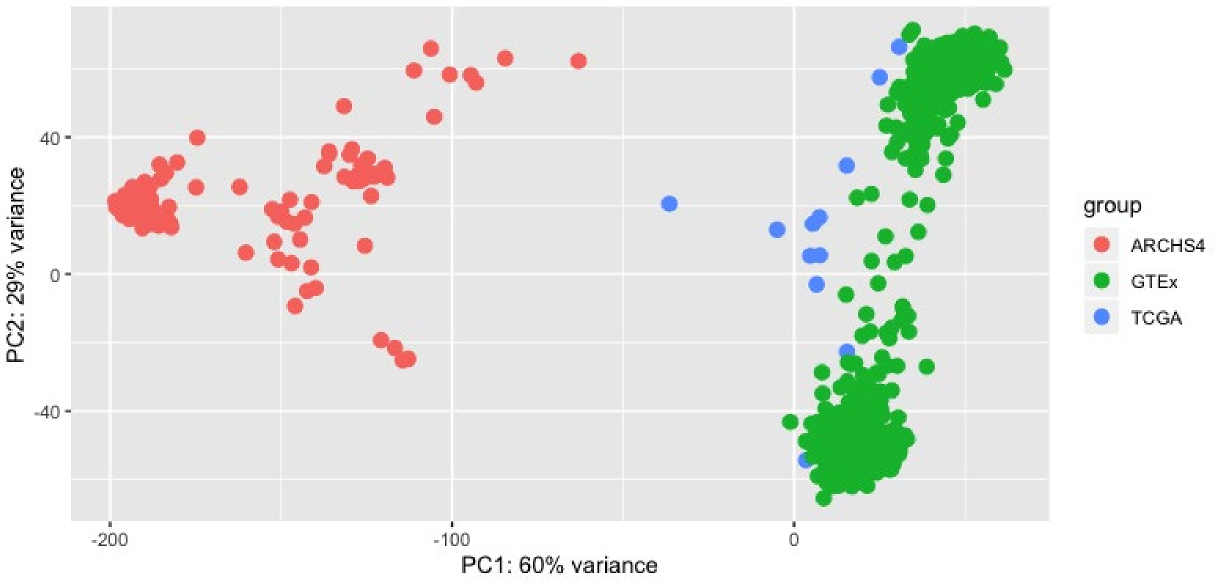


Breast

Figure 2: The first two principal components among liver, kidney and breast samples is derived from ARCHS4.

Uniform processing has the effect of removing the processing-method-specific variation; however, there remains uncertainty about how well this addresses other study-specific or systematic biases. I assumed batch effects would exist across data sources and selected tissues only present in each of the sources I investigated. I suspected batch effects may be appreciated visually using PCA plots and found a high degree of variation within tissue samples across data sources. The majority of within-tissue variation detected by considering the first two principal components among liver (~95%), kidney (~95%) and breast (~85%) samples is derived from ARCHS4 (Figure 2) while esophagus (~45%) and colon (~50%) within-tissue variation was not overtly driven by ARCHS4, with comparable variation observed in GTEx samples as well (Figure 3). Adjusting for data source had the effect of centering GTEx and TCGA samples within a cloud of more widely distributed ARCHS4 samples in liver, kidney and breast samples (Figure 4). Esophagus samples were shown to form a multimodal distribution with multiple distinct regions of density. In esophagus, this was recognised most clearly in samples from GTEx while colon samples formed a multimodal distribution with distinct regions of density observed in samples from both GTEx and ARCHS4 (Figure 5).

Esophagus



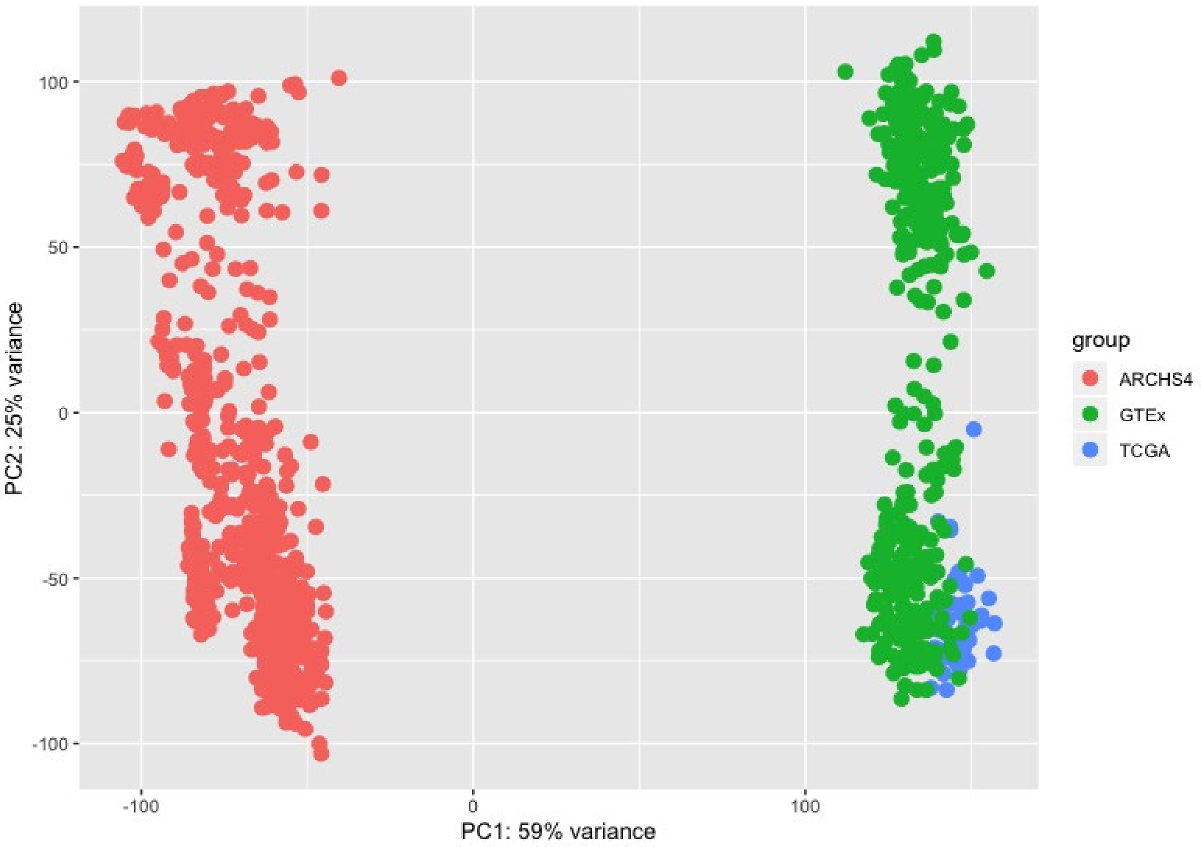
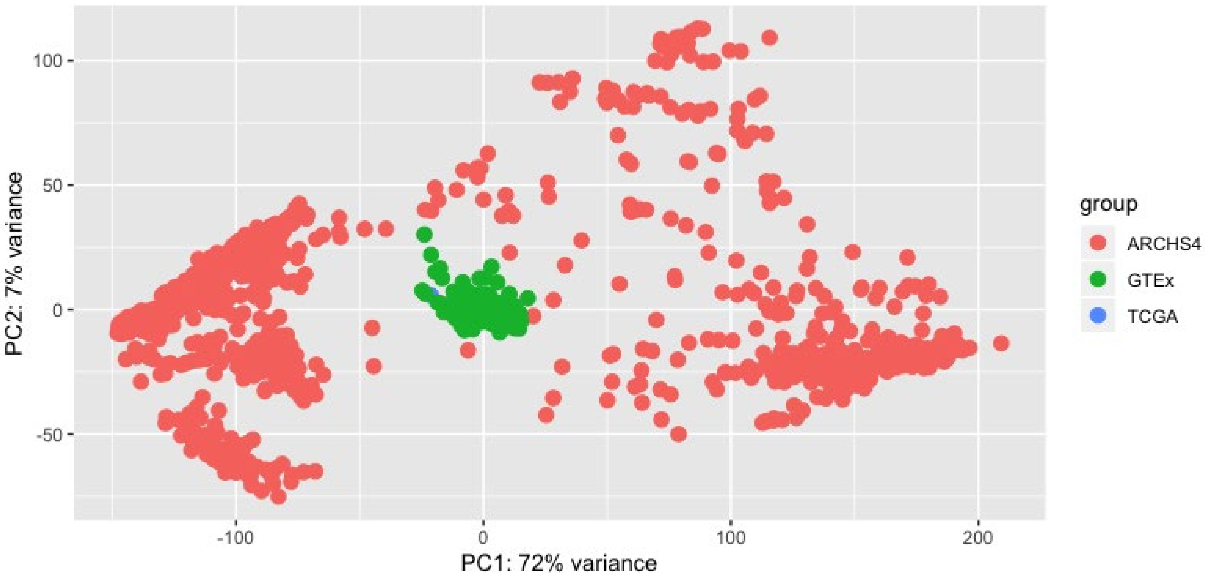
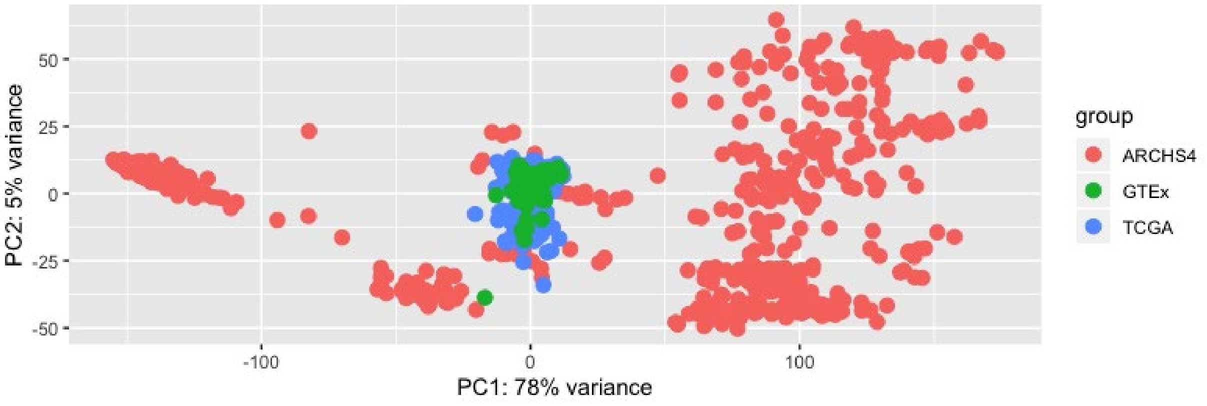
Colon


Figure 3: Esophagus and colon within-tissue variation was not overtly driven by ARCHS4, with comparable variation observed in GTEx samples as well.

Our supposition was that perhaps considering the entire repertoire of gene products to compare data sources was overzealous and that meaningful variation should consider only functional gene products—known to be involved in biochemical reactions. I performed complete-linkage agglomerative hierarchical clustering to assess tissue and data source contributions to gene expression variation among the top three most represented Reactome reactions within our dataset, each representing >= 50 unique transcripts detected in our data sources and, perhaps unexpectedly following insights aforementioned, found data source to be a strong confounding factor overall.

Among reactions I analyzed, R-HSA-381750: *Olfactory Receptor - G Protein olfactory trimer complex formation*, representing 75 transcripts, showed poorest separation across tissue both within and across data sources. However, other reactions, such as R-HSA-975040: *KRAB-ZNF / KAP Interaction*, representing 55 transcripts, and R-HSA-2730833: *Phosphorylation of TEC kinases by p-SYK*, representing 50 transcripts, showed improved separation across tissue within data sources though clustering samples across data sources remained poor (Figure 6). This analysis did not show evidence that allowed us to reject our prior supposition and left us with the conclusion that data source-specific batch effects remained in our data even following batch-effect adjustment.

Liver

Kidney



Breast

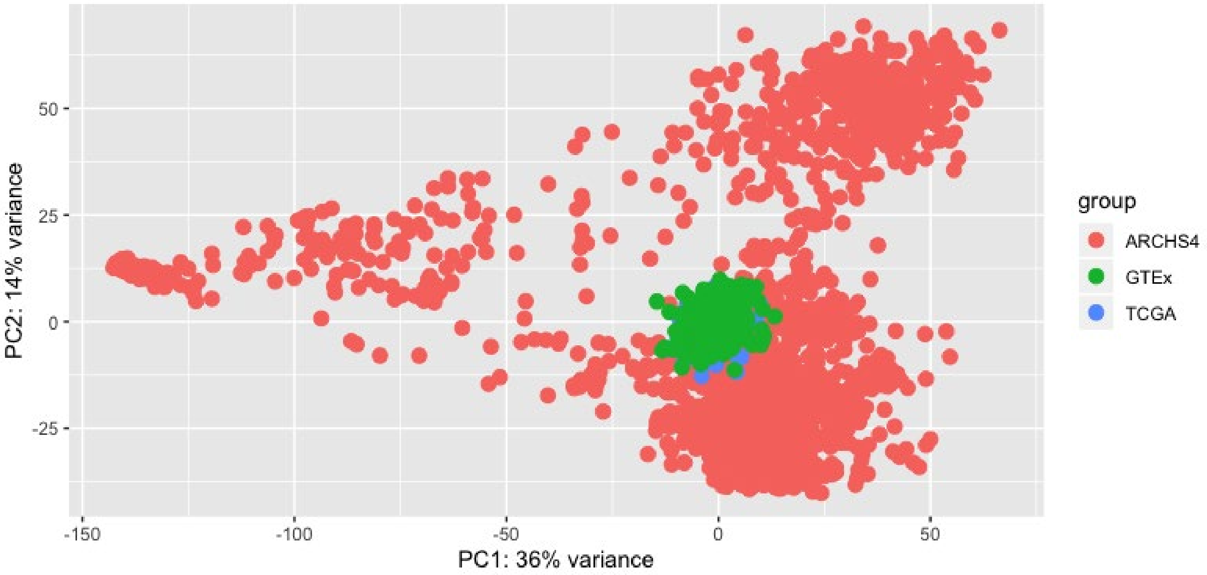
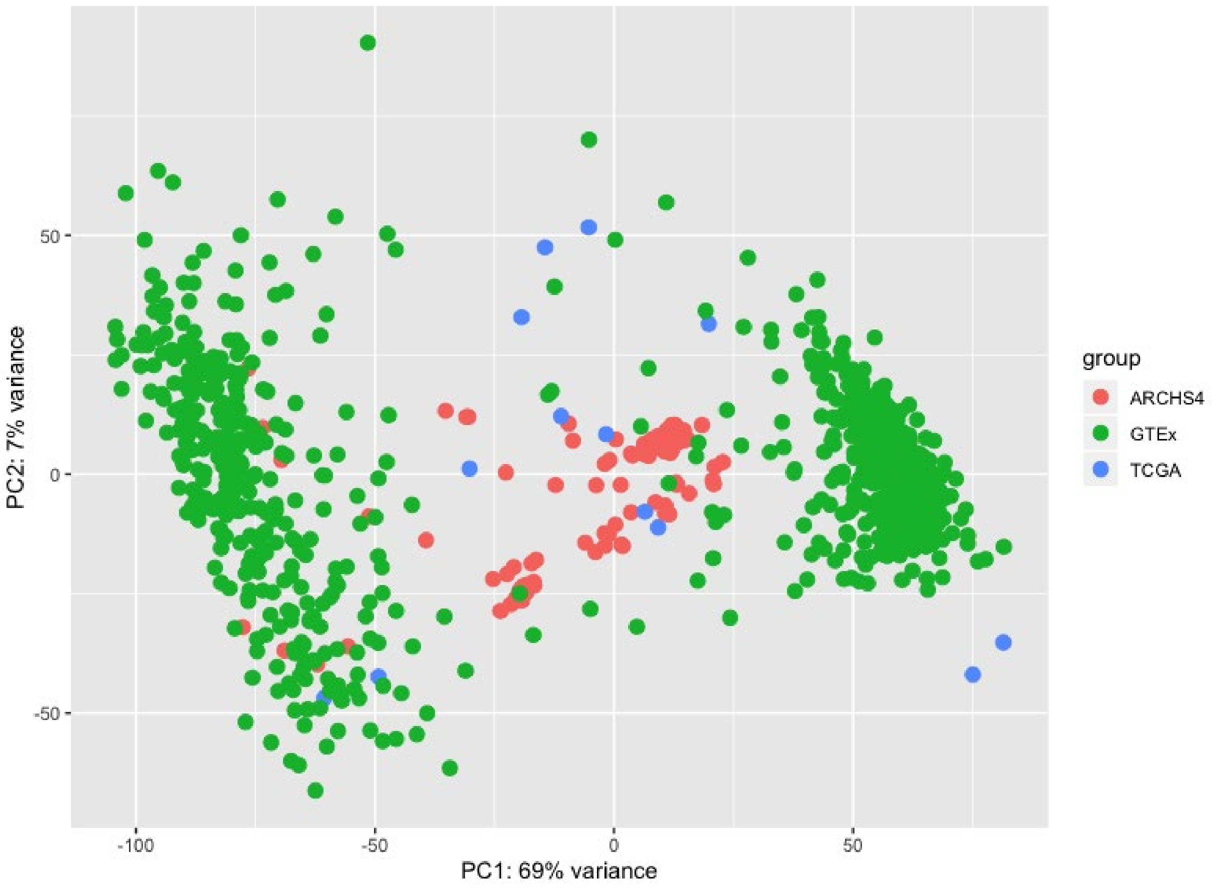


Figure 4: Adjusting for data source had the effect of centering GTEx and TCGA samples within a cloud of more widely distributed ARCHS4 samples in liver, kidney and breast samples.

Continuing our investigation into the variation across data sources I observed in both our PCA plots and our reaction clustering analysis, I noted Recount2 used Gencode v25 annotations in its pipeline while ARCHS4 used GRCh38.87 annotations and, noting many transcripts were dropped during our filtering, suspected there may be characteristic biases introduced either in the annotation process or the identifier conversion I performed with biomaRt (Durinck et al., 2009). I grouped transcripts by the data source from which they were dropped and whether they were dropped because they were either not detected among all samples within that data source or were unmappable by biomaRt. I performed overrepresentation analysis on Reactome pathways for each of the abovementioned groups of transcripts without finding significant enrichment using a significance threshold of false discovery rate (FDR) < 0.05; the strongest result was the overrepresentation of *Interleukin-2 family signaling* among transcripts not detected across all samples within GTEx, which had a non-significant FDR of merely 0.081.

Esophagus


50
0
-50
PC2: 7% variance
-100
-50
0
50
PC1: 69% variance
group
ARCHS4
GTEx
TCGA

Colon

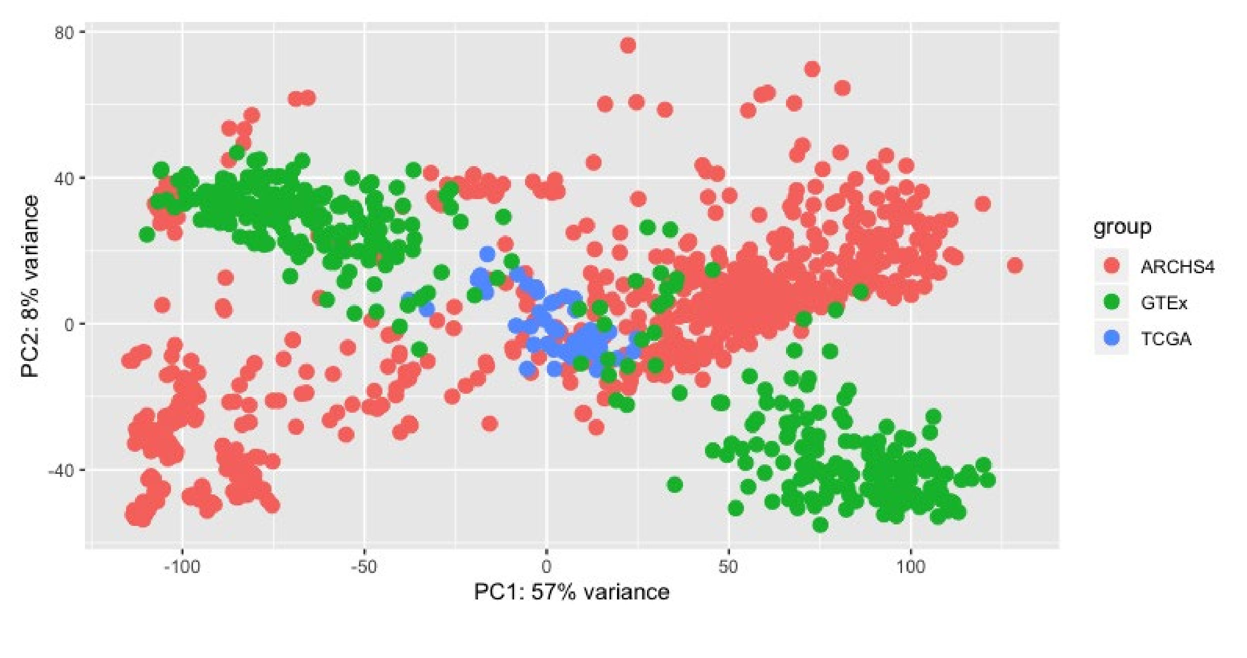
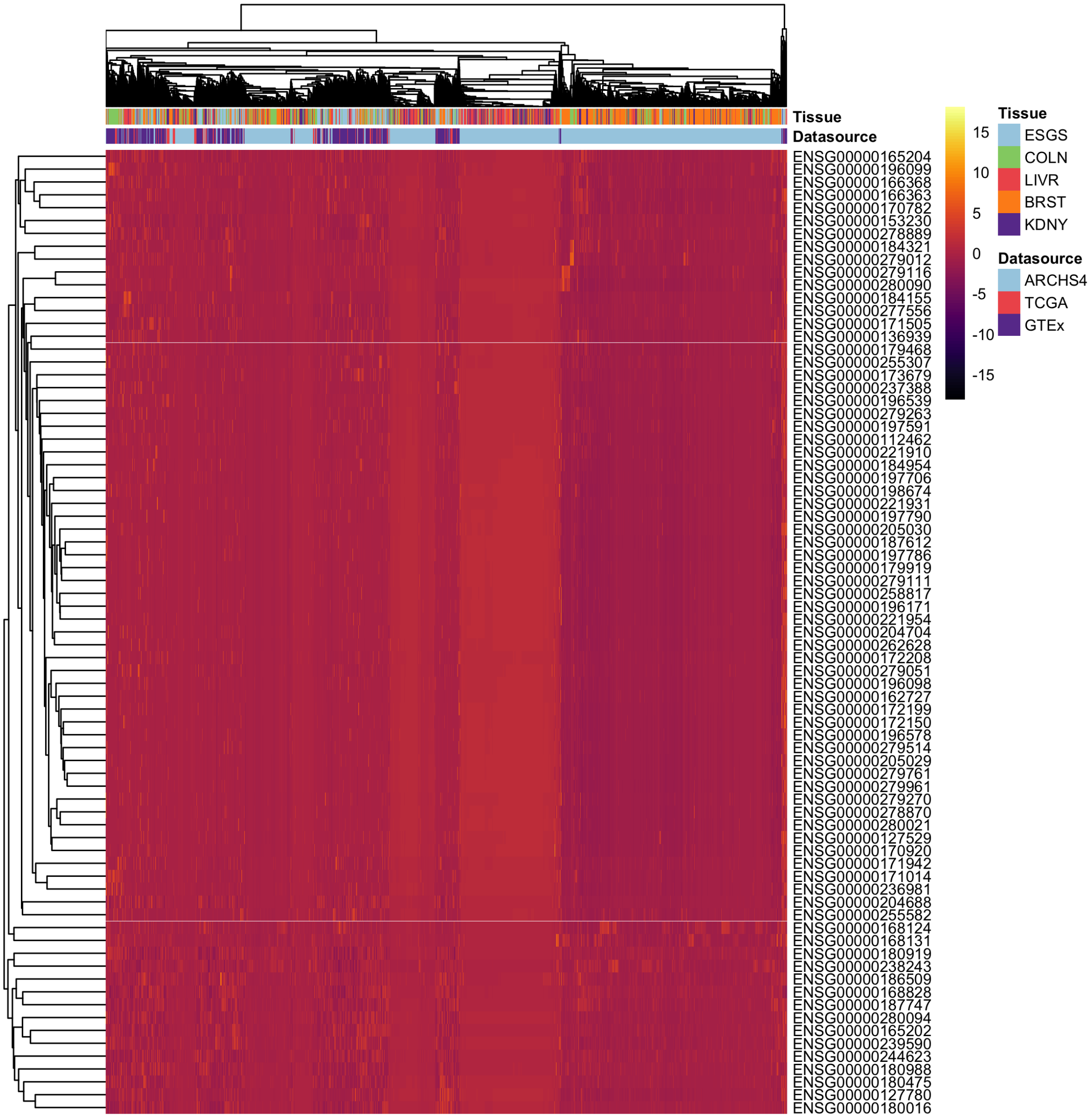


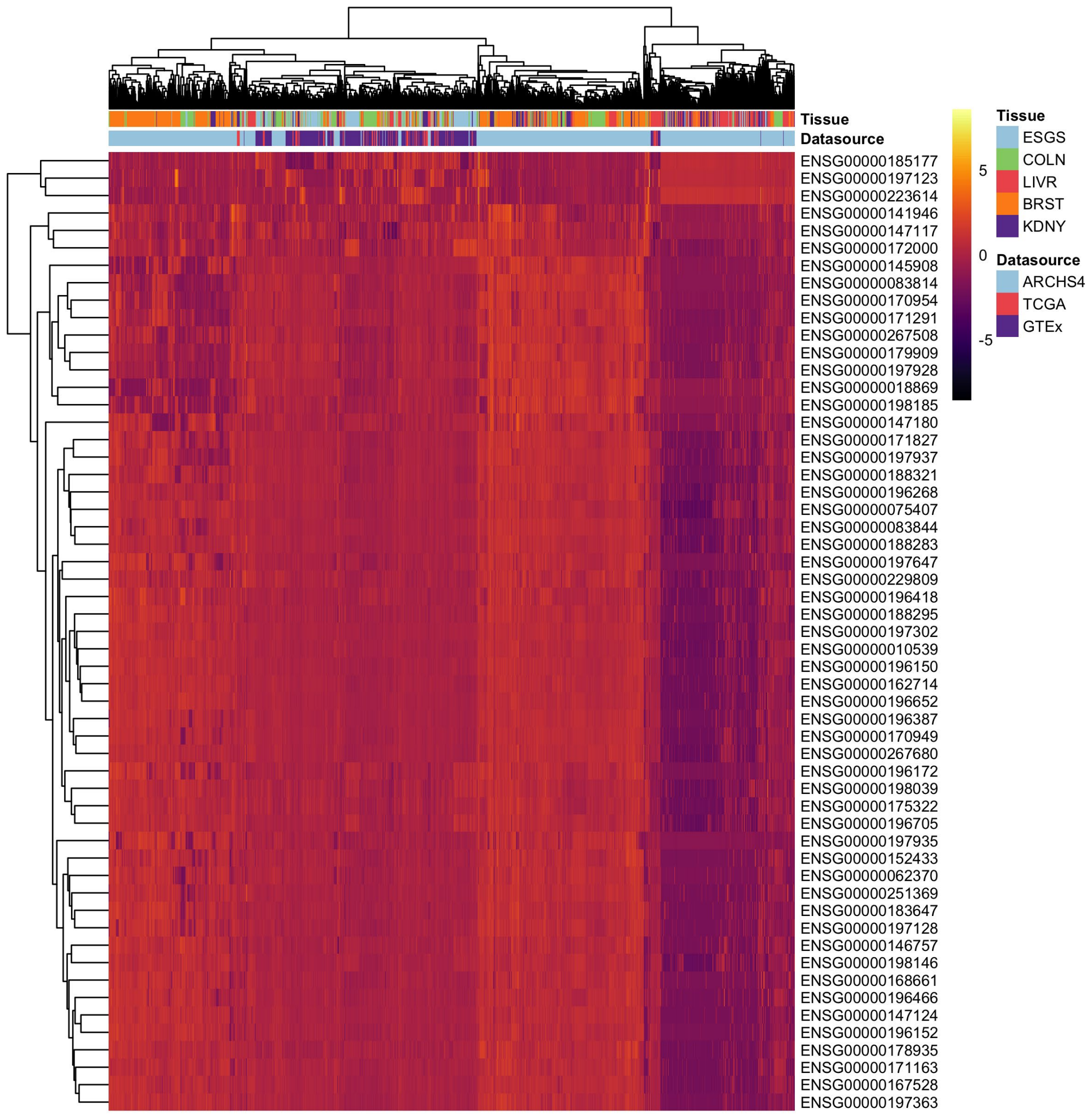
Figure 5: Esophagus samples were shown to form a multimodal distribution with multiple distinct regions of density. In esophagus, this was recognised most clearly in samples from GTEx while colon samples formed a multimodal distribution with distinct regions of density observed in samples from both GTEx and ARCHS4.

Following recognition of multimodal distributions within data sources, I hypothesized modes may represent different cellular lineages. I considered how GTEx tissue region *smtsd* annotations compared with whole tissue annotations from other data sources as well as GTEx itself. Among those tissues where multimodality was observed, distinct regions of density were clearly driven by GTEx tissue region, a course annotation associated with cell lineage. For example, I showed the bimodality observed within esophagus samples could be easily explained by separating *Gastroesophageal Junction*, *Mucosa* and *Muscularis* annotations (Figure 7).

A: R-HSA-381750



B: R-HSA-975040



C: R-HSA-2730833

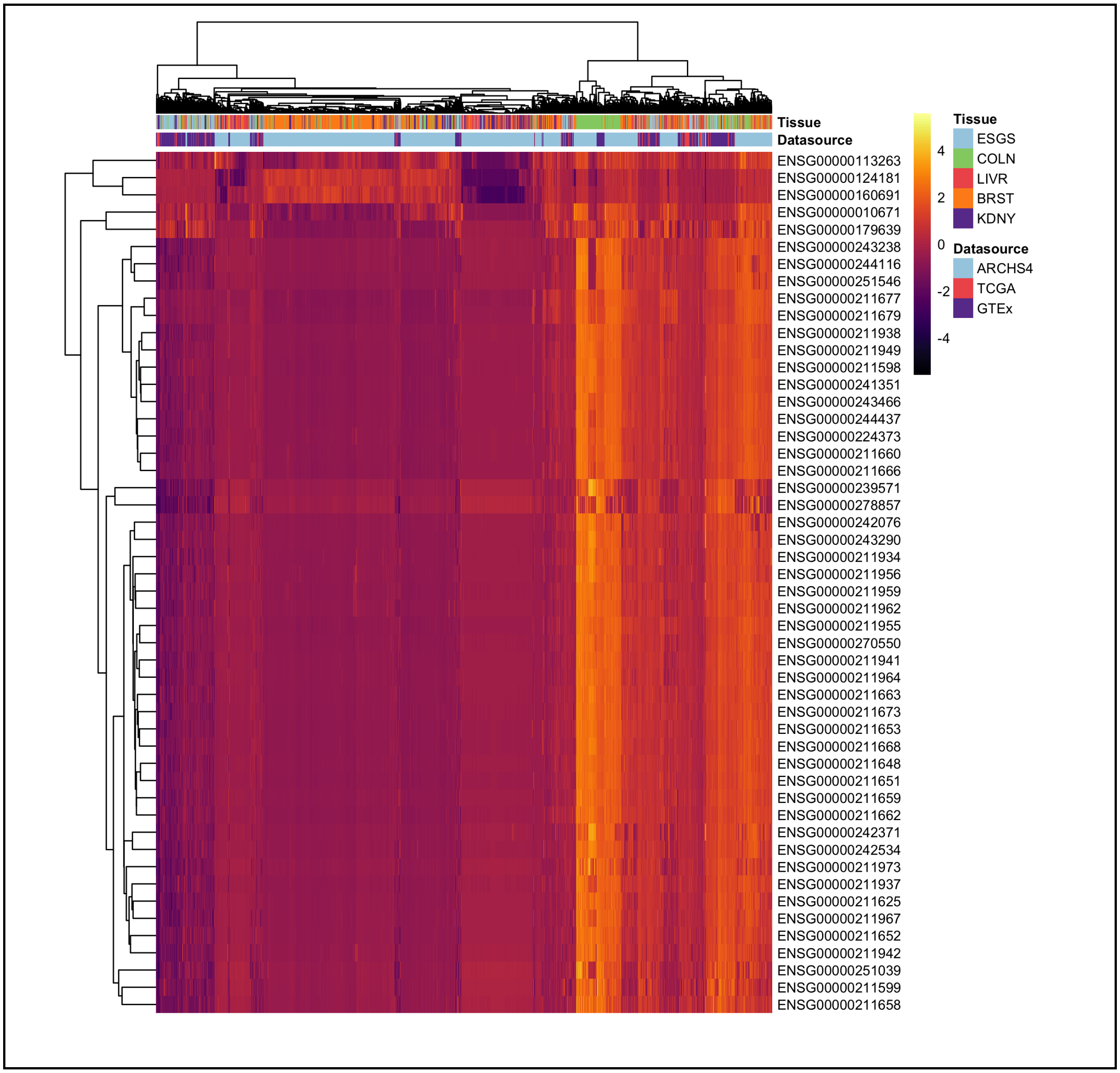
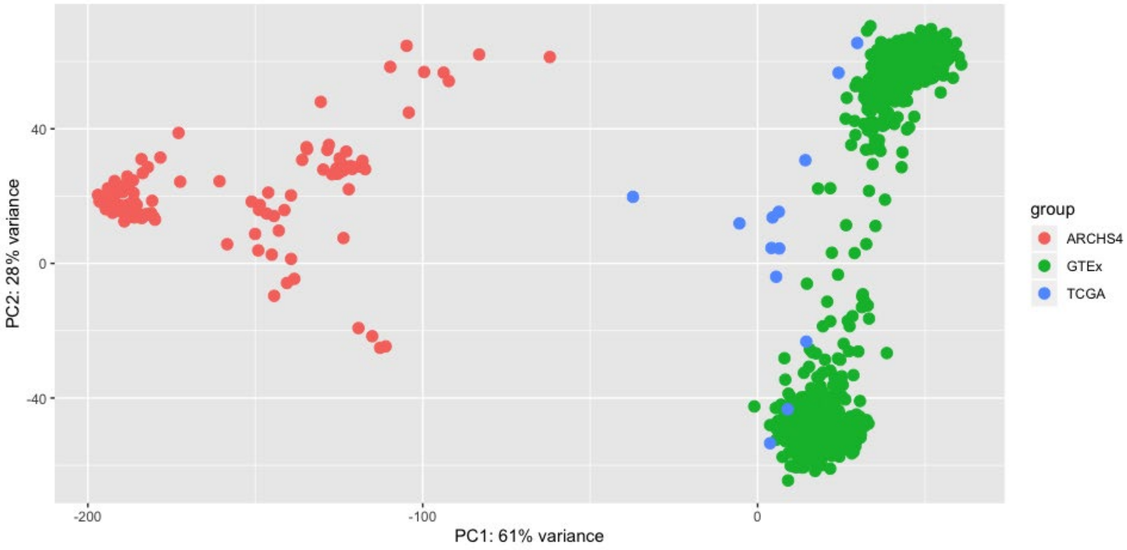


Figure 6: Panel A depicts R-HSA-381750: Olfactory Receptor - G Protein olfactory trimer complex formation, representing 75 transcripts, which showed poorest separation across tissue both within and across data sources. Other reactions, such as R-HSA-975040: KRAB-ZNF / KAP Interaction, representing 55 transcripts and depicted in Panel B, and R-HSA-2730833: Phosphorylation of TEC kinases by p-SYK, representing 50 transcripts and depicted in Panel C, showed improved separation across tissue within data sources though clustering samples across data sources remained poor.

In order to compare Recount2 and ARCHS4 processing effects more directly, I reviewed 168 breast tissue samples that had been processed by each pipeline separately (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE58135). The study that generated this

GEO series—in an effort to investigate the association of read-through fusion transcripts with breast cancer—conducted RNA-sequencing experiments using multiple breast tissue cell types including several breast cancer cell lines, triple negative breast cancer (TNBC) primary tumor, estrogen receptor positive (ER+) primary tumor, TNBC adjacent, ER+ adjacent as well as normal breast collected from breast reduction surgery and used as healthy controls (Varley et al., 2014). I discovered both technical variation between the two data sources as well as errors of annotation among ARCHS4 samples.

A: GTEx *smts* annotations



B: GTEx *smtsd* annotations

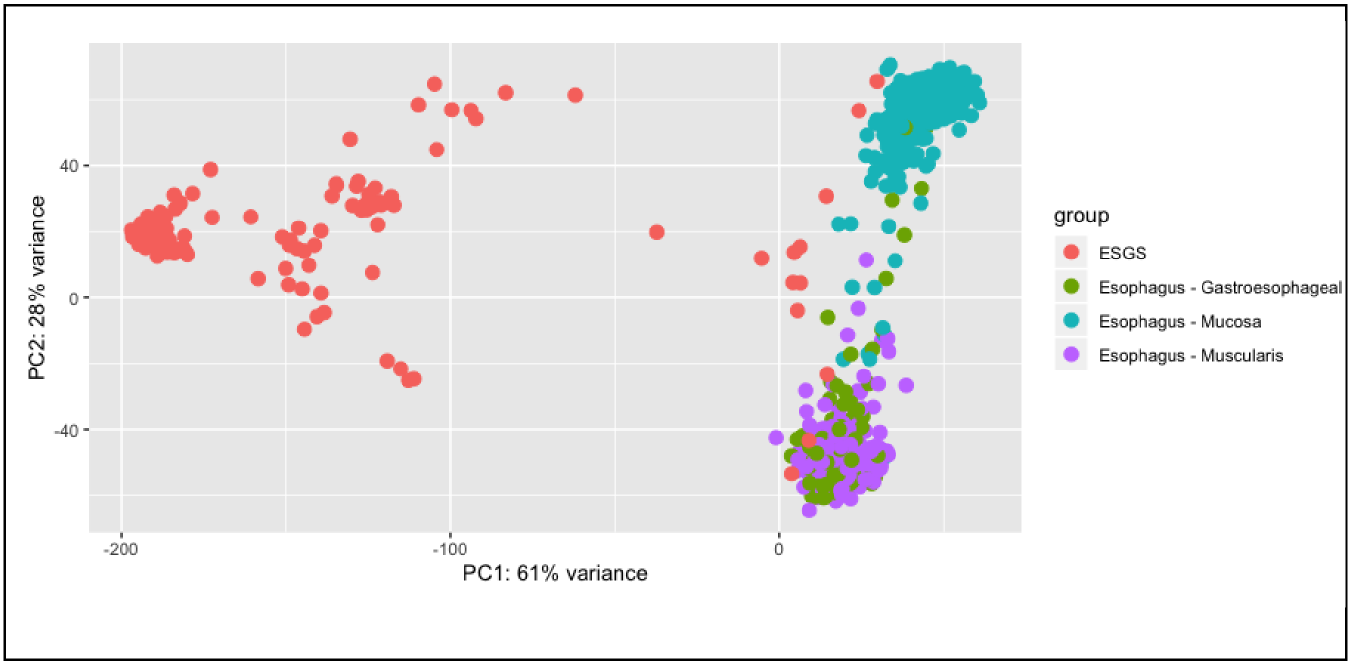


Figure 7: Distinct regions of density were clearly driven by GTEx tissue region, a course annotation associated with cell lineage. For example, relative to ARCHS4 and TCGA samples, bimodality observed within GTEx esophagus samples (Panel A) could be easily explained by separating Gastroesophageal Junction, Mucosa and Muscularis annotations (Panel B).

We used PCA to plot samples collected from this study that had been processed by both ARCHS4 and Recount2 and found the respective ARCHS4 samples clustered tightly with the rest of the ARCHS4 samples (collected from other studies) while the respective Recount2 samples clustered more closely with Recount2 samples from GTEx and TCGA (Figure 8). As the respective ARCHS4 and Recount2 sample sets are indeed drawn from the same underlying sample set, the observed separability is indicative of severe batch effects due to the differences in these pipelines' processing methods.

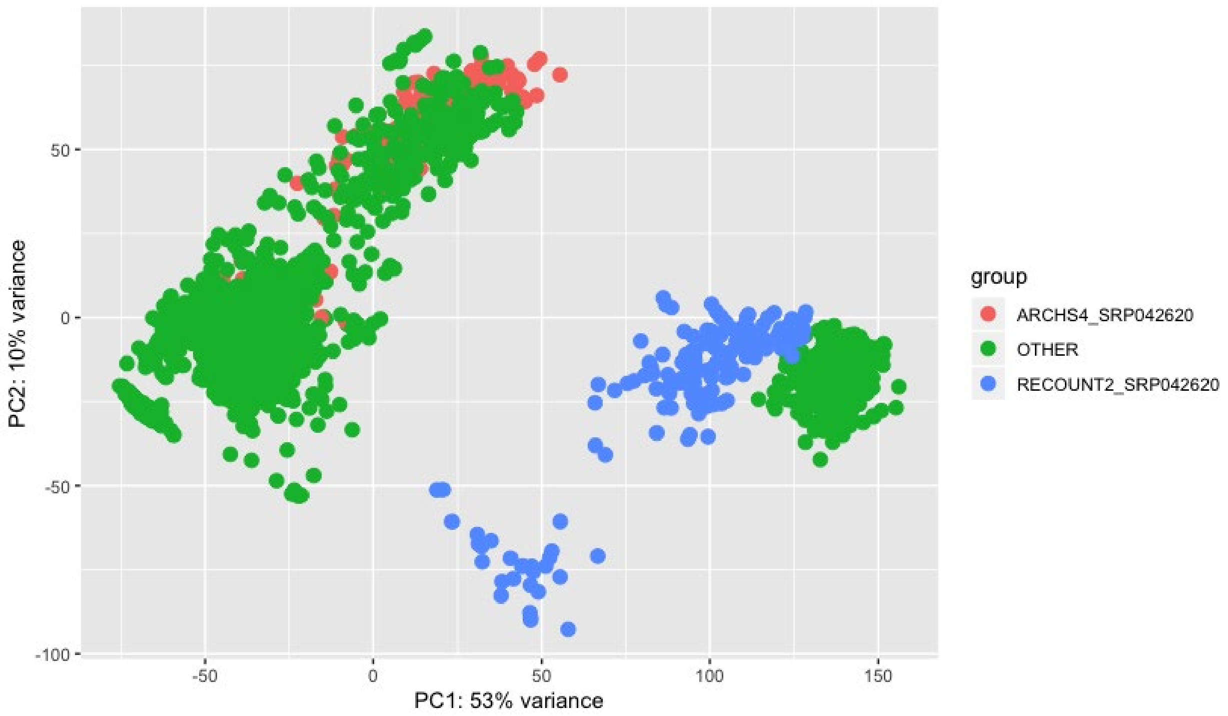


Figure 8: ARCHS4 samples clustered tightly with the rest of ARCHS4 samples (collected from other studies) while the respective Recount2 samples clustered more closely to those Recount2 samples from GTEx and TCGA.

In order to ensure our observations were not skewed simply by the fact that Varley et al. samples were largely transformed (by cancer) while our GTEx and TCGA samples represented healthy human tissues, I annotated the Varley et al. samples by cell type in our PCA as well (Figure 9). Surprisingly, ARCHS4 was found to include diseased samples (Terminal session log 0.2) but lack the healthy control samples from Varley et al. in its list of bulk breast tissue samples, despite the samples existing within the ARCHS4 database under the Varley et al. study identifier itself. Regarding the underlying cause of this issue, it appears a parsing issue fails to recognize and separate GEO/SRA samples precisely as intended; however the recognition of such an error suggests ARCHS4 bulk human tissue listings may not serve as representative samples of healthy human tissue.

With these observations potentiating difficulty integrating ARCHS4 samples with our other data sources, I considered the impact removing ARCHS4 samples would have on our tissue sample counts. Our tissue sample distribution with ARCHS4 covers five tissues with a mean sample count of 1329.2 and a total of 6646 samples (Table 1). I then considered the case where only GTEx and TCGA samples from Recount2 were used and found the tissue coverage increased from five to sixteen; however the average number of samples per tissue and the total number of samples decreased from 1329.2 to 384.25 and from 6646 to 6148, respectively (Table 2). I compared these cases to that of using GTEx data from Recount2 alone and found tissue coverage increased from 16 to 31 and total sample count from 6148 to 9662, while mean samples per tissue decreased only modestly from 384.25 to 311.7 (Table 3). Thus, I opted to use this GTEx data alone, forgoing integration of outside sources.

Table 1: Tissue sample counts for tissues shared among GTEx, TCGA (healthy samples) and ARCHS4

Tissue Label	GTEx	TCGA	ARCHS4	**Total Count**
Breast	218	112	2217	**2547**
Colon	376	51	797	**1224**
Esophagus	790	13	137	**940**
Kidney	36	129	681	**846**
Liver	136	50	903	**1089**
Total Count	**1556**	**355**	**4735**	**6646**

Table 2: Tissue sample counts for those tissues shared between GTEx and TCGA (healthy samples)

Tissue Label	GTEx	TCGA	**Total Count**
Adrenal Gland	159	3	**162**
Bladder	11	19	**30**
Brain	1409	5	**1414**
Breast	218	112	**330**
Cervix	11	3	**14**
Colon	376	51	**427**
Esophagus	790	13	**803**
Kidney	36	129	**165**
Liver	136	50	**186**
Lung	374	110	**484**
Pancreas	197	4	**201**
Prostate	119	52	**171**
Skin	974	1	**975**
Stomach	204	37	**241**
Thyroid	361	59	**420**
Uterus	90	35	**125**
Total Count	**5465**	**683**	**6148**

Table 3: Tissue sample counts for those tissues included in GTEx

Tissue Label	GTEx Sample Count
Adipose Tissue	620
Adrenal Gland	159
Bladder	11
Blood	595
Blood Vessel	750
Bone Marrow	102
Brain	1409
Breast	218
Cervix Uteri	11
Colon	376
Esophagus	790
Fallopian Tube	7
Heart	489
Kidney	36
Liver	136
Lung	374
Muscle	475
Nerve	335
Ovary	108
Pancreas	197

Pituitary	124
Prostate	119
Salivary Gland	70
Skin	974
Small Intestine	104
Spleen	118
Stomach	204
Testis	203
Thyroid	361
Uterus	90
Vagina	97
Total Count	**9662**

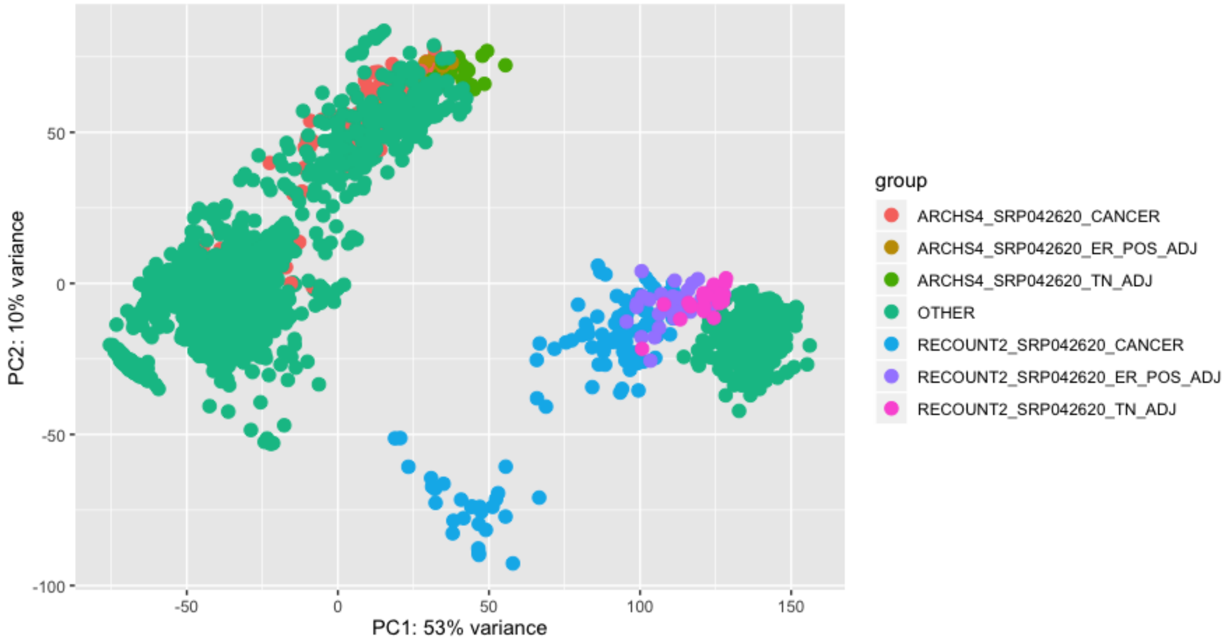


Figure 9: Varley et al. samples annotated by cell type and compared within our PCA.

Reactionwise Principal Component Coordinate Transformation

In order for any particular reaction to occur, its necessary reactants must be present. Some reaction components such as siRNA, metal ions and other organic and inorganic compounds are currently unquantified in a tissue-specific manner appropriate for integration with this work; however, biochemical reactions are mediated by enzymes and proteins resulting from translation of mRNA, which influence reaction rate. Across human tissues, mRNA abundance varies, affecting which and to what degree biochemical reactions take place. This differential reaction utilization is observed across tissues and is responsible for differential tissue-specific function. In order to generate values for each reaction which approximate reaction utilization, I applied principal component analysis across transcript sets annotated for each reaction as a conceptually simple, computationally efficient and outcome-naive aggregation strategy (Figure 10).

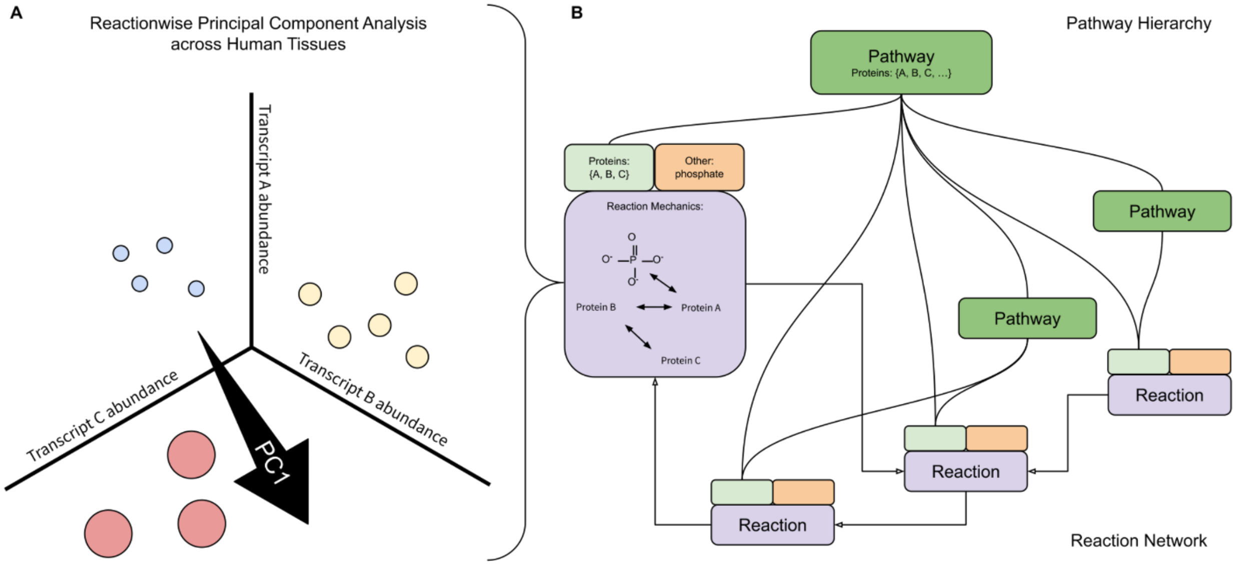


Figure 10: Panel A shows a first principal component (PC1) positioned within a space representing a reaction to which three proteins are annotated. Axes represent three transcript abundance levels corresponding to the reaction's three proteins. Circles represent tissue samples. Circle color represents sample tissue type. Circle size is varied to suggest image depth. Panel B shows reactions composed as three information sets: proteins (light green boxes), other components (orange boxes) and reaction mechanics (purple boxes). Protein sets are shown to be connected to a pathway hierarchy (dark green boxes) by curved connectors. Reaction mechanics are shown to be connected by open-headed arrows.

GTEx RNA-seq data files were downloaded from the Recount2 project website at https://jhubiostatistics.shinyapps.io/recount/ and records annotated with healthy tissue labels were retained. Samples removed included those annotated as *Cells - Transformed fibroblasts* (306), *Cells - Leukemia cell line (CML)* (102), *Cells - EBV-transformed lymphocytes* (139) and those with no tissue label (5). Transcripts whose identifiers were not annotated as participating in Reactome reactions were removed. Values were scaled up to the machine maximum integer value - 1 and added to 1, producing transcript pseudocounts (Love MI et al., 2019; Love MI et al., 2015), and processed using the variance stabilizing transform (Anders S et al., 2010) provided by DESeq2 (Love MI et al., 2014).

Transcripts were grouped in a many-to-many fashion according to the Reactome reactions in which their protein products participated. Principal component analysis was conducted using the prcomp function (https://stat.ethz.ch/R-manual/R-patched/library/stats/html/prcomp.html) from the R stats package (R Core Team, 2013) and resulting principal component result objects were stored. The distribution of proportion of variance explained by each principal component across reactions is shown in Figure 11 with the median first principal component explaining approximately 50% of variance, the median second principal component explaining approximately 25% of variance and the subsequent principal components tending to explain less variance, as would be expected. The first principal component value from each reaction was recorded for each sample, forming a matrix. This routine transformed a matrix representing 6,323 transcripts representing 6,323 unique genes across 9,115 tissue samples to a matrix representing 10,726 reactions across those same 9,115 samples.

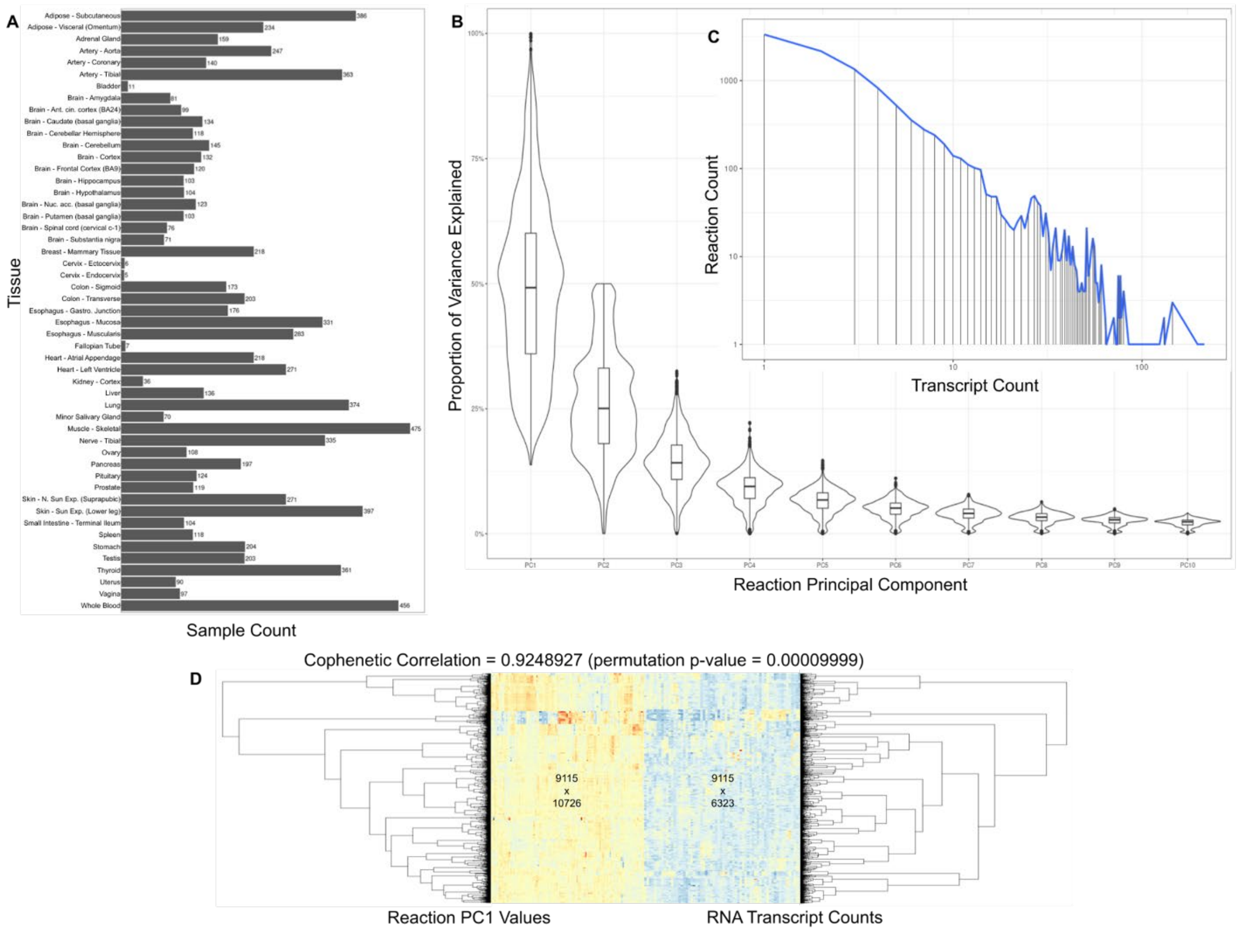


Figure 11: Panel A shows tissue sample counts across tissues. Panel B shows principal component distributions of proportions of variance explained across the first 10 principal components. Panel C shows a log-log plot of reaction vs transcript counts. Panel D shows the tissue sample dendrograms from both reaction PC1 (left) and transcript counts (right) are correlated significantly. Heatmaps shown are downsampled from the underlying matrices for visualization purposes.

Tissue sample counts ranged from 5 (Endocervix) to 475 (Skeletal muscle), indicated in Figure 11A. Reaction transcript counts ranged from 1 (multiple reactions) to 214 *Olfactory Receptor - G Protein olfactory trimer complex formation* (R-HSA-381750), indicated in Figure 11C.

Reactionwise Statistical Analysis

To demonstrate tissue-specific patterns across reactions, I applied a K=1 Nearest Neighbor (KNN) classifier (Fix & Hodges, 1951; Cover & Hart, 1967), calculating for each reaction and across all tissues, both the adjusted rand index (ARI)—which may be considered as an approximation of overall tissue classification accuracy (Hubert L et al., 1985)—and tissue-specific classification accuracy (ACC). Reactions with the highest ARI values are included in Table 4. If one were to assume some reactions acted simply as better predictors than others, we could place each reaction upon a diagonal whereby ACC for individual tissues equals ARI; however, tissuewise plots show reaction point clouds where ACC deviates from ARI, indicating distinct tissue-specific reaction states are assumed for some tissues while, at a system-wide scale, these same reactions are not reliable tissue predictors overall. To show this bias more clearly, I scaled both ACC and ARI to the interval [0,1] and subtract ARI from ACC, showing resulting plots for both Breast and Lung tissue (Figure 12). Reactions with the highest ACC values for Breast and Lung are included in Tables 5 and 6, respectively. To show these tissue-specific patterns across all tissues, I calculated the accuracy for each reaction for each tissue and generated a dendrogram where, generally, tissues grouped near each other are classified at similar rates by the same reactions whereas those grouped further apart are classified at differential rates by those reactions (Figure 12F).

Table 4: Reactions with highest ARI values

Reaction Name	Reaction ID	ARI
RAS GEFs promote RAS nucleotide exchange	R-HSA-5672965	0.8342
Liganded Gq/11-activating GPCRs act as GEFs for Gq/11	R-HSA-379048	0.8331

Liganded Gq-activating GPCRs bind inactive heterotrimeric Gq	R-HSA-749448	0.8331
The Ligand:GPCR:Gq complex dissociates	R-HSA-749452	0.8331
Liganded Gi-activating GPCR acts as a GEF for Gi	R-HSA-380073	0.8244
The Ligand:GPCR:Gi complex dissociates	R-HSA-749454	0.8234
Liganded Gi-activating GPCRs bind inactive heterotrimeric G-protein Gi	R-HSA-749456	0.8234
Liganded Gs-activating GPCR acts as a GEF for Gs	R-HSA-379044	0.8207
The Ligand:GPCR:Gs complex dissociates	R-HSA-744886	0.8207
Liganded Gs-activating GPCRs bind inactive heterotrimeric Gs	R-HSA-744887	0.8207

Table 5: Reactions with highest Breast ACC values

Reaction Name	Reaction ID	ARI	Breast ACC
PI3K inhibitors block PI3K catalytic activity	R-HSA-2400009	0.8099	0.9918
GPLD1 hydrolyses GPI-anchors from proteins	R-HSA-8940388	0.7760	0.9916
RAS GEFs promote RAS nucleotide exchange	R-HSA-5672965	0.8342	0.9914
The Ligand:GPCR:Gi complex dissociates	R-HSA-749454	0.8234	0.9912
Liganded Gi-activating GPCRs bind inactive heterotrimeric G-protein Gi	R-HSA-749456	0.8234	0.9912
Liganded Gq/11-activating GPCRs act as GEFs for Gq/11	R-HSA-379048	0.8331	0.9912
Liganded Gq-activating GPCRs bind inactive heterotrimeric Gq	R-HSA-749448	0.8331	0.9912

The Ligand:GPCR:Gq complex dissociates	R-HSA-749452	0.8331	0.9912
PI3K phosphorylates PIP2 to PIP3	R-HSA-2316434	0.8060	0.9911
Liganded Gi-activating GPCR acts as a GEF for Gi	R-HSA-380073	0.8244	0.9910

Table 6: Reactions with highest Lung ACC values

Reaction Name	Reaction ID	ARI	Lung ACC
Vesicle budding	R-HSA-203973	0.7265	0.9998
SEC31:SEC13 and v-SNARE recruitment	R-HSA-204008	0.7265	0.9998
SEC16 complex binds SAR1B:GTP:SEC23:SEC24	R-HSA-5694417	0.7165	0.9998
HSPA8-mediated ATP hydrolysis promotes vesicle uncoating	R-HSA-8868658	0.8178	0.9998
Auxilin recruits HSPA8:ATP to the clathrin-coated vesicle	R-HSA-8868660	0.8178	0.9998
CSNK1D phosphorylates SEC23	R-HSA-5694441	0.7294	0.9997
SNX9 recruits components of the actin polymerizing machinery	R-HSA-8868230	0.8171	0.9997
F- and N- BAR domain proteins bind the clathrin-coated pit	R-HSA-8867754	0.8116	0.9997
FAM20C phosphorylates FAM20C substrates	R-HSA-8952289	0.8112	0.9997
BAR domain proteins recruit dynamin	R-HSA-8868236	0.8181	0.9997

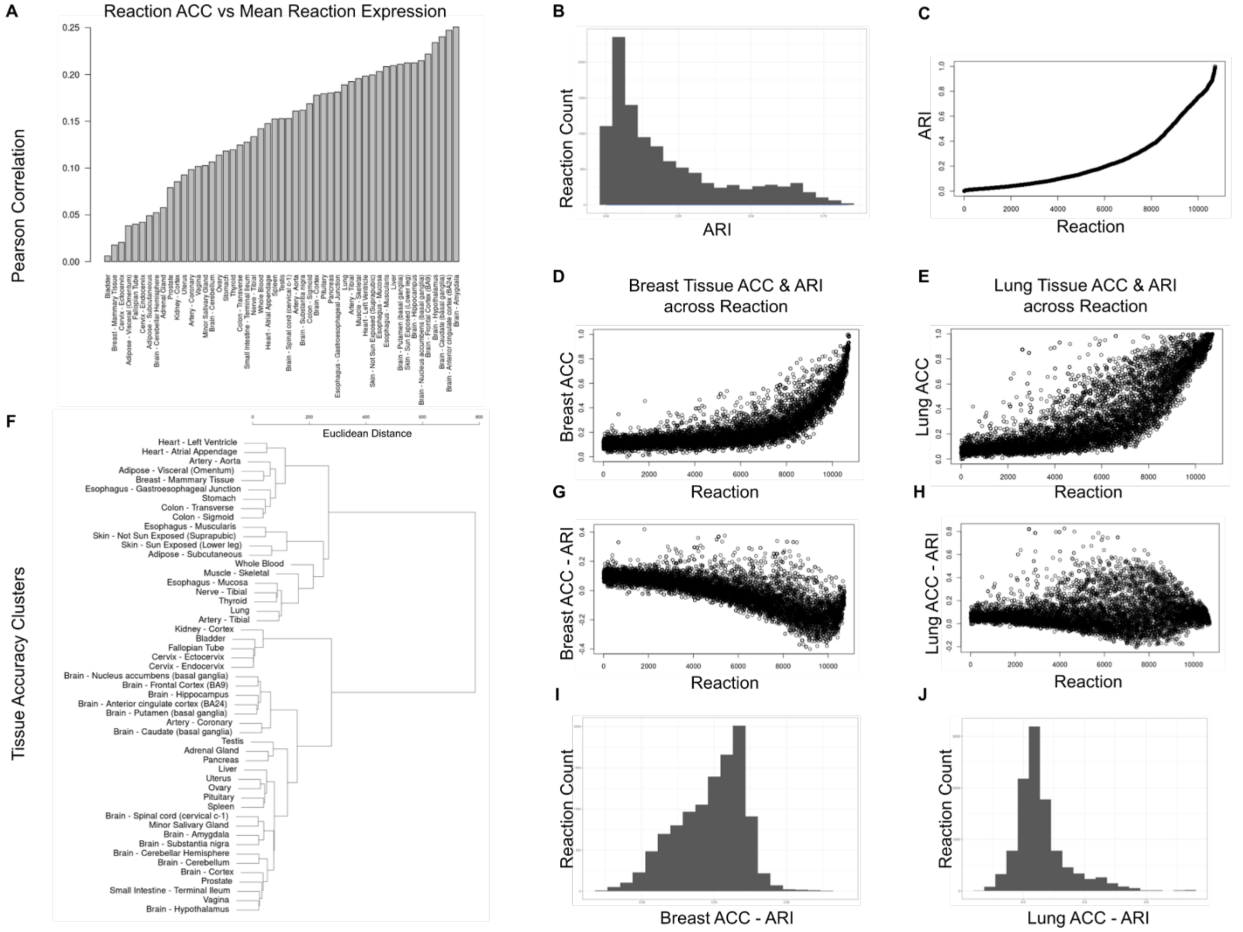


Figure 12: Panel A shows positive correlations between mean reaction expression difference and reactionwise KNN accuracy for each tissue. Panel B shows a histogram of ARI across all reactions. Panel C shows ARI for each reaction, scaled to the interval [0,1]. Panel D shows breast tissue KNN accuracy scaled to the interval [0,1]. Panel E shows lung tissue KNN accuracy scaled to the interval [0,1]. Panel F shows tissues hierarchically clustered by reactionwise KNN accuracy. Panel G shows scaled ARI (from C) subtracted from scaled breast tissue KNN accuracy (from D). Panel H shows scaled ARI (from C) subtracted from scaled lung tissue KNN accuracy (from E). Panel I shows a histogram of data in G. Panel J shows a histogram of data in H.

Recapitulation of mRNA Transcript Abundance Sample Structure

One common method to assess gene expression information across samples is to perform hierarchical clustering, which may reveal the structure among samples by considering overall samplewise similarity. In order to determine whether calculating reactionwise principal component summarization significantly degraded or otherwise influenced this structure, I calculated the Euclidean distance between samples and performed agglomerative hierarchical clustering using Ward's D (Ward Jr JH 1963; Murtagh F et al., 2014) on both reaction principal component coordinates and transcript counts across all samples and compared the resulting dendrograms. The Cophenetic correlation (Sokal RR et al., 1962; Sneath PH, 1973) between these dendrograms was calculated as 0.9248, which was shown to be significant by permuting over the reaction dendrogram labels as specified in the function documentation (https://rdrr.io/cran/dendextend/man/cor_cophenetic.html) 10,000 times (p-value < 1.0E-5). This demonstrates the reactionwise principal component coordinate values group tissue samples similarly to transcript counts, strengthening the hypothesis that the human Reactome reactions assume distinct tissue-specific states. The advantage of expressing this information as reaction states, rather than transcript counts, is that reactions form a network distinct from other networks formed by transcripts alone. The transformation of transcript count information into reactionwise information offers the opportunity to leverage gene expression information through a novel lens, that of the human biochemical network.

Phenotypic Validation

In order to validate our transformation, I sought to compare two groups of tissues with known biochemical variation. Using the results produced from a metaanalysis that examined cell proliferation among 31 tissues (Richardson et al., 2014), I separated our samples into three groups with a reported estimated representative mean value of cell turnover less than one month, between one month and one year and over one year in length. Removing sex-specific tissues and comparing high- and medium-turnover samples as well as medium- and low-turnover samples with two-tailed

Wilcoxon tests, I combine resulting p-values using Fisher's method and calculate a false discovery rate (FDR) for each reaction to adjust for multiple hypothesis testing. This analysis reveals the reactions with significantly differential first principal components between both high- and medium-turnover tissues and medium- and low-turnover tissues and reveals a list of reactions that—despite significant differential gene expression among these groups of samples—cannot be arrived upon using traditional differential gene expression analytical techniques.

Traditional pathway enrichment analysis considering reactions as genesets would not be appropriate here as genes are duplicated across reactions and attempting pathway enrichment with gene identifiers won't be able to capture this multiple-participation paradigm. Also, pathway enrichment using genes in cases where only a few (of many) genes who participate in a given reaction are present in this dataset would not show significant representation of that reaction's corresponding pathways, despite these few genes providing otherwise good tissue separability, leading to the potential for high false negative (Figure 13). This concept is exemplified in the difference between significant reaction and transcript pathway enrichment results.

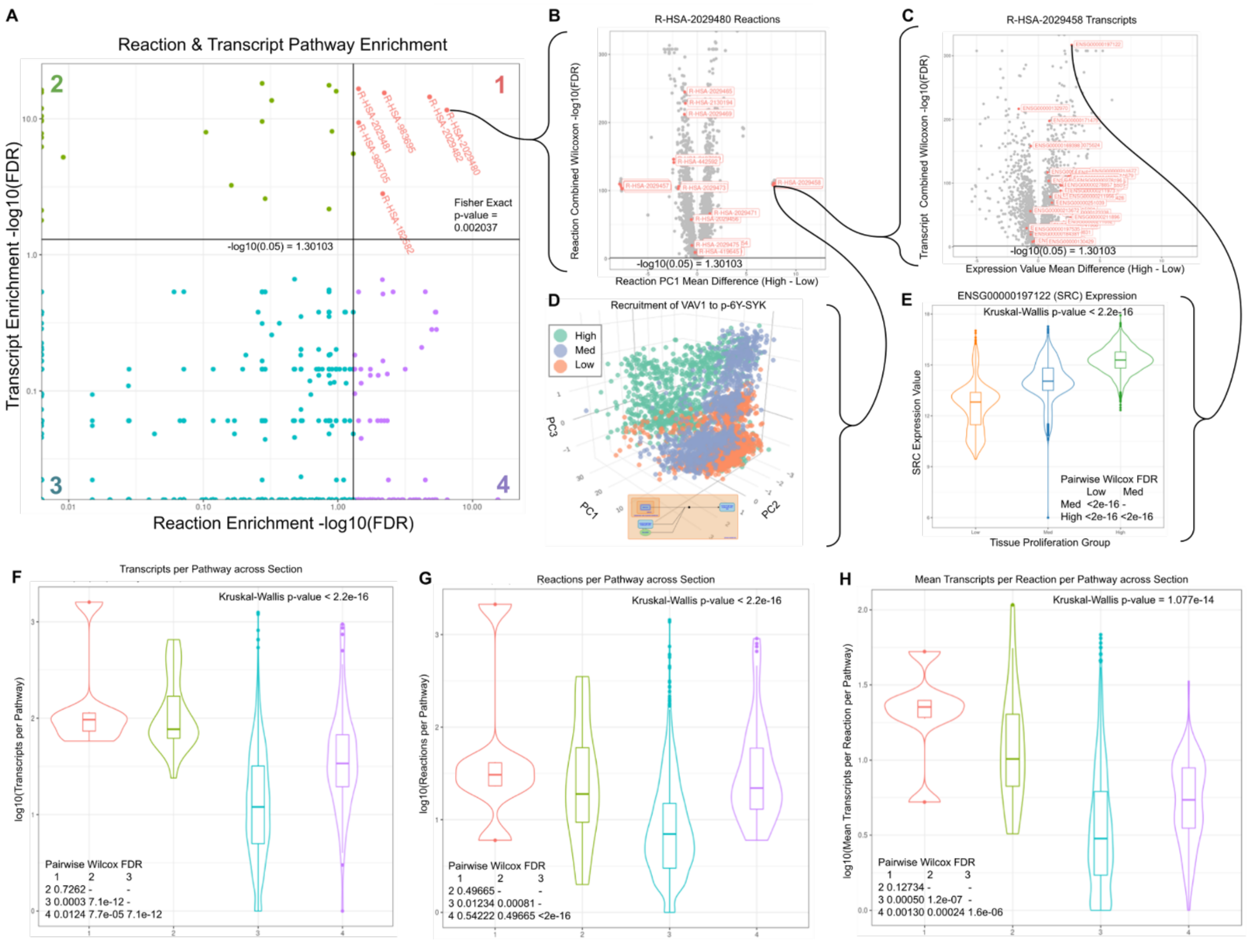


Figure 13: Panel A shows a log-log plot of reaction and pathway enrichment with four significance sections numbered and colored: 1/pink) pathways significant in both reaction and transcript enrichments, 2/green) pathways significant in transcript enrichment alone, 3/cyan) pathways not found to be significant in either enrichment analysis and 4/purple) pathways significant in reaction enrichment alone. Significance was defined as an FDR < 0.05 using hypergeometric test. The overlap set of pathways significant in both reaction and transcript enrichments is shown to be significant itself (Fisher exact p-value = 0.002037. Panel B shows reactions plotted by their combined Wilcoxon FDR and high vs low proliferation group mean PC1 difference with those from *Fcgamma receptor (FCGR) dependent phagocytosis* (R-HSA-2029480) the top-rightmost pathway (from A) labelled in pink. Panel C shows transcripts plotted by their combined Wilcoxon FDR and high vs low proliferation group mean expression difference with those from *Recruitment of VAV1 to p-6Y-SYK* (R-HSA-2029458) labelled in pink. Panel D shows tissue samples positioned within a space whose axes are the first three principal components of transcripts annotated to *Recruitment of VAV1 to p-6Y-SYK* (R-HSA-2029458) with high, medium and low proliferation samples labelled in green, blue and orange, respectively. A reactome diagram of the reaction is inset. Panel E shows expression distributions of SRC (ENSG00000197122) across the three proliferation groups with colors matching D. These distributions are shown to differ significantly with Kruskal-Wallis and Pairwise Wilcoxon FDR

results stated in the main text and shown in the panel itself. Panels F, G and H show distributions of transcripts, reactions and mean transcripts per reaction across pathways, respectively. Distributions represent pathways within significance sections from A with corresponding number labels and colors. These distributions are shown to differ significantly with Kruskal-Wallis and Pairwise Wilcoxon FDR results stated in the main text and shown in the panels themselves.

Setting an alpha = 0.05 and considering the overlapping enriched pathways between reaction and transcript enrichment approaches, I see these pathways are drawn from the same underlying distribution (Fisher Exact p-value = 0.002). Further, reviewing prior literature apropos of our significantly enriched pathways, I find support for cell proliferation-related mechanisms. Using Fisher's method to combine p-values, the most significant pathways have all been associated with proliferation: *Regulation of actin dynamics for phagocytic cup formation*, *Fcgamma receptor (FCGR) dependent phagocytosis* and *FCGR activation* (Luo Y et al., 2010); *CD22 mediated BCR regulation* (Matsubara N et al., 2018) and *Classical antibody-mediated complement activation* (Ghebrehiwet B. 2018). The 10 most significant pathways using the Fisher's method are reported in Table 7 along with results for reaction enrichment significance in Table 8 and transcript enrichment significance in Table 9.

Table 7: Enriched pathways with highest combined FDR

Pathway Name	Pathway ID	Combined FDR
Regulation of actin dynamics for phagocytic cup formation	R-HSA-2029482	1.731E-019
Fcgamma receptor (FCGR) dependent phagocytosis	R-HSA-2029480	6.187E-019
CD22 mediated BCR regulation	R-HSA-5690714	9.745E-019

Classical antibody-mediated complement activation	R-HSA-173623	9.745E-019
FCGR activation	R-HSA-2029481	1.289E-018
Antigen activates B Cell Receptor (BCR) leading to generation of second messengers	R-HSA-983695	2.520E-018
Role of LAT2/NTAL/LAB on calcium mobilization	R-HSA-2730905	3.258E-017
Creation of C4 and C2 activators	R-HSA-166786	1.002E-015
Chromatin organization	R-HSA-4839726	2.859E-015
Chromatin modifying enzymes	R-HSA-3247509	2.859E-015

Table 8: Enriched pathways with highest reaction FDR

Pathway Name	Pathway ID	Reaction FDR
Chromatin organization	R-HSA-4839726	3.836E-016
Chromatin modifying enzymes	R-HSA-3247509	3.836E-016
RUNX1 and FOXP3 control the development of regulatory T lymphocytes (Tregs)	R-HSA-8877330	0.0000003280
Fcgamma receptor (FCGR) dependent phagocytosis	R-HSA-2029480	0.0000003624
HDMs demethylate histones	R-HSA-3214842	0.000004028
O-linked glycosylation of mucins	R-HSA-913709	0.000004028
Interleukin-35 Signalling	R-HSA-8984722	0.000004485
Base Excision Repair	R-HSA-73884	0.000005247

Signaling by Non-Receptor Tyrosine Kinases	R-HSA-9006927	0.000005607
Signaling by PTK6	R-HSA-8848021	0.000005607

Table 9: Enriched pathways with highest transcript FDR

Pathway Name	Pathway ID	Transcript FDR
Classical antibody-mediated complement activation	R-HSA-173623	6.227E-019
CD22 mediated BCR regulation	R-HSA-5690714	2.327E-018
FCGR activation	R-HSA-2029481	2.446E-017
Creation of C4 and C2 activators	R-HSA-166786	5.442E-017
Role of LAT2/NTAL/LAB on calcium mobilization	R-HSA-2730905	1.132E-016
Scavenging of heme from plasma	R-HSA-2168880	2.451E-016
Antigen activates B Cell Receptor (BCR) leading to generation of second messengers	R-HSA-983695	3.193E-016
Regulation of actin dynamics for phagocytic cup formation	R-HSA-2029482	3.296E-015
Role of phospholipids in phagocytosis	R-HSA-2029485	0.00000000000002231
Initial triggering of complement	R-HSA-166663	0.000000000001223

Though pathway enrichment result overlap is significant, reaction and transcript results exhibit compositional differences. I investigated our enrichment results by separating pathways into significance sections 1) significant using both reaction and transcript enrichment approaches, 2) significant using the transcript approach only, 3) not significant using either approach and 4) significant using the reaction approach only. Transcript counts within pathways across these sections vary significantly (Kruskal-Wallis p-value < 2.2e-16) as well as between sections 2 and 3 (Pairwise Wilcoxon FDR = 7.1e-12), 2 and 4 (Pairwise Wilcoxon FDR = 7.7e-5) and 3 and 4 (Pairwise Wilcoxon FDR = 7.1e-12). Reaction counts within pathways vary significantly across section (Kruskal-Wallis p-value < 2.2e-16) as well as between sections 1 and 3 (Pairwise Wilcoxon FDR = 0.01234), 2 and 3 (Pairwise Wilcoxon FDR = 0.00081) and 3 and 4 (Pairwise Wilcoxon FDR < 2e-16). Investigating further, the mean transcript counts within reactions within pathways differ significantly across section (Kruskal-Wallis p-value = 1.077e-14) as well as between sections 1 and 3 (Pairwise Wilcoxon FDR = 0.00050), 1 and 4 (Pairwise Wilcoxon FDR = 0.00130), 2 and 3 (Pairwise Wilcoxon FDR = 1.2e-7), 2 and 4 (Pairwise Wilcoxon FDR = 0.00024) and 3 and 4 (Pairwise Wilcoxon FDR = 1.6e-6).

After performing pathway enrichment analysis using the reaction-based approach above, I investigated the significant pathway *Fcgamma receptor (FCGR) dependent phagocytosis* (R-HSA-2029480) at the reaction level. A volcano plot highlights the reaction first principal component difference between high and low tissue proliferation groups plotted against FDR values resulting from Wilcoxon p-values comparing high and medium and medium and low proliferation groups combined using Fisher's method (Figure 13B). Reactions in R-HSA-2029480 are labelled in pink. By selecting the significant reaction *Recruitment of VAV1 to p-6Y-SYK* (R-HSA-2029458), one can investigate further by showing sample structure within this reaction-defined space using the first three principal components where high, medium and low proliferation groups are indicated in green, blue and orange, respectively, along with an inset reaction diagram provided by Reactome

(Figure 13D). Next, in a similar fashion to above, a volcano plot shows transcript expression value mean difference between high and low tissue proliferation groups plotted against FDR values resulting from Wilcoxon p-values comparing high and medium and medium and low proliferation groups combined using Fisher's method. Transcripts in reaction R-HSA-2029458 are labelled in pink. I now select transcript ENSG0000017122 (SRC) and investigate further using by showing its expression distributions across proliferation group. SRC is a known proto-oncogene (https://www.genecards.org/cgi-bin/carddisp.pl?gene=SRC) and has been associated with cell proliferation in prior work (Sánchez-Bailón MP, 2012). Indeed, its expression is directly associated with proliferation and varies significantly among these three groups (Kruskal-Wallis p-value < 2.2e-16) as well as between pairs (Pairwise Wilcoxon FDR < 2e-16 between each pair).

Pathways significantly enriched by reactions or transcripts—but not both—are co-constituents of some top-level pathways but appear alone in others. The top-level pathways sharing results from both enrichment results are *Immune System*, *Signal Transduction* and *Hemostasis*. Reaction enrichment alone represents pathways belonging to *Chromatin organization*, *Gene expression*, *Metabolism of proteins*, *DNA Repair*, *Cell Cycle*, *Cellular responses to external stimuli*, *DNA Replication*, *Programmed Cell Death*, *Metabolism* and *Cell-Cell communication*. And transcript enrichment alone represents pathways belonging to *Vesicle-mediated transport*, *Sensory Perception* and *Muscle contraction*. The breadth of top-level pathway representation observed demonstrates the complementary nature of reaction-based pathway analysis performed on the underlying principal component data, relative to traditional enrichment analysis (see Appendix A).

Chapter 3: Infer Human Tissue-specific Reaction Networks

Geometric Deep Learning Architecture

In order to infer tissue-specific biochemical networks, I used known pathway hierarchy and reaction network structural annotations to generate geometric deep learning architectures with corresponding features (Figure 14). I then trained these architectures to classify tissue labels based on our reaction PC1 values and perform analysis on edge weightings arrived at by these architectures. I consider the biochemical network as a combination of an underlying preceding/following reaction network with an overarching pathway hierarchy. I separate these networks, generate deep learning architectures which represent them and train these architectures independently using the same 10,726 reaction PC1 values. Both architectures rely upon the 1-GNN described by Morris et al., (2019) and implemented in the Pytorch Geometric (Fey & Lenssen, 2019) *GraphConv()* function, using a summing aggregator and three layers of size 64, as described in the original article. This structure was benchmarked using an independent dataset, TU-molecular dataset (Morris et al., 2020), showing 64 the most performant batch size, which I used for our deep learning architectures. I select cross-entropy as our loss function and report accuracy as proportion of correct tissue classifications.

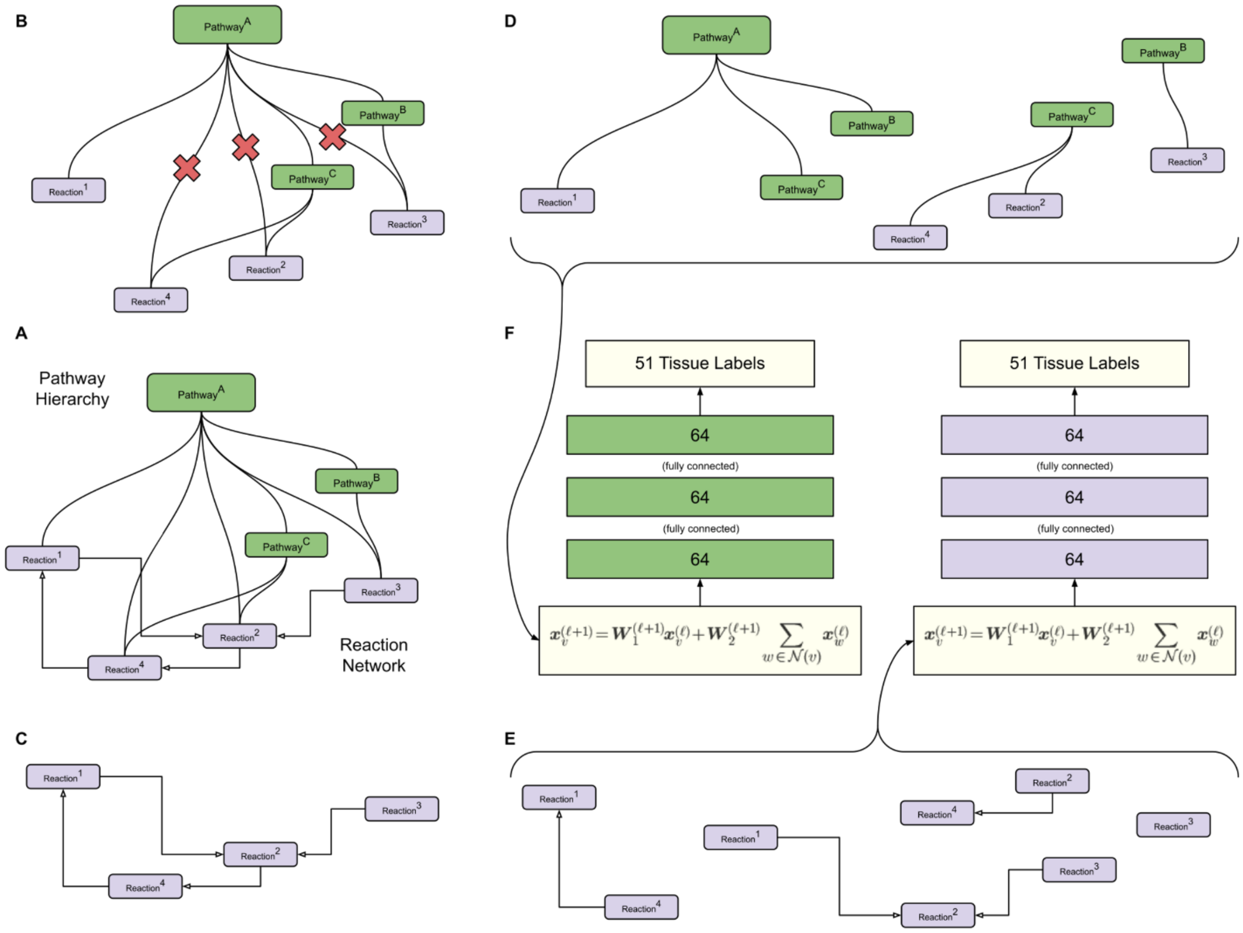


Figure 14: Panel A depicts the reactome pathway hierarchy and reaction network. Panel B shows the pathway hierarchy with edges connecting reactions to top-level pathways removed. Panel C shows the reaction network. Panel D shows the pathway hierarchy decomposition for the pathway hierarchy deep learning architecture. Panel E shows the reaction network decomposition for the reaction network deep learning architecture. Panel F shows both the pathway hierarchy and reaction network architectures with the default GraphConv aggregator function, passed through three layers of size 64 and trained to classify 51 tissue types.

The reaction network contains directed edges connecting pairs of reactions which is decomposed into 10,726 sub networks where each component is represented by a following reaction and its preceding reactions. The preceding reactions are considered the following reaction's neighborhood and their values are aggregated with a summation function. The pathway hierarchy was downloaded using the Reactome.org content service (available at https://reactome.org/dev/content-service) and differs from a pathway hierarchy used for pathway enrichment analysis in that its edges connect reactions to their nearest ancestral pathways, rather than pathways at all levels of the pathway hierarchy to which they belong. This results in input data consisting of the same 10,726 reaction PC1 values as well as 2,248 pathways. To add these pathways to our network structure in an unbiased way, I set each pathway vertex value to 1 across all samples.

Experimental Controls

To approximate an upper bound on the degree to which our reaction PC1 values explain tissue label, I used Resnet18 (He et al., 2016), a deep learning model recommended as a default selection (Pointer, 2019). Resnet18 was developed by Microsoft to classify images with 18 convolutional layers. Because Resnet18 is an image classifier, I reshaped our reaction PC1 values from a 1-dimensional vector of length 10,726 to two 2-dimensional vectors of height 173 and width 31 using the PyTorch (Paszke et al., 2017) *reshape()* function after redefining the first layer as a 2-channel, rather than the default Resnet18 3-channel used for RGB images. I also modified the final layer to output tissue labels. It may be possible for additional information about biological network structure to be subtly encoded in this input data by positioning biologically-related reaction PC1 values near each other in the resulting 2-channel matrix pairs. To test whether this was the case, I shuffled the reaction PC1 values and retrained our Resnet18 model; however, these results were indistinguishable, suggesting such additional information does not contribute to our Resnet18 model performance. To set lower bounds on our performance, I used our biologically-inspired deep learning architectures with randomly shuffled reaction PC1 values within samples and randomly shuffled tissue labels across samples. Curiously, I see randomly shuffling our reaction

PC1 values merely reduces accuracy to about 50% while randomly shuffling tissue labels reduces accuracy to about 5%, approximately random chance. I can imagine a case where the sum of reaction PC1 values for some tissue is high and the sum of reaction PC1 values for another tissue is low; such may be the case with the relatively high accuracy of the randomly shuffled reaction PC1 value negative control. Given our tissue sample sizes are uneven, one would not expect a random classifier to perform quite as poorly as 1/51 (< 2%); the approximately 5% accuracy of randomly shuffled tissue labels is a reflection of this sample size imbalance (Figure 15).

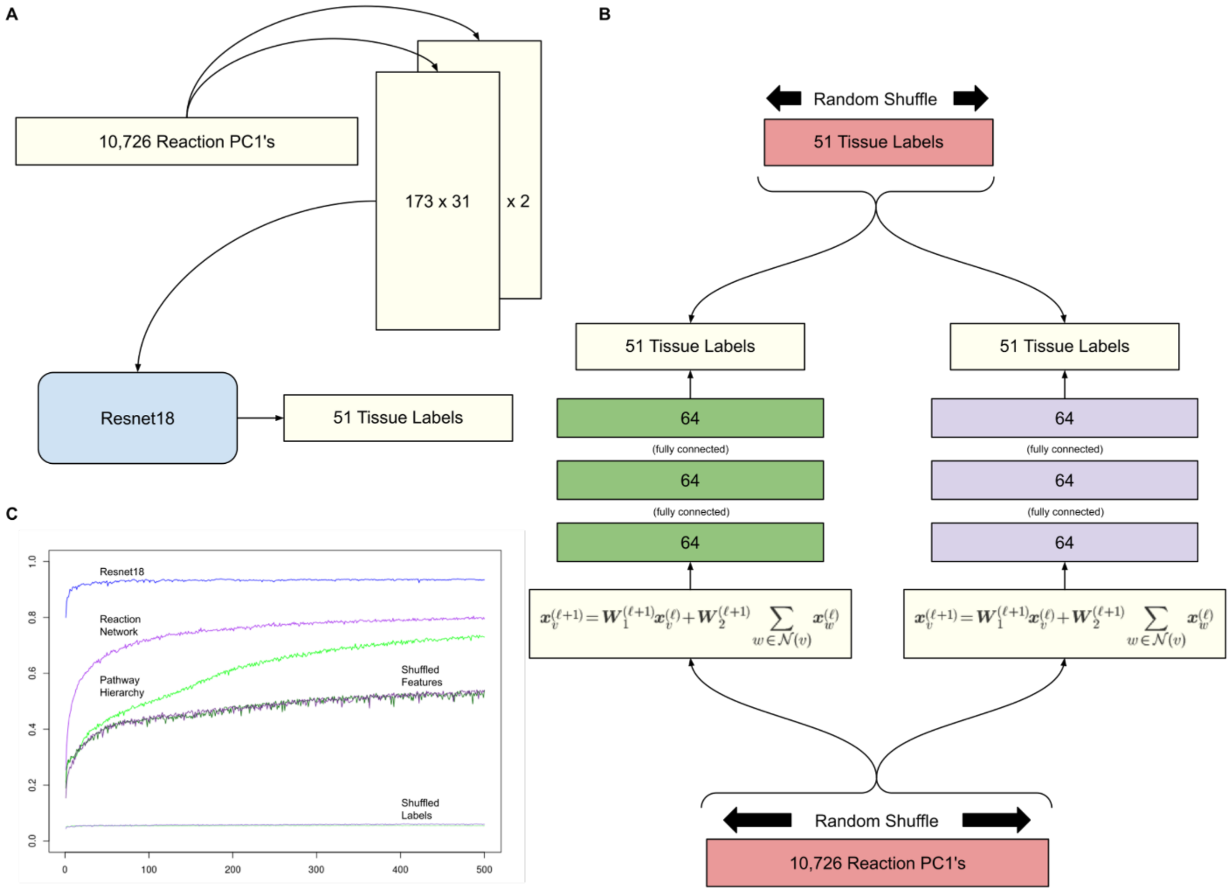


Figure 15: Panel A shows the reaction pc1 values reshaped into image-like tensors for Resnet18, used to classify 51 tissues as a positive control. Panel B shows both randomly shuffled reaction pc1 values and randomly shuffled tissue labels are used as features and targets, respectively, as negative controls. Panel C shows the performance as proportion correct for both our biologically-inspired deep learning architectures as well as our controls.

Results

In order to arrive at reaction networks specific to each tissue, I extracted edge weights for each tissue label using the *Integrated Gradients* technique described by Sundararajan et al. (2017) as implemented in the Captum PyTorch library (Kokhlikyan et al., 2009). Edge weights are arrived upon by averaging gradients from predicted output to input features. Across all tissues, there is a positive Pearson Correlation Coefficient (PCC) between edge weight and both ARI and ACC of the preceding reaction, calculated using our KNN classifier (Figure 16). Similarly, using the pathway hierarchy architecture, we see edge weights correlated with both ARI and ACC across edges whose preceding vertex is a reaction and we see edge weights correlated with tissue-specific pathway enrichment across edges whose preceding vertex is a pathway. For example, pathway *Fc epsilon receptor (FCERI) signaling* (R-HSA-2454202) is heavily weighted (~0.88), significantly enriched in lung tissue based on wilcoxon upper tail transcript expression FDR < 0.05 and hypergeometric enrichment FDR (FDR = 9.586511e-10) and is supported by the literature as a critical regulator in lung tissue (Gounni et al., 2006).

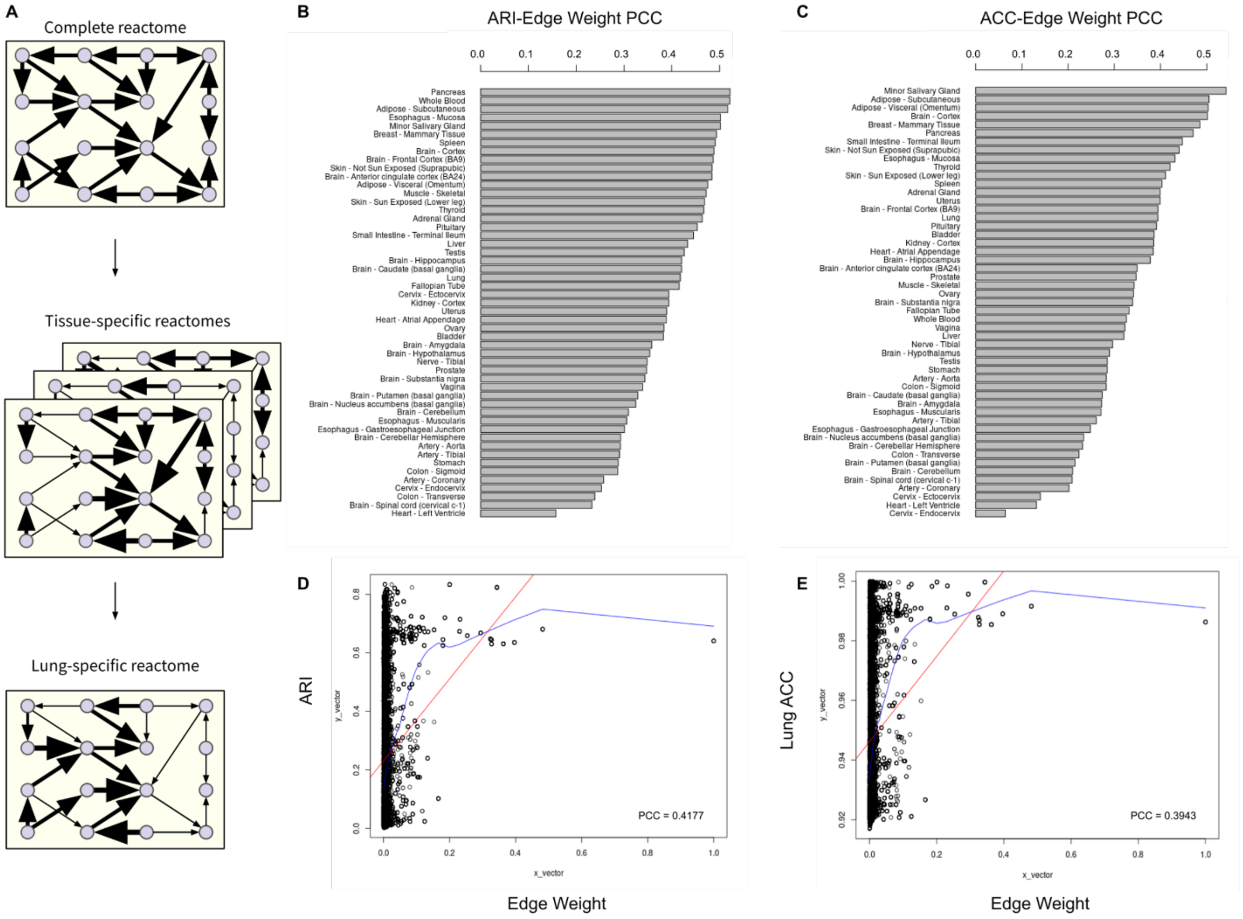


Figure 16: Panel A depicts the existing model of the reaction network with unweighted edges used to generate tissue-specific reactomes with weighted edges along with an example of a tissue-specific reactome. Panel B shows positive pearson correlation coefficients across all tissues of reaction network edge weights with the preceding reaction ARI, calculated using our KNN model. Panel C shows positive pearson correlation coefficients across all tissues of reaction network edge weights with the preceding reaction ACC, calculated using our KNN model. Panels D and E exemplify data underlying B and C for Lung tissue.

To investigate the superstructure from which our tissue-specific biochemical networks are derived, I generated 10 artificial reaction networks with degree-preserving randomization (Ray et al., 2012), each representing a random reaction network (Figure 17). I trained these artificial reaction networks to classify tissue using the same reaction PC1 values as our original reaction network only without any held-out samples. This allowed our artificial reaction network models to see all samples during training and overfit it, if they were able. I compared the training accuracy of these artificial results to the held-out test accuracy of our original reaction network. The original reaction network acted as a better learning structure (upper-tail wilcoxon p-value = 0.0014), which suggests the shape of the biochemical network itself is biologically meaningful and that tissue chemistry is able to take advantage of this structure to organize information across tissue. Deviation from this network structure negatively affects information content such that rewired deep learning architectures achieve significantly lower accuracy, though still performing better than our random data negative controls.

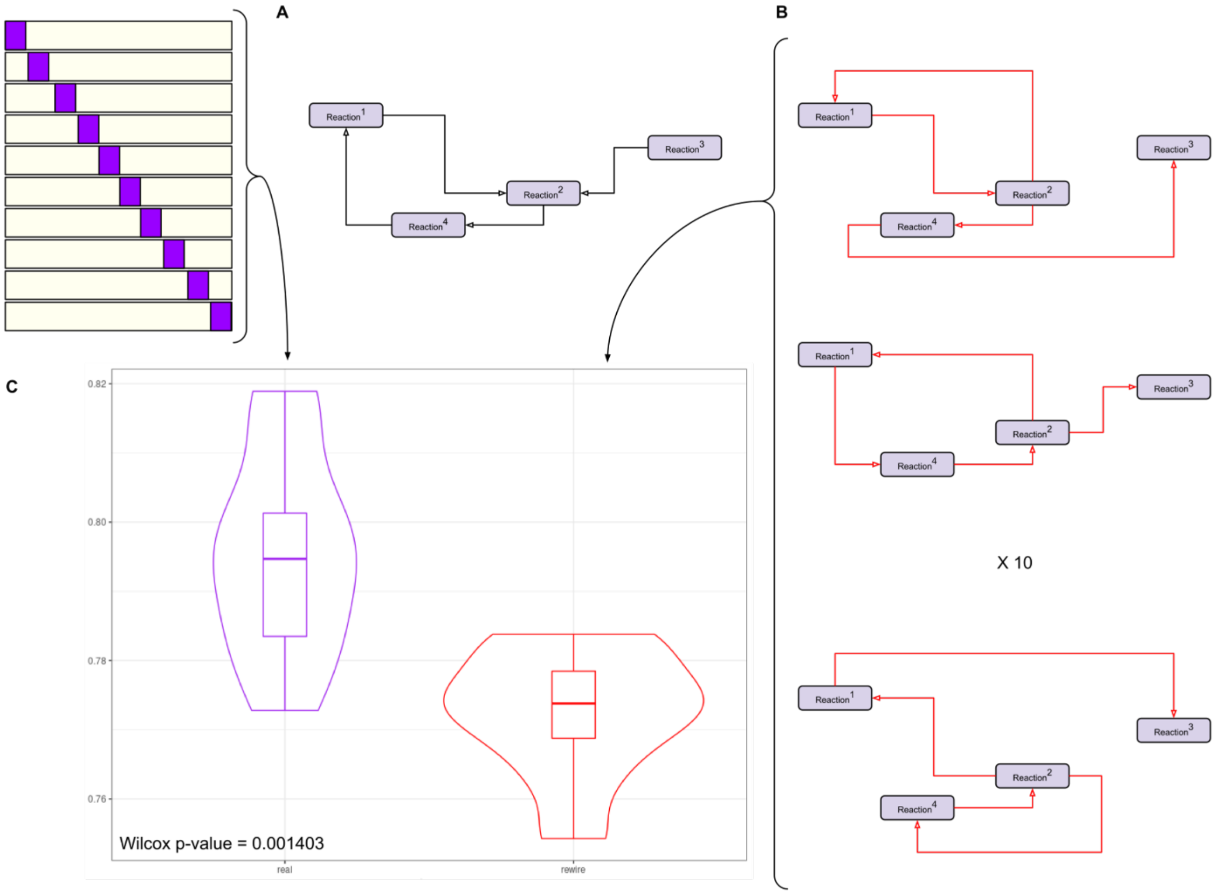
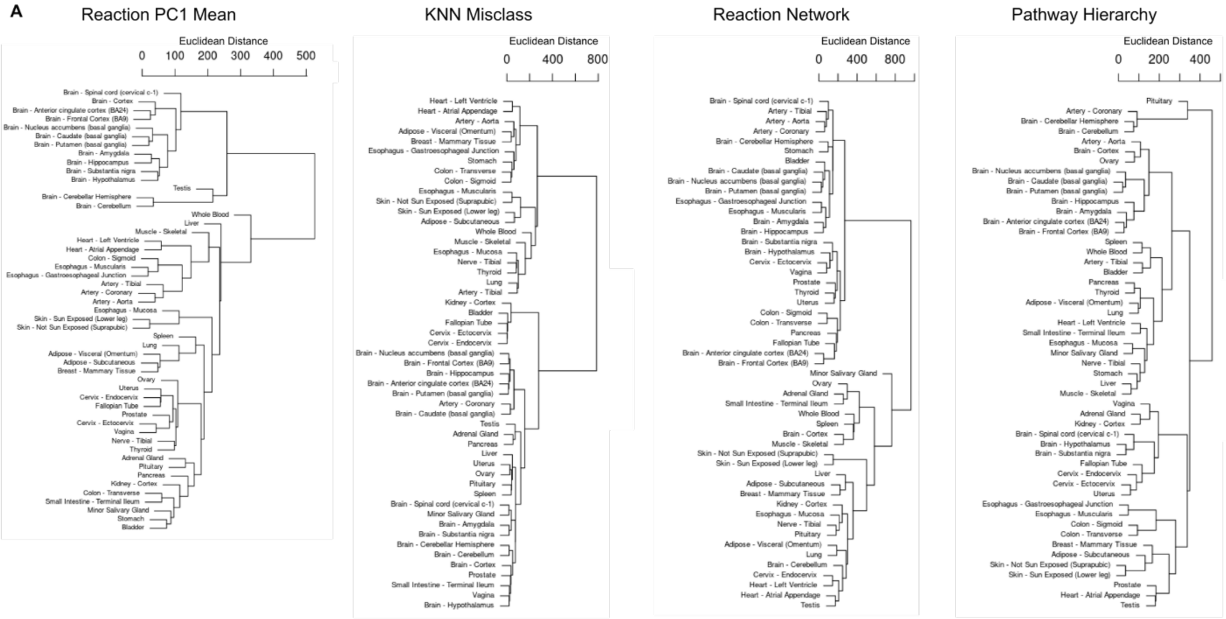


Figure 17: Panel A shows the real reaction network is used with real reaction pc1 values to generate a distribution of test accuracy values for held-out data from 10 cross-validation folds. Panel B shows 10 artificial reaction networks are generated using degree-

preserving rewiring with real data to generate a distribution of training accuracy values for the entire dataset. Panel C shows the real network performs significantly better on held-out test data than the artificial networks do on real training data.

To assess the reaction network and the pathway hierarchy structure more generally, I performed hierarchical clustering on each network's edge weights across tissue and quantitatively compare the results with hierarchical clusterings of our reaction PC1 values and our KNN misclassification rates across tissue (Figure 18). I calculate the cophenetic correlation of these hierarchies showing the reaction PC1 tissue clustering is more closely correlated with the pathway hierarchy and KNN tissue clustering results than the reaction network tissue clustering.



B

Cophenetic Correlations

	Reaction PC1 Mean	KNN Misclass	Reaction Network	Pathway Hierarchy
Reaction PC1 Mean	1	0.1232	0.03856	0.1731
KNN Misclass	0.1232	1	0.0460	0.0865
Reaction Network	0.03856	0.0460	1	0.0411
Pathway Hierarchy	0.1731	0.0865	0.0411	1

Figure 18: Panel A shows four tissue hierarchical clustering dendrograms giving a high-level overview of how each dataset organizes tissues. Panel B shows cophenetic correlation among the four dendrograms.

A tissue hierarchy published in prior literature and reproduced here in Figure 19 from Pierson et al., 2015 cannot be quantitatively compared in the same way as the data explicitly specifying their tree structure (e.g. underlying Euclidean distance matrix) is not provided but merely depicted in the figure. However, several patterns depicted are observed in our tissue hierarchies as well. The authors highlight clusters for several tissue groups including the cerebellum, other brain tissue, digestive tissue, artery, adipose tissue, heart and skin. In our reaction PC1 tissue clustering, cerebellum, other brain tissue, artery, adipose, heart and skin are grouped similarly while digestive tissue is largely grouped together except that Sigmoid Colon (not present in Pierson et al.) is grouped with esophagus. In our KNN tissue clustering, cerebellum is grouped together but other brain tissue differs from the literature and appears more widely distributed, digestive tissue is largely grouped together but Small Intestine (not present in Pierson et al.) is grouped separately. Artery tissue is not grouped together but adipose tissue, heart and skin are grouped similarly. In our reaction network tissue clustering, cerebellum and other brain tissues are not grouped similarly to Pierson et al. though basal ganglia is. Colon is grouped together but Small Intestine and Stomach are not. Artery tissue is grouped together and adipose tissue is grouped similarly as well as heart and skin tissues. In our pathway hierarchy tissue clustering, cerebellum is grouped together as is basal ganglia. Other brain tissues are more distributed. Colon tissues are grouped together but Stomach and Small Intestine are separate. Artery tissues are not grouped together and adipose and heart tissues are not grouped similarly but skin is.

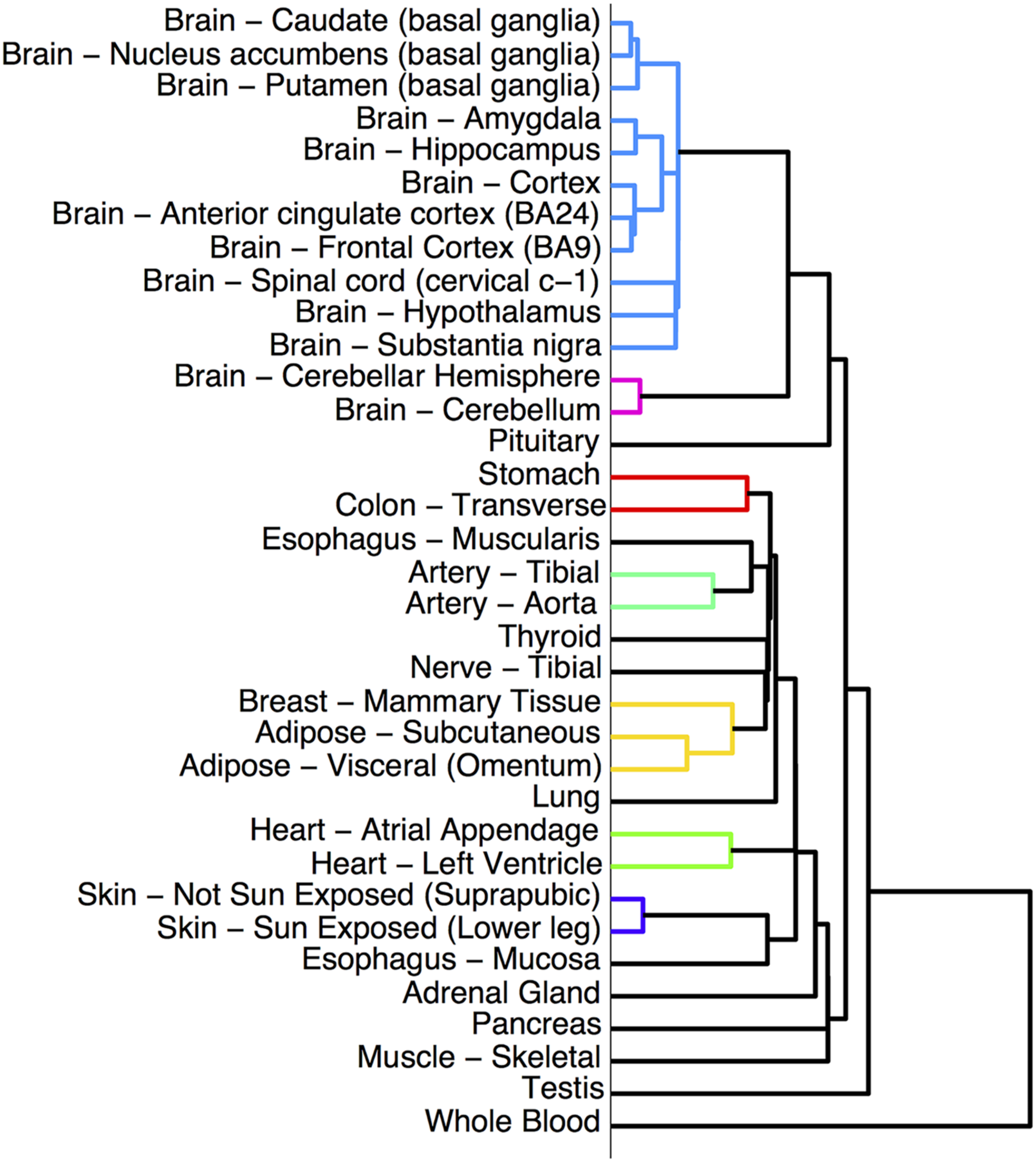


Figure 19: Tissue hierarchical clustering dendrogram from Pierson et al., 2015 (Reproduced with permission from the editor).

Chapter 4: Discussion and Future Directions

Summary

This project sought to identify and explain the biochemical reactions responsible for differential healthy human tissue functions by classifying healthy human tissues using biochemical reaction states inferred using machine learning models. In order to accomplish this, I used healthy human tissue-specific gene expression data, testing for batch effects and other sources of error, storing the data and our software in a common form, amenable to network analytic methods and accessible to the community. I arrived at a mapping from genes to biochemical reactions that the community may leverage to transform mRNA expression values and express within a network context. Constraining potential relationships to those found within the Reactome pathway hierarchy and the Reactome preceding/following reaction network, I generated deep learning architectures representing hierarchical pathway relationships and reactant-product relationships, respectively. I assessed the significance of our findings by controlling with positive and negative performance controls along with randomly rewiring in a degree-preserving way and compared final tissue hierarchies with those built using our input data and reported in prior literature. Finally, I conclude our work and suggest future directions of our project that may further enhance its contribution to translational medicine.

Technical Contributions

We consider three technological contributions developed in this project. The first is our reaction PCA pathway analysis method, which provides a multi-level perspective of pathway enrichment, complementing traditional gene expression-based analysis. Next, I consider our fully-trained deep learning models may be adapted for other classification, clustering or regression tasks related to human tissue or other phenotype associations. Finally, I believe our tissue-specific edge-sets may be used to construct refined geometric deep learning architectures, restricting degrees of freedom.

Findings

The Reactome preceding/following reaction network exhibits a non-random shape, suggesting a novel signature of evolution. It strikes us as natural to assume tissue biochemistry leverages differential subcomponents of the overall biochemical network to perform certain functions; however, at a global scale, it is not obvious that a significantly more accurate architecture must emerge. I look forward to further research that may shed light on this curious finding.

The novel human tissue hierarchies developed by our reaction PCA, nearest neighbor misclassification and geometric deep learning architectures largely recapitulate local tissue groupings observed in prior work but remote tissue groupings show variability. I envision future consideration of the tissue organizations I observe may drive future hypotheses.

Considerations

This project was limited by several factors and compromises that must be considered alongside our results. Firstly, the data I perform our analysis with exhibits several issues. Our sample size of merely 9,115 represents an infinitesimal fraction of the human population and is much smaller than I had set out to use. I chose to use this sample size after careful review of available data, finding no additional data suitable for our purposes at the time of writing. This data itself has several limitations. Consider, I use samples collected from cadaver tissue, indicating all donors experienced death due to some cause. The tissues from which the samples were collected may have appeared healthy; however, a donor may have suffered an unidentified illness, such as metastatic cancer, or experienced some temperature fluctuation or other environmental exposure perhaps leading to misrepresented healthy tissue gene expression patterns. The samples were collected using bulk-RNA sequencing, which provides only an average value of gene expression for each tissue. We know tissues are host to many cell types of varying lineage and it may be the case that only the primary cell types in these tissue samples are represented in our input data. I use RNA-seq data itself to approximate states of biochemical reactions but know transcript abundance only correlates with protein abundance, and not to reaction state itself. In the future, large-scale tissue-specific phosphoproteomic data may be able to address this shortcoming. Next, the biochemical networks described by Reactome are incomplete. I used only 6,323 transcripts for each sample that we believe act in reactions but know the human genome contains many more protein-coding genes. Reactome is increasing coverage over time but is not yet complete. Finally, our methods lose some information. I chose principal component analysis as a method for transcript set to reaction dimensionality reduction because it was a simple method, making no assumptions about our data, allowing us to maintain maximum naivete. Increasingly sophisticated methods are developed for studying RNA transcript values, utilization of which may explain more than a median of 50% of the variance our first principal component did. And parameters for our geometric deep learning architectures were left unaltered from the literature in which they were described. This choice likely translated into underperformance but allowed us to compare our results with random networks and negative controls in an unbiased way. Tuning parameters based on our biologically-inspired networks would have been dubious and brought results from our controls into question.

Future Directions

Breast tumor subtype—previously classified as belonging to one of either *ERPR*, *Her2 positive* or *TN*—is now thought to belong to an expanded set of the classes *Luminal A*, *Luminal B*, *Her2-enriched* and *basal-like* and may even be further considered as belonging to one of ten varieties identified using clustering analysis (Curtis et al., 2012 and Tyanova et al., 2016). By analyzing primary tumors from TCGA within the context of their tissue-specific biochemical networks, we may be able to identify cancer subtype-specific patterns within and among Reactome pathways. Crucially, because our final model will have been trained on mRNA expression values alone, we'll be able to compare results across samples with varying mutational burdens and protein abundancies and generate testable hypotheses about individual protein-mRNA relationships and individual mutations' downstream consequences.

With the scaling of scRNA-sequencing experiments (Svensson et al., 2018) and the aggregation of their results in publicly available databases (Cao et al., 2017), the generation of cell type- and cell subtype-specific Reactome pathway hierarchies and Reactome preceding/following reaction networks become near-term possibilities. In a single-cell sequencing setting, I may be able to apply this machine learning strategy to produce within-tissue cell population-specific network models, allowing us to better understand biological processes such as development, response to injury, innate immunity and cellular disease at finer resolution.

Our discriminative, *top-down* approach to tissue-specific biochemical network generation essentially down-weights connections between reactions considered less characteristic of that tissue. This approach succeeds in limiting the false positive rate to that of Reactome but provides no guidance regarding novel connections likely characteristic of a given tissue. Experimental approaches require time and resources. In a similar fashion to the way our approach addresses false positives by constraining our resulting tissue-specific networks' edge sets to subsets of each Reactome network's edge set, by leveraging our reaction mean expression, ACC, ARI and non-random network structure findings, I may be able to expand the network by adding *potential reaction relationship* edges, initiated with some small weights, that will allow machine learning algorithms to consider reaction connections not currently specified by Reactome.

References

Aldridge BB, Burke JM, Lauffenburger DA, Sorger PK. Physicochemical modelling of cell signalling pathways. Nature cell biology. 2006 Nov;8(11):1195.

Aleksandrov L, Maheshwari A, Sack JR. Approximation algorithms for geometric shortest path problems. InProceedings of the thirty-second annual ACM symposium on Theory of computing 2000 May 1 (pp. 286-295). ACM.

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Børresen-Dale AL, Boyault S. Signatures of mutational processes in human cancer. Nature. 2013 Aug;500(7463):415.

Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences. 1999 Jun 8;96(12):6745-50.

Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. Bioinformatics. 2010 Apr 12;26(10):1340-7.

Arnedos M, Vicier C, Loi S, Lefebvre C, Michiels S, Bonnefoi H, Andre F. Precision medicine for metastatic breast cancer—limitations and solutions. Nature reviews Clinical oncology. 2015 Dec;12(12):693.

Ashley EA. The precision medicine initiative: a new national effort. Jama. 2015 Jun 2;313(21):2119-20.

Ashley EA. Towards precision medicine. Nature Reviews Genetics. 2016 Sep;17(9):507.

Ayers M, Lunceford J, Nebozhyn M, Murphy E, Loboda A, Kaufman DR, Albright A, Cheng JD, Kang SP, Shankaran V, Piha-Paul SA. IFN-γ–related mRNA profile predicts clinical response to PD-1 blockade. The Journal of clinical investigation. 2017 Aug 1;127(8):2930-40.

Azimifar SB, Nagaraj N, Cox J, Mann M. Cell-type-resolved quantitative proteomics of murine liver. Cell metabolism. 2014 Dec 2;20(6):1076-87.

Baldwin RL. Temperature dependence of the hydrophobic interaction in protein folding. Proceedings of the National Academy of Sciences. 1986 Nov 1;83(21):8069-72.

Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A. NCBI GEO: archive for functional genomics data sets—update. Nucleic acids research. 2012 Nov 26;41(D1):D991-5.

Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, Joshi MB, Harpole D, Lancaster JM, Berchuck A, Olson Jr JA. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature. 2006 Jan;439(7074):353.

Blokzijl F, De Ligt J, Jager M, Sasselli V, Roerink S, Sasaki N, Huch M, Boymans S, Kuijk E, Prins P, Nijman IJ. Tissue-specific mutation accumulation in human adult stem cells during life. Nature. 2016 Oct;538(7624):260.

Botea A, Müller M, Schaeffer J. Near optimal hierarchical path-finding. Journal of game development. 2004 Mar;1(1):7-28.

Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nature biotechnology. 2016 May;34(5):525.

Burkhart J, TSAR: Time-Series Analysis tool for Respiratory Viral DREAM Challenge (syn7202976). Synapse. 2016. DOI: 10.7303/syn7202976.

Camacho DM, Collins KM, Powers RK, Costello JC, Collins JJ. Next-generation machine learning for biological networks. Cell. 2018 Jun 14;173(7):1581-92.

Cao Y, Zhu J, Jia P, Zhao Z. scRNASeqDB: a database for RNA-Seq based gene expression profiles in human single cells. Genes. 2017 Dec 5;8(12):368.

Carithers LJ, Ardlie K, Barcus M, Branton PA, Britton A, Buia SA, Compton CC, DeLuca DS, Peter-Demchok J, Gelfand ET, Guan P. A novel approach to high-quality postmortem tissue procurement: the GTEx project. Biopreservation and biobanking. 2015 Oct 1;13(5):311-9.

Ceccarelli M, Barthel FP, Malta TM, Sabedot TS, Salama SR, Murray BA, Morozova O, Newton Y, Radenbaugh A, Pagnotta SM, Anjum S. Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. Cell. 2016 Jan 28;164(3):550-63.

Changelian PS, Fearon DT. Tissue-specific phosphorylation of complement receptors CR1 and CR2. Journal of Experimental Medicine. 1986 Jan 1;163(1):101-15.

Chen X, Teichmann SA, Meyer KB. From Tissues to Cell Types and Back: Single-Cell Gene Expression Analysis of Tissue Architecture. Annual Review of Biomedical Data Science. 2018 Jul 20;1:29-51.

Cheriton D, Tarjan RE. Finding minimum spanning trees. SIAM Journal on Computing. 1976 Dec;5(4):724-42.

Ciriello G, Cerami E, Sander C, Schultz N. Mutual exclusivity analysis identifies oncogenic network modules. Genome research. 2012 Feb 1;22(2):398-406.

Coley CW, Jin W, Rogers L, Jamison TF, Jaakkola TS, Green WH, Barzilay R, Jensen KF. A graph-convolutional neural network model for the prediction of chemical reactivity. Chemical science. 2019;10(2):370-7.

Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, Hansen KD, Jaffe AE, Langmead B, Leek JT. Reproducible RNA-seq analysis using recount2. Nature biotechnology. 2017 Apr 11;35(4):319.

Costanzo M, VanderSluis B, Koch EN, Baryshnikova A, Pons C, Tan G, Wang W, Usaj M, Hanchard J, Lee SD, Pelechano V. A global genetic interaction network maps a wiring diagram of cellular function. Science. 2016 Sep 23;353(6306):aaf1420.

Coussens LM, Werb Z. Inflammation and cancer. Nature. 2002 Dec 19;420(6917):860.

Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, Gräf S. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012 Jun;486(7403):346.

Dainis AM, Ashley EA. Cardiovascular precision medicine in the genomics era. JACC: Basic to Translational Science. 2018 Apr 30;3(2):313-26.

De Raedt L, Guns T, Nijssen S. Constraint programming for data mining and machine learning. InTwenty-Fourth AAAI Conference on Artificial Intelligence 2010 Jul 5.

Dienstmann R, Vermeulen L, Guinney J, Kopetz S, Tejpar S, Tabernero J. Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. Nature Reviews Cancer. 2017 Feb;17(2):79.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan 1;29(1):15-21.

Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nature protocols. 2009 Aug;4(8):1184.

Eberwine J, Sul JY, Bartfai T, Kim J. The promise of single-cell sequencing. Nature methods. 2014 Jan 1;11(1):25.

Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research. 2002 Jan 1;30(1):207-10.

Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences. 1998 Dec 8;95(25):14863-8.

Ellis SE, Collado-Torres L, Jaffe A, Leek JT. Improving the value of public RNA-seq expression data by phenotype prediction. Nucleic acids research. 2018 Mar 5;46(9):e54-.

ENCODE Project Consortium. The ENCODE (ENCyclopedia of DNA elements) project. Science. 2004 Oct 22;306(5696):636-40.

Eroles P, Bosch A, Pérez-Fidalgo JA, Lluch A. Molecular biology in breast cancer: intrinsic subtypes and signaling pathways. Cancer treatment reviews. 2012 Oct 1;38(6):698-707.

Espinoza M. On Network Randomization Methods: A Negative Control Study. Department of Computer Science & Engineering. University of Connecticut. 2012.

Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, Haw R, Jassal B, Korninger F, May B, Milacic M. The reactome pathway knowledgebase. Nucleic acids research. 2017 Nov 14;46(D1):D649-55.

Fabregat A, Korninger F, Viteri G, Sidiropoulos K, Marin-Garcia P, Ping P, Wu G, Stein L, D'Eustachio P, Hermjakob H. Reactome graph database: Efficient access to complex pathway data. PLoS computational biology. 2018 Jan 29;14(1):e1005968.

Farris JS. On the cophenetic correlation coefficient. Systematic Zoology. 1969 Sep 1;18(3):279-85.

Fey M, Lenssen JE. Fast graph representation learning with PyTorch Geometric. arXiv preprint arXiv:1903.02428. 2019 Mar 6.

Fisher RA. On the interpretation of χ 2 from contingency tables, and the calculation of P. Journal of the Royal Statistical Society. 1922 Jan 1;85(1):87-94.

Formaneck MS, Ma L, Cui Q. Effects of temperature and salt concentration on the structural stability of human lymphotactin: Insights from molecular simulations. Journal of the American Chemical Society. 2006 Jul 26;128(29):9506-17.

Fourati S, Talla A, Mahmoudian M, Burkhart JG, Klen R, Henao R, Yu T, Aydın Z, Yeung KY, Ahsen ME, Almugbel R. A crowdsourced analysis to identify ab initio molecular signatures predictive of susceptibility to viral infection. Nature communications. 2018 Oct 24;9(1):4418.

Fowlkes EB, Mallows CL. A method for comparing two hierarchical clusterings. Journal of the American statistical association. 1983 Sep 1;78(383):553-69.

Friedman AA, Letai A, Fisher DE, Flaherty KT. Precision medicine for cancer with next-generation functional diagnostics. Nature Reviews Cancer. 2015 Dec;15(12):747.

Frühbeck G, Kiortsis DN, Catalán V. Precision medicine: diagnosis and management of obesity. The Lancet Diabetes & Endocrinology. 2018 Mar 1;6(3):164-6.

Fu J, Kammers K, Nellore A, Collado-Torres L, Leek JT, Taub MA. RNA-seq transcript quantification from reduced-representation data in recount2. BioRxiv. 2018 Jan 1:247346.

Galle PR, Forner A, Llovet JM, Mazzaferro V, Piscaglia F, Raoul JL, Schirmacher P, Vilgrain V. EASL clinical practice guidelines: management of hepatocellular carcinoma. Journal of hepatology. 2018 Jul 1;69(1):182-236.

Gastwirth JL. The estimation of the Lorenz curve and Gini index. The review of economics and statistics. 1972 Aug 1:306-16.

Gendelman R, Xing H, Mirzoeva OK, Sarde P, Curtis C, Feiler HS, McDonagh P, Gray JW, Khalil I, Korn WM. Bayesian network inference modeling identifies TRIB1 as a novel regulator of cell-cycle progression and survival in cancer cells. Cancer research. 2017 Jan 13.

Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. Nucleic acids research. 2016 Nov 28;45(D1):D331-8.

Gibson G. The environmental contribution to gene expression profiles. Nature reviews genetics. 2008 Aug;9(8):575.

Glass K, Huttenhower C, Quackenbush J, Yuan GC. Passing messages between biological networks to refine predicted interactions. PloS one. 2013 May 31;8(5):e64832.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. science. 1999 Oct 15;286(5439):531-7.

Gounni AS. The high-affinity IgE receptor (FcεRI): a critical regulator of airway smooth muscle cells?. American Journal of Physiology-Lung Cellular and Molecular Physiology. 2006 Sep;291(3):L312-21.

Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS, Zhang R, Hartmann BM, Zaslavsky E, Sealfon SC, Chasman DI. Understanding multicellular function and disease with human tissue-specific networks. Nature genetics. 2015 Jun;47(6):569.

Gremse M, Chang A, Schomburg I, Grote A, Scheer M, Ebeling C, Schomburg D. The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. Nucleic acids research. 2010 Oct 28;39(suppl_1):D507-13.

Hamilton W, Ying Z, Leskovec J. Inductive representation learning on large graphs. InAdvances in Neural Information Processing Systems 2017 (pp. 1024-1034).

Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. cell. 2011 Mar 4;144(5):646-74.

Han H, Shim H, Shin D, Shim JE, Ko Y, Shin J, Kim H, Cho A, Kim E, Lee T, Kim H. TRRUST: a reference database of human transcriptional regulatory interactions. Scientific reports. 2015 Jun 12;5:11432.

Han J, Zhu L, Kulldorff M, Hostovich S, Stinchcomb DG, Tatalovich Z, Lewis DR, Feuer EJ. Using Gini coefficient to determining optimal cluster reporting sizes for spatial scan statistics. International journal of health geographics. 2016 Dec;15(1):27.

Hanson E, Ballantyne J. Human Organ Tissue Identification by Targeted RNA Deep Sequencing to Aid the Investigation of Traumatic Injury. Genes. 2017 Nov 10;8(11):319.

He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. InProceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 770-778).

Hinton GE, Krizhevsky A, Wang SD. Transforming auto-encoders. InInternational Conference on Artificial Neural Networks 2011 Jun 14 (pp. 44-51). Springer, Berlin, Heidelberg.

Hinton GE. Representing part-whole hierarchies in connectionist networks. InProceedings of the Tenth Annual Conference of the Cognitive Science Society 1988 Aug (pp. 48-54).

Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, Bhai J, Billis K. Ensembl 2021. Nucleic Acids Research. 2021 Jan 8;49(D1):D884-91.

Huelsken J, Behrens J. The Wnt signalling pathway. Journal of cell science. 2002 Nov 1;115(21):3977-8.

Hutter F, Xu L, Hoos HH, Leyton-Brown K. Algorithm runtime prediction: Methods & evaluation. Artificial Intelligence. 2014 Jan 1;206:79-111.

Huttlin EL, Jedrychowski MP, Elias JE, Goswami T, Rad R, Beausoleil SA, Villén J, Haas W, Sowa ME, Gygi SP. A tissue-specific atlas of mouse protein phosphorylation and expression. Cell. 2010 Dec 23;143(7):1174-89.

Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167. 2015 Feb 11.

Jae Hwang S, Ravi SN, Tao Z, Kim HJ, Collins MD, Singh V. Tensorize, factorize and regularize: Robust visual relationship learning. InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018 (pp. 1014-1023).

Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007 Jan 1;8(1):118-27.

Karabulut NP, Frishman D. Sequence-and structure-based analysis of tissue-specific phosphorylation sites. PloS one. 2016 Jun 22;11(6):e0157896.

Keselj, S., Doshi, R., Nair, P. Capsule Network Experiments. skeselj. 2018 Jan 22. github.com. Accessed Nov. 28, 2018. https://github.com/skeselj/capsule-network-experiments

Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L. Human protein reference database—2009 update. Nucleic acids research. 2008 Nov 6;37(suppl_1):D767-72.

Ketkar N. Introduction to pytorch. InDeep learning with python 2017 (pp. 195-208). Apress, Berkeley, CA.

Kim BJ, Yoon CN, Han SK, Jeong H. Path finding strategies in scale-free networks. Physical Review E. 2002 Jan 23;65(2):027103.

Klipp E, Liebermeister W. Mathematical modeling of intracellular signaling pathways. BMC neuroscience. 2006 Oct;7(1):S10.

Kluyver T, Ragan-Kelley B, Pérez F, Granger BE, Bussonnier M, Frederic J, Kelley K, Hamrick JB, Grout J, Corlay S, Ivanov P. Jupyter Notebooks-a publishing format for reproducible computational workflows. InELPUB 2016 May 26 (pp. 87-90).

Kokhlikyan N, Miglani V, Martin M, Wang E, Alsallakh B, Reynolds J, Melnikov A, Kliushkina N, Araya C, Yan S, Reblitz-Richardson O. Captum: A unified and generic model interpretability library for pytorch. arXiv preprint arXiv:2009.07896. 2020 Sep 16.

Kopan R, Ilagan MX. The canonical Notch signaling pathway: unfolding the activation mechanism. Cell. 2009 Apr 17;137(2):216-33.

Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal. 2015 Jan 1;13:8-17.

Kramer M, Dutkowski J, Yu M, Bafna V, Ideker T. Inferring gene ontologies from pairwise similarity data. Bioinformatics. 2014 Jun 11;30(12):i34-42.

Krogan NJ, Lippman S, Agard DA, Ashworth A, Ideker T. The cancer cell map initiative: defining the hallmark networks of cancer. Molecular cell. 2015 May 21;58(4):690-8.

Kumar P, Tan Y, Cahan P. Understanding development and stem cells using single cell-based analyses of gene expression. Development. 2017 Jan 1;144(1):17-32.

Lachmann A, Ma'ayan A. KEA: kinase enrichment analysis. Bioinformatics. 2009 Jan 28;25(5):684-6.

Lachmann A, Torre D, Keenan AB, Jagodnik KM, Lee HJ, Wang L, Silverstein MC, Ma’ayan A. Massive mining of publicly available RNA-seq data from human and mouse. Nature communications. 2018 Apr 10;9(1):1366.

Lachmann A, Xu H, Krishnan J, Berger SI, Mazloom AR, Ma'ayan A. ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. Bioinformatics. 2010 Aug 13;26(19):2438-44.

Lazar C, Meganck S, Taminau J, Steenhoff D, Coletta A, Molter C, Weiss-Solís DY, Duque R, Bersini H, Nowé A. Batch effect removal methods for microarray gene expression data integration: a survey. Briefings in bioinformatics. 2012 Jul 31;14(4):469-90.

Lee S, Seo CH, Lim B, Yang JO, Oh J, Kim M, Lee S, Lee B, Kang C, Lee S. Accurate quantification of transcriptome from RNA-Seq data by effective length normalization. Nucleic acids research. 2010 Nov 8;39(2):e9-.

Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics. 2010 Oct;11(10):733.

Leek JT. Svaseq: removing batch effects and other unwanted noise from sequencing data. Nucleic acids research. 2014 Dec 1;42(21):e161-.

Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database Collaboration. The sequence read archive. Nucleic acids research. 2010 Nov 8;39(suppl_1):D19-21.

Leisch F. Sweave: Dynamic generation of statistical reports using literate data analysis. InCompstat 2002 (pp. 575-580). Physica, Heidelberg.

Lenz G, Wright GW, Emre NT, Kohlhammer H, Dave SS, Davis RE, Carty S, Lam LT, Shaffer AL, Xiao W, Powell J. Molecular subtypes of diffuse large B-cell lymphoma arise by distinct genetic pathways. Proceedings of the National Academy of Sciences. 2008 Sep 9;105(36):13520-5.

Li J, Jiang H, Wong WH. Modeling non-uniformity in short-read rates in RNA-Seq data. Genome biology. 2010 May;11(5):R50.

Li JJ, Bickel PJ, Biggin MD. System wide analyses have underestimated protein abundances and the importance of transcription in mammals. PeerJ. 2014 Feb 27;2:e270.

Li L, Wei Y, To C, Zhu CQ, Tong J, Pham NA, Taylor P, Ignatchenko V, Ignatchenko A, Zhang W, Wang D. Integrated omic analysis of lung cancer reveals metabolism proteome signatures with prognostic impact. Nature communications. 2014 Nov 28;5:5469.

Liu Y, Beyer A, Aebersold R. On the dependency of cellular protein levels on mRNA abundance. Cell. 2016 Apr 21;165(3):535-50.

Llovet JM, Montal R, Sia D, Finn RS. Molecular therapies and precision medicine for hepatocellular carcinoma. Nature Reviews Clinical Oncology. 2018 Oct;15(10):599.

Loskot P, Atitey K, Mihaylova L. Comprehensive review of models and methods for inferences in bio-chemical reaction networks. Frontiers in Genetics. 2019;10:549.

Ma J, Yu MK, Fong S, Ono K, Sage E, Demchak B, Sharan R, Ideker T. Using deep learning to model the hierarchical structure and function of a cell. Nature methods. 2018 Apr;15(4):290.

Macdonald PJ, Almaas E, Barabási AL. Minimum spanning trees of weighted scale-free networks. EPL (Europhysics Letters). 2005 Sep 14;72(2):308.

Mamoshina P, Vieira A, Putin E, Zhavoronkov A. Applications of deep learning in biomedicine. Molecular pharmaceutics. 2016 Mar 29;13(5):1445-54.

Manfredi C, Tindall JM, Hong JS, Sorscher EJ. Making precision medicine personal for cystic fibrosis. Science. 2019 Jul 19;365(6450):220-1.

Marguerat S, Schmidt A, Codlin S, Chen W, Aebersold R, Bähler J. Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. Cell. 2012 Oct 26;151(3):671-83.

Martini M, De Santis MC, Braccini L, Gulluni F, Hirsch E. PI3K/AKT signaling pathway and cancer: an updated review. Annals of medicine. 2014 Sep 1;46(6):372-83.

Mecham BH, Nelson PS, Storey JD. Supervised normalization of microarrays. Bioinformatics. 2010 Mar 31;26(10):1308-15.

Metzker ML. Sequencing technologies—the next generation. Nature reviews genetics. 2010 Jan;11(1):31.

Michalski RS. Understanding the nature of learning: Issues and research directions. Machine learning: An artificial intelligence approach. 1986;2(1):3-25.

Mitchell JS, Papadimitriou CH. The weighted region problem: finding shortest paths through a weighted planar subdivision. Journal of the ACM (JACM). 1991 Jan 3;38(1):18-73.

Morris C, Kriege NM, Bause F, Kersting K, Mutzel P, Neumann M. Tudataset: A collection of benchmark datasets for learning with graphs. arXiv preprint arXiv:2007.08663. 2020 Jul 16.

Morris C, Ritzert M, Fey M, Hamilton WL, Lenssen JE, Rattan G, Grohe M. Weisfeiler and leman go neural: Higher-order graph neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence 2019 Jul 17 (Vol. 33, No. 01, pp. 4602-4609).

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods. 2008 Jul;5(7):621.

Nakatsui M, Horimoto K, Okamoto M, Tokumoto Y, Miyake J. Parameter optimization by using differential elimination: a general approach for introducing constraints into objective functions. InBMC systems biology 2010 Sep (Vol. 4, No. 2, p. S9). BioMed Central.

Nellore A, Collado-Torres L, Jaffe AE, Alquicira-Hernández J, Wilks C, Pritt J, Morton J, Leek JT, Langmead B. Rail-RNA: scalable analysis of RNA-seq splicing and coverage. Bioinformatics. 2017 Dec 15;33(24):4033-40.

Newman AM, Gentles AJ, Liu CL, Diehn M, Alizadeh AA. Data normalization considerations for digital tumor dissection. Genome biology. 2017 Dec;18(1):128.

O'Connor LJ, Price AL. Distinguishing genetic correlation from causation across 52 diseases and complex traits. bioRxiv. 2018 Jan 1:205435.

Orton RJ, Sturm OE, Vyshemirsky V, Calder M, Gilbert DR, Kolch W. Computational modelling of the receptor-tyrosine-kinase-activated MAPK pathway. Biochemical Journal. 2005 Dec 1;392(2):249-61.

Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. Nature reviews genetics. 2011 Feb;12(2):87.

Pages F, Galon J, Dieu-Nosjean MC, Tartour E, Sautes-Fridman C, Fridman WH. Immune infiltration in human tumors: a prognostic factor that should not be ignored. Oncogene. 2010 Feb;29(8):1093.

Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, Lin Z, Desmaison A, Antiga L, Lerer A. Automatic differentiation in pytorch. 2017

Pearson K. LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. 1901 Nov 1;2(11):559-72.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J. Scikit-learn: Machine learning in Python. Journal of machine learning research. 2011;12(Oct):2825-30.

Pegram LM, Wendorff T, Erdmann R, Shkel I, Bellissimo D, Felitsky DJ, Record MT. Why Hofmeister effects of many salts favor protein folding but not DNA helix formation. Proceedings of the National Academy of Sciences. 2010 Apr 27;107(17):7716-21.

Phillips HS, Kharbanda S, Chen R, Forrest WF, Soriano RH, Wu TD, Misra A, Nigro JM, Colman H, Soroceanu L, Williams PM. Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. Cancer cell. 2006 Mar 31;9(3):157-73.

Pierson E, Koller D, Battle A, Mostafavi S, GTEx Consortium. Sharing and specificity of co-expression networks across 35 human tissues. PLoS computational biology. 2015 May 13;11(5):e1004220.

Pointer I. Programming PyTorch for Deep Learning: Creating and Deploying Deep Learning Applications. " O'Reilly Media, Inc."; 2019 Sep 20.

Ray J, Pinar A, Seshadhri C. Are I there yet? When to stop a Markov chain while generating random graphs. InInternational Workshop on Algorithms and Models for the Web-Graph 2012 Jun 22 (pp. 153-164). Springer, Berlin, Heidelberg.

Reinhard FB, Eberhard D, Werner T, Franken H, Childs D, Doce C, Savitski MF, Huber W, Bantscheff M, Savitski MM, Drewes G. Thermal proteome profiling monitors ligand interactions with cellular membrane proteins. Nature methods. 2015 Dec;12(12):1129.

Richardson RB, Allan DS, Le Y. Greater organ involution in highly proliferative tissues associated with the early onset and acceleration of ageing in humans. Experimental gerontology. 2014 Jul 1;55:80-91.

Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L. Improving RNA-Seq expression estimates by correcting for fragment bias. Genome biology. 2011 Sep;12(3):R22.

Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome biology. 2010 Mar;11(3):R25.

Rubinfeld H, Seger R. The ERK cascade. Molecular biotechnology. 2005 Oct 1;31(2):151-74.

Sabour S, Frosst N, Hinton GE. Dynamic routing between capsules. InAdvances in Neural Information Processing Systems 2017 (pp. 3856-3866).

Saha A, Kim Y, Gewirtz AD, Jo B, Gao C, McDowell IC, Engelhardt BE, Battle A, Aguet F, Ardlie KG, Cummings BB. Co-expression networks reveal the tissue-specific regulation of transcription and splicing. Genome research. 2017 Nov 1;27(11):1843-58.

Sanchez-Vega F, Mina M, Armenia J, Chatila WK, Luna A, La KC, Dimitriadoy S, Liu DL, Kantheti HS, Saghafinia S, Chakravarty D. Oncogenic Signaling Pathways in The Cancer Genome Atlas. Cell. 2018 Apr 5;173(2):321-37.

Sanner MF. Python: a programming language for software integration and development. J Mol Graph Model. 1999 Feb 1;17(1):57-61.

Saraçli S, Doğan N, Doğan İ. Comparison of hierarchical cluster analysis methods by cophenetic correlation. Journal of Inequalities and Applications. 2013 Dec 1;2013(1):203.

Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, Monfardini G. The graph neural network model. IEEE Transactions on Neural Networks. 2008 Dec 9;20(1):61-80.

Schwaederle M, Zhao M, Lee JJ, Eggermont AM, Schilsky RL, Mendelsohn J, Lazar V, Kurzrock R. Impact of precision medicine in diverse cancers: a meta-analysis of phase II clinical trials. Journal of clinical oncology. 2015 Nov 10;33(32):3817.

Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M. Global quantification of mammalian gene expression control. Nature. 2011 May;473(7347):337.

Seger R, Krebs EG. The MAPK signaling cascade. The FASEB journal. 1995 Jun;9(9):726-35.

Shargorodskiy A, Gururaj K, Naik M, Narvaez P, Srinivasa G. HPC Infrastructure for Genomic Workloads: A Look at Oregon Health & Science University's Exacloud. Intel White Paper. Intel Health and Life Sciences. 2015.

Sharma K, Schmitt S, Bergner CG, Tyanova S, Kannaiyan N, Manrique-Hoyos N, Kongi K, Cantuti L, Hanisch UK, Philips MA, Rossner MJ. Cell type–and brain region–resolved mouse brain proteome. Nature neuroscience. 2015 Dec;18(12):1819.

Shapiro E, Biezuner T, Linnarsson S. Single-cell sequencing-based technologies will revolutionize whole-organism science. Nature Reviews Genetics. 2013 Sep;14(9):618.

Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological science. 2011 Nov;22(11):1359-66.

Sonawane AR, Platig J, Fagny M, Chen CY, Paulson JN, Lopes-Ramos CM, DeMeo DL, Quackenbush J, Glass K, Kuijjer ML. Understanding tissue-specific gene regulation. Cell reports. 2017 Oct 24;21(4):1077-88.

Sun J, Ajwani D, Nicholson PK, Sala A, Parthasarathy S. Breaking cycles in noisy hierarchies. InProceedings of the 2017 ACM on Web Science Conference 2017 Jun 25 (pp. 151-160). ACM.

Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. InInternational Conference on Machine Learning 2017 Jul 17 (pp. 3319-3328). PMLR.

Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M. STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic acids research. 2014 Oct 28;43(D1):D447-52.

Söllner JF, Leparc G, Hildebrandt T, Klein H, Thomas L, Stupka E, Simon E. An RNA-Seq atlas of gene expression in mouse and rat normal tissues. Scientific data. 2017 Dec 12;4:170185.

Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade. Nature protocols. 2018 Apr;13(4):599.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols. 2012 Mar;7(3):562.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology. 2010 May;28(5):511.

Tridgell A, Mackerras P. The rsync algorithm. The Australian National University. 1996 Jun.

Tukey, J. W. 1977. *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Tung PY, Blischak JD, Hsiao CJ, Knowles DA, Burnett JE, Pritchard JK, Gilad Y. Batch effects and the effective design of single-cell gene expression studies. Scientific reports. 2017 Jan 3;7:39921.

Tyanova S, Albrechtsen R, Kronqvist P, Cox J, Mann M, Geiger T. Proteomic maps of breast cancer subtypes. Nature communications. 2016 Jan 4;7:10259.

Tyson JJ, Laomettachit T, Kraikivski P. Modeling the dynamic behavior of biochemical regulatory networks. Journal of theoretical biology. 2018 Nov 28.

Uhlen M, Zhang C, Lee S, Sjöstedt E, Fagerberg L, Bidkhori G, Benfeitas R, Arif M, Liu Z, Edfors F, Sanli K. A pathology atlas of the human cancer transcriptome. Science. 2017 Aug 18;357(6352):eaan2507.

Usaj M, Tan Y, Wang W, VanderSluis B, Zou A, Myers CL, Costanzo M, Andrews B, Boone C. TheCellMap. org: A web-accessible database for visualizing and mining the global yeast genetic interaction network. G3: Genes, Genomes, Genetics. 2017 May 1;7(5):1539-49.

Van Hoof D, Muñoz J, Braam SR, Pinkse MW, Linding R, Heck AJ, Mummery CL, Krijgsveld J. Phosphorylation dynamics during early differentiation of human embryonic stem cells. Cell stem cell. 2009 Aug 7;5(2):214-26.

Van Leeuwen J, Pons C, Mellor JC, Yamaguchi TN, Friesen H, Koschwanez J, Ušaj MM, Pechlaner M, Takar M, Ušaj M, VanderSluis B. Exploring genetic suppression interactions on a global scale. Science. 2016 Nov 4;354(6312):aag0839.

Vargas AJ, Harris CC. Biomarker development in the precision medicine era: lung cancer as a case study. Nature Reviews Cancer. 2016 Aug;16(8):525.

Varley KE, Gertz J, Roberts BS, Davis NS, Bowling KM, Kirby MK, Nesmith AS, Oliver PG, Grizzle WE, Forero A, Buchsbaum DJ. Recurrent read-through fusion transcripts in breast cancer. Breast cancer research and treatment. 2014 Jul 1;146(2):287-97.

Wagstaff K, Cardie C. Clustering with instance-level constraints. AAAI/IAAI. 2000 Jun 29;1097:577-84.

Wang M, Zhao Y, Zhang B. Efficient test and visualization of multi-set intersections. Scientific reports. 2015 Nov 25;5:16923.

Wang Q, Armenia J, Zhang C, Penson AV, Reznik E, Zhang L, Minet T, Ochoa A, Gross BE, Iacobuzio-Donahue CA, Betel D. Unifying cancer and normal RNA sequencing data from different sources. Scientific data. 2018 Apr 17;5:180061.

Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews genetics. 2009 Jan;10(1):57.

Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, Zheng H. Determination and inference of eukaryotic transcription factor sequence specificity. Cell. 2014 Sep 11;158(6):1431-43.

Wright G, Tan B, Rosenwald A, Hurt EH, Wiestner A, Staudt LM. A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. Proceedings of the National Academy of Sciences. 2003 Aug 19;100(17):9991-6.

Wu AR, Neff NF, Kalisky T, Dalerba P, Treutlein B, Rothenberg ME, Mburu FM, Mantalas GL, Sim S, Clarke MF, Quake SR. Quantitative assessment of single-cell RNA-sequencing methods. Nature methods. 2014 Jan;11(1):41.

Wu G, Dawson E, Duong A, Haw R, Stein L. ReactomeFIViz: a Cytoscape app for pathway and network-based data analysis. F1000Research. 2014;3.

Wu G, Feng X, Stein L. A human functional protein interaction network and its application to cancer data analysis. Genome biology. 2010 May;11(5):R53.

Ye QH, Qin LX, Forgues M, He P, Kim JW, Peng AC, Simon R, Li Y, Robles AI, Chen Y, Ma ZC. Predicting hepatitis B virus–positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning. Nature medicine. 2003 Apr;9(4):416.

Yu CH. Exploratory Data Analysis. Psychology. Oxford Bibliographies. 2017 Nov 29. DOI: 10.1093/OBO/9780199828340-0200

Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z, Chambers MC, Zimmerman LJ, Shaddox KF, Kim S, Davies SR. Proteogenomic characterization of human colon and rectal cancer. Nature. 2014 Sep;513(7518):382.

Zhang J. Protein-protein interactions in salt solutions. InProtein-protein interactions-computational and experimental tools 2012 Mar 30. IntechOpen.

Zhang K, Shasha D. Simple fast algorithms for the editing distance between trees and related problems. SIAM journal on computing. 1989 Dec;18(6):1245-62.

Zhan T, Rindtorff N, Boutros M. Wnt signaling in cancer. Oncogene. 2017 Mar;36(11):1461.

Zitnik M, Nguyen F, Wang B, Leskovec J, Goldenberg A, Hoffman MM. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. Information Fusion. 2019 Oct 1;50:71-91.

Appendices

Appendix A: Pathways Significantly Enriched Only by Reactions or Transcripts

Pathways enriched for mean cell proliferation rate tissue group by reactions only are shown below highlighted in purple. Those enriched by transcripts only are highlighted in green. Pathways are displayed within their hierarchical contexts.

- Chromatin organization (Homo sapiens) (R-HSA-4839726)
  - Chromatin modifying enzymes (Homo sapiens) (R-HSA-3247509)
    - HDMs demethylate histones (Homo sapiens) (R-HSA-3214842)
    - PKMTs methylate histone lysines (Homo sapiens) (R-HSA-3214841)
    - HATs acetylate histones (Homo sapiens) (R-HSA-3214847)
    - RMTs methylate histone arginines (Homo sapiens) (R-HSA-3214858)
- Gene expression (Transcription) (Homo sapiens) (R-HSA-74160)
  - RNA Polymerase II Transcription (Homo sapiens) (R-HSA-73857)
    - Generic Transcription Pathway (Homo sapiens) (R-HSA-212436)
      - FOXO-mediated transcription (Homo sapiens) (R-HSA-9614085)
      - Transcriptional regulation by RUNX1 (Homo sapiens) (R-HSA-8878171)
        - RUNX1 and FOXP3 control the development of regulatory T lymphocytes (Tregs) (Homo sapiens) (R-HSA-8877330)
        - RUNX1 regulates expression of components of tight junctions (Homo sapiens) (R-HSA-8935964)
      - Transcriptional regulation by RUNX2 (Homo sapiens) (R-HSA-8878166)
      - Transcriptional regulation by RUNX3 (Homo sapiens) (R-HSA-8878159)
- Metabolism of proteins (Homo sapiens)
  - Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs) (Homo sapiens) (R-HSA-381426)
  - Post-translational protein modification (Homo sapiens)
    - O-linked glycosylation (Homo sapiens) (R-HSA-5173105)
      - O-linked glycosylation of mucins (Homo sapiens) (R-HSA-913709)
- Immune System (Homo sapiens)
  - Adaptive Immune System (Homo sapiens)
    - Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell (Homo sapiens) (R-HSA-198933)
    - Signaling by the B Cell Receptor (BCR) (Homo sapiens)
      - CD22 mediated BCR regulation (Homo sapiens) (R-HSA-5690714)
  - Innate Immune System (Homo sapiens)
    - Fcgamma receptor (FCGR) dependent phagocytosis (Homo sapiens)
      - Role of phospholipids in phagocytosis (Homo sapiens) (R-HSA-2029485)
    - Fc epsilon receptor (FCERI) signaling (Homo sapiens) (R-HSA-2454202)
      - FCERI mediated NF-kB activation (Homo sapiens) (R-HSA-2871837)
      - FCERI mediated Ca+2 mobilization (Homo sapiens) (R-HSA-2871809)
      - Role of LAT2/NTAL/LAB on calcium mobilization (Homo sapiens) (R-HSA-2730905)
      - FCERI mediated MAPK activation (Homo sapiens) (R-HSA-2871796)
    - Complement cascade (Homo sapiens) (R-HSA-166658)
      - Regulation of Complement cascade (Homo sapiens) (R-HSA-977606)
      - Initial triggering of complement (Homo sapiens) (R-HSA-166663)
        - Creation of C4 and C2 activators (Homo sapiens) (R-HSA-166786)
          - Classical antibody-mediated complement activation (Homo sapiens) (R-HSA-173623)
    - Toll-like Receptor Cascades (Homo sapiens)
      - Toll Like Receptor 7/8 (TLR7/8) Cascade (Homo sapiens) (R-HSA-168181)
      - Toll Like Receptor 9 (TLR9) Cascade (Homo sapiens) (R-HSA-168138)
  - Cytokine Signaling in Immune system (Homo sapiens) (R-HSA-1280215)
    - Growth hormone receptor signaling (Homo sapiens)(R-HSA-982772)

- Signaling by Interleukins (Homo sapiens) (R-HSA-449147)
  - Interleukin-2 family signaling (Homo sapiens) (R-HSA-451927)
  - Interleukin-4 and Interleukin-13 signaling (Homo sapiens) (R-HSA-6785807)
  - Interleukin-12 family signaling (Homo sapiens) (R-HSA-447115)
    - Interleukin-35 Signalling (Homo sapiens) (R-HSA-8984722)

- DNA Repair (Homo sapiens) (R-HSA-73894)
  - Nucleotide Excision Repair (Homo sapiens)
    - Global Genome Nucleotide Excision Repair (GG-NER) (Homo sapiens) (R-HSA-5696399)
      - Formation of Incision Complex in GG-NER (Homo sapiens) (R-HSA-5696395)
  - Base Excision Repair (Homo sapiens) (R-HSA-73884)
    - Resolution of Abasic Sites (AP sites) (Homo sapiens) (R-HSA-73933)
      - Resolution of AP sites via the multiple-nucleotide patch replacement pathway (Homo sapiens)
        - PCNA-Dependent Long Patch Base Excision Repair (Homo sapiens) (R-HSA-5651801)
      - APEX1-Independent Resolution of AP Sites via the Single Nucleotide Replacement Pathway (Homo sapiens) (R-HSA-5649702)
  - Fanconi Anemia Pathway (Homo sapiens) (R-HSA-6783310)
  - DNA Damage Bypass (Homo sapiens) (R-HSA-73893)
    - Translesion synthesis by Y family DNA polymerases bypasses lesions on DNA template (Homo sapiens) (R-HSA-110313)
      - Termination of translesion DNA synthesis (Homo sapiens) (R-HSA-5656169)
  - DNA Double-Strand Break Repair (Homo sapiens)
    - Homology Directed Repair (Homo sapiens)
      - HDR through Homologous Recombination (HRR) or Single Strand Annealing (SSA) (Homo sapiens)
        - HDR through Homologous Recombination (HRR) (Homo sapiens) (R-HSA-5685942)
- Signal Transduction (Homo sapiens)
  - Intracellular signaling by second messengers (Homo sapiens)
    - PIP3 activates AKT signaling (Homo sapiens)
      - PTEN Regulation (Homo sapiens)
        - Regulation of PTEN stability and activity (Homo sapiens) (R-HSA-8948751)
  - Signaling by Receptor Tyrosine Kinases (Homo sapiens)
    - Signaling by FGFR (Homo sapiens)
      - Signaling by FGFR1 (Homo sapiens)
        - Downstream signaling of activated FGFR1 (Homo sapiens) (R-HSA-5654687)
          - PI-3K cascade:FGFR1 (Homo sapiens) (R-HSA-5654689)
  - Signaling by Non-Receptor Tyrosine Kinases (Homo sapiens) (R-HSA-9006927)
    - Signaling by PTK6 (Homo sapiens) (R-HSA-8848021)
      - PTK6 Regulates Proteins Involved in RNA Processing (Homo sapiens) (R-HSA-8849468)
      - PTK6 Regulates RHO GTPases, RAS GTPase and MAP kinases (Homo sapiens) (R-HSA-8849471)
  - Signaling by NOTCH (Homo sapiens) (R-HSA-157118)
    - Signaling by NOTCH3 (Homo sapiens) (R-HSA-9012852)
      - NOTCH3 Intracellular Domain Regulates Transcription (Homo sapiens) (R-HSA-9013508)
    - Signaling by NOTCH4 (Homo sapiens) (R-HSA-9013694)
      - NOTCH4 Intracellular Domain Regulates Transcription (Homo sapiens) (R-HSA-9013695)
  - Signaling by Leptin (Homo sapiens) (R-HSA-2586552)
  - Signaling by Nuclear Receptors (Homo sapiens)
    - ESR-mediated signaling (Homo sapiens) (R-HSA-8939211)
      - Estrogen-dependent gene expression (Homo sapiens) (R-HSA-9018519)
  - Signaling by GPCR (Homo sapiens) (R-HSA-372790)
    - GPCR downstream signalling (Homo sapiens) (R-HSA-388396)
      - G alpha (s) signalling events (Homo sapiens) (R-HSA-418555)
      - G alpha (z) signalling events (Homo sapiens) (R-HSA-418597)
      - G-protein beta:gamma signalling (Homo sapiens) (R-HSA-397795)

    - Signaling by Hedgehog (Homo sapiens)
        - Hedgehog 'on' state (Homo sapiens) (R-HSA-5632684)
- Cell Cycle (Homo sapiens) (R-HSA-1640170)
    - Chromosome Maintenance (Homo sapiens) (R-HSA-73886)
        - Telomere Maintenance (Homo sapiens)
            - Extension of Telomeres (Homo sapiens) (R-HSA-180786)
                - Telomere C-strand (Lagging Strand) Synthesis (Homo sapiens) (R-HSA-174417)
                    - Processive synthesis on the C-strand of the telomere (Homo sapiens) (R-HSA-174414)
    - Cell Cycle Checkpoints (Homo sapiens)
        - Mitotic Spindle Checkpoint (Homo sapiens) (R-HSA-69618)
        - G2/M Checkpoints (Homo sapiens) (R-HSA-69481)
            - G2/M DNA damage checkpoint (Homo sapiens) (R-HSA-69473)
    - Cell Cycle, Mitotic (Homo sapiens)
        - Mitotic G2-G2/M phases (Homo sapiens)
            - G2/M Transition (Homo sapiens)
                - Polo-like kinase mediated events (Homo sapiens) (R-HSA-156711)
        - S Phase (Homo sapiens)
            - Synthesis of DNA (Homo sapiens) (R-HSA-69239)
                - DNA strand elongation (Homo sapiens) (R-HSA-69190)
                    - Lagging Strand Synthesis (Homo sapiens) (R-HSA-69186)
- Cellular responses to external stimuli (Homo sapiens)
    - Cellular responses to stress (Homo sapiens)
        - Cellular response to starvation (Homo sapiens)
            - Amino acids regulate mTORC1 (Homo sapiens) (R-HSA-9639288)
- DNA Replication (Homo sapiens) (R-HSA-69306)
    - Synthesis of DNA (Homo sapiens) (R-HSA-69239)
        - DNA strand elongation (Homo sapiens) (R-HSA-69190)
            - Lagging Strand Synthesis (Homo sapiens) (R-HSA-69186)
- Hemostasis (Homo sapiens) (R-HSA-109582)
    - Platelet activation, signaling and aggregation (Homo sapiens)
        - Signal amplification (Homo sapiens) (R-HSA-392518)
    - Cell surface interactions at the vascular wall (Homo sapiens) (R-HSA-202733)
        - Tie2 Signaling (Homo sapiens) (R-HSA-210993)
- Programmed Cell Death (Homo sapiens) (R-HSA-5357801)
    - Apoptosis (Homo sapiens)
        - Apoptotic execution phase (Homo sapiens)
            - Apoptotic cleavage of cellular proteins (Homo sapiens) (R-HSA-111465)
                - Apoptotic cleavage of cell adhesion proteins (Homo sapiens) (R-HSA-351906)
- Metabolism (Homo sapiens)
    - Metabolism of vitamins and cofactors (Homo sapiens)
        - Metabolism of water-soluble vitamins and cofactors (Homo sapiens)
            - Cobalamin (Cbl, vitamin B12) transport and metabolism (Homo sapiens) (R-HSA-196741)
    - Integration of energy metabolism (Homo sapiens)
        - Regulation of insulin secretion (Homo sapiens)
            - Adrenaline,noradrenaline inhibits insulin secretion (Homo sapiens) (R-HSA-400042)
- Cell-Cell communication (Homo sapiens)
    - Cell junction organization (Homo sapiens)
        - Type I hemidesmosome assembly (Homo sapiens) (R-HSA-446107)
- Vesicle-mediated transport (Homo sapiens) (R-HSA-5653656)
    - Binding and Uptake of Ligands by Scavenger Receptors (Homo sapiens) (R-HSA-2173782)
        - Scavenging of heme from plasma (Homo sapiens) (R-HSA-2168880)
- Sensory Perception (Homo sapiens)
    - Olfactory Signaling Pathway (Homo sapiens) (R-HSA-381753)
- Muscle contraction (Homo sapiens)
    - Striated Muscle Contraction (Homo sapiens) (R-HSA-390522)

Appendix B: Data Availability

Recount2 GTEx RNA-seq data this project used is available online at https://jhubiostatistics.shinyapps.io/recount/.

Reactome data is available online at https://reactome.org/download-data and https://reactome.org/dev/content-service. This project used Reactome version 71, released December, 2019.

Appendix C: Source Code Availability

Source code used to perform analysis and generate figures is available by request for academic use and version-controlled in our online repository at https://github.com/joshuaburkhart/reticula.