Three Aspects of Gene Expression: Pathway Coexpression, Cross-Species Analysis of RNA-seq Data, and Bias in Gene Coexpression

Three Aspects of Gene Expression: Pathway Coexpression, Cross-Species Analysis of RNA-seq Data, and Bias in Gene Coexpression

A DISSERTATION PRESENTED
BY
YERED HAMMURABI PITA JUAREZ
TO
THE DEPARTMENT OF BIOSTATISTICS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
BIOSTATISTICS

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS

JULY 2017

Thesis advisor: Professor Rafael Irizarry Yered Hammurabi Pita Juarez

Three Aspects of Gene Expression: Pathway Coexpression, Cross-Species Analysis of RNA-seq Data, and Bias in Gene Coexpression

Abstract

In this dissertation, we propose methods for gene expression focused on three problems in functional genomics: describing relationships between biological pathways, comparing tissues from different species, and accounting for biases in gene coexpression.

In chapter 1, we present a pathway coexpression network that systematically quantifies and establishes a reference for high-level relationships between pathways. The method uses 3,207 microarrays from 72 normal human tissues and 1,330 of the most well established pathway annotations to describe global relationships between pathways. The pathway coexpression network accounts for shared genes to estimate correlations between pathway with related functions rather than with redundant annotations.

In chapter 2, we propose a method to adjust RNA-seq expression estimates from human and mouse tissues for differences between the genomic annotations. Previous studies using gene expression data to compare homologous genes across different species concluded that gene expression was more similar between homologous tissues of different species than between different tissues from the same species. Recently, the Mouse ENCODE consortium reached the opposite conclusion reporting that gene expression data from humans and mice samples cluster by species rather than by tissue. We showed that these results were driven by differences between species annotation. Our method uses ortholog probes, genomic regions within human-mouse orthologs with the

same length and almost identical sequences, to quantify gene expression data. The ortholog probes showed that the human and mouse samples cluster by tissue rather than by species.

In chapter 3, we used a linear model framework to estimates the correlation between genes taking into account the experimental factors from gene expression data sets. The correlation based on gene expression data has been a popular choice to describe relationships between genes. However, interpreting these correlation estimates is challenging since they can arise from biological as well as non-biological sources. We used a linear mixed model to quantify the influence of the variation within experimental factors on the observed correlation, and a linear model to estimate the correlation between the gene-specific effects of the experimental factors.

Contents

A mis padres, Ángel Pita Duque y Xóchitl Juárez Varela, y a mi hermano José Pita Juárez

Acknowledgments

I am much indebted to my advisors, Rafael Irizarry and Winston Hide, for not only their mentoring, but also their patience, support and understanding when I needed it most. I feel grateful and honored for the opportunity to work with them.

I sincerely thank my dissertation committee, JP Onnela and Curtis Huttenhower, for their many insightful contributions and guidance. I owe a debt of gratitude to Phoebe Hackett and Jelena Follweiler. Thank you for helping me during times of hardship as well as navigating through the department.

I am grateful for the generous financial support from CONACyT, Fundación México en Harvard, and the John F. and Virgina Taplin endowment.

I was very fortunate to have met wonderful friends who made me feel at home during graduate school. I thank Matey Neykov and Caleb Miles for their encouragement during the first years of the program and most importantly for staying true friends. I also want to thank Sixing Chen - for encouraging me to enjoy the snow, Benjamín Sánchez - for all the memorable bike rides, Carla Marquéz Luna - for all those wonderful coffee breaks. My deepest thanks go to my roommate Aleksandrina Goeva - for saving me, literally and figuratively, countless times.

Lastly, my heart goes out to my family. I would like to thank my parents, Ángel and Xóchitl, for their endless love and support. Thanks to them I was motivated to keep going no matter how hard

things got. I would like to thank my brother, José, for bringing art and music into my life, and for inspiring me to be a better person.

1

The Pathway Coexpression Network: Revealing Pathway Relationships

1.1 Introduction

A goal of functional genomics is to understand the relationships between biological processes. Pathways contribute to functional interplay within biological processes through complex but poorly

understood interactions. However, limited functional references for global pathway relationships exist. Pathways from databases such as KEGG and Reactome provide discrete annotations of biological processes. Their relationships are currently either inferred from gene set enrichment within specific experiments, or by simple overlap, linking pathway annotations that have genes in common. Here, we provide a unifying interpretation of functional interaction between pathways by systematically quantifying coexpression between 1,330 canonical pathways from the Molecular Signatures Database (MSigDB) to establish the Pathway Coexpression Network (PCxN). We estimated the correlation between canonical pathways valid in a broad context using a curated collection of 3,207 microarrays from 72 normal human tissues. PCxN accounts for shared genes between annotations to estimate significant correlations between pathways with related functions rather than with similar annotations. We demonstrate that PCxN provides novel insight into mechanisms of complex diseases using an Alzheimer's Disease (AD) case study. PCxN retrieved pathways significantly correlated with an expert curated AD gene list. These pathways have known associations with AD and were significantly enriched for genes independently associated with AD. As a further step, we show how PCxN complements the results of gene set enrichment methods by revealing relationships between enriched pathways, and by identifying additional highly correlated pathways. PCxN revealed that correlated pathways from an AD expression profiling study include functional clusters involved in cell adhesion and oxidative stress. PCxN provides expanded connections to pathways from the extracellular matrix. PCxN provides a powerful new framework for interrogation of global pathway relationships. Comprehensive exploration of PCxN can be performed at http://pcxn.org/.

The advancement of high throughput, high dimensional 'omic' technology has enabled quantifi-

cation of a vast array of cellular components. Inducing phenotypic changes, through mutations or perturbations, and observing their impact on genomic, proteomic and metabolomic assays has allowed us to assign roles to sets of genes and gene products[61,199,201]. We now appreciate that cell states are controlled by cascades of interactions coordinated into protein complexes and pathways[11,234,218]. Thus pathways have become the functional building blocks on which we base interpretation of cell state. However, systems approaches to interpret the relationships between omic components have focused upon development of gene based interrogation through gene-gene networks. Pathways drive biological processes through complex and poorly understood interactions, and only limited functional references for global pathway relationships exist. Mapping out pathway relationships is a fundamental challenge as we strive to influence cell development and disease[10,207].

1.1.1 Pathway Analysis

The development of databases such as KEGG[104], Reactome[104,44] and Biocarta[185] have provided curated lists of pathway membership. These gene lists enable systematic mapping of genomic scale data to biological processes. Gene expression profiling provides the most common basis for describing experimental changes in pathway terms. Usually, differentially expressed genes between a pair of conditions are used to highlight enriched pathways. Well established gene set enrichment methods such as GSEA[238], SAFE[14], PAGE[113] and GSA[113,58] produce lists of pathways that are significantly enriched in an individual experiment[181,98]. Gene set enrichment methods test relationships between phenotypes and pathways. These methods test if a pathway is overrepresented in genes differentially expressed between two phenotypes[181,98]. The results from pathway enrichment analysis do not pro-

vide insight into the relationships between pathways because they only determine association of each individual pathway with a particular phenotype change. Furthermore, the results are unique to the combination of samples compared[226] .

A characteristic of these approaches is that pathways are analyzed independently, the co-enrichment of other pathways considered only insofar as necessitating multiple hypothesis testing. Significant gene membership overlap exists between pathways; and similar but not identical names exist for equivalent, but differently constituted, pathways in separate databases. Describing the relationships between pathways with redundant annotations from different sources might capture high-content similarity rather than truly related biological mechanisms[210,252]. In hierarchical database structures such as GO[9], gene sets corresponding to one process may be fully contained within subset of a parent process. The development of multi-set approaches such as GenGO[150], Markov chain ontology analysis (MCOA)[71], model-based gene set analysis (MGSA)[16], and Selection via LASSO Penalized Regression (SLPR)[70] allows joint testing of pathways for enrichment. Multi-set methods alleviate problems relating to overlap and redundancy, and multifunctional, or pleiotropic, genes that play roles in different biological processes[206]. However, pathways are still treated as independent units without accounting for, or determining, expression correlation arising from biological interaction. Co-enrichment of pathways can either be a reflection of closely related functions or a consequence of overlapping annotation. Pathways also operate in networks, and so pathway-pathway relationships affect their constituent gene expression signatures.

1.1.2 Pathway Networks

A natural extension to gene-centric analysis is to consider the interactions between biological pathways, taking into account relationships between higher level systemic functions of the cell and the organism[79,105]. The key to existing approaches for mapping pathway relationships has been recognition that genes and their products interact with each other, resulting in combinations of gene network relationships, annotation, functional or semantic classification overlaps[117,57], protein interactions, and gene and network enrichment[1,169,255,187,2,51].

1.1.3 Networks Based on Annotation

Several methods for connecting pathways rely solely on annotation, using gene overlap to describe the relationships between gene sets. Methods such as Onto-Express[53] and BiNGO[156] use Gene Ontology (GO)[9] as their only source of curated gene sets and identify parent-child relationships of GO gene sets of interest via gene overlap. Since these methods were developed specifically for GO annotations, their applicability is limited to functional annotation within this hierarchical structure. More recent annotation-based methods such as the Molecular Concepts Maps (MCM)[212], the Enrichment Map[171,101] and the Constellation Map[241] are not restricted to GO. These methods build networks in which the nodes are gene sets and the edge weights are based on shared genes or an intra-experiment similarity score.

1.1.4 Networks Based on Curated Interactions

Pathway interaction networks can also be defined using distance measures based on aggregating curated gene level connections, such as protein-protein interactions (PPIs)[1,96,138,169] or empirically, based on gene coexpression data[255]. Methods based on PPI such as the pathway crosstalk network (PCN)[138] and the characteristic sub pathway network (CSPN)[96] determine relationships between pathways based on the assumption that two pathways are likely to interact if they share a significant number of PPIs. PCN identifies pathway relationships based on the number of shared interactions from a background PPI network to build a global network of pathway interactions[138]. CSPN identifies pathway interactions for a specific phenotype by counting the number of active PPIs defined from differentially expressed genes and a curated PPI background network[96]. Methods based on PPIs have important limitations; when two pathways share only a few PPIs between them but are still significantly related by other interactions, their functional relationship may be missed by the PPI approach. Moreover, these methods rely heavily on the background network structure, whose comprehensiveness, accuracy and importantly, context, bias the results. Issues with PPIs can be alleviated by integrating additional sources of curated relationships. Network Enrichment Analysis (NEA)[1] and CrossTalkZ[169] use a background gene network that complements PPIs with GO annotations and a network of functional coupling[2] to relate pathways based on the extent of their connectivity.

1.1.5 Networks Based on Gene Expression

Systems approaches to interpret the relationships between differentially expressed genes have focused upon development of gene coexpression networks, where these genes are related to each other by known coexpression in extensive large scale assays [275,256]. These methods have been adapted to quantify pathway correlations. For instance, the gene-set coexpression level (GSCoL) method establishes pathway interactions based on sparse canonical correlation analysis of fold change levels derived from gene expression data [2,255]. The Constellation Map provides an enhanced visualization of GSEA results, by defining a distance between pathway pairs. This distance is based on the per-sample similarity of their enrichments across the experimental data. The similarity is based on normalized mutual information rather than the correlation coefficient to capture nonlinear associations. A limiting issue in these methods is that results are unique to the combination of samples compared, restricting conclusions to a specific context, usually a single experiment. Also, experimental and platform biases can drown out changes in biological signal [226,132] and complicate cross experiment comparison. Thus far, only limited pathway networks have been constructed and existing approaches are not designed for creating a global reference network that can be used for discovery and mining of pathway relationships. Public omics data archives such as the Gene Expression Omnibus (GEO) [40] and ArrayExpress [25] contain genome-wide gene expression data from a growing number of experiments [219]. These large collections of microarray data allow meta analyses on gene expression that extend the use of thousands of data sets beyond their initial experimental design [228,152,223]. Harnessing the scope of these repositories is increasingly being realised as a powerful tool for identifying

universal genomic features[165,13,5].

1.1.6 The Pathway Coexpression Network

In this work, we address the need for a consistent functional map of pathway interactions. A reference network of global relationships between pathways serves two purposes: it allows deeper exploration of basic cell biology, and serves as a tool to discover novel mechanisms and targets in disease while building testable models of pathway interaction. Our aim has been to create a network that delineates the global relationships between canonical pathways in as broad a context as possible.
To achieve this goal, we have developed the Pathway Coexpression Network (PCxN). For each experiment from a curated collection of normal human tissue microarrays[165] from publicly available experiments in GEO, we estimated the correlation between pathway summaries based on the mean expression ranks of their gene members along with the corresponding p-value. In the presence of shared genes between the pathway annotations, we adjusted the correlation using the mean expression ranks of the shared genes. Finally, we combined the experiment-level correlation estimates and their corresponding p-values to determine which correlations were significant across all experiments. PCxN significantly expands the scope of pathway methods by estimating global relationships between a wide range of curated pathway annotations, based on coexpression across an expansive gene expression collection. The growing number of available pathway annotations from different sources extends their coverage of biological processes. However, as pathway collections get larger and more complex, the redundancy between the contents of the pathway annotations increases. Pathway coexpression based relationships are often dominated by shared genes. Thus, we have taken into account

the shared genes between pathways so the pathway relationships reflect actual related functions rather than similarities in annotations.

Here we report how PCxN effectively captures intra-pathway relationships within known pathways such as the ribosome pathway. Then, we show how PCxN finds pathways associated with a complex disease: Alzheimer's disease (AD). PCxN determines well known pathways related to AD, including those that influence amyloid pathology and innate immune response. Finally, we show how use of PCxN can complement and expand the results of gene set enrichment analysis within an AD gene expression profiling study. PCxN helps to interpret the results by describing the relationships between the enriched pathways, and provides the opportunity to discover novel relationships by revealing pathways which are highly correlated with the enrichment results. PCxN addresses the need to describe relationships between pathways present across diverse tissues and conditions. These relationships provide a pathway interaction model for a biologically driven phenotype, provide a reference to prioritize targets of biological processes, and provide a powerful enhancement for interpretation of results from gene set enrichment methods. We have built a comprehensive web tool for PCxN to explore novel relationships and to aid with the interpretation of results from gene set enrichment methods (`http://pcxn.org/`). In addition, PCxN is available as a Bioconductor package (`http://bioconductor.org/packages/pcxn/`).

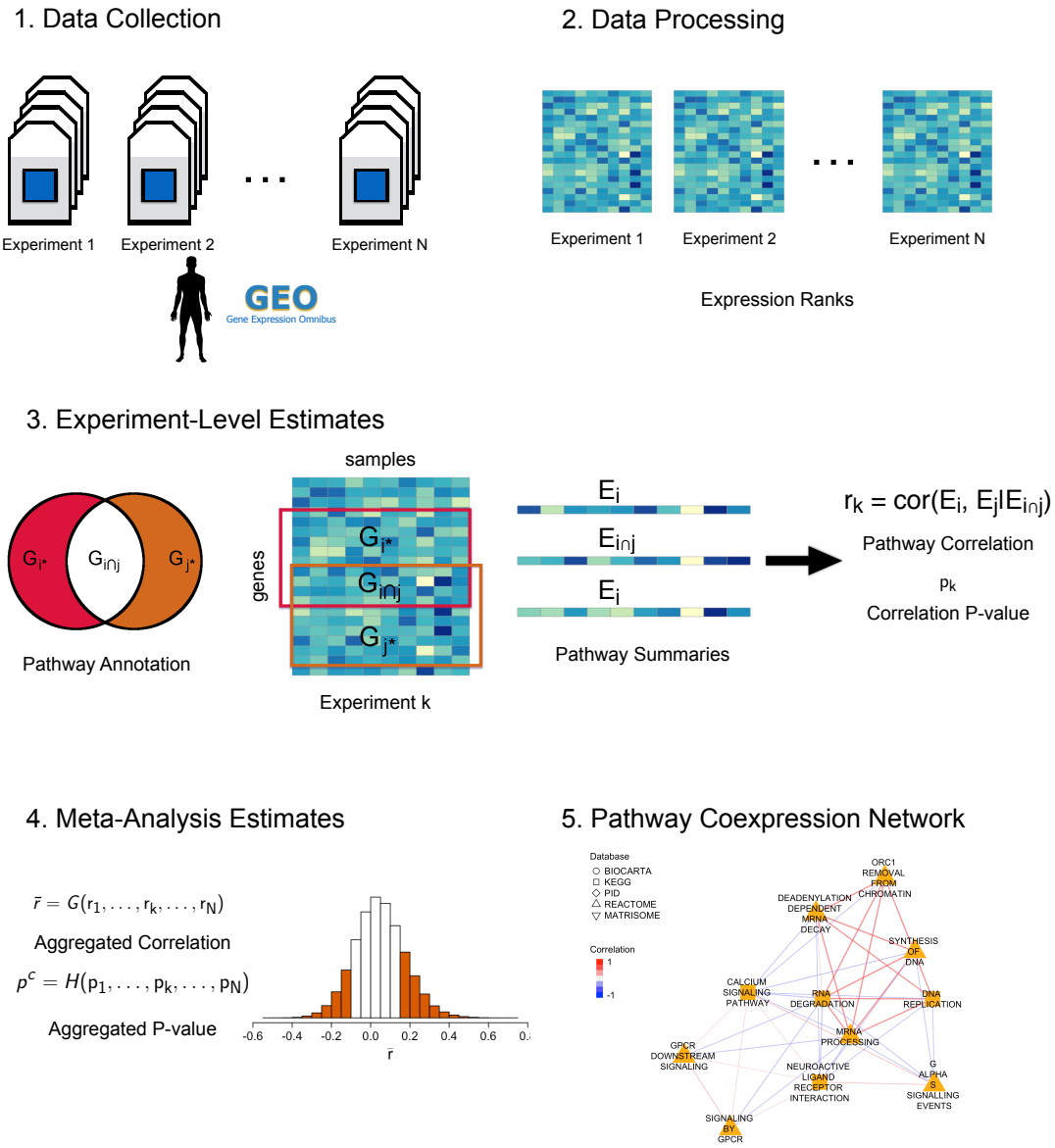
1.2 Results

1.2.1 PCxN Overview

PCxN is a weighted undirected network where the nodes represent pathways and the edges are based on the correlation between the expression of the pathways. We built PCxN using 1,330 pathways from the Molecular Signatures Database (MSigDB v.5.1)[143] and 3,207 human microarrays from 72 normal human tissues from GEO curated in Barcode 3.0[143,40,165]. The network was created by first ranking normalized gene expression levels to provide a uniform scale for all samples, an approach similar to `pathprint`[5]. Ranks provide robust summary statistics to calculate expression scores that do not depend on the dynamic range of an array[130,72]. Pathways were assigned an expression summary in each array based on the mean rank of its constituent genes. Since our gene expression background is composed of several experiments representing different tissues, for each pair of canonical pathways we estimated the correlation between their expression summaries and tested for significance in every experiment. Then we combined the experiment-level estimates into global estimates. Two pathways are connected in the coexpression network if the correlation coefficient between them is significant after adjusting for multiple comparison. Our goal is to describe the relationships between canonical pathways when their functions are related, rather than when their annotations have similar content. The pathway correlations in the network were adjusted to account for the shared genes between pathway pairs. If a pathway pair shares genes, we estimate the correlation between the pathway summaries conditioned on the summary for the shared genes (Figure 1.1).

Figure 1.1 *(following page).* **Pathway Coexpression Network (PCxN) Overview.** (1) Human gene expression arrays for normal human tissues curated from GEO in Barcode 3.0 (2) The gene expression levels were replaced by their ranks so all arrays share a common scale. (3) For each microarray experiment, we first estimated the pathway expression based on the mean of the expression ranks, then the pathway correlation adjusted for shared genes, and tested the significance of the correlation. (4) We aggregated the experiment-level estimates to get the global pathway correlation and its corresponding significance. (5) We built a pathway coexpression network based on the significant pathway correlations.

Figure 1.1. (continued)



Significant Correlations within the Ribosome Pathway

To determine how effectively PCxN captures tightly related biological functions we analysed the ribosome pathway (KEGG accession hsa03010). The KEGG *Ribosome* pathway is a gene set that represents a well characterized, meaningful and ubiquitous biological function [47,245,205,59]. We compared the pathway correlation coefficients and the corresponding p-values estimates from permuted gene sets generated from within the ribosome pathway with estimates from random gene sets. Since our method accounts for the contribution of shared genes to estimate the pathway correlation, we considered cases where the gene sets shared no genes, and cases with different degrees of gene overlap. In the no overlap case, we created ribosome gene sets by permuting the genes in the ribosome pathway (126 genes) and splitting them into two separate gene sets. The corresponding random gene sets were created by sampling 126 genes at random and splitting them into two. For the overlap cases, the gene sets were split into two gene sets sharing genes. We used the overlap coefficient to describe the overlap between gene sets represented as pathways. The overlap coefficient between two sets is the size of the intersection divided by the size of the smaller of the two sets. Unlike other measures of set overlap, the overlap coefficient between two sets is always 1 whenever one of the sets is a subset of the other, and always 0 whenever the two sets are disjoint. A key feature of PCxN is to estimate the correlation between gene sets taking into account their shared genes, so we decided to use the overlap coefficient to describe the degree of overlap between the pathway annotations. We considered 9 different overlap cases, ranging from low overlap (overlap coefficient $o_{AB} = 0.0469$) to high overlap (overlap coefficient $o_{AB} = 0.8532$).

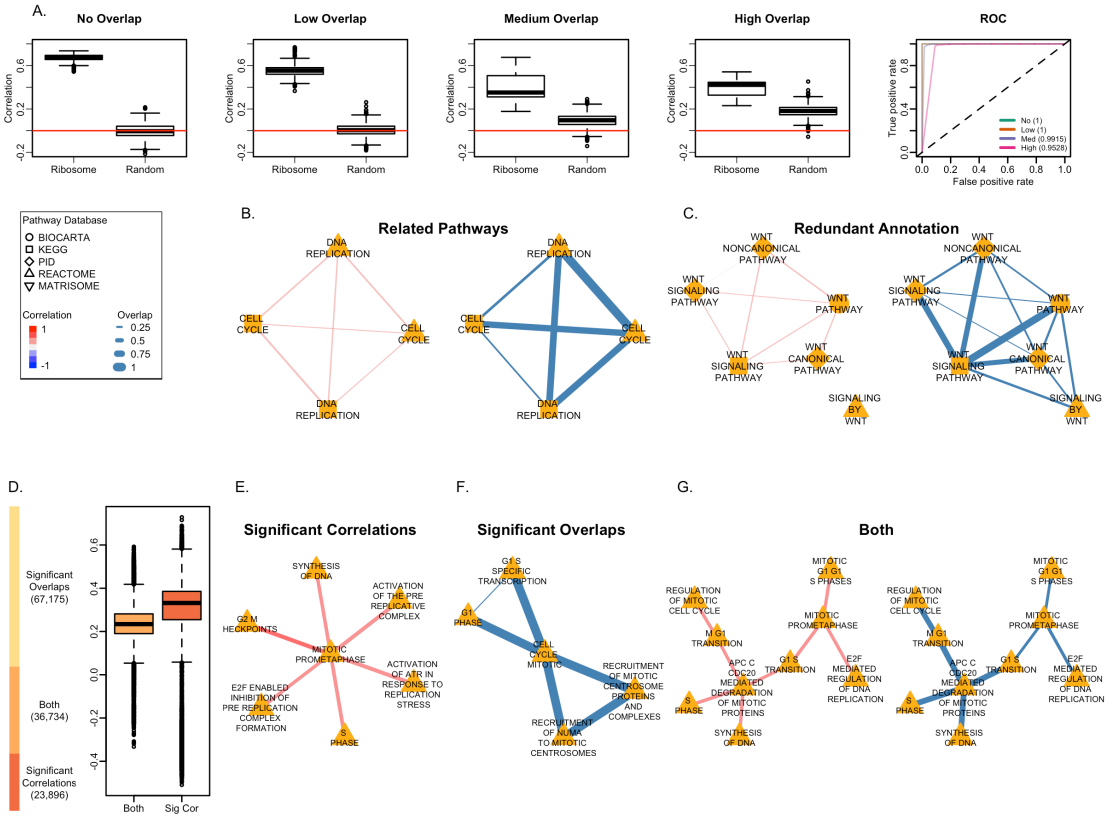


Figure 1.2. Significant Correlations between the Ribosome Pathway and Impact of Gene Overlap. (A) Boxplots of the correlation estimates between the Ribosome gene sets and random gene sets, and receiver operating characteristic (ROC) curves under different degrees of overlap: no overlap, low overlap (overlap coefficient 0.0469, AUC = 1), medium overlap (overlap coefficient 0.5517, AUC = 0.9915) and high overlap (overlap coefficient 0.8532, AUC = 0.9528). The shape of the node in the following networks corresponds to the pathway database. For coexpression networks, the edge color indicates the value of the correlation and edge width is proportional to the correlation magnitude. For the overlap networks, the edge width is proportional to the overlap coefficient. (B) Pathway coexpression and overlap network for the KEGG and Reactome annotations of the *Cell Cycle* and *DNA Replication* pathways. These pathways have related functions and share genes between them. (C) Pathway coexpression network and overlap network for different versions of the *Wnt Signaling* pathway. In the coexpression network, missing edges correspond to correlations that are not significant. These pathway annotations are redundant and represent the same function (D) The stacked bar plot shows the number of pathways pairs with only significant correlations in red, with only significant overlaps in yellow, and with both in orange. The boxplots show the distribution of the correlation coefficients with pathway pairs with only significant correlations (red) and with both significant overlaps and significant correlations (orange). (E) Pathway coexpression network for the Reactome pathways related to the mitotic metaphase of the cell cycle with significant correlations but no shared genes. (F) Overlap network for Reactome pathways related to the mitotic cell cycle with significant overlaps but no significant correlations. (G) Pathway coexpression network and overlap network for cell cycle phases and related processes from Reactome with both significant correlations and significant overlaps.

The correlation estimates from the ribosome gene sets are positive while the estimates for the random gene sets are smaller in magnitude and closer to zero (Figure 1.2A). Under the assumption that a significant p-value for ribosome gene sets is a true positive while a significant p-value for random gene sets is a false positive, we assessed the ability of our method to identify truly significant correlation coefficients. All the p-values from the ribosome gene sets were significant, while most of the p-values for the random gene sets were not significant. This trend is evident in the receiver operating characteristic (ROC) curves for the no overlap and overlap cases (Figure 1.2A).

1.2.2 Accounting for Gene Overlap

Pathway annotations from different sources present challenges when relating pathways: equivalent pathways with different annotations have similar but not identical names, annotations exist for equivalent but differently constituted pathways in separate databases, and pathways with completely different names share genes[210,252]. The MSigDB canonical pathways collection is a curated selection of pathway annotations from other databases: Reactome[160], KEGG[106], the Pathway Interaction Database (PID)[220], Biocarta[185], and the Matrisome Project[180].

PCxN and Redundant Pathways

An example of pathway annotation redundancy within MSigDB includes annotations from Reactome and KEGG for both the *Cell Cycle* and the *DNA Replication* pathways (Figure 1.2B). These pathways share genes between each other because they represent the same processes, and DNA replication is a function related to the cell cycle. In the Reactome annotations, the *DNA Replication*

pathway is a subset of the *Cell Cycle* pathway. The pathway correlation is significant and positive for these pathways. In other cases, there is more than one annotation for the same pathway. MSigDB has annotations from KEGG, Biocarta, Reactome and the Pathway Interaction Database (PID) for the *Wnt signaling* pathway. These annotations share genes among each other. Unlike the previous example, the correlation estimates between the Wnt signaling pathways have a small magnitude and most of them are not significant (Figure 1.2C). Our motivation to account for shared genes between pathways is to assign significant correlation coefficients between pathways representing related functions and non-significant correlation coefficients for pathways with redundant annotations representing the same function.

Impact of Gene Overlap

In order to understand the trade-offs resulting from discarding shared genes in estimating the correlation in PCxN, we compared significantly correlated pathways with pathways where the amount of shared genes is significant according to Fisher's exact test. We decided to use Fisher's exact test because this test has been widely used to describe relationships between gene sets based on shared genes in methods such as POSOC[135] , Ontologizer[83], GOstats[63]. Furthermore, the Molecular Concepts Map (MCM)[213] uses Fisher's exact test as similarity score between gene sets to build networks for gene sets. Of all canonical pathway pairs, 19% have only significant correlation coefficients, 52% have only significant overlaps and 29% have both (Figure 1.2D).

PCxN has an advantage over overlap based approaches when we consider pathways with related functions but without shared genes. For example considering the Reactome pathways, the *Mitotic*

Prometaphase pathway describes a function related to the cell cycle, is significantly correlated with other Reactome pathways involved in cell cycle, but does not have genes in common with them (Figure 1.2E). On the other hand, the correlation from PCxN is not significant between pathways with a very high gene overlap even though these pathways might represent closely related functions. For instance, pathway annotations from Reactome representing different aspects of the mitotic cell cycle as well as other closely related cell cycle processes have a significant gene overlap with the general *Cell Cycle Mitotic* pathway but are not significantly correlated (Figure 1.2F). However, some pathways with related functions have both significant correlations and significant overlaps. For instance, we identified Reactome pathways for mitotic cell cycle phases and related processes that are significantly correlated and have significant overlap among them (Figure 1.2G). The *APC/C CDC20 Mediated Degradation of Mitotic Proteins* pathway is both significantly correlated and has significant overlaps with the *Synthesis of DNA*, *S Phase*, *M/G1 Transition* and *G1/S Transition* pathways. The ubiquitin ligase anaphase-promoting complex or cyclosome (APC/C) initiates chromatid separation and entrance into anaphase[102], and the cell-division cycle protein 20 (CDC20) is an essential regulator of cell division that activates APC/C[257,258]. The *E2F Mediated Regulation of DNA Replication* pathway is significantly correlated and has a significant overlap with the *Mitotic Prometaphase* pathway which in turn is significantly correlated and has a significant overlap with the *G1/S Transition* pathway. The E2F family of transcription factors play a major role during the G1/S transition in mammalian and plant cell cycle[77].

1.2.3 Case Study: Alzheimer's Disease (AD)

With the goal of determining the value of our approach in understanding pathway relationships in complex disease, we chose an important disease for which there is abundant transcriptomic data, established genetic associations, and the need for better understanding of the roles of pathways and their relationships is fundamental to the prioritisation of drugs and drug targets. AD is a progressive multifarious neurodegenerative disorder[118,28] and the most common type of dementia. AD is one of the great health-care challenges of the 21st century[222]. Pathologically it is characterized by intracellular neurofibrillary tangles and extracellular amyloidal protein deposits contributing to senile plaques[118]. While the neuropathological features of AD are recognized, little is known about the causes of the disease and no curative treatments are available[222,118]. We chose this disease to illustrate how the PCxN can reveal important or even novel functional relationships underlying a complex pathological phenotype. We performed a series of additional analyses that bring together genes that have been identified by totally independent assays: genetic and transcriptomic surveys associated with AD.

We used genes within an AD curated list (ADCL) as the disease gene signature. The ADCL is a set of association-derived and experimental-derived genes related to AD. Consisting of 68 genes of which 61 genes were present in the PCxN gene expression background (Appendix A.8). The ADCL is the result of expert assessment of the current understanding of AD from a combination of key genes from genome-wide association studies and from functional analyses. We integrated the ADCL to PCxN first by estimating all the pairwise correlations between the summary for its constituent

genes and the summaries for the canonical pathways adjusted for overlap across each experiment in the gene expression background along with the corresponding p-values. Then, we aggregated the experiment level correlation estimates and combined the p-values. Finally, we adjusted the combined p-values from the correlations with the ADCL with the rest of the combined p-values from the correlations between the canonical pathways for multiple comparison using FDR. PCxN allowed us to identify canonical pathways significantly correlated with the curated AD gene list. The top 10 correlated pathways (Figure 1.3A) are all known to be related to Alzheimer's disease or amyloid pathology[197,272,170,147,175,268,124,37,21] and the majority of the top 25 correlated pathways (Appendix A.9) are related to immune responses. The top correlated pathway to ADCL, *GPVI Mediated Activation Cascade*, is associated with regulation of Amyloid beta (Aβ). GPVI and FCER1 initiate platelet activation that leads to activation of Syk. Syk enhances the formation of stress granules that are prevalent in AD affected brains. The stress granules produce reactive oxygen and nitrogen species that are toxic to neuronal cells. Downregulation of Syk expression reduces Aβ production and increases the clearance of Aβ across the blood-brain barrier[197]. Since PCxN does not rely on shared genes, PCxN uncovers relationships that would have been missed by methods that rely only on gene overlap to describe the relationships between pathways. All of the top ten correlated pathways (Figure 1.3B) have no genes in common with the ADCL (Appendix A.9).

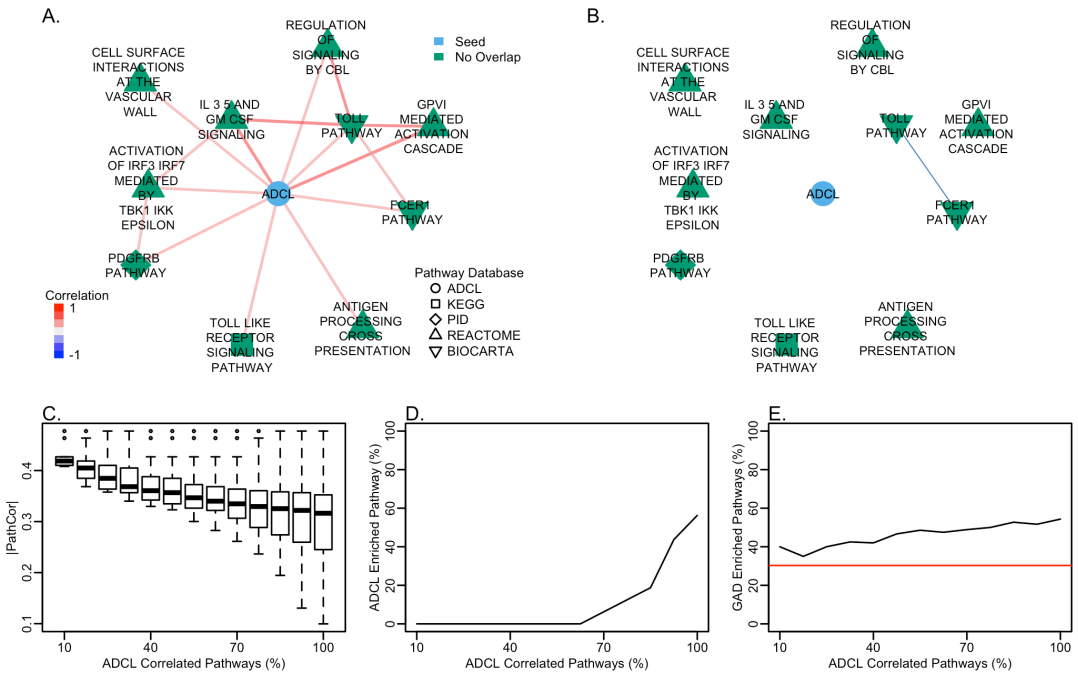


Figure 1.3. Canonical Pathways Correlated with the Alzheimer's Disease Curated List. The ADCL is colored in blue. Neighbors without genes in common with the ADCL are highlighted in green. The shape of the node corresponds to the pathway database. For the coexpression network, the edge color indicates the value of the correlation and the edge width is proportional to the correlation magnitude. For the overlap network, the edge width is proportional to the overlap coefficient. (A) Pathway coexpression network for the top pathways correlated with the ADCL (by correlation magnitude). All correlated pathways have established associations with AD: *GPVI Mediated Activation Cascade* [197], IL-3, 5 and GM-CSF signalling [272], *Antigen Processing Cross Presentation* [170], *PDGFRB Pathway* [175], *Toll Pathway* [268], *Regulation of Signaling by CBL* [147], *Toll-like Receptor Signaling* [124], *Activation of IRF3/IRF7 Mediated by TBK1/IKK Epsilon* [124], *Cell Surface Interactions at the Vascular Wall* [37], *FCER1 Pathway* [21]. (B) Shared genes (overlap coefficient) between the top pathways correlated with the ADCL. (C) Correlation magnitude of all canonical pathways correlated with the ADCL sorted by the magnitude of their correlation and split in bins of increasing size. (D) Proportion of canonical pathways enriched for the genes within the ADCL ($p < 0.001$, adjusted with FDR) present in the canonical pathways correlated with the ADCL (E) Proportion of canonical pathways enriched for genes associated with AD from the Genetic Association Database present in the pathways correlated with the ADCL ($p < 0.001$, adjusted with FDR). The red line indicates the proportion of all 1,330 canonical pathways enriched for genes within the ADCL.

To explore novel insights resulting from the use of PCxN, and as a complement to enrichment methods based on gene overlap, we compared the top ADCL correlated pathways with pathways significantly enriched for genes in the ADCL. First, we ordered all pathways correlated with the ADCL

(Appendix A.9) by the magnitude of their correlation and split the pathways into bins of increasing size (Figure 1.3C). We began with a bin including the 10 most correlated pathways. Every following bin includes 10 additional correlated pathways, so the last bin contains all pathways correlated with the ADCL. For each bin, we calculated the proportion of pathways significantly enriched for the ADCL. As we move across bins, the proportion of ADCL enriched pathways increases (Figure 1.3D). Furthermore, none of the top 30 correlated pathways was enriched for genes in the ADCL.

Enrichment for AD associated Genes in ADCL Correlated Pathways

To assess the validity of the ADCL correlation results, we tested the enrichment of genes associated with AD in pathways correlated with the ADCL using independent methods[17,233]. We assessed relationships using genetic association by retrieving genes inferred to be associated with AD from the Genetic Association Database (updated August 18, 2014). The Genetic Association Database (GAD) is a comprehensive archive of published genetic association studies that provides a repository of genetic association by data aggregation from genome-wide association and other genetic association studies[17]. We retrieved 668 genes associated with Alzheimer's disease of which 534 are present in the gene expression data from GEO (Appendix A.10). We used Fisher's exact test to determine which of the canonical pathways in PCxN correlated with the ADCL are significantly enriched for genes associated with Alzheimer's. The ADCL has 14 genes in common with genes associated with Alzheimer's in GAD, and the overlap is highly significant ($p = 7.34 \times 10^{-9}$).

Of the top 10 pathways correlated with the ADCL, 6 out of 10 were significantly enriched with genes related to Alzheimer's found by genetic association. We sorted the ADCL neighbors by the

magnitude of their correlation with the ADCL and split them into bins of increasing size (Figure 1.3C). As we move across the bins, the proportion of pathways significantly enriched for genes related to Alzheimer's in the neighbors of the Alzheimer's curated list was higher compared to all of canonical pathways; out of 1330 canonical pathways, 403 (30%) were significantly enriched after adjusting for multiple comparison using FDR and p-value cut-off of 0.001 (Appendix A.11, Figure 1.3E). The enrichment results demonstrate a significant link between the correlation of pathways with curated AD genes and genes found independently by genetic association with Alzheimer's.

1.2.4 Complement to GSEA: Revealing Relationships Between Enriched Pathways

PCxN can be used effectively to determine relationships between pathways as a complement to interpret gene set enrichment (GSE) methods. A typical GSE result is a list of gene sets that are significantly enriched by a list of query genes. PCxN can describe the relationships between the enriched gene sets using the global pathway correlation estimates. To explore correlation between gene sets enriched with a set of query genes, we used Gene Set Enrichment Analysis (GSEA)[238] to find pathways from the MSigDB canonical pathways collection enriched for genes differentially expressed in an AD expression dataset (GSE5281) consisting of genes expressed in post mortem samples of AD in the superior frontal gyrus (Appendix A.12). The expression data set consisted of 34 superior frontal gyrus samples: 11 controls (clinically and histopathologically normal aged human brains) and 23 affected with AD[139] (Appendix A.13).

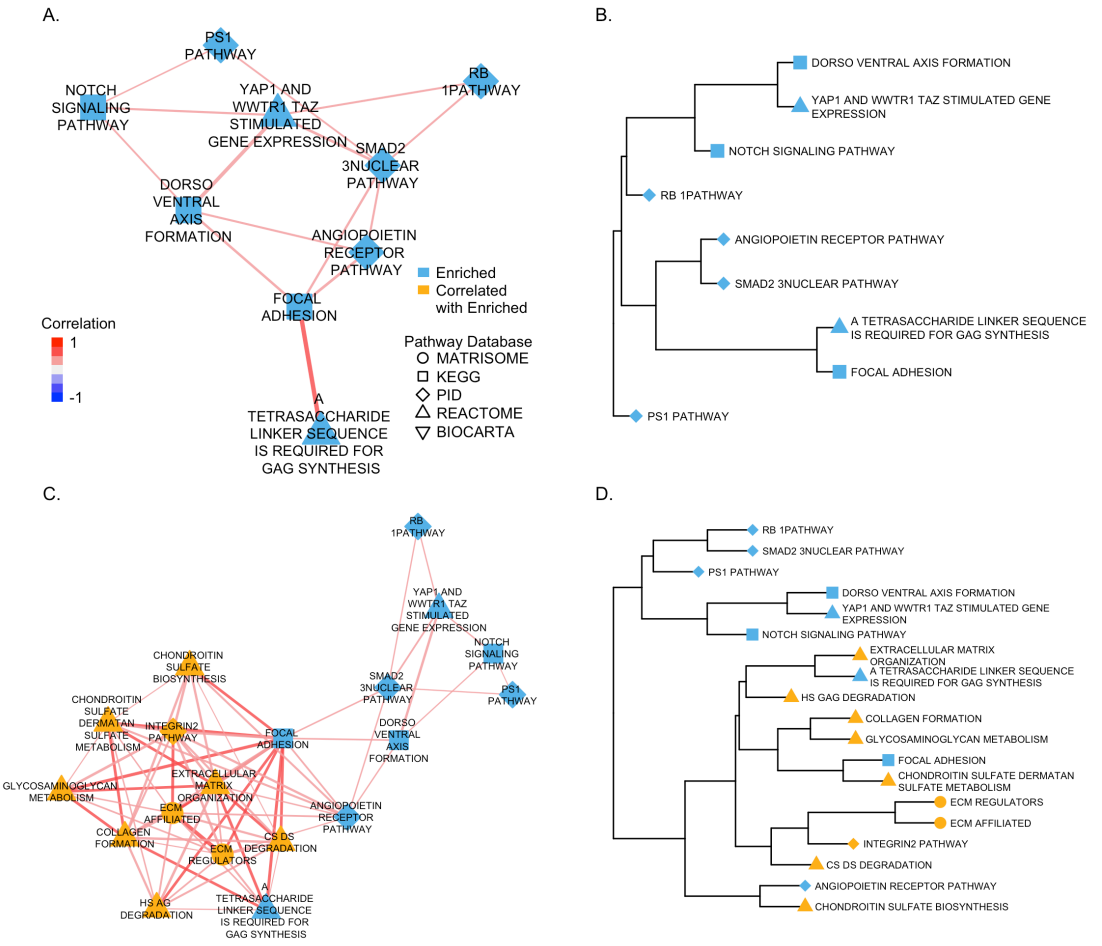


Figure 1.4. Pathway Coexpression for GSEA Enriched Canonical Pathways. GSEA enriched pathways are colored in blue, correlated pathways are yellow. The shape of the node corresponds to the pathway database, the edge color indicates the value of the correlation and the edge width is proportional to the correlation magnitude. (A) Pathway coexpression network for the top 10 GSEA enriched canonical pathways. (B) Hierarchical clustering using average linkage and $1 - |\text{PathCor}|$ as the distance between the top 10 GSEA enriched canonical pathways. (C) Pathway coexpression network for the GSEA enriched pathways and their top 10 correlated pathways (by $|\text{PathCor}|$). (D) Hierarchical clustering using average linkage and $1 - |\text{PathCor}|$ as the distance between the top 10 GSEA enriched canonical pathways and their top 10 correlated pathways.

We chose to examine the functional relationships among the top ten enriched pathways identified by GSEA. Functionally, they all appear to be consistently associated with the AD literature

(e.g. the *PS1 Pathway* role in AD[229]). We retrieved significant correlations between the enriched pathways to explore their functional relationships as revealed by PCxN (Figure 1.4A). To explore the most closely functionally related pathways, we clustered the enriched pathways based on their correlations (Figure 1.4B). The cluster containing the highest correlations consists of pathways involved in cell adhesion and oxidative stress response (*Focal Adhesion*, *A Tetrasaccharide Linker Sequence is Required for GAG Synthesis*, *Angiopoietin Receptor* and *SMAD2/3 Nuclear Pathway* (Appendix A.14)). These pathways shared reported functions. Focal adhesions have been implicated in regulating Aβ signalling and cell death in AD[30]. As part of cell adherence to the extracellular matrix (ECM), integrins are activated and the focal adhesion pathway is activated. The ECM/integrin/focal adhesion pathway is involved in the regulation of anchorage-dependent cell survival. Cell adhesion to ECM and overexpressing FAK (focal adhesion kinase), member of *Focal Adhesion Pathway*, is protective against oxidative stress, which has been observed in AD brains[159]. FAK also has the ability to regulate several other cell-death or survival pathways[30]. Members of *A Tetrasaccharide Linker Sequence is Required for GAG Synthesis* are also involved in cell adhesion, which plays an important role in cell death/survival. Members of this pathway include neurocan and brevican, whose expression is mostly restricted to neuronal tissues[264]. Loss of brevican is associated with loss of synapses[178], while Aβ has been shown to increase neurocan expression in astrocytes[265]. In addition to adhesion molecules, angiopoietins (members of the clustered *Angiopoietin Receptor Pathway*) share function as they are activated in response to oxidative stress. Elevated Angiopoietin-1 serum levels can be observed in patients with AD[224]. The closely clustered *SMAD2/3 Nuclear Pathway* contains SMAD3 which regulates expression of angiogenic molecules in tumor cells and vascularization in tumor le-

sions[149]. SMADs transduce extracellular signals from Transforming Growth Factor β (TGFβ) to the nucleus[155]. SMAD3, one of the key members of the SMAD2/3 nuclear pathway, is down regulated in AD[253], while TGFβ is upregulated. The imbalance between SMAD3 and TGFAβ signalling, shifts the regulatory signalling towards a dysregulated inflammatory activation potentially leading to neurodegenerative changes, such as decreased Aβ clearing[253].

The other top ten pathways identified in this GSEA have also been associated with AD and some show documented functional relationships. PS1 is well known as a common cause of familial AD[107]. *Dorso-ventral Axis Formation* has been suggested as one of the pathways regulated by miRNAs identified in a bioinformatics study of Drosophila AD models[115]. Notch is coexpressed with PS1 and altered in AD affected brains[19], YAP1 and WWTR1/TAZ mediate gene transcription induced by the Aβ protein precursor and its paralogues[191]. Finally, increased levels of hyperphosphorylated RB protein have been observed in AD[211] indicating that neurons in AD attempt to re-enter the cell cycle[244].

1.2.5 Complement to GSEA: Expanded Enriched Gene Sets

In addition to providing relationships between the GSEA results, PCxN can provide potentially novel relationships by retrieving canonical pathways significantly correlated with the pathways identified as enriched. We retrieved the top 10 canonical pathways which were the most correlated with the AD GSEA enriched gene sets, and clustered the correlated pathways along with the results from GSEA (Figure 1.4C-D). Most of the top correlated neighbors are components of extracellular matrix (ECM) and form a highly-correlated cluster (Figure 1.4D) with the top correlated GSEA pathways. The ECM components revealed by PCxN have been highly studied in relation to

Alzheimer's [250,54,259,178,46]. The ECM changes significantly during the early stages of AD [134], but only a limited number of individual ECM components have been studied so far [227].

1.2.6 Exploring PCxN

We created a user-friendly webtool (http://pcxn.org) that can be used to interactively explore and visualise pathway relationships found in PCxN. The tool allows a user to query the various pathway databases using one or more pathways and retrieve correlation estimates, p-values and overlap coefficients. Since the correlations adjusted for shared genes are a complementary perspective to relationships based on gene overlap, the webtool also provides the option to view coexpression networks based on correlation coefficients not adjusted for shared genes in addition to the PCxN coexpression network that is based on the adjusted correlation. The results are presented through heatmaps (which also offer clustering of pathways), interactive networks (with multiple pre-made structures) and data tables. Pathway members are also retrievable along with their descriptions. In addition, PCxN is available as Bioconductor software (http://bioconductor.org/packages/pcxn/) and data (http://bioconductor.org/packages/pcxnData/) packages which contain the same exploratory/visualization functionality and data as the webtool.

1.3 Discussion

We have developed and described PCxN, a coexpression method to describe global relationships between pathways. PCxN estimates the correlation between 1,330 canonical pathways using a curated collection of 3,207 microarrays in 134 experiments from 72 normal human tissues. We integrated a

wide range of experiments by estimating the correlation between summaries of the pathway expression, testing their significance in every experiment, and then aggregating the experiment-level estimates into global estimates. We used gene sets derived from permutations of the *Ribosome* pathway (KEGG) and random gene sets to show that PCxN effectively captures relationships between gene sets with related functions while discarding relationships from random gene sets. The correlation estimates between the ribosome gene set were positive and significant, while the correlation estimates for random gene sets were not significant and with a magnitude close to zero. These results suggest that the correlation between two pathways with related functions is significant.

The influence of redundant annotations across pathways databases is often overlooked. Pathway databases often include pathways that share genes with one another to varying degrees. Shared genes between pathways can either be a consequence of closely related functions or redundant annotation from different sources. Ignoring such redundancies during pathway analysis can lead to identifying pathways relationships due to high content-similarity, rather than truly related biological mechanisms. PCxN adjusts the correlation between pathways by conditioning on the shared genes. The correlations between redundant annotations for the *Wnt signaling* pathway had a small magnitude and were mostly not significant. When pathways share genes due to related functions, the correlations between them might be significant depending on the degree of the overlap. For instance, we found pathways for mitotic cell cycle and related processes that were significantly correlated and had significant overlaps between them. The significant correlations and significant overlaps between these pathways revealed known relationships between ADC/C, CDC20 and the E2F family of transcription factors with the mitotic cell cycle. However, the correlations between a different set of

pathways representing other aspects of the mitotic cell cycle, such as the *Mitotic Cell Cycle* and the *G1 Phase* pathways and related processes, such as the *Recruitment of Mitotic Centromere Proteins and Complexes*, were not significant while the overlap was highly significant. PCxN was successful in uncovering relationships between the *Mitotic Prometaphase* pathway and other cell cycle related pathways such as the *G2/M Checkpoints* and the *S Phase* that do not have genes in common.

PCxN provides powerful means to generate models for complex diseases by providing pathways significantly correlated with an assay-independent disease gene signature. We used PCxN to identify key processes related to Alzheimer's disease (AD) using an AD curated list (ADCL). The top pathways correlated with the ADCL have known relationships with AD or amyloid pathology. Furthermore, the correlated pathways were significantly enriched for genes associated with AD independently derived from genome wide association studies. These results show the value of PCxN in finding biological processes associated with complex diseases using gene signatures. PCxN provides a powerful contribution to the interpretation of the gene set enrichment methods by describing the relationships between enriched pathways independent of gene overlap. We used PCxN to describe the relationships between pathways identified as enriched by GSEA in a published microarray gene expression experiment profiling the effect of AD in the superior frontal gyrus. We expanded the scope of gene set enrichment results by retrieving pathways correlated with the enriched pathways. The top pathways correlated with the enriched pathways are components of extracellular matrix (ECM) and form a highly correlated cluster. We note that the ECM undergoes significant changes during the early stages of AD, but only a few ECM components have been studied. The relationships between the ECM pathways from PCxN could provide leads to future studies of the

individual ECM components.

PCxN relies on the completeness and correctness of pathway annotations to relate biological processes. Also, PCxN only considers a pathway as a gene list, omitting any knowledge of the interaction between its members. PCxN is also limited by the gene expression data used to estimate the correlations. The current implementation only uses one microarray platform and a curated expression background. It is widely accepted that pathway activation is phenotype dependent. Using the PCxN approach it will be possible to explore whether pathway-pathway relationships change in relationship to a phenotype, or if consistent functional links prevail irrespective of cell state. Further work is required to investigate how network topology changes with expression background, and in particular into whether pathway networks are significantly disrupted in disease. This implementation of PCxN does not take advantage of the growing number of publicly available RNA-seq data. In future, the method will be expanded to include a wider range of pathway annotations and to use gene expression data from other platforms such as RNA-seq.

PCxN establishes the utility of describing relationships between pathways in a broad context. By using a diverse set of gene expression experiments, PCxN leverages correlation estimates across various human tissues effectively capturing relationships regardless of shared genes. We expect that PCxN can serve as a basis for a high-level map of the relationships between biological process. We built an interactive web-tool that provides a user-friendly portal to explore the PCxN at `http://pcxn.org/`, as well as a Bioconductor software (`http://bioconductor.org/packages/pcxn/`) and data (`http://bioconductor.org/packages/pcxnData/`) package.

1.4 Materials and Methods

1.4.1 Data Collection

Gene Expression Data Retrieval

We used 134 experiments with 3,207 Affymetrix Human Genome U133 Plus 2.0 microarrays from 72 normal human tissues manually curated in Barcode 3.0[165] (Appendix A.5). The curated microarrays in Barcode 3.0 were filtered to exclude poor quality samples[165,166]. We used the R package GEOquery[45] to retrieve raw CEL files from the Gene Expression Omnibus (GEO)[13]. We processed the raw data with fRMA[161]. We obtained the annotation for the array platform from[31]. To resolve redundancies, multiple probes were mapped to unique Entrez Gene IDs by their mean expression level.

Pathway Annotations

We retrieved the C2: Canonical Pathways collection from MSigDB[238] (v5.1 updated January 2016, Appendix A.6). The canonical pathways collection from MSigDB is a curated selection of pathway annotations from other databases: Reactome[160], Kyoto Encyclopedia of Genes and Genomes (KEGG)[106], the Pathway Interaction Database (PID)[220], Biocarta[185], the Matrisome Project[180], the IUBMB-Sigma-Nicholson Metabolic Pathway Charts[184], UCSD Signaling Gateway[137], Science's Signal Transduction Knowledge Environment[81], and the annotation for a Wnt Signaling Pathway PCR array from QIAGEN (Table 1.1).

Table 1.1. Pathway Annotation Sources for MSigDB Canonical Pathways Collection.

Source	Pathways
Reactome	674
Biocarta	217
Pathway Interaction Database	196
Kyoto Encyclopedia of Genes and Genomes	186
Science's Signal Transduction Knowledge Environment	28
Matrisome Project	10
IUBMB-Sigma-Nicholson Metabolic Pathway Charts	10
UCSD Signaling Gateway	8
QIAGEN	1

Sources for the pathway annotations in MSigDB Canonical Pathways collection.

1.4.2 Experiment-Level Estimates

Since the microarrays from the gene expression background belong to different experiments representing different tissues, pooling the microarrays to estimate the correlation between pathways would ignore the underlying structure of the data. Even if the correlations are homogeneous, pooling the data is not a valid procedure in general. The pooled estimates may be severely biased due to the heterogeneity of the experiments[3,91]. Instead of pooled estimates, we first estimated the pathway correlation coefficients and their corresponding p-values for each experiment, and then we com-

bined the experiment-level estimates into global estimates.

1.4.3 Pathway Expression

We represent an experiment with L samples as the $K \times L$ matrix X where K is the total number of genes in the array. Thus, the element x_{kl} of the matrix X corresponds to the expression for gene k in array l. For each array, the genes were ranked by their expression level. Rank normalizations do not depend on the dynamic range of an array and provide a common range.

$$\text{genes} \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1L} \\ x_{21} & x_{22} & \cdots & x_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ x_{K1} & x_{K2} & \cdots & x_{KL} \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \end{matrix} \begin{matrix} \text{arrays} \\ \\ \\ \\ \end{matrix} \longrightarrow \text{genes} \begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1L} \\ R_{21} & R_{22} & \cdots & R_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ R_{K1} & R_{K2} & \cdots & R_{KL} \end{bmatrix} \begin{matrix} \text{arrays} \\ \\ \\ \\ \end{matrix}$$

We represent the expression ranks as the $K \times L$ matrix S, where L is the total number of arrays and K is the total number of genes in the array. Since within each array the genes are ranked by expression level, from 1 (low expression) to K (high expression), the entries of the matrix S are

$$S_{kl} = \operatorname*{rank}_{1 \le l \le L}(x_{kl})$$

where x_{kl} is the expression level for gene k in array l.

In this approach pathways are represented as gene sets: groups of functionally related genes. Thus, a pathway is represented by its gene set annotation $G = \{g_1, \ldots, g_n\}$. The pathway expres-

sion E is a gene set summary statistic based on the expression ranks of the pathway genes; the pathway expression E is the mean of the expression ranks of the pathway genes.

$$\begin{array}{c} \\ \text{genes} \end{array} \overset{\text{arrays}}{\begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1L} \\ R_{21} & R_{22} & \cdots & R_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ R_{m1} & R_{n2} & \cdots & R_{nL} \\ \vdots & \vdots & \ddots & \vdots \\ R_{K1} & R_{K2} & \cdots & R_{KL} \end{bmatrix}} \longrightarrow \overset{\text{Mean Ranks}}{\begin{bmatrix} E_1 & E_2 & \cdots & E_L \end{bmatrix}}$$

Consider an experiment with L samples, the experiment-level summary for pathway G is given by the $L \times 1$ vector E with entries

$$E_l = \frac{1}{n} \sum_{g \in G} S_{gl}$$

To calculate E, first we take the rows from S corresponding to the genes $\{g_1, \ldots, g_n\}$ to get the matrix of ranks of the pathway constituent genes, and then we take the mean across the columns of this matrix, producing the $L \times 1$ vector E.

Compared to other summary statistics, the mean is fast to compute and easy to interpret. We considered several approaches for the pathway summary statistic, but we found that in most cases the mean performed well. For instance, we considered a summary based on principal components

analysis (PCA) but the variance explained by the first principal component was less than 50% for all canonical pathways in the majority of the gene expression experiments from the curated collection of normal human tissues (Appendix A.2).

1.4.4 Pathway Correlation

Shrinkage Estimator

We used a shrinkage estimator to compute the experiment-level pathway correlation coefficients. In our setting, a shrinkage estimator will give more reliable experiment-level correlation estimates for experiments with few samples and will set correlation coefficients with a small magnitude to 0 [221]. The shrinkage estimator R^* is a linear combination of the standard correlation estimator R and a restricted submodel of the correlation matrix

$$R^* = \lambda T + (1 - \lambda)R$$

where $0 \leq \lambda \leq 1$, R is the empirical correlation matrix and T is identity matrix.

The restricted submodel T assumes that all of the variables are uncorrelated. The optimal λ is found by minimizing the mean squared error $L(\lambda)$ between the shrinkage estimator R^* and the true correlation matrix P.

$$L(\lambda) = ||R^* - P||_F^2 = ||\lambda T - (1 - \lambda)R - P||_F^2 = \sum_{i=1}^{p} \sum_{j=1}^{p} (\lambda t_{ij} + (1 - \lambda)r_{ij} - \varrho_{ij})^2$$

The analytical solution λ^* for the optimal λ[131]

$$\lambda^* = \text{argmin } L(\lambda)$$

is guaranteed to exist and minimize the mean squared error $L(\lambda)$. The solution[221] is given by

$$\lambda^* = \frac{\sum_{k \neq l} \text{Var}(r_{kl})}{\sum_{k \neq l} r_{kl}^2}$$

Gene Overlap

Since genes can be involved in more than one biological process and often pathways share genes, we accounted for the gene overlap between pathways to determine the coexpression between two pathways. Our goal is to describe relationships between patwhays representing related functions rather than pathways with similar annotations. For pathway i with gene set G_i and pathway j with gene set G_j there are two possible cases for shared genes: the gene sets overlap or do not overlap.

Non-Overlapping Gene Sets

First we calculated the expression summary E_i and E_j for pathways i and j respectively. Then, we estimated the pathway correlation as the Spearman correlation between the two pathway expression summaries

$$\text{PathCor}(i,j) = \text{cor}(E_i, E_j)$$

Overlapping Gene Sets

Our approach to deal with overlapping pathway gene sets was to condition the correlation between the summaries for the pathways G_i and G_j on the summary for the genes common to both pathways ($G_{i \cap j} = G_i \cap G_j$).

First, we calculated the summaries E_i,E_j , and $E_{i \cap j}$ corresponding to pathway G_i , pathway G_j and the shared genes $G_{i \cap j}$. Then we estimated the partial correlation between the pathway summaries conditional on the summary for the shared genes

$$\text{PathCor}(i,j) = \text{cor}(E_i, E_j | E_{i \cap j})$$

Hypothesis Testing

We used a t-test to determine which experiment-level correlation coefficients were significantly different from 0.

$$H_0 : \text{PathCor}(i,j) = 0 \qquad H_1 : \text{PathCor}(i,j) \neq 0$$

For the correlation coefficients between pathways without shared genes, the t-test is given by

$$t = r\sqrt{\frac{n-2}{1-r^2}} \sim t_{n-2}$$

where r is the experiment-level correlation estimate.

For the correlation coefficients between pathways with shared genes, the t-test is given by

$$t = r\sqrt{\frac{n-3}{1-r^2}} \sim t_{n-3}$$

where r is the experiment-level conditional correlation estimate.

1.4.5 Meta-Analysis Estimates

Hunter-Schmidt Estimator

We used the experiment-level correlation estimates to compute the overall correlation between two gene sets with a weighted average

$$\bar{r} = \frac{\sum_{i=1}^{N} n_i r_i}{\sum_{i=1}^{N} n_i}$$

where n_i is the number of samples for experiment i , r_i is the correlation estimate for experiment i and N is the total number of experiments[3].

Liptak P-value Aggregation

Since we estimated the correlation coefficients at the experiment level, we first obtained a p-value from each of the experiments by testing if the experiment-level correlation was significant. In order to determine the significance of the overall correlation coefficient we combined the p-values from

each experiment using Liptak's method[148,146]. The combined p-values across all experiments are given by

$$p^c = 1 - \varphi(Y)$$

where

$$Y = \frac{\sum_{i=1}^{N} n_i \Phi^{-1}(1 - p_i)}{\sqrt{\sum_{i=1}^{N} n_i^2}}$$

φ is the standard normal probability density function, Φ^{-1} is the standard normal inverse cumulative distribution function, n_i is the number of samples for experiment i , p_i is the p-value for experiment i and N is the total number of experiments.

After aggregating the experiment-level p-values for all pathway pairs, we adjust the combined p-values for multiple comparison using the Benjamini–Hochberg FDR method[18].

1.4.6 Overlap Coefficient

The overlap coefficient is a similarity measure for the overlap between two sets. For two sets G and H, the overlap coefficient is given by

$$o_{GH} = \frac{|G \cap H|}{\min\{|G|, |H|\}}$$

where $0 \leq o_{GH} \leq 1$. The overlap coefficient is simply the size of the intersection divided by the size of the smaller of the two sets. We chose the overlap coefficient instead of other measures of overlap like the Jaccard index because it highlights whenever a pathway is a fully contained within another pathway. If a set G is a subset of H, the overlap coefficient is always 1. On the other hand, if the sets G and H are disjoint, the overlap coefficient is always 0.

1.4.7 Ribosome Gene Sets

The annotation for the Ribosome pathway was retrieved from the KEGG REST server using the KEGGREST package (v. 1.10.1) [Tenenbaum]. We ran 1000 iterations for the no overlap and each overlap case using gene sets derived from the ribosome pathway annotation and random gene sets.

No Overlap Case

For the no overlap case, the KEGG Ribosome pathway was split in half. The ribosome pathway annotation, composed of 126 genes, was split into two non overlapping gene sets with 63 genes each with the following steps

1. Permute indexes of the genes belonging to the ribosome pathway
2. Split the gene set into two non overlapping gene sets A and B
3. Calculate the pathway summaries E_A and E_B for gene sets A and B respectively
4. Calculate the pathway correlation using the pathway summaries E_A and E_B

For the random gene set, we sampled 126 genes present in the gene expression background, and split them with the following steps

1. Sample 126 genes from the background
2. Split the genes into two non overlapping gene sets A^r B^r with 63 genes each
3. Calculate the pathway summaries E^r_A and E^r_B for gene sets A^r and B^r respectively
4. Calculate the pathway correlation using the pathway summaries E^r_A and E^r_B

Overlap Cases

We created representative cases of gene overlap between two gene sets. In particular, we created two overlapping sets s_1 and s_2 from n distinct elements. In the first step, the two sets s_1 and s_2 share all but one element. In each consecutive step, we shift the indexes of one of the sets to decrease the number of shared elements between s_1 and s_2 until the last step when the two sets s_1 and s_2 do not have any elements in common.

Step 1 $s_1 = \{1, \ldots, \overbrace{(n-1)}^{\leftarrow}\}$

$s_2 = \{1, \ldots, (n-1), n\}$

Step 2 $s_1 = \{1, 2, \ldots, (n-1)\}$

$s_2 = \{\underset{\rightarrow}{2}, \ldots, (n-1), n\}$

Step 3 $s_1 = \{1, 2, \ldots, \overbrace{(n-2)}^{\leftarrow}\}$

$s_2 = \{2, \ldots, (n-2), (n-1), n\}$

Step 4 $s_1 = \{1, 2, 3, \ldots, (n-2)\}$

$s_2 = \{\underset{\rightarrow}{3}, \ldots, (n-2), (n-1), n\}$

$\vdots$

Step n $s_1 = \{1, \ldots, (n - \lceil n/2 \rceil)\}$

$s_2 = \{(n - \lceil n/2 \rceil + 1), \ldots, n\}$

In order to consider different scenarios for the amount of shared genes between pathways, we built 9 different configurations of overlapping gene sets. These 9 overlap cases ranged from low overlap ($o_{AB} = 0.0469$) to high overlap ($o_{AB} = 0.8532$).

For the overlap cases, we split the KEGG Ribosome pathway was split into overlapping gene sets.

1. Permute indexes of the genes belonging to the ribosome pathway.
2. Split the gene set into two overlapping gene sets $\mathcal{A}$ and B.
3. Get the shared genes $\mathcal{A} \cap B$ between sets $\mathcal{A}$ and B.
4. Calculate the pathway summaries $E_{\mathcal{A}}$, E_B, and $E_{\mathcal{A} \cap B}$.
5. Calculate the partial correlation between the summaries for the genes sets $\mathcal{A}$ and B, conditional on the shared genes $E_{\mathcal{A} \cap B}$.

For the random gene sets, we sampled 126 genes present in the gene expression background and then split them into overlapping gene sets.

1. Sample 126 genes from the background.
2. Split the gene set into two overlapping gene sets $\mathcal{A}^r$ and B^r .
3. Get the shared genes $\mathcal{A}^r \cap B^r$ between the gene sets $\mathcal{A}^r$ and B^r.
4. Calculate the pathway summaries $E_{\mathcal{A}^r}$, E_{B^r} and $E_{\mathcal{A}^r \cap B^r}$.
5. Calculate the partial correlation between the summaries for the genes sets $\mathcal{A}^r$ and B^r, conditional on the shared genes $E_{\mathcal{A}^r \cap B^r}$.

ROC Curves Based on P-values

We generated a set of p-values based on the random gene sets and another set of p-values based on the ribosome gene sets. Assuming that a significant p-value for ribosome gene sets is a true positive while a significant p-value for random gene sets is a false positive, we assessed the ability of our

method to identify truly significant correlation coefficients (Table 1.2). We used different p-value cut-offs for significance to build a receiver operating characteristic (ROC) curve.

Table 1.2. Confusion Matrix for the Ribosome and the Random Gene Sets.

	Ribosome Gene Set	Random Gene Set
Significant	True Positive (TP)	False Positive (FP)
Not significant	False Negative (FN)	True Negative (TN)

Assignment of true positive (TP) and false positives (FP) based on the different p-value cut-offs for significance from the ribosome and the random gene sets.

1.4.8 Significant Pathway Overlap

We used Fisher's exact test to identify significant overlaps between all pathway pairs. For pathway i with gene set G_i and pathway j with gene set G_j, we used a contingency table based on their shared genes to perform an one-sided Fisher's exact test (Table 1.3).

Table 1.3. Contigency Table for Shared Genes between Pathways.

	Genes in G_i	Genes in G_i^c
Genes in G_j	$G_i \cap G_j$	$G_i^c \cap G_j$
Genes in G_j^c	$G_i \cap G_j^c$	$G_i^c \cap G_j^c$

The contigency table splits the gene sets of pathways i and j into four disjoint sets. The shared genes between the two pathways, $G_i \cap G_j$, the genes unique to pathway i, $G_i \cap G_j^c$, the genes unique to pathway j, $G_i^c \cap G_j$, and the genes that do not belong to either pathway i or j, $G_i^c \cap G_j^c$.

Then we adjusted the corresponding p-values for multiple comparison using FDR, and considered an overlap significant if $p < 0.05$.

1.4.9 PCxN Webtool and Bioconductor Packages

The PCxN webtool is available at http://pcxn.org/. The webtool was built using open source software and libraries. The back-end of the website was developed using JSP(JavaServer Pages) powered by a Tomcat (http://tomcat.apache.org/, version 7.0.52) HTTP-server. MySQL (https://www.mysql.com/, version 5.5.46) was used to manage a relational database containing pathway correlation coefficients. The front-end user interface was developed using HTML and specialized libraries. The Jquery.js library (http://jquery.com/, version 2.1.1) was used to handle events. The canvasXpress.js library (https://canvasxpress.org/, version 13.5) was used to

build heatmaps. The cytoscape.js library (http://js.cytoscape.org/, version 2.7.11) was used to build networks. PCxN is also available through Bioconductor as two distinct but interacting R packages. The `pcxn` package (http://bioconductor.org/packages/pcxn/) contains exploration and visualization wrapper functions that use data matrices stored in the `pcxnData` package (http://bioconductor.org/packages/pcxnData/).

2

Cross-Species Analysis of Gene Expression Data

2.1 Introduction

2.1.1 Mouse as a Model Organism

For decades the mouse has been the most widely used organism to model human physiology and disease[232]. Despite the divergence of mice and humans from a common ancestor approximately 90 million years ago, mice have close evolutionary affinities with humans[92]. On average, the protein coding regions of the human and mouse genome are 85% identical[232,269]. Furthermore, mice have numerous properties that facilitate their handling relative to other mammals and vertebrates. Mice have fast reproduction, short life spans, are inexpensive, easy to handle and can be manipulated at the molecular level[60]. Hence, the mouse has been used as an animal model in biomedical research to study mammalian development, disease, and to test drugs for over 50 years[247,248,39,6]. Under the assumption that if two proteins have similar sequences they share similar properties and functions[261,88], common functions and the elements of the genome that encode them are conserved and comparable between humans and mice. Despite great progress in understanding the genetics, anatomy and physiology of the mouse, the attrition rate of compounds tested in Phase II clinical trials is still high [5], evidencing the lack of a comprehensive knowledge of the molecular differences between mice and humans that limit the translation of mouse studies to humans[48].

2.1.2 Orthologous Genes

To compare the gene expression data between different species, genes across different species have to be matched and paired based on shared ancestry. A homolog is a gene inherited in two species by

a common ancestor[64,74]. There are two fundamentally distinct types of homologous relationships between genes according to their mode of descent from their common ancestor: orthologs and paralogs[66,65]. Orthologs are genes in different species that have evolved trough speciation events only while paralogs arise by duplication events[66,74,64]. The definition of orthologs is formally based on evolutionary criteria, but is often taken to imply functional conservation[73,237,242]. The assumption of conserved function between orthologs has been supported even between relatively distant species, by observations of conserved phenotypic effects when orthologs were subject to knock-in experiments[69,80] or in situ[208,182], clarifying the role of genes involved in human diseases. On the other hand, duplication makes room for paralogs to evolve new functions in paralogs[188]. Therefore, genes are usually paired by orthologs to compare gene expression profiles across different species.

Methods to identify orthologs can be classified as similarity-based methods and phylogeny-based methods[64]. Similarity-based methods are based on the bidirectional best hit (BBH) assumption. The BBH assumption is that if a certain function is required in two different species, it is most likely that this function is carried out by the pair of the most similar genes from these two species in terms of sequence similarity[74,64,99,242]. On the other hand, phylogeny-based methods try to reconcile inferred relationships between genes based on sequence similarity with phylogenetic trees that describe the evolutionary relationships between different species[64].

2.1.3 Gene Expression Comparisons across Species

Phenotypic differences between species are often driven by evolutionary adaptations in gene expression. However, many developmental programs are deeply conserved across species. Gene expression

comparisons among homologous genes across vertebrate species and tissues have been explored using microarrays[167,140,38] and RNA-seq[24,172,12]. All of these early studies concluded that gene expression was more similar between homologous tissues of different species than between different tissues of the same species. These results have been interpreted as a reflection of evolutionarily conserved transcriptional programs driving the production of major proteins that define specific tissues such as heart, lung or liver[26,239]. Furthermore, these results support the assumption that vertebrate animal models, such as mouse, serve as useful models of the physiology of specific human organs despite millions of years of evolutionary divergence.

Early gene expression comparisons between human and mouse tissues were very challenging since the gene expression was measured with a different platforms for human and mouse samples. In 2005, Yanai et al. analyzed human and mouse gene expression microarray data from different tissues and concluded that "any tissue is more similar to any other human tissue examined than to its corresponding mouse tissue"[266]. Two follow-up papers showed that platform-specific variability was driving the observed differences between species[141,262]. The results from Yanai et al. suffered from a serious confounding issue, the gene expression was measured using different platforms for human and mouse samples. The gene expression for the human samples were measured using the Affymetrix GeneChip Human Genome U95A while the mouse samples were measured with the Affymetrix GeneChip Mouse Genome U74A. Therefore, the observed differences between human and mouse tissues could be due to technical differences between the two different microarray platforms rather than actual differences between the two species[133].

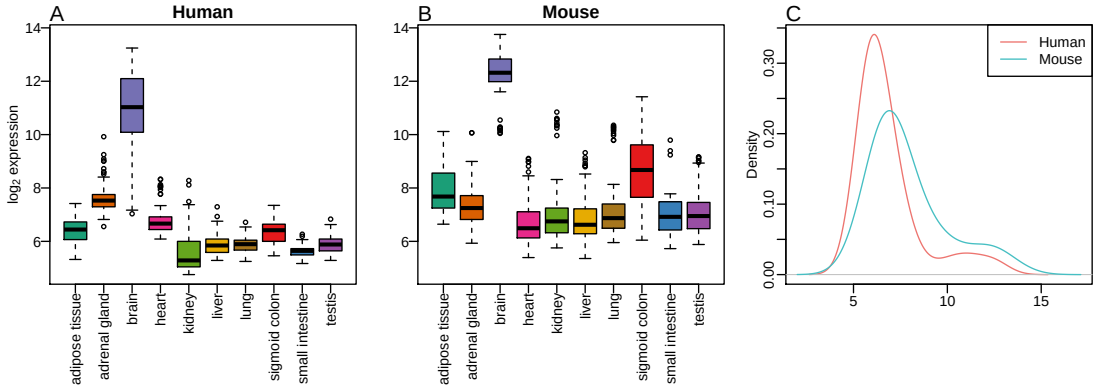


Figure 2.1. Probe Effects in Microarrays. Probe effects can result in clustering by species. Raw gene expression values for neurotransmitter receptor genes for several publicly available human and mouse microarray samples. Box plots for the (A) human gene GRIA2 (glutamate ionotropic receptor AMPA type subunit 2) raw expression values from several human tissues, and for the (B) mouse gene Gria2 (glutamate receptor, ionotropic, AMPA2 (alpha 2)) from several mouse tissues. (C) Density estimator for the distribution of the raw expression values. As expected, these genes appear to be expressed mainly in the brain samples. Note the overall expression levels for the rest of the tissues are higher in the mouse samples. This may be because the probes used in the mouse microarray have higher affinity levels not due to higher expression levels in mouse tissues.

Microarrays measure gene expression using single stranded DNA molecules, referred to as *probes*, that hybridize to specific DNA sequences that are proportional to the RNA transcripts in the samples being assayed[168]. However, it is well known that the sequence of these probes greatly affects the observed measurement[136,100]. Different types of microarrays are used to measure gene expression from different species, and these use different probes to measure homologous genes. As a result, expression levels that are the same in mouse and human samples can result in substantially different measurements due only to the difference in probe sequence (Figure 2.1). McCall et al. demonstrated that if one accounts for probe effects, the distance between different tissues from the same species was much larger than to their corresponding tissue across species[168].

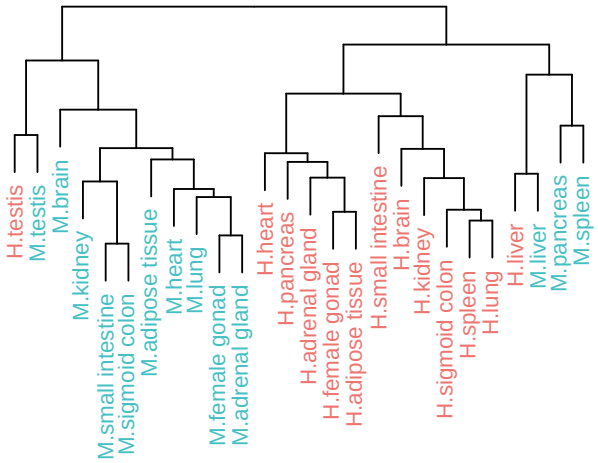


Figure 2.2. Hierarchical Clustering of ENCODE Data Set. Dendrogram for the ENCODE data set based on euclidean distance and complete linkage. The color corresponds to species (Human samples, Mouse samples).

The ENCODE consortium collected RNA-seq data from 13 tissues from human and mouse. Unlike the previous studies with microarrays, the ENCODE consortium used the same platform, RNA-seq, to measure the gene expression for both species. The ENCODE consortium concluded that different tissues within the same species are more similar in gene expression than homologous tissues in different species[144] (Figure 2.2). They acknowledged that this result is somewhat unexpected since previous comparative studies reported that gene expression was more similar between homologous tissues of different species than between different tissues of the same species. The ENCODE consortium proposed that previous studies might have been biased focusing on a few specialized tissues expressing mostly tissue-specific genes, while the overall pattern supports less tissue specificity.

Despite using the same platform to measure gene expression for both species, Gilad et al. showed that the RNA-seq data from the ENCODE consortium were collected using a sequencing design

which confounded the assignment of sequencing flowcells and lanes with species[78] (Table 2.1).

Table 2.1. Sequencing Design for the ENCODE Data Set.

D87PMJN1:253 D2GUAACXX:7	D87PMJN1:253 D2GUAACXX:8	D4LHBFN1:276 C2HKJACXX:4	MONK:312 C2GR3ACXX:6	HWI-ST373:375 C3172ACXX:7
heart	adipose	adipose	heart	brain
kidney	adrenal	adrenal	kidney	pancreas
liver	sigmoid	sigmoid	liver	brain
spleen	lung	lung	small_bowel	spleen
small_bowel	ovary	ovary	testis	
testis		pancreas		

The color corresponds to species (Human samples, Mouse samples). The labels for each sequence batch are ordered as `[machine]:[run]:[flow cell]:[lane]`

Even with the confounded design, Breschi et al. reported that restricted to five tissues (brain, liver, kidney, heart, and testes), the observed clustering is by tissue and not by species[26]. Furthermore, Breschi et al. analyzed matched samples from six homologous organs (brain, liver, kidney, heart, cerebellum, and testis) in seven vertebrate species (chicken, chimpanzee, human, mouse, opossum, platypus, rhesus)[12] using a linear model to quantify the amount of expression variation that originates from variation across tissue and from variation across species. More than 70% of the variance in gene expression can be explained by either organ or species; the contribution of organ (41%) being larger than the contribution of species (31%) consistent with the observed global dominated tissue-dominated clustering. Sudmant et al.[239] conducted a meta-analysis of four datasets, including the re-sequenced ENCODE data set, encompassing 5 homologous tissues from 11 vertebrate species[24,172,144], supplemented by 51 human tissues sequenced by the GTEx consortium[42]. Sudmant et al. concluded that the majority of samples clustered with homologous tissues of different species rather than with

different tissues of the same species. This observation is consistent with the idea that many developmental gene expression programs are conserved across mammals and supports the utility of rodents as models of human tissue physiology. The ENCODE consortium argued that these 5 tissues express a high number of tissue specific genes. Furthermore, the ENCODE consortium reported clustering by tissue if the data set restricted to these 5 tissues[144]. However, the clustering was by species if they included all 13 tissues.

Table 2.2. Sequencing Design for the Resequenced ENCODE Data Set.

D93Z3NS1:226 C75WFACXX:2	D93Z3NS1:226 C75WFACXX:3	D93Z3NS1:226 C75WFACXX:4	HWI-ST689:310 C7764ACXX:6
heart	sigmoid	adipose	brain
kidney	lung	adrenal	liver
small_bowel	spleen	pancreas	testis
heart	sigmoid	adipose	brain
kidney	lung	adrenal	liver
small_bowel	spleen	pancreas	testis

The labels for each sequence batch are ordered as `[machine]:[run]:[flow cell]:[lane]`. The color corresponds to species (Human samples, Mouse samples).

Since the confounding of the sequencing design with species was near perfect in the ENCODE data set, we cannot determine if the observed differences between human and mouse samples is due to differences between species or due to differences between sequencing batches. The ENCODE consortium repeated a smaller version of the experiment described in the original PNAS paper[144]. In the revisited experiment, the new experimental design does not confound the assignment to sequencing flowcells and lanes with species (Table 2.2). Lin et al. argued that their original result was not due to sequencing batch since in the resequenced data the samples separated by species (Fig-

ure 2.3).

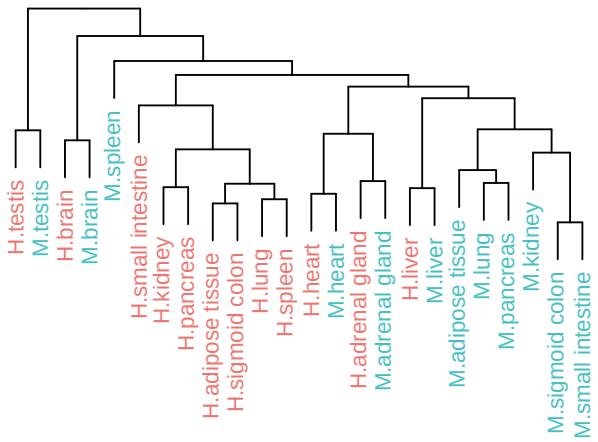


Figure 2.3. Hierarchical Clustering of Resequenced ENCODE Data Set. Dendrogram for the ENCODE data set described by Lin et al based on euclidean distance and complete linkage. The color corresponds to species (Human samples, Mouse samples).

The observed differences in gene expression between human and mouse tissues reported by Lin et al. were driven by differences between the human and mouse annotations. The human and mouse annotations used to quantify gene expression have important differences confounded by species. First, we provide an overview of the differences between the human and mouse annotations. Then we show the impact of the differences between the human and mouse annotations on the observed differences in gene expression. In order to account for the differences between the species annotations we propose using ortholog probes, genomic regions within the human-mouse orthologs with the same length and almost identical sequence, to quantify gene expression. In general, the gene expression estimates based on the orthologs probes are more similar between homologous tissues from different species than between different tissues from the same species. Then, we identify

differentially expressed orthologs between human and mouse tissues using the gene expression estimates from the ortholog probes and corroborating with microarrays. Finally, we show that the observed differences in histone marks, an independent assay, between human and mouse tissues are driven by the difference in number of annotated transcripts per gene.

2.2 Results

2.2.1 Differences Between Human-Mouse Orthologs

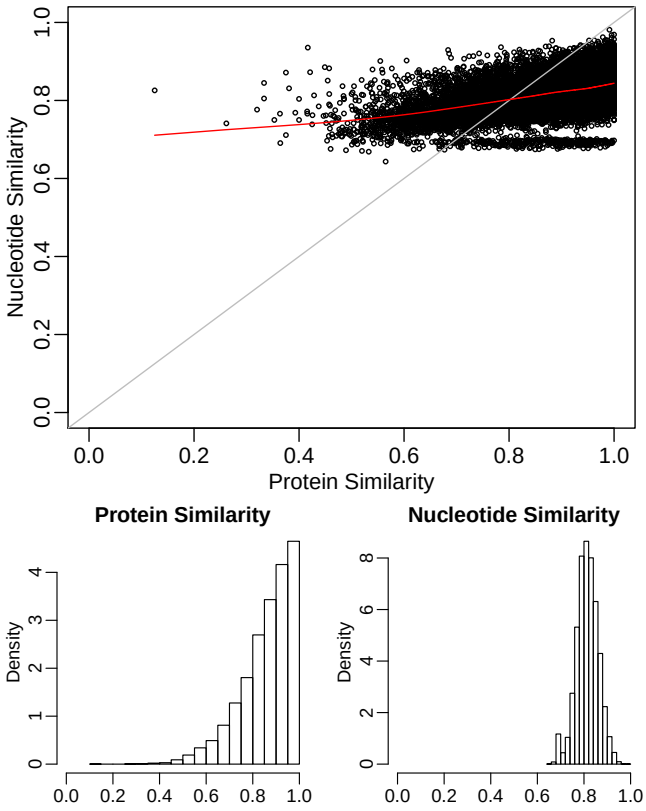


Figure 2.4. Protein vs Nucleotide Similarity. Scatter plot of protein vs. nucleotide similarity between the modENCODE human-mouse orthologs. The grey line is the identity line, and the red line is the LOESS fit regressing the nucleotide similarity on the protein similarity. Histograms of protein and nucleotide similarity between the modENCODE human-mouse orthologs.

The modENCODE phylogenomics approach to identify the human-mouse orthologs relied on the alignment of the longest protein translation for each human and mouse gene pair[261]. On average when it comes to protein-coding genes, mice are 85% similar to humans[157,269]. However, RNA-seq

relies on nucleotide sequences rather than protein sequences to quantify gene expression. We observed differences between the protein similarity and the nucleotide similarity between orthologs (Figure 2.4). The distribution of protein similarity skewed towards higher values (mean =0.8587, median=0.8848) compared to the distribution of nucleotide similarity (mean = 0.8143, median = 0.8144). The observed differences reflect the rapid accumulation of nucleotide changes in synonymous (third base) codon sites followed by a slower mutation rate in non-synonymous sites[157].

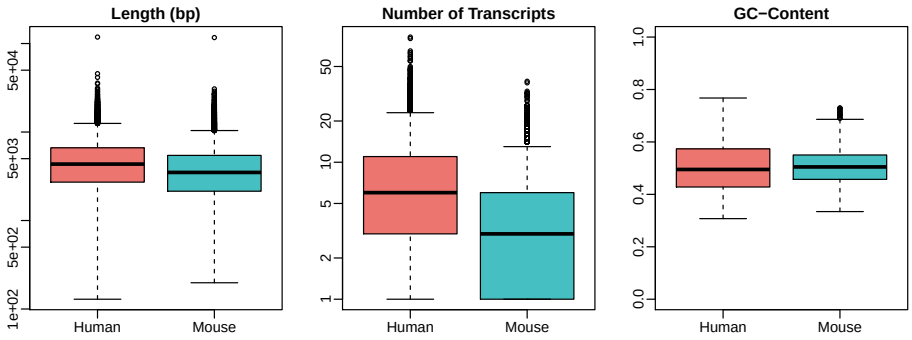
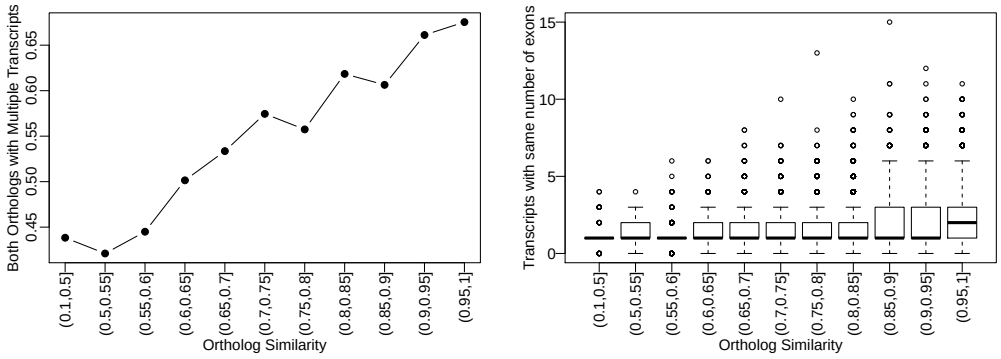


Figure 2.5. Differences between Human and Mouse Orthologs. Boxplots for the length of the union of transcript exons in base pairs, the number of annotated transcripts per gene, and the GC-content of the union of transcript exons for the human (in red) and the mouse (in blue) orthologs.

Despite the high average nucleotide and protein similarity between human and mouse orthologs, these genes have important differences. Human genes on average (5,100 bp) are longer than mouse genes (4,200 bp), have more annotated transcripts per gene on average (8 transcripts) than mouse genes (4 transcripts), and have different GC-content than mouse genes (Figure 2.5). These differences are likely to impact the relationships between the human-mouse orthologs. For instance, conserved proteins tend to be longer than non-conserved ones[145,177].

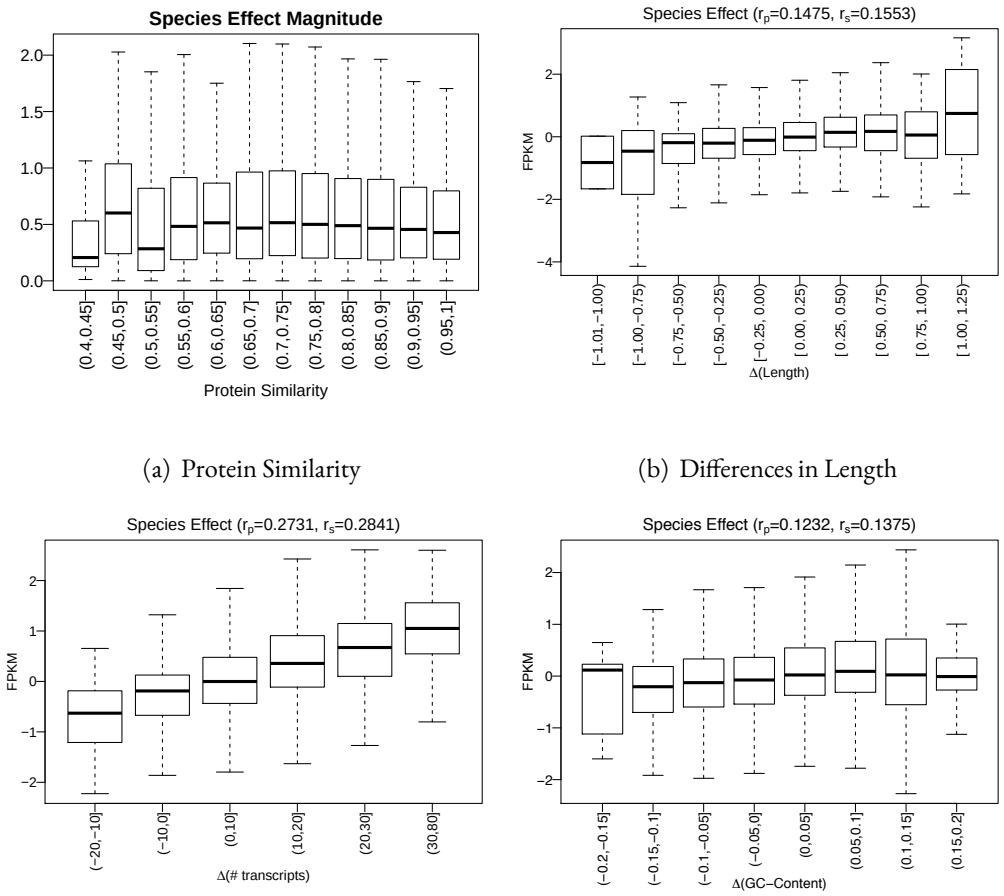


(a) Multiple Transcripts (b) Transcripts with the Same Number of Exons

Figure 2.6. Transcript Annotation vs. Ortholog Similarity. (a) Proportion of human-mouse ortholog pairs where both have more than one annotated transcript binned by ortholog similarity measured as the protein sequence similarity between orthologs. (b) Number of annotated transcripts from human-mouse ortholog pairs with the same number of exons binned by ortholog similarity measured as the protein sequence similarity between orthologs.

The number of annotated transcripts is related with ortholog similarity and the number of annotated transcripts, and with the number of exons as previously described by Morata et al.[177]. As the ortholog similarity increases, the number of ortholog pairs where both genes have more than one annotated transcript increases. Also as the ortholog similarity increases, ortholog pairs have more annotated transcripts with the same number of exons (Figure 2.6).

2.2.2 Species Effect Correlated with Differences Between Orthologs



(a) Protein Similarity

(b) Differences in Length

(c) Differences in Number of Transcripts

(d) Differences in GC-Content

Figure 2.7. **Species Effect vs. Differences between Orthologs.** The species effect is the observed difference in means between the human and mouse normalized FPKM values from the resequenced ENCODE RNA-seq data set. (a) Absolute value of the species effect binned by the protein similarity between the human-mouse orthologs. In the following boxplots, r_p denotes the Pearson correlation coefficient and r_s denotes the Spearman correlation coefficient. Species effect binned by (b) the differences in the log_{10} lengths of the union of transcript exons, (c) the difference in number of annotated transcripts, (d) the difference in GC-content of the union of transcript exons between the human and mouse orthologs.

The species effect, the observed differences in expression between the human and mouse tissues, is correlated with the differences between the human-mouse orthologs. We used the FPKM values processed by Lin et al. for the resequenced ENCODE RNA-seq data set to estimate the species effect. The magnitude of the species effect decreases as the similarity between the orthologs increases (Figure 2.7). Moreover, the species effect is positively correlated with differences in length, number of transcripts, and GC-content between the human-mouse orthologs (Figure 2.7).

2.2.3 Accounting for Differences Between Orthologs Reduces Systematic Biases

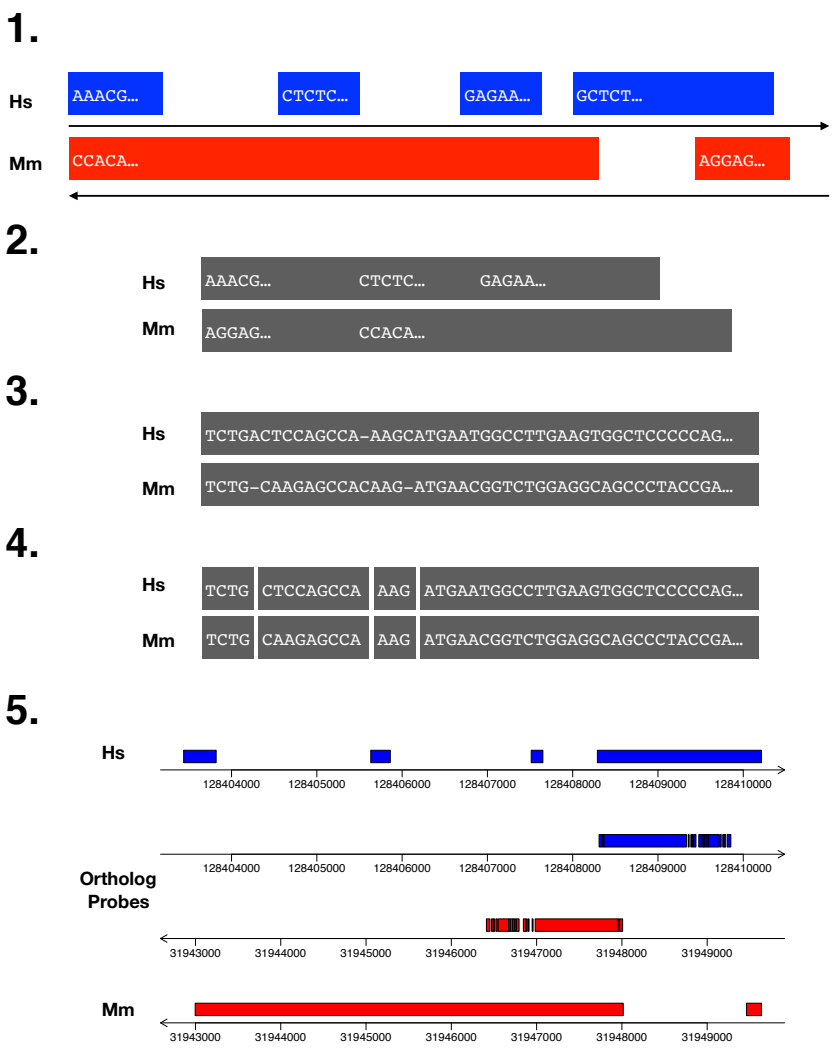


Figure 2.8. Ortholog Probes. Procedure to align the human-mouse orthologs to find the ortholog probes. (1) Get the sequences for the union of exons from the human (Hs) and mouse (Mm) genes. The color of corresponds to the strand orientation; blue for positive (+), and red for negative (-). (2) Concatenate the human (Hs) and mouse (Mm) sequences after arranging them in the same orientation. (3) Align the concatenated sequences using the Smith-Waterman local alignment algorithm. (4) Split the alignments to remove all gaps (insertions/deletions) in the aligned sequences. (5) Map the alignment segments to their corresponding genomic positions. The mapped alignment segments are the ortholog probes. The ortholog probes are genomic regions with the same length and similar sequences for both species.

As a consequence of the inherent differences between the human and mouse annotations, the ENCODE consortium used genomic regions with different characteristics between the two species to quantify the gene expression for each pair of human-mouse orthologs. In this manner, the gene expression estimates are confounded by the differences between the human and mouse annotations. In order to account for the differences between species annotations, we revisit the idea of using probes for gene expression to account for the differences between the human-mouse orthologs. In microarrays, a probe is a small fragment of DNA sequence (25 bp in Affymetrix microarrays) used to quantify gene expression. We use genomic regions within the union of transcript exons from the human-mouse orthologs with the same length and almost identical sequence as probes to quantify the gene expression (Figure 2.8). We use these probes to account for the differences between the human and mouse annotations in estimating the gene expression for the resequenced ENCODE RNA-seq data set.

For instance, in the human-mouse ortholog pair LARS2/Lars2 (leucyl-tRNA synthetase 2) the human gene is longer (5,393 bp) than the mouse gene (3,881 bp) (Figure 2.9A-B). Moreover, the human gene has 8 annotated transcripts while the mouse gene has only 1 annotated transcript. Despite the differences between the human and mouse annotations, both genes have the same function. The human LARS2 gene encodes for the mitochondrial leucyl-tRNA synthetase enzyme. This enzyme plays an important role in the synthesis of proteins in the mitochondria[116,203,225]. The LARS2 gene is conserved in the mouse as well as in chimpanzee, Rhesus monkey, dog, rat, chicken, zebrafish, fruit fly, mosquito, C.elegans, S.cerevisiae, K.lactis, E.gossypii, S.pombe, M.oryzae, N.crassa, A.thaliana[120,119,270,85,23,7,183,20]. The mouse Lars2 gene has the same function as its corresponding hu-

man LARS2 ortholog[176,114].

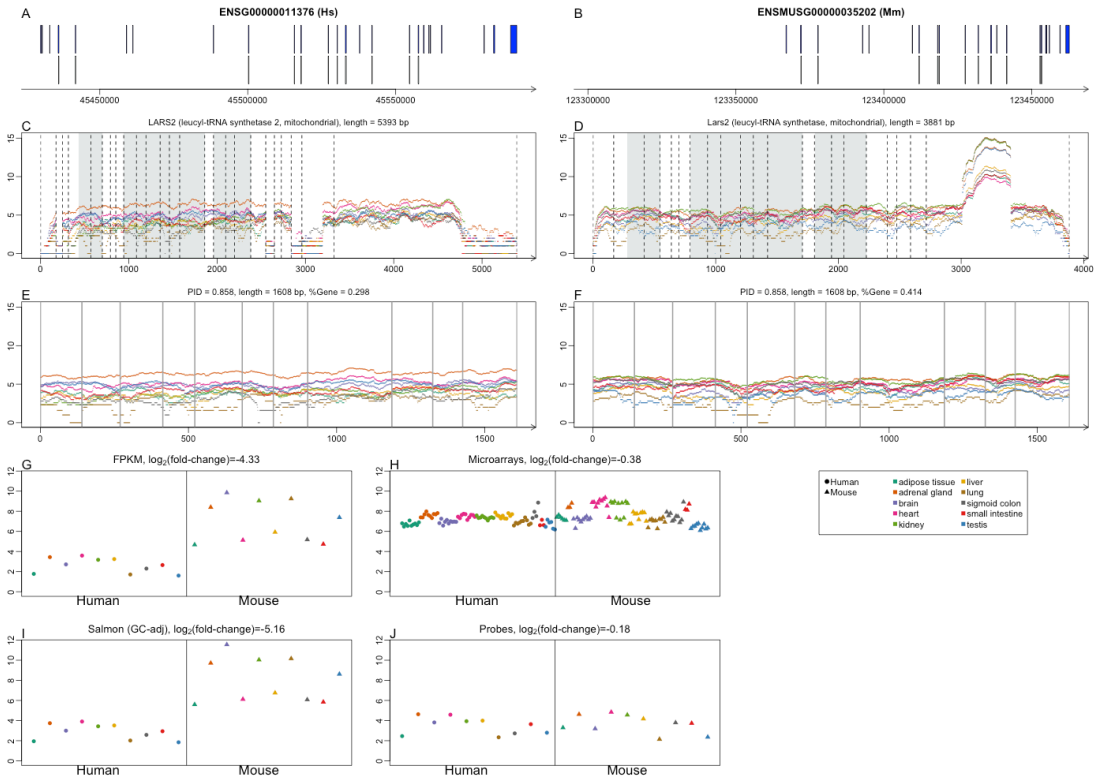


Figure 2.9. Ortholog Probe for LARS2/Lars2. Diagram for the union of transcript exons of the human (A) and mouse (B) genes. The regions in gray below the exons correspond to the ortholog probe. Coverage plots for the union of transcript exons excluding the intronic regions for the human (C) and mouse (D) genes. The dashed lines are the exon boundaries and the shaded areas correspond to the ortholog probes. Coverage lots for the human (E) and mouse (F) ortholog probes. Scatter plots for the normalized FPKM values (G), the normalized Salmon estimates corrected for GC-content bias (I), and the normalized probe values (J) for the resequenced ENCODE RNA-seq data set. Scatter plot for the normalized Barcode 3.0 microarray values.

Based on the normalized FPKM values, the expression for the human-mouse ortholog LARS2/Lars2 is higher in the mouse tissues than in the human tissues. However, with the normalized probe values the difference in expression between the human and mouse tissues is lower (Figure 2.9). Furthermore, the species effect estimate from the normalized probe values (-0.18) is closer to the species effect effect estimates from a collection of curated microarray samples[168] (-0.38).

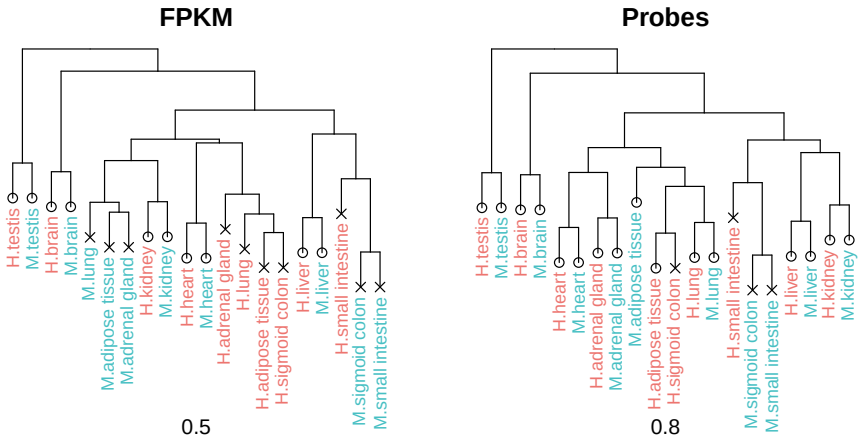


Figure 2.10. Hierarchical Clustering of Resequenced ENCODE Data Set. Dendrogram for the resequenced ENCODE RNA-seq data set based on hierarchical clustering with complete linkage and Euclidean distance. The dendrogram on the right (FPKM) is based on the normalized FPKM values, and the dendrogram on the right (Probes) is based on the normalized ortholog probe counts. The circles (○) correspond to pairs of human and mouse tissues where the type match, while the crosses (×) correspond to pairs of human and mouse tissues where the type does not match. The numbers at the bottom of each dendrogram are the proportion of human and mouse tissue pairs where the tissue type matches.

We compared the clustering of the resequenced ENCODE RNA-seq data set using the normalized FPKM values with the normalized probe values (Figure 2.10). The clustering with the normalized FPKM values resembles the results reported by the ENCODE consortium; the samples cluster by species rather than by tissue. However, in the clustering with the normalized probe values accounting for the annotation differences, the samples pair by tissue type rather than by species.

We also considered the influence of tissue type in the clustering results by leaving out one tissue type and repeating the clustering (Figure B.11). In the resulting clusterings we observed the same pattern; in the clustering based on the FPKM values the samples cluster by mostly by species rather than by tissue and in the clustering based on the probe values the samples cluster by tissue rather than by species.

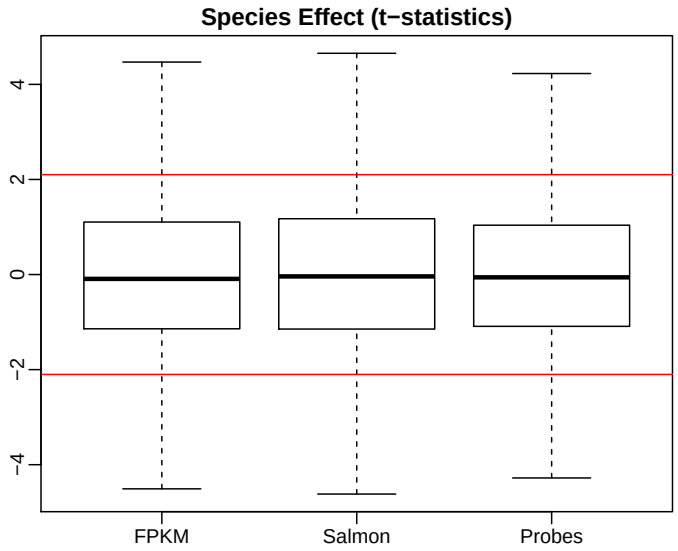
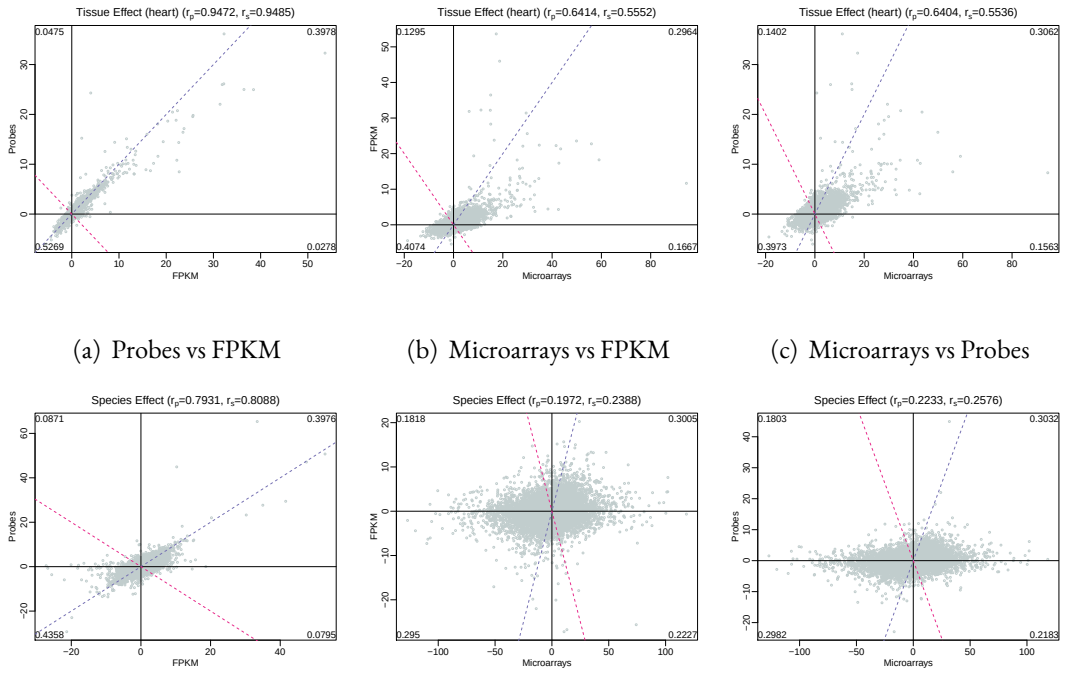


Figure 2.11. Species and Effect T-statistics. Boxplots of the t-statistics comparing the observed difference in means between the human and mouse tissues in the resequenced ENCODE RNA-seq data set using the normalized FPKM values, the normalized Salmon estimated adjusted for GC-content bias and the normalized probe values. The red lines correspond to the significance threshold at a 0.05 level.

We computed the t-statistics for the species effect using the normalized FPKM values, expression estimates adjusted for GC-content bias using Salmon[198], and the normalized probe values to adjust the for the differences in variances between the different tissue and species effect estimates. Adjusting for systematic biases such as GC-content reduces the observed differences between species as expected from the hierarchical clustering results (Figure 2.11).

2.2.4 Differences Between Tissues Replicated in Microarrays, Differences Between Species Not Replicated



(a) Probes vs FPKM (b) Microarrays vs FPKM (c) Microarrays vs Probes

(d) Probes vs FPKM (e) Microarrays vs FPKM (f) Microarrays vs Probes

Figure 2.12. **Species and Tissue Effect T-statistics Comparison.** The species t-statistics are the t-statistics comparing the observed difference in means between the human and mouse normalized expression values from the resequenced ENCODE RNA-seq data set. In this case, the tissue t-statistics for heart are the t-statistics comparing the observed difference in means between the human and mouse heart samples and the rest of the human and mouse tissues. Comparison between the tissue t-statistics from the (a) FPKM and the probe normalized values, the (b) FPKM and the microarray normalized values, and the (c) probe and microarray normalized values. Comparison between the species t-statistics from the (d) FPKM and the probe normalized values, the (e) FPKM and the microarray normalized values, and the (f) probe and microarray normalized values. The purple dashed line is $x = y$, and the pink dashed line is $x = -y$. The number in each corner indicates the proportion of points in each quadrant defined by $x = 0$ and $y = 0$.

For each tissue we computed the difference in expression for each gene between that given tissue and all others. Note that this difference will be large for genes that are uniquely expressed in that

tissue. If these differences are due to real biological signal, we should see a similar pattern using an assay other than RNA-seq. We therefore estimated these differences using a microarray data set (Table B.3). The difference between tissues with both platforms correlate strongly. In this section, we use as an example the tissue effect for the heart samples (Figure 2.12). The difference in expression for the other tissues also correlate strongly in both microarrays and RNA-seq (Figure B.13). On the other hand, the correlation between the species effect in both platforms is lower than the tissue effect (Figure 2.12).

2.2.5 Differentially Expressed Orthologs

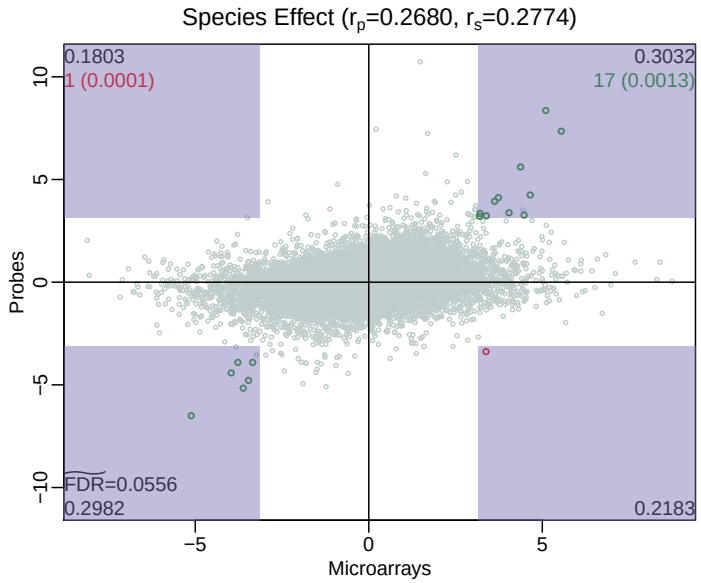


Figure 2.13. Differentially Expressed Orthologs. The species effect is the observed difference in means between the human and mouse normalized expression values. Species effect estimates based on the normalized microarray values (d_g^{mcr}) and the probe normalized values from the resequenced ENCODE RNA-seq data set (d_g^{probes}). The shaded region is defined by the threshold on the species effect estimates from microarrays ($|d_g^{\text{mcr}}| > C$) and from the probes ($|d_g^{\text{probes}}| > C$). The points in green and red are orthologs with $q < 0.05$ and species effect estimates above the threshold C in both platforms. The points in green correspond to the differentially expressed orthologs. On the other hand, we consider the points in red as false positives since the sign of the species effect estimates depends on the platform. The number of differentially expressed orthologs and false positives are colored accordingly, with the proportion in parenthesis. The number in each corner is the proportion of points in each quadrant defined by $d_g^{\text{mcr}} = 0$ and $d_g^{\text{probes}} = 0$.

We used a collection of curated microarray samples from normal human and mouse tissues[168] to estimate the species effect and to identify the differentially expressed orthologs between the human and mouse tissues. Since the platform effect from the microarray samples is independent from RNA-seq, we used them to corroborate the differentially expressed orthologs. Thus, we have three sets of species effect estimates from two different platforms: from RNA-seq the estimates from the normal-

ized FPKM values and from the normalized probe values, and the estimates from the normalized microarray values.

Table 2.3. Differentially Expressed Orthologs in both the Resequenced ENCODE RNA-seq Data Set and the Barcode 3.0 Microarrays for Different FDR Cut-Offs.

Cut-Off	$\|\|\mathcal{A}\|\|$	$\|\|B\|\|$	Differentially Expressed	$\widetilde{\text{FDR}}$
0.9400	940	187	753	0.1989
2.4500	60	6	54	0.1000
3.1500	18	1	17	0.0556
3.3800	12	0	12	< 0.01

$\mathcal{A}$ is the set of orthologs with $q < 0.05$ and $|d_g| > C$ in both the normalized probe values from the resequenced ENCODE RNA-seq data set and the Barcode 3.0 normalized microarray values, B is the set of orthologs where the sign of the species effect d_g is different between the platforms, and $\widetilde{\text{FDR}}$ is the false discovery rate estimate based on $\mathcal{A}$ and B.

We use the normalized probe values from the resequenced RNA-seq data set and the normalized microarray values to identify the differentially expressed orthologs in the human and mouse tissues. We select the human-mouse orthologs where the species effect was significant and had the same sign in both platforms (Figure 2.13).

Table 2.4. Differentially Expressed Orthologs.

Ensembl (Hs:Mm)	Symbol (Hs)	Name (Hs)	Symbol (Mm)	Name (Mm)
ENSG00000067560:ENSMUSG00000036463	RHOA	ras homolog family member A	4930544G11Rik	RIKEN cDNA 4930544G11 gene
ENSG00000129538:ENSMUSG00000035896	RNASE1	ribonuclease A family member 1, pancreatic	Rnase1	ribonuclease, RNase A family, 1 (pancreatic)
ENSG00000136943:ENSMUSG00000021477	CTSV	cathepsin V	Ctsl	cathepsin L
ENSG00000100814:ENSMUSG00000071470	CCNB1IP1	cyclin B1 interacting protein 1	Ccnb1ip1	cyclin B1 interacting protein 1
ENSG00000051620:ENSMUSG00000019853	HEBP2	heme binding protein 2	Hebp2	heme binding protein 2
ENSG00000115361:ENSMUSG00000026003	ACADL	acyl-CoA dehydrogenase long chain	Acadl	acyl-Coenzyme A dehydrogenase, long-chain
ENSG00000197446:ENSMUSG00000052974	CYP2F1	cytochrome P450 family 2 subfamily F member 1	Cyp2f2	cytochrome P450, family 2, subfamily f, polypeptide 2
ENSG00000139610:ENSMUSG00000023031	CELA1	chymotrypsin like elastase family member 1	Cela1	chymotrypsin-like elastase family, member 1
ENSG00000164039:ENSMUSG00000028167	BDH2	3-hydroxybutyrate dehydrogenase 2	Bdh2	3-hydroxybutyrate dehydrogenase, type 2
ENSG00000118418:ENSMUSG00000066456	HMGN3	high mobility group nucleosomal binding domain 3	Hmgn3	high mobility group nucleosomal binding domain 3
ENSG00000165568:ENSMUSG00000045410	AKR1E2	aldo-keto reductase family 1 member E2	Akr1e1	aldo-keto reductase family 1, member E1
ENSG00000176903:ENSMUSG00000054383	PNMA1	PNMA family member 1	Pnma1	paraneoplastic antigen MA1
ENSG00000196419:ENSMUSG00000022471	XRCC6	X-ray repair cross complementing 6	Xrcc6	X-ray repair complementing defective repair in Chinese hamster cells 6
ENSG00000147536:ENSMUSG00000031546	GINS4	GINS complex subunit 4	Gins4	GINS complex subunit 4 (Sld5 homolog)
ENSG00000074935:ENSMUSG00000019845	TUBE1	tubulin epsilon 1	Tube1	epsilon-tubulin 1
ENSG00000142856:ENSMUSG00000028549	ITGB3BP	integrin subunit beta 3 binding protein	Itgb3bp	integrin beta 3 binding protein (beta3-endonexin)
ENSG00000196465:ENSMUSG00000039824	MYL6B	myosin light chain 6B	Myl6b	myosin, light polypeptide 6B

Differentially expresed orthologs based in both the normalized probe values from the resequenced ENCODE RNA-seq data set and the Barcode 3.0 normalized microarray values at an $\widetilde{\text{FDR}} \approx 0.05$.

We indentify 17 differentially expressed orthologs using the normalized probe values and corroborated by microarrays at an FRD of 0.05 (Table 2.4). Our analysis results in less differentially

expressed orthologs between species less than originally reported by the ENCODE consortium and identified using the normalized FPKM values. They originally reported 4,767 differentially expressed human-mouse orthologs. Furthermore, we tested the differences in means between the human and mouse normalized FPKM values with LIMMA [215] adjusting for multiple comparison with `qvalue` [260] and identified 2,719 differentially expressed human-mouse orthologs ($q < 0.05$).

Table 2.5. Enriched GO Slim Terms in Differentially Expressed Orthologs.

		FDR ≈ 0.05				
GO ID	Term	Annotated	Significant	Expected	Odds Ratio	p-value
GO:0019748	secondary metabolic process	39	1	0.0500	24.7549	0.0484
GO:0051276	chromosome organization	977	5	1.2400	6.0515	0.0061
		FDR < 0.01				
GO ID	Term	Annotated	Significant	Expected	Odds Ratio	p-value
GO:0019748	secondary metabolic process	39	1	0.0300	36.0191	0.0344
GO:0048646	anatomical structure formation involved in morphogenesis	912	3	0.8200	5.2017	0.0437
GO:0002376	immune system process	2158	5	1.9400	4.2933	0.0319

GO Slim terms significantly enriched in the differentially expressed orthologs at different FDR cut-offs. The enrichment test was conducted using Fisher's exact test.

We examined the differentially expressed orthologs at an FDR of 0.05 and 0.01 for enrichments in Gene Ontology (GO) biological process terms [8] to determine the type of genes that are differentially expressed between species (Table 2.5).

The ENCODE consortium reported more than 50 enriched GO terms related to basic cellular

function[144]. On the other hand, we identify 2 enriched GO terms at an FDR of 0.05 and 3 enriched GO terms at an FDR of 0.01. These GO terms are related to well known differences between humans and mice. In both sets of differentially expressed orthologs, the top enriched term was *secondary metabolic process* (GO:0019748). As mice are significantly smaller than humans, their basal metabolic rate, the rate of energy production over a set period of time under constant conditions, is much less than that of humans, because mice simply have less body mass and less total energy production. However, the basal metabolic rate per gram of body weight is seven times greater in mice than in humans[50].

The other term enriched in the differentially expressed orthologs at an FDR of 0.05 was *chromosome organization* (GO:0051276). Mice have 20 pairs of chromosomes while humans have 23 pairs, but the mouse haploid genome is about 3 picograms similar to the human haploid genome[217]. The gene order (synteny) of the mouse and human genomes is conserved although there are several rearrangements per chromosome[194]. However, mouse chromosomes have undergone an unusually high number of genomic rearrangements per unit of evolutionary time[82].

With a more stringent FDR cut-off, the terms *anatomical structure formation involved in morphogenesis* (GO:0048646) and *immune system process* (GO:0002376) are significantly enriched. Although mice share genes, organ systems and systemic physiology with humans, the two species differ significantly in terms of morphometry, physiology and life history. Humans are about 3,000 times larger than mice, and this size difference imposes constraints on physiology and life history with significant effects on the species' ability to adapt to environmental conditions[50,75]. One of the most obvious differences between mouse and human morphogenesis is the time of birth. The mouse em-

bryo is born almost immediately after all the organs develop (19–20 days post conception). On the other hand at the end of organogenesis, the human embryo has a disproportionally large head relative to the whole body and other organs. The embryo continues to stay in the uterus for a few more months, a period termed as the fetal stage. During this stage, many organs continue to grow and eventually develop into their proper sizes for birth [263].

The overall structure of the immune system in mice and humans is quite similar [87]. Despite this conservation there exist significant differences between mice and humans in immune system development, activation, and response to challenge. Such differences should not be surprising as the two species diverged somewhere between 65 and 75 million years ago, differ hugely in both size and lifespan, and have evolved in quite different ecological niches where widely different pathogenic challenges need to be met [173]. These are not trivial differences. For instance, leukocyte transit times may be quite different in mice and humans, and a larger, broader repertoire of B and T cells must be maintained for many years in humans (up to 50 mouse lifetimes) [173]. Neutrophils are a rich source of leukocyte defensins in humans, but defensins are not expressed by neutrophils in mice (14). In contrast, Paneth cells, which are present in the crypts of the small intestine, express more than 20 defensins (cryptdins) in mice but only two in human, likely reflecting different evolutionary pressures related to microorganism exposure through food intake [214].

2.2.6 Histone Peaks Difference Correlated with Difference in Annotated Transcripts

The ENCODE consortium used ChIP-seq as an independent assay to confirm the differences in expression between human and mouse tissues. The observed differences between the ChIP-seq are also affected by differences between the human and mouse annotations. We used histone modifications ChIP-seq data from the human REMC and the mouse ENCODE projects as an independent assay to examine the influence of the differences between the human and mouse annotation on the observed differences between human and mouse tissues (Table B.3). Following Lin et al.[144], we used H3K4me3 and H3K27ac histone modification levels in the 1 kb flanking regions of the transcription start sites. The H3K4me3 (trimethylation of kysine 4 on the histone H3 protein subunit) modification is commonly associated with active transcription[84]. The H3K27ac (acethylation at the 27th lysene residue of the histone H3 protein subunit) modification is associated with higher activation of transcription[43]. Hence, the H3K4me3 and the H3K27ac modifications are active marks associated with gene expression.

The ENCODE consortium tested the differences in histone peak intensities for the set of orthologs reported as differentially expressed using a Wilcoxon test. They split the differentially expressed orthologs into two separate sets: where the gene expression is higher in the human tissues than in the mouse tissues (Hs $>$ Mm) and where the gene expression is higher in the mouse tissues than in the human tissues (Hs $<$ Mm). The differences between the gene-associated peak intensities for the differentially expressed orthologs in the original ENCODE RNA-seq data set were signifi-

cant for both H3K4me3 and H3K27ac across all human-mouse tissue pairs (Table B.5, Figure B.21).

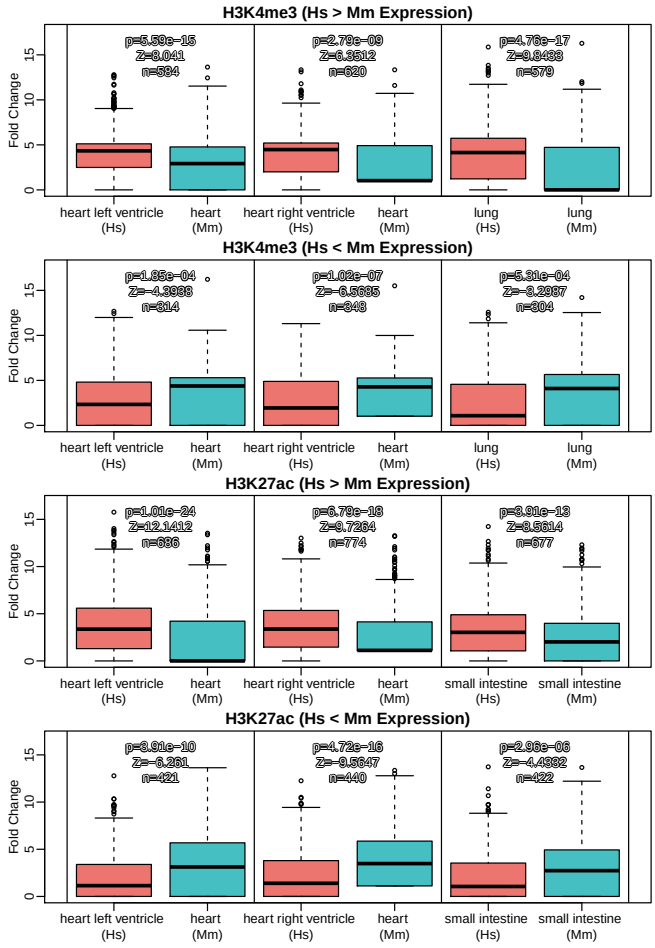


Figure 2.14. Histone Peak Intensities for Differentially Expressed Orthologs in the Resequenced ENCODE RNA-seq Data Set. Fold Enrichment over control of H3K4me3 and H3K27ac present at promoters of the differentially expressed orthologs based on the normalized FPKM values from the resequenced ENCODE RNA-seq data set. The differentially expressed orthologs are separated into orthologs where the gene expression is higher in the human tissues than in the mouse tissues (Hs $>$ Mm), and where the gene expression is higher in the mouse tissues than in the human tissues (Hs $<$ Mm). The p-values and Z-statistics were generated by the nonparametric paired Wilcoxon test between the human and mouse gene-associated histone peak intensities, n is the number of human-mouse orthologs where at least one of them has a gene-associated peak intensity.

In our analysis, we excluded the ChIP-seq spleen samples to match our analysis of the resequenced ENCODE RNA-seq data set. The differences between the histone peaks for the differentially ex-

pressed orthologs based on the normalized FPKM values from the resequenced ENCODE RNA-seq data set were also significant for both H3K4me3 and H3K27ac across all human-mouse tissue pairs (Table B.6, Figure 2.14).

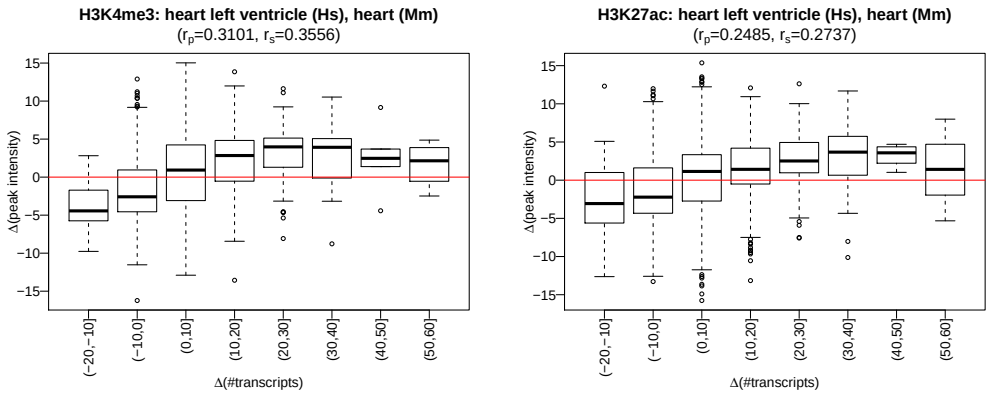


Figure 2.15. Differences between Histone Peak Intensities vs. Differences between the Number of Transcripts. Differences between the gene-associated histone peak intensities and the number of annotated transcripts for the human-mouse orthologs in the human (Hs) heart (left ventricle) and the mouse (Mm) heart samples. r_p denotes the Pearson correlation and r_s denotes the Spearman correlation.

We found that the observed differences between the gene-associated histone peak intensities for the active promoter marks H3K4me3 and H3K27ac were correlated with the differences between the number of annotated transcripts per gene (Figure 2.15). The number of transcription start sites depends on the number of annotated transcripts per gene. Since the human genes and the mouse genes from each ortholog pair do not always have the same number of annotated transcripts, the number of transcription start sites is confounded with species. However, if we only consider the human-mouse orthologs with the same number of annotated transcripts to account for the differences between the human and mouse annotations, the differences between the histone peaks for the differentially expressed orthologs reported by Lin et al. are not significant for neither H3K4me3 and

H3K27ac (Table B.8, Figure B.23).

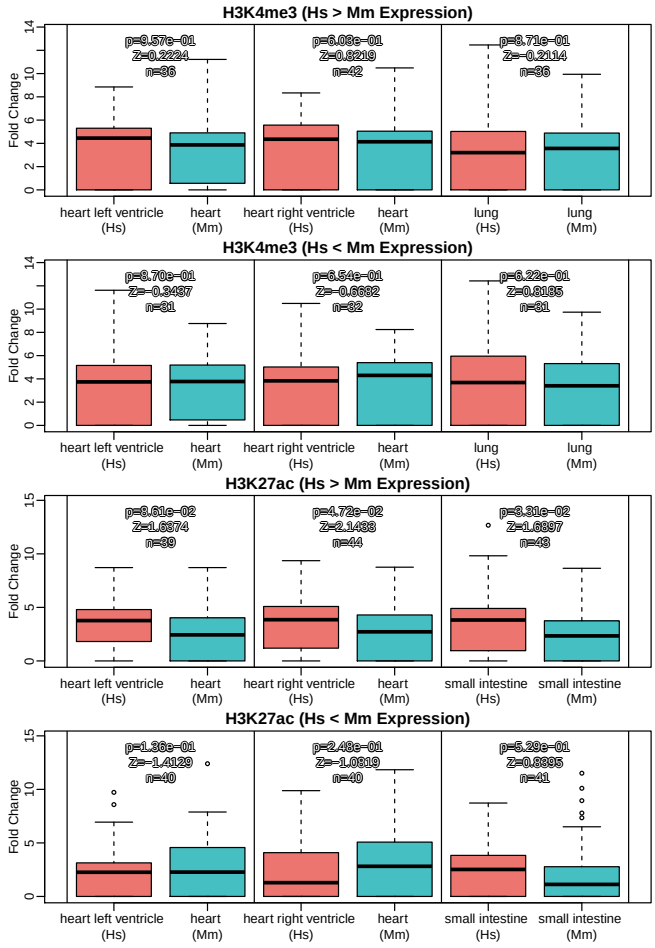


Figure 2.16. Histone Peak Intensities for Differentially Expressed Orthologs in the Resequenced ENCODE RNA-seq Data Set with the Same Number of Annotated Transcripts Fold Enrichment over control of H3K4me3 and H3K27ac present at promoters of the differentially expressed orthologs with the same number of transcripts based on the normalized FPKM values from the resequenced ENCODE RNA-seq data set. The differentially expressed orthologs are separated into orthologs where the gene expression is higher in the human tissues than in the mouse tissues (Hs $>$ Mm), and where the gene expression is higher in the mouse tissues than in the human tissues (Hs $<$ Mm). The p-values and Z-statistics were generated by the nonparametric paired Wilcoxon test between the human and mouse gene-associated histone peak intensities, n is the number of human-mouse orthologs where at least one of them has a gene-associated peak intensity.

For the differentially expressed orthologs based on the FPKM values from the resequenced ENCODE RNA-seq data set, only in 2 of the 12 human-mouse tissue pairs the difference between the histone peaks is significant considering only human-mouse orthologs with the same number of annotated transcripts (Table B.9, Figure 2.16).

We could not determine whether the differences in histone peaks for the differentially expressed orthologs based on both the normalized probe values and the normalized microarray values where the human-mouse orthologs have the same number of transcripts was significant due to the low number of histone peaks present. Even with a higher FDR cut-off for the differentially expressed orthologs ($\widetilde{\text{FDR}} \approx 0.20$), only a small number of orthologs had gene-associated peak intensities (Table B.10, Figure 2.17).

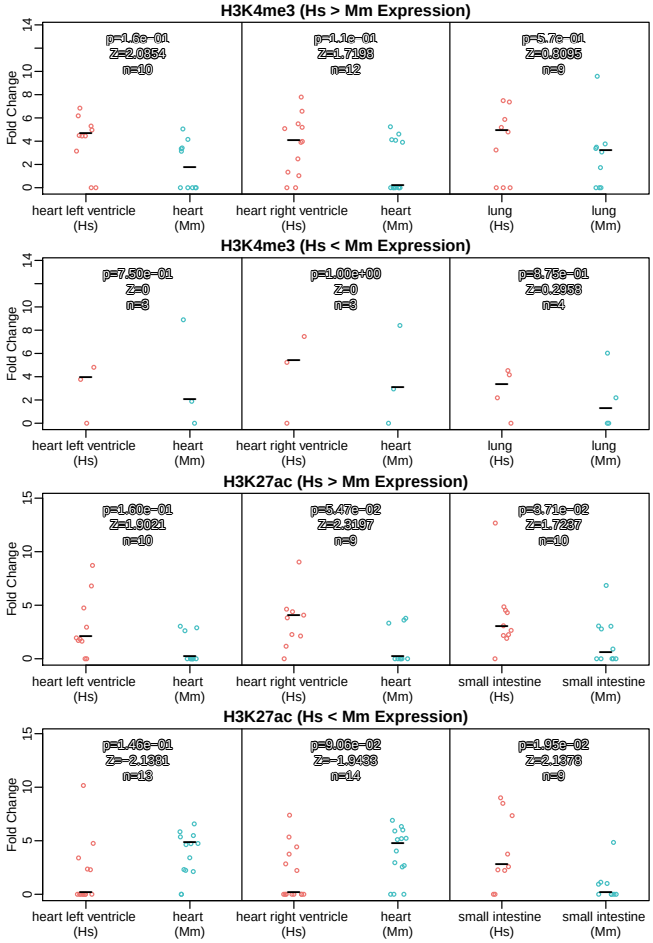


Figure 2.17. Histone Peak Intensities for Differentially Expressed Orthologs in both the Resequenced ENCODE RNA-seq Data Set and the Microarrays from Barcode 3.0 with the Same Number of Annotated Transcripts. Fold enrichment over control of H3K4me3 and H3K27ac present at promoters of the differentially expressed orthologs with the same number of transcripts based on both the normalized probe values from the resequenced ENCODE RNA-seq data set and the normalized microarray values from Barcode 3.0. The black lines are the median peak intensities. The differentially expressed orthologs are separated into orthologs where the gene expression is higher on average in the human tissues than in the mouse tissues (Hs $>$ Mm), and where the gene expression is higher on average in the mouse tissues than in the human tissues (Hs $<$ Mm). The p-values and Z-statistics were generated by the nonparametric paired Wilcoxon test between the human and mouse gene-associated histone peak intensities, n is the number of human-mouse orthologs where at least one of them has a gene-associated peak intensity.

2.3 Discussion

The modENCODE consortium relied on protein alignments to identify human-mouse orthologs. However, gene expression quantification relies on nucleotide sequences rather than protein sequences to estimate gene expression. The similarity between the orthologs based on protein sequences (mean=0.8587,median=0.8848) is skewed towards higher values than the similarity based on nucleotide sequences (mean=0.8143, median=0.8144). Moreover, the gene annotations are different between species. For instance, on average human genes are longer (5,100 bp) than mouse genes (4,200 bp), and have more transcripts annotated per gene (8 transcripts) then mouse genes (4 transcripts). As a consequence the genomic regions used to quantify gene expression are inherently different between species. Hence, the gene expression estimates are confounded by differences between the species annotations.

The observed differences between human and mouse tissues in gene expression are driven in part by differences between the human and mouse annotations. The species effect is positively correlated with differences in length, number of annotated transcripts per gene, and GC-content between the human-mouse orthologs. We found that the differences in expression between tissues in the resequenced ENCODE RNA-seq data set correlate strongly with the estimates from our microarray data set. This result suggests that the differences between tissues are real biological signal rather than technical artifacts. On the other hand, the correlation between the species effect in the resequenced ENCODE RNA-seq data set and in the microarray data set was lower than the tissue effect.

We proposed using ortholog probes, genomic regions within the union of transcript exons from

the human-mouse orthologs with the same length and similar sequence, to account for the differences between the species annotations in gene expression quantification. We clustered the resequenced ENCODE data set using the gene expression estimates from our ortholog probes. In our clustering results, the samples paired by homologous tissues rather than by species in agreement with previous studies with microarrays[167,140,38] and RNA-seq[24,172,12]. Our results support the assumption that the mouse is a useful model of the physiology of specific human organs despite millions of years of evolutionary divergence and the existence of evolutionarily conserved transcriptional programs driving the production of major proteins that define specific tissues[26,239].

We identified orthologs differentially expressed between the human and mouse tissues using the gene expression estimates from the ortholog probes and corroborating with microarrays. In our analysis, we identified significantly less differentially expressed orthologs (17) than originally reported by Lin et al. (4,767), and less than we identified using the normalized FPKM values from the resequenced ENCODE data set (2,719). We used microarrays because the platform effect of microarrays is independent from RNA-seq, and the sample size of the resequenced RNA-seq data set is small. In the resequenced ENCODE data set, each species there is only one sample per tissue.

Lin et al. used histone marks as an independent assay as supporting evidence for differentially expressed orthologs between species. We showed that the differences in histone marks between human and mouse tissues were biased by the difference in number of annotated transcripts per gene between the human and mouse annotations. The observed differences between the gene-associated histone peak intensities for the active promoter marks H3K4me3 and H3K27ac are correlated with the differences between the number of annotated transcripts per gene. The assignment of gene-

associated histone peak intensities is based on transcription start sites which depends on the number of annotated transcripts per gene. The human and mouse genes from each ortholog pair do not always have the same number of annotated transcripts.

The ortholog probes relied on alignments between the nucleotide sequences of the human-mouse orthologs. In some instances, poor alignments result in a ortholog probes which inflate the observed differences between human and mouse tissues. For instance, the ortholog probe for BLOC1S6/Bloc1S6 is very short (110 bp) compared to the human gene (6,812 bp) and the mouse gene (3,899 bp). The ortholog probe increases the species effect (4.77) compared to the species effect based on FPKM (-0.27) and microarrays (-0.90) because the region covered by the mouse ortholog probes includes very few reads compared to the human ortholog probe (Figure B.10). Moreover the ortholog probes rely on the accuracy and completeness of the ortholog annotation, the gene annotation and the genome build.

Similar to union exon based approaches for RNA-seq gene expression quantification, the ortholog probes ignore the complexity of genes with multiple transcript isoforms[277]. We noticed that as the ortholog similarity increases, the number of ortholog pairs where both genes have more than one annotated transcript increase and have more annotated transcripts with the same number of exons. Another related major issue is that functional and evolutionary analyses of genes are usually performed on one or few representatives of their expression products, i.e. transcripts and proteins.[271]. Alternative splicing give rise to multiple transcript isoforms from the same gene. Alternative splicing affects over 90% of genes in humans and mice[186,195,36,196,254], and accounts for the increase of at least one order of magnitude in transcriptomic and proteomic complexity. A typical human gene,

then, produces multiple transcript isoforms that can differ both in their coding and untranslated regions. Different transcript isoforms may play different, and even antagonistic, functional roles that can also be species-specific[240]. Differences between organisms in the fraction of genes with multiple transcript isoforms may contribute to explain the observed phenotypic differences[111,123,27,112]. Moving towards ortholog relationships at the transcript level is challenging. The functionality of gene isoforms and their amount is still an unsolved problem[110,93,202]. Moreover, recent proteomics studies[62] show that a fraction of transcripts do not reach the protein level, and for this reason are less likely to be functional.

The human-mouse orthologs have different characteristics depending on the species annotation. The differences between the species annotations can drive the observed differences between species in gene expression and histone marks. Because the human and mouse annotations used to quantify gene expression and histone peak intensities have important differences confounded by species. We revisited the idea of probes to take into account the differences between species annotations to estimate the gene expression and compare samples from different species.

2.4 Materials and Methods

2.4.1 Human-Mouse Ortholog Annotation

Following Lin et al.[144], we used 15,106 protein coding human-mouse orthologs generated by by the modENCODE and mouse ENCODE consortia[261]. The modENCODE and mouse ENCODE consortia identified the orthologs using a phylogenomics-based approach.

2.4.2 Genome and Gene Annotation

We used the same genome build and gene annotation as Lin et al.[144]. For the human samples we used the ENSEMBL[267] genome build Homo sapiens GRCh37.58 and the GENCODE Release 14 transcript annotation. We downloaded the human genome build from the Illumina iGenomes page (`http://support.illumina.com/sequencing/sequencing_software/igenome.html`), and the transcript annotation from GENCODE[89] (`ftp://ftp.sanger.ac.uk/pub/gencode/release_14/gencode.v14.annotation.gtf.gz`). For the mouse samples, we used the ENSEMBL[267] genome build Mus musculus GRCm38.68 and its corresponding transcript annotation. We downloaded both the mouse genome build (`ftp://ftp.ensembl.org/pub/release-68/fasta/mus_musculus/dna/Mus_musculus.GRCm38.68.dna_sm.toplevel.fa.gz`) and the transcript annotation (`ftp://ftp.ensembl.org/pub/release-68/gtf/mus_musculus/Mus_musculus.GRCm38.68.gtf.gz`) from the ENSEMBL page.

2.4.3 Ortholog Similarity

We estimated the ortholog similarity using the same definition as the modENCODE consortium: the fraction of paired residues with a positive BLOSUM62 score from a BLASTP alignment[261]. First, we retrieved the protein sequences from ENSEMBL with `biomaRt` v.2.28.0[56,55] using the ENSEMBL gene identifiers from the ortholog annotation; the mouse protein sequences from `mmusculus_gene_ensembl` and the human protein sequences from `hsapiens_gene_ensembl`. Following Wu et al., whenever an ENSEMBLE gene ID mapped to more than one protein sequence,

we keep the longest protein sequence.

Then we aligned the human and mouse protein sequences corresponding to the human-mouse ortholog pairs with BLASTP using the BLOSUM62 scoring matrix, and the optimal penalties (gap opening cost of 11 and gap extension cost of 1)[200,121]. The ortholog protein alignment based similarity[209] is given by

$$\text{PID} = \frac{\text{identical positions}}{\text{aligned positions}}$$

We also estimated the ortholog similarity based on nucleotide alignments. First, we took the union of the transcript exons as the genomic region corresponding to each gene. Then, we retrieved the genomic sequences for each gene. We aligned the mouse and human nucleotide sequences using BLASTN[200] with the default parameters, and estimated the similarity based on the resulting nucleotide alignments.

2.4.4 RNA-seq Data Sets

Lin et al.[144] analyzed previously published and newly generated RNA-seq data from human and mouse tissues. The previously published data consisted of data from ENCODE, the Roadmap to Epigenomics Mapping Consortium, and the Illumina BodyMap 2.0. These data sets have a clear batch effect to compare species, since the human and mouse samples were analyzed by different laboratories at different times. Lin et al. addressed this limitation of the previously published data sets by focusing on the analysis of the newly collected data, RNA-seq data from 13 human and mouse

tissues collected by the same lab, using the same processing protocol (Table B.1).

As described earlier, Gilad et al.[78] described the confounding between sequencing batch (the assignment of samples to sequencing flowcells and lanes) with species present in the data used by Lin et al. Lin et al resequenced 12 of the 13 tissues with the sequencing batches balanced by species. We focused on the gene expression data from the resequenced original library preparations from Lin et al.[144] (Table B.2).

2.4.5 Quality Assessment of RNA-seq Data Sets

We used FastQC version 0.10.1 (`https://www.bioinformatics.babraham.ac.uk/projects/fastqc/`) to assess the quality of the individual FASTQ files, and the mapping statistics from the STAR alignments (version 2.5.0c[52]).

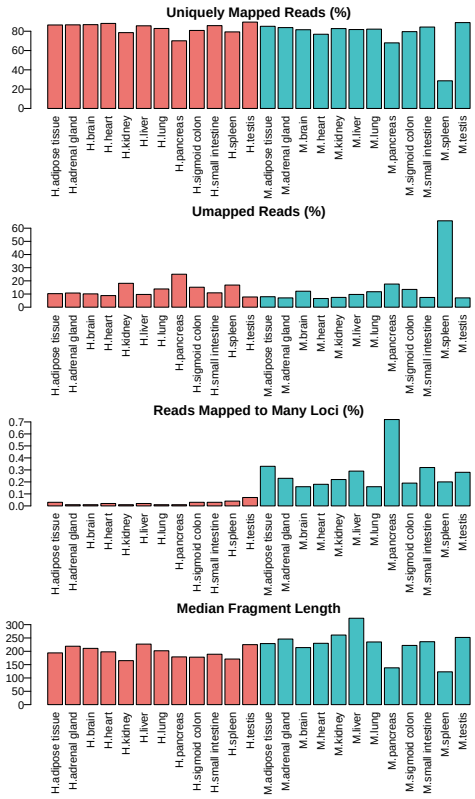


Figure 2.18. STAR Mapping Statistics for the Resequenced ENCODE Data Set.

We only considered 10 out of the 12 tissues resequenced by Lin et al.[144], because the mouse pancreas and spleen samples had issues with GC-content bias and short fragment lengths. Both FASTQ files (ENCFF859JTH, ENCFF432KKN) from the mouse spleen sample had the highest number of QC fails reported by FastQC, followed by one FASTQ file from the human pancreas (ENCFF849WWH) and from the mouse pancreas (ENCFF541KUW) (Table B.4). Furthermore, the mouse spleen sample had the lowest percentage of uniquely mapped reads among the mouse samples, a high percentage of unmapped reads, and a high percentage of reads mapped to too many loci (Figure 2.18). Both the mouse pancreas and spleen samples had a lower median fragment length compared to the rest of

the mouse samples (Figure 2.18).

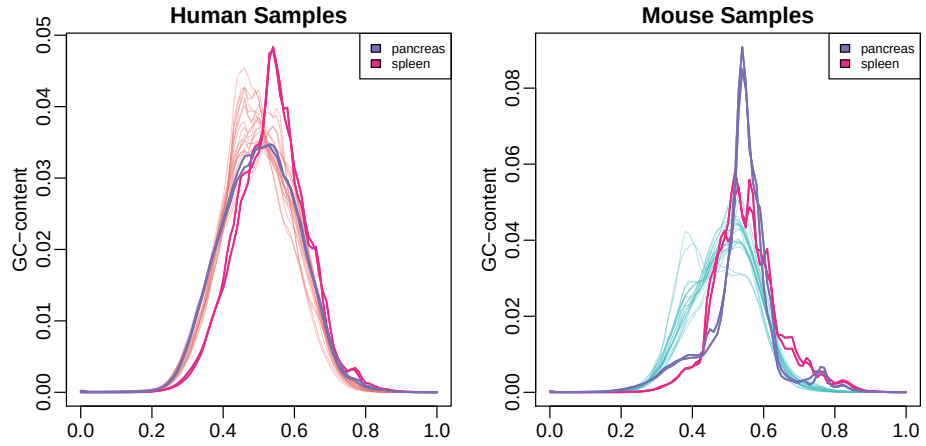


Figure 2.19. GC-Content Distribution for Resequenced RNA-seq Data Set. Distribution of GC-content over all reads for the human and mouse tissues. The pancreas and spleen samples are colored in purple and pink respectively.

Furthermore, the distribution of GC-content over all the reads from the FASTQ files from the human and mouse spleen samples, as well as the mouse pancreas sample had unusually shaped distributions (Figure 2.19).

Gene Expression for ENCODE RNA-seq Data Sets

The FPKM values for the original and resequenced ENCODE RNA-seq data sets were kindly provided by Lin et al.[144]. In the ENCODE consrotium pipeline, the FASTQ files were aligned using Tophat[246] to ENSEMBL genome build Homo sapiens GRCh37.58 for the human samples and to Mus musculus GRCm38.68[68] for the mouse samples. The FPKM values were assigned using Cufflinks[246] with the GTF files Gencode Release 14[89] or `Mus_musculus.GRCm38.68.gtf`[67] for humans and mice, respectively.

Following Lin et al.[144] we took the $\log_2$ of the FPKM values after adding pseudo count of 1 ($\log_2(\text{FKPM} + 1)$). After pairing the log-transformed FPKM values for the human and mouse samples using the ortholog annotation, we normalized the transformed FPKM values using quantile normalization with the function `normalize.quantiles` from the Bioconductor `preprocessCore` version 1.34.0[22].

Additionally, we quantified the expression adjusted for fragment GC-content bias using Salmon v.0.8.0[198] of the resequenced ENCODE RNA-seq data set with the following settings: `--gcBias -l ISR`. We then used `tximport` v.1.6.0[3] to summarize the transcript-level abundances into gene-level abundances. As with the FPKM values, we normalized the gene-level abundances using quantile normalization with the function `normalize.quantiles` from the Bioconductor `preprocessCore` version 1.34.0[22].

2.4.6 Ortholog Probes

We used `BSgenome` version 1.40.1[192] to import the genome annotations in R, and `GenomicFeatures` version 1.40.1[128] to import the gene annotation in R. For each ortholog, we retrieved its trancripts, and we took the union of the transcript exons to represent the gene. Then, we retrieved the genomic sequence for each gene. If the gene has a positive strand orientation, then we concatenated the sequences for the union of exons into a single sequence. If the gene has a negative strand orientation, we first reverse the orientation of the sequences from each exon, and then we concatenated. In this manner, the concatenated sequence as a whole will match the gene negative strand orientation. Then, we used the Smith-Waterman algorithm implemented in `Biostrings` version 2.40.2[193] to

align the concatenated sequences for each pair of orthologs. Since we want to have the same number of bases in both human and mouse alignments, we split the resulting alignments into segments excluding all insertions and deletions. Then, we removed the segments less than 10 base pairs long. Finally, we mapped the alignments to their corresponding genomic locations to get the ortholog probes.

2.4.7 Alignment Parameters Selection

The Smith-Waterman algorithm uses a score function where matches increase the overall score of an alignment while mismatches decrease it. The score is given by the number of matches, the number of mismatches, and the number of insertions or deletions[251]. Specifically, out score is defined as

$$\mathrm{No}\{\text{matches}\} - \mu \times \mathrm{No}\{\text{mismatches}\} - \partial \times \mathrm{No}\{\text{insertions/deletions}\}$$

where μ is the mismatch penalty and ∂ the gap penalty. Thus, a good alignment has a positive score and a poor alignment a negative score. The local algorithm finds an alignment with the highest score by considering only alignments that score positive and picking the best one from those[251,179]. The optimal alignment maximizes this score.

Since there is considerable disagreement on the choice of alignment parameters[86,179], we considered the following settings for the alignment parameters

$$1 \leq \delta \leq 10$$

$$1 \leq \mu \leq 10$$

We chose the alignment parameters based on the following features of the alignments: the number of aligned positions excluding the gaps (Figure B.3), the proportion of the human and mouse gene present in the alignment (Figures B.4,B.5)

$$\frac{\text{length of gene}}{\text{aligned positions}}$$

and the following similarity metrics[209] (Figures B.6,B.7,B.8,B.9)

$$\text{PID}_1 = \frac{\text{identical positions}}{\text{aligned positions} + \text{internal gap positions}}$$

$$\text{PID}_2 = \frac{\text{identical positions}}{\text{aligned positions}}$$

$$\text{PID}_3 = \frac{\text{identical positions}}{\text{length of shorter sequence}}$$

$$\text{PID}_4 = \frac{\text{identical positions}}{\text{average length of two sequences}}$$

For a given set of alignment parameters, we aligned all human-mouse orthologs, and took the median of each of the features for all the resulting alignments. Then for each set of alignment parameters (μ, δ) we took the rank of the median of every feature. We picked the alignment parameters with the highest mean rank across all features. The highest scoring alignment parameters were $\mu = 1$, $\delta = 2$ (Figure 2.20).

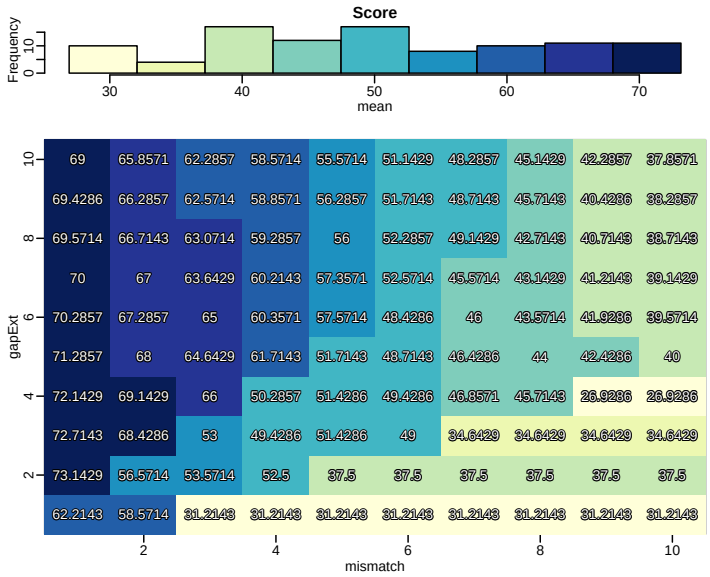


Figure 2.20. Alignment Parameters Score. Score for each set of alignment parameters, `gapExt` (δ) and `mismatch` (μ), based on the mean ranks of the median length, median proportion of human gene and mouse gene present in the alignment, and median sequence similarity. (Top) Histogram with the color palette for alignment parameters score, and (bottom) heat map for the score for all alignment parameters considered.

2.4.8 Gene Expression with Ortholog Probes

We aligned the RNA-seq reads to their respective genomes using STAR version 2.5.0c[52] with the default parameters. From these alignments, we obtained the coverage using the BAM files with `bamsignals` v.1.10.0[3]. Then, we used the function `featureCounts` from `Rsubread` version

1.22.3[142] to assign the read counts to the ortholog probes with the following settings: `useMetaFeatures=TRUE`, `allowMultiOverlap=FALSE`, `isPairedEnd=TRUE`, and `requireBothEndsMapped=TRUE`.

We normalized the probe read counts for sequencing depth and length based on the transcripts per kilobase million (TPM) normalization[277,216]. For each probe p,

$$TPM_p = \frac{RPK_p}{\left(\frac{\sum_p RPK_p}{10^6}\right)}$$

where

$$RPK_p = \frac{r_p}{l_p}$$

r_p is the read counts for probe p, and l_p is the length in base pairs for probe p.

Finally, we took the $\log_2$ of the probe TPM values after adding a pseudo count of 1 ($\log_2(TPM + 1)$). After pairing the log-transformed probe TPM values for the human and mouse samples using the ortholog annotation, we normalized the transformed probe TPM values using quantile normalization with the function `normalize.quantiles` from the Bioconductor `preprocessCore` version 1.34.0[22].

2.4.9 Human and Mouse Microarrays

We used 83 Affymetrix Human Genome U133 Plus 2.0 microarrays from 10 normal human tissues and 82 Affymetrix Mouse Genome 430 2.0 Array from 10 normal mouse tissues manually curated

in Barcode 3.0[165] (Table B.3). The curated microarrays in Barcode 3.0 were filtered to exclude poor quality samples[165,166]. We used the R package `GEOquery`[45] to retrieve raw CEL files from the Gene Expression Omnibus (GEO)[13]. We processed the raw data with fRMA[161]. We obtained the annotation for the human array platform from `hgu133plus2.db`[31] and the mouse annotation from `mouse4302.db`[32]. We only keep the probe sets that mapped uniquely to one ortholog. We matched the normalized values for the human and mouse microarrays using the ortholog annotation, and normalized them using quantile normalization with the function `normalize.quantiles` from the Bioconductor `preprocessCore` version 1.34.0[22].

2.4.10 Species Effect

Let $Y_{s,g,t,k}$ denote the quantile normalized log (base 2) transformed expression values where $s \in \{\text{human, mouse}\}$ indexes the species, $g \in \{1, \ldots, G\}$ indexes the human-mouse orthologs, $t \in \{1, \ldots, T\}$ indexes the tissues and $k \in \{1, \ldots, K\}$ indexes the replicates within tissue. The species effect is the observed difference between the human and mouse samples for each ortholog. We defined this difference as

$$d_g = \bar{Y}_{\text{human},g} - \bar{Y}_{\text{mouse},g}$$

where

$$\bar{Y}_{s,g} = \frac{1}{KT} \sum_{k=1}^{K} \sum_{t=1}^{T} Y_{s,g,t,k}$$

2.4.11 Tissue Effect

For a given tissue $t_\circ$, the tissue effect is the difference in expression between that tissue and all others for each ortholog. We defined this difference as

$$d_{g,t_\circ} = \bar{Y}_{g,t_\circ} - \bar{Y}_{g,-t_\circ}$$

where

$$\bar{Y}_{g,t_\circ} = \frac{1}{2K} \sum_{k=1}^{K} \sum_{s \in \{\text{human,mouse}\}} Y_{s,g,t_\circ,k}$$
$$\bar{Y}_{g,-t_\circ} = \frac{1}{2K} \sum_{k=1}^{K} \sum_{t \neq t_\circ} \sum_{s \in \{\text{human,mouse}\}} Y_{s,g,t_\circ,k}$$

2.4.12 T-Statistics

We also computed the t-statistics corresponding to the species and tissue effect to account for the differences in variance between the estimates based on FPKM, ortholog probes, and microarrays. The t-statistic for the species effect is

$$z_g = \frac{d_g}{\widehat{\text{SE}}(d_g)}$$

using the usual estimate for the pooled standard deviation $\widehat{\text{SE}}(d_g)$.

Similarly, the t-statistic for the tissue effect is

$$z_{g,t_\circ} = \frac{d_{g,t_\circ}}{\widehat{\mathrm{SE}}(d_{g,t_\circ})}$$

where $\widehat{\mathrm{SE}}(d_{g,t_\circ})$ is also the pooled standard deviation estimate.

2.4.13 Differentially Expressed Orthologs

To identify the differentially expressed orthologs we used both the resequenced ENCODE RNA-seq data set and the Barcode 3.0 microarray samples. First, we estimated the species effect, the difference in means between the human and mouse samples, separately using the microarray samples and the normalized probe counts from the resequenced ENCODE data set.

Then we identified the differentially expressed orthologs first by testing the difference in means between the human and mouse samples using LIMMA[215] with both platforms, and adjusting for multiple comparison with `qvalue` version 2.4.2[260]. Then we selected orthologs with $q < 0.05$ and species effect estimates above a threshold C in both platforms

$$A = \{g : |d_g^{\text{mcr}}| > C, |d_g^{\text{probes}}| > C, q_{\text{mcr}} < 0.05, q_{\text{probes}} < 0.05\}$$

We considered an ortholog differentially expressed if the species effect estimates have the same sign in

both platforms. Thus, the set of differentially expressed orthologs is

$$D = \{g : |d_g^{\text{mcr}}| > C, |d_g^{\text{probes}}| > C, q_{\text{mcr}} < 0.05, q_{\text{probes}} < 0.05, \text{sign}(d_g^{\text{mcr}}) = \text{sign}(d_g^{\text{probes}})\}$$

On the other hand, we considered orthologs whose species effect estimates have different signs depending on the platform as false positives. We define this set of false positives as

$$B = \{g : |d_g^{\text{mcr}}| > C, |d_g^{\text{probes}}| > C, q_{\text{mcr}} < 0.05, q_{\text{probes}} < 0.05, \text{sign}(d_g^{\text{mcr}}) \neq \text{sign}(d_g^{\text{probes}})\}$$

We picked the threshold C to control for the false discovery rate based on the set B and the estimate

$$\widetilde{\text{FDR}} = \frac{||B||}{||A||}$$

2.4.14 Gene Ontology (GO) Enrichment Tests

We retrieved the Gene Ontology (GO) biological process annotation from `GO.db` v.3.5.0[3], and `org.Hs.eg.db` v.3.5.0 to map the human Ensembl gene identifiers to their corresponding GO terms. We restricted the GO annotation to the GO slim terms. GO slims are cut-down versions of the GO ontologies containing a subset of the terms in the whole GO. They give a broad overview of the ontology content without the detail of the specific fine grained terms. We downloaded the GO

slim generic annotation from the Gene Ontology Consortium[8] (http://www.geneontology.org/ontology/subsets/goslim_generic.obo). We used GSEABase v.1.40.1[3] to map the GO annotation from GO.db to their corresponding GO Slim terms. We performed the enrichment tests using Fisher's exact test implemented in topGO v.2.30.1[3] .

2.4.15 ChIP-seq Data Sets

We downloaded bed narrowPeak files with the peaks called with MACS[276] from the ENCODE page (https://www.encodeproject.org/)[41]. Following the recommendations from the ENCODE project (https://www.encodeproject.org/chip-seq/histone/), we downloaded replicated peaks for the mouse tissues and stable peaks for the human tissues. The mouse tissues have two replicates, the replicated peaks are the set of peak calls from the pooled replicates. These peaks are observed in both replicates. On the other hand, the human tissues do not have replicates. The stable peaks for the human tissues are the set of peak calls from two partitions, or *pseudoreplicates*. A pseudoreplicate is a subsample of reads, chosen without replacement, from a single replicate used as a substitute for replication in the absence of true biological replicates[125]. We paired the human and mouse peak calls by histone mark and tissue type (Table B.3).

We assigned peak intensities to each gene following the same procedure as Lin et al.[144]. First we defined the promoter regions as 1 kb before and after the transcription start site (TSS). We assigned the log2 fold changes of enrichment over control from the peaks overlapping with the promoter region as the promoter-associated peak intensity. We used the function findOverlaps from GenomicRanges version 1.24.3[129] to find the peaks overlapping with the promoter regions. If more

than one peak overlapped with the promoter region, we assigned the sum of the log2 fold changes corresponding to all the overlapping peaks as the promoter-associated peak intensity. For genes with multiple promoters, we selected the highest promoter-associated peak intensity.

For genes without overlapping peaks, we assigned a peak intensity of 0. For our analysis, we only considered ortholog pairs where at least one of the genes had an assigned peak intensity. For each histone mark, we normalized the peak intensities using quantile normalization with the function `normalize.quantiles` from the Bioconductor `preprocessCore` version 1.34.0 [22] for each given pair of human and mouse tissues.

We also considered only the peak intensities associated with human-mouse orthologs with the same number of annotated transcripts. First, we subset the gene-associated peak intensities for the human-mouse orthologs with the same number of annotated transcripts. We assigned a peak intensity of 0 to genes without overlapping peaks. Then, we selected the human-mouse ortholog pairs where at least one of them had an assigned peak intensity. Finally for each histone mark, we normalized the peak intensities using quantile normalization with the function `normalize.quantiles` from the Bioconductor `preprocessCore` version 1.34.0 [22] for each given pair of human and mouse tissues.

3

Linear Models with Applications to Gene Coexpression

3.1 Introduction

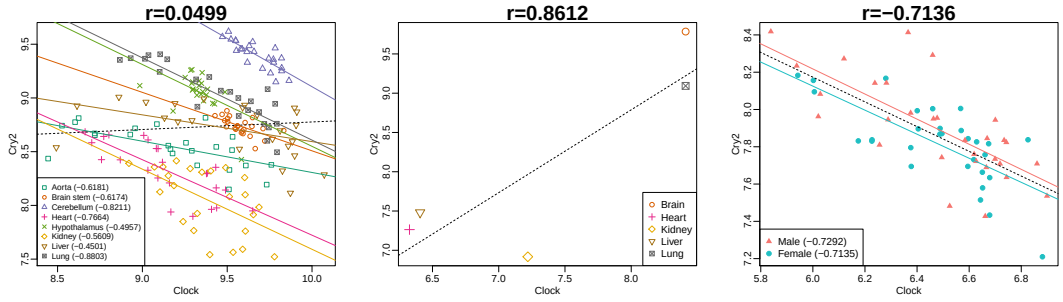
A major goal of functional genomics is to describe relationships between genes. The advent of microarrays led to numerous gene expression studies involving large numbers of samples. The increas-

ing amount of data from gene expression studies motivated the analysis of relationships between genes using their correlation based on gene expression data. When the mRNA expression of two or more genes is correlated across multiple samples, these genes are said to be coexpressed[76]. Gene co-expression networks[249] are networks where the nodes correspond to genes, and the edges are based on a measure of similarity between the gene expression profiles such as correlation. Studies using gene coexpression networks usually follow three major steps[273,235,158,29]. First, estimate the correlation between all genes. Second, build a network where the genes are nodes and the edges between them are based on their correlation. Finally, find clusters of genes in the network. Previous work has shown that many coexpression clusters are conserved across phylogeny[236,174,189,231], enriched with protein-protein interactions[174,190,95], and enriched with specific functional categories of genes including ribosomal, mitochondrial, synaptic, immune, hypoxic, mitotic among other categories[174,190,95,94]. However, interpreting correlations based on gene expression data is challenging because these correlations can arise from biological as well as non-biological sources. In this work, we use the framework of linear models to quantify the influence of the variation from experimental factors on the observed correlation, and to correct the observed expression values for the effect of experimental factors.

3.2 Results

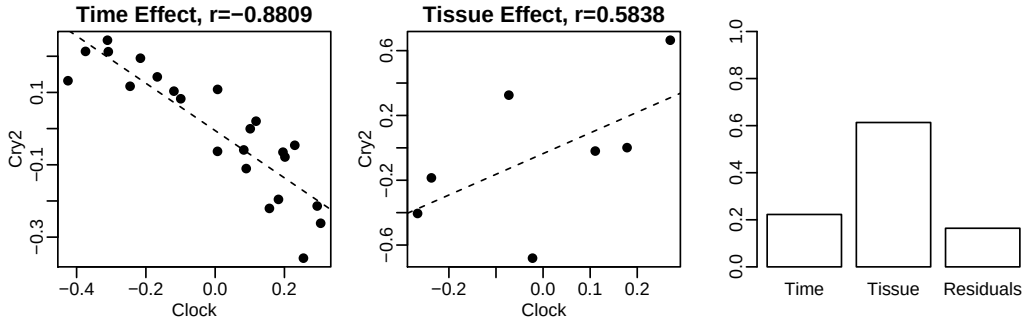
3.2.1 Spurious correlation from ignoring experimental factors

Ignoring experimental factors from gene expression studies can lead to spurious correlations[49]. To illustrate the effects of ignoring the underlying structure of the data, we estimated the correlation between two known circadian genes, Cry2 and Clock, using data from a time course data set (Table C.3)[274]. The time course data set consists of 8 mouse tissues sampled every 2 hours for 2 days (24 time points). If we pool the gene expression profiles for each gene from all tissues, the observed correlation is small and positive (0.0449). However, the correlation stratified by tissue across time points is negative, ranging from -0.4501 in liver to -0.8803 in lung (Figure 3.1a). This is a clear example of Simpson's paradox. By ignoring the presence of different groups, in this case tissues, the sign of the correlation changes[4,230].

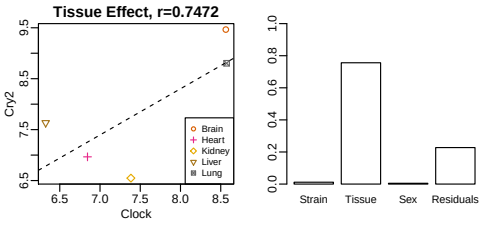


(a) Time Course (b) Tissue Panel (c) Liver Biological Replicates

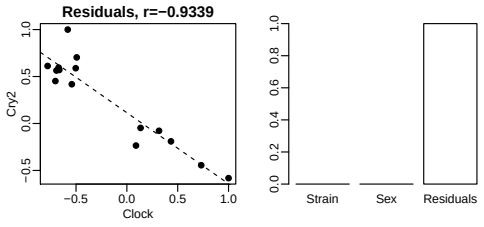
Figure 3.1. **Correlation between Cry2 and Clock.** Correlation between circadian genes Cry2 and Clock in different data sets. In the scatter plots, the black dashed line corresponds to the pooled correlation. (a) **Time Course.** The color and shape represent the tissue type, and each colored line corresponds to the correlation stratified by tissue. (b) **Tissue Panel.** The expression profiles for Cry2 and Clock were aggregated in each tissue by taking the mean. The color and shape represent the tissue type. (c) **Liver Biological Replicates.** The color and shape represent the sex of the mice. The colored lines correspond to the correlation stratified by sex.



(a) Time Course



(b) Tissue Panel



(c) Liver Biological Replicates

Figure 3.2. **Correlation between Cry2 and Clock under different contexts.** (a) **Time Course.** Scatter plots for the time effect estimates and tissue effect estimates from the time course data set. Pairwise gene expression variance partition for the time course data set. (b) **Tissue panel.** Scatter plot for the tissue effect estimate from the tissue panel data set. Pairwise gene expression variance partition for the curated tissues data set. (c) **Liver Biological Replicates.** Scatter plot for the residuals from the liver biological replicates data set. Pairwise gene expression variance partition for the liver biological replicates data set.

Furthermore, the correlation between Cry2 and Clock changes under different contexts. The correlation between different tissues is positive while the correlation across time within the same tissue is negative. We used a different data set consisting of 5 different tissues from biological replicates to estimate correlation between Cry2 and Clock (Table C.4). As we observed in the time course data set, the correlation driven by the differences between tissues is positive (Figure 3.1b). On the other hand, without the context of different tissue types the correlation changes sign. We used a data set

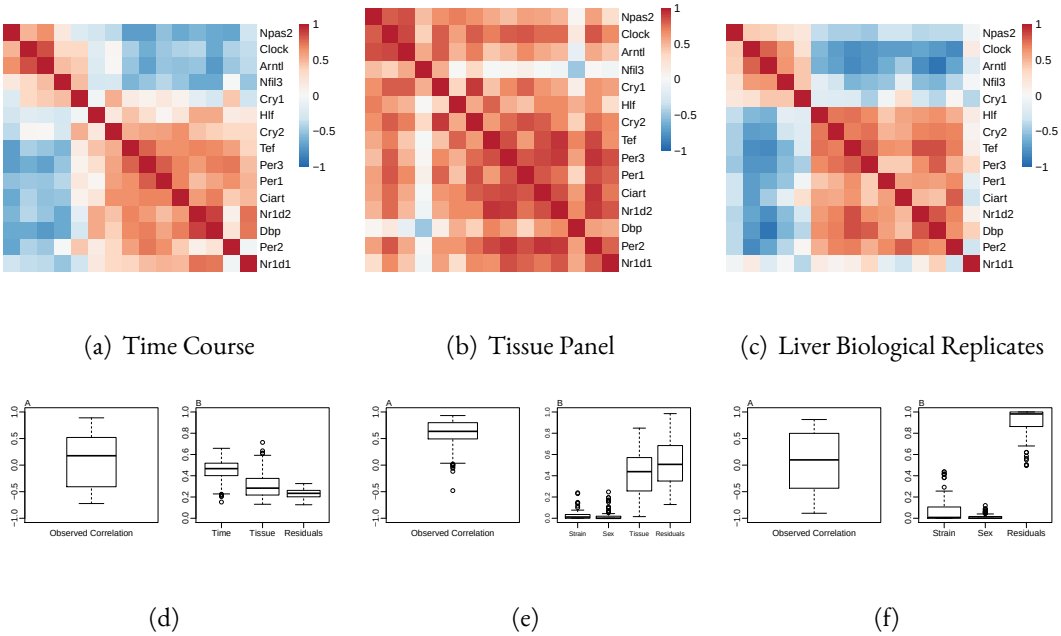
consisting of liver samples from male and female mice from 10 different strains (Table C.5). The correlation between Cry2 and Clock is negative (-0.7136) across all samples as well as stratified by sex (Figure 3.1c).

We estimate the gene-specific effect of tissue and time in the time course data set using a linear model. The correlation between the Cry2 and Clock time effect estimates is negative (-0.8809) while the correlation between the tissue effect estimates is positive (0.5838). Moreover, the observed positive pooled correlation was driven by the variability between tissues (Figure 3.2a). We also estimate the gene-specific tissue effect in the curated tissues data set. The correlation between the Cry2 and Clock tissue effect estimates is also positive (0.8612) and the observed positive pooled correlation is also driven by tissue variability (Figure 3.2b). In the liver biological replicates, the observed negative correlation between Cry2 and Clock was not driven by the differences between tissues since all the samples come from the same tissue. After adjusting with our model the expression values of Cry2 and Clock for the effects of strain and sex, the correlation is negative (-0.9339). Moreover, the negative observed pooled correlation was not driven by the variability between sexes or strains (Figure 3.2c). The negative correlation in the liver biological replicates might reflect the differences in circadian phase between Cry2 and Clock (Table C.1).

3.2.2 Gene correlation structure conserved in different data sets

We used circadian genes to estimate correlation between genes know to be related[274]. The expression of these genes oscillates with the circadian rhythm. The data sets that we consider have different sources of variation that affect the observed correlation between the circadian genes. The observed

pooled correlation between the circadian genes is similar in the time course data set and in the liver biological replicates but different in the tissue panel data set (Figure 3.3a-c).



(a) Time Course (b) Tissue Panel (c) Liver Biological Replicates

(d) (e) (f)

Figure 3.3. **Correlation between circadian genes.** Heat maps for the observed correlation in the (a) time course experiment, (b) tissue panel, and (c) liver biological replicates. Pairwise gene expression variance partition for the (d) time course data set, the (e) tissue panel, and the (f) liver biological replicates. The box plots are the proportion of variance explained across all pairs of circadian genes.

In the time course data set the observed pooled correlation is driven by the variability across time (Figure 3.3d). In the curated tissue data set, the observed pooled correlation is not driven entirely by tissue variability (Figure 3.3e). Likewise, in the liver biological replicates the correlation is not sorely driven by the variability between strains or sexes (Figure 3.3).

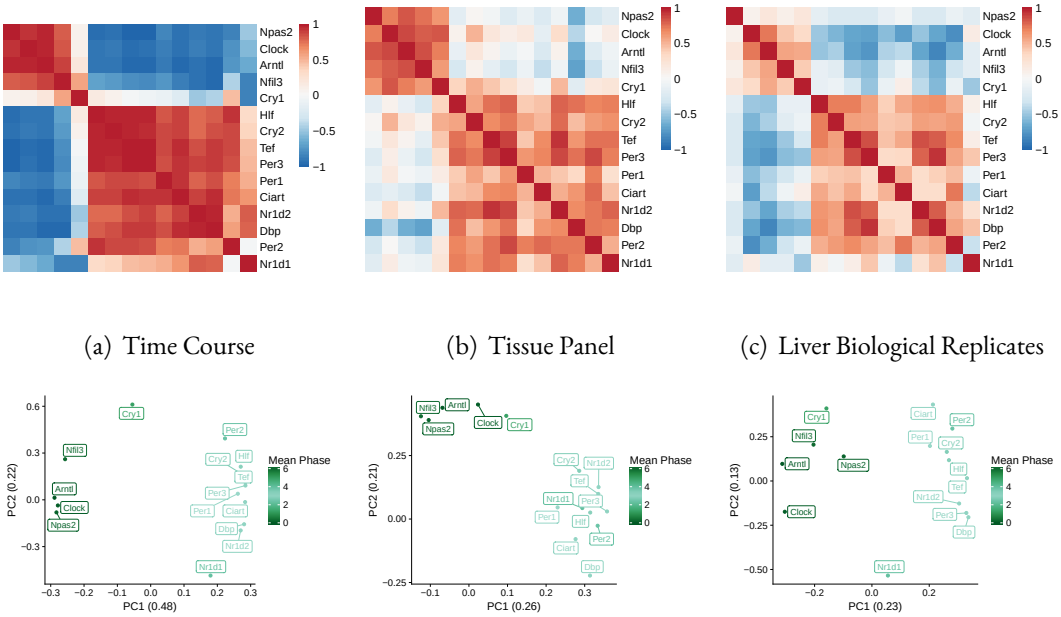
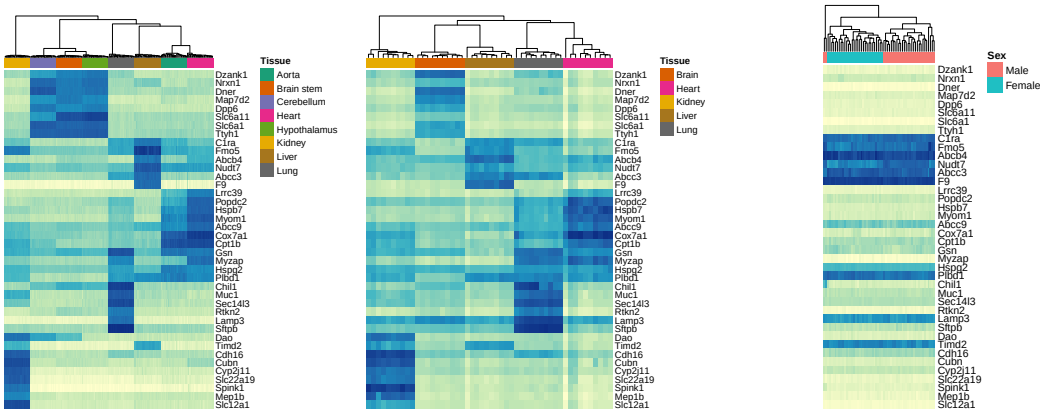


Figure 3.4. **Circadian gene-specific effects.** Heat maps for the correlation between the (a) time effect estimates from the time course data set, the (b) residuals from the tissue panel, and the (c) residuals from the liver biological replicates. PCA plots based on the (d) time effect estimates from the time course data set, the (e) residuals from the tissue panel, the (f) residuals from the liver biological replicates. The color corresponds to the mean phase of the circadian genes.

We use linear models to take into account context in the correlation between the circadian genes. The correlation between the gene-specific time effect estimates from the time course data set shows clearly two groups of circadian genes (Figure 3.4a). We use the residuals from linear models to adjust for the gene-specific tissue effect in the tissue panel data set and the gene-specific effects of strain and sex in the liver biological replicates. The patterns in the correlation between the residuals for the circadian genes resemble the patterns in the time course data set (Figure 3.4b,c). The groups of circadian genes in the time course data set have a clear difference in their mean phase (Figure 3.4d, Ta-

ble C.1). Similarly, the groups of circadian genes in the residuals from the curated tissues data set and the liver biological replicates resemble the groups observed in the time course data set (Figure 3.4e,f).



(a) Time Course (b) Tissue Panel (c) Liver Biological Replicates

Figure 3.5. **Tissue-specific genes.** Heat maps of the tissue-specific genes for the (a) time course data set, the (b) tissue panel and the (c) liver biological replicates.

Additionally we used tissue-specific genes; genes highly expressed in brain, heart, lung, kidney and liver (Figure 3.5). The tissue-specific genes separate the samples by tissues in the time course data set (Figure 3.5a) and in the tissue panel data set (Figure 3.5b). In the liver biological replicates, the tissue-specific genes separate almost completely the samples by sex (Figure 3.5c). The patterns in the pooled correlation between the tissue-specific genes are similar in the time course data set (Figure 3.6a) and in the tissue panel data set (Figure 3.6b) since the similar tissue types are present in both data sets. On the other hand, there is no clear pattern in the correlation from the liver biological replicates (Figure 3.6c).

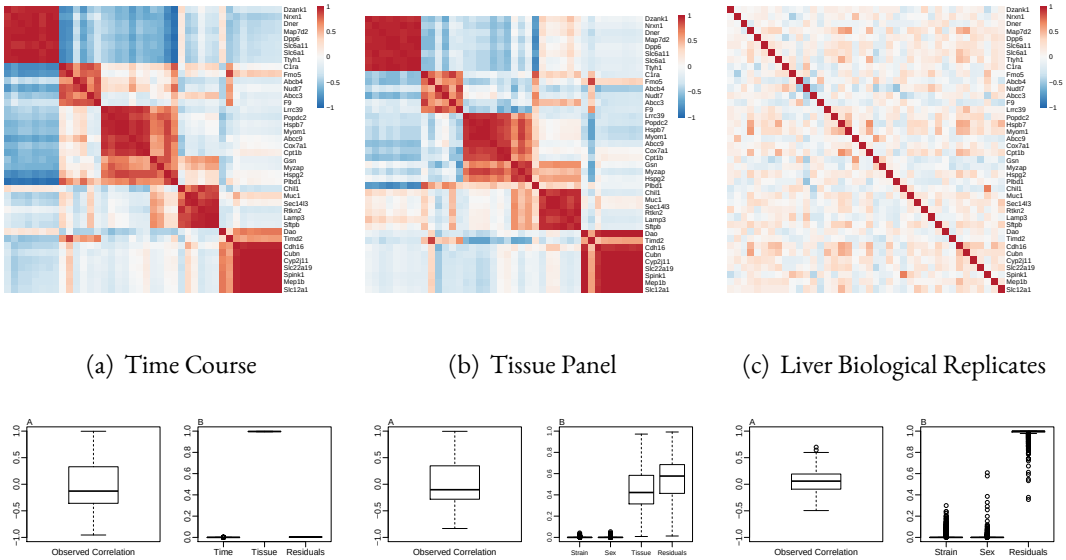
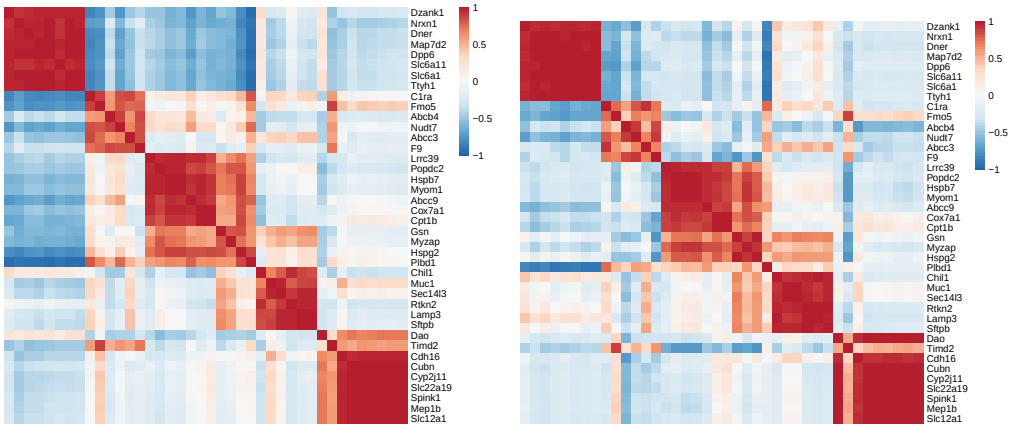


Figure 3.6. **Correlation between tissue-specific genes.** Heat maps for the observed correlation in the (a) time course data set. the (b) tissue panel, and the (c) liver biological replicates. Pairwise gene expression variance partition for the (d) time course data set, the (e) tissue panel, and the (f) liver biological replicates. The box plots are the proportion of variance explained across all pairs of circadian genes.

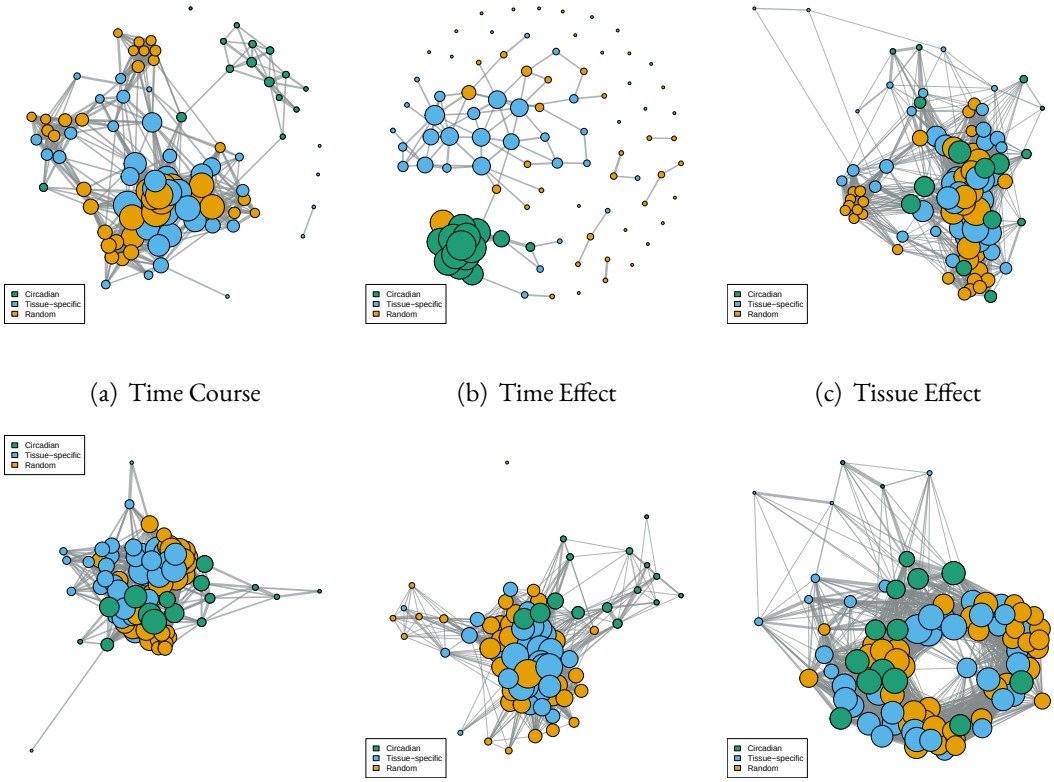
In the time course data set, the correlation is completely driven by the variability between tissues (Figure 3.6d). However, in the tissue panel data set the correlation is not driven mainly by the tissue variability (Figure 3.6e). In the absence of different tissues, the correlation in the liver biological replicates is driven by the residual variability, variability not due to sex or strain (Figure 3.6f). The patterns observed in the correlation between the gene-specific tissue effect from the time course data set (Figure 3.7a) and from the tissue panel data set (Figure 3.7b) resemble the patterns observed in the pooled correlation (Figure 3.6a,b).



(a) Time Course (b) Tissue Panel

Figure 3.7. **Correlation between tissue-specific gene-specific effects.** Heat maps for the correlation between the gene-specific tissue effect estimates from the (a) time course data set and the (b) tissue panel data set.

3.2.3 Gene coexpression networks under different contexts.



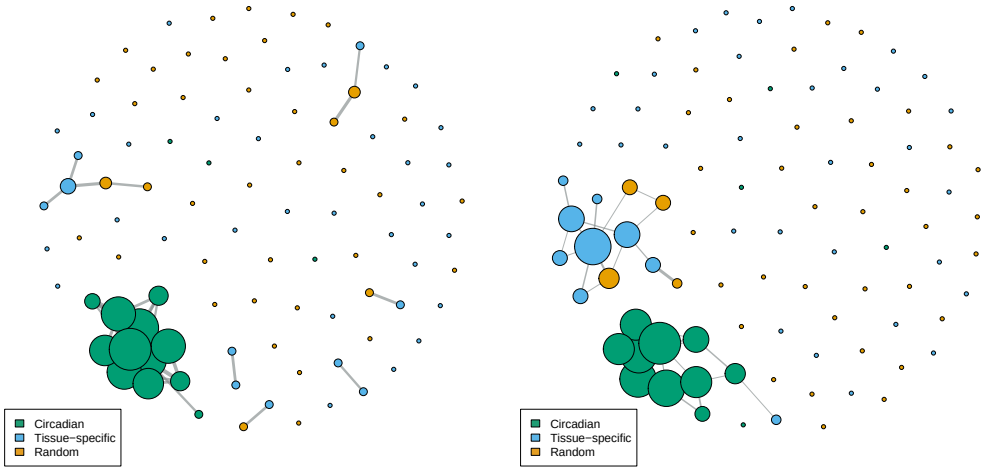
(a) Time Course (b) Time Effect (c) Tissue Effect

(d) Tissue Panel (e) Residuals (f) Tissue Effect

Figure 3.8. **Coexpression network for the circadian and tissue-specific genes.** The size of the vertex is proportional to the degree of the gene (connections to other genes). The color of the genes correspond to their class: green for circadian genes, blue for tissue specific genes and yellow for randomly selected genes. Coexpression network based on the (a) observed correlation, the (b) correlation between the gene-specific time effect estimates, and the (c) gene-specific tissue effect estimates from the time course data set. Coexpression network based on the (d) observed correlation, the (e) residuals, and the (f) gene-specific tissue effect estimates from the tissue panel data set.

We used WGCNA[126] to build coexpression networks based on the observed correlation and the correlation between the gene-specific estimates from the linear models. In the networks, we include the circadian genes, the tissue-specific genes as well as 40 randomly selected genes. We consider

the degree of the genes, the number of connections to other genes in the network, as a measure of importance of a gene in a coexpression network. In the coexpression networks based on the observed correlation and the correlation between the gene-specific tissue effect estimates from the time course experiment, the randomly selected genes and the tissue-specific genes have higher degrees than the circadian genes (Figure 3.8a,c). However, in the coexpression network based on the correlation between the gene-specific time effect estimate the circadian genes have the highest degree (Figures 3.8b,and C.4). In the coexpression networks based on the tissue panel, the randomly selected genes and the tissue-specific genes dominate the network (Figures 3.8d-f, and C.5). On the other hand, in all coexpression networks based on the liver biological replicates the circadian genes have the highest degree (Figures 3.9a,b, and C.6).



(a) Liver Biological Replicates (b) Residuals

Figure 3.9. **Coexpression network for the circadian and tissue-specific genes.** The size of the vertex is proportional to the degree of the gene (connections to other genes). The color of the genes correspond to their class: green for circadian genes, blue for tissue specific genes and yellow for randomly selected genes. Coexpression network based on the (a) observed correlation and the (b) residuals from the liver biological replicates.

3.3 Discussion

The correlation between gene expression profiles has been widely used as a measure of gene coexpression to infer relationships between genes. We need to take into account the drivers of gene expression variation to interpret correlation from gene expression data. The interpretation of gene coexpression relationships depends heavily on context[127]. In a data set consisting of samples from multiple tissues, coexpressed modules (modules defined by coexpression similarity) will often distinguish genes that are expressed in tissue-specific patterns[189,103]. On the other hand, in a data set consisting of samples from a single tissue type, coexpression modules may distinguish genes that are preferentially expressed in distinct cell types that comprise that tissue[190]. Furthermore, in a data set consisting of samples from a homogeneous cellular population, coexpression modules may correspond more directly to sets of genes that work in tandem to perform various intracellular functions[127].

We used a linear mixed model to partition the observed expression variation among experimental factors such as time points and tissue types. The linear mixed models offer a framework to describe the influence of the sources of variation on the observed correlation between a pair of genes. We used a linear model to estimate and correct for the effect of experimental factors on the expression profiles. We recovered the correlation structure of known mouse circadian genes driven by changes in expression across time by accounting for the differences between tissues in three distinct data sets with different sources of variation using linear models. We proposed an approach to estimate and interpret the correlation under different contexts. Estimating the correlation without taking into

account the experimental factors from a given data set might produce spurious correlations. Our approaches are limited to settings where the experimental factors are known in advance. Furthermore, the estimates for variance partition only take into account one pair of genes at a time. However, our results highlight the importance of context for the estimation and interpretation of correlation based on gene expression data.

3.4 Materials and Methods

3.4.1 Time Course Data Set

We used 192 Affymetrix MoGene 1.0 ST arrays from 9 C57/BL6 mouse tissues (aorta, brain stem, cerebellum, heart, hypothalamus, kidney, liver, and lung) sampled every 2 hours for 2 days (24 samples per tissue)[274] (Table C.3). We downloaded the CEL files from the Gene Expression Omnibus (GEO)[13] and imported them in R using `oligo` v.1.42.0[35]. We obtained the microarray annotation from `pd.mogene.1.0.st.v1` v.3.14.1[34] and processed the CEL files using the frozen RMA (fRMA) implementation[164] in `fRMA` v.1.30.1[162,McCall & Irizarry].

3.4.2 Liver Biological Replicates and Tissue Panel Data Sets

We used 60 Affymetrix Mouse Genome 430 2.0 arrays from the Novartis 12 Strain Diet Sex Survey control group (Table C.5). The data set consists of liver tissue harvested from 3 males and 3 females from 10 mice strains (129S1/SvImJ, A/J, C57BL/6J, BALB/cJ, C3H/HeJ, DBA/2J, I/LnJ, MRL/MpJ-Tnfrsf6lpr/J, NZB/BINJ, and SM/J).

For the tissue panel data set, we used 50 Affymetrix Mouse Genome 430 2.0 arrays from 5 mouse tissues (brain, heart, kidney, liver, and lung) curated in Barcode 3.0[165] (Table C.4).

We used the R package `GEOquery`[45] to retrieve raw CEL files from GEO[13], and processed the raw data with the fRMA implementation[162] in `fRMA` v.1.30.1[162] using the annotation from `mouse4302.db`[33].

3.4.3 Mapping Probe Sets to Genes

We mapped the probe sets to their corresponding genes with `mouse4302.db` v.3.2.3[33] for the Affymetrix Mouse Genome 430 2.0 arrays, and with `mogene10stprobeset.db` v.8.7.0[153] for the Affymetrix MoGene 1.0 ST arrays. When a gene mapped to multiple probe sets, we assigned the average of the normalized probe set expression value as the gene-level expression value.

3.4.4 Circadian Gene Set

The set of circadian genes are core clock genes than oscillate with circadian rhythm in various mouse tissues reported by Zhang et al[274]. We also retrieved the mean phase estimates from JTK_CYCLE[97]. We fitted a harmonic regression model to the normalized circadian gene expression values from the time course experiment using the R package `HarmonicRegression` v.1.0[151].

3.4.5 Tissue Specific Genes

We used the Gene Expression Barcode 3.0[165] estimates to select probe sets from the Affymetrix MoGene 1.0 ST array that were highly expressed in the following tissues: brain, heart, lung, kidney and

liver (Figure C.2). For each tissue, we mapped the probe sets to their corresponding genes using `mogene10sttranscriptcluster.db` v.8.7.0[154] and selected 8 genes per tissue.

3.4.6 Linear Mixed Model Framework for Variance Estimates

We used a linear mixed model[15,122,204] for each pair of genes to describe the influence of the sources of variation on the observed correlation. Consider the model

$$Y_{ijk} = \mu + \alpha_i + b_j + c_k + \varepsilon_{ijk}$$

where i indexes the genes, j and k index covariates of interest such as tissue type, α_i is the fixed gene effect,

$$b_j \sim N(0, \sigma_b^2)$$

$$c_k \sim N(0, \sigma_c^2)$$

are random effects, and $\varepsilon_{ijk} \sim N(0, \sigma_e^2)$ is the error term. Under this model and assuming independence between the random effects, we have

$$\mathrm{Var}(Y_{ijk}) = \sigma_b^2 + \sigma_c^2 + \sigma_e^2$$

Now consider a pair of genes $Y_{i'jk'}$ and $Y_{i''jk''}$ sharing the same effect b_j. Then, their covariance is given by

$$\begin{aligned}
\text{Cov}(Y_{i'jk'}, Y_{i''jk''}) &= \text{Cov}(\mu + \alpha_{i'} + b_j + c_{k'} + \varepsilon_{i'jk'}, \mu + \alpha_{i''} + b_j + c_{k''} + \varepsilon_{i''jk''}) \\
&= \text{Cov}(b_j, b_j) \\
&= \sigma_b^2
\end{aligned}$$

Thus,

$$\text{Cor}(Y_{i'jk'}, Y_{i''jk''}) = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_c^2 + \sigma_e^2}$$

We can interpret the variance partition above as the contribution of the effect b to the observed correlation. Likewise, for two genes sharing the same c_k

$$\text{Cor}(Y_{i'j'k}, Y_{i''j''k}) = \frac{\sigma_c^2}{\sigma_b^2 + \sigma_c^2 + \sigma_e^2}$$

can be interpreted as the contribution of the effect c to the observed correlation.

We estimated the variance terms for the random effects with restricted maximum likelihood[90] to describe the influence of the sources of variation on the observed correlation.

Time Course Model

$$Y_{ijk} = \mu + \alpha_i + b_j + c_k + \varepsilon_{ijk}$$

where i indexes the genes, j indexes the tissues, k indexes the time point, α_i is the fixed gene effect, $b_j \sim N(0, \sigma_b^2)$ is the random effect for tissue type, $c_k \sim N(0, \sigma_c^2)$ is the random effect for time, and $\varepsilon_{ijk} \sim N(0, \sigma_e^2)$ is the error term. For a given pair of genes, we estimated the contribution of tissue type to the observed correlation as

$$\frac{\hat{\sigma}_b^2}{\hat{\sigma}_b^2 + \hat{\sigma}_c^2 + \hat{\sigma}_e^2}$$

and the contribution of time as

$$\frac{\hat{\sigma}_c^2}{\hat{\sigma}_b^2 + \hat{\sigma}_c^2 + \hat{\sigma}_e^2}$$

Liver Biological Replicates Model

$$Y_{ijkl} = \mu + \alpha_i + b_j + c_k + \varepsilon_{ijkl}$$

where i indexes the genes, j indexes the strain, k indexes the sex, l indexes the replicates, α_i is the fixed gene effect, $b_j \sim N(0, \sigma_b^2)$ is the random effect for strain, $c_k \sim N(0, \sigma_c^2)$ is the random effect for sex,

and $\varepsilon_{ijkl} \sim N(0, \sigma_e^2)$ is the error term. For a given pair of genes, we estimated the contribution of strain to the observed correlation as

$$\frac{\hat{\sigma}_b^2}{\hat{\sigma}_b^2 + \hat{\sigma}_c^2 + \hat{\sigma}_e^2}$$

and the contribution of sex as

$$\frac{\hat{\sigma}_c^2}{\hat{\sigma}_b^2 + \hat{\sigma}_c^2 + \hat{\sigma}_e^2}$$

Tissue Panel Model

$$Y_{ijkl} = \mu + \alpha_i + b_j + c_k + d_l + \varepsilon_{ijkl}$$

where i indexes the genes, j indexes the tissues, k indexes the strain, l indexes the sex, α_i is the fixed gene effect, $b_j \sim N(0, \sigma_b^2)$ is the random effect for tissue type, $c_k \sim N(0, \sigma_c^2)$ is the random effect for strain, $d_l \sim N(0, \sigma_d^2)$ is the random effect for sex, and $\varepsilon_{ijkl} \sim N(0, \sigma_e^2)$ is the error term. For a given pair of genes, we estimated the contribution of tissue type to the observed correlation as

$$\frac{\hat{\sigma}_b^2}{\hat{\sigma}_b^2 + \hat{\sigma}_c^2 + \hat{\sigma}_d^2 + \hat{\sigma}_e^2}$$

the contribution of strain as

$$\frac{\hat{\sigma}_c^2}{\hat{\sigma}_b^2 + \hat{\sigma}_c^2 + \hat{\sigma}_d^2 + \hat{\sigma}_e^2}$$

and the contribution of sex as

$$\frac{\hat{\sigma}_d^2}{\hat{\sigma}_b^2 + \hat{\sigma}_c^2 + \hat{\sigma}_d^2 + \hat{\sigma}_e^2}$$

3.4.7 Linear Model Framework for Effect Estimates

Linear models have been used to model the effect from the different batches[109,108]. We use this framework to estimate relationships between genes in a particular context (e.g. across time, between different tissue types).

Time Course Model

$$Y_{ijk} = \alpha_i + \beta_{ij} + \gamma_{ik} + \varepsilon_{ijk}$$

where i indexes the genes, j indexes the tissues, k indexes the time point, α_i is the gene-specific mean, β_{ij} is the gene-specific tissue effect, γ_{ik} is the gene-specific time effect, and $\varepsilon_{ijk} \sim N(0, \sigma^2)$ is the error term.

Liver Biological Replicates Model

$$Y_{ijkl} = \alpha_i + \beta_{ij} + \gamma_{ik} + \varepsilon_{ijkl}$$

where i indexes the genes, j indexes the strains, k indexes the sex, l indexes the replicates, α_i is the gene-specific mean, β_{ij} is the gene-specific strain effect, γ_{ik} is the gene-specific sex effect, and $\varepsilon_{ijkl} \sim \mathcal{N}(0, \sigma^2)$ is the error term.

Tissue Panel Model

$$Y_{ijkl} = \alpha_i + \beta_{ij} + \gamma_{ik} + \lambda_{il} + \varepsilon_{ijkl}$$

where i indexes the genes, j indexes the tissues, k indexes the strain, l indexes the sex, α_i is the gene-specific mean, β_{ij} is the gene-specific tissue effect, γ_{ik} is the gene-specific strain effect, λ_{il} is the gene-specific sex effect, and $\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$ is the error term.

3.5 Coexpression Networks

We generated the coexpression network from the correlation matrices using the function `adjacency.fromSimilarity` from the R package `WGCNA` v.1.63. The edges between a pair of genes i and j in the coexpression net-

work are given by

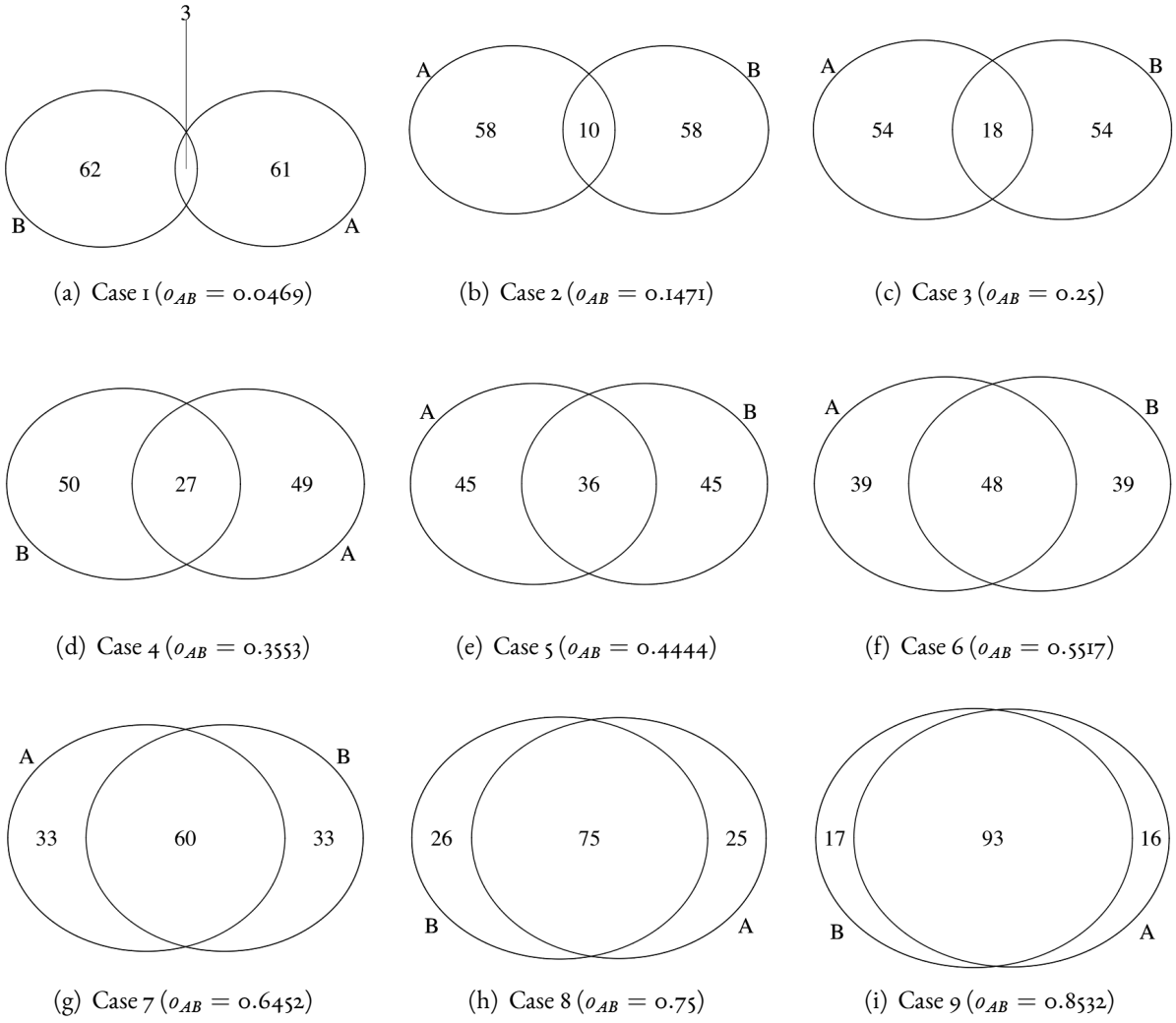
$$|r_{ij}|$$

where r_{ij} is the correlation between genes i and j. We used the default value of β ($\beta = 6$) to generate the coexpression network.

A

Supplementary Materials for Chapter 1

A.1 Supplementary Figures



(a) Case 1 ($o_{AB} = 0.0469$) (b) Case 2 ($o_{AB} = 0.1471$) (c) Case 3 ($o_{AB} = 0.25$)

(d) Case 4 ($o_{AB} = 0.3553$) (e) Case 5 ($o_{AB} = 0.4444$) (f) Case 6 ($o_{AB} = 0.5517$)

(g) Case 7 ($o_{AB} = 0.6452$) (h) Case 8 ($o_{AB} = 0.75$) (i) Case 9 ($o_{AB} = 0.8532$)

Figure A.1. Venn Diagrams for No Overlap and Overlap Cases of the Ribosome Gene Sets. Venn diagrams between the partitions used to separate the (KEGG) Ribosome pathway and random genes into gene sets with different degrees of shared genes.

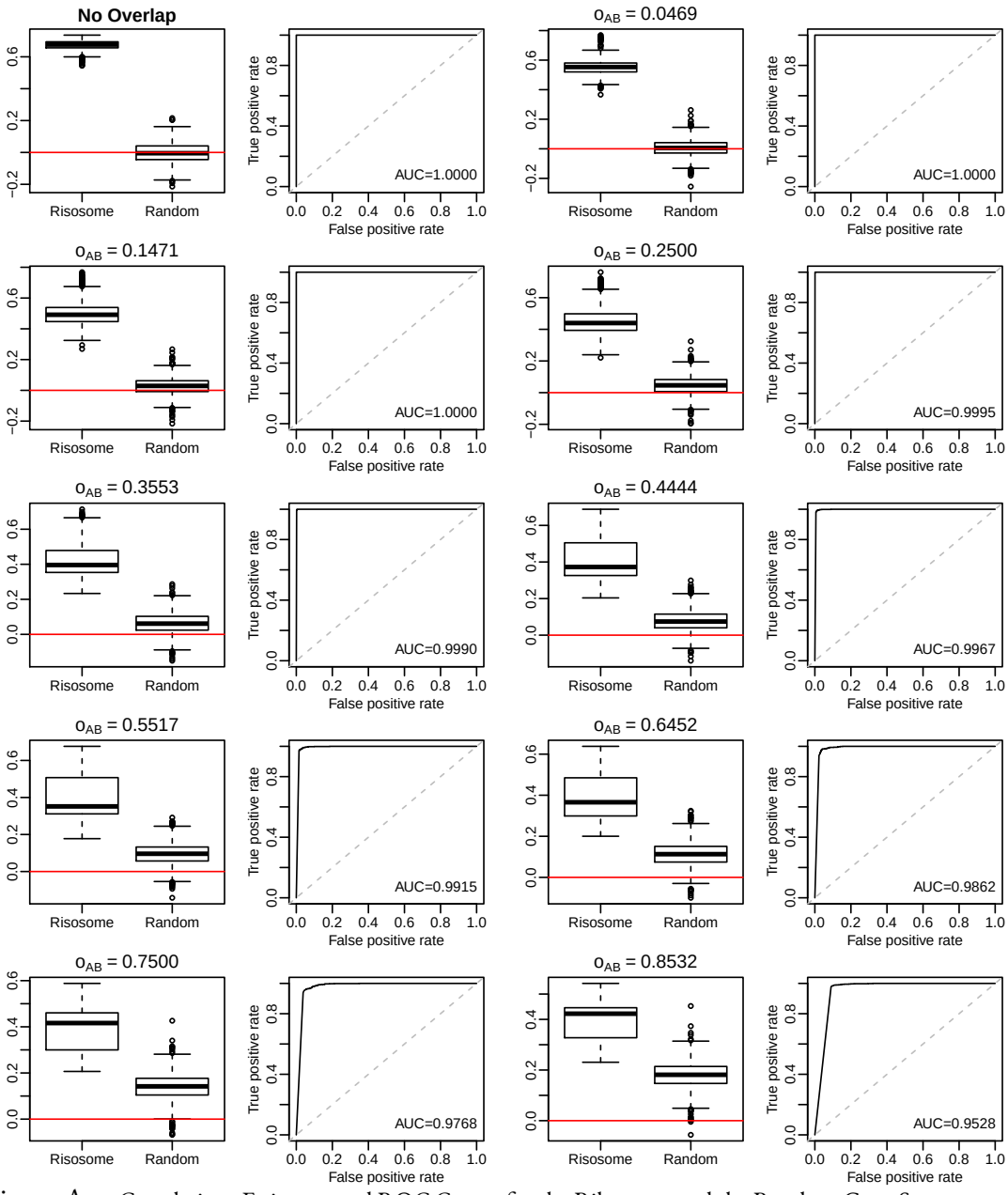


Figure A.2. Correlations Estimates and ROC Curves for the Ribosome and the Random Gene Sets.
Boxplots of the correlation estimates between the ribosome gene sets and random gene sets under different cases of gene overlap. ROC curves for the ribosome gene sets and random gene sets under different cases of shared genes. We assume that a significant p-value ($p < 0.05$) for a ribosome gene set is a true positive, while a significant p-value for a random gene set is a false positive.

A.2 Pathway Summary Statistic

In PCxN we use the mean rank as the pathway summary statistic. We considered using the projection into the first principal component as the summary statistic. However, the variance explained by the first component was low in most of the curated experiments from normal human tissues. For each experiment from the curated collection, we estimated the percentage of variance explained by the first principal component for each of the 1,330 canonical pathways from the MSigDB CP v 5.1 collection. The barplot below shows the proportion of pathway for which the first principal component explains more than 50% of the variance.

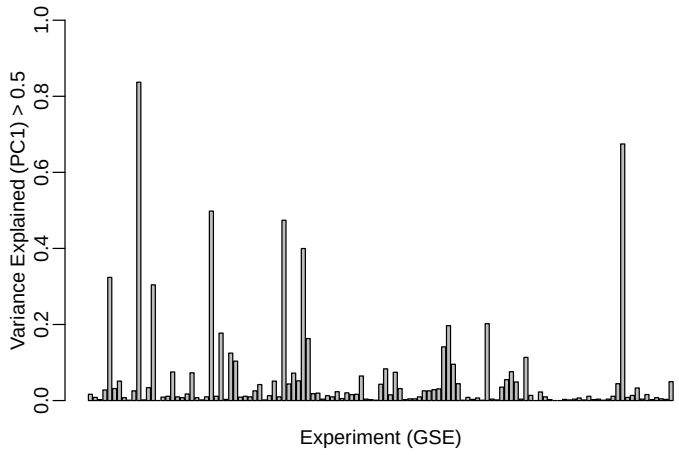


Figure A.3. Proportion of canonical pathways for which the first principal component explains more than 50% of the variance across all experiments.

A.3 Impact of Gene Overlap (GO:BP)

We compared the number of significantly correlated pathways with the number of pathways which share a significant number of genes according to Fisher's exact test for the MSigDB GO:BP v 5.1 gene

set collection.

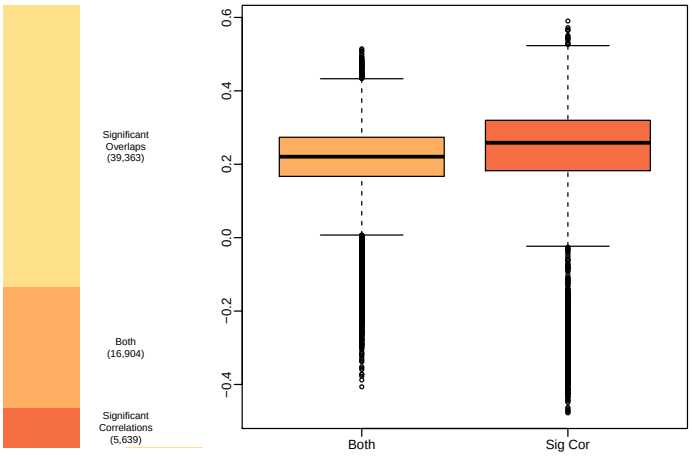


Figure A.4. The stacked bar plot shows the number of pathways pairs with only significant correlations in red, with only significant overlaps in yellow, and with both in orange. The boxplots show the distribution of the correlation coefficients with pathway pairs with only significant correlations (red) and with both significant overlaps and significant correlations (orange).

The results for the GO:BP gene set collection are similar to the results that we reported for the Canonical Pathways in Figure 2D. We also observed more significant overlaps than significant correlations, and about 30% of the pathway pairs have both significant overlap and significant correlations. For pathway pairs with both significant overlaps and significant correlations, the correlations are lower on average that for pathway pairs with significant correlations only.

A.4 Robustness of the Correlation Estimates

We used Jackknife statistics to estimate the bias of the aggregated correlation estimates to assess the robustness of the coexpression network. The Jackknife statistics corresponding to the weighted

average of the experiment-level correlation estimated leaving out one experiment at a time

$$\bar{r}_{(k)} = \frac{\sum_{i \neq k} n_i r_i}{n_i}$$

where r_i is the correlation estimate for experiment i, and n_i is the number of samples in experiment i.

The estimate for the bias is given by

$$\widehat{\text{Bias}} = (N-1)(\bar{r}_{(\cdot)} - \bar{r})$$

where N is the total number of experiments, $\bar{r}$ is the aggregated correlation estimate, and

$$\bar{r}_{(\cdot)} = \frac{1}{N} \sum_{k=1}^{N} \bar{r}_{(k)}$$

The estimate for the bias reflects the influence of each experiment i from the curated collection of normal human tissues on the aggregated correlation estimate $\bar{r}$.

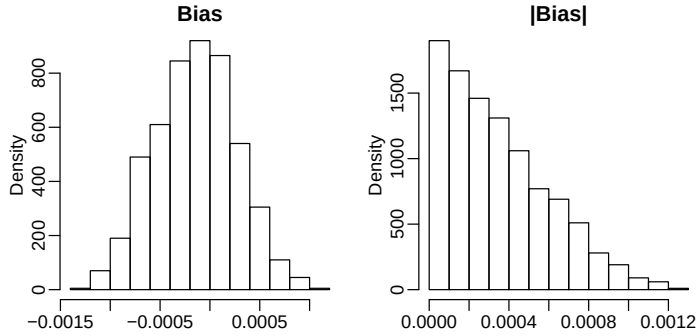


Figure A.5. Histogram for the bias estimates (left), and the magnitude of the bias (right).

The overall bias is very low (mean -0.0001462, median -0.0001378, mean magnitude 0.0003480, median magnitude 0.0002962). The small bias of the correlation estimates of PCxN, demonstrate the robustness of the pathway coexpression network.

A.5 GSM and GSE Accessions of Gene Expression Data.

https://zenodo.org/record/1214588/files/pcbi.1006042.s004.xlsx

A.6 Canonical Pathways Annotation.

https://zenodo.org/record/1214588/files/pcbi.1006042.s005.xlsx

A.7 Gene Overlap and Correlation Estimates for the Canonical Pathways in Fig 1.2B–G.

https://zenodo.org/record/1214588/files/pcbi.1006042.s006.xlsx

A.8 Alzheimer's Disease Curated List.

https://zenodo.org/record/1214588/files/pcbi.1006042.s007.docx

Domain expert curated list of genes associated with Alzheimer's disease identified via genome wide association studies (GWAS).

A.9 Canonical Pathways Correlated with the Alzheimer's Disease Curated List, and Canonical Pathways Enriched for Genes within the Alzheimer's Disease Curated List.

https://zenodo.org/record/1214588/files/pcbi.1006042.s008.xlsx

A.10 Genes Associated with Alzheimer's Disease from the Genetic Association Database.

https://zenodo.org/record/1214588/files/pcbi.1006042.s009.xlsx

A.11 Canonical Pathways Enriched for Genes Associated with Alzheimer's Disease from the Genetic Association Database.

https://zenodo.org/record/1214588/files/pcbi.1006042.s010.xlsx

A.12 Results from Gene Set Enrichment Analysis on an Alzheimer's Disease Profiling Experiment.

https://zenodo.org/record/1214588/files/pcbi.1006042.s011.xlsx

A.13 GEO Accessions for the Alzheimer's Disease Profiling Experiment.

https://zenodo.org/record/1214588/files/pcbi.1006042.s012.xlsx

A.14 Correlations between Canonical Pathways identified as Enriched by Gene Set Enrichment Analysis and Canonical Pathways Correlated with Pathways identified as Enriched by Gene Set Enrichment Analysis.

https://zenodo.org/record/1214588/files/pcbi.1006042.s013.xlsx

B

Supplementary Materials for Chapter 2

B.1 Supplementary Figures

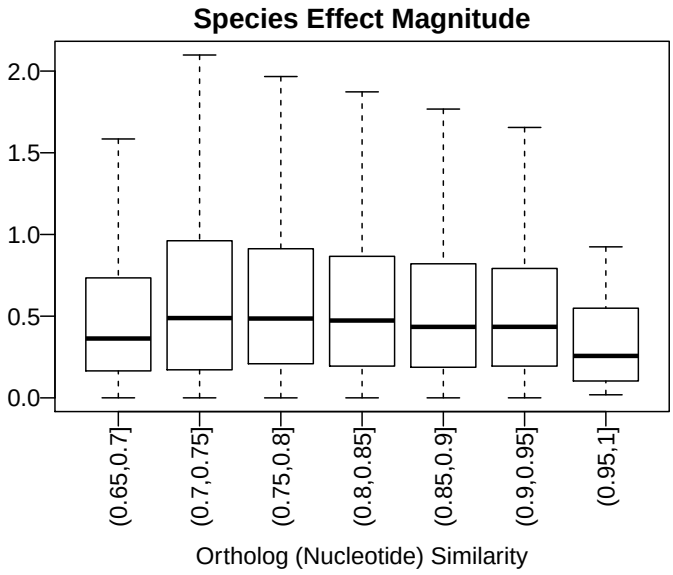


Figure B.1. Species Effect Magnitude vs. Nucleotide Similarity. The species effect is the observed difference in means between the human and mouse normlized FPKM values from the resequenced ENCODE RNA-seq data set. Absolute value of the species effect binned by the nucleotide similarity between the human-mouse orthologs.

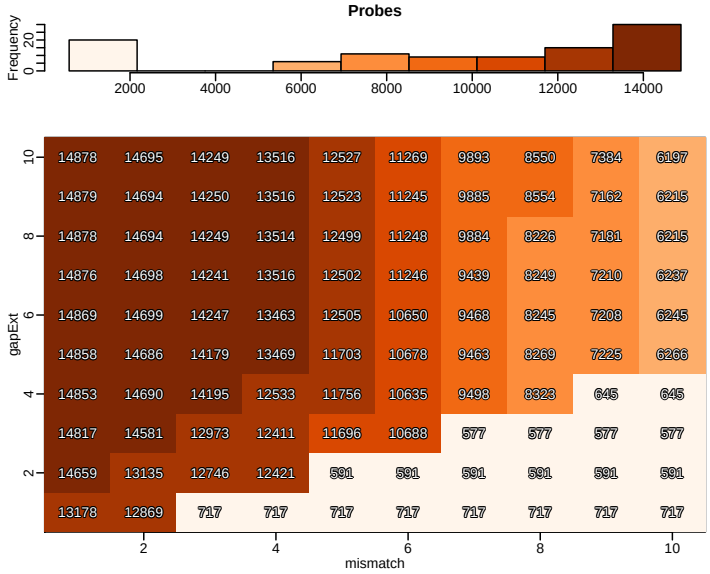


Figure B.2. Alignment Parameters and Number of Probes. Median number of valid alignments for each set of alignment parameters, `gapExt` (δ) and `mismatch` (μ). (Top) Histogram with the color palette and (bottom) heat map for the number of valid alignments.

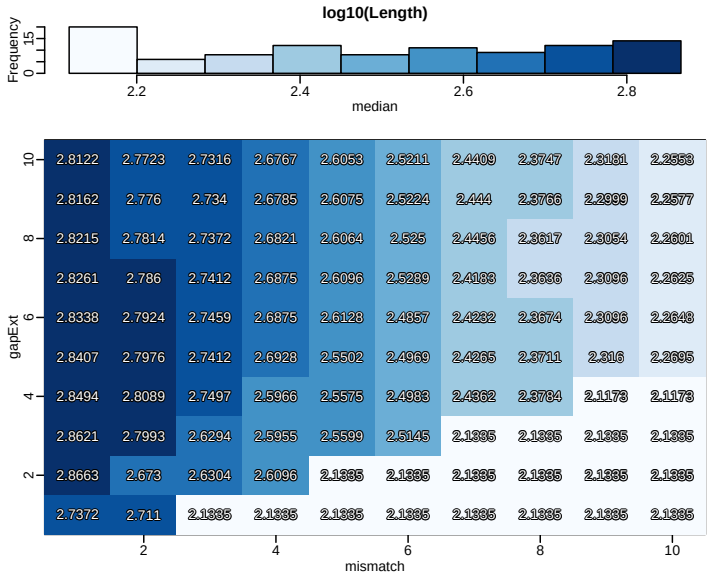


Figure B.3. Alignment Parameters and Probe Length. Median alignment length for each set of alignment parameters, `gapExt` (δ) and `mismatch` (μ). (Top) Histogram with the color palette and (bottom) heat map for the alignment length.

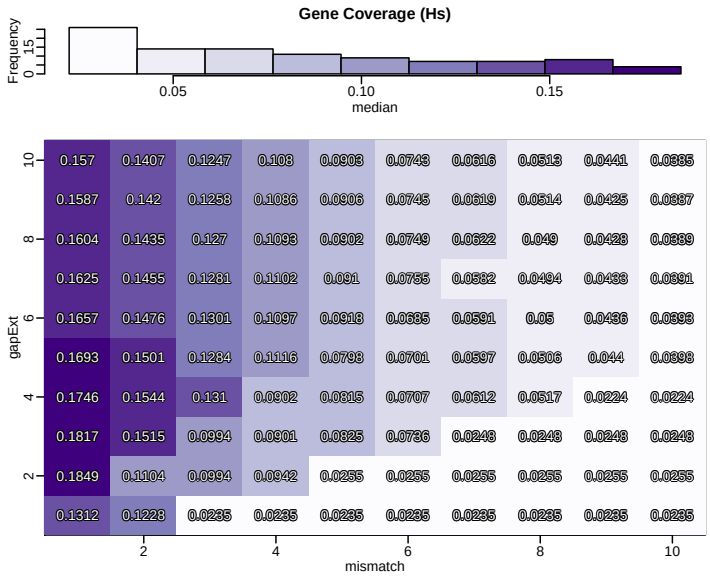


Figure B.4. Alignment Parameters and Human Ortholog Coverage. Median proportion of the human gene present for each set of alignment parameters, `gapExt` (δ) and `mismatch` (μ). (Top) Histogram with the color palette and (bottom) heat map for the proportion of the human gene present.

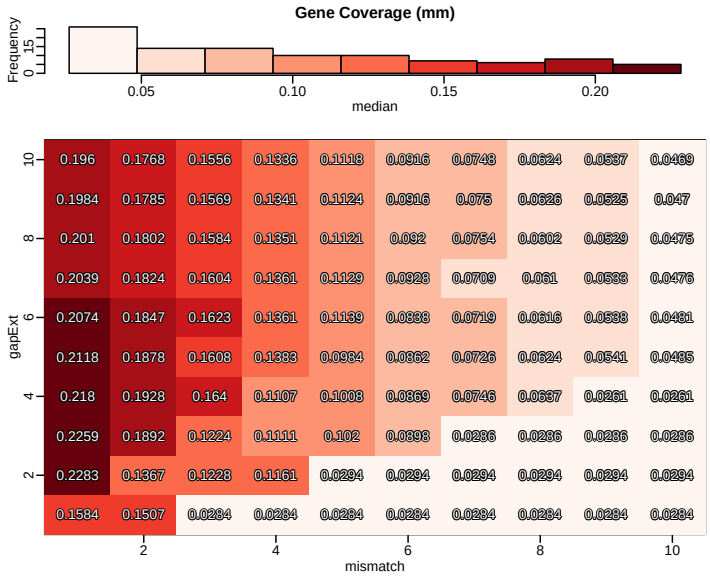


Figure B.5. Alignment Parameters and Mouse Ortholog Coverage. Median proportion of the mouse gene present for each set of alignment parameters, `gapExt` (δ) and `mismatch` (μ). (Top) Histogram with the color palette and (bottom) heat map for the proportion of the mouse gene present.

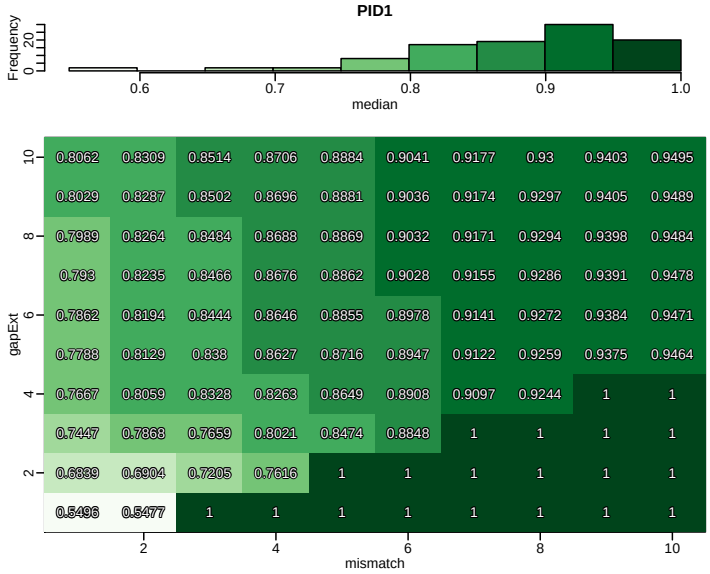


Figure B.6. Alignment Parameters and Probe Similarity (PID1). Median probe similarity (PID1) for each set of alignment parameters, `gapExt` (δ) and `mismatch` (μ). (Top) Histogram with the color palette and (bottom) heat map for the probe similarity (PID1).

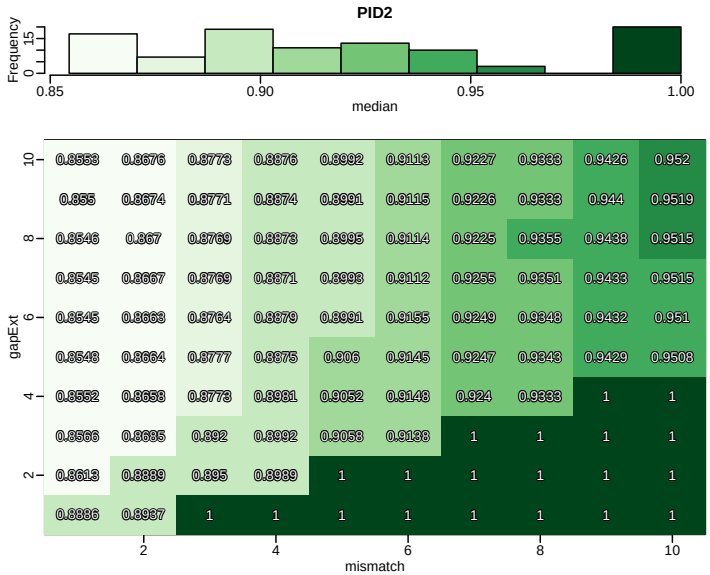


Figure B.7. Alignment Parameters and Probe Similarity (PID2). Median probe similarity (PID2) for each set of alignment parameters, `gapExt` (δ) and `mismatch` (μ). (Top) Histogram with the color palette and (bottom) heat map for the probe similarity (PID2).

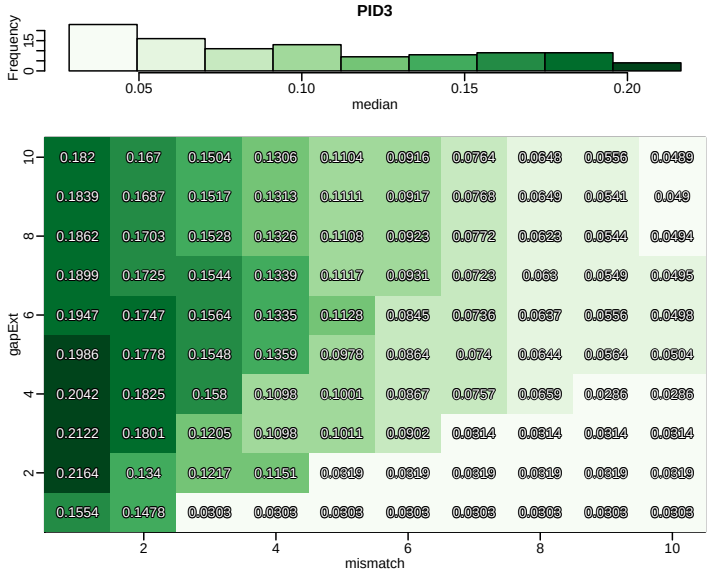


Figure B.8. Alignment Parameters and Probe Similarity (PID3). Median probe similarity (PID3) for each set of alignment parameters, `gapExt` (∂) and `mismatch` (μ). (Top) Histogram with the color palette and (bottom) heat map for the probe similarity (PID3).

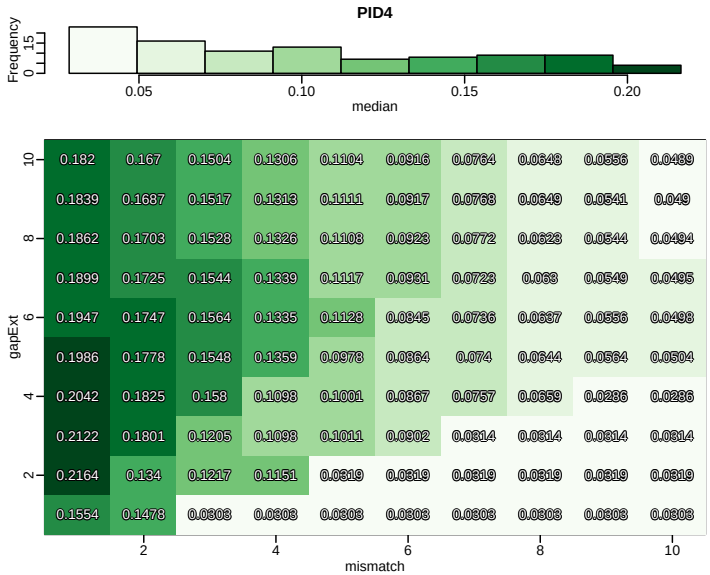


Figure B.9. Alignment Parameters and Probe Similarity (PID4). Median probe similarity (PID4) for each set of alignment parameters, `gapExt` (∂) and `mismatch` (μ). (Top) Histogram with the color palette and (bottom) heat map for the probe similarity (PID4).

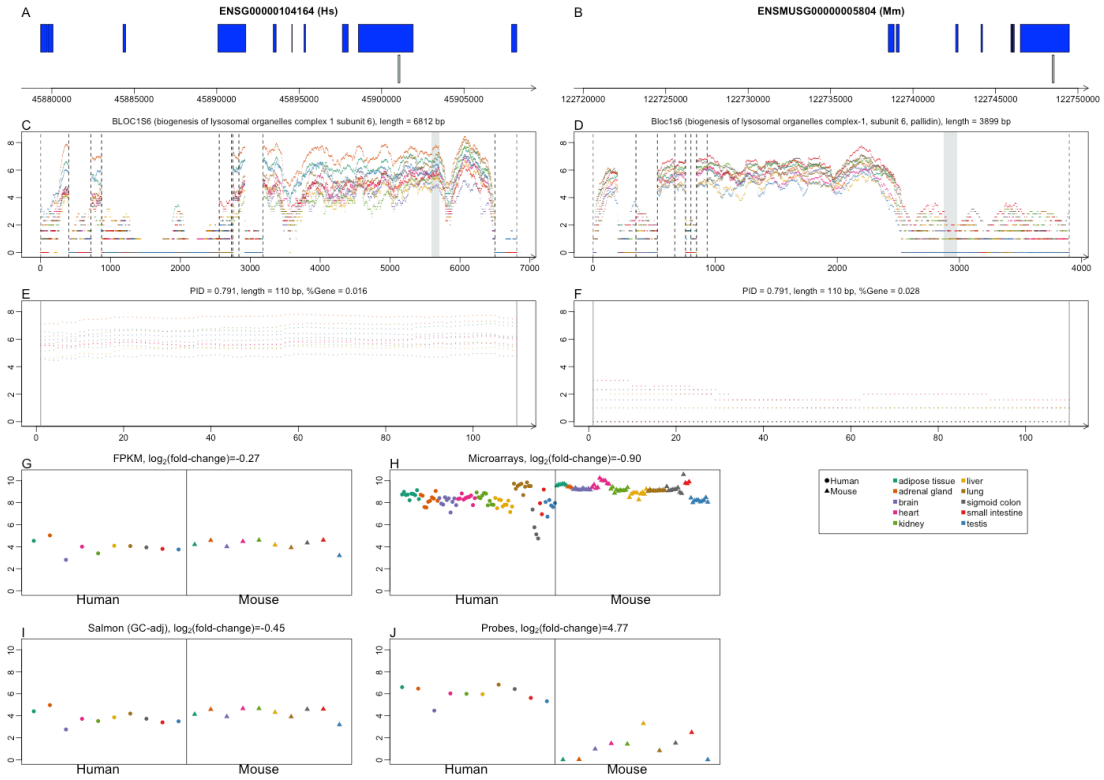


Figure B.10. Ortholog Probe for BLOC1S6/Bloc1S6. Diagram for the union of transcript exons of the human (A) and mouse (B) genes. The regions in gray below the exons correspond to the ortholog probe. Coverage plots for the union of transcript exons excluding the intronic regions for the human (C) and mouse (D) genes. The dashed lines are the exon boundaries and the shaded areas correspond to the ortholog probes. Coverage lots for the human (E) and mouse (F) ortholog probes. Scatter plots for the normalized FPKM values (G), the normalized Salmon estimates corrected for GC-content bias (I), and the normalized probe values (J) for the resequenced ENCODE RNA-seq data set. Scatter plot for the normalized Barcode 3.0 microarray values.

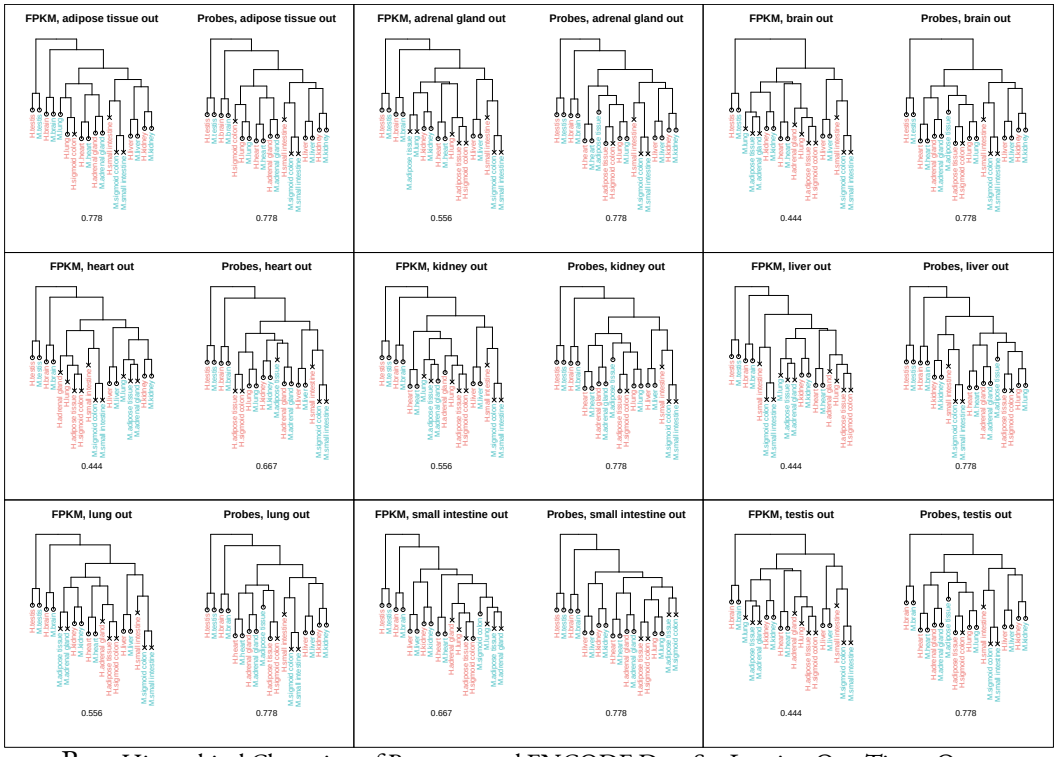


Figure B.11. Hierarchical Clustering of Resequenced ENCODE Data Set Leaving One Tissue Out. Dendrogram for the resequenced ENCODE RNA-seq data set based on hierarchical clustering with complete linkage and Euclidean distance leaving one tissue type out at a time. The dendrogram on the right (FPKM) is based on the normalized FPKM values, and the dendrogram on the right (Probes) is based on the normalized ortholog probe counts. The circles (○) correspond to pairs of human and mouse tissues where the type match, while the crosses (×) correspond to pairs of human and mouse tissues where the type does not match. The numbers at the bottom of each dendrogram are the proportion of human and mouse tissue pairs where the tissue type matches.

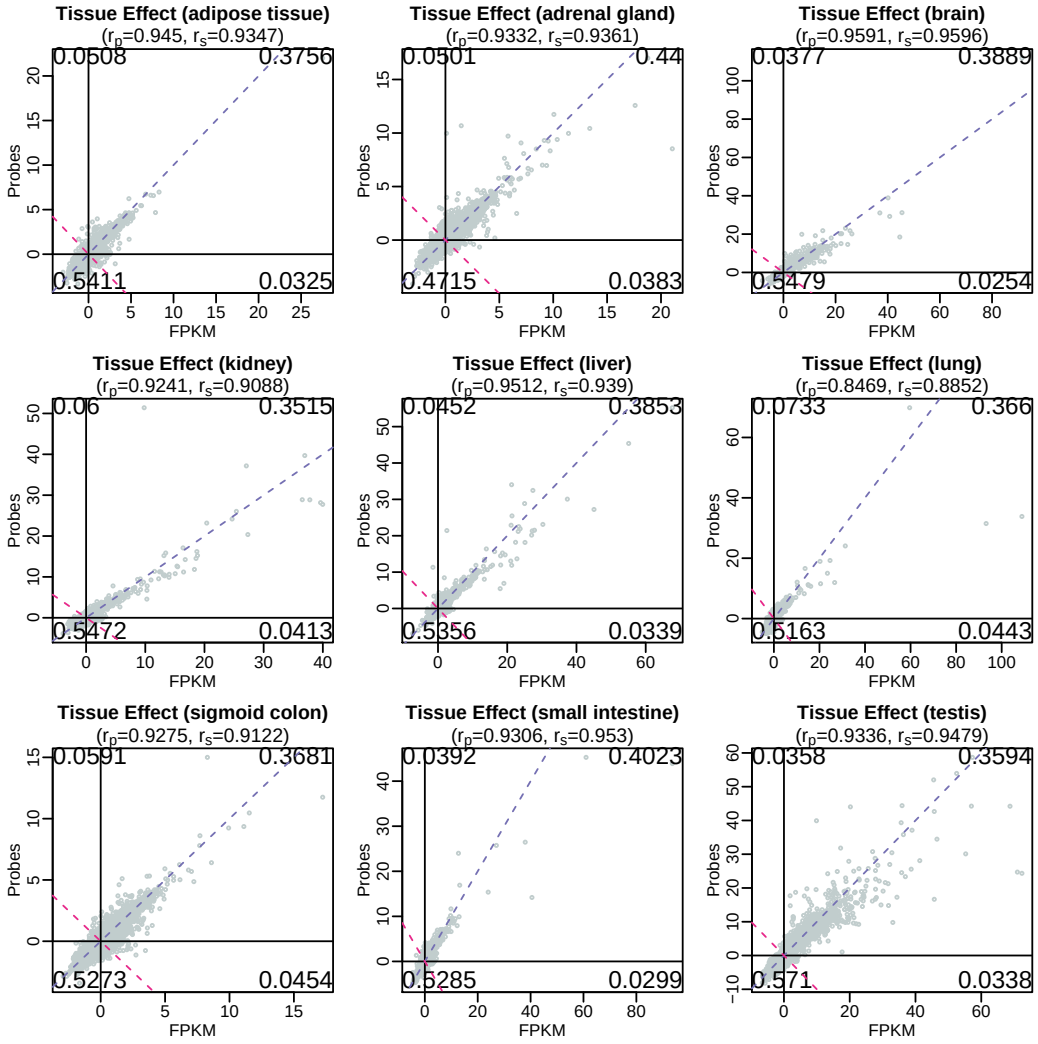


Figure B.12. Tissue T-statistics Comparison between Probes and FPKM. The tissue t-statistics are the t-statistic comparing the observed difference in means between the human and mouse samples from a given tissue type and the rest of the human and mouse tissues. Comparison between the tissue t-statistics from the FPKM and the probe normalized values. The purple dashed line is $x = y$, and the pink dashed line is $x = -y$. The number in each corner indicates the proportion of points in each quadrant defined by $x = 0$ and $y = 0$.

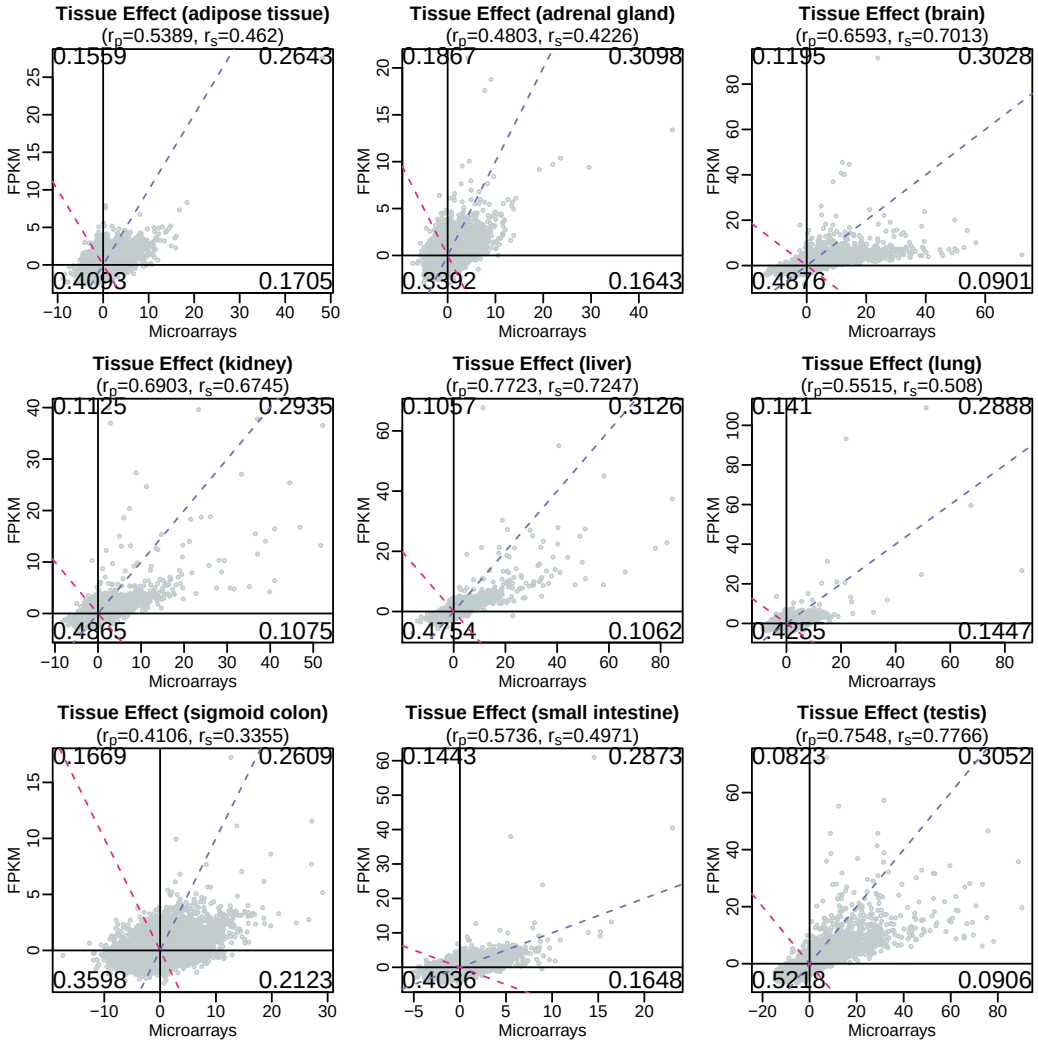


Figure B.13. Tissue T-statistics Comparison between RNA-seq (FPKM) and Microarrays. The tissue t-statistics are the t-statistic comparing the observed difference in means between the human and mouse samples from a given tissue type and the rest of the human and mouse tissues. Comparison between the tissue t-statistics from the FPKM and the microarray normalized values. The purple dashed line is $x = y$, and the pink dashed line is $x = -y$. The number in each corner indicates the proportion of points in each quadrant defined by $x = 0$ and $y = 0$.

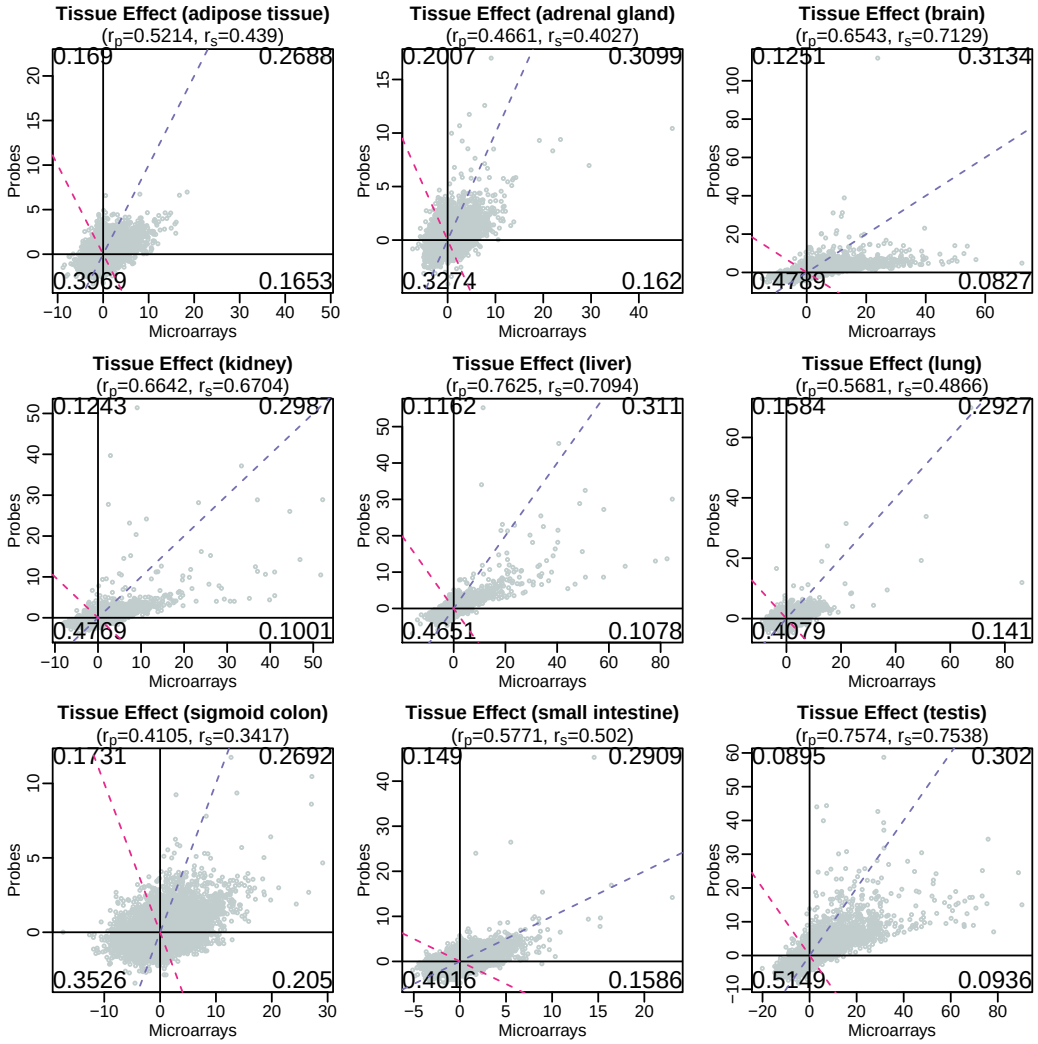
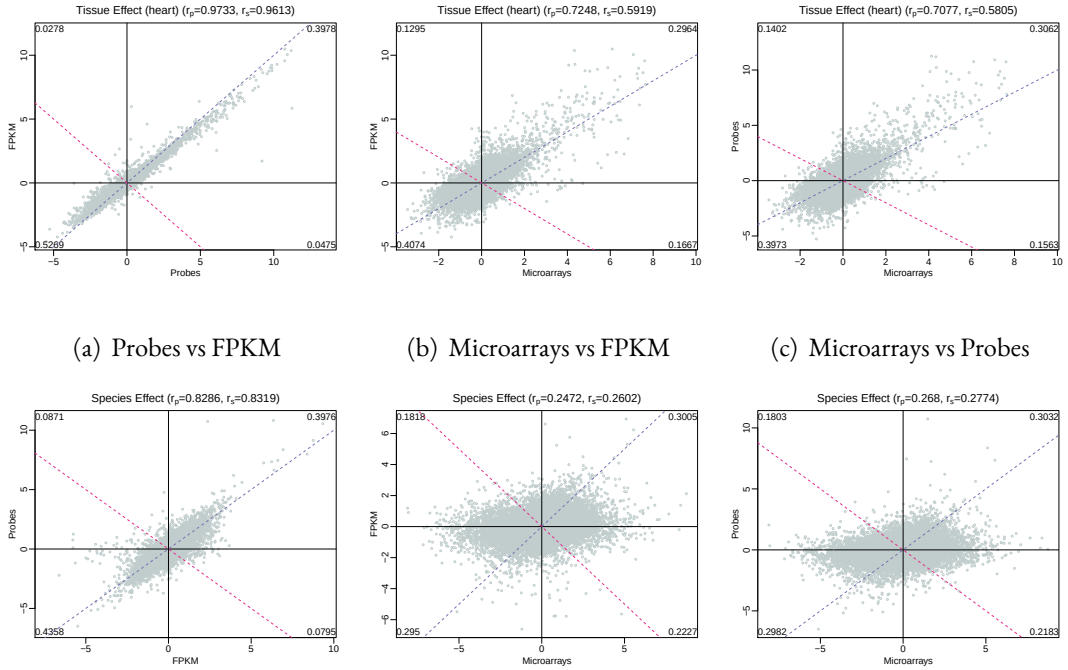


Figure B.14. Tissue T-statistics Comparison between RNA-seq (Probes) and Microarrays. The tissue t-statistics are the t-statistic comparing the observed difference in means between the human and mouse samples from a given tissue type and the rest of the human and mouse tissues. Comparison between the tissue t-statistics from the probe and microarray normalized values. The purple dashed line is $x = y$, and the pink dashed line is $x = -y$. The number in each corner indicates the proportion of points in each quadrant defined by $x = 0$ and $y = 0$.



(a) Probes vs FPKM (b) Microarrays vs FPKM (c) Microarrays vs Probes

(d) Probes vs FPKM (e) Microarrays vs FPKM (f) Microarrays vs Probes

Figure B.15. **Species and Tissue Effect Comparison.** The species effect is the observed difference in means between the human and mouse normalized expression values from the resequenced ENCODE RNA-seq data set. In this case, the tissue effect for heart is the observed difference in means between the human and mouse heart samples and the rest of the human and mouse tissues. Comparison between the tissue effect estimates from the (a) FPKM and the probe normalized values, the (b) FPKM and the microarray normalized values, and the (c) probe and microarray normalized values. Comparison between the species effect estimates from the (d) FPKM and the probe normalized values, the (e) FPKM and the microarray normalized values, and the (f) probe and microarray normalized values. The purple dashed line is $x = y$, and the pink dashed line is $x = -y$. The number in each corner indicates the proportion of points in each quadrant defined by $x = 0$ and $y = 0$.

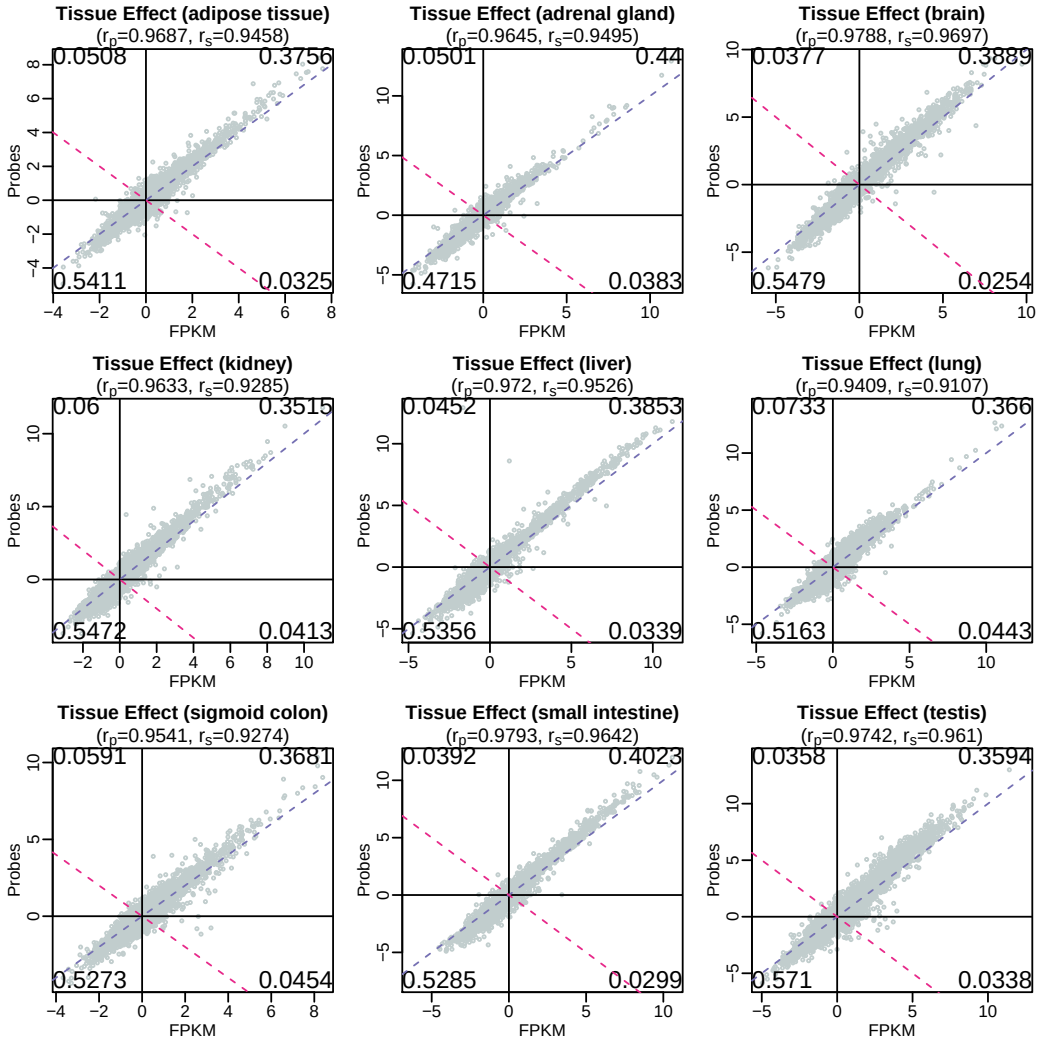


Figure B.16. Tissue Effect Comparison between Probes and FPKM. The tissue effect is the observed difference in means between the human and mouse samples from a given tissue type and the rest of the human and mouse tissues. Comparison between the tissue effect estimates from the FPKM and the probe normalized values. The purple dashed line is $x = y$, and the pink dashed line is $x = -y$. The number in each corner indicates the proportion of points in each quadrant defined by $x = 0$ and $y = 0$.

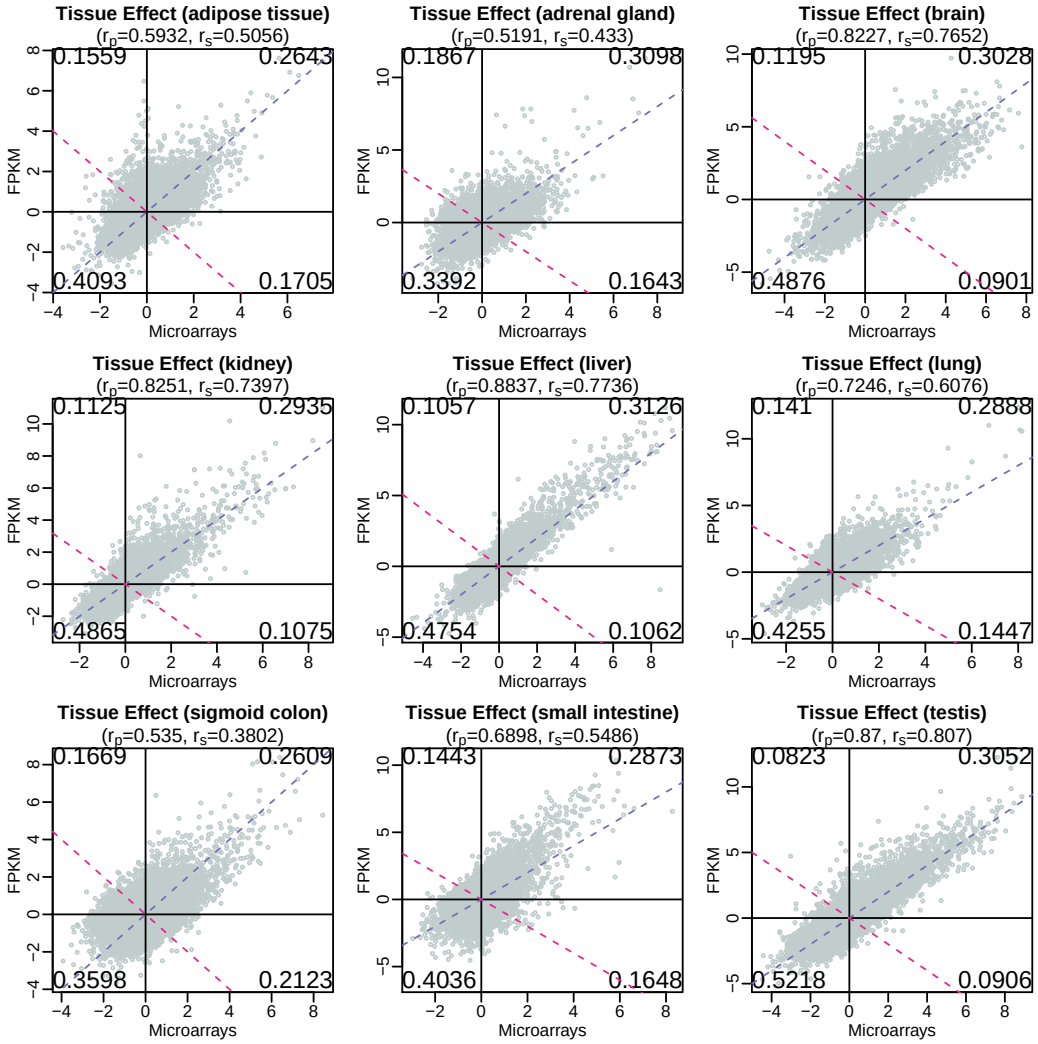


Figure B.17. Tissue Effect Comparison between RNA-seq (FPKM) and Microarrays. The tissue effect is the observed difference in means between the human and mouse samples from a given tissue type and the rest of the human and mouse tissues. Comparison between the tissue effect estimates from the FPKM and the microarray normalized values. The purple dashed line is $x = y$, and the pink dashed line is $x = -y$. The number in each corner indicates the proportion of points in each quadrant defined by $x = 0$ and $y = 0$.

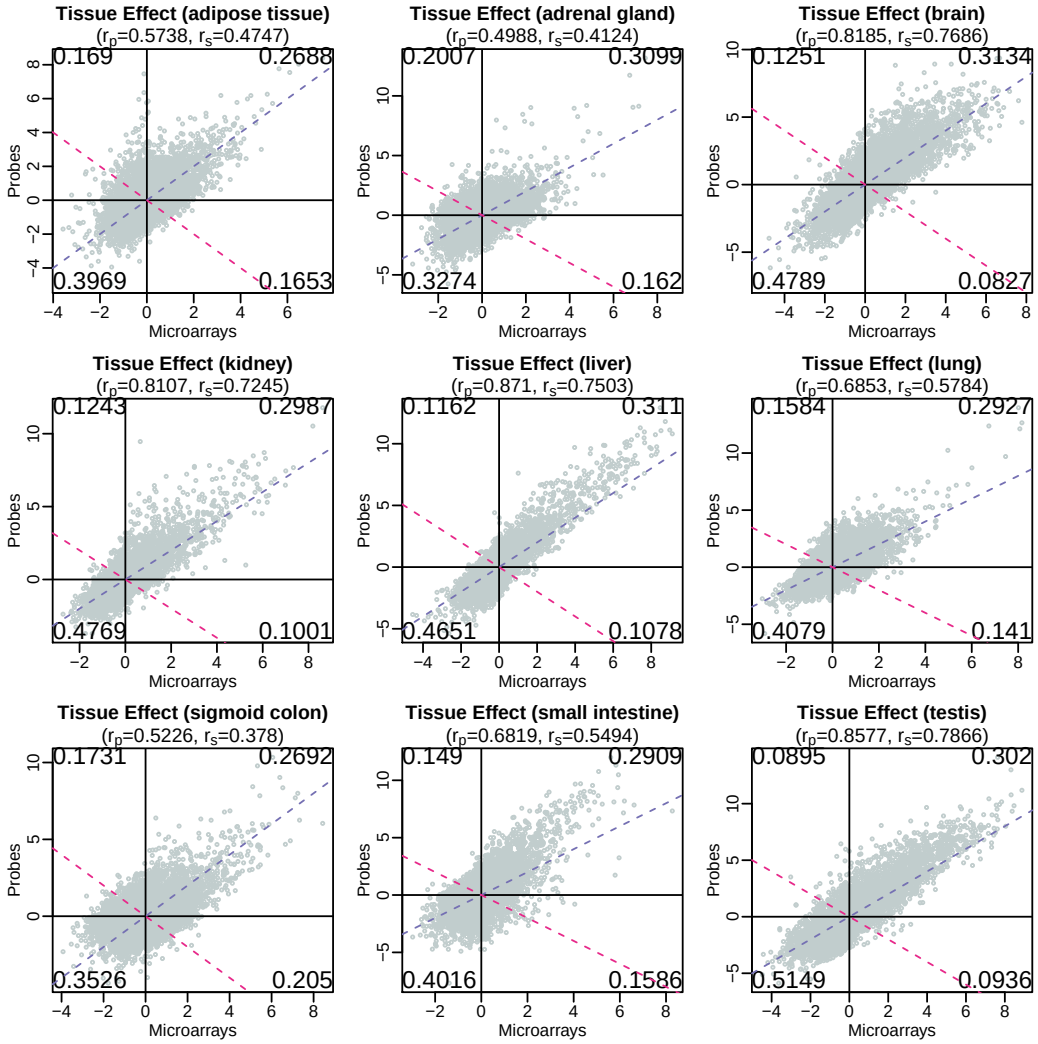


Figure B.18. Tissue Effect Comparison between RNA-seq (Probes) and Microarrays. The tissue effect is the observed difference in means between the human and mouse samples from a given tissue type and the rest of the human and mouse tissues. Comparison between the tissue effect estimates from the probe and microarray normalized values. The purple dashed line is $x = y$, and the pink dashed line is $x = -y$. The number in each corner indicates the proportion of points in each quadrant defined by $x = 0$ and $y = 0$.

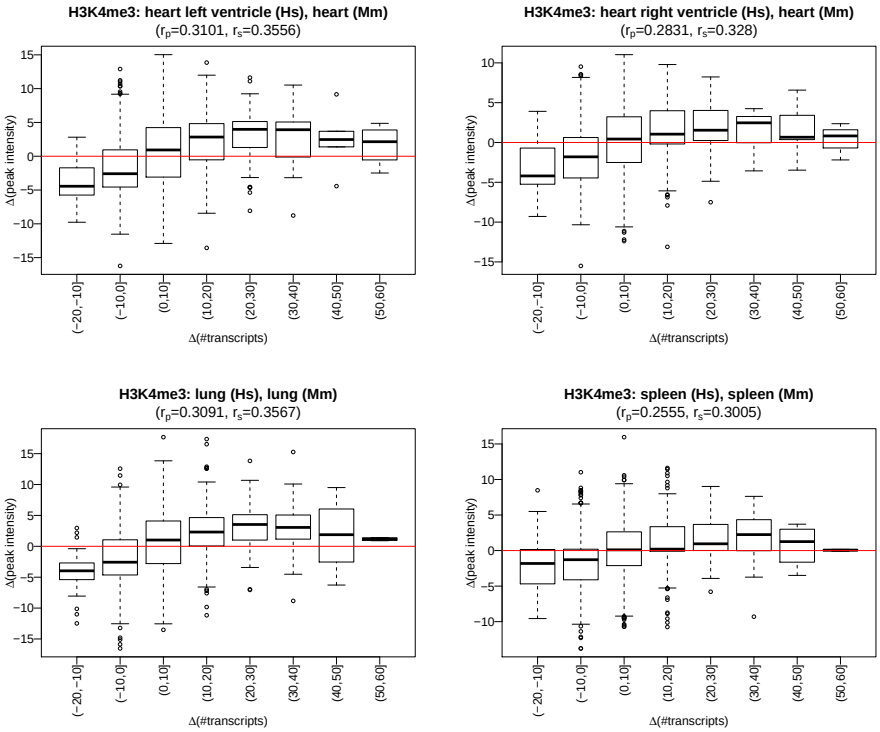


Figure B.19. **Differences between H3K4me3 Peak Intensities vs. Differences between the Number of Transcripts.** Differences between the gene-associated H3K4me3 peak intensities from the REMC and mouse ENCODE projects and the number of annotated transcripts for the human-mouse orthologs.

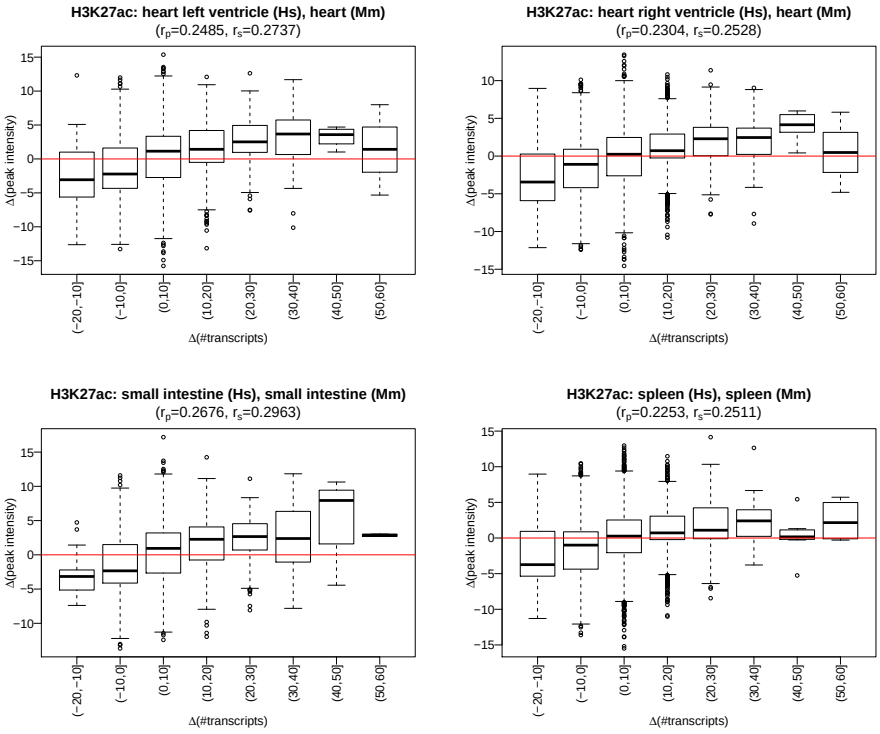


Figure B.20. **Differences between H3K27ac Peak Intensities vs. Differences between the Number of Transcripts.** Differences between the gene-associated H3K27ac peak intensities from the REMC and mouse ENCODE projects and the number of annotated transcripts for the human-mouse orthologs.

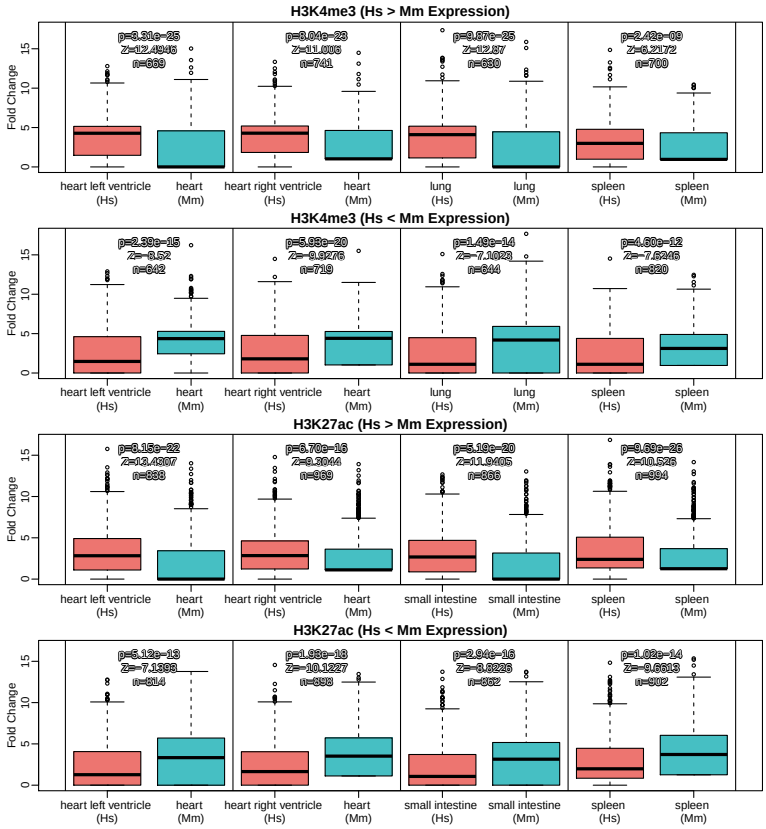


Figure B.21. Histone Peak Intensities for Differentially Expressed Orthologs in the ENCODE RNA-seq Data Set. Fold Enrichment over control of H3K4me3 and H3K27ac present at promoters of the differentially expressed orthologs based on the normalized FPKM values from the ENCODE RNA-seq data set. The differentially expressed orthologs are separated into orthologs where the gene expression is higher in the human tissues than in the mouse tissues (Hs $>$ Mm), and where the gene expression is higher in the mouse tissues than in the human tissues (Hs $<$ Mm). The p-values and Z-statistics were generated by the nonparametric paired Wilcoxon test between the human and mouse gene-associated histone peak intensities, n is the number of human-mouse orthologs where at least one of them has a gene-associated peak intensity.

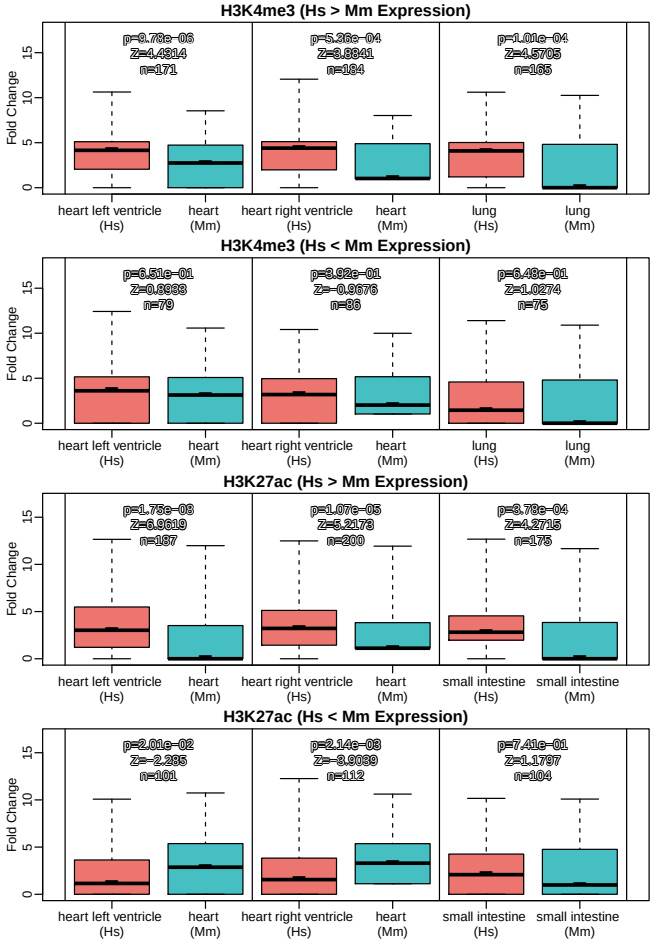


Figure B.22. Histone Peak Intensities for Differentially Expressed Orthologs in both the Resequenced ENCODE RNA-seq Data Set and the Microarrays from Barcode 3.0. Fold enrichment over control of H3K4me3 and H3K27ac present at promoters of the differentially expressed orthologs based on both the normalized probe values from the resequenced ENCODE RNA-seq data set and the normalized microarray values from Barcode 3.0. The differentially expressed orthologs are separated into orthologs where the gene expression is higher on average in the human tissues than in the mouse tissues (Hs $>$ Mm), and where the gene expression is higher on average in the mouse tissues than in the human tissues (Hs $<$ Mm). The p-values and Z-statistics were generated by the nonparametric paired Wilcoxon test between the human and mouse gene-associated histone peak intensities, n is the number of human-mouse orthologs where at least one of them has a gene-associated peak intensity.

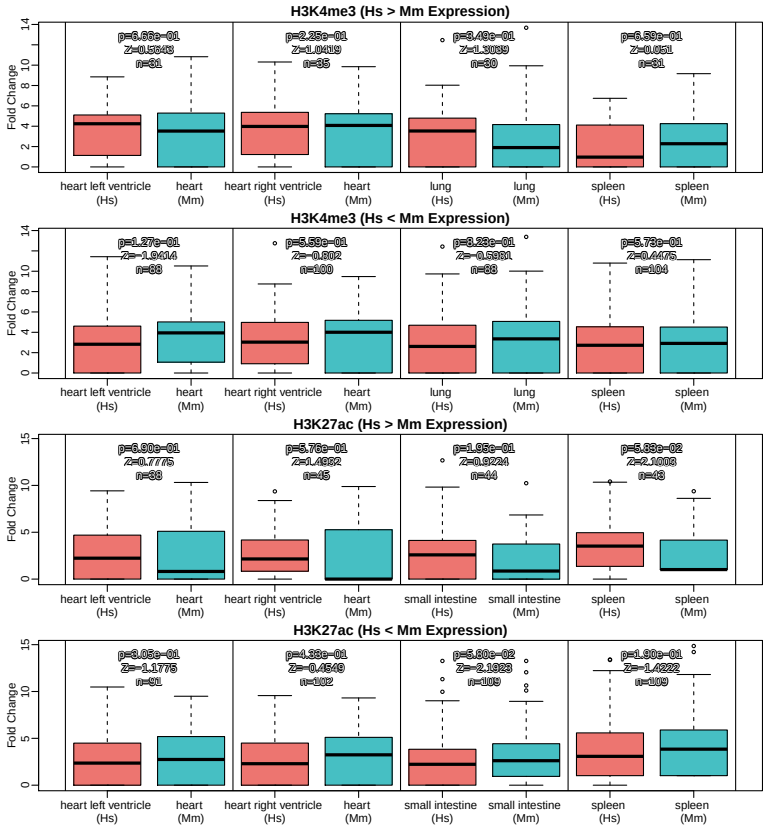


Figure B.23. Histone Peak Intensities for Differentially Expressed Orthologs in the ENCODE RNA-seq Data Set with the same Number of Annotated Transcripts. Fold Enrichment over control of H3K4me3 and H3K27ac present at promoters of the differentially expressed orthologs with the same number of transcripts based on the normalized FPKM values from the ENCODE RNA-seq data set. The differentially expressed orthologs are separated into orthologs where the gene expression is higher in the human tissues than in the mouse tissues (Hs $>$ Mm), and where the gene expression is higher in the mouse tissues than in the human tissues (Hs $<$ Mm). The p-values and Z-statistics were generated by the nonparametric paired Wilcoxon test between the human and mouse gene-associated histone peak intensities, n is the number of human-mouse orthologs where at least one of them has a gene-associated peak intensity.

B.2 Supplementary Tables

Table B.1. **ENCODE RNA-seq Data.**

Experiment accession	Species	Tissue
ENCSR612HYR	Homo sapiens	small intestine
ENCSR270OKS	Homo sapiens	sigmoid colon
ENCSR448VSW	Homo sapiens	spleen
ENCSR129KCJ	Homo sapiens	lung
ENCSR085HNI	Homo sapiens	liver
ENCSR274JRR	Homo sapiens	brain
ENCSR046XHI	Homo sapiens	female gonad
ENCSR693GGB	Homo sapiens	testis
ENCSR001UXR	Homo sapiens	pancreas
ENCSR680AAZ	Homo sapiens	adrenal gland
ENCSR635GTY	Homo sapiens	heart
ENCSR236OON	Homo sapiens	adipose tissue
ENCSR071ZMO	Homo sapiens	kidney
ENCSR394YLM	Mus musculus	kidney
ENCSR216KLZ	Mus musculus	liver
ENCSR266ESZ	Mus musculus	testis
ENCSR170SVO	Mus musculus	small intestine
ENCSR518GDK	Mus musculus	sigmoid colon
ENCSR248XKS	Mus musculus	pancreas
ENCSR288TLO	Mus musculus	adipose tissue
ENCSR966JPL	Mus musculus	spleen
ENCSR164BAZ	Mus musculus	heart
ENCSR870AQU	Mus musculus	lung
ENCSR516UNF	Mus musculus	female gonad
ENCSR713OCQ	Mus musculus	adrenal gland
ENCSR554PHF	Mus musculus	brain

Experiment accessions for the human and mouse tissues from [144].

Table B.2. **ENCODE Resequenced RNA-seq Data.**

Experiment accession	Species	Tissue
ENCSR236OON	Homo sapiens	adipose tissue
ENCSR680AAZ	Homo sapiens	adrenal gland
ENCSR274JRR	Homo sapiens	brain
ENCSR635GTY	Homo sapiens	heart
ENCSR071ZMO	Homo sapiens	kidney
ENCSR085HNI	Homo sapiens	liver
ENCSR129KCJ	Homo sapiens	lung
ENCSR270OKS	Homo sapiens	sigmoid colon
ENCSR612HYR	Homo sapiens	small intestine
ENCSR693GGB	Homo sapiens	testis
ENCSR288TLO	Mus musculus	adipose tissue
ENCSR713OCQ	Mus musculus	adrenal gland
ENCSR554PHF	Mus musculus	brain
ENCSR164BAZ	Mus musculus	heart
ENCSR394YLM	Mus musculus	kidney
ENCSR216KLZ	Mus musculus	liver
ENCSR870AQU	Mus musculus	lung
ENCSR518GDK	Mus musculus	sigmoid colon
ENCSR170SVO	Mus musculus	small intestine
ENCSR266ESZ	Mus musculus	testis

Experiment accessions for the human and mouse tissues resequenced from[144].

Table B.3. **Human and Mouse Histone Peak Calls.**

Experiment (Hs)	Tissue (Hs)	Experiment (Mm)	Tissue (Mm)	Histone Mark
ENCFF124AQU	heart right ventricle	ENCFF733HUI	heart	H3K27ac
ENCFF134ZIJ	heart left ventricle	ENCFF733HUI	heart	H3K27ac
ENCFF755CRA	small intestine	ENCFF047ETI	small intestine	H3K27ac
ENCFF529DJT	spleen	ENCFF976VFY	spleen	H3K27ac
ENCFF491YHG	heart right ventricle	ENCFF599BFW	heart	H3K4me3
ENCFF814QSX	heart left ventricle	ENCFF599BFW	heart	H3K4me3
ENCFF076SMI	spleen	ENCFF769BAR	spleen	H3K4me3
ENCFF898EJC	lung	ENCFF508WEP	lung	H3K4me3

Experiment accessions for the matched human (Hs) and mouse (Mm) tissues histone peak calls for H3K4me3 and H3K27ac. The peak calls are based on the hg19 (GRCh37) genome annotation for the human samples, and on the mm10 (GRCm38) genome annotation for the mouse samples.

Table B.4. Enriched GO Slim Terms in Differentially Expressed Orthologs.

$\widetilde{\text{FDR}} \approx 0.20$						
GO ID	Term	Annotated	Significant	Expected	Odds Ratio	p-value
GO:0006399	tRNA metabolic process	169	15	8.5500	1.8740	0.0245
GO:0006605	protein targeting	312	27	15.7900	1.8358	0.0046
GO:0034655	nucleobase-containing compound catabolic process	395	31	19.9900	1.6501	0.0103
GO:0006259	DNA metabolic process	848	63	42.9100	1.5781	0.0013
GO:0051276	chromosome organization	977	72	49.4400	1.5711	0.0007
GO:0006790	sulfur compound metabolic process	315	23	15.9400	1.5172	0.0496
GO:0006412	translation	540	39	27.3300	1.5102	0.0159
GO:0034641	cellular nitrogen compound metabolic process	5200	321	263.1500	1.4429	0.0000
GO:0009058	biosynthetic process	5113	294	258.7500	1.2672	0.0025
GO:0009056	catabolic process	2061	121	104.3000	1.2250	0.0402
$\widetilde{\text{FDR}} = 0.10$						
GO ID	Term	Annotated	Significant	Expected	Odds Ratio	p-value
GO:0071554	cell wall organization or biogenesis	10	1	0.0400	31.5367	0.0382
GO:0051276	chromosome organization	977	9	3.8000	2.9099	0.0121

GO Slim terms significantly enriched in the differentially expressed orthologs at different FDR cut-offs. The enrichment test was conducted using Fisher's exact test.

Table B.5. **Histone Peak Differences between the Differentially Expressed Orthologs in reported by Lin et al.**

H3K4me3 (Hs > Mm)					
Human tissue	Mouse tissue	Differentially expressed	n	p.value	Z
heart left ventricle	heart	2569	669	3.31×10^{-25}	12.4946
heart right ventricle	heart	2569	741	8.04×10^{-23}	11.0060
lung	lung	2569	630	9.87×10^{-25}	12.8700
spleen	spleen	2569	700	2.42×10^{-9}	6.2172
H3K4me3 (Hs < Mm)					
Human tissue	Mouse tissue	Differentially expressed	n	p.value	Z
heart left ventricle	heart	2198	642	2.39×10^{-15}	-8.5200
heart right ventricle	heart	2198	719	5.93×10^{-20}	-9.9276
lung	lung	2198	644	1.49×10^{-14}	-7.1023
spleen	spleen	2198	820	4.60×10^{-12}	-7.6246
H3K27ac (Hs > Mm)					
Human tissue	Mouse tissue	Differentially expressed	n	p.value	Z
heart left ventricle	heart	2569	838	8.15×10^{-22}	13.4307
heart right ventricle	heart	2569	969	6.70×10^{-16}	9.3044
small intestine	small intestine	2569	866	5.19×10^{-20}	11.9405
spleen	spleen	2569	994	9.69×10^{-26}	10.5260
H3K27ac (Hs < Mm)					
Human tissue	Mouse tissue	Differentially expressed	n	p.value	Z
heart left ventricle	heart	2198	814	5.12×10^{-13}	-7.1393
heart right ventricle	heart	2198	898	1.93×10^{-18}	-10.1227
small intestine	small intestine	2198	862	2.94×10^{-16}	-8.8226
spleen	spleen	2198	902	1.02×10^{-14}	-9.6613

Wilcoxon test results for the difference between the fold enrichment over control of H3K4me3 and H3K27ac present at promoters of the differentially expressed orthologs reported by Lin et al. The differentially expressed orthologs were separated into orthologs where the gene expression is higher in the human tissues than in the mouse tissues (Hs > Mm), and where the gene expression is higher in the mouse tissues than in the human tissues (Hs < Mm). The p-values and Z-statistics were generated by the nonparametric paired Wilcoxon test between the human and mouse gene-associated histone peak intensities, *n* is the number of human-mouse orthologs where at least one of them has a gene-associated peak intensity.

Table B.6. **Histone Peak Differences between the Differentially Expressed Orthologs in the Resequenced ENCODE Data Set.**

H3K4me3 (Hs > Mm)					
Human tissue	Mouse tissue	Differentially expressed	n	p.value	Z
heart left ventricle	heart	1422	584	5.59×10^{-15}	8.0410
heart right ventricle	heart	1422	620	2.79×10^{-9}	6.3512
lung	lung	1422	579	4.76×10^{-17}	9.8433
H3K4me3 (Hs < Mm)					
Human tissue	Mouse tissue	Differentially expressed	n	p.value	Z
heart left ventricle	heart	1297	314	1.85×10^{-4}	-4.3938
heart right ventricle	heart	1297	348	1.02×10^{-7}	-6.5685
lung	lung	1297	304	5.31×10^{-4}	-3.2987
H3K27ac (Hs > Mm)					
Human tissue	Mouse tissue	Differentially expressed	n	p.value	Z
heart left ventricle	heart	1422	686	1.01×10^{-24}	12.1412
heart right ventricle	heart	1422	774	6.79×10^{-18}	9.7264
small intestine	small intestine	1422	677	3.91×10^{-13}	8.5614
H3K27ac (Hs < Mm)					
Human tissue	Mouse tissue	Differentially expressed	n	p.value	Z
heart left ventricle	heart	1297	421	3.91×10^{-10}	-6.2610
heart right ventricle	heart	1297	440	4.72×10^{-16}	-9.5647
small intestine	small intestine	1297	422	2.96×10^{-6}	-4.4332

Wilcoxon test results for the difference between the fold enrichment over control of H3K4me3 and H3K27ac present at promoters of the differentially expressed orthologs based on the normalized FPKM values from the resequenced ENCODE RNA-seq data set. The differentially expressed orthologs were separated into orthologs where the gene expression is higher in the human tissues than in the mouse tissues (Hs > Mm), and where the gene expression is higher in the mouse tissues than in the human tissues (Hs < Mm). The p-values and Z-statistics were generated by the nonparametric paired Wilcoxon test between the human and mouse gene-associated histone peak intensities, *n* is the number of human-mouse orthologs where at least one of them has a gene-associated peak intensity.

Table B.7. **Histone Peak Differences between the Differentially Expressed Orthologs in both the Resequenced ENCODE Data Set and the Microarrays from Barcode 3.0.**

H3K4me3 (Hs > Mm)					
Human tissue	Mouse tissue	Differentially expressed	n	p.value	Z
heart left ventricle	heart	391	171	9.78×10^{-6}	4.4314
heart right ventricle	heart	391	184	5.36×10^{-4}	3.8841
lung	lung	391	165	1.01×10^{-4}	4.5705
H3K4me3 (Hs < Mm)					
heart left ventricle	heart	362	79	6.51×10^{-1}	0.8933
heart right ventricle	heart	362	86	3.92×10^{-1}	-0.9676
lung	lung	362	75	6.48×10^{-1}	1.0274
H3K27ac (Hs > Mm)					
Human tissue	Mouse tissue	Differentially expressed	n	p.value	Z
heart left ventricle	heart	391	187	1.75×10^{-8}	6.9619
heart right ventricle	heart	391	200	1.07×10^{-5}	5.2173
small intestine	small intestine	391	175	3.78×10^{-4}	4.2715
H3K27ac (Hs < Mm)					
Human tissue	Mouse tissue	Differentially expressed	n	p.value	Z
heart left ventricle	heart	362	101	2.01×10^{-2}	-2.2850
heart right ventricle	heart	362	112	2.14×10^{-3}	-3.9039
small intestine	small intestine	362	104	7.41×10^{-1}	1.1797

Wilcoxon test results for the difference between the fold enrichment over control of H3K4me3 and H3K27ac present at promoters of the differentially expressed orthologs based on both the normalized probe values from the resequenced ENCODE RNA-seq data set and the normalized microarrays values from Barcode 3.0 where the human gene and the mouse gene have the same number of annotated transcripts. The differentially expressed orthologs were separated into orthologs where the gene expression is higher in the human tissues than in the mouse tissues (Hs > Mm), and where the gene expression is higher in the mouse tissues than in the human tissues (Hs < Mm). The p-values and Z-statistics were generated by the nonparametric paired Wilcoxon test between the human and mouse gene-associated histone peak intensities, *n* is the number of human-mouse orthologs where at least one of them has a gene-associated peak intensity.

Table B.8. **Histone Peak Differences between the Differentially Expressed Orthologs in reported by Lin et al. with the same Number of Annotated Transcripts.**

H3K4me3 (Hs > Mm)					
Human tissue	Mouse tissue	Differentially expressed	n	p.value	Z
heart left ventricle	heart	158	31	6.66×10^{-1}	0.5643
heart right ventricle	heart	158	35	2.25×10^{-1}	1.0419
lung	lung	158	30	3.49×10^{-1}	1.3039
spleen	spleen	158	31	6.59×10^{-1}	0.0510
H3K4me3 (Hs < Mm)					
Human tissue	Mouse tissue	Differentially expressed	n	p.value	Z
heart left ventricle	heart	338	88	1.27×10^{-1}	-1.9414
heart right ventricle	heart	338	100	5.59×10^{-1}	-0.8020
lung	lung	338	88	8.23×10^{-1}	-0.5981
spleen	spleen	338	104	5.73×10^{-1}	0.4475
H3K27ac (Hs > Mm)					
Human tissue	Mouse tissue	Differentially expressed	n	p.value	Z
heart left ventricle	heart	158	38	6.90×10^{-1}	0.7775
heart right ventricle	heart	158	45	5.76×10^{-1}	1.4932
small intestine	small intestine	158	44	1.95×10^{-1}	0.9224
spleen	spleen	158	43	5.83×10^{-2}	2.1003
H3K27ac (Hs < Mm)					
Human tissue	Mouse tissue	Differentially expressed	n	p.value	Z
heart left ventricle	heart	338	91	3.05×10^{-1}	-1.1775
heart right ventricle	heart	338	102	4.33×10^{-1}	-0.4549
small intestine	small intestine	338	109	5.80×10^{-2}	-2.1923
spleen	spleen	338	109	1.90×10^{-1}	-1.4222

Wilcoxon test results for the difference between the fold enrichment over control of H3K4me3 and H3K27ac present at promoters of the differentially expressed orthologs reported by Lin et al. where the human gene and the mouse gene have the same number of annotated transcripts. The differentially expressed orthologs were separated into orthologs where the gene expression is higher in the human tissues than in the mouse tissues (Hs > Mm), and where the gene expression is higher in the mouse tissues than in the human tissues (Hs < Mm). The p-values and Z-statistics were generated by the nonparametric paired Wilcoxon test between the human and mouse gene-associated histone peak intensities, *n* is the number of human-mouse orthologs where at least one of them has a gene-associated peak intensity.

Table B.9. **Histone Peak Differences between the Differentially Expressed Orthologs in the Resequenced ENCODE Data Set with the same Number of Annotated Transcripts.**

H3K4me3 (Hs > Mm)					
Human tissue	Mouse tissue	Differentially expressed	n	p.value	Z
heart left ventricle	heart	92	36	9.57×10^{-1}	0.2224
heart right ventricle	heart	92	42	6.03×10^{-1}	0.8219
lung	lung	92	36	8.71×10^{-1}	-0.2114
H3K4me3 (Hs < Mm)					
Human tissue	Mouse tissue	Differentially expressed	n	p.value	Z
heart left ventricle	heart	132	31	8.70×10^{-1}	-0.3437
heart right ventricle	heart	132	32	6.54×10^{-1}	-0.6682
lung	lung	132	31	6.22×10^{-1}	0.8185
H3K27ac (Hs > Mm)					
Human tissue	Mouse tissue	Differentially expressed	n	p.value	Z
heart left ventricle	heart	92	39	8.61×10^{-2}	1.6374
heart right ventricle	heart	92	44	4.72×10^{-2}	2.1433
small intestine	small intestine	92	43	3.31×10^{-2}	1.6897
H3K27ac (Hs < Mm)					
Human tissue	Mouse tissue	Differentially expressed	n	p.value	Z
heart left ventricle	heart	132	40	1.36×10^{-1}	-1.4129
heart right ventricle	heart	132	40	2.48×10^{-1}	-1.0819
small intestine	small intestine	132	41	5.29×10^{-1}	0.8395

Wilcoxon test results for the difference between the fold enrichment over control of H3K4me3 and H3K27ac present at promoters of the differentially expressed orthologs based on the normalized FPKM values from the resequenced ENCODE RNA-seq data set where the human gene and the mouse gene have the same number of annotated transcripts. The differentially expressed orthologs were separated into orthologs where the gene expression is higher in the human tissues than in the mouse tissues (Hs > Mm), and where the gene expression is higher in the mouse tissues than in the human tissues (Hs < Mm). The p-values and Z-statistics were generated by the nonparametric paired Wilcoxon test between the human and mouse gene-associated histone peak intensities, *n* is the number of human-mouse orthologs where at least one of them has a gene-associated peak intensity.

Table B.10. **Histone Peak Differences between the Differentially Expressed Orthologs in both the Resequenced ENCODE Data Set and the Microarrays from Barcode 3.0 with the same Number of Annotated Transcripts.**

H3K4me3 (Hs > Mm)					
Human tissue	Mouse tissue	Differentially expressed	n	p.value	Z
heart left ventricle	heart	29	10	1.6×10^{-1}	2.0854
heart right ventricle	heart	29	12	1.1×10^{-1}	1.7198
lung	lung	29	9	5.7×10^{-1}	0.8095
H3K4me3 (Hs < Mm)					
heart left ventricle	heart	39	3	7.50×10^{-1}	0.0000
heart right ventricle	heart	39	3	1.00×10^{0}	0.0000
lung	lung	39	4	8.75×10^{-1}	0.2958
H3K27ac (Hs > Mm)					
Human tissue	Mouse tissue	Differentially expressed	n	p.value	Z
heart left ventricle	heart	29	10	1.60×10^{-1}	1.9021
heart right ventricle	heart	29	9	5.47×10^{-2}	2.3197
small intestine	small intestine	29	10	3.71×10^{-2}	1.7237
H3K27ac (Hs < Mm)					
Human tissue	Mouse tissue	Differentially expressed	n	p.value	Z
heart left ventricle	heart	39	13	1.46×10^{-1}	-2.1381
heart right ventricle	heart	39	14	9.06×10^{-2}	-1.9433
small intestine	small intestine	39	9	1.95×10^{-2}	2.1378

Wilcoxon test results for the difference between the fold enrichment over control of H3K4me3 and H3K27ac present at promoters of the differentially expressed orthologs based on both the normalized probe values from the resequenced ENCODE RNA-seq data set and the normalized microarrays values from Barcode 3.0 where the human gene and the mouse gene have the same number of annotated transcripts. The differentially expressed orthologs were separated into orthologs where the gene expression is higher in the human tissues than in the mouse tissues (Hs > Mm), and where the gene expression is higher in the mouse tissues than in the human tissues (Hs < Mm). The p-values and Z-statistics were generated by the nonparametric paired Wilcoxon test between the human and mouse gene-associated histone peak intensities, *n* is the number of human-mouse orthologs where at least one of them has a gene-associated peak intensity.

B.3 Human and Mouse Microarray Samples.

https://zenodo.org/record/1242623/files/barcode_mcr_samples.xlsx

B.4 FastQC Report for the Resequenced ENCODE Data Set.

https://zenodo.org/record/1242623/files/lin2_fastQC_report.xlsx

C

Supplementary Materials for Chapter 3

C.1 Supplementary Figures

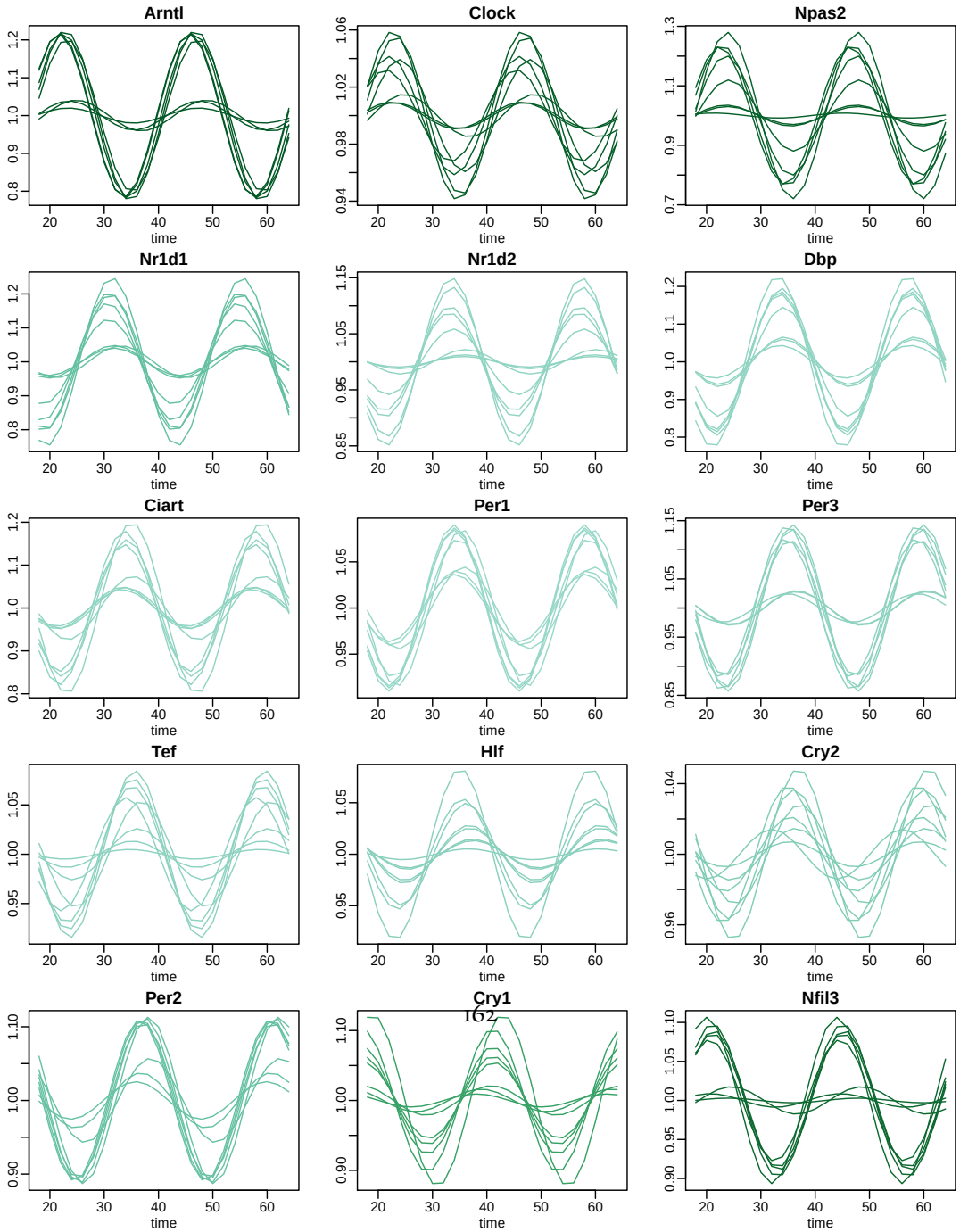


Figure C.1. Harmonic regression fits for circadian genes. The regression is based on the time course data set.

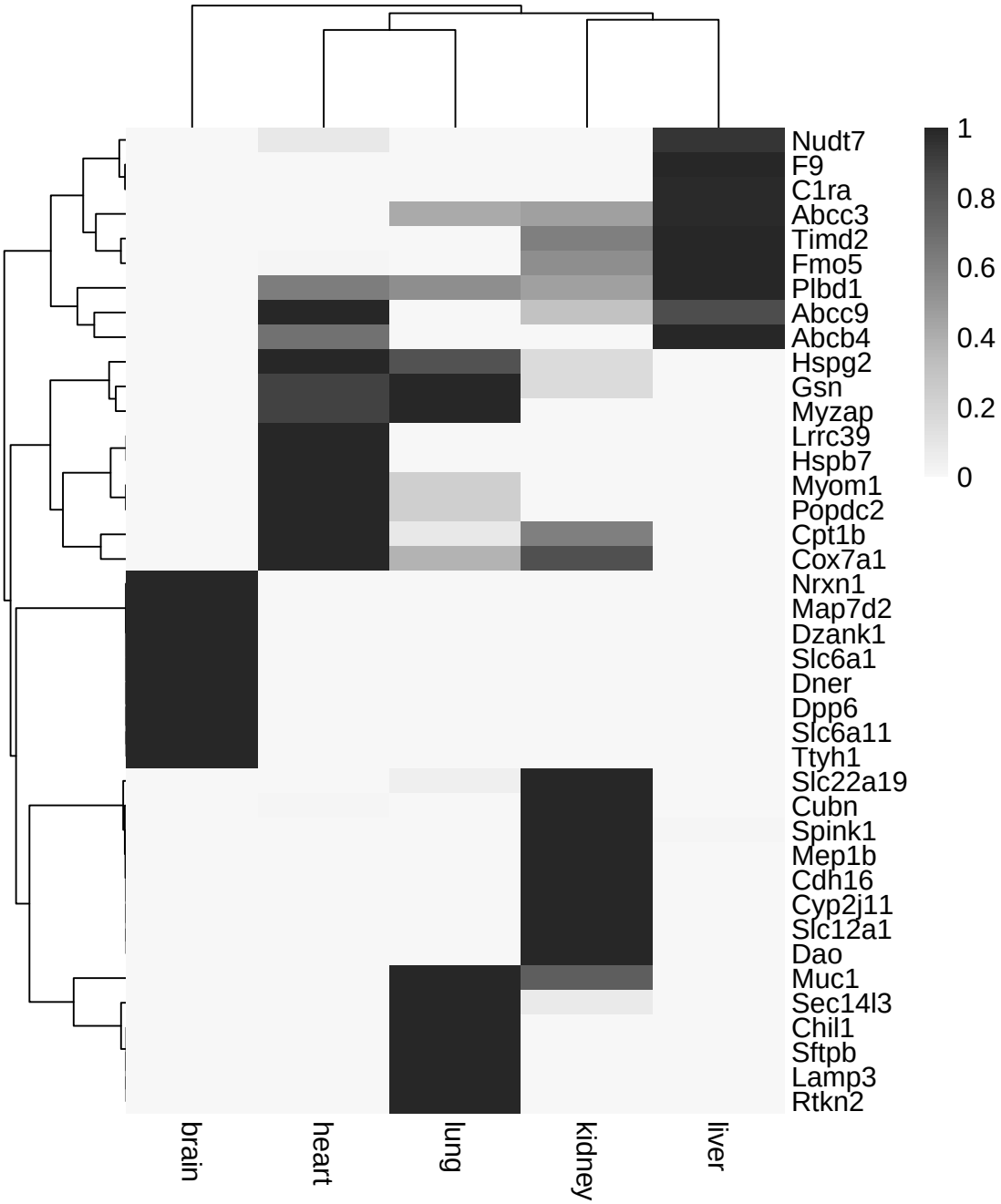


Figure C.2. Tissue Specific Gene Set. Gene Expression Barcode estimates for the set of tissue specific genes.

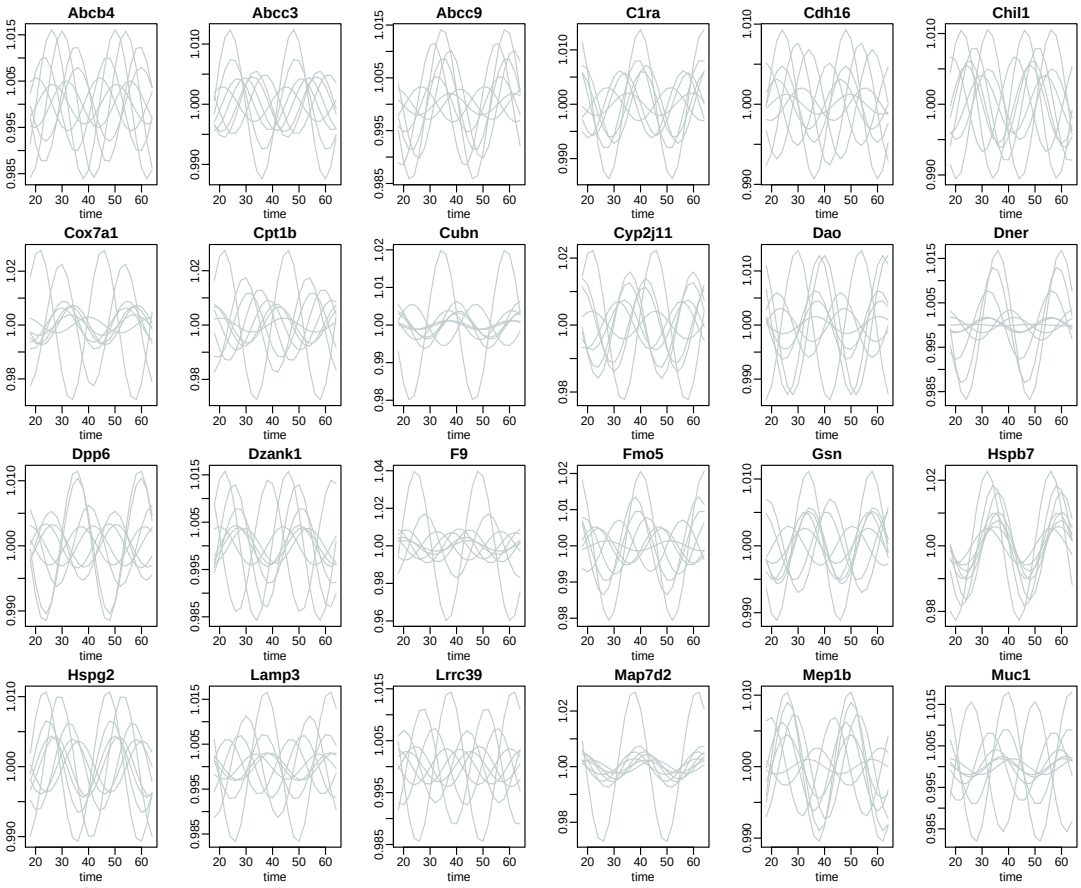


Figure C.3. Harmonic regression fits for tissue-specific genes. The regression is based on the time course data set.

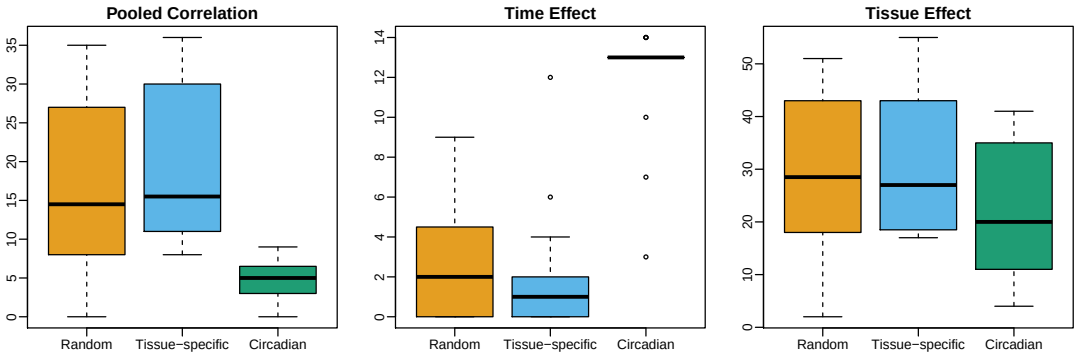


Figure C.4. **Time course coexpression network degree.** Box plots with the degree split by gene class for the coexpression networks based on the pooled observed correlation, the correlation between the gene-specific time effect estimates, and the correlation based on the gene-specific tissue effect.

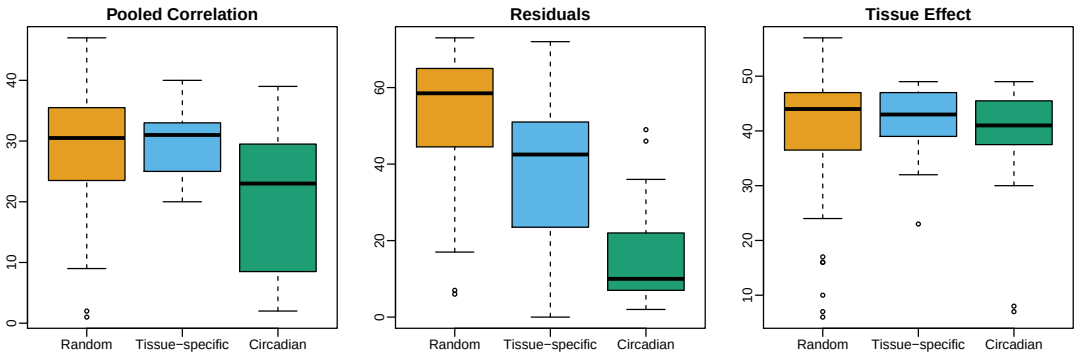


Figure C.5. **Tissue panel coexpression network degree.** Box plots with the degree split by gene class for the coexpression networks based on the pooled observed correlation, the correlation between the residuals, and the correlation based on the gene-specific tissue effect.

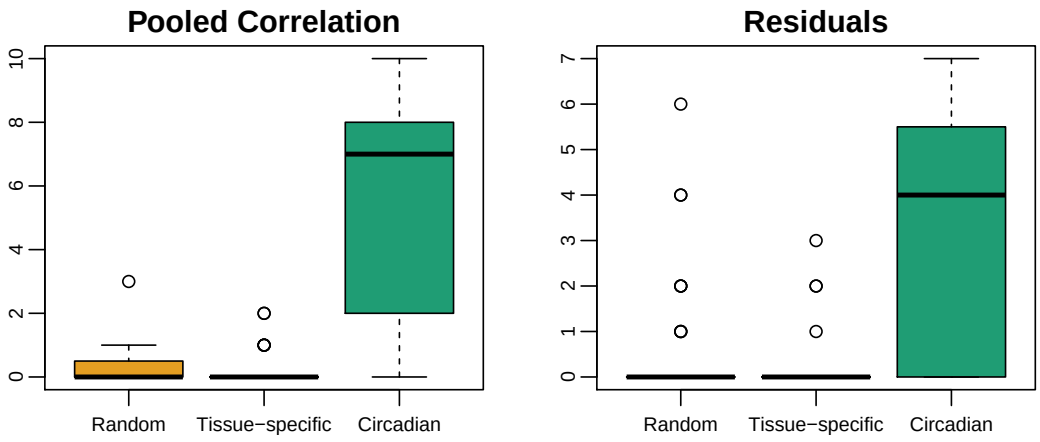


Figure C.6. **Liver biological replicates coexpression network degree.** Box plots with the degree split by gene class for the coexpression networks based on the pooled observed correlation and the correlation between the residuals.

C.2 Supplementary Tables

Gene	Mean.Phase
Arntl	0.0000
Npas2	0.1514
Cry1	4.9948
Nfil3	6.0077
Clock	0.0791
Rorc	5.0495
Nr1d1	2.2956
Cry2	3.5008
Bhlhe41	2.9090
Hlf	3.4256
Per2	3.8866
Dbp	2.9915
Tef	3.3136
Per1	3.1220
Nr1d2	2.9871
Ciart	3.0077
Per3	3.3059

Table C.1. **Circadian Genes Mean Phase.** Mean phase estimates for the mouse circadian genes.

C.3 Time Course Data Set GEO Accessions.

https://zenodo.org/record/1243315/files/LiverBiologicalReplicates_samples.xlsx

C.4 Tissue Panel Data Set GEO Accessions.

https://zenodo.org/record/1243315/files/TissuePanel_samples.xlsx

C.5 Liver Biological Replicates Data Set GEO Accessions.

https://zenodo.org/record/1243315/files/LiverBiologicalReplicates_samples.xlsx

References

[1] Alexeyenko, A., Lee, W., Pernemalm, M., Guegan, J., Dessen, P., Lazar, V., Lehtiö, J., & Pawitan, Y. (2012). Network enrichment analysis: extension of gene-set enrichment analysis to gene networks. *BMC Bioinformatics*, 13, 226.

[2] Alexeyenko, A. & Sonnhammer, E. L. L. (2009). Global networks of functional coupling in eukaryotes from comprehensive data integration. *Genome Res.*, 19(6), 1107–1116.

[3] Almeida-de Macedo, M. M., Ransom, N., Feng, Y., Hurst, J., & Wurtele, E. S. (2013a). Comprehensive analysis of correlation coefficients estimated from pooling heterogeneous microarray data. *BMC Bioinformatics*, 14, 214.

[4] Almeida-de Macedo, M. M., Ransom, N., Feng, Y., Hurst, J., & Wurtele, E. S. (2013b). Comprehensive analysis of correlation coefficients estimated from pooling heterogeneous microarray data. *BMC bioinformatics*, 14(1), 214.

[5] Altschuler, G. M., Hofmann, O., Kalatskaya, I., Payne, R., Ho Sui, S. J., Saxena, U., Krivtsov, A. V., Armstrong, S. A., Cai, T., Stein, L., & Hide, W. A. (2013). Pathprinting: An integrative approach to understand the functional basis of disease. *Genome Med.*, 5(7), 68.

[6] Arrowsmith, J. (2011). Trial watch: Phase ii failures: 2008–2010. *Nature reviews Drug discovery*, 10(5), 328–329.

[7] Arsham, A. M. & Neufeld, T. P. (2009). A genetic screen in Drosophila reveals novel cytoprotective functions of the autophagy-lysosome pathway. *PLoS ONE*, 4(6), e6068.

[8] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000a). Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1), 25.

[9] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G.

(2000b). Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat. Genet.*, 25(1), 25–29.

[10] Barabási, A.-L., Gulbahce, N., & Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.*, 12(1), 56–68.

[11] Barabási, A.-L. & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, 5(2), 101–113.

[12] Barbosa-Morais, N. L., Irimia, M., Pan, Q., Xiong, H. Y., Gueroussov, S., Lee, L. J., Slobodeniuc, V., Kutter, C., Watt, S., Colak, R., et al. (2012). The evolutionary landscape of alternative splicing in vertebrate species. *Science*, 338(6114), 1587–1593.

[13] Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I. F., Soboleva, A., Tomashevsky, M., & Edgar, R. (2007). NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res.*, 35(Database), D760–D765.

[14] Barry, W. T., Nobel, A. B., & Wright, F. A. (2005). Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, 21(9), 1943–1949.

[15] Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, Articles*, 67(1), 1–48.

[16] Bauer, S., Gagneur, J., & Robinson, P. N. (2010). GOing bayesian: model-based gene set analysis of genome-scale data. *Nucleic Acids Res.*, 38(11), 3523–3532.

[17] Becker, K. G., Barnes, K. C., Bright, T. J., & Wang, S. A. (2004). The genetic association database. *Nat. Genet.*, 36(5), 431–432.

[18] Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.*, (pp. 289–300).

[19] Berezovska, O., Xia, M. Q., & Hyman, B. T. (1998). Notch is expressed in adult brain, is coexpressed with presenilin-1, and is altered in alzheimer disease. *J. Neuropathol. Exp. Neurol.*, 57(8), 738–745.

[20] Berg, M., Rogers, R., Muralla, R., & Meinke, D. (2005). Requirement of aminoacyl-tRNA synthetases for gametogenesis and embryo development in Arabidopsis. *Plant J.*, 44(5), 866–878.

[21] Bodea, L.-G., Wang, Y., Linnartz-Gerlach, B., Kopatz, J., Sinkkonen, L., Musgrove, R., Kaoma, T., Muller, A., Vallar, L., Di Monte, D. A., Balling, R., & Neumann, H. (2014). Neurodegeneration by activation of the microglial Complement–Phagosome pathway. *J. Neurosci.*, 34(25), 8546–8556.

[22] Bolstad, B. M. (2016). *preprocessCore: A collection of pre-processing functions*. R package version 1.34.0.

[23] Boniecki, M. T., Rho, S. B., Tukalo, M., Hsu, J. L., Romero, E. P., & Martinis, S. A. (2009). Leucyl-tRNA synthetase-dependent and -independent activation of a group I intron. *J. Biol. Chem.*, 284(39), 26243–26250.

[24] Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., et al. (2011). The evolution of gene expression levels in mammalian organs. *Nature*, 478(7369), 343–348.

[25] Brazma, A., Kapushesky, M., Parkinson, H., Sarkans, U., & Shojatalab, M. (2006). [20] data storage and analysis in ArrayExpress. In *Methods in Enzymology* (pp. 370–386).

[26] Breschi, A., Djebali, S., Gillis, J., Pervouchine, D. D., Dobin, A., Davis, C. A., Gingeras, T. R., & Guigó, R. (2016). Gene-specific patterns of expression variation across organs and species. *Genome biology*, 17(1), 151.

[27] Brett, D., Pospisil, H., Valcárcel, J., Reich, J., & Bork, P. (2002). Alternative splicing and genome complexity. *Nature genetics*, 30(1), 29–30.

[28] Burns, A. & Iliffe, S. (2009). Alzheimer's disease. *BMJ*, 338(feb05 1), b158–b158.

[29] Butte, A. J. & Kohane, I. S. (1999). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In *Biocomputing 2000* (pp. 418–429). World Scientific.

[30] Caltagarone, J., Jing, Z., & Bowser, R. (2007). Focal adhesions regulate abeta signaling and cell death in alzheimer's disease. *Biochim. Biophys. Acta*, 1772(4), 438–445.

[31] Carlson, M. (2016a). hgu133plus2.db: Affymetrix human genome U133 plus 2.0 array annotation data (chip hgu133plus2).

[32] Carlson, M. (2016b). *mouse4302.db: Affymetrix Mouse Genome 430 2.0 Array annotation data (chip mouse4302)*. R package version 3.2.3.

[33] Carlson, M. (2016c). *mouse4302.db: Affymetrix Mouse Genome 430 2.0 Array annotation data (chip mouse4302)*. R package version 3.2.3.

[34] Carvalho, B. (2015). *pd.mogene.1.0.st.v1: Platform Design Info for Affymetrix MoGene 1.0 ST v1*. R package version 3.14.1.

[35] Carvalho, B. S. & Irizarry, R. A. (2010). A framework for oligonucleotide microarray preprocessing. *Bioinformatics*, 26(19), 2363–7.

[36] Castrignanò, T., D'Antonio, M., Anselmo, A., Carrabino, D., D'Onorio De Meo, A., D'erchia, A., Licciulli, F., Mangiulli, M., Mignone, F., Pavesi, G., et al. (2008). Aspicdb: a database resource for alternative splicing analysis. *Bioinformatics*, 24(10), 1300–1304.

[37] Catricala, S., Torti, M., & Ricevuti, G. (2012). Alzheimer disease and platelets: how's that relevant. *Immun. Ageing*, 9(1), 20.

[38] Chan, E. T., Quon, G. T., Chua, G., Babak, T., Trochesset, M., Zirngibl, R. A., Aubin, J., Ratcliffe, M. J., Wilde, A., Brudno, M., et al. (2009). Conservation of core gene expression in vertebrate tissues. *Journal of biology*, 8(3), 33.

[39] Cheon, D.-J. & Orsulic, S. (2011). Mouse models of cancer.

[40] Clough, E. & Barrett, T. (2016). The gene expression omnibus database. *Methods Mol. Biol.*, 1418, 93–110.

[41] Consortium, E. P. et al. (2011). A user's guide to the encyclopedia of dna elements (encode). *PLoS biology*, 9(4), e1001046.

[42] Consortium, G. et al. (2015). The genotype-tissue expression (gtex) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235), 648–660.

[43] Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A., et al. (2010). Histone h3k27ac

separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences*, 107(50), 21931–21936.

[44] Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., Jupe, S., Kalatskaya, I., Mahajan, S., May, B., Ndegwa, N., Schmidt, E., Shamovsky, V., Yung, C., Birney, E., Hermjakob, H., D'Eustachio, P., & Stein, L. (2011). Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, 39(Database issue), D691–7.

[45] Davis, S. & Meltzer, P. S. (2007). GEOquery: a bridge between the gene expression omnibus (GEO) and BioConductor. *Bioinformatics*, 23(14), 1846–1847.

[46] de Jager, M., van der Wildt, B., Schul, E., Bol, J. G. J. M., van Duinen, S. G., Drukarch, B., & Wilhelmus, M. M. M. (2013). Tissue transglutaminase colocalizes with extracellular matrix proteins in cerebral amyloid angiopathy. *Neurobiol. Aging*, 34(4), 1159–1169.

[47] de Jonge, H. J. M., Fehrmann, R. S. N., de Bont, E. S. J. M., Hofstra, R. M. W., Gerbens, F., Kamps, W. A., de Vries, E. G. E., van der Zee, A. G. J., te Meerman, G. J., & ter Elst, A. (2007). Evidence based selection of housekeeping genes. *PLoS One*, 2(9), e898.

[48] de Magalhaes, J. P. (2014). Why genes extending lifespan in model organisms have not been consistently associated with human longevity and what it means to translation research. *Cell Cycle*, 13(17), 2671–2673.

[49] De Veaux, R. D. & Hand, D. J. (2005). How to lie with bad data. *Statistical Science*, (pp. 231–238).

[50] Demetrius, L. (2005). Of mice and men. When it comes to studying ageing and the means to slow it down, mice are not just small humans. *EMBO Rep.*, 6 Spec No, 39–44.

[51] Di Lena, P., Martelli, P. L., Fariselli, P., & Casadio, R. (2015). NET-GE: a novel NETwork-based gene enrichment for detecting biological processes associated to mendelian diseases. *BMC Genomics*, 16 Suppl 8, S6.

[52] Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1), 15–21.

[53] Draghici, S., Khatri, P., Bhavsar, P., Shah, A., Krawetz, S. A., & Tainsky, M. A. (2003). Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Res.*, 31(13), 3775–3781.

[54] Duits, F. H., Hernandez-Guillamon, M., Montaner, J., Goos, J. D. C., Montañola, A., Wattjes, M. P., Barkhof, F., Scheltens, P., Teunissen, C. E., & van der Flier, W. M. (2015). Matrix metalloproteinases in alzheimer's disease and concurrent cerebral microbleeds. *J. Alzheimers. Dis.*, 48(3), 711–720.

[55] Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., & Huber, W. (2005). Biomart and bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21(16), 3439–3440.

[56] Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nature protocols*, 4(8), 1184–1191.

[57] Dutkowski, J., Kramer, M., Surma, M. A., Balakrishnan, R., Cherry, J. M., Krogan, N. J., & Ideker, T. (2013). A gene ontology inferred from molecular networks. *Nat. Biotechnol.*, 31(1), 38–45.

[58] Efron, B. & Tibshirani, R. (2007). On testing the significance of sets of genes. *Ann. Appl. Stat.*, 1(1), 107–129.

[59] Eisenberg, E. & Levanon, E. Y. (2013). Human housekeeping genes, revisited. *Trends Genet.*, 29(10), 569–574.

[60] Elso, C., Lu, X., Morrison, S., Tarver, A., Thompson, H., Thurkow, H., Yamada, N. A., & Stubbs, L. (2008). Germline translocations in mice: unique tools for analyzing gene function and long-distance regulatory mechanisms. *Journal of the National Cancer Institute Monographs*, 2008(39), 91–95.

[61] Ewing, R. M., Chu, P., Elisma, F., Li, H., Taylor, P., Climie, S., McBroom-Cerajewski, L., Robinson, M. D., O'Connor, L., Li, M., Taylor, R., Dharsee, M., Ho, Y., Heilbut, A., Moore, L., Zhang, S., Ornatsky, O., Bukhman, Y. V., Ethier, M., Sheng, Y., Vasilescu, J., Abu-Farha, M., Lambert, J.-P., Duewel, H. S., Stewart, I. I., Kuehl, B., Hogue, K., Colwill, K., Gladwish, K., Muskat, B., Kinach, R., Adams, S.-L., Moran, M. F., Morin, G. B., Topaloglou,

T., & Figeys, D. (2007). Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol. Syst. Biol.*, 3, 89.

[62] Ezkurdia, I., del Pozo, A., Frankish, A., Rodriguez, J. M., Harrow, J., Ashman, K., Valencia, A., & Tress, M. L. (2012). Comparative proteomics reveals a significant bias toward alternative protein isoforms with conserved structure and function. *Molecular biology and evolution*, 29(9), 2265–2283.

[63] Falcon, S. & Gentleman, R. (2006). Using GOstats to test gene lists for GO term association. *Bioinformatics*, 23(2), 257–258.

[64] Fang, G., Bhardwaj, N., Robilotto, R., & Gerstein, M. B. (2010). Getting started in gene orthology and functional analysis. *PLoS Comput Biol*, 6(3), e1000703.

[65] Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Systematic Biology*, 19(2), 99–113.

[66] Fitch, W. M. (2000). Homology: a personal view on some of the problems. *Trends in genetics*, 16(5), 227–231.

[67] Flicek, P., Ahmed, I., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., et al. (2012). Ensembl 2013. *Nucleic acids research*, 41(D1), D48–D55.

[68] Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., et al. (2011). Ensembl 2012. *Nucleic acids research*, 40(D1), D84–D90.

[69] Franco, A. T., Malaguarnera, R., Refetoff, S., Liao, X.-H., Lundsmith, E., Kimura, S., Pritchard, C., Marais, R., Davies, T. F., Weinstein, L. S., et al. (2011). Thyrotrophin receptor signaling dependence of braf-induced thyroid tumor initiation in mice. *Proceedings of the National Academy of Sciences*, 108(4), 1615–1620.

[70] Frost, H. R. & Amos, C. I. (2017). Gene set selection via LASSO penalized regression (SLPR). *Nucleic Acids Res.*, 45(12), e114.

[71] Frost, H. R. & McCray, A. T. (2012). Markov chain ontology analysis (MCOA). *BMC Bioinformatics*, 13, 23.

[72] Fujibuchi, W., Kiseleva, L., Taniguchi, T., Harada, H., & Horton, P. (2007). CellMontage: similar expression profile search server. *Bioinformatics*, 23(22), 3103–3104.

[73] Gabaldón, T. (2008). Large-scale assignment of orthology: back to phylogenetics? *Genome biology*, 9(10), 235.

[74] Gabaldón, T. & Koonin, E. V. (2013). Functional and evolutionary implications of gene orthology. *Nature Reviews Genetics*, 14(5), 360–366.

[75] Gabdoulline, R., Kaisers, W., Gaspar, A., Meganathan, K., Doss, M., Jagtap, S., Hescheler, J., Sachinidis, A., & Schwender, H. (2015). Differences in the early development of human and mouse embryonic stem cells. *PloS one*, 10(10), e0140803.

[76] Gaiteri, C., Ding, Y., French, B., Tseng, G. C., & Sibille, E. (2014). Beyond modules and hubs: the potential of gene coexpression networks for investigating molecular mechanisms of complex brain disorders. *Genes, Brain and Behavior*, 13(1), 13–24.

[77] Gaubatz, S., Lindeman, G. J., Ishida, S., Jakoi, L., Nevins, J. R., Livingston, D. M., & Rempel, R. E. (2000). E2F4 and E2F5 play an essential role in pocket protein-mediated G1 control. *Mol. Cell*, 6(3), 729–735.

[78] Gilad, Y. & Mizrahi-Man, O. (2015). A reanalysis of mouse encode comparative gene expression data. *F1000Research*, 4.

[79] Glazko, G. V. & Emmert-Streib, F. (2009). Unite and conquer: univariate and multivariate approaches for finding differentially expressed gene sets. *Bioinformatics*, 25(18), 2348–2354.

[80] Goh, A. M., Coffill, C. R., & Lane, D. P. (2011). The role of mutant p53 in human cancer. *The Journal of pathology*, 223(2), 116–126.

[81] Gough, N. R. (2002). Science's signal transduction knowledge environment. *Annals of the New York Academy of Sciences*, 971(1), 585–587.

[82] Graves, J. A. M. (1996). Mammals that break the rules: genetics of marsupials and monotremes. *Annual review of genetics*, 30(1), 233–260.

[83] Grossmann, S., Bauer, S., Robinson, P. N., & Vingron, M. (2007). Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis. *Bioinformatics*, 23(22), 3024–3031.

[84] Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R., & Young, R. A. (2007). A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, 130(1), 77–88.

[85] Gunderson, F. Q. & Johnson, T. L. (2009). Acetylation by the transcriptional coactivator Gcn5 plays a novel role in co-transcriptional spliceosome assembly. *PLoS Genet.*, 5(10), e1000682.

[86] Gusfield, D., Balasubramanian, K., & Naor, D. (1994). Parametric optimization of sequence alignment. *Algorithmica*, 12(4-5), 312.

[87] Haley, P. J. (2003). Species differences in the structure and function of the immune system. *Toxicology*, 188(1), 49–71.

[88] Hardison, R. C. (2003). Comparative genomics. *PLoS Biol.*, 1(2), E58.

[89] Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). Gencode: the reference human genome annotation for the encode project. *Genome research*, 22(9), 1760–1774.

[90] Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358), 320–338.

[91] Hassler, U., Uwe, H., & Thorsten, T. (2003). Nonsensical and biased correlation due to pooling heterogeneous samples. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(3), 367–379.

[92] Hedges, S. B., Dudley, J., & Kumar, S. (2006). Timetree: a public knowledge-base of divergence times among organisms. *Bioinformatics*, 22(23), 2971–2972.

[93] Hon, C.-C., Weber, C., Sismeiro, O., Proux, C., Koutero, M., Deloger, M., Das, S., Agrahari, M., Dillies, M.-A., Jagla, B., et al. (2012). Quantification of stochastic noise of splicing and polyadenylation in entamoeba histolytica. *Nucleic acids research*, 41(3), 1936–1952.

[94] Horvath, S., Zhang, B., Carlson, M., Lu, K., Zhu, S., Felciano, R., Laurance, M., Zhao, W., Qi, S., Chen, Z., et al. (2006). Analysis of oncogenic signaling networks in glioblastoma identifies aspm as a molecular target. *Proceedings of the National Academy of Sciences*, 103(46), 17402–17407.

[95] Huang, Y., Li, H., Hu, H., Yan, X., Waterman, M. S., Huang, H., & Zhou, X. J. (2007). Systematic discovery of functional modules and context-specific functional annotation of human genome. *Bioinformatics*, 23(13), i222–i229.

[96] Huang, Y. & Li, S. (2010). Detection of characteristic sub pathway network for angiogenesis based on the comprehensive pathway network. *BMC Bioinformatics*, 11 Suppl 1, S32.

[97] Hughes, M. E., Hogenesch, J. B., & Kornacker, K. (2010). Jtk_cycle: an efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets. *Journal of biological rhythms*, 25(5), 372–380.

[98] Hung, J.-H., Yang, T.-H., Hu, Z., Weng, Z., & DeLisi, C. (2012). Gene set enrichment analysis: performance evaluation and usage guidelines. *Brief. Bioinform.*, 13(3), 281–291.

[99] Huynen, M. A. & Bork, P. (1998). Measuring genome evolution. *Proceedings of the National Academy of Sciences*, 95(11), 5849–5856.

[100] Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., & Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2), 249–264.

[101] Isserlin, R., Merico, D., Voisin, V., & Bader, G. D. (2014). Enrichment map – a cytoscape app to visualize and explore OMICs pathway enrichment results. *F1000Res.*

[102] Izawa, D. & Pines, J. (2011). How APC/C-Cdc20 changes its substrate specificity in mitosis. *Nat. Cell Biol.*, 13(3), 223–233.

[103] Jordan, I. K., Mariño-Ramírez, L., Wolf, Y. I., & Koonin, E. V. (2004). Conservation and coevolution in the scale-free human gene coexpression network. *Molecular biology and evolution*, 21(11), 2058–2070.

[104] Kanehisa, M. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28(1), 27–30.

[105] Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., & Tanabe, M. (2011). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, 40(D1), D109–D114.

[106] Kanehisa, M., Minoru, K., Yoko, S., Masayuki, K., Miho, F., & Mao, T. (2015). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, 44(D1), D457–D462.

[107] Kelleher, R. J. & Shen, J. (2017). Presenilin-1 mutations and alzheimer's disease. *Proceedings of the National Academy of Sciences*, 114(4), 629–631.

[108] Kerr, M. K. & Churchill, G. A. (2001a). Experimental design for gene expression microarrays. *Biostatistics*, 2(2), 183–201.

[109] Kerr, M. K. & Churchill, G. A. (2001b). Statistical design and the analysis of gene expression microarray data. *Genetics Research*, 77(2), 123–128.

[110] Kim, E., Goren, A., & Ast, G. (2008). Alternative splicing: current perspectives. *Bioessays*, 30(1), 38–47.

[111] Kim, E., Magen, A., & Ast, G. (2006). Different levels of alternative splicing among eukaryotes. *Nucleic acids research*, 35(1), 125–131.

[112] Kim, H., Klein, R., Majewski, J., & Ott, J. (2004). Estimating rates of alternative splicing in mammals and invertebrates. *Nature genetics*, 36(9), 915–916.

[113] Kim, S.-Y. & Volsky, D. J. (2005). PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics*, 6, 144.

[114] Kiss, H., Darai, E., Kiss, C., Kost-Alimova, M., Klein, G., Dumanski, J. P., & Imreh, S. (2002). Comparative human/murine sequence analysis of the common eliminated region 1 from human 3p21. 3. *Mammalian genome*, 13(11), 646–655.

[115] Kong, Y., Wu, J., & Yuan, L. (2014). MicroRNA expression analysis of adult-onset drosophila alzheimer's disease model. *Curr. Alzheimer Res.*, 11(9), 882–891.

[116] Konovalova, S. & Tyynismaa, H. (2013). Mitochondrial aminoacyl-trna synthetases in human disease. *Molecular genetics and metabolism*, 108(4), 206–211.

[117] Kramer, M., Dutkowski, J., Yu, M., Bafna, V., & Ideker, T. (2014). Inferring gene ontologies from pairwise similarity data. *Bioinformatics*, 30(12), i34–42.

[118] Kumar, A., Singh, A., & Ekavali (2015). A review on alzheimer's disease pathophysiology and its management: an update. *Pharmacol. Rep.*, 67(2), 195–203.

[119] Labouesse, M. (1990). The yeast mitochondrial leucyl-tRNA synthetase is a splicing factor for the excision of several group I introns. *Mol. Gen. Genet.*, 224(2), 209–221.

[120] Labouesse, M., Netter, P., & Schroeder, R. (1984). Molecular basis of the 'box effect', A maturase deficiency leading to the absence of splicing of two introns located in two split genes of yeast mitochondrial DNA. *Eur. J. Biochem.*, 144(1), 85–93.

[121] Ladunga, I. S. (2009). Finding homologs in amino acid sequences using network blast searches. *Current protocols in bioinformatics*, (pp. 3–4).

[122] Laird, N. M. & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, (pp. 963–974).

[123] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822), 860–921.

[124] Landreth, G. E. & Reed-Geaghan, E. G. (2009). Toll-like receptors in alzheimer's disease. *Curr. Top. Microbiol. Immunol.*, 336, 137–153.

[125] Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B. E., Bickel, P., Brown, J. B., Cayting, P., et al. (2012). Chip-seq guidelines and practices of the encode and modencode consortia. *Genome research*, 22(9), 1813–1831.

[126] Langfelder, P. & Horvath, S. (2008). Wgcna: an r package for weighted correlation network analysis. *BMC Bioinformatics*, (1), 559.

[127] Langfelder, P., Luo, R., Oldham, M. C., & Horvath, S. (2011). Is my network module preserved and reproducible? *PLoS computational biology*, 7(1), e1001057.

[128] Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M., & Carey, V. (2013a). Software for computing and annotating genomic ranges. *PLoS Computational Biology*, 9.

[129] Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M., & Carey, V. (2013b). Software for computing and annotating genomic ranges. *PLoS Computational Biology*, 9.

[130] Le, H.-S., Oltvai, Z. N., & Bar-Joseph, Z. (2010). Cross-species queries of large gene expression databases. *Bioinformatics*, 26(19), 2416–2423.

[131] Ledoit, O., Olivier, L., & Michael, W. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10(5), 603–621.

[132] Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., & Irizarry, R. A. (2010a). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.*, 11(10), 733–739.

[133] Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., & Irizarry, R. A. (2010b). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10), 733–739.

[134] Lepelletier, F.-X., Mann, D. M. A., Robinson, A. C., Pinteaux, E., & Boutin, H. (2017). Early changes in extracellular matrix in alzheimer's disease. *Neuropathol. Appl. Neurobiol.*, 43(2), 167–182.

[135] Lewin, A. & Grieve, I. C. (2006). Grouping gene ontology terms to improve the assessment of gene set enrichment in microarray data. *BMC Bioinformatics*, 7, 426.

[136] Li, C. & Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proceedings of the National Academy of Sciences*, 98(1), 31–36.

[137] Li, J., Ning, Y., Hedley, W., Saunders, B., et al. (2002). The molecule pages database. *Nature*, 420(6916), 716.

[138] Li, Y., Agarwal, P., & Rajagopalan, D. (2008). A global pathway crosstalk network. *Bioinformatics*, 24(12), 1442–1447.

[139] Liang, W. S., Dunckley, T., Beach, T. G., Grover, A., Mastroeni, D., Walker, D. G., Caselli, R. J., Kukull, W. A., McKeel, D., Morris, J. C., Hulette, C., Schmechel, D., Alexander, G. E., Reiman, E. M., Rogers, J., & Stephan, D. A. (2007). Gene expression profiles in anatomically and functionally distinct regions of the normal aged human brain. *Physiol. Genomics*, 28(3), 311–322.

[140] Liao, B.-Y. & Zhang, J. (2005). Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Molecular biology and evolution*, 23(3), 530–540.

[141] Liao, B.-Y. & Zhang, J. (2006). Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Molecular biology and evolution*, 23(3), 530–540.

[142] Liao, Y., Smyth, G. K., & Shi, W. (2013). The subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Research*, 41, e108.

[143] Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P., & Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27(12), 1739–1740.

[144] Lin, S., Lin, Y., Nery, J. R., Urich, M. A., Breschi, A., Davis, C. A., Dobin, A., Zaleski, C., Beer, M. A., Chapman, W. C., et al. (2014). Comparison of the transcriptional landscapes between human and mouse tissues. *Proceedings of the National Academy of Sciences*, 111(48), 17224–17229.

[145] Lipman, D. J., Souvorov, A., Koonin, E. V., Panchenko, A. R., & Tatusova, T. A. (2002). The relationship of protein conservation and sequence length. *BMC Evolutionary Biology*, 2(1), 20.

[146] Liptak, T. (1958). On the combination of independent tests. *Magyar Tud Akad Mat Kutato Int Kozl*, 3, 171–197.

[147] Liu, Y., Liu, F., Grundke-Iqbal, I., Iqbal, K., & Gong, C.-X. (2011). Deficient brain insulin signalling pathway in alzheimer's disease and diabetes. *J. Pathol.*, 225(1), 54–62.

[148] Loughin, T. M. (2004). A systematic comparison of methods for combining p-values from independent tests. *Comput. Stat. Data Anal.*, 47(3), 467–485.

[149] Lu, S., Lee, J., Revelo, M., Wang, X., Lu, S., & Dong, Z. (2007). Smad3 is overexpressed in advanced human prostate cancer and necessary for progressive growth of prostate cancer cells in nude mice. *Clin. Cancer Res.*, 13(19), 5692–5702.

[150] Lu, Y., Rosenfeld, R., Simon, I., Nau, G. J., & Bar-Joseph, Z. (2008). A probabilistic generative model for GO enrichment analysis. *Nucleic Acids Res.*, 36(17), e109.

[151] Lueck, S., Thurley, K., Thaben, P. F., & Westermark, P. O. (2014). Rhythmic degradation explains and unifies circadian transcriptome and proteome data. *Cell Reports*, 9, 741–751.

[152] Lukk, M., Kapushesky, M., Nikkilä, J., Parkinson, H., Goncalves, A., Huber, W., Ukkonen, E., & Brazma, A. (2010). A global map of human gene expression. *Nat. Biotechnol.*, 28(4), 322–324.

[153] MacDonald, J. W. (2017a). *mogene10stprobeset.db: Affymetrix mogene10 annotation data (chip mogene10stprobeset)*. R package version 8.7.0.

[154] MacDonald, J. W. (2017b). *mogene10sttranscriptcluster.db: Affymetrix mogene10 annotation data (chip mogene10sttranscriptcluster)*. R package version 8.7.0.

[155] Macias, M. J., Martin-Malpartida, P., & Massagué, J. (2015). Structural determinants of smad function in TGF-β signaling. *Trends Biochem. Sci.*, 40(6), 296–308.

[156] Maere, S., Heymans, K., & Kuiper, M. (2005). BiNGO: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, 21(16), 3448–3449.

[157] Makałowski, W., Zhang, J., & Boguski, M. S. (1996). Comparative analysis of 1196 orthologous mouse and human full-length mrna and protein sequences. *Genome Research*, 6(9), 846–857.

[158] Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., & Califano, A. (2006). Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. In *BMC bioinformatics*, volume 7 (pp.S7).: BioMed Central.

[159] Markesbery, W. R. (1997). Oxidative stress hypothesis in alzheimer's disease. *Free Radic. Biol. Med.*, 23(1), 134–147.

[160] Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B., Kanapin, A., Lewis, S., Mahajan, S., May, B., Schmidt, E., Vastrik, I., Wu, G., Birney, E., Stein, L., & D'Eustachio, P. (2009). Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.*, 37(Database issue), D619–22.

[161] McCall, M. N., Bolstad, B. M., & Irizarry, R. A. (2010a). Frozen robust multiarray analysis (fRMA). *Biostatistics*, 11(2), 242–253.

[162] McCall, M. N., Bolstad, B. M., & Irizarry, R. A. (2010b). Frozen robust multiarray analysis (frma). *Biostatistics*, 11(2), 242–253.

[McCall & Irizarry] McCall, M. N. & Irizarry, R. A. *mogene.1.0.st.v1frmavecs: Vectors used by frma for microarrays of type mogene.1.0.st.v1frmavecs*. R package version 1.1.0.

[164] McCall, M. N., Jaffee, H. A., & Irizarry, R. A. (2012). frma st: frozen robust multiarray analysis for affymetrix exon and gene st arrays. *Bioinformatics*, 28(23), 3153–3154.

[165] McCall, M. N., Jaffee, H. A., Zelisko, S. J., Sinha, N., Hooiveld, G., Irizarry, R. A., & Zilliox, M. J. (2014). The gene expression barcode 3.0: improved data processing and mining tools. *Nucleic Acids Res.*, 42(Database issue), D938–43.

[166] McCall, M. N., Murakami, P. N., Lukk, M., Huber, W., & Irizarry, R. A. (2011a). Assessing affymetrix GeneChip microarray quality. *BMC Bioinformatics*, 12, 137.

[167] McCall, M. N., Uppal, K., Jaffee, H. A., Zilliox, M. J., & Irizarry, R. A. (2010c). The gene expression barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. *Nucleic acids research*, 39(suppl_1), D1011–D1015.

[168] McCall, M. N., Uppal, K., Jaffee, H. A., Zilliox, M. J., & Irizarry, R. A. (2011b). The gene expression barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. *Nucleic acids research*, 39(suppl 1), D1011–D1015.

[169] McCormack, T., Frings, O., Alexeyenko, A., & Sonnhammer, E. L. L. (2013). Statistical assessment of crosstalk enrichment between gene groups in biological networks. *PLoS One*, 8(1), e54945.

[170] McGeer, P. L., Itagaki, S., Boyes, B. E., & McGeer, E. G. (1988). Reactive microglia are positive for HLA-DR in the substantia nigra of parkinson's and alzheimer's disease brains. *Neurology*, 38(8), 1285–1291.

[171] Merico, D., Isserlin, R., Stueker, O., Emili, A., & Bader, G. D. (2010). Enrichment map: A Network-Based method for Gene-Set enrichment visualization and interpretation. *PLoS One*, 5(11), e13984.

[172] Merkin, J., Russell, C., Chen, P., & Burge, C. B. (2012). Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science*, 338(6114), 1593–1599.

[173] Mestas, J. & Hughes, C. C. (2004). Of mice and not men: differences between mouse and human immunology. *The Journal of Immunology*, 172(5), 2731–2738.

[174] Miller, J. A., Horvath, S., & Geschwind, D. H. (2010). Divergence of human and mouse brain transcriptome highlights alzheimer disease pathways. *Proceedings of the National Academy of Sciences*, 107(28), 12698–12703.

[175] Miners, J. S., Schulz, I., & Love, S. (2017). Differing associations between Aβ accumulation, hypoperfusion, blood-brain barrier dysfunction and loss of PDGFRB pericyte marker in the precuneus and parietal white matter in alzheimer's disease. *J. Cereb. Blood Flow Metab.*, (pp. 271678X17690761).

[176] Mootha, V. K., Bunkenborg, J., Olsen, J. V., Hjerrild, M., Wisniewski, J. R., Stahl, E., Bolouri, M. S., Ray, H. N., Sihag, S., Kamal, M., et al. (2003). Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria. *Cell*, 115(5), 629–640.

[177] Morata, J., Béjar, S., Talavera, D., Riera, C., Lois, S., de Xaxars, G. M., & de la Cruz, X. (2013). The relationship between gene isoform multiplicity, number of exons and protein divergence. *PloS one*, 8(8), e72742.

[178] Morawski, M., Brückner, G., Jäger, C., Seeger, G., Matthews, R. T., & Arendt, T. (2012). Involvement of perineuronal and perisynaptic extracellular matrix in alzheimer's disease neuropathology. *Brain Pathol.*, 22(4), 547–561.

[179] Mount, D. W. (2008). Using gaps and gap penalties to optimize pairwise sequence alignments. *Cold Spring Harbor Protocols*, 2008(6), pdb–top40.

[180] Naba, A., Clauser, K. R., Hoersch, S., Liu, H., Carr, S. A., & Hynes, R. O. (2011). The matrisome: In silico definition and in vivo characterization by proteomics of normal and tumor extracellular matrices. *Mol. Cell. Proteomics*, 11(4), M111.014647–M111.014647.

[181] Naeem, H., Zimmer, R., Tavakkolkhah, P., & Küffner, R. (2012). Rigorous assessment of gene set enrichment tests. *Bioinformatics*, 28(11), 1480–1486.

[182] Nagao, T., Leuzinger, S., Acampora, D., Simeone, A., Finkelstein, R., Reichert, H., & Furukubo-Tokunaga, K. (1998). Developmental rescue of drosophila cephalic defects by the human otx genes. *Proceedings of the National Academy of Sciences*, 95(7), 3737–3742.

[183] Nawaz, M. H., Pang, Y. L., & Martinis, S. A. (2007). Molecular and functional dissection of a putative RNA-binding region in yeast mitochondrial leucyl-tRNA synthetase. *J. Mol. Biol.*, 367(2), 384–394.

[184] Nicholson, D. E. (2001). Iubmb-nicholson metabolic pathways charts. *Biochemistry and Molecular Biology Education*, 29(2), 42–44.

[185] Nishimura, D. & Darryl, N. (2001). BioCarta. *Biotech Software & Internet Report*, 2(3), 117–120.

[186] Nurtdinov, R. N., Artamonova, I. I., Mironov, A. A., & Gelfand, M. S. (2003). Low conservation of alternative splicing patterns in the human and mouse genomes. *Human Molecular Genetics*, 12(11), 1313–1320.

[187] Ogris, C., Guala, D., Helleday, T., & Sonnhammer, E. L. L. (2017). A novel method for crosstalk analysis of biological networks: improving accuracy of pathway annotation. *Nucleic Acids Res.*, 45(2), e8.

[188] Ohno, S. (2013). *Evolution by gene duplication*. Springer Science & Business Media.

[189] Oldham, M. C., Horvath, S., & Geschwind, D. H. (2006). Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proceedings of the National Academy of Sciences*, 103(47), 17973–17978.

[190] Oldham, M. C., Konopka, G., Iwamoto, K., Langfelder, P., Kato, T., Horvath, S., & Geschwind, D. H. (2008). Functional organization of the transcriptome in human brain. *Nature neuroscience*, 11(11), 1271.

[191] Orcholski, M. E., Zhang, Q., & Bredesen, D. E. (2011). Signaling via amyloid precursor-like proteins APLP1 and APLP2. *J. Alzheimers. Dis.*, 23(4), 689–699.

[192] Pagès, H. (2016). *BSgenome: Infrastructure for Biostrings-based genome data packages and support for efficient SNP representation*. R package version 1.40.1.

[193] Pagès, H., Aboyoun, P., Gentleman, R., & DebRoy, S. (2016). *Biostrings: String objects representing biological sequences, and matching algorithms*. R package version 2.40.2.

[194] Painter, T. S. (1928). A Comparison of the Chromosomes of the Rat and Mouse with Reference to the Question of Chromosome Homology in Mammals. *Genetics*, 13(2), 180–189.

[195] Pan, Q., Bakowski, M. A., Morris, Q., Zhang, W., Frey, B. J., Hughes, T. R., & Blencowe, B. J. (2005). Alternative splicing of conserved exons is frequently species-specific in human and mouse. *Trends in Genetics*, 21(2), 73–77.

[196] Pan, Q., Shai, O., Lee, L. J., Frey, B. J., & Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature genetics*, 40(12), 1413–1415.

[197] Paris, D., Ait-Ghezala, G., Bachmeier, C., Laco, G., Beaulieu-Abdelahad, D., Lin, Y., Jin, C., Crawford, F., & Mullan, M. (2014). The spleen tyrosine kinase (syk) regulates alzheimer amyloid-β production and tau hyperphosphorylation. *J. Biol. Chem.*, 289(49), 33927–33944.

[198] Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods*, 14(4), 417.

[199] Pearson, H. (2007). Meet the human metabolome. *Nature*, 446(7131), 8.

[200] Pearson, W. R. (2013). An introduction to sequence similarity ("homology") searching. *Current protocols in bioinformatics*, (pp. 3–1).

[201] Pevsner, J. (2015). *Bioinformatics and Functional Genomics*. John Wiley & Sons.

[202] Pickrell, J. K., Pai, A. A., Gilad, Y., & Pritchard, J. K. (2010). Noisy splicing drives mrna isoform diversity in human cells. *PLoS genetics*, 6(12), e1001236.

[203] Pierce, S. B., Gersak, K., Michaelson-Cohen, R., Walsh, T., Lee, M. K., Malach, D., Klevit, R. E., King, M.-C., & Levy-Lahad, E. (2013). Mutations in lars2, encoding mitochondrial leucyl-trna synthetase, lead to premature ovarian failure and hearing loss in perrault syndrome. *The American Journal of Human Genetics*, 92(4), 614–620.

[204] Pinheiro, J. C. & Bates, D. M. (2000). Mixed-effects models in s and s-plus springer. *New York*.

[205] Popovici, V., Goldstein, D. R., Antonov, J., Jaggi, R., Delorenzi, M., & Wirapati, P. (2009). Selecting control genes for RT-QPCR using public microarray data. *BMC Bioinformatics*, 10, 42.

[206] Pritykin, Y., Ghersi, D., & Singh, M. (2015). Genome-Wide detection and analysis of multifunctional genes. *PLoS Comput. Biol.*, 11(10), e1004467.

[207] Pujol, A., Mosca, R., Farrés, J., & Aloy, P. (2010). Unveiling the role of network and systems biology in drug discovery. *Trends Pharmacol. Sci.*, 31(3), 115–123.

[208] Quiring, R., Walldorf, U., Kloter, U., & Gehring, W. J. (1994). Homology of the eyeless gene of drosophila to the small eye gene in mice and aniridia in humans. *Science*, 265(5173), 785–789.

[209] Raghava, G. P. & Barton, G. J. (2006). Quantification of the variation in percentage identity for protein sequence alignments. *BMC bioinformatics*, 7(1), 415.

[210] Ramanan, V. K., Shen, L., Moore, J. H., & Saykin, A. J. (2012). Pathway analysis of genomic data: concepts, methods, and prospects for future development. *Trends Genet.*, 28(7), 323–332.

[211] Ranganathan, S., Scudiere, S., & Bowser, R. (2001). Hyperphosphorylation of the retinoblastoma gene product and altered subcellular distribution of E2F-1 during alzheimer's disease and amyotrophic lateral sclerosis. *J. Alzheimers. Dis.*, 3(4), 377–385.

[212] Rhodes, D. R., Kalyana-Sundaram, S., Tomlins, S. A., Mahavisno, V., Kasper, N., Varambally, R., Barrette, T. R., Ghosh, D., Varambally, S., & Chinnaiyan, A. M. (2007a). Molecular concepts analysis links tumors, pathways, mechanisms, and drugs. *Neoplasia*, 9(5), 443–454.

[213] Rhodes, D. R., Kalyana-Sundaram, S., Tomlins, S. A., Mahavisno, V., Kasper, N., Varambally, R., Barrette, T. R., Ghosh, D., Varambally, S., & Chinnaiyan, A. M. (2007b). Molecular concepts analysis links tumors, pathways, mechanisms, and drugs. *Neoplasia*, 9(5), 443–454.

[214] Risso, A. (2000). Leukocyte antimicrobial peptides: multifunctional effector molecules of innate immunity. *Journal of leukocyte biology*, 68(6), 785–792.

[215] Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7), e47.

[216] Robert, C. & Watson, M. (2015). Errors in rna-seq quantification affect genes of relevance to human disease. *Genome biology*, 16(1), 177.

[217] Romanenko, S. A., Perelman, P. L., Serdukova, N. A., Trifonov, V. A., Biltueva, L. S., Wang, J., Li, T., Nie, W., O'Brien, P. C., Volobouev, V. T., Stanyon, R., Ferguson-Smith, M. A., Yang, F., & Graphodatsky, A. S. (2006). Reciprocal chromosome painting between three laboratory rodent species. *Mamm. Genome*, 17(12), 1183–1192.

[218] Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M.,

Milstein, S., Rosenberg, J., Goldberg, D. S., Zhang, L. V., Wong, S. L., Franklin, G., Li, S., Albala, J. S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R. S., Vandenhaute, J., Zoghbi, H. Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M. E., Hill, D. E., Roth, F. P., & Vidal, M. (2005). Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062), 1173–1178.

[219] Rung, J. & Brazma, A. (2013). Reuse of public genome-wide gene expression data. *Nat. Rev. Genet.*, 14(2), 89–99.

[220] Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., & Buetow, K. H. (2009). PID: the pathway interaction database. *Nucleic Acids Res.*, 37(Database issue), D674–9.

[221] Schäfer, J. & Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.*, 4, Article32.

[222] Scheltens, P., Blennow, K., Breteler, M. M. B., de Strooper, B., Frisoni, G. B., Salloway, S., & Van der Flier, W. M. (2016). Alzheimer's disease. *Lancet*, 388(10043), 505–517.

[223] Schena, M. & Knudsen, S. (2004). *Guide to Analysis of DNA Microarray Data, 2nd Edition and Microarray Analysis Set*. Wiley-Liss.

[224] Schreitmüller, B., Leyhe, T., Stransky, E., Köhler, N., & Laske, C. (2012). Elevated angiopoietin-1 serum levels in patients with alzheimer's disease. *Int. J. Alzheimers. Dis.*, 2012, 324016.

[225] Schwenzer, H., Zoll, J., Florentz, C., & Sissler, M. (2013). Pathogenic implications of human mitochondrial aminoacyl-trna synthetases. In *Aminoacyl-tRNA Synthetases in Biology and Medicine* (pp. 247–292). Springer.

[226] Seita, J., Sahoo, D., Rossi, D. J., Bhattacharya, D., Serwold, T., Inlay, M. A., Ehrlich, L. I. R., Fathman, J. W., Dill, D. L., & Weissman, I. L. (2012). Gene expression commons: an open platform for absolute gene expression profiling. *PLoS One*, 7(7), e40321.

[227] Sethi, M. K. & Zaia, J. (2017). Extracellular matrix proteomics in schizophrenia and alzheimer's disease. *Anal. Bioanal. Chem.*, 409(2), 379–394.

[228] Sharov, A. A., Schlessinger, D., & Ko, M. S. H. (2015). ExAtlas: An interactive online tool for meta-analysis of gene expression data. *J. Bioinform. Comput. Biol.*, 13(6), 1550019.

[229] Sherrington, R., Rogaev, E. I., Liang, Y., Rogaeva, E. A., Levesque, G., Ikeda, M., Chi, H., Lin, C., Li, G., Holman, K., Tsuda, T., Mar, L., Foncin, J. F., Bruni, A. C., Montesi, M. P., Sorbi, S., Rainero, I., Pinessi, L., Nee, L., Chumakov, I., Pollen, D., Brookes, A., Sanseau, P., Polinsky, R. J., Wasco, W., Da Silva, H. A., Haines, J. L., Perkicak-Vance, M. A., Tanzi, R. E., Roses, A. D., Fraser, P. E., Rommens, J. M., & St George-Hyslop, P. H. (1995). Cloning of a gene bearing missense mutations in early-onset familial alzheimer's disease. *Nature*, 375(6534), 754–760.

[230] Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, (pp. 238–241).

[231] Snel, B., Van Noort, V., & Huynen, M. A. (2004). Gene co-regulation is highly conserved in the evolution of eukaryotes and prokaryotes. *Nucleic acids research*, 32(16), 4725–4731.

[232] Stamatoyannopoulos, J. A., Snyder, M., Hardison, R., Ren, B., Gingeras, T., Gilbert, D. M., Groudine, M., Bender, M., Kaul, R., Canfield, T., et al. (2012). An encyclopedia of mouse dna elements (mouse encode). *Genome biology*, 13(8), 418.

[233] Stelzer, G., Plaschkes, I., Oz-Levi, D., Alkelai, A., Olender, T., Zimmerman, S., Twik, M., Belinky, F., Fishilevich, S., Nudel, R., Guan-Golan, Y., Warshawsky, D., Dahary, D., Kohn, A., Mazor, Y., Kaplan, S., Iny Stein, T., Baris, H. N., Rappaport, N., Safran, M., & Lancet, D. (2016). VarElect: the phenotype-based variation prioritizer of the GeneCards suite. *BMC Genomics*, 17 Suppl 2, 444.

[234] Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksöz, E., Droege, A., Krobitsch, S., Korn, B., Birchmeier, W., Lehrach, H., & Wanker, E. E. (2005). A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122(6), 957–968.

[235] Steuer, R., Kurths, J., Daub, C. O., Weise, J., & Selbig, J. (2002). The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics*, 18(suppl_2), S231–S240.

[236] Stuart, J. M., Segal, E., Koller, D., & Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *science*, 302(5643), 249–255.

[237] Studer, R. A. & Robinson-Rechavi, M. (2009). How confident can we be that orthologs are similar, but paralogs differ? *Trends in Genetics*, 25(5), 210–216.

[238] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, 102(43), 15545–15550.

[239] Sudmant, P. H., Alexis, M. S., & Burge, C. B. (2015). Meta-analysis of rna-seq expression data across species, tissues and studies. *Genome biology*, 16(1), 287.

[240] Takeda, J.-i., Suzuki, Y., Sakate, R., Sato, Y., Seki, M., Irie, T., Takeuchi, N., Ueda, T., Nakao, M., Sugano, S., et al. (2008). Low conservation and species-specific evolution of alternative splicing in humans and mice: comparative genomics analysis using well-annotated full-length cdnas. *Nucleic acids research*, 36(20), 6386–6395.

[241] Tan, Y., Wu, F., Tamayo, P., Haining, W. N., & Mesirov, J. P. (2015). Constellation map: Downstream visualization and interpretation of gene set enrichment results. *F1000Res.*, 4, 167.

[242] Tatusov, R. L., Koonin, E. V., & Lipman, D. J. (1997). A genomic perspective on protein families. *Science*, 278(5338), 631–637.

[Tenenbaum] Tenenbaum, D. *KEGGREST: Client-side REST access to KEGG*.

[244] Thakur, A., Siedlak, S. L., James, S. L., Bonda, D. J., Rao, A., Webber, K. M., Camins, A., Pallàs, M., Casadesus, G., Lee, H.-G., Bowser, R., Raina, A. K., Perry, G., Smith, M. A., & Zhu, X. (2008). Retinoblastoma protein phosphorylation at multiple sites is associated with neurofibrillary pathology in alzheimer disease. *Int. J. Clin. Exp. Pathol.*, 1(2), 134–146.

[245] Thorrez, L., Van Deun, K., Tranchevent, L.-C., Van Lommel, L., Engelen, K., Marchal, K., Moreau, Y., Van Mechelen, I., & Schuit, F. (2008). Using ribosomal protein genes as reference: a tale of caution. *PLoS One*, 3(3), e1854.

[246] Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., & Pachter, L. (2012). Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nature protocols*, 7(3), 562–578.

[247] Ueda, Y., Okano, M., Williams, C., Chen, T., Georgopoulos, K., & Li, E. (2006). Roles for dnmt3b in mammalian development: a mouse model for the icf syndrome. *Development*, 133(6), 1183–1192.

[248] Van Dam, D. & De Deyn, P. P. (2011). Animal models in the drug discovery pipeline for alzheimer's disease. *British journal of pharmacology*, 164(4), 1285–1300.

[249] van Dam, S., Võsa, U., van der Graaf, A., Franke, L., & de Magalhães, J. P. (2017). Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in bioinformatics*, (pp. bbw139).

[250] Végh, M. J., Heldring, C. M., Kamphuis, W., Hijazi, S., Timmerman, A. J., Li, K. W., van Nierop, P., Mansvelder, H. D., Hol, E. M., Smit, A. B., & van Kesteren, R. E. (2014). Reducing hippocampal extracellular matrix reverses early memory deficits in a mouse model of alzheimer's disease. *Acta Neuropathol Commun*, 2, 76.

[251] Vingron, M. & Waterman, M. S. (1994). Sequence alignment and penalty choice: Review of concepts, case studies and implications. *Journal of molecular biology*, 235(1), 1–12.

[252] Vivar, J. C., Pemu, P., McPherson, R., & Ghosh, S. (2013). Redundancy control in pathway databases (ReCiPa): An application for improving Gene-Set enrichment analysis in omics studies and "big data" biology. *OMICS*, 17(8), 414–422.

[253] von Bernhardi, R., Cornejo, F., Parada, G. E., & Eugenín, J. (2015). Role of TGFβ signaling in the pathogenesis of alzheimer's disease. *Front. Cell. Neurosci.*, 9, 426.

[254] Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P., & Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221), 470–476.

[255] Wang, T., Gu, J., Yuan, J., Tao, R., Li, Y., & Li, S. (2013). Inferring pathway crosstalk networks using gene set co-expression signatures. *Mol. Biosyst.*, 9(7), 1822–1828.

[256] Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., Grouios, C., Kazi, F., Lopes, C. T., Maitland, A., Mostafavi, S., Montojo, J., Shao, Q., Wright, G., Bader, G. D., & Morris, Q. (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.*, 38(Web Server issue), W214–20.

[257] Weinstein, J. (1997). Cell cycle-regulated expression, phosphorylation, and degradation of p55cdc. a mammalian homolog of CDC20/Fizzy/slp1. *J. Biol. Chem.*, 272(45), 28501–28511.

[258] Weinstein, J., Jacobsen, F. W., Hsu-Chen, J., Wu, T., & Baum, L. G. (1994). A novel mammalian protein, p55CDC, present in dividing cells is associated with protein kinase activity and has homology to the saccharomyces cerevisiae cell division cycle proteins cdc20 and cdc4. *Mol. Cell. Biol.*, 14(5), 3350–3363.

[259] Wilhelmus, M. M. M., Bol, J. G. J. M., van Duinen, S. G., & Drukarch, B. (2013). Extracellular matrix modulator lysyl oxidase colocalizes with amyloid-beta pathology in alzheimer's disease and hereditary cerebral hemorrhage with amyloidosis–dutch type. *Exp. Gerontol.*, 48(2), 109–114.

[260] with contributions from Andrew J. Bass, J. D. S., Dabney, A., & Robinson, D. (2015). *qvalue: Q-value estimation for false discovery rate control*. R package version 2.4.2.

[261] Wu, Y.-C., Bansal, M. S., Rasmussen, M. D., Herrero, J., & Kellis, M. (2014). Phylogenetic identification and functional characterization of orthologs and paralogs across human, mouse, fly, and worm. *bioRxiv*, (pp. 005736).

[262] Xing, Y., Ouyang, Z., Kapur, K., Scott, M. P., & Wong, W. H. (2007). Assessing the conservation of mammalian gene expression using high-density exon arrays. *Molecular biology and evolution*, 24(6), 1283–1285.

[263] Xue, L., Cai, J.-Y., Ma, J., Huang, Z., Guo, M.-X., Fu, L.-Z., Shi, Y.-B., & Li, W.-X. (2013). Global expression profiling reveals genetic programs underlying the developmental divergence between mouse and human embryogenesis. *BMC genomics*, 14(1), 568.

[264] Yamaguchi, Y. (2000). Lecticans: organizers of the brain extracellular matrix. *Cell. Mol. Life Sci.*, 57(2), 276–289.

[265] Yan, H., Zhu, X., Xie, J., Zhao, Y., & Liu, X. (2016). β-amyloid increases neurocan expression through regulating sox9 in astrocytes: A potential relationship between sox9 and chondroitin sulfate proteoglycans in alzheimer's disease. *Brain Res.*, 1646, 377–383.

[266] Yanai, I., Graur, D., & Ophir, R. (2004). Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. *Omics: a journal of integrative biology*, 8(1), 15–24.

[267] Yates, A., Akanni, W., Amode, M. R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., et al. (2015). Ensembl 2016. *Nucleic acids research*, 44(D1), D710–D716.

[268] Yu, Y. & Ye, R. D. (2015). Microglial Aβ receptors in alzheimer's disease. *Cell. Mol. Neurobiol.*, 35(1), 71–83.

[269] Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., Pope, B. D., et al. (2014). A comparative encyclopedia of dna elements in the mouse genome. *Nature*, 515(7527), 355–364.

[270] Zagorski, W., Castaing, B., Herbert, C. J., Labouesse, M., Martin, R., & Slonimski, P. P. (1991). Purification and characterization of the Saccharomyces cerevisiae mitochondrial leucyl-tRNA synthetase. *J. Biol. Chem.*, 266(4), 2537–2541.

[271] Zambelli, F., Pavesi, G., Gissi, C., Horner, D. S., & Pesole, G. (2010). Assessment of orthologous splicing isoforms in human and mouse orthologous genes. *BMC genomics*, 11(1), 534.

[272] Zambrano, A., Otth, C., Mujica, L., Concha, I. I., & Maccioni, R. B. (2007). Interleukin-3 prevents neuronal death induced by amyloid peptide. *BMC Neurosci.*, 8, 82.

[273] Zhang, B. & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1).

[274] Zhang, R., Lahens, N. F., Ballance, H. I., Hughes, M. E., & Hogenesch, J. B. (2014a). A circadian gene expression atlas in mammals: implications for biology and medicine. *Proceedings of the National Academy of Sciences*, 111(45), 16219–16224.

[275] Zhang, W., Zang, Z., Song, Y., Yang, H., & Yin, Q. (2014b). Coâ□□expression network analysis of differentially expressed genes associated with metastasis in prolactin pituitary tumors. *Mol. Med. Rep.*, 10(1), 113–118.

[276] Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., et al. (2008). Model-based analysis of chip-seq (macs). *Genome biology*, 9(9), R137.

[277] Zhao, S., Xi, L., & Zhang, B. (2015). Union exon based approach for rna-seq gene quantification: To be or not to be? *PloS one*, 10(11), e0141910.