Instituto Tecnologico y de Estudios Superiores de Monterrey

Monterrey Campus

School of Engineering and Sciences

Computational estimation of system-level gene coexpression across human tissues

A thesis presented by

Miguel Ángel Cortés Guzmán

Submitted to the
School of Engineering and Sciences
in partial fulfillment of the requirements for the degree of

Master of Science

in

Computer Science

Monterrey, Nuevo León, October, 2020

Dedication

To my mother and father, who believed in me and supported my new aspirations and goals. Thank you for everything.
To my friends: A, who listened when I needed to talk. N, who kept me a beautiful company and F, who inspired me to embark on this journey in the first place.

Acknowledgements

The author of this document wishes to thank Tecnológico de Monterrey for tuition support and CONACyT for living expenses support. Both institutions have made it possible for me to work on my research and to continue my preparation for becoming a more proficient researcher.

I also wish to thank the professors of the Master in Computer Science program at Tecnológico de Monterrey for all their invaluable lessons and interest in the preparation of young researchers like my fellow trainees and myself. Special thanks to Dr. Victor Treviño, my thesis advisor, who pushed me into learning many new things and abilities, as well as into always working hard during my research.

Computational estimation of system-level gene coexpression across human tissues
by
Miguel Ángel Cortés Guzmán

Abstract

Large-scale gene coexpression projects have been a valuable resource for researchers involved in bioinformatics, molecular biology and biomedical sciences as they provide support for formulating hypotheses regarding gene functions and interactions, as well as for prioritizing genes in experimental designs. Such projects however, contain results calculated from all sorts of samples including healthy, disease and experimental condition specimens in addition to many of them not being based on sequencing technologies. The understanding of normality in the context of human gene coexpression is pivotal as this helps uncovering new functional associations for previously known or unknown genes and it serves as a comparison point when studying disease states. Other tools besides the Pearson Correlation Coefficient have not been traditionally explored for large-scale coexpression, potentially letting more complex non-linear associations between genes pass. In this computer science master thesis, a system-level coexpression estimation across a variety of normal human tissues is proposed. The objective is not only improve on the current areas of opportunity that exist in the large-scale coexpression research domain, but to also provide the scientific community with a novel and useful resource of system-level human coexpression data. Results comprise the first large-scale coexpression estimation in the literature that exclusively considers normal samples in the input data that were profiled with sequencing technologies in combination with 3 distinct coexpression metrics considered for calculation: the Pearson Correlation Coefficient, the Spearman Rank Correlation Coefficient and the highly interpretable Chi-square test of independence.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

In the course of the last decade, revolutionary advances in the field of molecular biology have allowed for the characterization of human genes and features associated with them through sequencing technologies [1]. These technologies make it possible for researchers to obtain the sequence of building blocks constituting complex molecules such as Deoxyribonucleic Acid (DNA) and Ribonucleic Acid (ARN) which in turn allows for the study of genes at the molecular level. Despite the high specialization of these sequencing technologies, every day they are becoming more accessible and affordable to researchers around the world. As a result, massive amounts of genomic data generated from these studies have been deposited in public scientific repositories at a growing rate across the years [2]. This gives rise to the invaluable opportunity of systematically processing and analyzing these experiments by using computational tools and techniques to uncover useful biological knowledge from the initial raw data.

The measurement of a gene's activity is one capability that sequencing technologies have made possible at a higher resolution than ever before, this measurement is known as gene expression and it can be quantified in biological samples for thousands of genes at the same time [1]. A widely applied analysis consists of comparing gene expression between samples in different conditions (e.g. Healthy against diseased) while looking for statistically significant genes differing in their expression [3]. This classic differential expression analysis provides insight into, for example, what genes are involved in disease or experimental states. This allows for the generation of hypotheses proposing new diagnostic biomarkers and genes with patient survival prediction potential [4, 5]. Differential expression analysis is an example of the usefulness of gene expression data even when considering genes just at an individual level and not studying the relationships between them [6]. However, many other analyses for gene expression data exist. In some of them, the relationships between genes are pivotal in helping the discovery of new gene functions and inter-regulations.

Genes do not act alone, but rather they are organized in complex networks of interactions that require groups of genes to be expressed in a coordinated way to carry out biological functions [7]. Therefore expression levels of genes may be related directly or indirectly to one another. This way of studying gene expression provides insight into how genes behave dynamically in living organisms. It is inspired in the systems biology philosophy where living organisms can only be understood as a complex system of interactions between its parts [8].

One way to study interactions between genes for large datasets such as the ones generated by gene expression profiling technologies is by performing gene coexpression analysis [9–14], a technique which uses computational and statistical tools to analyze the patterns that exist between the expression of pairs of genes.

Over the years, coexpression databases have been developed to provide researchers in molecular biology and bioinformatics with an accessible and quick reference to a large amount of coexpression data for many different genes [15–17]. This information is valuable for many applications including formulating hypotheses regarding gene interactions and functions of poorly understood genes as well as aiding the design of both laboratory and computational experiments targeting genes associated with specific biological functions. Researchers may prioritize the study of genes based on the strength of the relationships present in a network of coexpressed genes enriched with biological functions of interest. Databases allow researchers from all areas to access data of this nature without having to select and process expression data themselves, a task which can quickly turn into a time-consuming process depending on the volume of the data and that requires considerable programming skills [15–17]. Databases bring the benefits of coexpression analysis to a broader audience in the scientific community and propel downstream research.

Although existing human coexpression databases in the literature have been very useful to investigate interesting relationships between genes, none of them consider only normal samples profiled with sequencing technologies as their source data [9, 15, 17–20]. Such databases use gene expression measurements obtained with non-state-of-the-art profiling technologies such as the microarray [21], or include a combination of normal, disease, developmental, and under experimental conditions samples which could not be representative when trying to study gene coexpression as it is in normal human biology. Additionally, these existing databases have focused only on the Pearson Correlation Coefficient (PCC) to characterize the coexpression associations between genes [22]. While the PCC is useful as a metric in coexpression, other metrics have been applied in different fields to find associations between variables in large datasets. These metrics could have the potential of uncovering complex associations between the expression of genes and should be explored in the context of coexpression [23].

In this work, relationships between pairs of a vast collection of human genes are quantitatively estimated by using gene expression measurements from a variety of healthy human samples from different tissues as well as a collection of robust statistical and computational techniques. To the best understanding of the author of this work, this would be the first time that this kind of estimation is made at a system-level by using only samples from healthy donors and human gene expression data obtained with sequencing technologies. This estimation is important because it provides a reference baseline data source of human gene coexpression which does not exist yet and that can be used for comparative purposes, and for guiding the discovery of the function of previously uncharacterized genes in the context of normal human biology.

In this chapter, an introduction to coexpression analysis is given as well as the motivation, research questions, objectives, and general methodology adopted in this master thesis project. A revision of important theoretical concepts to understand the work and related work in literature is also addressed. Chapter 2 documents all the statistical and computational techniques required to carry out this work. It also includes the rationale behind novel concepts proposed in this project for coexpression estimation at the system-level. Both Chapter 3 and 4 report results obtained in this work. The former deals with results that needed to be obtained beforehand to make the system-level coexpression estimation proposed robust. The latter contains the main results pertaining coexpression calculation perse. Finally, Chapter 5 discusses the findings and future work stemming from this project in addition to closing remarks.

1.1 Gene coexpression: concept and use

Gene coexpression analysis is an important tool in systems biology as it allows for studying genes considering them as interconnected members of biological networks instead of standalone entities. It is believed that two genes are coexpressed when they exhibit coordinated behaviors in their expression as evidenced by looking at their expression levels across a common set of samples [24]. Many of these coordinated behaviors when they exist are commonly linear (refer to Figure 1.1 for examples), but other kinds of non-linear relationships may also exist [23].

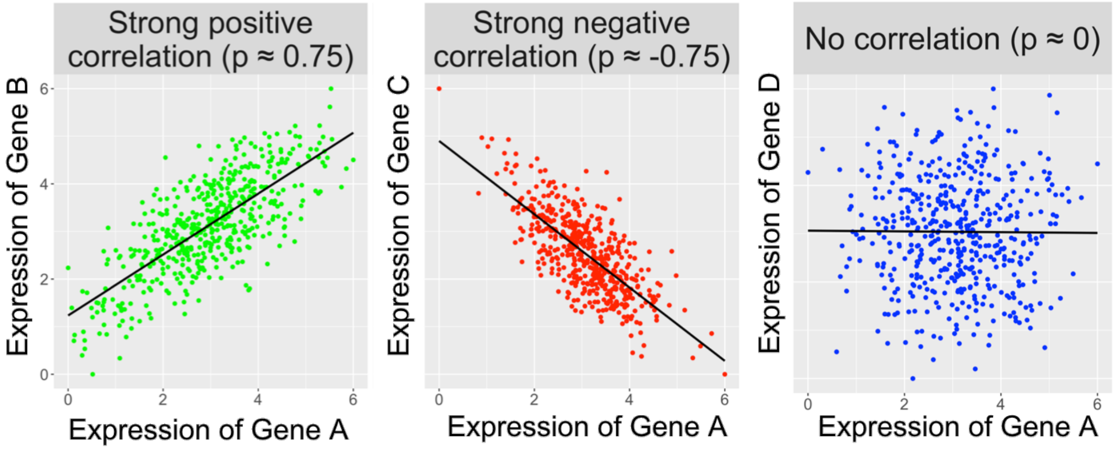


Figure 1.1: Example of 3 different patterns of relative expression between pairs of genes quantified with the Pearson Correlation Coefficient p for simulated data (2.5.2). From left to right: linear positive coexpression in green (as the expression of one gene increases, so does the expression of the other proportionally), linear negative coexpression in red (as the expression of one gene increases, the expression of the other proportionally) and no coexpression in blue (expression of the genes relative to one another is distributed randomly).

The idea behind coexpression analysis is fairly simple in essence: analyze every possible combination of two genes present in an expression dataset. The latter is done by computing a coexpression measure whose result is determined by the observed expression across samples of each pair of genes. Once all pairs have been analyzed, the results can be interpreted as a graph in which nodes are genes and edges exist between any two genes if a strong coexpression measure is observed between them. Densely interconnected clusters of nodes in these coexpression graphs are of great interest as coexpression is correlated with cofunctionality [25], or in other words, coexpressed genes have a higher chance of being functionally related than genes chosen at random. This property of coexpressed genes is especially useful when trying to get an idea of what an experimentally unstudied gene does or when looking for the best candidates to invest in for a study at the lab investigating novel genes participating in biological processes.

The formulation of hypotheses regarding the function of poorly understood genes based on their neighbors in the coexpression graph with well-characterized functions is known as Guilt-By-Association analysis (GBA) [9], a well-known application of coexpression. Another application would be to study potential disease mechanisms when comparing how gene relationships estimated through coexpression have changed between healthy controls and disease states [26]. Disease gene prioritization to save time and resources at the lab is also possible by looking at a gene's coexpressed partners involved in the disease of interest and estimating a likelihood that the candidate gene is also involved [27]. The concept of coexpression can be extended to not just diseases, but the study of any biological function or feature related to sets of genes. Coexpression data can be integrated with other information from molecular profiling experiments such as those characterizing regulatory elements in the human genome, methylation, and genetic variants to boost the ability of coexpression networks of inferring gene functions [9].

Important applications of coexpression are summarized in Figure 1.2 where there are also some basic intuitions into what these applications achieve. More details on coexpression are given in Section 1.3.2 of theoretical background in this chapter.

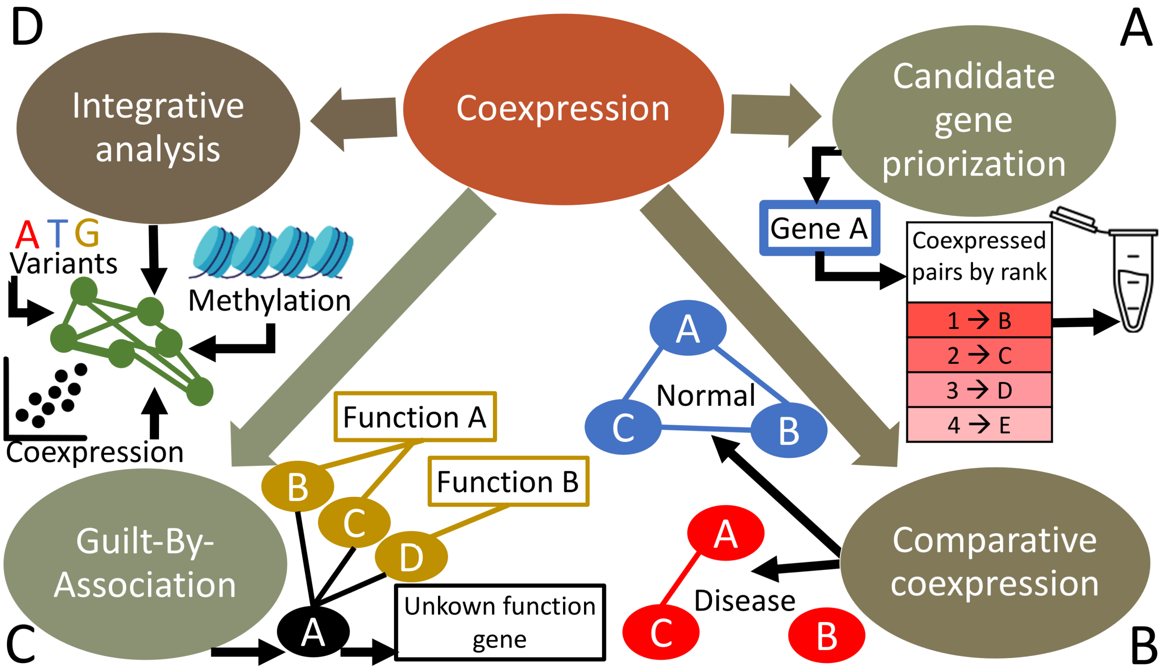


Figure 1.2: Some important applications of coexpression analysis. A: genes that are strongly coexpressed with a gene known to be implicated in biological functions, diseases, or processes of interest are good candidates for further research [27]. Coexpression serves as a guide to try and conduct downstream experiments that are more likely to have interesting findings (e.g. At the wet lab) potentially saving time and resources over an unguided search for genes associated with a particular function, disease, etc. B: comparison of the coexpression network between two conditions. Coexpression relationships may appear, disappear, or change their strength across conditions [28]. Investigating this provides insight into what gene associations are behind the observed phenotypes. C: GBA allows for assigning putative functions to poorly characterized genes based on their coexpressed gene pairs with known functions. Many implementations exist for this, but popular choices include variations of voting algorithms [29]. D: coexpression data is flexible and can be integrated with other kinds of genomic data to build hybrid networks [9].

1.2 Proposal

This section outlines the scientific needs that this project is attempting to answer as well as the research questions raising in their development. The objectives of the project are also mentioned along with the methodological steps to achieve these objectives.

1.2.1 Motivation and justification

This project addresses the following areas of opportunity and needs that currently exist in the domain of large-scale human coexpression:

- No publicly available coexpression data exists that is calculated only based on healthy human tissues to serve as a reference of normal state coexpression, and that is derived from the most recent gene expression profiling technology

- Only one coexpression metric has been traditionally used as the base for large-scale gene coexpression projects, trying new metrics may help discover other kinds of coexpression patterns between genes that the traditional method cannot detect while also boosting interpretability of the results

- Current coexpression databases hosting data derived from the most recent gene expression profiling technology only provide the raw data, without options to further explore coexpressions through visualizations and analysis tools

In this work, a robust system-level human gene coexpression estimation is carried out. This would be the first time in literature that an estimation of this nature is obtained from Ribonucleic Acid Sequencing (RNA-seq) data and only from reference human samples across a variety of tissues that overall capture the normal healthy state gene expression observed in humans. The study of normal states is important as it provides a baseline for later comparisons against disease or experimental conditions [30]. The data used derives from RNA-seq, a technology that provides the best quality for gene expression profiling to this day [31]. This sets this work in a very reduced group of large-scale coexpression projects that have been built with RNA-seq data. However, none of these previous projects provide comprehensive visualizations for such data, obligating researchers that wish to check simple analyses to process the data themselves.

Large-scale coexpression projects have also typically used the PCC [22] as the main tool for their calculations [15–17]. This well known statistical tool can detect biologically relevant relationships between genes, but it is limited to reporting only linear correlations. In this project, another well known statistical tool is applied for the first time to the context of gene coexpression: the Chi-Square Test of Independence [32]. With this tool, pairs of genes are investigated in a different way that provides insight into whether if the expression of one gene is dependant on the expression of the other. The use of the Chi-square statistic is inspired by the work of Reshef and collaborators where a generalization for multi-level associations is introduced [23]. As it will be discussed later (1.4.2), the computation time required to apply the exact method described in the aforementioned work would be very costly, so the Chi-square statistic is proposed as a faster alternative with the same intuitions.

The use of the Chi-square test is implemented in coexpression analysis by converting the expression of genes to discrete categories and looking at the distributions of observed and expected at random counts of these categories. This approach does not assume a linear correlation between continuous gene expression values to hint at a possible biological relationship between genes and also provides a more intuitive way to interpret coexpression relationships. This work considers both the Chi-square statistic and the PCC for a vast collection of unique gene pairs in addition to the Spearman Rank Correlation Coefficient (SRCC) [33], another very popular correlation metric outside the domain of coexpression that has not been hosted in any large-scale human coexpression project.

Another improvement that this project seeks over other already existent resources stems from the possibility of making gene coexpression results more interpretable than ever before. By using the Chi-square test, it is possible to identify gene expression categories that are responsible for an interesting Chi-square statistic when the latter is observed and even test their significance using a simple post hoc test [34, 35]. It is also possible to identify exactly which samples contribute the most to the observed Chi-square statistic and if they exhibit specific characteristics (e.g. Tissue type).

In this work, the tissue of origin of the samples is of particular interest because the identification of genes that exhibit tissue-specific expression patterns has important implications in the understanding of certain diseases and potential novel therapeutic strategies to treat them [36]. The results of this work will make it possible to go one level deeper into the study of tissue-specific genes by considering tissue-specific coexpression instead of just single gene tissue-specific expression. Sample contribution scores for microarray data have been implemented in other databases [37], but this project will be the first resource that aids the investigation of the role that different normal human tissues play in particular coexpressions calculated from RNA-seq data.

1.2.2 Hypothesis and research questions

The hypothesis of this work is that the Chi-square statistic can be used to produce a robust estimation of normal human system-level coexpression that improves the interpretability achieved by the current coexpression metrics used in literature. Some additional research questions that are related to this hypothesis are listed here:

- How does the Chi-Square statistic compare to the PCC or the SRCC in terms of calculated coexpressions?
- What is an effective way of creating a dataset that is representative of human gene expression at the system-level to serve as an input for coexpression analysis?
- How can the contribution of each human tissue or expression state in the system-level coexpression be estimated?

1.2.3 Objectives

The main objective of this project is to build a comprehensive high-quality human gene coexpression resource focusing on normal human samples profiled with RNA-seq. To achieve this, the following operative objectives must be met:

1. Obtain high-quality RNA-seq gene expression data from a variety of normal human tissues. See Chapter 2 Section 2.1 and Chapter 3 Section 3.1
2. Create a system-level dataset in which a variety of normal human gene expression states are represented. See Chapter 3 Section 3.2
3. Compare different strategies for coexpression calculation that utilize the system-level expression data as input. See Chapter 3 Section 3.3.
4. Compute the coexpression data for all possible unique pairs of genes in the input expression data using the chi-square, PCC, and SRCC coexpression metrics. See Chapter 4 Section 4.1

5. Validate the calculations using specific examples of coexpressed genes, comparisons between metrics, permutation-based tests, literature searches, and comparisons of co-expression across different types of samples. See Chapter 4 Sections 4.2, 4.4, 4.5 and 4.6

6. Deposit the results in a web-accessible database that incorporates visualization and analysis tools to add interpretability and complement the raw data which will be at the disposal of the scientific community. Chapter 4 Section 4.7

1.2.4 Methodology

Figure 1.3 summarizes the general methodology of this project. The process starts with data collection by fetching the gene expression data and associated metadata from the Genotype-Tissue Expression (GTEx) [38] project corresponding to operative objective number 1. The heterogeneity in this dataset is explored and characterized in step 2 of the methodology (second operative objective) using data mining tools to find different gene expression states that are represented in GTEx.

Different strategies for system-level coexpression calculation are then tested in step 3 of the methodology to add robustness to the final database calculation and fulfill operative objective 3. The coexpression data is then computed using both the Chi-square statistic and classic correlation coefficients for all gene pairs as established in methodology step 4 and operational objective 4. In step 5 of the methodology, validation analyses, and examples are presented according to operative objective 5. The final data results obtained are deposited in a publicly available web resource according to operational objective 6 in the last part of the methodology.

All analyses will be carried out using the R programming language for statistical computing [39]. The visualizations presented throughout this work are constructed with base R functions and the *ggplot* package [40]. Other used R packages are cited throughout the document.

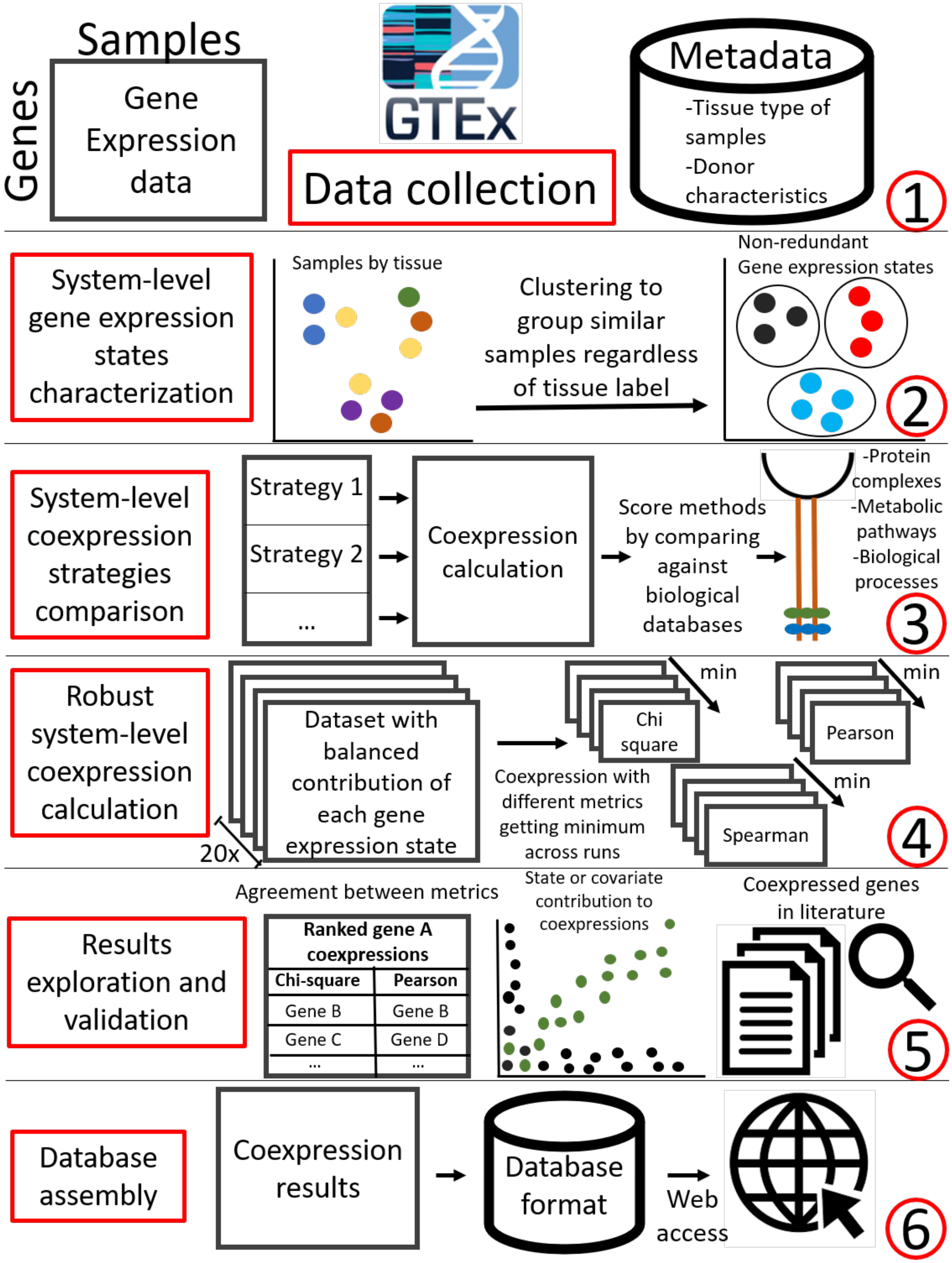

Samples
Genes
Gene Expression data
GTEx
Data collection
Metadata
-Tissue type of samples
-Donor characteristics
1
System-level gene expression states characterization
Samples by tissue
Clustering to group similar samples regardless of tissue label
Non-redundant Gene expression states
2
System-level coexpression strategies comparison
Strategy 1
Strategy 2
...
Coexpression calculation
Score methods by comparing against biological databases
-Protein complexes
-Metabolic pathways
-Biological processes
3
Robust system-level coexpression calculation
Dataset with balanced contribution of each gene expression state
20x
Chi square
Pearson
Spearman
min
min
min
Coexpression with different metrics getting minimum across runs
4
Results exploration and validation
Agreement between metrics
Ranked gene A coexpressions
Chi-square
Pearson
Gene B
Gene B
Gene C
Gene D
...
...
State or covariate contribution to coexpressions
Coexpressed genes in literature
5
Database assembly
Coexpression results
Database format
Web access
6

Figure 1.3: Graphic summary of project methodology.

1.3 Theoretical background

In this section, some important concepts for understanding and carrying out this work are discussed.

1.3.1 Gene expression

1.3.1.1 Concept

The concept of gene expression originates from the interaction between the three major informational macromolecules in molecular biology: DNA, RNA, and proteins as discussed in the context of the Central Dogma of Molecular Biology [41]. DNA contains sequences of the nitrogenous bases adenine, cytosine, thymine, and guanine (nucleotides) whose particular length and order codify specific genes. This code provides a structure for a process known as transcription, which involves the creation of RNA transcripts from the DNA mold of a gene.

Transcription happens when cells in living organisms need the function provided by the protein encoded in a particular gene or set of genes to carry out some process (e.g. Basic metabolism, response to environmental stimuli, etc.). Each gene in the DNA can be used to produce an analogous RNA, or very frequently more than one since a single gene mold may lead to the production of different transcript variants. RNA is a temporal molecule with many interesting functions, but by far the most important and well-characterized one is to serve as an information intermediary between DNA and proteins. The translation of RNA to proteins is carried out by structures known as the ribosomes.

Proteins are in charge of directly carrying out the biological functions that living cells need to survive, but it is only until a gene is transcribed from DNA into RNA that proteins can be generated. A key moment in this process is therefore when a gene's activation is actually needed by the cell or organism. In this situation, the gene must be expressed from its DNA mold into RNA molecules. Gene expression in the context of bioinformatics and computational biology refers to a quantitative measurement of the transcription of genes. Note that in the intuition given here, protein-coding genes were used to illustrate the concept and importance of gene expression, but there are about 84,485 protein-coding transcripts which only represent 36.79% of the total known transcripts (229,580) in the human transcriptome (the set of all human transcripts) [42]. This means that gene expression can also be quantified for transcripts with functions other than being messengers for protein translation, or very frequently with unknown functions.

1.3.1.2 Microarray Chips

The first technology that allowed for gene expression quantification and that became available for many researchers was the microarray [21]. It consists of a chip that contains attached DNA strands (called probes) corresponding to known gene sequences. The idea is to measure how strongly the RNA extracted from samples of interest binds to these reference probes in the chip. The process begins by extracting RNA from a sample and converting it to complementary cDNA (i.e. Reverse transcribe the RNA) to improve its stability because RNA is a temporal molecule and it can be easily degraded on accident if not converted to a more stable molecule. cDNAs effectively correspond to the molds from which the studied RNA can be produced. If a particular cDNA is present in the collection of cDNAs that were obtained from biological samples (i.e. the RNA from which it arises was expressed), it is expected that this cDNA couples with its complementary strand in the chip (a reaction known as hybridization) proportionally to its abundance in the sample. Before the hybridization reaction, cDNA is labeled with a fluorescent dye and the final expression for each probe is reported as an intensity of the fluorescent signal detected by a specialized automated analyzer.

1.3.1.3 Ribonucleic Acid Sequencing

The next innovation in the field of gene expression profiling came from massive DNA sequencing adapted to RNA, a technique known as RNA-seq [43]. This technique has many variations, but a popular method that is broadly used consists of obtaining cDNA from the RNA that is extracted from the samples of interest similar to what is done for microarrays. cDNA is fragmented to specific lengths and additional nucleotide sequences known as adapters are added at both ends of each cDNA molecule. This is done so that cDNAs can be recognized by the sequencing machine. Once inside the machine, these cDNAs are amplified producing multiple copies of each molecule. Afterward and for the sequencing perse, each nucleotide of each molecule is sequentially exposed to free nucleotides. As expected, only the nucleotide that compliments the one observed in the target cDNA can bind to the molecule. Every time a new nucleotide binds to the target cDNA, the machine can recognize the identity of the binding nucleotide via a color-coding system, so it can keep track of the sequence of nucleotides that are being used during this process.

The operation of sequentially exposing each nucleotide of a cDNA molecule to free nucleotides can be carried out in parallel for many cDNA molecules. Once the process finishes, the machine outputs a file containing strings of text depicting the order and identity of the nucleotides used to compliment each of the input cDNA molecules. These strings of text are known as reads and typically there will be several millions of them per sample. These reads can be mapped to a reference transcriptome, which is a file that has a curated transcript sequence-to-gene relation. This is done to know which reads arise from the expression of which genes. Once this last step is finished, the measurement of expression simply corresponds to the count of reads that match with a particular transcript in the reference transcriptome.

RNA-seq and microarrays produce correlated results when measuring gene expression. However, RNA-seq has been shown to provide higher resolution for low expression transcripts in addition to providing better coverage of the transcriptome [44,45]. Microarrays are limited to quantifying only the expression of the gene probes that are built into the chip and whose exact sequence has to be known beforehand. Specifically for gene coexpression, the use of RNA-seq has also been shown to provide substantial advantages over microarrays [9]. The use of RNA-seq is not straightforward though. There are quite a few sources of bias to be accounted for when quantifying gene expression via RNA-seq, which will be described next.

1.3.1.4 Read normalization and Transcripts Per Million

As mentioned, one of the steps of the RNA-seq protocol involves the fragmentation of cDNAs into smaller chunks to meet the sequencer equipment requirements. The latter implies that longer genes transcribing into longer RNA molecules will be represented by more numerous fragmented cDNA molecules in the sequencing experiment when compared to smaller genes [46]. This causes the final read counts of transcripts to depend not only on expression level, which is the real target of quantification, but also on the length of the gene. This is an example of one of the biases that can occur during RNA-seq. To account for these kinds of bias, several read normalization methods have been proposed. One of these methods applies a transformation to read counts known as Transcripts Per Million (TPM) which has been shown to be more consistent with the theoretical concepts behind read normalization than other transformations and to be robust to gene length [47]. It is also more useful for capturing gene expression variation among transcripts that is attributable to biological sources as opposed to noisy variations when compared to other normalization methods [48]. TPM read normalization is defined in a two-step process as:

1. $$\vec{r}^{\,L} = \frac{\vec{r}}{\vec{L}} \tag{1.1}$$

   Where $\vec{r} \in \mathbb{Z}^n_{\geq 0}$ is a non-negative integers read counts vector of n genes quantified for 1 sample. Each element of this vector must be scaled (element-wise) by the corresponding lengths of the genes quantified contained in vector $\vec{L}$ (the number of nucleotides in their biological sequences) to obtain the scaled vector $\vec{r}^{\,L}$.

2. $$\vec{r}^{\,TPM} = \frac{\vec{r}^{\,L}}{(\sum \vec{r}^{\,L})/10^3} \tag{1.2}$$

   Where $\vec{r}^{\,TPM}$ is the final TPM normalized vector of the sample in question.

Repeating the process for all sample vectors in a dataset will yield a TPM normalized expression matrix. TPM read normalization is performed to counter the gene length and sequencing depth bias existent in RNA-seq protocols. However, there are still other considerations when working with gene expression data such as the fact that the distribution of TPMs is very skewed. This makes additional data transformations useful for downstream analysis [49]. Other frequent problems are the so-called batch effects, which consist of technical variations in RNA-seq experiments that introduce unwanted variance to the measurements. An example is when some samples of an experiment are prepared for sequencing by different laboratory personnel. To assess these and other kinds of batch effects, another level of normalization techniques focused on the samples of the expression matrix may be employed before moving on to analyze the gene expression data [50]. A powerful and widely used method belonging to this group is Quantile Normalization (QN) (2.3.1) [51].

1.3.1.5 Availability of gene expression data

Over the years, both microarray and RNA-seq technologies have become increasingly available and accessible to researchers around the world. The latter has led to the generation of a huge amount of gene expression data which is used in publications and is shared with the scientific community through repositories such as the Gene Expression Omnibus (GEO) [52], a resource which nowadays hosts results for over 100,000 expression profiling experiments.

Despite this, the availability of gene expression data specifically from reference samples (i.e. Normal healthy samples) is limited in terms of quantity when compared to the numerous experiments obtaining data from disease states or samples under specific artificial stimuli [53]. Attempting to maximize the number of reference samples by compiling and combining data from different expression profiling projects sounds attractive to obtain sufficient samples for some downstream analysis, but differences between equipment and methods across experiments can be troublesome [53].

Thankfully, there exist some human expression profiling projects which focus on producing gene expression data for a large number of reference samples under standardized conditions. The biggest project to this day doing this is the GTEx project, which has been profiling samples from post-mortem donors with RNA-seq since 2013 [38]. This is the source of the gene expression data used in this work which will be described thoroughly in Chapter 2 (2.1).

1.3.2 Details on gene coexpression

Gene coexpression was introduced in this chapter as a way to study the relative trends of expression that exist between gene pairs and to estimate coordinated expression behaviors between groups of genes. Here the process to carry out this analysis is described in more detail.

1.3.2.1 Gene coexpression as networks

The idea behind coexpression analysis is to compute a correlation or association metric (most commonly the PCC in literature) between all possible unique pairs of genes. So for a $M \in \mathbb{R}^{nxm}$ gene expression matrix with n genes and m samples, the total number of coexpression metrics to compute $ncoexp$ corresponds to all possible combinations without repetition of 2 elements that can be obtained from a n elements set (see Figure 1.4 for a small example). This is a case of the general combinations formula which can be written as:

$$ncoexp = \frac{n \cdot (n-1)}{2} \tag{1.3}$$

This number of metrics corresponds exactly to the number of values in the lower (or upper) triangle without the diagonal of a $X \in \mathbb{R}^{nxn}$ gene by gene square matrix which many computational tools for correlation calculation return on a matrix input. Figure 1.5 shows how the number of coexpressions to calculate increases with the number of genes considered in a gene expression dataset.

Coexpression databases are focused precisely on providing researchers with the results of large-scale coexpression calculations comprising massive coexpression matrices. In these kinds of projects, a large number of genes and samples are considered and the computation of the coexpression matrix may be done several times for robustness [15]. Databases provide the scientific community with an already calculated knowledgebase of coexpression which otherwise would be logistically and computationally difficult to obtain from scratch.

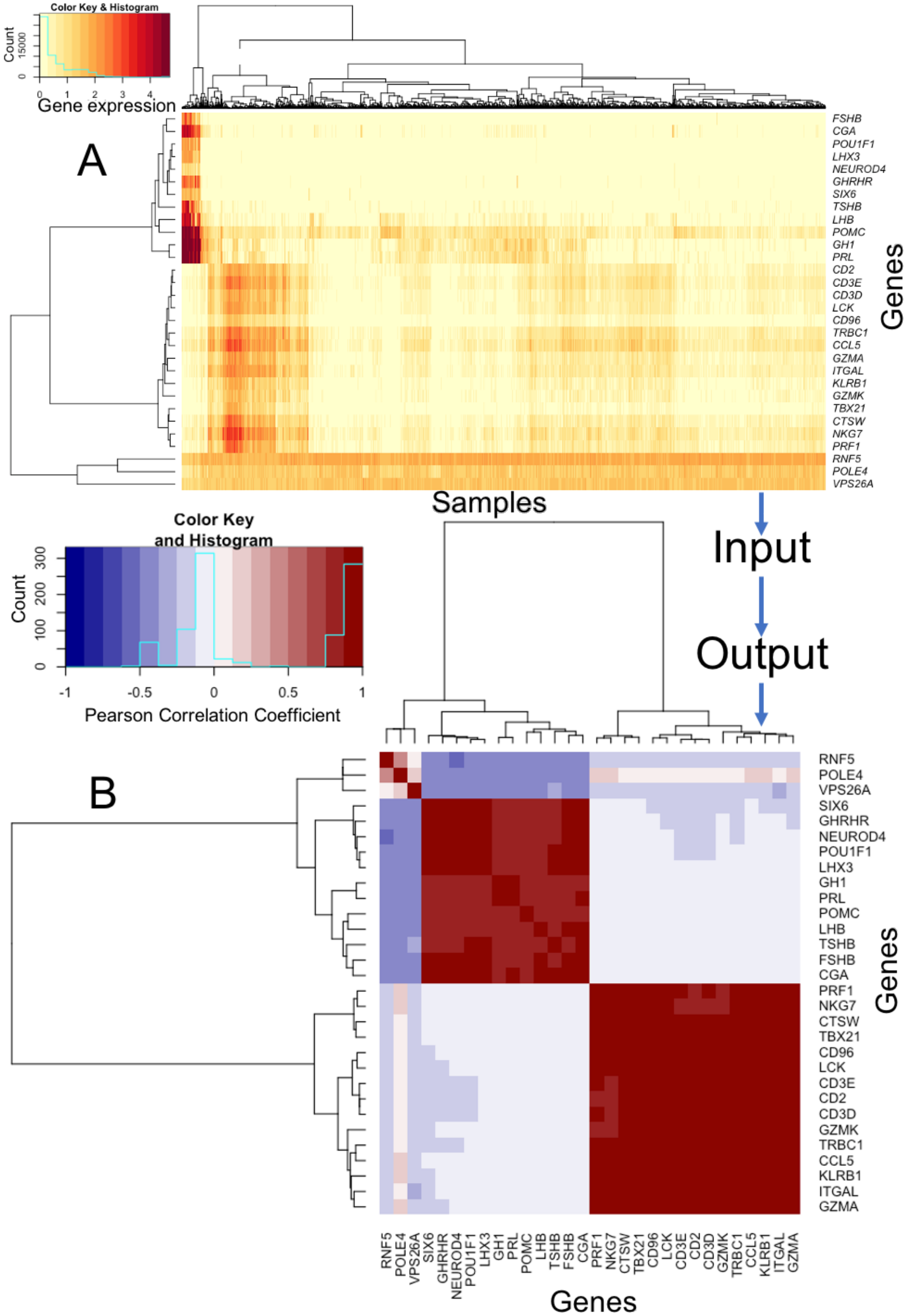


Figure 1.4: Example of input and output in a coexpression analysis. A: expression matrix of 30 genes (rows) by 2000 samples (columns). The matrix is ordered using Hierarchical Clustering (HC) with correlation distance (2.4.3.1). Groups of genes that have similar expression profiles across samples can already be identified here. B: resulting coexpression matrix which is symmetric with genes in both columns and rows. Color scales are shown in each panel.

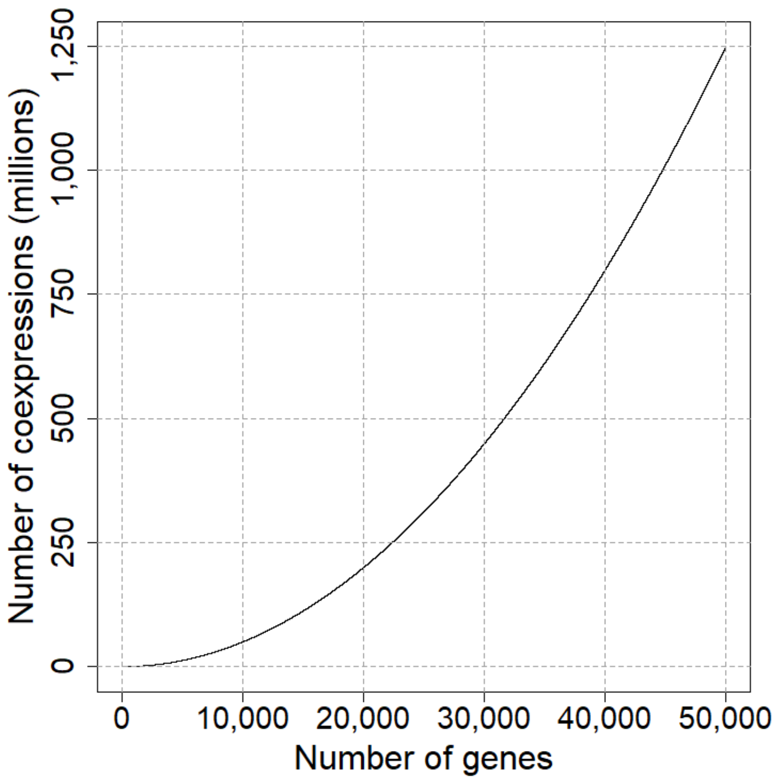


Figure 1.5: Quadratic order growth in number of possible coexpressions relative to the number of input genes.

Coexpression matrices may be downloaded from a database or obtained manually from processing a small gene expression experiment considering a few genes and some control/treatment samples. Regardless of the origin, a usual generic coexpression workflow is as follows:

1. *Establish criteria for considering a given pair of genes as truly coexpressed*: many different approaches are possible in this step varying from the very simple hard threshold method, which consists of keeping only measurements whose absolute value is greater than a certain value (see Figure 1.6 for a small example), to more sophisticated statistical methods [54]

2. *Interpret true coexpression values as an adjacency matrix*: a network may already be built from this interpretation where nodes are genes and edges exist between a pair of nodes if the coexpression between the genes was considered true in the last step. The edges of the network may be weighted according to the intensity of the coexpression measure as shown in Figure 1.6

3. *Find clusters of densely connected genes (called modules in coexpression terminology) in the coexpression graph*: these modules effectively represent groups of genes that had an overall high coexpression measurement between them by the metric used in coexpression analysis. Genes belonging to a module, therefore, have a higher chance of being functionally related to one another than genes chosen at random. For this step, techniques in graph theory become invaluable tools to characterize these coexpressed

gene modules. Many tools exist in the literature that facilitate this analysis including the popular Weighted Gene Coexpression Analysis (WGCNA) [55]. This tool finds modules of highly coexpressed genes based on hub genes which are nodes in the graph that have many edges connected to them. WGCNA performs additional transforms to the coexpression matrix which require the initial coexpressions to be in the range $[-1, 1]$ (as it is the case for any PCC calculation), but many other techniques are available to use for module detection that will work with any coexpression metric [56]

4. *Analyze coexpression modules for biological significance*: use gene set enrichment analysis or gene overrepresentation analysis to test if the genes in the modules have a statistically significant overlap with well-characterized sets of genes with known biological functions [57, 58]. From here it is possible to generate hypotheses, for example, that maybe uncharacterized genes in the modules are associated with certain functions or that maybe a known gene is associated with additional and previously undescribed processes for that gene

The extent of the gene coexpression network can also be variable. For instance, one may choose to analyze only the top coexpressed genes with a gene of interest and the significant connections between them to generate a local network relative to the chosen gene (i.e. Only using parts of the coexpression matrix). It could be that the objective is to analyze the data globally by constructing a network with all significant coexpressions between all possible pairs of genes (i.e. Using the complete coexpression matrix). This flexibility of being able to control the depth of the gene coexpression network analysis is a nice property of gene coexpression data.

1.3.2.2 Large-scale coexpression

The term ”large-scale coexpression” does not have a formal definition in the scientific literature pertaining to coexpression nor it considers a specific number of genes/samples for a project to be classified as large-scale. The term is used here to refer to projects whose aim is to compute coexpression data for tens of thousands of genes across thousands of samples with a database-driven focus. This implies these projects deal with a total number of gene pairs in the scale of the hundreds of millions (like this master thesis work) or even in the scale of billions [16]. Large-scale coexpression projects distinguish themselves from other works whose main focus is to study coexpression for specific sets of genes that are perhaps relevant for particular diseases or biological functions. A typical approach in these smaller projects is to find the differentially expressed genes between a control and a phenotype of interest. Coexpression analysis is then carried out only on these genes which are usually in the scale of thousands and the number of samples used also tends to be more limited [59].

Carrying out coexpression analysis when working with a few thousand genes and a few hundred samples is usually not a problem in a standard computer nowadays. Tools such as the previously described WGCNA are very user friendly for this purpose as they consist of full pipelines that take an expression matrix as input and output clustered modules of coexpressed genes. Modules may be directly sent to gene set enrichment or overrepresentation analysis. However, many conceptual and computational complications arise when the idea is to obtain coexpression analysis results from a large number of genes and samples.

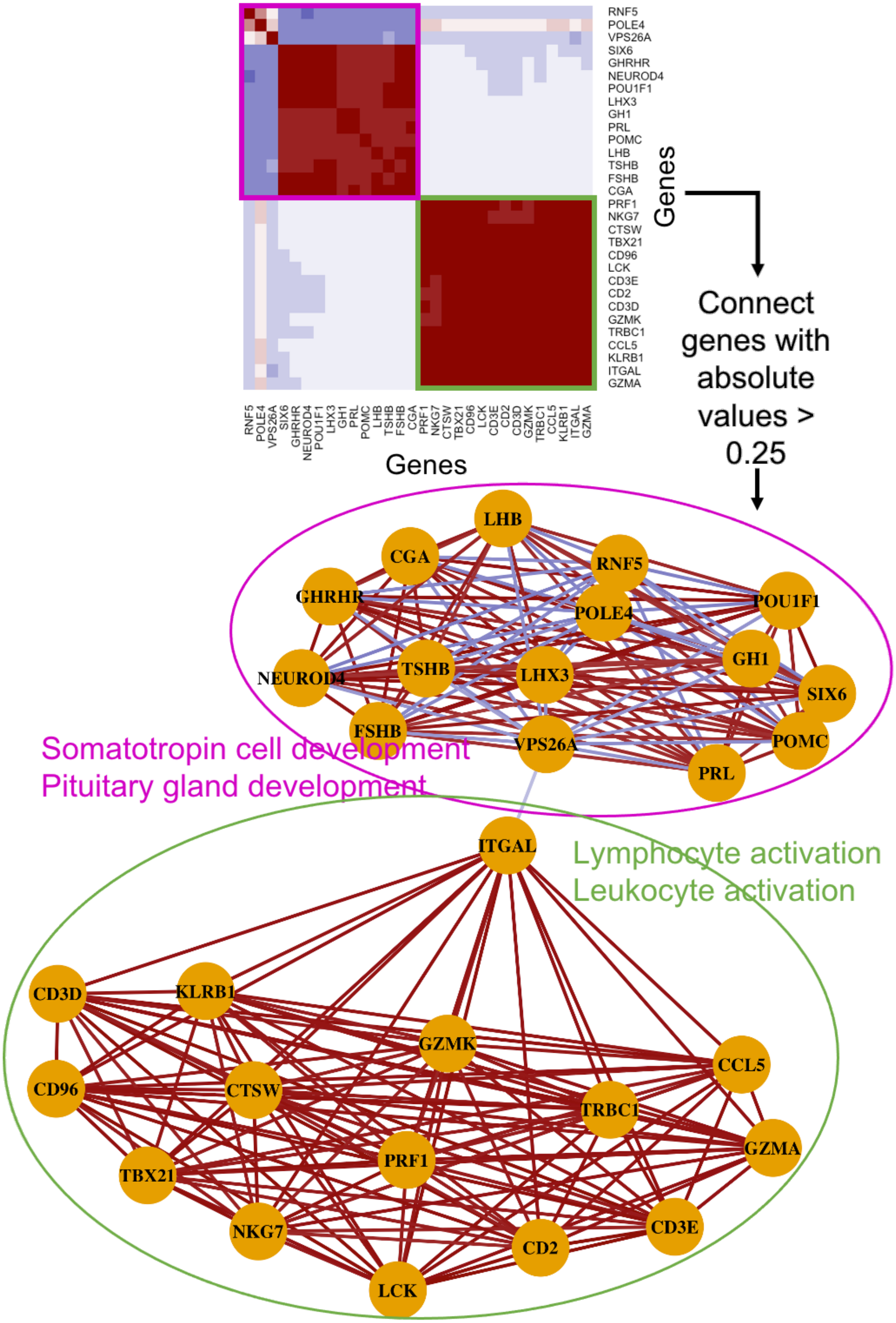


Figure 1.6: Example of a gene coexpression network obtained from applying a hard threshold to a gene coexpression matrix (continuation of Figure 1.4). Groups of nodes which are densely connected are frequently involved in interesting biological functions.

Large-scale coexpression projects frequently have as an objective to provide their robust estimations of coexpression data to other researchers in the scientific community. This is done via databases, which have a format based on gene lists. These lists display all the results for a single gene (i.e. All metrics for that gene when paired with all others) while sorting them in descending order according to the strength of the coexpression. The idea is to provide all possible gene pairs of a query gene with their corresponding coexpression evaluation on demand. This way users can check the associations with their genes of interest in an efficient manner. If they require several genes for their analyses, a bulk download of already calculated coexpression is also typically available.

1.3.2.3 The problem of sample type overrepresentation

When working with large-scale coexpression, input samples will likely be unbalanced regarding the gene expression state they represent (e.g. A tissue or experimental condition) because some types of samples are more easily obtained than others. This prevents the objective of obtaining a coexpression estimation that approximates the overall trends across all types of samples. This problem has also been characterized as a situation of high redundancy in the input samples. Consider a hypothetical dataset that has 100 samples total: 50% of the samples are from skin, muscle, fat, brain, and liver (10 each). The remaining 50 samples are all blood samples. If one naively proceeds to coexpression analysis using the full dataset as it is, the resulting coexpression patterns can be biased towards the patterns observed in blood.

In general, for the example given previously, it can be said that those blood samples are redundant among themselves (i.e. It is expected that they are very similar in terms of gene expression and hence contribute the same kind of information to the coexpression analysis). The same is expected for the other tissues, but the culprit for an overall high ”redundancy sum” in the dataset is blood as its higher sample counts contribute more to the overall redundancy. The latter is true if all the blood samples are indeed very similar among them. However, if there were two significantly distinct groups of blood samples, then members of one group would not be redundant when compared with members of the other. In pioneer studies using expression data from simpler organisms like the bacteria *Escherichia Coli*, it has been demonstrated that high sample redundancy in a dataset can be detrimental for the accurate building of coexpression networks [60].

Some strategies to tackle the problem of sample type overrepresentation are discussed in chapter 2. They include a literature approach, as well as ideas proposed in this work (2.6).

1.3.2.4 Gene coregulation

Gene coregulation is a term that is very frequently mentioned in coexpression analysis because it is of great interest to investigate this via coexpression data. Coregulation and coexpression are related but comprise different kinds of biological phenomena.

The concept of coregulated genes refers to a set of genes with shared underlying mechanisms that control their rates of expression in a coordinated manner such as Transcription Factors (TFs) [25]. TFs are genes that regulate the expression of others at a DNA level and that frequently are hub genes in coexpression networks [61]. If two genes that are coexpressed are known to be regulated by a common TF, then they are not only coexpressed but also potentially coregulated.

Note that a pair of genes may share a common TF, but what truly makes them coregulated is the fact that the TF influences their expression rates in a synchronized manner rather than separately for each one of them. Evidence of the latter (although not definitive proof) would be that the analyzed pair of genes are coexpressed among themselves, otherwise, their expression levels would behave randomly relative to one another despite having a common TF. Note as well that one may come across the statistical finding that a pair of genes are coexpressed without necessarily having to attribute that to a common TF. Therefore coregulation entails coexpression, but not the other way around. Investigating coregulation requires the triangulation of coexpressions with TFs or other regulatory mechanisms. Some large-scale coexpression projects have provided separate gene lists in their databases focusing on the coexpression measures between query genes and TFs to facilitate the study of this phenomena [16].

1.4 Related work

1.4.1 Large-scale human coexpression projects

In the literature, there have quite a few large-scale coexpression projects working with human gene expression data. Three of them stand out for providing robust estimations of human coexpression in general by considering samples of a variety of types and working with either RNA-seq data or only normal tissue data. These characteristics are pivotal in this master thesis work and hence of great interest during literature revision.

1.4.1.1 COXPRESSdb

COXPRESSdb is a large-scale coexpression project with its own publicly available database which has seen 4 major updates since its original release in 2008 [15,62–65]. This database is focused on coexpression from an evolutionary point of view as it currently hosts results for 12 organisms including *Homo Sapiens*. Many of their result validation experiments are based on searching for conserved coexpression edges across species (a concept known as supportability) which makes a given pair of genes more likely to be truly coexpressed. In the beginning, the database was built solely off microarray data, but nowadays it also hosts human coexpression calculated from RNA-seq experiments accounting for 17,067 genes. This database is the most complete in the literature in terms of data delivered to the user. Additional features are provided apart from gene lists of coexpressed gene pairs and general gene information. Notable mentions of these extra features are:

- It is possible to visualize the scatter plot of every coexpression in the gene lists (only for microarray data)
- A list of Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways in which the queried gene is present is given along with the gene list [66]

- A score of the contribution to the coexpression by the top contributing samples is given along with the origin (name of source expression experiment) of such samples (only available for microarray data)
- For each gene in a gene's coexpressed genes list, it is possible to see in which Gene Ontology (GO) biological classes and KEGG pathways the gene belongs [67]
- For each gene in a gene's coexpressed genes list, one can ask for a local coexpression network considering the gene in question and all its top coexpressed neighbors with their respective involvement in KEGG pathways noted
- It is possible to visualize the average expression of a pair of coexpressed genes across the different tissues (samples include normal, disease, and experimentally treated tissues and cells) considered in COXPRESSdb

It is worth noting that COXPRESSdb is not dedicated exclusively to capturing coexpression in normal healthy tissues or on samples that have not been put under artificial stimuli. As noted in their expression pages [68], the microarray data (the data for which scatter plots and other visuals are available) include cancer (e.g. Breast, flash-frozen, microdissected ductal tumor cells) and experimentally treated (e.g Skin taken 96 hours after exposure to 5% nickel sulfate) samples in addition to normal samples. COXPRESSdb uses the PCC as the base for computing coexpression. However, they propose the transformation of the PCC to the Mutual Rank (MR) measure. For a pair of coexpressed genes A and B, the MR represents the geometric mean arising from the rank of B in A's gene list and the rank of A in B's gene list [69]. The MR is the only metric available when querying a gene and obtaining its gene list in COXPRESSdb.

COXPRESSdb has traditionally addressed the problem of sample type overrepresentation with the use of a Weighted PCC (1.3.2, 2.6). This technique was used in the first 3 major versions of the project, but as of the latest version of COXPRESSdb, the authors have moved from the Weighted PCC strategy to a Principal Component Analysis (PCA) based one for the computation of their most recent coexpression data [15]. PCA is a very popular technique in data science which basically consists of producing linear combinations of variables in a dataset [70]. The combinations are called Principal Components (PCs) and they maximize the variance observed in the dataset's objects across variables (2.4.1).

The most recent process followed in COXPRESSdb consisted of applying PCA to a 17,067 (genes) by 10,485 (samples) matrix of human RNA-seq expression data obtained through Sequence Read Archive queries [71]. 10,485 PCs were obtained representing linear combinations of the sample vectors. Only the first 1000 PCs accounting for a large proportion of the total variance were retained and 1000 submatrices were assembled containing 100 randomly selected PCs from the original big matrix of 1000 PCs. Coexpression analysis is carried out on each of the submatrices. The final value for the coexpression between gene x and gene y is the minimum observed for these genes across the 100 realizations of coexpression analysis.

The idea behind the described strategy is to summarize different gene expression states that exist in the input data by considering the variance between them as the new input datum. More intuitively, this corresponds to performing coexpression analysis ”on the variance” across samples in the initial dataset. One problem that can arise from this strategy is that it becomes more difficult to pinpoint what kinds of samples are contributing to the observed coexpression. This observation is derived from what is discussed by some literature on PCA analysis for expression data [72]. Using this technique could therefore make it difficult to later tell exactly which tissue or condition is contributing to a particular coexpression. This is not one of the main interests in COXPRESSdb, so this technique is more attractive for them.

1.4.1.2 Genefriends

Genefriends was published in 2012 and saw its most recent major update in 2014 [16, 27]. The latest version of this database is inspired by COXPRESSdb as it adopted the Weighted PCC strategy and MR metric for the computation of the coexpression data. For Genefriends, human expression RNA-seq data was fetched from SRA resulting in measurements for 44,248 genes across 4,133 samples after pre-processing. Similarly to COXPRESSdb, Genefriends includes samples challenged with some stimuli or under disease/developmental conditions (e.g. Embryonic cells). The focus of Genefriends is to be a GBA tool to identify potential novel genes involved in complex diseases [9]. Genefriends provides fewer visualization tools to the user compared to COXPRESSdb, but it features some things that COXPRESSdb does not: they include the raw PCC in the coexpressed gene lists in addition to the MR, they implement a separate coexpressed TFs gene list for each queried gene and they allow to carry out enrichment analysis of coexpressed genes via the Database for Annotation, Visualization and Integrated Discovery (DAVID) [73].

1.4.1.3 Human Gene Correlation Analysis

Human Gene Correlation Analysis (HGCA) is another large-scale expression project with its own public database. This one is based completely on microarray data. The database was released in 2012 without any major updates justifying a new publication since then [17]. Nevertheless, it is an important pioneer work because it is the only database (to the best knowledge gained while reviewing literature related to this work) that is specifically built using only normal healthy samples from a diversity of human tissues. Such samples (4,732) were manually selected from initial queries yielding 62,000 samples in GEO. The focus of this database is similar to that of Genefriends. It is based on the PCC computation between 54,000 microarray probes. No specific strategy for countering sample type overrepresentation and redundancy was applied in this work.

1.4.1.4 Comparative summary of principal large-scale coexpression projects

Table 1.1 summarizes the main features studied during the literature revision performed for this work regarding existent large-scale projects focused on human coexpression. Features are defined as follows:

- *Available raw data source*: type of gene expression profiling technology used to obtain the input data

- *Available explorable data source*: the gene expression profiling technology for which it is possible to visualize expression distributions and/or gene pair scatter plots inside the project's database
- *Available analyzable data source*: the gene expression profiling technology from which the coexpression data derives. These data can be used to construct coexpression networks and/or perform enrichment/overrepresentation analysis inside the project's database.
- *Base coexpression metric*: statistical measure used for calculating coexpressions
- *Sample characteristics*: spectrum of sample conditions used to estimate coexpression

Table 1.1: Comparison of principal large-scale coexpression projects in literature

Feature	COXPRESSdb	Genefriends	HGCA
Available raw data source	Microarray and RNA-seq	RNA-seq	Microarray
Available explorable data source	Microarray	None	None
Available analyzable data source	Microarray	RNA-seq	Microarray
Base coexpression metric	PCC	PCC	PCC
Sample characteristics	Normal + disease + experimental conditions + developmental stages	Normal + disease + experimental conditions + developmental stages	Normal

1.4.1.5 Other large-scale coexpression projects

Quite a few more interesting large-scale coexpression projects exist that have focused on producing results for *Homo sapiens*, but according to the literature review performed, all of these are based solely on microarray data like COEXPEDIA, STARNET or GENEVESTIGATOR [18–20]. Others are focused on a single type of samples such as MIrExpress [74], a project that can be considered as large-scale coexpression since it considers 20,283 human genes and 6,909 samples. However, all of the samples correspond to immune cells as the project is solely dedicated to researching coexpressions in this particular human system.

1.4.2 Discovering associations in large datasests

While revising the concepts behind coexpression analysis, it has been established that the idea of the technique goes about finding trends or relationships between the expression of pairs of genes. Outside the context of gene expression, the essence of this problem is encountered in many domains that deal with large datasets. The PCC is also a popular measure in these domains to find relationships of dependence between two random variables. It is no surprise that this tool has been widely applied in coexpression analysis.

Despite the popularity of the PCC, research has been done to investigate other options for finding associations in large datasets. One of these ideas, which was developed in the work of Reshef and collaborators [23], has inspired for the first time the use of the Chi-square statistic for coexpression as proposed in this master thesis. In the cited paper, a grid-based strategy using mutual information is presented. The idea is that for each pair of random variables, a matrix of possible grid bins dividing the scatter plot defined by the variables is constructed. The matrix is defined by varying both the number of cutoffs and values of the cutoffs indicating the bins. The best way to bin the scatter plot is determined by the configuration that gives the maximum information gain. This is reported as the Maximal Information Coefficient (MIC), the measure of association proposed in the discussed paper.

The work by Reshef and collaborators answers to the need for a general and equitable association measure [23]. General refers to the capability of the measure of identifying relationships driven by any kind of function and even relationships that cannot be well characterized by a single function such as superpositions of functions. Equitable means that the measure should not be more prone to favor the identification of associations when they are driven by a certain type of function at equal levels of noise. The PCC is an example of a measure that is equitable but not general as it fails to report interesting associations that are not linear. On the other hand, measures such as the SRCC and mutual information-based estimators that do not perform the steps of MIC tend to be more general, but favor the detection of certain functions over others even at equal levels of noise (e.g. Linear over sinusoid functions, etc.)

Although the use of MIC for gene coexpression would be interesting, in this work it was decided that the first approximation in that direction would be done with the Chi-square test of independence. This is done to avoid increased computation times required for a workflow using MIC. In such a scenario, one would have to explore different values to bin gene expression and different numbers of bins for each pair of genes. The Chi-square test is mathematically related to mutual information and also uses a grid-based strategy by binning the scatter plot of the random variables tested [75]. It is expected that this test can give a good proxy for tackling coexpression from the point of view discussed here. MIC has proved to be useful at identifying associations that other tools cannot in different sets of real data including genomic and microbiological data [23]. It will be interesting to see what the Chi-square proxy can do in coexpression analysis.

1.5 Summary of introductory concepts

To finalize the introduction to this master thesis work, here are summarized the key points that have been discussed in this chapter:

- Gene expression is a quantitative measurement of how active a gene is in a biological sample.

- The state-of-the-art technology for quantifying gene expression is RNA-seq. The output of RNA-seq consists of read counts for different transcripts in the profiled samples. These read counts need to be normalized to account for factors such as gene length and batch effects.

- Gene coexpression analysis looks at the expression of genes in pairs and tries to determine if there exists a coordinated behavior between the expression of the pair

- Genes that are coexpressed have a higher chance of being functionally related than genes picked at random. This allows for the opportunity to use coexpression to hypothesize about the function of poorly characterized genes and to prioritize genes for further research.

- Large-scale coexpression projects are frequently dedicated to building databases comprising a huge amount of coexpression data derived from tens of thousands of genes and thousands of samples. These databases are valuable because they allow researchers from all fields to access robust coexpression results for validation or secondary analyses.

- Current large-scale human coexpression projects hosting databases have not focused on generating data from normal samples profiled with RNA-seq at the same time. These projects have also only used the PCC metric as their sole mean to explore coexpression between genes.

- Other measures of association exist besides the classic PCC that could help to identify meaningful coexpressions between genes. Some of these measures consider the expression of genes in discrete bins and can detect non-linear associations and relationships between pairs of variables which are genes in the context of coexpression.

Chapter 2

Materials and methods

In this chapter, the gene expression data used in this work for the estimation of normal human gene coexpression proposed is revised in detail. All the computational techniques and statistical tools used to obtain the results presented in Chapters 3 and 4 are also reviewed.

2.1 The Genotype-Tissue Expression project data

The GTEx project is the source of all gene expression data used in this work. It comprises a high-quality collection of samples from post-mortem human donors whose expression has been profiled using RNA-seq. The first publication of the project was made in 2013 [38]. It is founded by the National Health Institute (NIH) of the United States. The initiative had in mind the study of expression Quantitative Trait Loci (eQTLs), which are variations in the DNA sequence that have an effect on gene expression. However, the data generated for this purpose provides invaluable opportunities to study human gene expression in the context of many other analyses. Over the years GTEx has been growing in terms of the number of samples they consider becoming the largest unified resource of normal human RNA-seq.

2.1.1 Donor exclusion criteria

The GTEx project has attempted to assemble a large resource of human gene expression while also seeking to obtain it from overall healthy individuals. Data in GTEx represents a good proxy to what is observed in normal human biology.

Exclusion criteria for donors of the GTEx project are summarized in Table 2.1. Donors from both genders and any ancestry group were included. While it is not possible to guarantee that all samples from all donors in GTEx represent an ideal normal and healthy state, this is the biggest expression profiling project which has searched for the collection of non-diseased samples across a variety of human tissues [76].

Table 2.1: GTEx donors exclusion criteria

Criteria	Parameters
Post-mortem interval	More than 24 hours
Age	Less than 21 or more than 70
History of viral infections	Human immunodeficiency virus or hepatitis
History of cancer	Metastatic cancer or chemotherapy/radiotherapy in the last 2 years before death
Blood transfusion	In the last 48 hours before death
Body Mass Index	Less than 18.5 or more than 35
Expert pathologist examination	Abnormal findings in the histological analysis

2.1.2 Tissues and profiled genes

In this master thesis, the version 8 release of GTEx is used. The ”Gene TPMs” expression matrix file was downloaded from `https://www.gtexportal.org/home/datasets` in March 2019. This file was confirmed as not changed or updated since then as of September 14th, 2020. The file contains TPM normalized RNA-seq gene expression read counts for a collapsed gene model based on the Encyclopedia Of DNA Elements (ENCODE) human genome release 26 (GRCh38.p10) (1.3.1.4) [42]. A single gene can be expressed through different transcript variants, so this collapsed gene model produces only one expression value per gene through merging and ambiguity resolving [77].

From the original expression data, 16,704 samples obtained from 981 distinct donors representing 52 normal human tissues were considered for downstream analysis (Figure 2.1). GTEx includes Brain Cortex and Brain Frontal Cortex tissue types but as documented in their webpage [78], these tissues are essentially the same (e.i. Replicates) except for the time of sample taking from the donors. Brain cortex was taken together with the other non-brain samples of the donors (e.i. Earlier) while Brain Frontal Cortex was sampled later at a specialized center. The same situation applies to the Brain Cerebellum and Brain Cerebellar Hemisphere tissues.

Note that in the distribution shown in Figure 2.1, it is evident that the sample tissue types are unbalanced. This creates an instance of the sample type overrepresentation problem (1.3.2.3). This is important because the coexpression estimation proposed in this work intends to cover in a uniform way the variety of human tissues provided in GTEx.

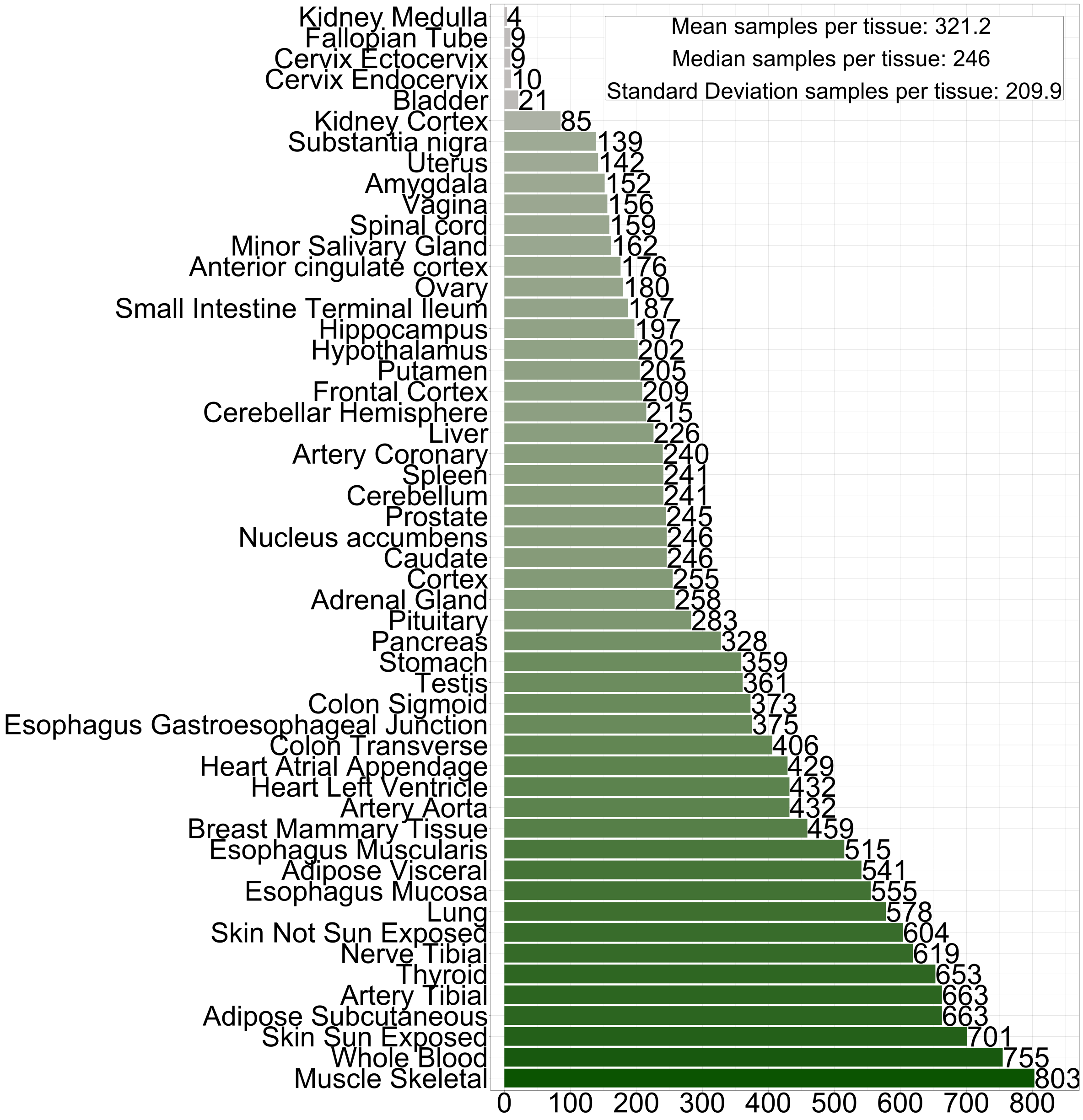


Figure 2.1: Tissue type distribution of GTEx data used in this work.

Originally, the expression matrix contains measurements for 52,600 genes across all samples yielding the distribution shown in Figure 2.2. However, as it is common when working with gene expression data, low expression genes must be identified and filtered out as their measurements may be indistinguishable from noise that affects downstream analyses [79]. It is frequent to refer to these low expressed genes as simply ”not expressed” if they have consistently low expression values across a set of samples. In the following section, a criterium to identify these genes is described.

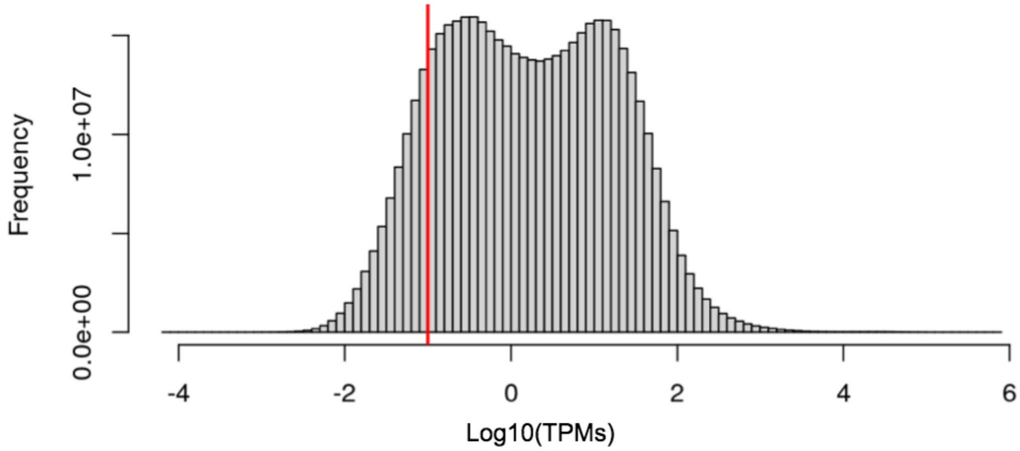


Figure 2.2: Distribution of logarithmically transformed TPMs (for visualization as TPMs in their raw scale are typically skewed) in the original GTEx expression matrix. Zero valued TPMs in raw scale are dropped as a result of the logarithmic transform. The red line indicates a value of 0.1 TPM in raw scale (-1 in loagrithmic scale), a frequently used value for considering genes as expressed in these data (2.1.3.)

2.1.3 Consideration of expressed genes

Due to its quality and importance, GTEx data has been used in many publications. There have been good results in different kinds of analyses considering a low expression criteria of 0.1 TPM. More specifically, a gene has been considered expressed for downstream analysis if it has values greater than the 0.1 threshold in at least 20% of the samples. This approach is used by GTEx authors themselves in the construction of an eQTLs database [76,77]. It is also used by several authors of recent works performing different analyses on GTEx expression data as depicted in Table 2.2. In this work, the criterium of greater than or equal to 0.1 TPM in 20% of the total samples is taken into account as an initial filter for considering genes as expressed similarly to the works cited. In this specific case, 20% of the total samples in the dataset represents 3,341 samples.

Table 2.2: List of recent works using the GTEx data and criteria for gene filtering $>$ 0.1 TPM in at least 20% of samples

Paper	Journal and Citation
The GTEx Consortium atlas of genetic regulatory effects across human tissues	Science [80].
eQTLMAPT: Fast and Accurate eQTL Mediation Analysis With Efficient Permutation Testing Approaches	Frontiers in Genetics [54].
Pseudogenes Provide Evolutionary Evidence for the Competitive Endogenous RNA Hypothesis	Molecular Biology and Evolution [81].
Hepatocyte gene expression and DNA methylation as ancestry-dependent mechanisms in African Americans	npj Genomic Medicine [82].
Transcriptome Analysis of the Human Tibial Nerve Identifies Sexually Dimorphic Expression of Genes Involved in Pain, Inflammation, and Neuro-Immunity	Frontiers in Molecular Neuroscience [83].
Molecular insights into genome-wide association studies of chronic kidney disease-defining traits	Nature Communications [84].

2.2 Identification of tissue-specific genes

Using exclusively the consideration of expressed genes described in the last section could come across some issues. Consider the situation in which a gene is not expressed globally in more than 20% of the samples but maybe when looking at individual tissues, the gene is clearly specifically expressed for some tissue. Figure 2.3 depicts some descriptive examples. In panel A, the Glyceraldehyde-3-Phosphate Dehydrogenase (GAPDH) gene is expressed in all samples of all tissues as all measurements are more than 0.1 TPM on the raw scale or -1 on the logarithmic scale. This gene has no problems meeting the expression criteria described in the last section. In panel B, the Uromodulin (UMOD) gene is not expressed in the majority of samples in GTEx as most TPMs fall below the red dotted line at 0.1 TPM raw scale or -1 logarithm. In fact, after doing the calculations, UMOD would not be considered as expressed by the criterium defined in the previous section. However, in Figure 2.3 it is clear that in some tissues UMOD is indeed expressed even at more than 10-fold the initially considered expression threshold (green dotted line at 1 TPM raw scale or 0 in logarithm). This is especially obvious in the Kidney Cortex tissue.

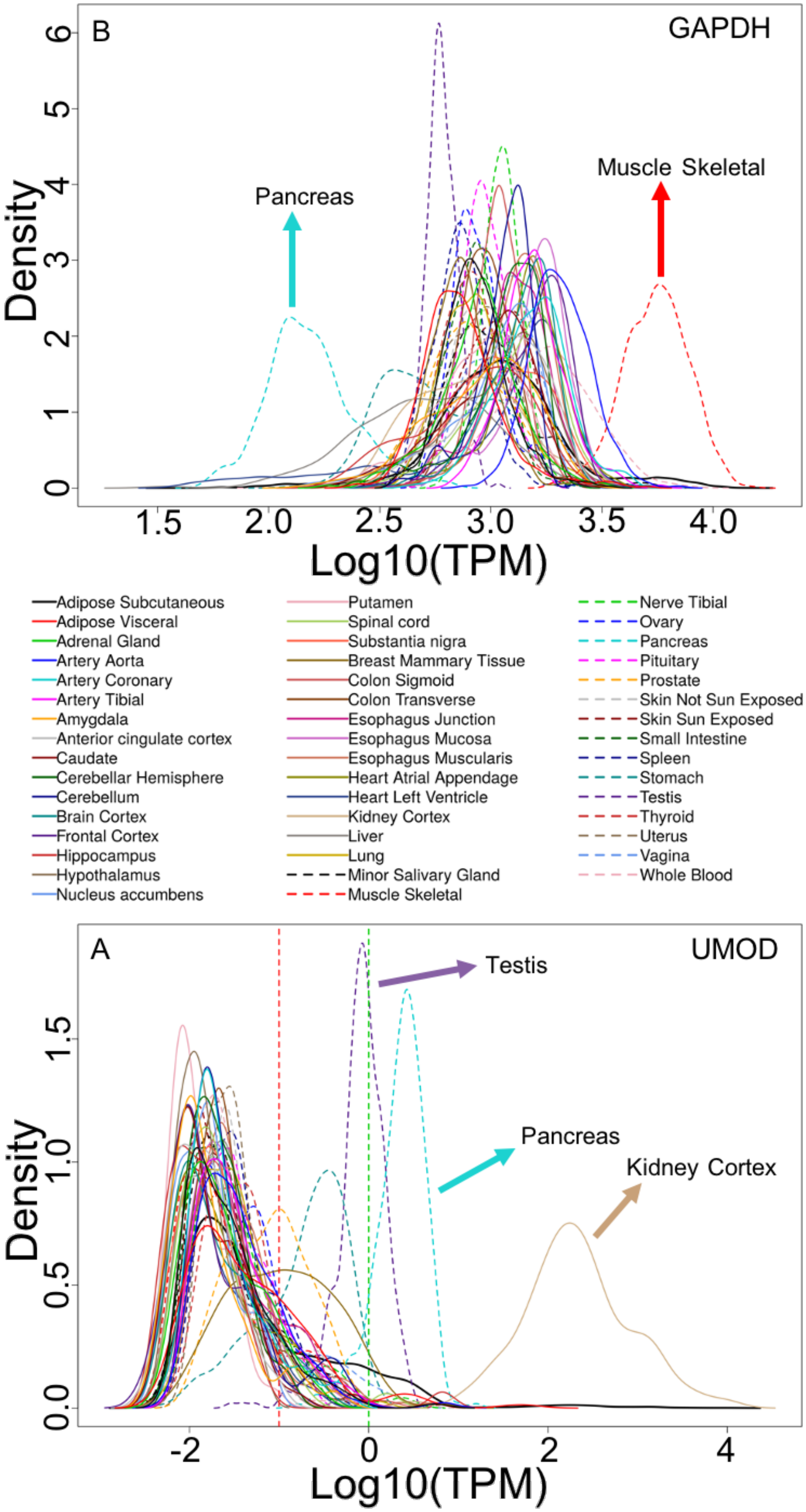


Figure 2.3: TPM distributions of two genes in different GTEx tissues. A: example of a gene expressed across all GTEx samples. B: example of a gene expressed only in a subset of samples.

To investigate and recover genes expressed only in particular tissues, in this work it is proposed to use two metrics which can be calculated for all genes. The first one is the Max Percentage of Expression (MPE) defined as:

$$MPE^g = max_{1 \leq i \leq T}(\vec{x}^g) \tag{2.1}$$

Where $\vec{x}^g$ is a vector of length T (the number of tissues present in the dataset) containing the percentage of samples in which gene g is expressed for each tissue. This metric requires setting a TPM expression threshold to consider the gene expressed in a sample. The idea is to identify what is the tissue in which the gene is more frequently expressed relative to the number of samples in the tissue.

The second metric is the Logarithmic Max Mean Expression (LMME) defined as:

$$LMME^g = log(max_{1 \leq i \leq T}(\vec{y}^g)) \tag{2.2}$$

Where $\vec{y}^g$ is a vector of length T containing the average gene expression of gene g for each tissue in the dataset. The idea is to locate the tissue in which the highest average expression is observed for gene g.

By visualizing the distributions of MPE and LMME for a set of candidate tissue-specific genes, it is possible to identify genes that have both a high percentage of expression and high expression levels for at least one tissue since the maximum of each metric is taken. The logarithm in LMME is precisely applied for visualization as the distribution of TPMs in the raw scale is highly skewed.

2.3 Gene expression data preparation

The gene expression data preparation procedures described in this section are carried out after the identification of expressed genes and tissue-specific genes (2.1.3, 3.1). They aim to improve the distributional properties of the data and to account for possible batch effects affecting the experiments.

2.3.1 Logarithmic transform

When measured, gene expression distributions frequently contain lots of values close to 0 and some very high values in comparison. These high values can be on the scale of 10 thousand TPMs or even more. To address this skewness in the distribution of gene expression measurements, it is common to transform the original measurements using a logarithmic operator. In this work, a box-cox transform with parameters $\lambda_1 = 0$ and $\lambda_2 = 1$ is used to prepare the expression data for downstream analysis [85]. This transform has the form:

$$Y_{ij} = log_{10}(X_{ij} + 1) \tag{2.3}$$

Where X_{ij} is each cell in a gene expression matrix and Y_{ij} is the corresponding cell in the transformed matrix. A logarithmic base of 10 is commonly used as it allows to more easily interpret the data after the transform. To convert back to the raw scale, one only has to think of a particular measurement as 10 to the power of any given transformed datum. Adding one (or any number) before taking the logarithm is known as adding a pseudocount in the context of gene expression. It ensures that measurements that are originally 0 are not undefined after the transform. This procedure has the property that expression values that were initially zero remain as zero after applying the formula.

2.3.2 Quantile Normalization

Gene expression data frequently suffers from the so-called batch effects which introduce unwanted variations in the measurements. In this work, the QN technique is used to diminish the impact of these effects [51]. Algorithm 1 shows the steps to perform QN on a gene expression matrix. QN assumes that samples should have approximately the same statistical distributions across genes. Deviations from this, therefore, are considered to be due to technical rather than biological variation. To correct this, QN forces the distribution of every sample to be the same by taking the average expression of each quantile as a reference [50]. Quantiles can be conceptualized as genes ranked equally by expression value across samples. The resulting normalized expression matrix after applying QN has higher comparability between samples than the original matrix.

Algorithm 1: Quantile Normalization

```
input  : An expression matrix $E$ with $n$ genes in the rows and $m$ samples in the columns
output: A normalized matrix $Q$ with same dimensions as $E$
Function QN($E$):
    // Sort the sample vectors of $E$ in ascending order
    $Q \leftarrow$ initializeMatrix($n, m$);
    for $j \leftarrow 1$ to $m$ do
        $Q[,j] \leftarrow$ sort($E[,j]$);
    end
    // Average each row vector after sorting and assign the mean as the
       value of all elements in the row
    $a \leftarrow$ initializeVector($n$);
    for $i \leftarrow 1$ to $n$ do
        $a[i] \leftarrow$ mean($Q[i,]$);
        $Q[i,] \leftarrow a[i]$;
    end
    // Reorder the column vectors so that the individual values match
       their original positions in $E$
    $b \leftarrow$ initializeVector($m$);
    for $j \leftarrow 1$ to $m$ do
        $b \leftarrow$ reorder($Q[,j]$, $E[,j]$);
        $Q[,j] \leftarrow b$;
    end
    return $Q$
```

2.4 Gene expression heterogeneity characterization

In this work, some data mining techniques are used to characterize the different gene expression states that are captured in GTEx data. It is of interest to find groups of samples whose expression profiles are similar to consider all the different gene expression states with equal importance in the downstream coexpression analysis. Details on this can be found in the next Chapter (3.2). This section only contains a technical overview of the used methods.

2.4.1 Principal Component Analysis

PCA is a linear algebra-based technique that allows for the creation of linear combinations that successively maximize the variance across the p variables of a dataset X [70]. X consists of a matrix of p column vectors and n row observations. The ”new variables” captured in the linear combinations are the PCs and they have the property of being uncorrelated with each other. PCA is popular in domains working with large datasets as it allows for efficient and meaningful data exploration due to its variance maximization property. While PCA does not reduce the dimensionality of datasets perse, it allows for prioritizing the strongest PCs (those capturing most of the variance) and for discarding a large number of PC dimensions that have smaller effects in the trends describing the data.

The idea behind PCA is to find the direction along which the variance of the observations in X is maximized after they have been projected into a line or plane which reduces the dimensionality of X by one. To achieve this, the concept of eigenvalues and eigenvectors is used. It is proven result that the unit eigenvectors of the covariance matrix of dataset X represent the lines or planes with the desired characteristic of successive variance maximization after they have been sorted descendingly by their eigenvalue [70]. The latter can be applied repeatedly to know what lines or planes are adequate to successively project the data into new low-dimensional spaces.

Eigenvectors are special vectors whose direction does not change after being passed through a linear transformation. Such a transformation in this case is represented by the covariance matrix of X. The only effect that the transform has over these vectors is that it can scale them (e.i. change their longitude) by a constant factor λ which is also the eigenvalue of the eigenvector. These constants in PCA directly represent the variance of the observations after projection into its corresponding eigenvectors, so they are the objective of maximization.

When all eigenvectors have been found, many computational tools for PCA calculation consider them all in a pxp matrix W where the first column represents the eigenvector with the largest eigenvalue and the last column the eigenvector with the smallest eigenvalue. After this, the final observations-by-PCs matrix Y can be calculated by computing:

$$Y = (W^T X^T)^T \tag{2.4}$$

Note that Y has the same dimensions as X, but now it is expressed in terms of the eigenvectors of the covariance matrix of X. From here, the strongest PCs may be selected for downstream analysis to take advantage of the properties of PCA. One will usually dispense with the PCs that give very little information about the data.

2.4.2 t-distribution Stochastic Neighbor Embedding

t-distribution Stochastic Neighbor Embedding (t-SNE) is a technique originally developed by van der Maaten and Hinton [86]. It allows for the visualization of high-dimensional datasets by producing low-dimensional embeddings that preserve the local and global structure of the original data in general. In PCA the priority is keeping dissimilar observations far apart by maximizing variance during the projection of points in lower dimensions. In t-SNE, the priority is keeping similar points together in non-linear low-dimensional representations of the data. This makes t-SNE more appealing for visualizing potential groups of observations that could be present in the data.

t-SNE starts by computing the euclidean distances between all pairs of observations in a dataset X with n row observations and p column features or variables. x_i indicates the ith row vector in X. These distances are transformed into probabilities using the Gaussian probability density function centered at each x_i. These probabilities indicate the likelihood of x_i picking any x_j as its nearest neighbor. The standard deviation of the Gaussians is determined relative to a user-provided parameter called perplexity which intuitively controls the number of points which will be considered close to x_i. These points, therefore, are more likely to be considered nearest neighbors of x_i. The idea is to randomly project the data points into a lower-dimensional $p - 1$ space and then recalculate the probabilities between all data points using a t-distribution density function. The resulting probabilities will initially not match their analogs in the higher-dimensional space. However, the data points in the embedding may be re-arranged iteratively to fix this. The t-distribution is used in the embedding to avoid excessive crowding of points which is a problem with Gaussians.

The probabilities in high-dimensional space should be maintained in low-dimensional embeddings of X if the embedding is correctly modeled. The difference between these probabilities is measured with a metric called Kullback-Leibler (KL) divergence which t-SNE minimizes:

$$KL(P||Q) = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{q_{ij}} \tag{2.5}$$

Where P is the joint probability distribution in high-dimensional space and Q the joint probability distribution is low-dimensional space. p_{ij} can be interpreted as the probability of x_j being selected as a nearest neighbor of x_i in the high-dimensional space. The same applies to q_{ij} in the low-dimensional space. t-SNE may be used successively until a target number of dimensions (usually 2) is reached. It is worth mentioning that contrary to PCA, it is not advisable to use the results of t-SNE as input for other analyses. The reason for this is that t-SNE preserves the relative nearest neighbors structure in the original data, but not necessarily distance or density [87]. t-SNE is therefore primarily a visualization tool.

2.4.3 Clustering

Clustering is the problem of grouping objects with similar characteristics together while keeping dissimilar objects in different groups [88]. It is an instance of unsupervised classification as no prior labels of any sort are given along with the objects. More formally, for a dataset X with n objects represented as row vectors and p feature column vectors, k subsets of the objects in X should be found such that:

$$S_1 \cap S_2 \cap \ldots \cap S_k = \emptyset \tag{2.6}$$

Any object x_i belonging to some subset S belongs only to that subset and no others. This means the intersection of all S is empty thus making the collection of subsets a true partition. Many ways to cluster objects in a dataset exist. Different distance functions may be used to characterize the similarity between objects across the features of the dataset. Also, many algorithms for clustering have been described. Two of these algorithms used in this work are reviewed in this section.

2.4.3.1 Hierarchical clustering

Hierarchical clustering (HC) algorithms have many variants. Here only agglomerative HC (AHC) is described which is the algorithm that is used in this work. In AHC, every object starts in its own cluster, so there are as many clusters as data points. The objective is to merge similar clusters iteratively until all objects converge into a single cluster [88]. By keeping track of every instance of two clusters merging, it is possible to construct a dendrogram that indicates which clusters merged and at what point in the algorithm this happened. Typically, it will be up to the user to decide at what point to ”cut” the dendrogram to get a specific partition of objects.

During the first iteration of AHC, it is easy to figure out how similar any cluster is to every other cluster because these only consist of one data point. A routine computation of the distance between points in the corresponding feature dimension of the dataset will do in this first step. Euclidean distance is the most commonly used distance function. For later iterations, clusters will start having more than one object giving rise to the question of how to measure the distance between clusters. This part of the algorithm is known as linkage. There are many ways to do linkage including measuring the distance between the farthest points of the two clusters, between the nearest, or even by taking the average distance between all combinations of points.

The particular linkage method chosen in this work is Ward’s method [89]. It optimizes an objective function to choose how to merge clusters. The standard objective function returns the difference of the variances among objects in a cluster after and before merging. This value is minimized in order to keep clusters as compact as possible in every iteration. AHC is very popular in genomics as it allows for visualizing objects (e.g. Genes, samples, or both at the same time) in dendrogram-ordered heatmaps. In these visualizations, highly similar objects can be grouped together and the patterns of their measurements can be clearly seen graphically across the colored heatmap cells.

2.4.3.2 Louvain-Jaccard clustering

The Louvain community detection algorithm is a graph-based clustering technique [90]. It has been used successfully in genomics to find groups of similar samples and cells [91]. The input is a k-Nearest Neighbors (kNN) graph constructed from a distance matrix computed for all possible pairs of objects in the dataset using some distance function chosen by the user. The algorithm begins by putting each node (object) in the graph in its own cluster at first. Louvain clustering works with a measure known as modularity [92], that intuitively represents the number of observed edges in a cluster minus the number of edges expected at random.

Modularity is maximized iteratively until no gain in modularity is possible by reassigning nodes to different clusters. All nodes in a cluster are then aggregated into a single node in a new graph and the process is repeated. Edges of the input graph may be weighted to influence modularity according to domain knowledge. In this work, Jaccard weights are used (Louvain-Jaccard clustering) [91]. These weights are defined as:

$$w_{E(i,j)} = \frac{\mid \{\forall k \mid E(k,i)\} \cap \{\forall k \mid E(k,j)\} \mid}{\mid \{\forall k \mid E(k,i)\} \cup \{\forall k \mid E(k,j)\} \mid} \tag{2.7}$$

Where $E(i,j)$ denotes an edge between sample i and sample j in the kNN graph. The numerator corresponds to the cardinality of the intersection between all nodes connected to sample i and all connected to sample j (e.i. The shared neighbors). The denominator is similar but for the union of these sets instead. The effect of applying these weights to all edges in the graph is that samples with a high number of shared neighbors relative to their total number of neighbors will tend to cluster together.

2.4.3.3 Clustering validation indices

Results of clustering algorithms are to some extent dependant on the parameters used to run them. For instance, for the algorithms reviewed in this section, one could vary the cut points in the HC dendrogram or the number of nearest neighbors used as a base for the Louvain-Jaccard clustering input graph. To help estimate what parameters result in better clustering partitions, many validation indices have been designed to assess the quality of clustering.

Two very popular internal validation indices are the Silhouette index [93] and Dunn's index [93, 94]. They receive the denomination of internal because they rely only on the clustering itself and not on external labels or data for their evaluations. The Silhouette index for any object i in a clustering partition is defined as:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{2.8}$$

Where $a(i)$ is the average distance (usually euclidean) from object i to the other objects that have been assigned to the same cluster as i. $b(i)$ is the distance from object i to its closest neighbor that has been assigned to a different cluster. This index will vary in the range $[-1, 1]$ where -1 indicates a very likely misassignment because objects in other clusters are more similar to i than the objects in the same cluster as i. 1 indicates that the cluster assignment is very likely to be correct. To obtain a global Silhouette index for the partition, the index of all objects is averaged. Dunn's index is defined as:

$$D(\mathcal{C}) = \frac{\min_{C_k, C_l \in \mathcal{C}, C_k \neq C_l} \left(\min_{i \in C_k, j \in C_l} \operatorname{dist}(i,j)\right)}{\max_{C_m \in \mathcal{C}} \operatorname{diam}\left(C_m\right)} \tag{2.9}$$

Where $\mathcal{C}$ is the clustering partition to evaluate consisting in m clusters C. The numerator indicates the minimum distance between any two objects i and j belonging to different clusters. The denominator indicates the maximum diameter $diam$ found between any two objects belonging to the same cluster. The idea is that the clustering algorithm should produce a partition with high inter-cluster variability and low intra-cluster variability. Dunn's index evaluates this by only looking at the extreme cases in the partition. The result is a value in the range $[0, \infty]$ which should be maximized.

2.5 Coexpression metrics

In this section, 3 different statistical tools for studying and estimating associations between two random variables are reviewed. These measures are by no means the only tools that can be used to characterize these associations or even specifically coexpression, but they represent the ones used for downstream coexpression analysis in this work.

2.5.1 Chi-square test of independence

One of the most useful tools in statistics to compare groups consisting of different categories is the Chi-square statistic which can be obtained through a Chi-square test of independence [95]. A basic example of the use of this test would be comparing two groups of patients with the same illness. One group receives treatment and the other does not. Once the treatment period for the first group is finalized, a survey is carried out where each patient from both groups answers if whether or not they notice improvement over their illness. Naturally, one would anticipate patients that received treatment to feel better in general. In other words, one would anticipate that the observed frequency of patients who received treatment and that felt better (e.i. The intersection) was greater than that expected by random chance. Moreover, one could also anticipate the same for those patients who did not receive treatment and that did not feel better. The other two cases, which are feeling better with no treatment or not feeling better with treatment would probably have less than expected counts as the patients have distributed more prominently in the first two mentioned groups. The Chi-square test analyses this information and summarizes it in a statistic.

The Chi-square test can easily be extended to analyze many categories. Take for instance the proposal that is being made in this work: consider two vectors $\vec{x}$ and $\vec{y}$ containing gene expression values for two different genes across a common set of n samples. These vectors, which initially contain real numbers, may be discretized using various criteria and techniques. Consider the case where 3 categories have been independently defined for each vector: *low* expression, *medium* expression and *high* expression. Each value in the vectors is classified into one and only one of these categories.

Table 2.3 shows an example of a contingency table, which is the structure from which one may compute a Chi-square statistic. Observed counts (Obs) can be found highlighted in yellow. these are simply the cardinalities of the intersections of the corresponding categories for both genes. A useful intuition for interpreting these tables and the observed counts consist of interpreting it by rows. For example, in the first row, the distribution of the samples that have a low expression for gene x is conceptualized in terms of all possible expression levels of y [32]. The same concept applies to columns (gene y) relative to the rows. The green cell shows the total number of samples n (length of both gene vectors) which one would obtain by summing the observed counts from all cells. To calculate the expected counts for each cell i, j the following expression is used:

$$Expec_{i,j} = \frac{M_i^R \cdot M_j^C}{n} \tag{2.10}$$

Where M_i^R are the row marginals for row i and M_j^C are the column marginals for column j. These marginals are shown in blue in the example Table 2.3. They are simply the sum of the row or column observed counts. Expected counts show the anticipated cell frequency if the counts of the table were distributed randomly while taking into account the observable number of samples in the studied categories. For example, when computing the expected count for low expression on gene x and high expression on gene y, the total number of low expression counts for x across all categories and the total number of high expression counts for y across all categories are taken into account in the marginals. Once the observed and expected counts for each cell i, j have been obtained, the Chi-square of the cell shall be obtained by the expression:

$$\chi^2_{i,j} = \frac{(Obs_{i,j} - Expeci, j)^2}{Expec_{i,j}} \tag{2.11}$$

The Chi-square of a cell will increase with bigger differences between observed and expected counts. It will tend to 0 when these values are similar. The final Chi-square statistic for the pair of genes in question is simply the sum of all individual Chi-square cell values:

$$\chi^2 = \sum \chi^2_{i,j} \tag{2.12}$$

Example Table 2.3 shows the final Chi-square in red. In this case, it is indeed a very high Chi-square when compared to statistics arising from random data with the same structure. This suggests that the counts of expression levels observed for this pair of genes as a whole are shifted compared to a random distribution. The latter at the same time suggests the existence of some sort of biological interaction between the genes that causes the expression levels to not behave randomly, but in a rather coordinated way. A given expression level in gene x tends to correspond with a specific level of expression of gene y. Notice how it is possible to tell the exact contribution of each one of the cells to the final statistic. In the example, it is easy to see that most of the contribution comes from samples that have low expression in gene x and high expression in y.

$\sum \chi^2_{i,j} = 213.23$	Gene y				x Marginals
	Categories	Low Expression	Medium Expression	High Expression	*Obs* Row sums
Gene x	Low Expression	$Obs = 342$ $Expec = 384$ $\chi^2_{1,1} = 4.5$	$Obs = 16$ $Expec = 17$ $\chi^2_{1,2} = 0$	$Obs = 53$ $Expec = 11$ $\chi^2_{1,3} = 163.5$	411
	Medium Expression	$Obs = 931$ $Expec = 911$ $\chi^2_{2,1} = 0.4$	$Obs = 45$ $Expec = 39$ $\chi^2_{2,2} = 0.8$	$Obs = 0$ $Expec = 26$ $\chi^2_{2,3} = 25.8$	976
	High Expression	$Obs = 599$ $Expec = 578$ $\chi^2_{3,1} = 0.8$	$Obs = 20$ $Expec = 25$ $\chi^2_{3,2} = 1$	$Obs = 0$ $Expec = 16$ $\chi^2_{3,3} = 16.4$	619
y Marginals	*Obs* Column sums	1872	81	53	$\sum Obs_{i,j} = n = 2006$

Table 2.3: Example of contingency table and the information extracted from it to compute the Chi-square statistic. The data displayed is real data obtained in this study.

The significance of each cell in a contingency table may be subject to a statistical test. There are several approaches for this including a Fisher exact test which delivers more exact p-values when the sample size is low and the observed values in the cells are also low [35]. However, an asymptotic approach is usually enough if the sample size is not reduced [34]. With the asymptotic method, the significance can be computed by finding the standard residuals of each cell:

$$stdres_{i,j} = \frac{Obs_{i,j} - Expec_{i,j}}{Expec_{i,j}} \tag{2.13}$$

Each standard residual $stdres$ can be used as a z-score to assign a p-value to the cell using the Standard Normal Distribution. Because of this, the Chi-square test is highly informative about the source of the statistic. This is one of the nice properties that this test has. It is also a non-parametric test that does not have special requirements regarding variance or homoscedasticity of the input data [95]. A p-value for the final Chi-square statistic may be obtained by looking at the random Chi-square distribution that matches the degrees of freedom (DFs) of the contingency table used in the test. DFs are given by:

$$DFs = (|cat^R| + |cat^C|) - 2 \tag{2.14}$$

Where $|cat^R|$ is the number of categories in the rows and $|cat^C|$ the number of categories in the columns. For the example in Table 2.3, the random distribution arising from $df = 4$ would be used to compute the area under the curve between 0 and the obtained statistic ($\chi^2 = 213.23$). Taking the complement of this probability is the p-value of the test, which in this case is virtually 0. From here, a significance level of 5% may be used to reject the null hypothesis in this specific example. The null hypothesis is that the tested contingency table follows a random distribution, implying that the expression levels of the genes are independent of one another.

2.5.1.1 K-means for gene expression discretization

Gene expression needs to be discretized to carry out the computation of the Chi-square statistic for coexpression. In this work, the k-means clustering algorithm is used for this purpose [96]. Very much like other clustering algorithms (2.4.3), k-means finds groups of objects in a dataset that have high similarity among themselves and low similarity with other groups. In this context, however, the goal is only to find the values in a gene expression vector x that bin the vector into k bins such that the within-bin variance is minimized and the between-bin variance is maximized. For this task, what k-means has to do simplifies to a breaks optimization problem where only the values that separate extreme objects between clusters are of interest.

K-means begins by initializing k random or user-provided centroids. Each data point is assigned to its closest centroid determined by some distance function. The centroids are then updated with the mean of the data points that have been assigned to them. The process is repeated iteratively until some stopping criterion is met. The algorithm is usually applied to multi-dimensional data, but in this work specifically, it is applied to each gene expression vector separately. This means that k-means will work with the gene expression values in a 1-dimensional line similar to what is done in breaks optimization.

To discretize the GTEx gene expression vectors, a parameter of $k = 3$ centroids to obtain 3 expression categories per gene is used. These categories represent *low*, *medium* and *high* expression. Since k-means can be sensitive to data outliers, the extreme values of every gene vector are ignored when finding the values that bin the vector into the desired categories. More specifically, values with a gene expression magnitude greater than 99.8% of the rest of the values in the vector or lesser than 0.02% are ignored for the k-means runs. The centroids are not initialized randomly, but rather at the values of the quantiles 0, 0.5, and 1 for each gene vector (without the extreme values). This allows for a deterministic discretization of each vector into the desired gene expression categories.

2.5.2 Pearson Correlation Coefficient

The most widely used statistical tool for coexpression analysis is the PCC [22]. The PCC has been shown to be a good estimator of known confirmed gene functional associations and hence it is also expected to be trustworthy in the discovery of potential novel associations. Regarding the performance of the PCC, it is known that features derived from PCC coexpression can predict the relationships existent between well-characterized genes with known shared biological functions [97]. This has been tested with genes involved in the same metabolic pathways (e.g. Glycolysis, immunoglobulin synthesis, etc.) as documented in projects that organize curated biological pathway annotations such as GO [67]. The typical definition of the PCC of a population's sample can be used to calculate the correlation of expression between a pair of genes:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.15}$$

Where x_i is the *ith* element (sample) of the expression vector of gene x and $\bar{x}$ is the mean of all values in the vector. The same definitions apply for gene y. Both vectors have length n which is the number of samples used for calculation (expression is measured for both genes in the same samples). The mathematical expression in the numerator corresponds to the covariance between the genes and the denominator standardizes it by the product of the standard deviations of each of the genes.

The PCC can take any value between $[-1, 1]$ and it does not have units. As the PCC approaches 1, it is said that the analyzed correlation is positive (as gene x increases in expression so does y). As it approaches -1, a negative correlation is said to take place (as gene x increases in expression, y decreases, or vice versa). If the result is close to 0 there is evidence of no correlation. In other words, the behavior of one gene can be described as random with respect to the other. The computational calculation of the PCC is easy and well suited for large datasets, but as addressed before (1.4.2), the measure is not general. Only linear relationships may be detected with the PCC.

2.5.3 Spearman Rank Correlation Coefficient

The SRCC can also be used to estimate coexpression [33]. Traditionally, it has not being used for large-scale coexpression projects, but it is a popular choice in many domains for correlation analysis. For calculating the SRCC, the pair of input gene vectors must be ranked before the correlation calculation (increasing computing time). However, it allows for detecting other non-linear relationships between the variables such as exponential trends [23]. The definition of the SRCC for a population's sample is given here:

$$r_{xy}^{s} = \frac{\sum_{i=1}^{n}(x_i^s - \bar{x}^s)\cdot(y_i^s - \bar{y}^s)}{\sqrt{\sum_{i=1}^{n}(x_i^s - \bar{x}^s)^2 \cdot (\sum_{i=1}^{n}(y_i^s - \bar{y}^s)^2}} \tag{2.16}$$

Where x_i^s in the context of gene expression is the rank of the *ith* biological sample in the expression vector of gene x relative to all other values in x. $\bar{x}^g$ is the mean of the ranks of x. The same definitions apply for gene y. Note that this formulation is correct even in the case of tied ranks when one must assign a fractional rank to the tied observations (average tie-breaking method). The interpretation of the SRCC has the same intuition as the PCC.

2.6 System-level coexpression calculation strategies

The problem of sample type overrepresentation has already been discussed as an important consideration for coexpression calculation in the large-scale context (1.3.2.3). Three strategies for mitigating this problem are considered and investigated. These are described in this section and summarized graphically in Figure 2.4.

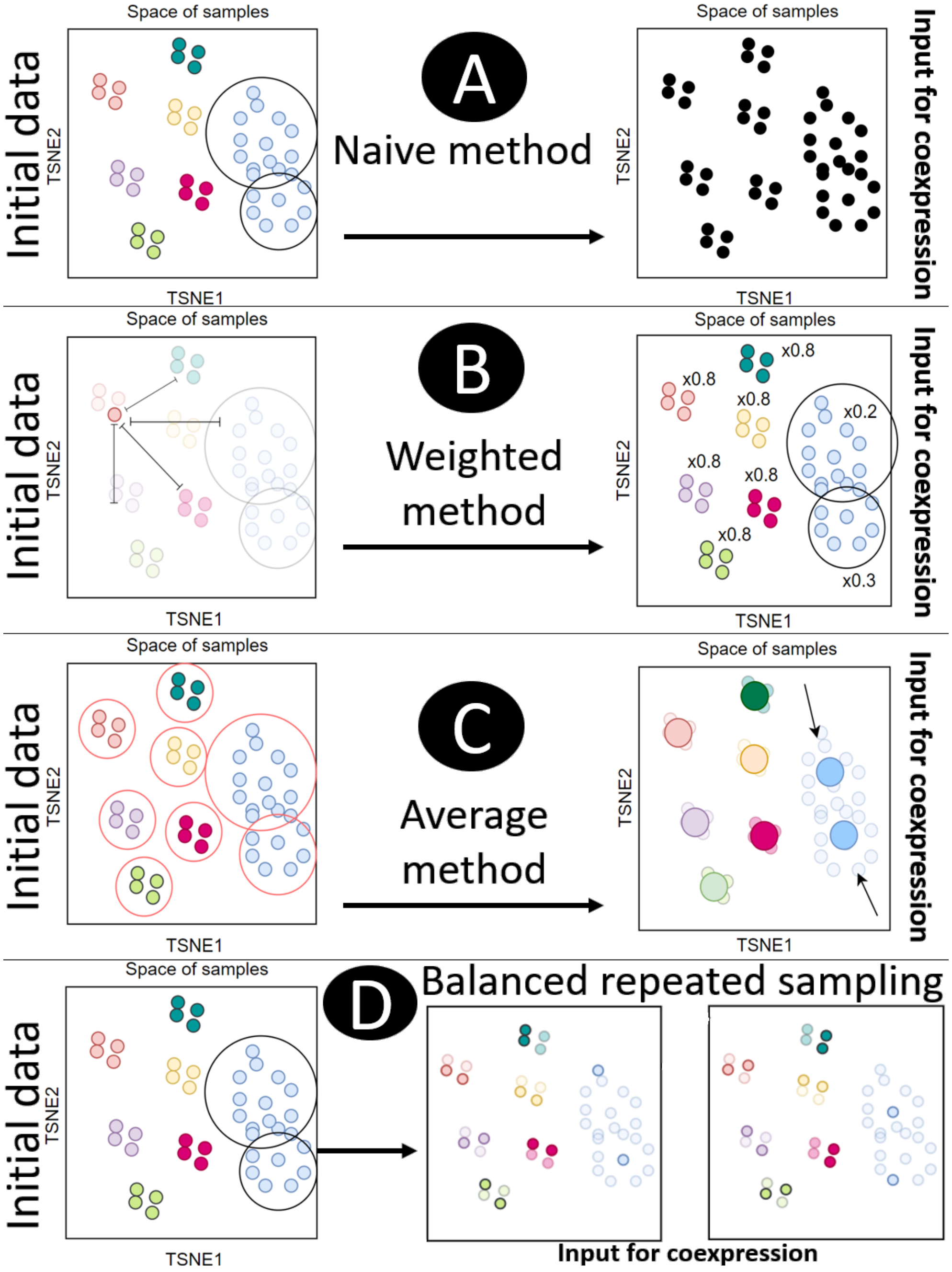


Figure 2.4: Coexpression methods studied to address sample type overrepresentation and redundancy in the input gene expression data. For simplicity, it is assumed that samples are being visualized in 2 dimensions via some tool such as t-SNE (2.4.2). A: each color represents a tissue type or high-level cluster made of similar tissue types. Subclusters inside these groups may be present. All samples are taken for coexpression analysis without further considerations in the naive method. B: calculate the similarity of each sample with all other samples. Weight down samples if there are more samples alike (redundant samples are weighted down more). At the moment of computing coexpression, sample importance is shrinked based on weights deriving from sample similarities (redundancy). C: start by clearly identifying all tissues/clusters and also subclusters if there are any. Average each cluster and compute coexpression. The new ”samples” are centroids. D: sample each tissue/cluster n times, taking k samples in each iteration. In the example shown $n = 2$ and $k = 2$. To get final coexpressions take quantiles (e.g. quantile 0 or minimum) or some summary statistic like average.

2.6.1 Naive method

This strategy is really just for comparison purposes as this method ignores the existence of possible bias in coexpression calculation due to certain types of samples being overrepresented in the input data. It simply uses all samples available for coexpression calculation without further considerations (See Figure 2.4 panel 1).

2.6.2 Weighted coexpression

This is the strategy that has been used in the literature of large-scale coexpression [62]. The idea is to reduce the importance of samples that are very similar to each other and whose characteristics are overrepresented in the data at the moment of calculating the coexpression (see Figure 2.4 panel 2). For clarity, the procedure is described here in terms of the PCC, but it can technically be adapted to any coexpression metric. The Weighted Pearson Correlation Coefficient (WPCC) is defined as:

$$r_{xy}^{w} = \frac{\sum_{i=1}^{n} w_i \cdot (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} w_i \cdot (x_i - \bar{x})^2 \cdot \sum_{i=1}^{n} w_i \cdot (y_i - \bar{y})^2}} \tag{2.17}$$

The WPCC computation is very similar to that of the regular PCC described in Equation 2.15. However, the WPCC weights the covariance and variances with the term w_i which refers to the weight of each sample i in the gene vectors x and y. The weights are calculated in a three-step process:

1. *Find the PCC between all unique pairs of samples*: here Equation 2.15 can be used taking x and y as sample vectors whose elements are n genes. The output of this step is a triangular matrix which is rotated and copied to the other triangle of a $B \in \mathbb{R}^{mxm}$ square matrix. m is the number of samples in the expression dataset. The diagonal of this matrix is composed of only ones representing the correlation of each sample with itself

2. *Compute a redundancy matrix from* B *using*:

$$J_{i,j} = \frac{max(0, B_{i,j} - C)}{1 - C} \tag{2.18}$$

   Where $J_{i,j}$ is each cell of a square $J \in \mathbb{R}^{mxm}$ redundancy matrix. C is a user threshold that indicates the minimum value of PCC at which two samples will start being considered as redundant. Any PCC value equal or lesser than the threshold will be transformed into a redundancy of 0 while greater values will be scaled to $(0, 1]$. 1 indicates that sample i and sample j have maximum redundancy among themselves

3. *Compute the weight of each sample* i *by applying*:

$$w_i = \frac{1}{\sqrt{\sum_{j=1}^{m} J_{i,j}}} \tag{2.19}$$

   The latter expression indicates that for each row vector in J, one must sum all the values of these vectors including the redundancy between a sample and itself. Taking the reciprocal of the square root of the sum is the weight for the sample in question. Once this is done, w_i can be applied in Equation 2.17 to compute the WPCC of all possible unique gene pairs

Since the weight for any sample i in the WPCC becomes smaller as the redundancies sum for the same sample becomes larger, the samples with larger redundancy will contribute less to the PCC at the moment of calculating gene coexpression (See Figure 2.4 panel 2).

2.6.3 Average coexpression

Average coexpression is a strategy proposed in this work. It consists of simply computing high dimensional gene expression centroids for each tissue. Coexpression is then calculated with the resulting ”samples” which are vectors containing the mean gene expression of each tissue. The idea is to balance the dataset by summarizing all samples of each tissue in a single expression vector. In fact, this strategy can generalize to summarizing any partition of samples and not just a tissue-based partition (see 2.4 panel 3).

2.6.4 Balanced Repeated Sampling

The Balanced Repeated Sampling (BRS) method is a strategy proposed in this work. It tries to mitigate sample redundancy by randomly sampling with uniform probability and without replacement exactly k samples of each tissue or cluster of a sample partition for coexpression calculation. The procedure is repeated s times (realizations) to obtain s gene coexpression matrices. The final coexpression matrix can be obtained by taking the minimum or the mean of each cell with the same indices across the s matrices. Samples may be drawn more than once across the s realizations, but only once during a particular realizations.

Algorithm 2 follows the BRS process. It is also documented in the algorithm that any order statistic may be used to obtain the final coexpression matrix and not just the quantile 0 (minimum). Just as the average method, this strategy can work with any partition of samples and not just a tissue-based one. The number of samples that can be taken from each group of samples has an upper bound given by the size of the smallest group (see Figure 2.4 panel 4).

Algorithm 2: Balanced Repeated Sampling coexpression calculation

input : An expression matrix E with n genes in the rows and m samples in the columns, a partition P of p classes over the row samples of E, the number of objects k to sample per class in P per sampling repetition, the number of sampling repetitions s, the desired quantile q across repetitions to compute the final result

output: A triangular coexpression matrix of all unique combinations of 2 genes from a set of n genes

```
Function BRS($E$, $P$, $k$, $s$, $q$):
    // Obtain $s$ submatrices of $E$, each with $p * k$ samples
    $M_1, M_2, \ldots, M_s \leftarrow$ initializeMatrix($n$, $p * k$);
    for $i \leftarrow 1$ to $s$ do
        // Initialize set of object identifiers which will be sampled in
           this iteration
        $K \leftarrow \emptyset$;
        for $j \leftarrow 1$ to $p$ do
            // Sample $k$ elements from class $j$
            $K \leftarrow K \cup$ sample($\{x \in P \mid x = j\}$, $k$);
        end
        $M_i \leftarrow$ subsetMatrix($E$, $K$);
    end
    // Obtain $s$ square coexpression matrices, where only the lower or
       upper triangle is used
    $X_1, X_2, \ldots, X_s \leftarrow$ initializeMatrix($n$, $n$);
    for $i \leftarrow 1$ to $s$ do
        $X_i \leftarrow$ coexpression($M_i$);
    end
    // Obtain the final coexpression matrix $Y$ by finding the desired
       quantile across equally indexed cells in the lower (or upper)
       triangular matrices
    $Y \leftarrow$ initializeMatrix($n$, $n$);
    for $i \leftarrow 2$ to $n$ do
        for $j \leftarrow 1$ to $i - 1$ do
            $Y[i,j] \leftarrow$ getQuantile($[\,M_1[i,j], M_2[i,j], \ldots, M_s[i,j]\,]$, $q$);
        end
    end
    return $Y$;
```

2.7 Gene expression data transformations for coexpression

Additionally to the choice of possible coexpression measures and the choice of coexpression strategies that can be used (2.5, 2.6), there is also the possibility to transform the gene expression data before coexpression calculation to obtain certain properties that may help the analysis. The transforms reviewed in this section change the gene expression datum after routine normalization (2.3). They are used for some experiments presented in Chapter 3.

2.7.1 Z-score normalization

Z-score normalization is a popular data transformation in several domains including genomics [98]. It consists in performing the following for each gene vector $\vec{x}$ in the expression matrix:

$$z_i = \frac{x_i - \mu_x}{\sigma_x} \tag{2.20}$$

Where z_i is the z-score transformed version of the ith element x_i of gene vector $\vec{x}$. μ_x is the mean of vector $\vec{x}$ and σ_x the standard deviation. The idea of z-score is to center the distribution of values in $\vec{x}$ around 0 to describe gene expression in terms of deviations from the mean expression. The transform works best when the original expression distribution approaches a normal distribution. In this case, the transformed distribution will also be approximately symmetric. The latter means that z-scores with the same magnitude but different signs are equally likely. On the other hand, for initially skewed distributions this is not necessarily true, thus making z-scores less comparable between them.

2.7.2 Principal Component Analysis regression

PCA is a technique that allows for capturing the variance observed across the features of a dataset through linear combinations known as PCs (2.7.2). In the gene expression context, consider a group of samples belonging to the same tissue or cluster. One could say that it is expected that such samples have very little biological variation between them as they all have the same biological origin. A great proportion of variation between the samples, for example, characterized through PCA, could therefore be attributable to other unwanted sources such as technical variation or batch effects representing confounding artifacts in the gene expression data.

Parsana and collaborators have investigated the effect of removing potential sources of unwanted variation from expression data in subsequent coexpression analysis by subtracting strong PCs from the data [99]. PCA in these experiments is performed with samples as observations and genes as features. The transformation consists of finding the regression residuals for each gene vector in the expression matrix as indicated by:

$$\widehat{E_i} = E_i - [\mu_i + (\beta_i \times L_{1:p})] \tag{2.21}$$

Where $\widehat{E_i}$ is the residualized expression vector of gene i resulting from subtracting the predicted values vector of a linear model (in square brackets) from the original expression vector E_i. In the linear model [100], μ_i is the intersect and $L_{1:p}$ represent the top p PCs. Each PC will be weighted by their corresponding fitted coefficients β. The ”top” principal components in the cited work are determined with a permutation-based approach [101], but can also be decided by a captured cumulative variation threshold criteria.

In the source paper of this transformation, authors were not working with large scale coexpression and only restricted their analyses to the 5000 most variable genes. To translate the concept to large-scale coexpression, in this work the most variable genes per group of samples are first found to guide the transformation. To do this for each group of samples, a scatter plot of the Median Absolute Deviation (MAD) and the ratio MAD over median of each gene is analyzed via a Local Regression (LOESS) [102]. The ratio MAD over median expression is analogous to the coefficient of variation for standard deviation and mean. LOESS is used to fit a smooth curve of local quadratic polynomials to the scatter plot (see Figure 2.5 for an example).

Using a fitted LOESS model as a reference helps to identify genes with high variability relative to their expression levels across samples of the analyzed tissue or cluster. From there, another filter is applied that discards genes with low overall expression by simply looking at the distribution of median expression values as shown in Figure 2.6. Similar patterns across groups of samples were found in this analysis when looking for the most variable guide genes, so the same parameters of cutoffs are considered for all groups of samples.

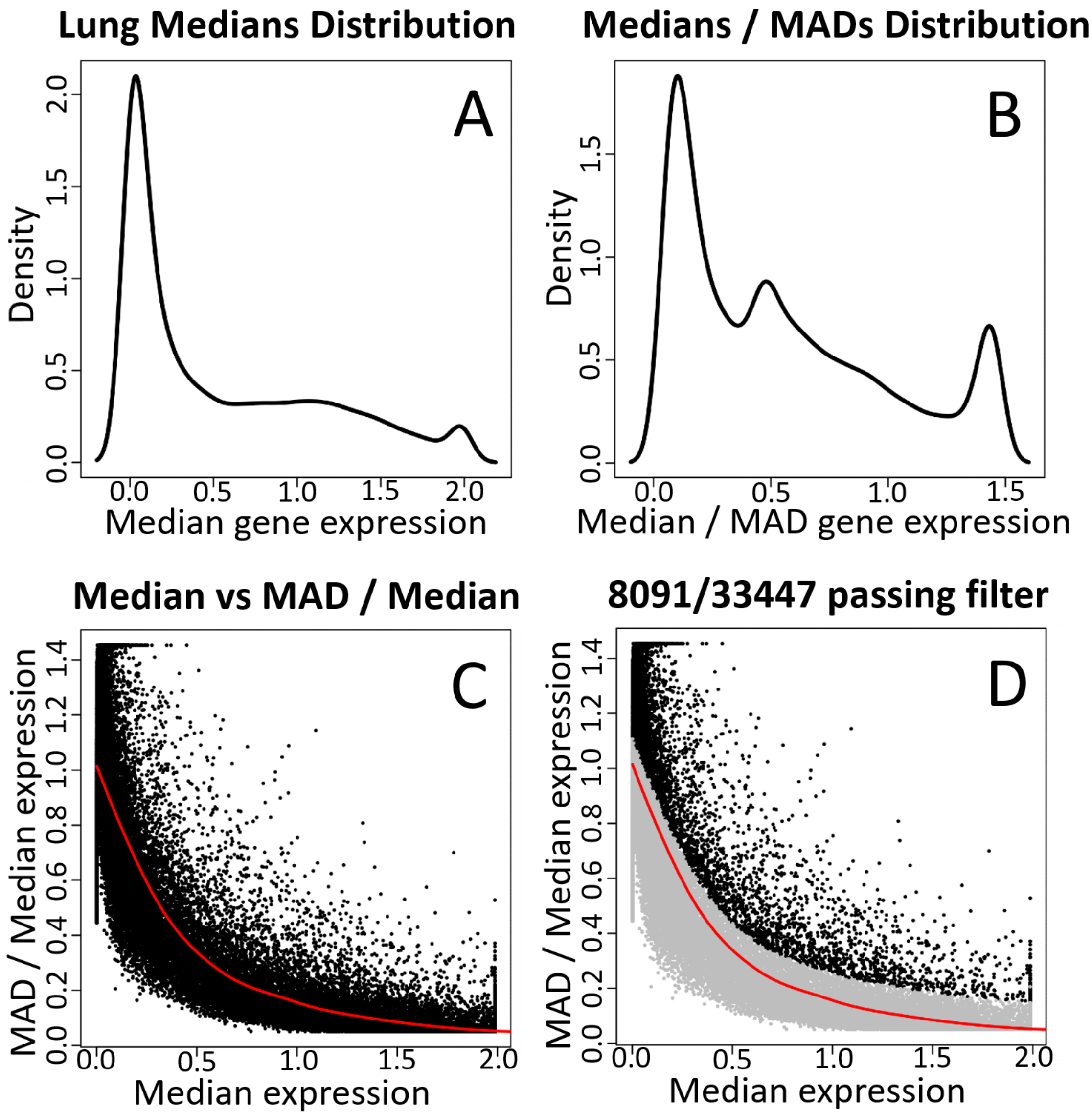


Figure 2.5: Example identification of most variable guide genes by LOESS on a group of Lung samples. A: distribution of the expression medians for each gene across Lung samples. B: distribution of the ratio expression MAD over expression median for each gene across Lung samples. C: scatter plot of genes with their median expression in the x-axis and ratio MAD over median in the y-axis. The fitted LOESS curve is shown in red. C: shaded white area indicates the area under the fitted curve and the 10% of genes with the smallest measurements that surpass the fitted curve. Genes falling in this area are filtered out.

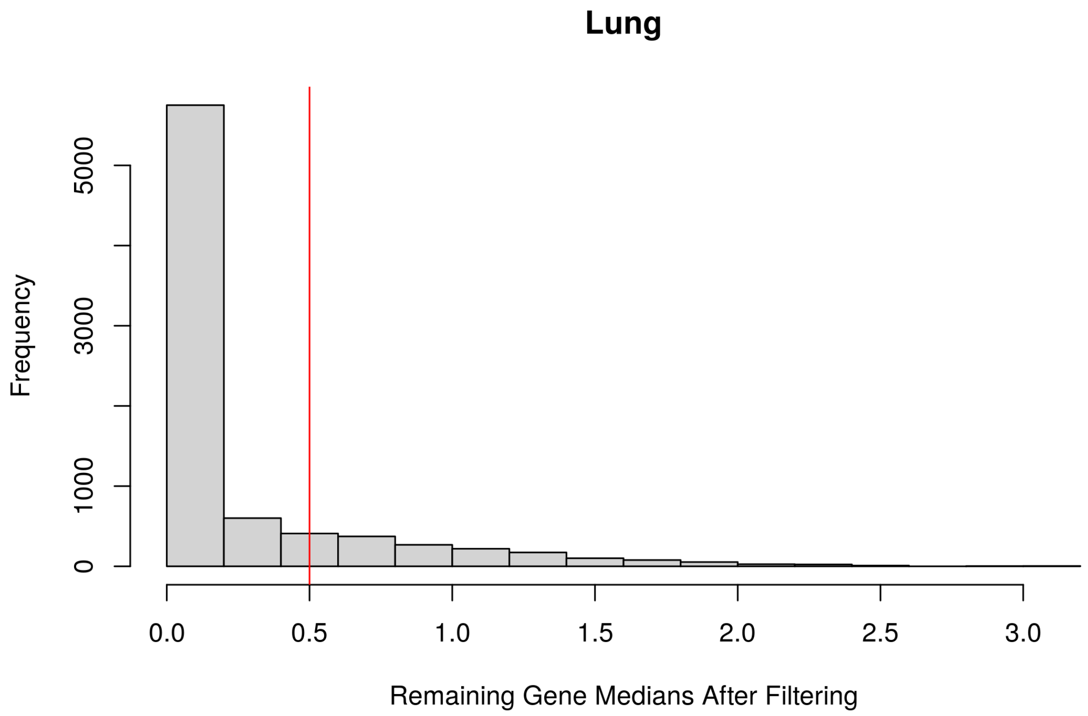


Figure 2.6: Some of the genes remaining after applying the filter shown in Figure 2.5 have very low median expression. They are filtered out here if their values lay to the left of the red line shown in this histogram. The final result comprises a set of genes with overall high median expression and high MAD over median expression ratio for the analyzed group of samples (Lung in this example).

2.8 Biological pathways validation of coexpression

It is expected that all reviewed system-level coexpression strategies (2.6) but the naive one can do a good job at preventing the sample type overrepresention problem (1.3.2.3). However, the effect that each method can have on the discovery of coexpressions is not obvious. A way to compare this effect is required. Comparing coexpression results for several methods is traditionally difficult as the ground truth is not known. In other words, the set of all true coexpressions between genes is not fully known. Despite the latter, there is a way to get an approximation of the ground truth by using biological knowledge and database resources. This is a very frequently used strategy to evaluate coexpression pipelines relative to one another [99]. It consists of assuming that any pair of genes in the same biological pathway are coexpressed.

By gathering curated biological pathways information as the observed data and considering the output of the different coexpression methods as the predicted data, a strategy similar to what is done when evaluating a binary classifier in machine learning can be applied [103]. This can be attempted because coexpression is correlated with shared functionality between genes [25].

2.8.1 Evaluation of coexpression performance

For the analysis based on biological pathways, the True Positives (TPs) and False Positives (FPs) are of particular interest as evaluation indicators. TPs constitute pairs of genes that have at least one biological pathway in common and that come up as coexpressed in a method's analysis. The FPs are pairs that do not share biological pathways but that are detected as coexpressed. These indicators are important because they provide information on how many correct (or at least highly likely to be correct) coexpressions are detected. they also speak about how many potential mistakes could have been made. More specifically, the False Discovery Rate (FDR) is analyzed and defined as:

$$FDR = \frac{FP}{FP + TP} \tag{2.22}$$

The FDR is in the range $[0, 1]$. It represents the proportion of FPs relative to the total positives (coexpressed genes) predicted by a method. FDR has been chosen to evaluate this kind of analysis based on biological pathways in the literature before [60,99]. The complement of the FDR (e.i. $1 - FDR$) is the Positive Predictive Value (PPV), also known as precision. The PPV is preferred in this work for intuition reasons as the evaluation with the PPV is better when the value is higher while with FDR it would be better if the value is lower.

The objective of the analysis is to compute the PPV for a series of thresholds set across the range of the coexpression metric. For instance, in the context of the PCC, absolute value coexpressions greater than a certain value will be considered as true coexpressions and the rest discarded. The PPV is computed when calling the retained gene coexpressions as the predicted positives. The process is repeated for increasingly stricter thresholds allowing for the creation of a PPV-PCC thresholds curve which depicts the performance of the evaluated method under different settings.

Another important metric to pay attention to is the total number of pairs of coexpressed genes found. This can be thought of as the number of edges in the coexpression graph. One of the objectives in large-scale coexpression analyses is to favor the discovery of novel associations between genes. Discovery is more likely when more coexpressed genes are found. The PPV can also increase solely because fewer coexpressions are being called as true by a given method. It is less likely to call FPs if there are fewer overall positives, so it is important to assess the number of coexpression edges along with the PPV to fairly evaluate methods.

2.8.2 Biological databases

For the experiments performed to compare coexpression strategies based on biological pathways information, the following databases are queried using the *MIGSA* R package [104] to access Enrichr's [105] repository:

- BioCarta 2016 [106]
- GO Biological Process 2018 [67]
- KEGG 2019 Human [66]
- Comprehensive Resource of Mammalian Protein Complexes (CORUM) [107]

- Panther 2016 [108]
- Reactome 2016 [109]
- WikiPathways 2019 Human [110]

A simple filter for the words ”mouse”, ”disease”, ”defect” and ”cancer” was applied to all the pathways and gene sets fetched from these databases. Some of them contained identified pathological gene interactions and mouse data. Pathways with more than 50 genes are discarded to keep only the most specific ones, an idea applied by Obayashi and collaborators [15]. Potential coexpressions between members of smaller pathways or gene sets are more likely to be true due to the specificity of the function. Some of the pathways are just different versions of the same pathway for different databases, but this is not a problem for the analysis.

2.9 Coexpression data analysis

In this section, some tools used to explore and validate the coexpression data once calculated are reviewed.

2.9.1 Permutation tests

A permutation or randomization test can be broadly defined as a non-parametric technique that compares an observed distribution of test statistics with a distribution obtained for the same statistic and the same data but with rearranged labels (null distribution) [111]. When significant associations are observed in the non-permuted data, it is expected that upon rearranging the labels and recalculating the statistic, that these associations will disappear. This implies that the significant associations observed are not produced by chance. Distributions of permuted and non-permuted input data used to calculate the target statistics are the same since no foreign values are introduced in the process. The latter makes both of the resulting distributions of statistics comparable.

In this work, a permutation-based approach is used to demonstrate that a large number of coexpression associations resulting from the system-level human coexpression estimation are not driven by chance. Gene-wise permutation-based tests are also used to estimate the weight of specific groups of samples in the system-level human coexpression calculated. This is done by recalculating coexpression once the gene expression measurements for the group of samples in question have been substituted by randomly selected values in other sample groups (4.5).

2.9.2 Statistical tests

2.9.2.1 Hypergeometric test

The hypergeometric test is useful to check the statistical significance of set intersections [58]. In this work, it is applied to characterize the coincidences between coexpressed genes found using distinct coexpression metrics. The test computes a p-value based on the size of the intersection k between the two sets tested (the value that a random variable X takes). The probability mass function of the hypergeometric distribution given by:

$$P(X = k) = \frac{\begin{pmatrix} K \\ k \end{pmatrix}\begin{pmatrix} N-K \\ n-k \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}} \tag{2.23}$$

Where K is the size of the first set. In this context, such set is a collection of genes found to be coexpressed with some gene. N is the size of the universe (total genes) and n is the size of the second set. The parameters K, N, and n determine the mass function under which the intersection size k is evaluated. A small p-value (usually $<$ 0.05) would indicate that the observed overlap between sets of coexpressed genes is very likely not due to chance.

2.9.2.2 Correction of p-values by False Discovery Rate

When performing several related hypotheses tests, the probability that some of the significant results represent wrong rejections of the null hypothesis (type 1 errors) increases naturally due to the number of tests performed. An example of this would be testing the overlaps between coexpressed genes given by different metrics for many genes. To account for this, several p-value correction methods exist. In this work, a correction by FDR or Benjamini-Hochberg method is used [112]. It consists of computing a vector of adjusted p-values $\vec{q}$ using:

$$q_i = \begin{cases} p_i, & if \;\; r_i = n \\ q_{i-1}, & if \;\; q_{i-1} <= p_i\left(\frac{n}{r_i}\right) \\ \frac{n}{r_i}, & if \;\; q_{i-1} > p_i\left(\frac{n}{r_i}\right) \end{cases} \tag{2.24}$$

Where p_i is the ith p-value of a vector $\vec{p}$ containing the unadjusted p-values sorted in descending order. n is the total number of p-values. r_i is the rank of p_i so that the biggest unadjusted p-value has a rank n and the smallest rank 1. Once computed, the p-values in $\vec{q}$ shall be interpreted as their unadjusted counterparts in terms of significance.

2.9.2.3 Wilcoxon signed-rank test

The Wilcoxon signed-rank test is a paired non-parametric statistical test that helps to determine if two samples come from different populations by looking at their mean ranks [113]. A null distribution of the test's statistic W has mean 0 and variance $\frac{n(n+1)(2n+1)}{6}$. n is the number of paired observations in the tested samples whose difference is not equal to 0. The observed statistic can be calculated with:

$$W = \sum_{i=1}^{n} \left[\text{sign}\left(x_{2,i} - x_{1,i}\right) \cdot r_i\right] \tag{2.25}$$

Where $x_{2,i}$ and $x_{1,i}$ are pairs of observations from each of the two samples tested sorted in increasing order by the values resulting from their absolute differences. $sign()$ is a function that returns the sign of the number used as input. r_i is the corresponding rank of a pair of observations where the pair that has the smallest absolute difference has rank 1. When obtained, the observed W statistic may be compared to its matching null distribution by using a critical value to obtain a p-value. In this work, the Wilcoxon signed-rank test is used to compare paired results between coexpression metrics (4.4.3).

Chapter 3

Preparation for system-level coexpression

In this chapter, results obtained before proceeding to the calculation of system-level coexpression are presented. These results had to be obtained before the main results of the project that are presented in Chapter 4 as they provide the base for subsequent robust coexpression calculation. Details on the data, methods, and algorithms used are documented in Chapter 2.

3.1 Tissue-specific genes

The original GTEx expression matrix was processed with a general low expressed genes filter resulting in a new matrix of 31,494 genes that remained from the initial 56,200 (2.3). The 24,706 genes that were initially filtered out were investigated to see if they were tissue-specific and worth to keep for downstream analysis. The metrics used for this purpose were the MPE and LMME (2.2). This is done because the initial consideration of expressed genes does not account for tissue-specific genes which are only expressed in specific subsets of samples. In this work, the term ”tissue-specific gene” is used loosely since it is not required for a gene to be expressed in exclusively one tissue for it to be considered specific. Instead, the term is used here to refer to genes that are expressed in *some* subset of samples where subsets represent tissue types. If a gene is clearly expressed in at least one tissue (could be more), and it is not expressed in more than 20% of samples across all tissues, then it is termed tissue-specific in this work.

For the analysis, tissues with very low sample counts (less than 50 samples) were not considered. For each of the 24,706 tissue-specific candidate genes, the MPE and LMME metrics were computed. In the case of the MPE, the TPM threshold to consider a gene expressed in any sample was increased 10-fold to 1 TPM compared to the 0.1 threshold used in the first general gene filter (2.3). This is done in order to have higher confidence when considering a gene expressed for a particular tissue or tissues only.

Figure 3.1 shows the scatter plot of the paired MPE and LMME metrics for each of the analyzed genes. Histograms of the distributions of both metrics are also shown along with a density plot. There is particular interest in the upper-right area of the plot as genes there have a high average expression in at least one tissue. Genes with a high percentage of expression in at least one tissue are also in this area.

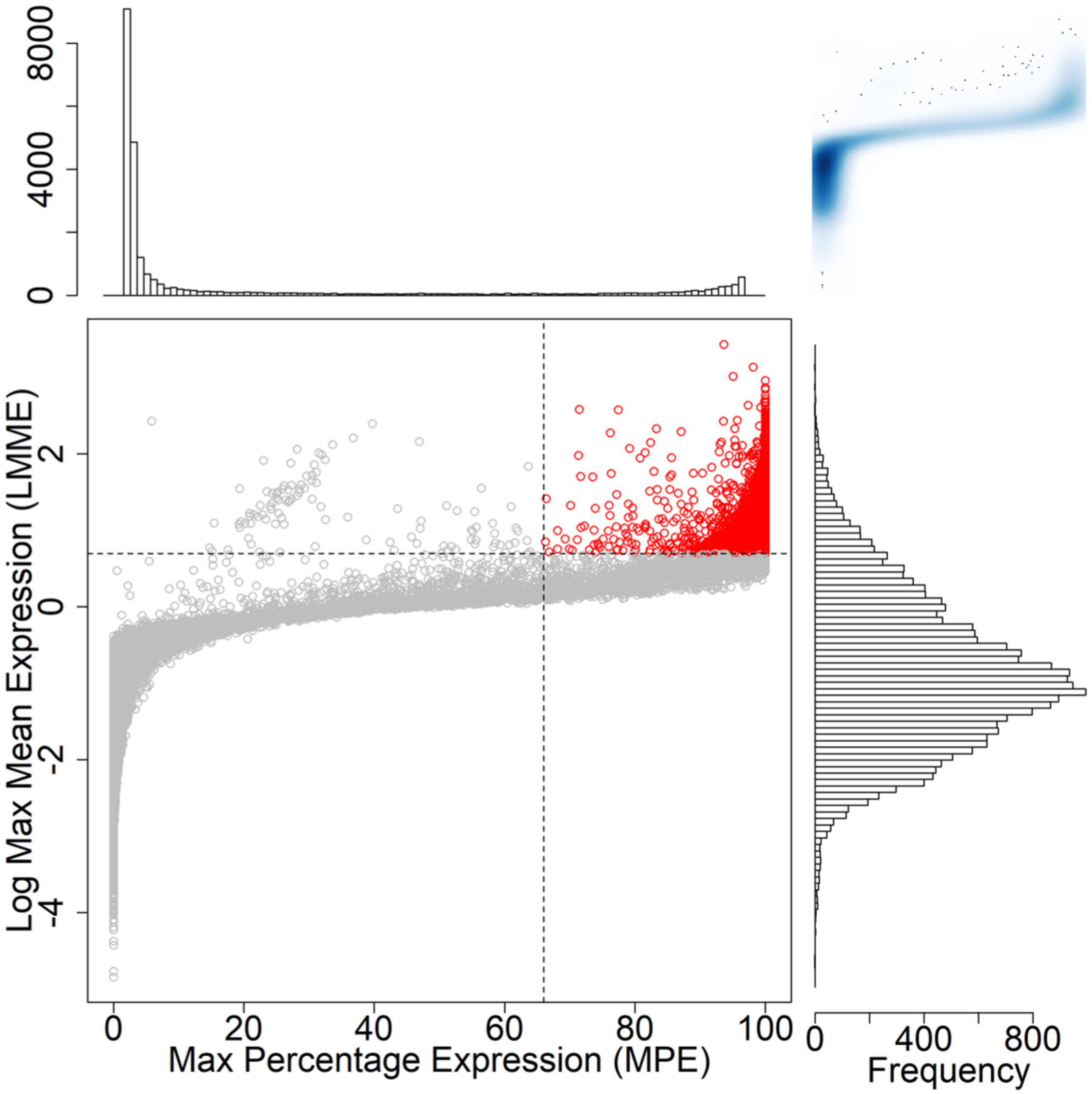


Figure 3.1: Investigated genes for tissue specificity. Genes highlighted in red are rescued from the initial general gene filter due to good evidence that they are tissue specific. Detail on these genes can be seen in figure 3.2.

Based on the observed trend, it is proposed that in order to consider a gene as tissue-specific and include it in the project for downstream analysis, it must have an MPE greater than or equal to 66% and a LMME greater than or equal to 0.69897 (TPM of 5 in the raw scale). These criteria are visually estimated to capture those genes which are tissue-specific with high probability. Only 1,953 genes out of the 24,706 genes investigated fulfilled the requirements established through visualization of the MPE and LMME metrics. These genes are rescued from the initial gene filter based on all samples in the dataset which had left them out. There is enough evidence to suggest that these genes are expressed, just not globally, and rather on specific tissues. This quality distinguishes these genes from non-expressed noisy genes. In the end, a grand total of 33,447 genes are considered for downstream analysis.

As shown in Figure 3.2, it was detected that the majority of rescued genes exhibit their

MPE and LMME in the testis tissue. This is consistent with literature observations in which testis is regarded as a tissue with a high number of tissue-specific genes [114].

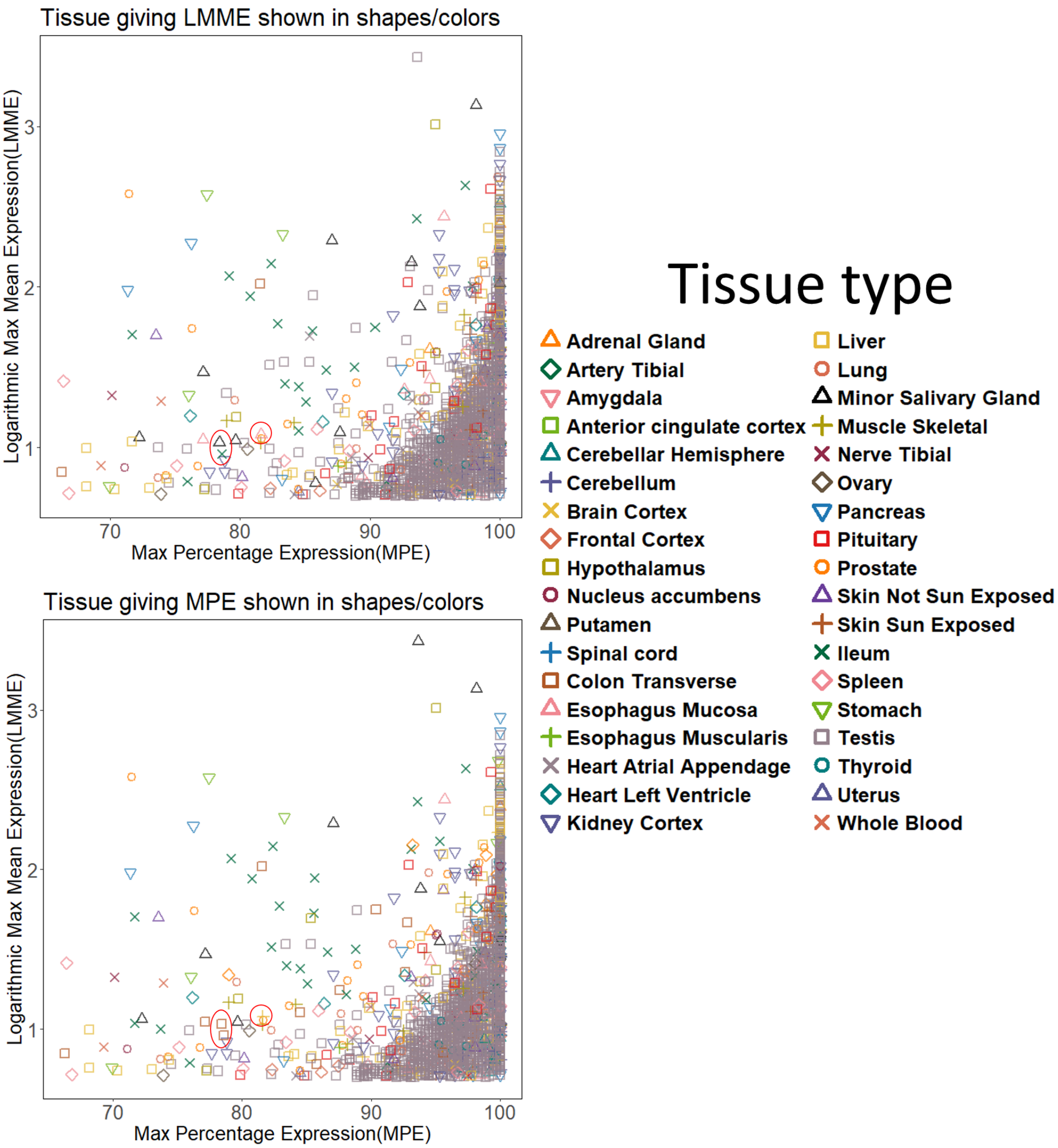


Figure 3.2: Tissue type in which the MPE or LMME is observed for GTEx genes recovered from the initial general low expressed genes filter. The tissue of MPE and LMME tends to be the same across genes except for a few exceptional cases (circled in red). All genes shown are rescued due to them having a high mean expression in some tissue (> 5 TPM in raw scale) and a high percentage of expressed samples in some tissue (> 66%). This does not have to be exclusively for one tissue. The MPE and LMME of one gene passing the threshold may correspond to different tissues each.

3.2 Characterization of GTEx gene expression states

The present work has as an objective to capture coexpression relationships between genes that are representative of normal human biology as a whole. For this purpose, several types of samples across a variety of tissues are considered. Some strategies are proposed to make sure that samples from certain types that are more numerous in the input data do not become overrepresented in the coexpression calculations (2.6). However, some of these strategies assume that samples are already categorized into non-redundant groups. In other words, that the tissue partition of the GTEx samples has highly similar samples labeled with the same tissue and dissimilar samples with different tissues.

To check the latter assumption that GTEx tissues represent well defined and distinct gene expression states, Louvain-Jaccard clustering (as implemented by the *igraph* R package) was performed (2.4.3.2) [115]. If tissues represent a nice partition of GTEx samples, then it is expected that clustering the samples would render a similar partition. To assign nearest neighbors to each sample in the dataset for clustering, it is necessary to calculate the similarity between the gene expression profiles of all possible unique pairs of samples. Finding these similarities using, for example, euclidean distance, is very computationally expensive in the original gene expression high dimensional space. To simplify this task, PCA was used to generate linear combinations of gene expression values that successively maximize the variance across samples (2.4.1) [70].

PCA is useful in this setting because it captures the main differences between samples. Even when considering a cumulative variance captured by the PCs as high as 99% of the total dataset variance, a dimensionality reduction from 33,447 genes (original dimensionality across the genes) to 6,947 PCs is achieved. The input kNN graph for the Louvain-Jaccard clustering is therefore calculated from a 16,704- samples-by-6,947 PCs matrix.

A t-SNE was also computed to be able to visualize the sample's high dimensional PCs in a 2-dimensional embedding [86]. The Barnes-Hut implementation of t-SNE was used to calculate an exact embedding (e.i Theta parameter of 0) with perplexity 50, a learning rate of 200, and 5000 iterations [116]. Figure 3.3 shows the t-SNE scatter plot with points labeled by tissue type. Interesting observations include that many samples from certain tissues seem to have very similar expression profiles to those observed in a different tissue at least in t-SNE space. Notable examples include a clear mix of Breast and Adipose Subcutaneous samples, a very heterogeneous mix of the Skin tissues (sun and not sun-exposed), a mix of the intestinal tissues Colon Transverse, Colon Sigmoid, and Illeum (the last portion of the small intestine), a mix of esophagus and stomach samples (potentially corresponding to the gastroesophageal junction region) and a mix of samples from several parts of the central nervous system among others.

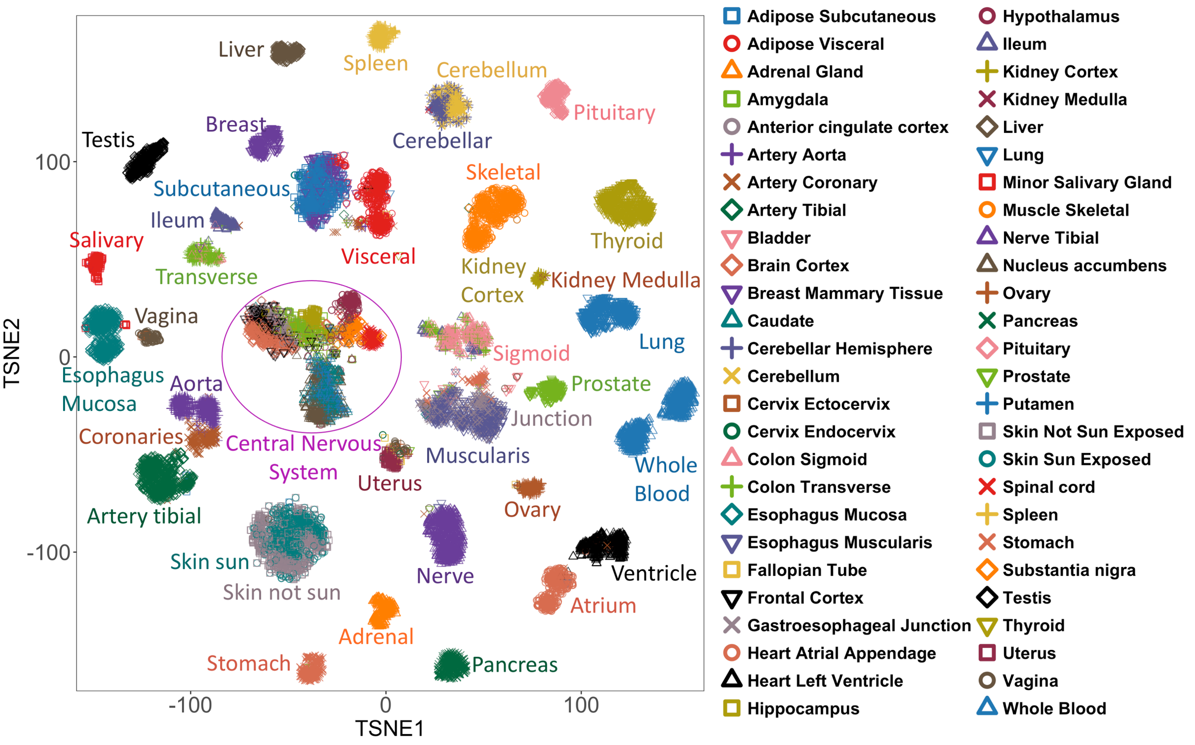


Figure 3.3: GTEx samples t-SNE plot based on 6,947 PCs with tissue type point aesthetics shown.

For deciding the nearest-neighbors parameter for the kNN graph which has to be computed prior to clustering, all possible values between 10 and 150 were explored. The Silhouette Index [93], Dunn Index [94], modularity [92], number of clusters and size of the smallest cluster were used as error estimators as presented in Figure 3.4. Partitions across all tested ranges have very similar Silhouette and Dunn indices as evidenced by the range of the values obtained. However, there are some k values that offer slightly better performance.

For the minimum cluster size estimator, it was sought that this would not exceed a value of 89. Thanks to the t-SNE plot with the tissue labels in Figure 3.3, it was a known intuition that the Kidney Cortex (85 samples) and Kidney Medulla (4 samples) tissues would probably form a very well defined group and that this would be the smallest cluster. On the other hand, smaller sizes of smallest cluster were associated with worse validation indexes. Another observation from the t-SNE plot was that there is a tendency of different tissues, in general, to merge together as opposed to samples in one tissue creating subgroups. A notable exception is Whole Blood, which clearly separates into 2 subgroups potentially due to the moment in which samples were taken [117]. Despite this, in general, fewer clusters are expected than initial tissue types. Modularity has a decreasing trend as the number of clusters increases, but for the value proposed (dotted lines in Figure 3.3), it is still bigger than 0.95.

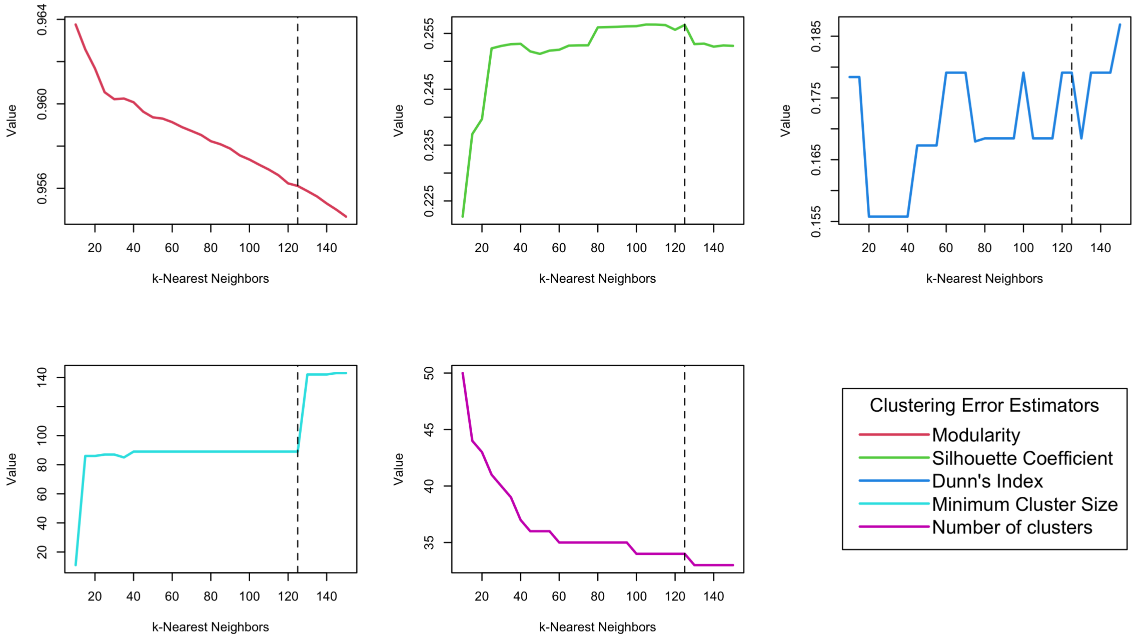


Figure 3.4: Clustering error estimators for different partitions of the GTEx samples obtained with the Louvain-Jaccard algorithm. Dotted lines indicate a $k = 125$ parameter.

Once the clustering partition was found, the previously computed t-SNE was plotted again, but labeling the points with cluster memberships instead of tissue types. As shown in Figure 3.5, the t-SNE visualization agrees very well with the partition found for the samples through clustering despite the fact that t-SNE information is not used for the clustering in any way. This is encouraging because a clearer separation of samples has been achieved when compared to the tissue-type partition seen in Figure 3.3. The names of the clusters shown in the legend of Figure 3.5 correspond to manual annotations of the clusters based on the main tissue-types found in each cluster. The latter is done to have a more descriptive description of clusters instead of simply numeric labels (i.e. Just 1, 2, ... 34). A detailed description of the tissue type composition of each cluster is given in Table 3.1.

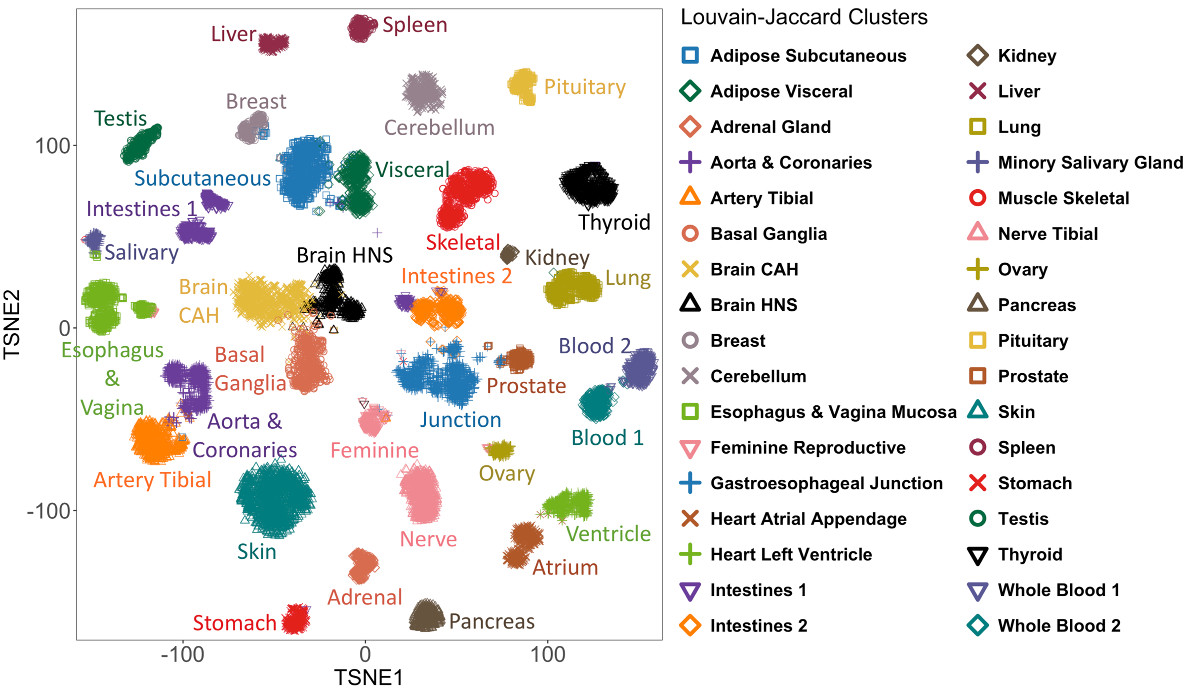


Figure 3.5: GTEx samples t-SNE plot based on 6,947 PCs with Louvain-Jaccard clustering point aesthetics shown.

Table 3.1: Tissue compositions of Louvain-Jaccard clusters found in the GTEx data.

Cluster	Total samples	Composition
Adipose Subcutaneous	909	Adipose Subcutaneous: 72.39% Breast Mammary Tissue: 22.22% Adipose Visceral: 2.97% Artery Coronary: 0.77% Skin Sun Exposed: 0.44% & 7 more
Adipose Visceral	533	Adipose Visceral: 96.25% Artery Coronary: 1.88% Artery Aorta: 0.56% Colon Transverse: 0.38% Lung: 0.38% & 3 more
Adrenal Gland	258	Adrenal Gland: 100%
Aorta & Coronaries	661	Artery Aorta: 64.75% Artery Coronary: 31.77% Artery Tibial: 2.72% Esophagus Gastroesophageal Junction: 0.3% Adipose Visceral: 0.15% & 2 more
Artery Tibial	651	Artery Tibial: 98.31% Artery Coronary: 0.77% Adipose Subcutaneous: 0.46% Uterus: 0.31% Ovary: 0.15%
Basal Ganglia	690	Caudate: 35.07% Nucleus accumbens: 33.62% Putamen: 29.28% Hypothalamus: 0.58% Amygdala: 0.43% & 4 more
Brain CAH	958	Brain Cortex: 26.41% Frontal Cortex: 21.5% Anterior cingulate cortex: 18.16% Hippocampus: 17.22% Amygdala: 14.93% & 6 more
Brain HNS	536	Hypothalamus: 35.26% Spinal cord: 29.48% Substantia nigra: 25.37% Hippocampus: 5.6% Nucleus accumbens: 2.05% & 5 more
Breast	262	Breast Mammary Tissue: 97.33% Skin Sun Exposed: 1.15% Bladder: 0.76% Adipose Subcutaneous: 0.38% Cervix Endocervix: 0.38%
Cerebellum	458	Cerebellum: 52.62% Cerebellar Hemisphere: 46.72% Brain Cortex: 0.22% Hypothalamus: 0.22% Spinal cord: 0.22%
Esophagus & Vagina Mucosa	679	Esophagus Mucosa: 81.59% Vagina: 15.17% Minor Salivary Gland: 2.95% Cervix Ectocervix: 0.29%
Feminine Reproductive	218	Uterus: 64.22% Vagina: 20.18% Cervix Endocervix: 4.13% Ovary: 4.13% Cervix Ectocervix: 3.21% & 3 more
Gastroesophageal Junction	1013	Esophagus Muscularis: 50.44% Esophagus Gastroesophageal Junction: 36.43% Stomach: 9.08% Prostate: 1.78% Bladder: 1.38% & 4 more
Heart Atrial Appendage	433	Heart Atrial Appendage: 97.92% Heart Left Ventricle: 1.15% Artery Coronary: 0.69% Lung: 0.23%
Heart Left Ventricle	433	Heart Left Ventricle: 98.38% Heart Atrial Appendage: 1.15% Artery Coronary: 0.46%
Intestines 1	506	Colon Transverse: 60.28% Small Intestine Terminal Ileum: 32.02% Colon Sigmoid: 6.32% Thyroid: 0.59% Artery Coronary: 0.4% & 2 more
Intestines 2	461	Colon Sigmoid: 73.97% Colon Transverse: 20.17% Small Intestine Terminal Ileum: 5.21% Stomach: 0.43% Esophagus Gastroesophageal Junction: 0.22%
Kidney	89	Kidney Cortex: 95.51% Kidney Medulla: 4.49%
Liver	226	Liver: 100%
Lung	577	Lung: 99.65% Amygdala: 0.17% Whole Blood: 0.17%
Minory Salivary Gland	142	Minor Salivary Gland: 99.3% Breast Mammary Tissue: 0.7%
Muscle Skeletal	806	Muscle Skeletal: 99.63% Artery Tibial: 0.12% Minor Salivary Gland: 0.12% Skin Sun Exposed: 0.12%
Nerve Tibial	626	Nerve Tibial: 98.88% Esophagus Muscularis: 0.48% Vagina: 0.32% Artery Coronary: 0.16% Prostate: 0.16%
Ovary	169	Ovary: 100%
Pancreas	328	Pancreas: 100%
Pituitary	282	Pituitary: 100%
Prostate	231	Prostate: 96.97% Bladder: 1.73% Vagina: 1.3%
Skin	1300	Skin Sun Exposed: 53.31% Skin Not Sun Exposed: 46.46% Adipose Subcutaneous: 0.08% Artery Tibial: 0.08% Vagina: 0.08%
Spleen	241	Spleen: 100%
Stomach	264	Stomach: 99.62% Colon Transverse: 0.38%
Testis	361	Testis: 100%
Thyroid	649	Thyroid: 99.85% Fallopian Tube: 0.15%
Whole Blood 1	400	Whole Blood: 100%
Whole Blood 2	354	Whole Blood: 100%

Additionally, HC was performed on a matrix containing the percentage of samples that each tissue (columns) contributes to each cluster (rows) to graphically visualize how samples from each tissue are distributed across found clusters (2.4.3.1). This is presented in a heatmap in Figure 3.6 where it is possible to see how samples in the original partition of GTEx tissues have rearranged into a different representation of gene expression states found via clustering. The HC dendrogram is manually edited for a better visualization. The partition of samples found through clustering is not perfect as evidenced by some potential outliers in some clusters which are detectable in Figure 3.5. However, the clustering partition represents a less redundant depiction of the gene expression states represented in the GTEx data when compared to the tissue-based partition.

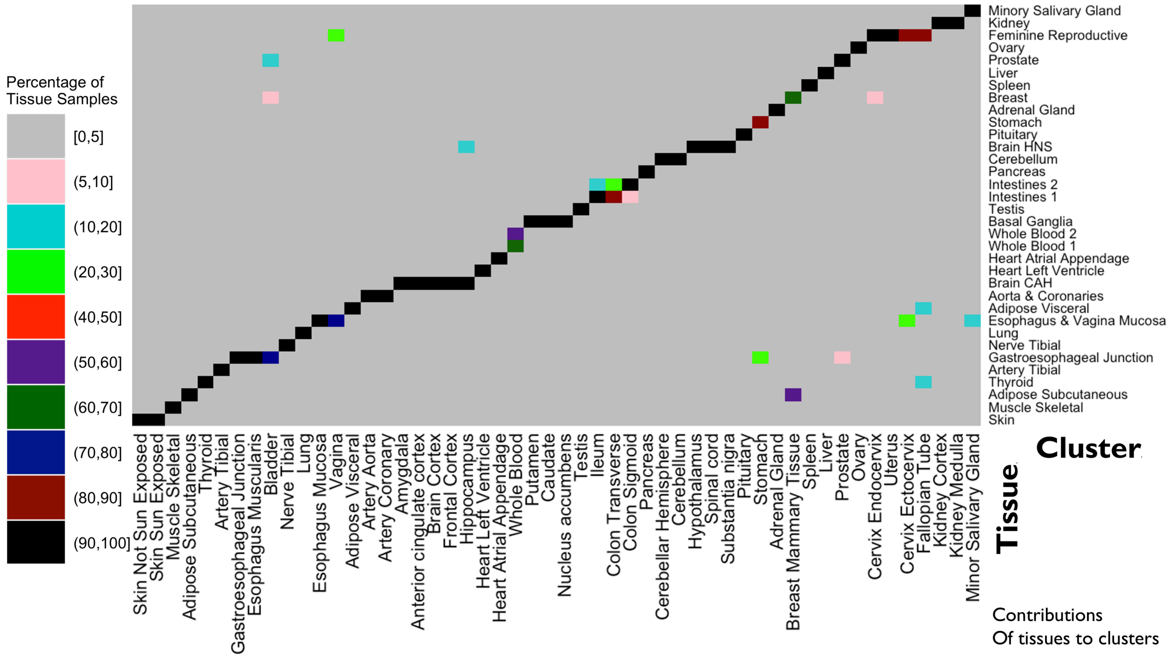


Figure 3.6: Distribution of tissue sample counts across Louvain-Jaccard clusters found in GTEx data. Sample percentages were clustered using HC and the resulting ordering of samples was edited for a cleaner visualization.

3.3 Performance of different system-level coexpression strategies

Results comparing distinct system-level coexpression strategies to tackle the problem of sample type overrepresentation are presented here (2.6, 1.3.2.3). Although these strategies address the aforementioned problem, it is initially unclear which one should be preferred over the others for coexpression calculation in this project. An analysis based on a biological pathways evaluation framework is presented (2.8). This sheds light on what methods are more robust at the moment of identifying coexpressed genes.

The experiment is set up with a total of 7,568 biological pathways and gene sets retrieved from biological databases (2.8.2). A total of 11,759 unique genes participating in one or more of these pathways were also retrieved. For testing each of the coexpression methods without exceedingly high computational costs, the experiment was carried out on a representative random sample of 5000 genes. Since the obtained genes in the pathways use ENTREZ identifiers [118], the ENSEMBL identifiers [119] that GTEx uses were mapped to ENTREZ using the *biomaRt* tool [120]. After filtering out genes that failed to map to ENTREZ, ambiguously mapped genes (i.e. A single ENTREZ mapped to multiple ENSEMBL) were assigned the expression of the ENSEMBL gene with the highest mean expression from within all ENSEMBL that mapped to the ambiguous ENTREZ. 20,619 genes remained with proper ENTREZ identifiers in the GTEx data. 11,180 of these were present in the genes from the retrieved pathways. In the end, the random sample of 5000 genes was drawn from the overlap of 11,180 genes.

All possible gene pairs in the set of 5000 test genes that had pathways in common according to the assembled reference pathway compendium were registered. 130,135 gene pairs with this characteristic were found. These pairs will be used as an approximation to the ground truth in order to evaluate the resulting coexpressions from the different investigated methods. For this analysis, only the PCC was used with the purpose of avoiding extensive computation time. PPVs were calculated for each possible absolute value PCC threshold in the range $[0, 0.9]$ with a step of 0.05 (2.8.1). When a particular coexpression was greater than or equal to the threshold, it was regarded as a discovery (predicted positive) and tested against the pathways compendium.

Sample weights for the weighted coexpression method were computed before subsetting the 5000 genes used in the experiment in order to obtain a representative result for this method. For the same algorithm, a correlation threshold of 0.4 for the calculation of sample redundancy is used as reported by the authors of this method in COXPRESSDB's documentation (2.6.2) [121]. For the average coexpression method and the BRS method, both the tissue-based partition of samples as given originally by GTEx (only tissues with more than 50 samples are considered) and the clustering-based partition found in the last section were used.

For BRS, values of $k = \lfloor(66.6 * min(|\{x \in P \,|\, x = p\}|))/100\rfloor$ (i.e. Two thirds the size of the tissue or cluster with the least samples) and of $s = 20$ are used as sampling size and number of realizations respectively. This is done to make it highly likely that all samples are considered at least once for coexpression calculation in some of the 20 realizations (2.6.4). For computing the final coexpressions when using BRS, taking the minimum across realizations was tested. A randomly selected realization of BRS was also analyzed.

3.3.1 Expression without further transformations

First, the results obtained by the researched large-scale coexpression calculations strategies when working with routine normalized expression without further transformations are presented (2.3). The PPV values for each threshold and for each method were used to build the curves shown in Figure 3.7. The same plot also shows how many discoveries each method achieved independently from if they were TPs of FPs. In order to summarize the performance of the methods in general across all thresholds, the area under each of these curves is calculated and shown in Table 3.2.

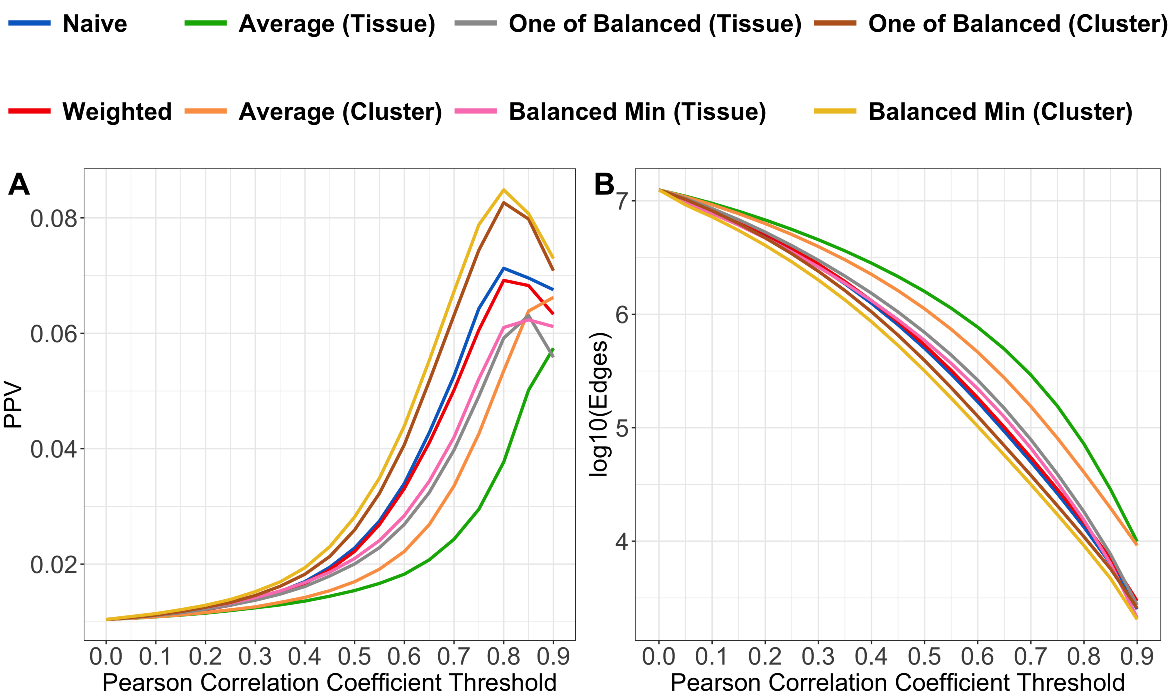


Figure 3.7: Results of pathways based evaluation of different system-level coexpression calculation strategies when presented with gene expression data without further transformations aside from routine normalization. A: PPV across different PCC thresholds. B: number of coexpression edges in a logarithmic scale returned by each method at different thresholds. Balanced method refers to the BRS algorithm.

Results show that the best overall area under the PPV curve was achieved by the BRS method working with a cluster-based partition of samples. Comparing BRS cluster against BRS tissue is particularly interesting as both methods employ a conservative strategy by taking the minimum of 20 realizations for coexpression calculation. Despite this, there is a notable difference in the PPV of both methods. Initially, it seems that differences could be driven by the number of samples considered for coexpression calculation. BRS tissue considers a total of $47 * 56 = 2,912$ samples (the number of tissues with more than 50 samples times two-thirds of the smallest tissue) and BRS cluster considers only $34 * 59 = 2,006$ (the number of clusters times two-thirds of the smallest cluster). However, the PPVs of BRS tissue are worse than the naive or weighted method which considers all of the samples for calculation. The average coexpression method also performs worse when working with the tissue partition as opposed to the clustering partition.

Table 3.2: Overall Predictive Positive Value results for a biological pathways-based evaluation of system-level coexpression strategies.

Method	Area under PPV curve
Naive	0.02739
Weighted	0.02655
Average Tissue	0.01780
Average Cluster	0.02144
One Realization of BRS Tissue	0.02337
Min BRS Tissue	0.02428
One Realization of BRS Cluster	0.03105
Min BRS Cluster	0.03256

There is evidence that suggests that the cluster partition may be making a difference independently from the number of samples used for analysis. From the results shown in the last section, it can be said that some samples that are in different tissues as defined by GTEx have high redundancy between them. Although not rigorously proven, high redundancy on input datasets for coexpression has been empirically linked to worse performance in this kind of analysis based on biological pathways [60]. What is observed in this particular analysis fits the same pattern. Clustering of the input samples before coexpression analysis has been done before in the literature [122], resulting in improvements in downstream coexpression analyses.

Other remarks of this analysis include the number of found coexpressions at different thresholds shown in Figure 3.7 panel B. BRS cluster seems to be the most conservative method, predicting the least amount of coexpressed genes in general. For example, at a PCC threshold of 0.4, BRS predicts about 10^6 coexpressions taking the minimum across realization and a bit more looking at just one realization of the 20. The next method with the least predicted coexpressions is the naive method with about $10^{6.25}$ coexpressions. The amount of predicted coexpressions is clearly related to the PPV of each method as more conservative methods have fewer chances of calling FPs. This kind of relationship is also observed in other studies of this spirit in literature [99].

3.3.2 Z-score normalized and Principal Component regressed expression

The performance obtained for different gene expression transformations additional to routine normalization was also tested in the context of the biological pathways-based coexpression performance analysis. For this, the same experimental setup described in the previous section was used with the only difference being that this time only the BRS strategy working with a cluster-based partition of samples was evaluated. Only this algorithm is tested as it seemed to be the best performing method on routine normalized data without further transforms (referred to here as ”untransformed data” for practical purposes).

The gene expression transformations tested consist of a z-score normalization performed in a cluster-wise fashion (i.e. the transform is applied to the gene vectors of each cluster separately) and a cluster-wise PCA regression (2.7). The regression is tested by choosing the top PCs dynamically with a permutation-based approach and at different cumulative variance thresholds (70%, 80%, and 90%).

Figure 3.8 shows the results of the analysis. Transforming the input gene expression data before coexpression calculations seems to improve the overall PPVs obtained compared to not performing any transform. This is observed especially for the z-score transform which seems to improve the PPVs dramatically. However, when examining the number of predicted coexpressions, this transform also finds a notably smaller number of coexpressions compared to the numbers observed when using untransformed data or PCA regression transformed data. Regarding the PCA regression transform, it also seems to improve PPV values over the ones observed when not transforming, but this improvement is not observed for PCC values in which it is harder to decide if a coexpression should be considered as true or not (less than 0.7 PCC for example).

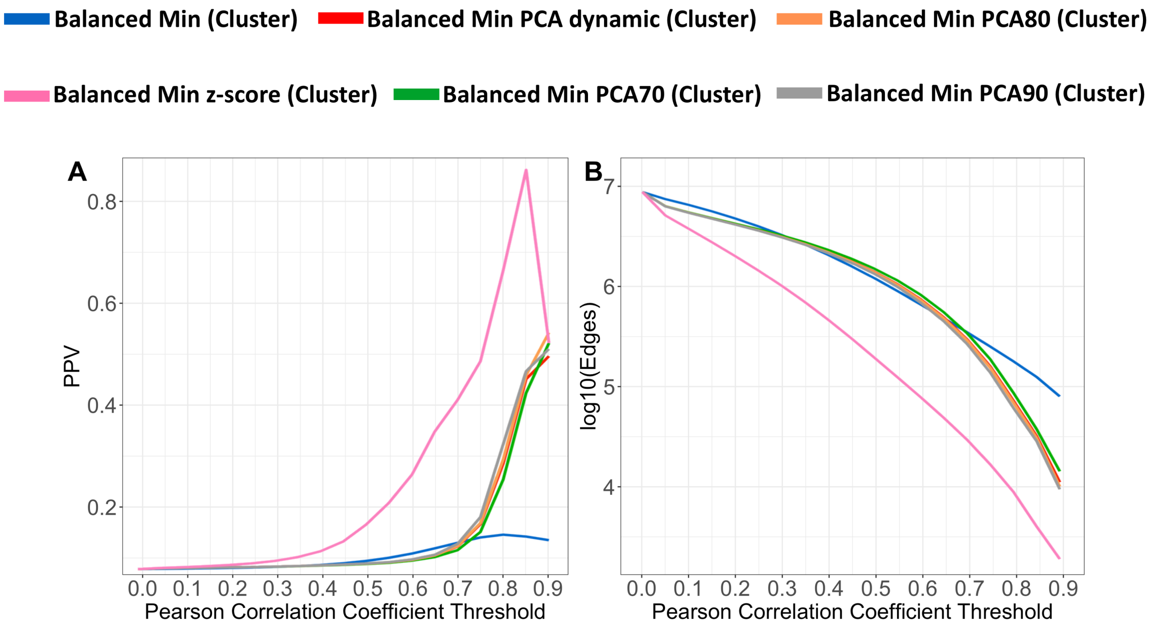


Figure 3.8: Biological pathways-based performance of the BRS algorithm working with additional expression transformations. Results for the untransformed data are also included for comparison and are now shown in blue. A: PPV across different PCC thresholds. B: number of coexpression edges in a logarithmic scale returned by each method at different thresholds. Balanced method refers to the BRS algorithm.

Additionally, the effect of the tested transformations on the gene expression data was characterized via HC in a representative sample of the expression matrix. 1000 random genes and 300 random samples total were taken from different random clusters as seen in Figure 3.9. Notably, sample clusters found using the untransformed data are not respected by the transforms in spite of these being applied to each cluster separately. HC dendrograms were recomputed for each transform in Figure 3.9. It was expected that the expression data would change scale due to the transforms, but it was also theorized that a similar clustering to that observed in the untransformed data would be obtained. This last assumption turned out to be wrong as the transforms clearly modify gene expression to the point were previously defined clusters become unmeaningful.

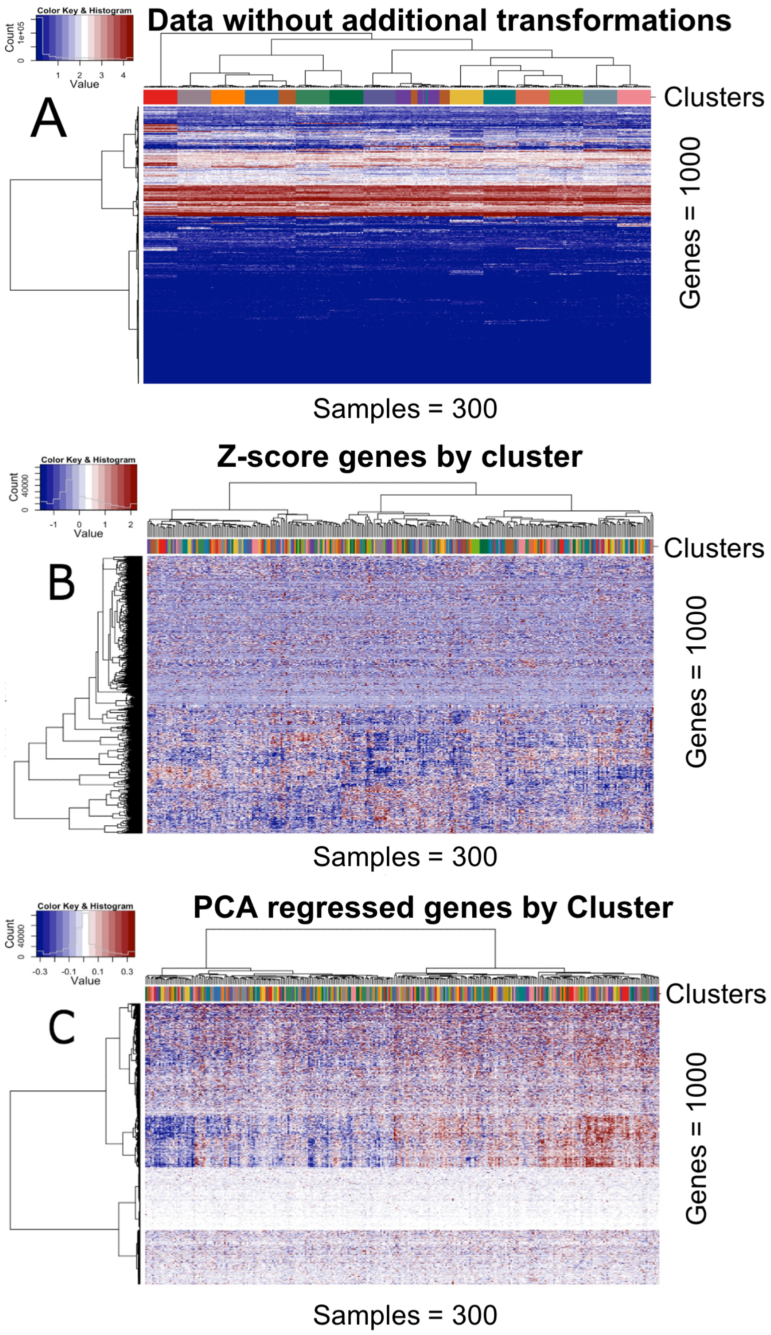


Figure 3.9: Effect of z-score and PCA regression transformations on gene expression data. A: untransformed data (only routine normalized). As expected, HC clusters agree well with the Louvain-Jaccard clusters found before (3.2). Clusters represented in the random sample taken for this visualization are showns in the heatmap's column colorbar. B: HC of cluster-wise z-score transformed gene expression. C: HC of cluster-wise PCA regressed gene expression data. In this case top PCs are selected by a 80% cumulative variance criterium.

Chapter 4

Estimation of system-level coexpression

Results covered so far in Chapter 3 include the estimation of gene expression states in the input data, evaluations of the discovery of coexpressions by different system-level calculation strategies, and the effect of some data transformations on the gene expression. After reviewing them, it was decided that the first version of the computational estimation of system-level human coexpression proposed in this work would be done using the BRS procedure based on the clustering partition without further transformations to the gene expression apart from routine normalization. The reasoning behind not using some of the studied additional gene expression transforms at least for now is that gene expression states seemed to become less well defined upon applying them. They caused substantial changes to the gene expression that did not maintain the original clustering of the data. For the specific case of the cluster-wise z-score transform, applying this technique to the input gene expression also seemed to greatly reduce the amount of found coexpressions as demonstrated in the last section.

4.1 Implementation

BRS was applied using the same parameters as in the preliminary tests which consist of a sampling size of $k = 59$ samples per cluster and number of realizations $s = 20$ (2.6.4). This is done to calculate system-level coexpression for a total of 33,447 human genes using the Chi-square, PCC, and SRCC metrics (2.5). Each realization of BRS considered a balanced dataset consisting of 33,447 genes and 2,006 samples resulting from taking 59 samples from each of the 34 clusters representing distinct gene expression states in the GTEx data. Just as in the preliminary tests, the final robust estimation of coexpression for any pair of genes is computed by taking the minimum measurement observed for the pair in question across the 20 realizations. In order to optimize computation times, C++ functions to calculate the coexpression metrics were implemented and executed from R using the *Rcpp* package [39, 123]. To optimize disk space and memory usage, integer versions of the metrics were handled instead of their floating-point representations by keeping enough digits to specify the value of a measure to 2 decimal points of precision.

In the case of the Chi-square calculations, some special cases were programmed to handle situations in which the expected counts of a given cell in the test were extremely low. The latter can lead to misleadingly large Chi-square statistics. If any expected value was lesser than $\frac{1}{n+1}$ where $n = 2006$ (number of samples used), it was automatically set to the aforementioned ratio. Another precaution consisted of checking if the Chi-square contribution of a given cell was greater than 5 in combination with an expected value less than 0.1. When this was true, an expected value of exactly 0.1 would be considered and the Chi-square contribution of the cell was recalculated. These special cases do not affect the obtained distributions of Chi-squared values significantly, but they do help in getting rid of some inconsistencies that could arise during a minority of calculations.

4.2 Distributions of observed and expected Chi-square statistics

As a result of the coexpression estimation, 559,334,181 measurements for each of the three coexpression metrics were obtained. Note that the actual process of calculation involves computing this amount of coexpression measurements times 20 for each metric as specified by the BRS strategy (2.6.4).

The distribution arising from the robust BRS calculation can be seen in Figure 4.1 (blue line). Some other relevant distributions are also shown for comparative purposes. The gray curve corresponds to the resulting distribution after carrying out a gene-wise permutation of discretized gene expression values in 1 of the 20 expression matrices used in BRS. Only values within the same gene vector are shuffled and coexpression is recalculated for this permuted realization. The gray curve almost perfectly fits the distribution in black which lies behind it. The black curve corresponds to the random Chi-square distribution that agrees with the parameters of the coexpression estimation (4 DFs). The fact that the null distribution of the coexpression estimation fits the random reference distribution suggests that strong Chi-square values observed in one of 20 realizations of the estimation (red line) and the robust minimum estimation are not due to chance.

The FDR line (in green) in Figure 4.1 was estimated by dividing the area under the curve of the null distribution (gray) over the area of the minimum robust estimation (blue) distribution. Areas are bounded by different increasing Chi-squared cutoff points and the highest value considered in the histogram. When considering all Chi-squared values for the computation, the FDR is large because the null distribution is a lot denser for non-significant statistics close to 0. When stricter cutoffs are set (i.e. by moving them to the right of the x-axis), the distribution of observed statistics quickly becomes denser and the FDR drops. If one was to consider all coexpressions above a certain Chi-squared value as significant, then about $FDR * 100$ percent of those coexpressions would be mistaken rejections of the null hypothesis (type 1 error). Figure 4.2 shows the discussed distributions in terms of cumulative counts.

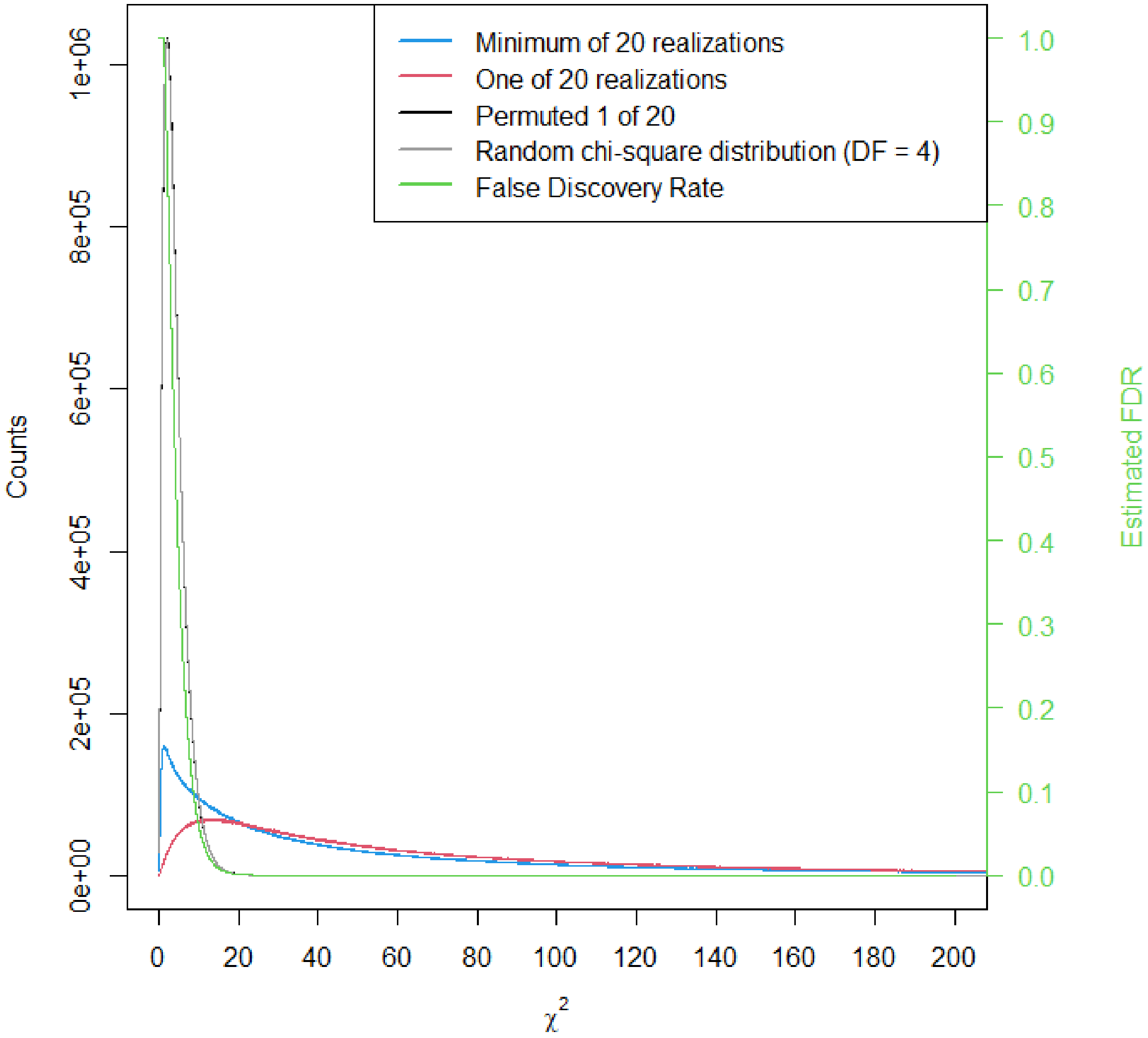


Figure 4.1: Chi-squared distributions related to the system-level coexpression calculation proposed in this work. The histogram is represented in a lineplot style to make the visualization of multiple distributions clearer. Blue line: distribution of robust minimum chi-squared statistics across 20 realizations. Red line: distribution of one of 20 realizations (randomly selected) computed as part of the BRS process. Gray line: null distribution estimated by permuting the labels of the discretized gene expression vectors prior to calculating 1 realization of BRS (same realization index used for red line). Black line: random Chi-square distribution with a parameter of 4 DFs (mostly concealed behind the gray line due to a great fit).

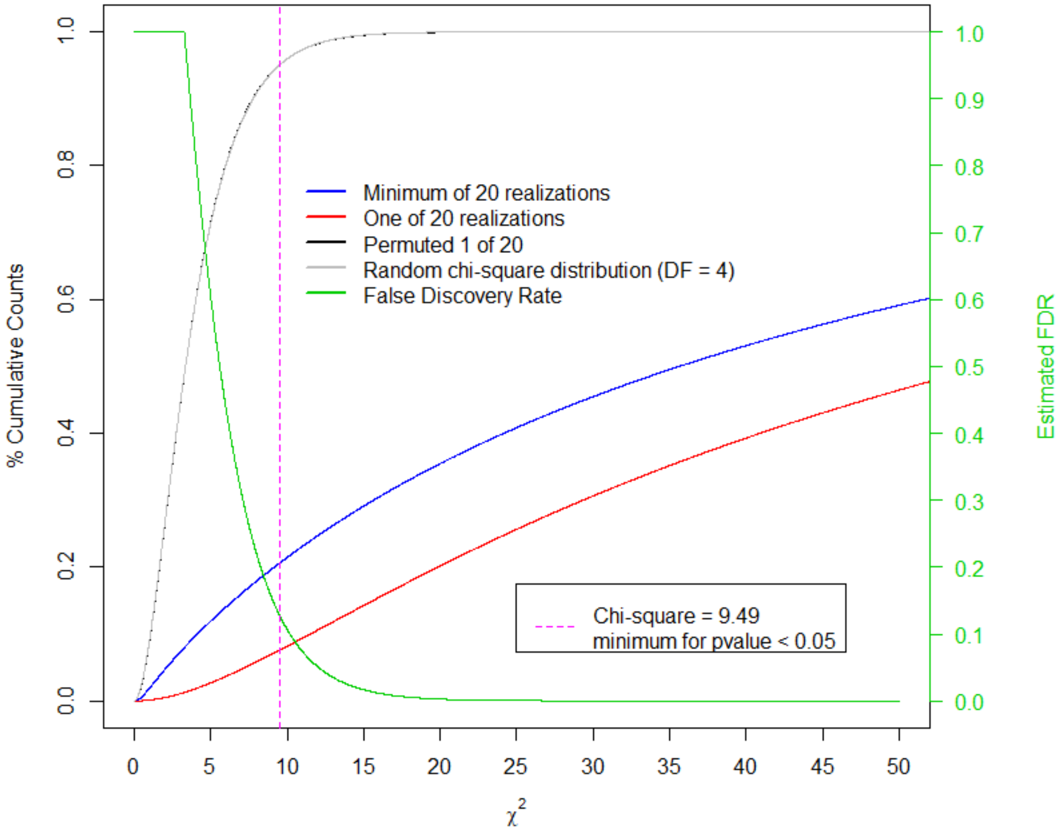


Figure 4.2: Zoomed-in cumulative distributions version of Figure 4.1. This figure puts in context just how dense the observed distributions are at stronger Chi-square values when compared to the reference random distributions or the one arising from the permuted input data. The dotted line marks the minimum (approximate) Chi-square value needed to obtain a p-value lesser than 0.05 if the random distribution at 4 DFs is used as a reference. Using such criteria is unrealistic for the distributions obtained as even in the case of the minimum robust estimation, about 80% of the total calculated coexpressions would be significant. Statistical significance was not taken as a good way of determining if two genes were truly coexpressed in this work. A more rank-based approach of interpreting the results is used. The issue is further discussed in Chapter 5.

4.3 Comprehensive coexpression visualizations

In this section, the most basic unit of obtained results is presented: the coexpression scatter plot. Thanks to the implementation of the Chi-square test in the coexpression calculations, it is possible to add a lot of useful information to such plot. This helps evaluating a coexpression at the individual level and extract meaningful intuitions from it. In this work, this plot is referred to as the contingency scatter plot.

Below is a detailed explanation of the components of the contingency scatter plot visualization for gene coexpression shown in Figure 4.3:

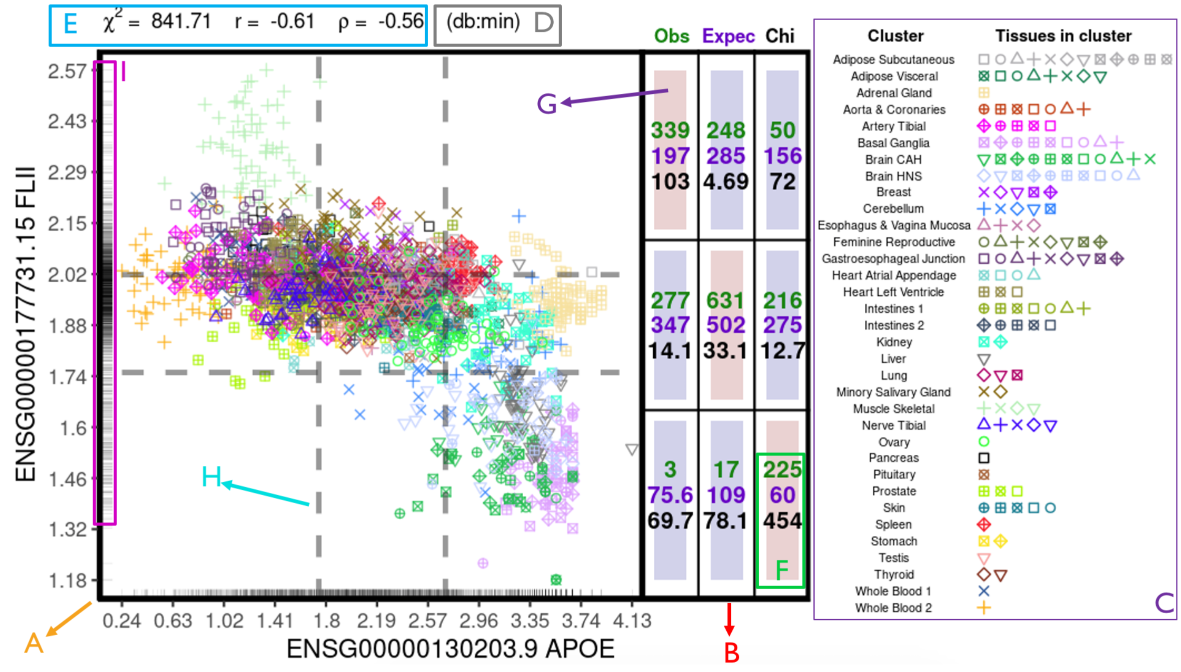


Figure 4.3: Contingency scatter plot for single coexpression visualization. The example presented is chosen as it appears to feature a relationship which can be decently modeled by both a linear and non-linear function. It exemplifies the variety and complexity that exist among gene expression data in the context of coexpression. APOE: apolipoprotein E gene, FLII: actin remodeling protein FLII gene. This coexpression is interesting from the literature point of view as APOE has been implicated in cytoskeleton biological pathways [124]. A: scatter box area with samples plotted as the points. B: contingency table of Chi-squared test. C: legend mapping of gene expression states to plotted samples. D: calculation source of the plotted data. E: coexpression measurements by 3 different metrics (Chi-square is χ^2, PCC r and SRCC ρ). F: cells of contingency table contain color-coded observed and expected counts in addition to their individual Chi-square values. G: color background of table cells indicate the results of a post-hoc asymptotic test for the significance of the cell (red denotes significantly more observed counts than expected and blue the opposite). H (dotted lines): gene expression values that bin each gene into 3 discretized categories. I: rug plot.

A *Scatter box*: plotting area for the points which represent a sample each. The shapes and labels of points indicate the cluster of origin of the sample and the original GTEx tissue label that is contained within the cluster. Expression measurements for both genes are uniformized (min-max scaling) to the range $[0, 1]$ so that it is easy to obtain a more visually comparable square plot. The latter does not affect the results of any of the coexpression measures. Moreover, the labels of values shown along the x and y axes still correspond to the logarithm base 10 of the TPMs for each gene. Since genes were not further transformed apart from routine normalization during calculations, these values may be easily interpreted.

B *Contingency table*: same table as used during the actual Chi-square test for the two genes in question. The table is accommodated to match the gene expression levels shown in the scatter box. The down-right corner considers the samples that are low-expressed for both genes. Expression of the genes in the x and y axes increases as one moves right or up respectively.

C *Gene expression states legend*: maps the samples plotted in the scatter box through colors to one of the gene expression states characterized with clustering as (3.2). Additionally, a shape mapping was also added where different shapes signify different tissues in the original GTEx sample partition that are captured into one of the clusters.

D *Calculation source*: as a result of using the BRS algorithm for the robust calculation of system-level coexpression, it is possible to plot any of the 20 realizations that exist for any individual coexpression. Usually, it is of interest plotting the realization giving the minimum coexpression measurement (denoted as *db:min*). This part of the plot clearly states this to avoid confusion. Naturally, the minimum realization is determined by a particular measurement out of the 3 available. Throughout this document, the Chi-square measurement is used as a reference to plot the source data in the case of the minimum. The other two metrics will correspond to the realization plotted for consistency, but may not necessarily be the global minimum across all realizations for those specific metrics.

E *Coexpression measurements*: these are the results of the computation of the 3 different coexpression measurements considered in the estimation. Chi-square is denoted χ^2, the PCC (pearson) as r and the SRCC (spearman) as ρ.

F *Contents of contingency table cells*: each one of these cells contains the observed counts or number of samples falling in the corresponding 2-dimensional bins of the scatter box. They also have the expected counts if the samples were distributed at random according to the row and column marginals. The Chi-square contribution to the overall statistic of the test is also shown. A color mapping to each one of these numbers can be found at the top of the contingency table.

G *Contingency table cells background*: this is colored according to the results of a post-hoc asymptotic test that determines if there is significantly more (red) or less (blue) observed than expected counts in the cell. The significance of the test is considered at a standard value of 0.5. When the test is not significant, the background will appear without color.

H *Values binning gene expression*: these dotted lines represent the cutoffs calculated via the k-means algorithm that discretize the gene expression measurements of each gene (2.5.1.1). The resulting 2-dimensional binning of samples enters the Chi-square calculation as the observed counts.

I *Rug plot*: one appears for each gene. They summarize the density of the distribution of the continuous gene expression data. Rug plots consist of 1-dimensional visualizations that simply draw a line at the values corresponding with points in the scatter box. Lines become crowded in areas with greater densities of points.

The contingency scatter plots allow for detecting exactly what combinations of gene expression categories are responsible for the observed overall Chi-squared statistic. For example, in Figure 4.3 it is clear that in terms of the Chi-square, the coexpression is actively happening when APOE and FLII vary inversely in their expressions. If one gene is low-expressed, the other one is high-expressed and vice versa. This is easily verifiable by the fact that it is exactly the cells in the diagonal of the contingency table that significantly describe this behavior. Note that this specific pattern suggests what is precisely described by a linear relationship. This is why the PCC is also at the higher end of its scale in this example.

Contingency scatter plots also aid at detecting which gene expression states contribute more to the overall coexpression between genes. In Figure 4.3 it is interesting that the cell with the most contribution to the overall Chi-square contains samples exclusively from brain gene expression states (cell is labeled F). This is another example of the additional information that the contingency scatter plot gives which can be very insightful regarding what is actually happening and where. APOE is a widely researched gene in the nervous system because of its connections to neurodegenerative disorders. The particular coexpression shown in Figure 4.3 has a high chance of being biologically meaningful as APOE has been shown to downregulate several genes associated with actin (FLII remodels this protein), myosin, and microtubules in the nervous system [124]. The association detected in the plot is strong and was observed primarily in brain gene expression states.

4.4 Chi-square and correlation coefficients comparison

In this section, some examples comparing coexpressions obtained with different metrics during the system-level coexpression estimation are shown. This also serves as a way to validate coexpression estimation by assessing the agreement that different metrics have at the moment of detecting coexpressed gene pairs. For all these comparisons, the minimum across 20 realizations of each metric is used.

4.4.1 Patterns of coexpressed gene pairs

4.4.1.1 Chi-square and Pearson Correlation Coefficient

Three genes were carefully selected as examples for this comparison. They represent cases whose coexpression gene lists ordered by estimation strengths have distinct levels of overlap between the Chi-square and PCC. The latter will be further demonstrated later in another analysis (4.4.2).

Figure 4.4 clearly shows that Chi-square and PCC are correlated coexpression metrics. The parabola pattern observed is stereotypical when comparing different association measures with the PCC [23]. Here it is confirmed that such a pattern is in fact observed for the Chi-square in the context of coexpression. The parabola shape is given by a high density of points with both a low Chi-square and a close to 0 PCC (zone A or parabola vertex), as well as by points exhibiting a positive or negative correlation which tend to correspond to high Chi-square values (point patterns or parabola arms heading towards zones F and E).

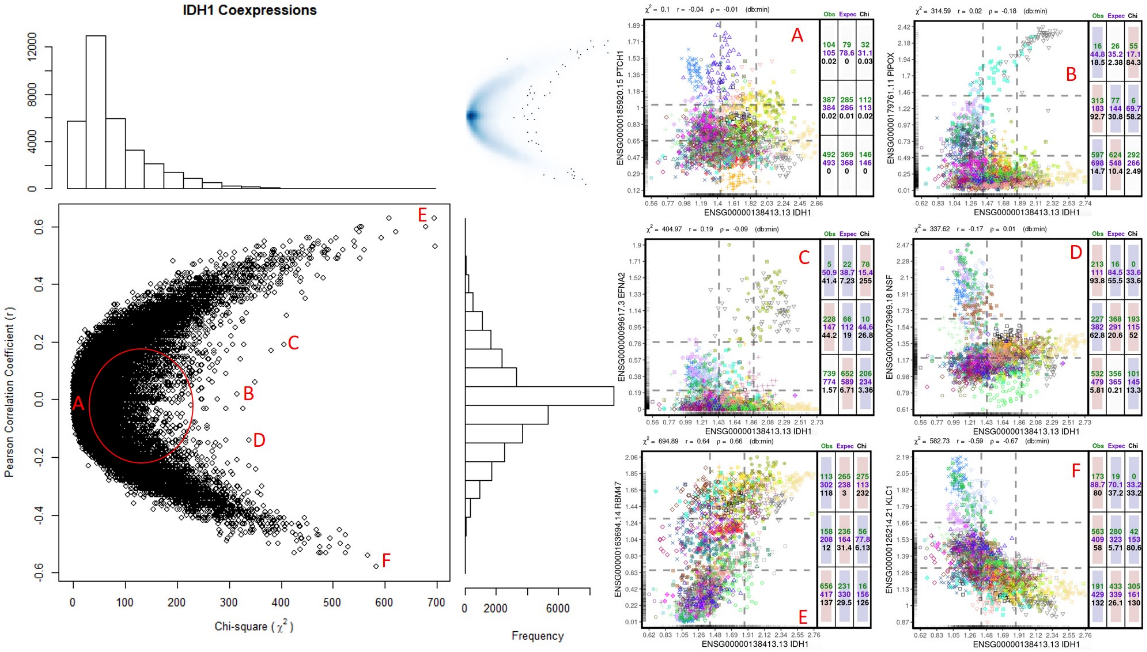


Figure 4.4: Comparison of isocitrate dehydrogenase 1 gene (IDH1) calculated coexpressions by Chi-square and PCC. The main scatter plot shows a gene pair (composed of IDH1 and every other possible gene in the estimation) at each point along with the distributions of their respective Chi-square and PCC. A point density plot is also shown in the top-right corner of the main scatter plot. The red circle/ellipse covers values of around 100-500 Chi-square and -0.2 to 0.2 PCC. Its purpose is to indicate the denser zone of disagreement between the two metrics. Smaller labeled scatter plots correspond to an example coexpression found at the corresponding labeled zone in the main scatter plot. To consult legend mappings of the samples represented in the labeled scatter plots see Figure 4.3.

Zone A of Figure 4.4 corresponds to calculations that indicate no effective coexpression from both the Chi-square and the PCC point of view as seen in the example shown between IDH1 and the protein patched homolog 1 gene (PTCH1). Zones B-D are of particular interest as they represent gene pairs with probable non-linear coexpression relationships among them as they have a high Chi-square measurement, but low PCC. Zones E and F are where strong coexpressions that can be modeled with linear functions between the gene pairs lie as they render both a strong Chi-squared and PCC. The Chi-squared/PCC patterns observed for this gene should be representative of what happens for most genes (4.4.2).

The next example is depicted in Figure 4.5. This time around, BRCA2 should be representative of extreme cases in which Chi-squared and PCC fit the stereotypical parabola trend better than average (4.4.2). Note the augmented density and length of the parabola arms in the BRCA2 example as compared to the IDH1 example in Figure 4.4. This implies more linear relationships between BRCA2 and its gene pairs. Density in the zone where combinations of high Chi-square values and close to 0 PCC appear (implying non-linear relationships) is also decreased when comparing it to the corresponding zone for IDH1.

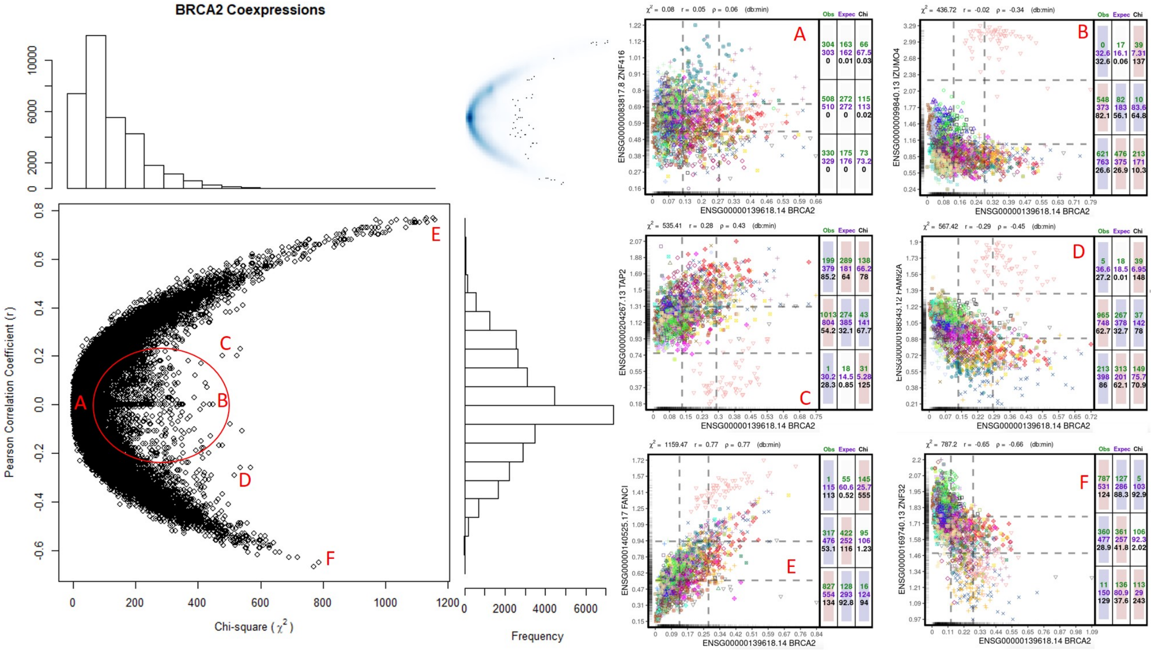


Figure 4.5: Comparison of DNA repair associated BRCA 2 gene (BRCA2) calculated coexpressions by Chi-square and PCC. The main scatter plot shows a gene pair (composed of BRCA2 and every other possible gene in the estimation) at each point along with the distributions of their respective Chi-square and PCC. A point density plot is also shown in the top-right corner of the main scatter plot. The red circle/ellipse covers values of around 100-500 Chi-square and -0.2 to 0.2 PCC. Its purpose is to indicate the denser zone of disagreement between the two metrics. Smaller labeled scatter plots correspond to an example coexpression found at the corresponding labeled zone in the main scatter plot. To consult legend mappings of the samples represented in the labeled scatter plots see Figure 4.3.

The last example for comparing Chi-squared and PCC is depicted in Figure 4.6. From observations derived from this work, ANG should be representative of extreme cases in which Chi-squared and the PCC fit the stereotypical parabola trend worse than average (4.4.2). Note the greatly diminished length of the parabola arms in the ANG example as compared to IDH1 in Figure 4.4. This implies less linear relationships between ANG and its gene pairs. Density in the zone where combinations of high Chi-square values and close to 0 PCCs lie is augmented. The parabola shape seems to be infiltrated by a series of points around zone C which do not vary much in terms of PCC but that vary greatly in Chi-square. It is worth mentioning that the example coexpressions are shown for ANG evidence a strong component of a particular cluster (Liver) driving high Chi-squared values in zones of high disagreement between Chi-square and PCC.

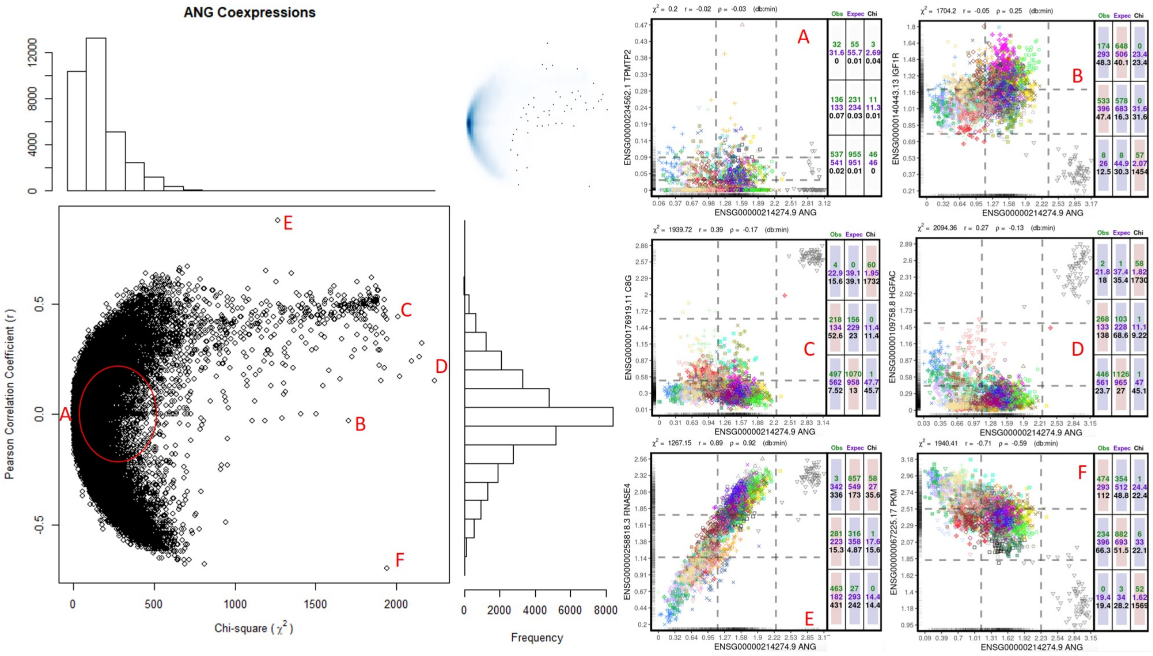


Figure 4.6: Comparison of angiogenin gene (ANG) calculated coexpressions by Chi-square and PCC. The main scatter plot shows a gene pair (composed of ANG and every other possible gene in the estimation) at each point along with the distributions of their respective Chi-square and PCC. A point density plot is also shown in the top-right corner of the main scatter plot. The red circle/ellipse covers values of around 100-500 Chi-square and -0.2 to 0.2 PCC. Its purpose is to indicate the denser zone of disagreement between the two metrics. Smaller labeled scatter plots correspond to an example coexpression found at the corresponding labeled zone in the main scatter plot. To consult legend mappings of the samples represented in the labeled scatter plots see Figure 4.3.

4.4.1.2 Chi-square and Spearman Rank Correlation Coefficient

Similar to what was done in the case of the Chi-square statistic and the PCC, here are presented carefully chosen examples that depict distinct levels of similarities and differences between the Chi-square statistic and the SRCC.

Figure 4.7 resembles what was observed between the Chi-square and PCC metrics in the average case. The parabola shape is maintained in this example with the SRCC as well. For the case of the SRCC, this trend can not be considered as representative of the average case (4.4.2), but it can be seen as an example in which there is a moderately good agreement between these metrics.

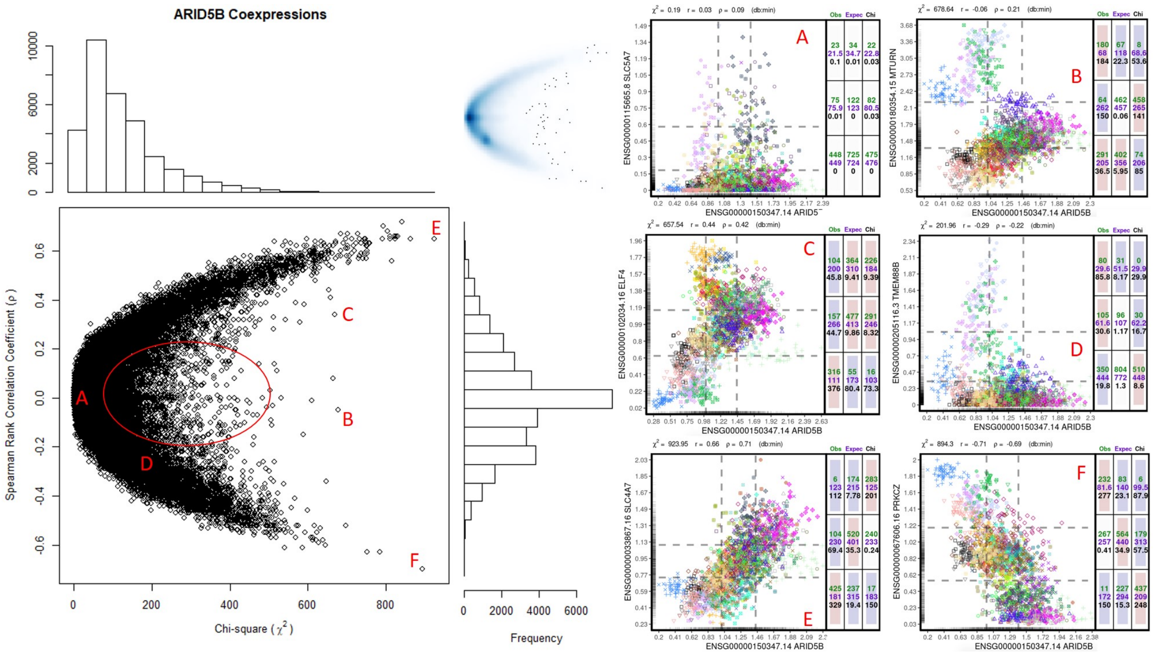


Figure 4.7: Comparison of AT-rich interactive domain-containing protein 5B gene (ARID5B) calculated coexpressions by Chi-square and SRCC. The main scatter plot shows a gene pair (composed of ARID5B and every other possible gene in the estimation) at each point along with the distributions of their respective statistics. A point density plot is also shown in the top-right corner of the main scatter plot. The red circle/ellipse covers values of around 100-500 Chi-square and -0.2 to 0.2 PCC. Its purpose is to indicate the denser zone of disagreement between the two metrics. Smaller labeled scatter plots correspond to an example coexpression found at the corresponding labeled zone in the main scatter plot. To consult legend mappings of the samples represented in the labeled scatter plots see Figure 4.3.

Figure 4.8 illustrates a case in which there is a high agreement between the Chi-square and SRCC measures. The parabola trend is present, but it has widened, causing the density of points in the disagreement zone (red ellipse) to be lower when compared to the ARID5B case. Labeled coexpressions are very interesting in this example as some of them correspond to immunoglobulin related genes which hold both linear (panel E) and non-linear (panel D) associations with IGHA1.

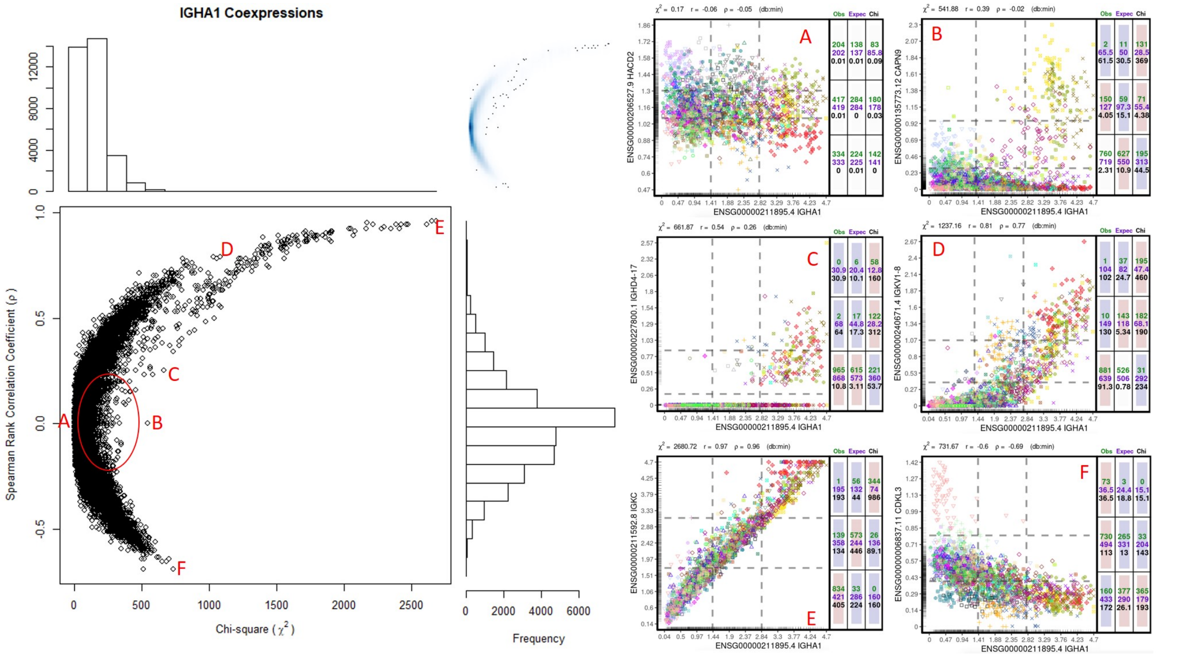


Figure 4.8: immunoglobulin heavy constant alpha 1 gene (IGHA1) calculated coexpressions by Chi-square and SRCC. The main scatter plot shows a gene pair (composed of IGHA1 and every other possible gene in the estimation) at each point along with the distributions of their respective statistics. A point density plot is also shown in the top-right corner of the main scatter plot. The red circle/ellipse covers values of around 100-500 Chi-square and -0.2 to 0.2 PCC. Its purpose is to indicate the denser zone of disagreement between the two metrics. Smaller labeled scatter plots correspond to an example coexpression found at the corresponding labeled zone in the main scatter plot. To consult legend mappings of the samples represented in the labeled scatter plots see Figure 4.3.

Figure 4.9 depicts a case with a very low correlation between the Chi-square statistic and the SRCC. For the first time, the stereotypical parabola shape becomes almost nonexistent. It would seem as if many coexpressions stayed invariant for Chi-square while varying greatly for the SRCC. Upon inspecting the labeled coexpressions, it becomes clear that the BSND gene, very much like the case of ANG in Figure 4.6, exhibits cluster specificity (for Kidney in this case). It seems like in these kinds of coexpressions driven by tissue or cluster-specific genes, both the PCC and the SRCC differ more from Chi-square.

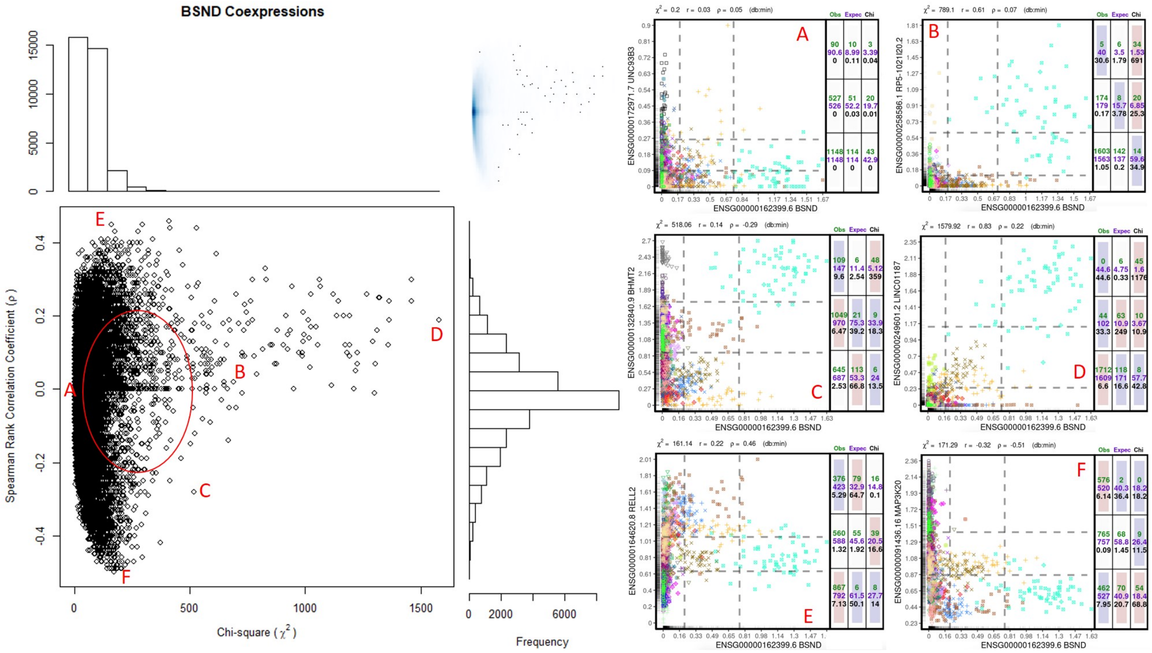


Figure 4.9: Comparison of Barttin gene (BSND) calculated coexpressions by Chi-square and SRCC. The main scatter plot shows a gene pair (composed of BSND and every other possible gene in the estimation) at each point along with the distributions of their respective statistics. A point density plot is also shown in the top-right corner of the main scatter plot. The red circle/ellipse covers values of around 100-500 Chi-square and -0.2 to 0.2 PCC. Its purpose is to indicate the denser zone of disagreement between the two metrics. Smaller labeled scatter plots correspond to an example coexpression found at the corresponding labeled zone in the main scatter plot. To consult legend mappings of the samples represented in the labeled scatter plots see Figure 4.3.

4.4.1.3 Pearson Correlation Coefficient and Spearman Rank Correlation Coefficient

It is known that the PCC and SRCC produce similar results when the input data complies with certain distributional properties [125]. However, during the construction of the examples that have been addressed in this section, some situations in which the PCC and SRCC were quite different did occur.

Before taking a look at a case with notable differences, first in Figures 4.10 and 4.11, examples of moderate and high agreement between PCC and SRCC are presented. The metrics are clearly correlated linearly, but can frequently produce differently ordered coexpressed pairs in terms of coexpression strength (4.4.2).

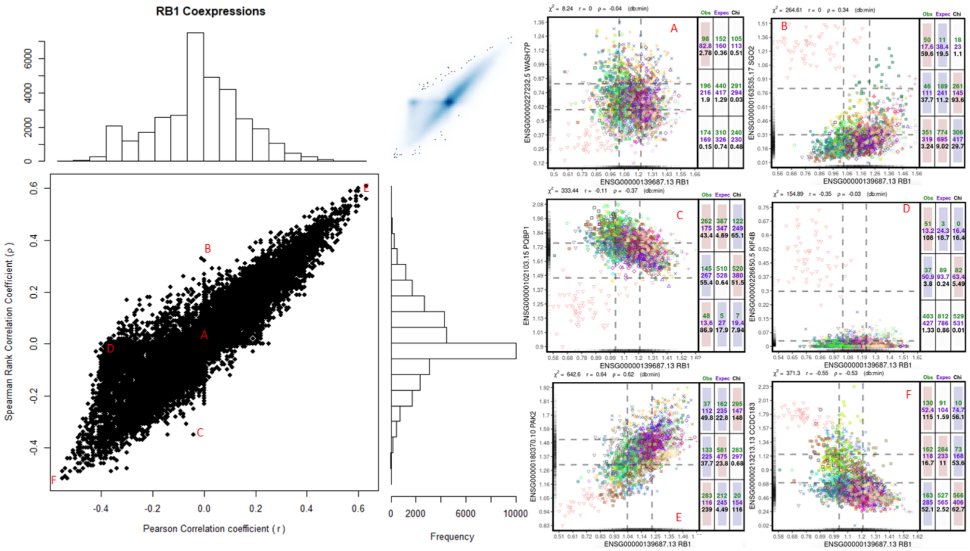


Figure 4.10: Comparison of retinoblastoma transciptional corepresor 1 gene (RB1) calculated coexpressions by PCC and SRCC. The main scatter plot shows a gene pair (composed of RB1 and every other possible gene in the estimation) at each point along with the distributions of their respective statistics. A point density plot is also shown in the top-right corner of the main scatter plot. Smaller labeled scatter plots correspond to an example coexpression found at the corresponding labeled zone in the main scatter plot. To consult legend mappings of the samples represented in the labeled scatter plots see Figure 4.3.

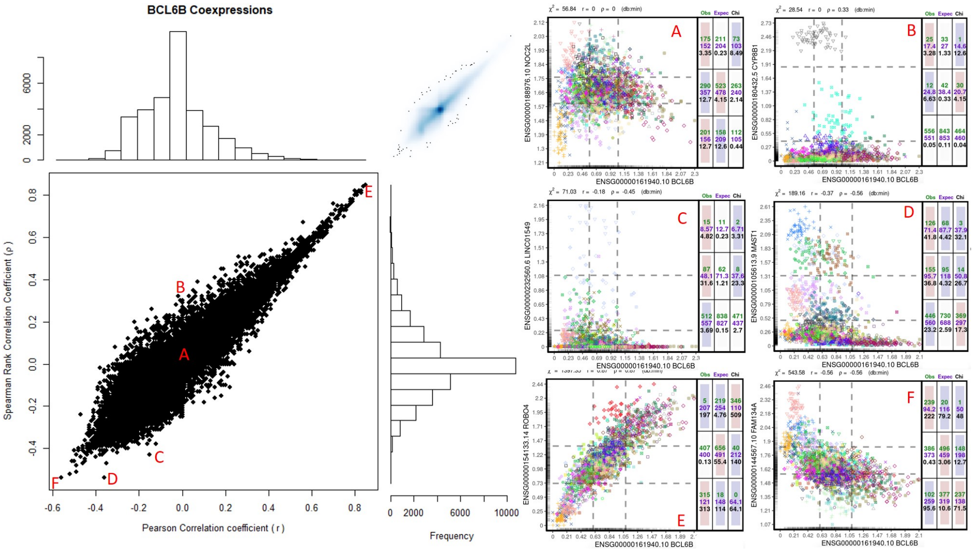


Figure 4.11: Comparison of B-cell lymphoma 6 member B gene (BCL6B) calculated coexpressions by PCC and SRCC. The main scatter plot shows a gene pair (composed of RB1 and every other possible gene in the estimation) at each point along with the distributions of their respective statistics. A point density plot is also shown in the top-right corner of the main scatter plot. Smaller labeled scatter plots correspond to an example coexpression found at the corresponding labeled zone in the main scatter plot. To consult legend mappings of the samples represented in the labeled scatter plots see Figure 4.3.

Figure 4.12 describes a case in which PCC and SRCC produce rather different results. From this example, it seems yet again like the possible culprits of these disagreements between metrics are cluster-specific genes. HIST3N is in fact expressed with selectivity on the Testis cluster. Moreover, with this observation, the involvement of these kinds of cluster-specific genes was spotted for all combinations of metrics at situations where they produced the most different results between them.

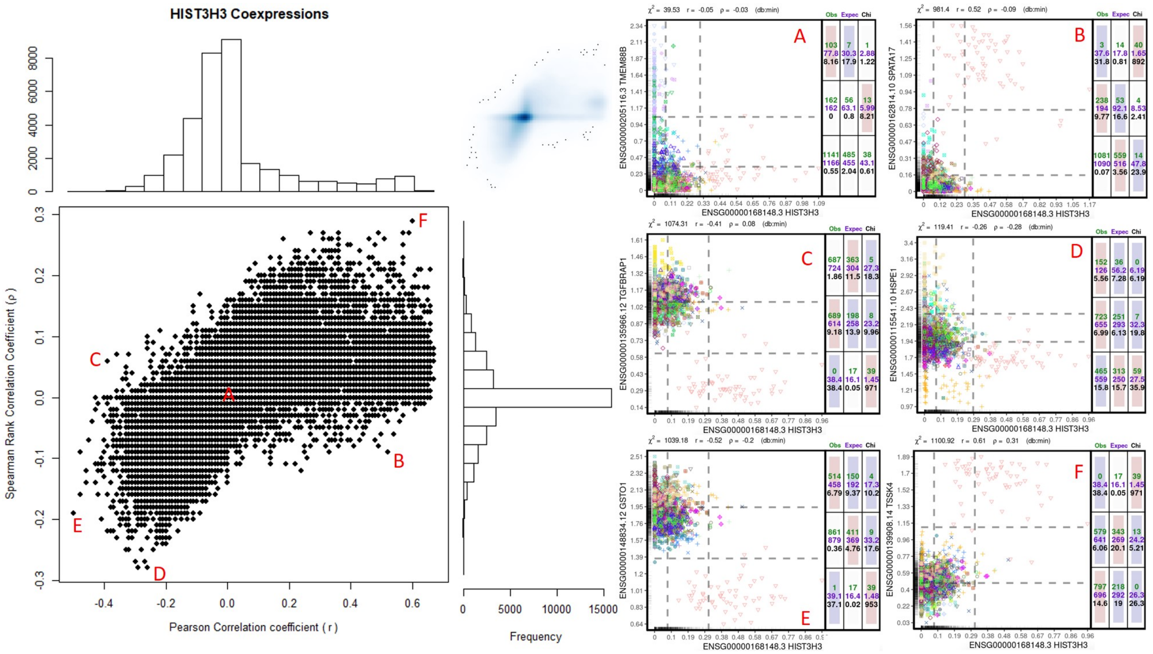


Figure 4.12: Comparison of histone cluster 3 gene (HIST3H3) calculated coexpressions by PCC and SRCC. The main scatter plot shows a gene pair (composed of HIST3H3 and every other possible gene in the estimation) at each point along with the distributions of their respective statistics. Note that the decimal precision used during coexpression estimations makes the points look oddly arranged in the plot, but this does not affect the analysis. A point density plot is also shown in the top-right corner of the main scatter plot. Smaller labeled scatter plots correspond to an example coexpression found at the corresponding labeled zone in the main scatter plot. To consult legend mappings of the samples represented in the labeled scatter plots see Figure 4.3.

4.4.2 Coexpressed gene pairs rank analysis

Here the coexpression measures are compared in terms of the top associations they detect. The top 100 coexpressed genes with each gene in the estimation were retrieved for each metric. The size of the intersection between these sets of 100 genes was computed. Each gene's top coexpressed pairs according to one metric are compared with their similes from the two other measures separately. The significance of the intersection sizes according to the size of the two intersecting sets (always 100) and the universe of genes (33,447) is assessed via a hypergeometric test (2.9.2.1).

Starting with the comparison of Chi-square and PCC, Figure 4.13 shows the distribution of intersection sizes resulting from the aforementioned experiment. In order to obtain a significant result by the hypergeometric test after a Benjamini-Hochberg correction (2.9.2.2), at least 2 genes should appear in both the top 100 of Chi-square and the top 100 of PCC for any given gene. The distribution reveals that these measures agree on 60-70 genes on average, something expected due to the correlation seen between these metrics in the last section. Note that the intersection size of the genes used as examples in the last section is indicated here. They correspond to extreme values in the distribution and one average case.

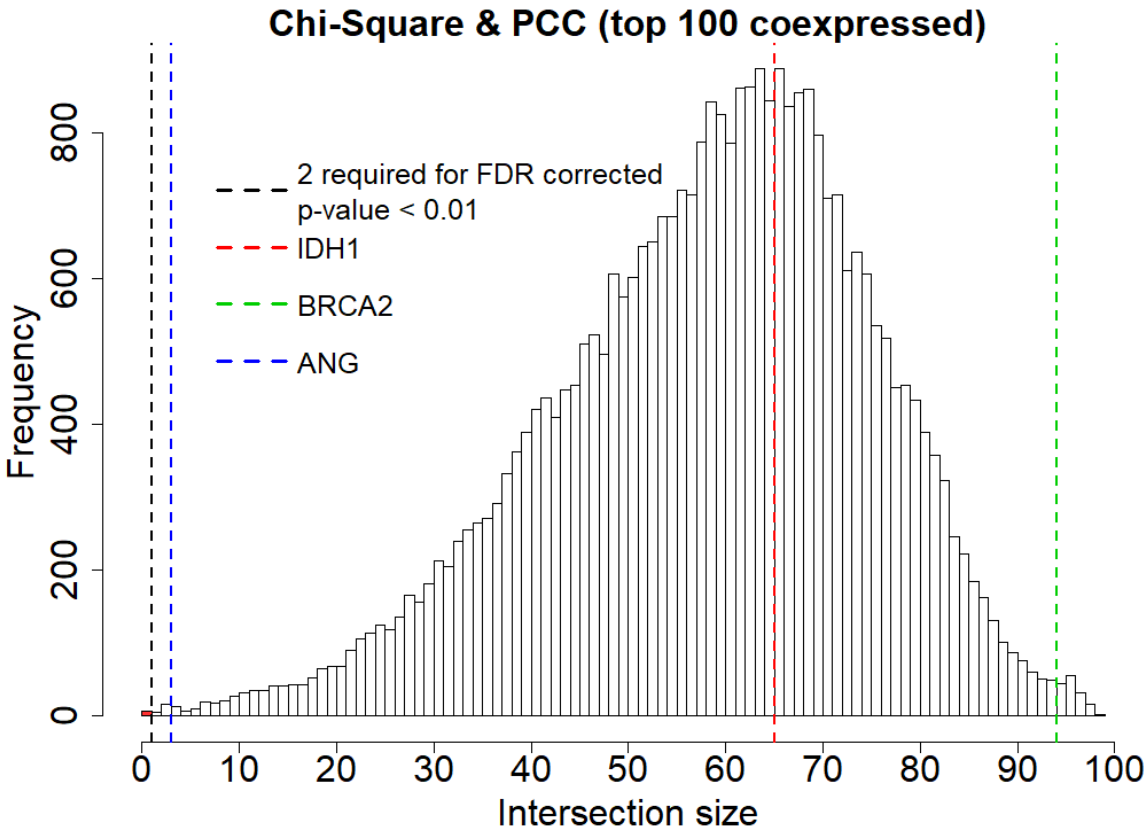


Figure 4.13: Distribution of overlaps between top 100 coexpressed pairs of each gene calculated by Chi-Square and PCC. Lines show where the overlaps of some example genes lie as well as the minimum number of overlaps for the intersections to be significant. The red bar indicates intersection sizes which do not result on a significant hypergeometric test.

Figure 4.14 shows the results for the Chi-square/SRCC case. The distribution is very different from that observed in the Chi-squared/PCC experiment. There are quite a few cases in which the overlap between the metrics is very low and in many even 0. The average number of overlaps, in this case, is not very informative as there could even be a mixture of distributions consisting of a logarithmic curve near zero and another one resembling a Gaussian centered at about 65. Colored lines indicate the intersection size values for the genes used as examples in the last section.

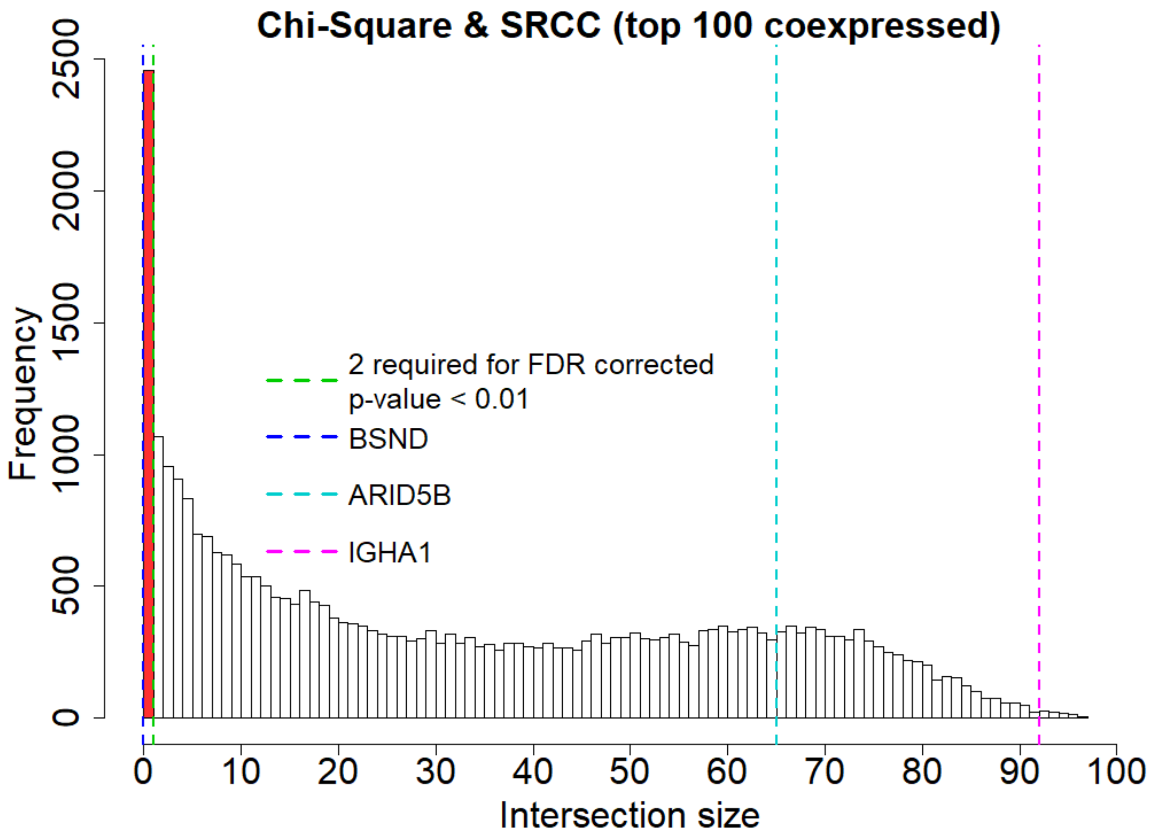


Figure 4.14: Distribution of overlaps between top 100 coexpressed pairs of each gene calculated by Chi-Square and SRCC. Lines show where the overlaps of some example genes lie as well as the minimum number of overlaps for the intersections to be significant. The red bar indicates intersection sizes which do not result on a significant hypergeometric test.

The results for the PCC/SRCC comparison can be seen in Figure 4.15. The obtained distribution is quite similar to that observed for the Chi-square/SRCC case, only with a greater number of significant overlaps. Observations gathered so far indicate that the SRCC is the measure that produces the most different results from all computed measures at least in terms of strongly coexpressed genes.

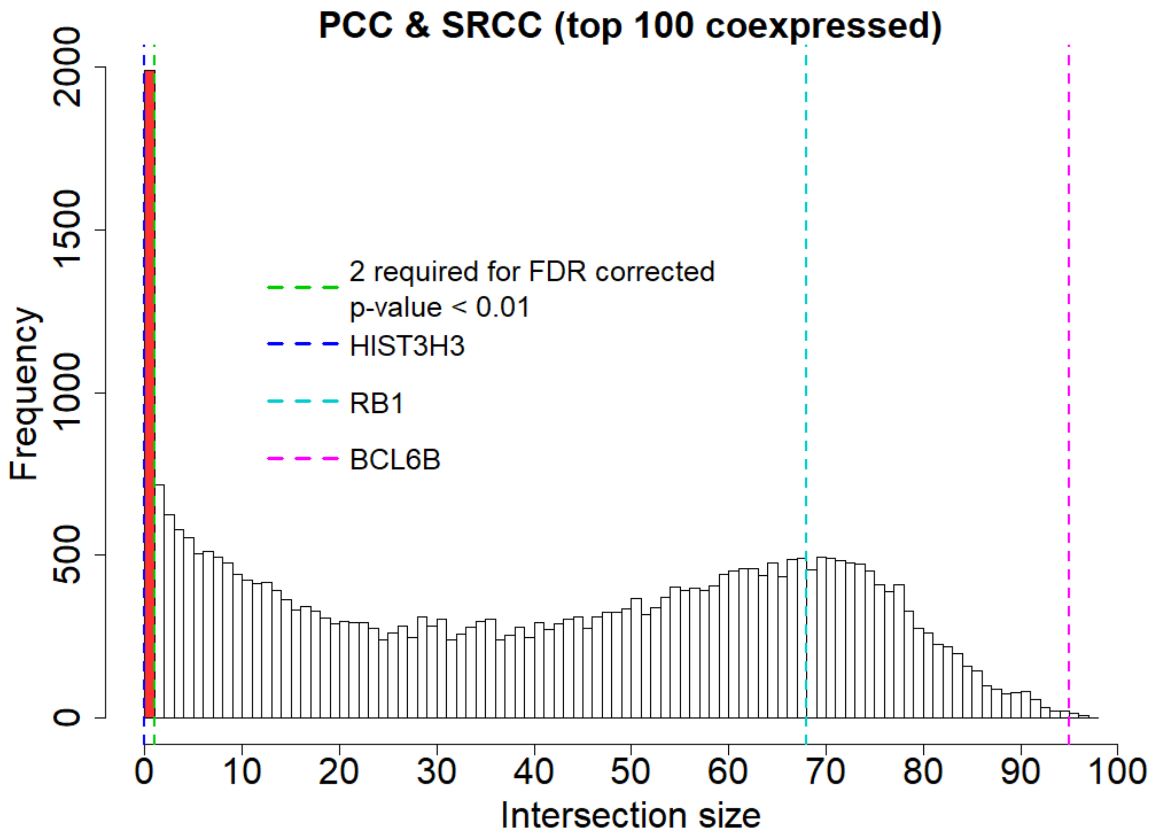


Figure 4.15: Distribution of overlaps between top 100 coexpressed pairs of each gene calculated by PCC and SRCC. Lines show the minimum number of overlaps for the intersections to be significant. The red bar indicates intersection sizes which do not result on a significant hypergeometric test.

4.4.3 Coexpressed gene pairs in literature

Looking at agreements between metrics and examples of coexpressed genes helps to validate the calculations. However, it is also desirable to test the found coexpressions against reported interactions between genes for further validation. If strong associations between genes found via coexpression agree with gene pairs that are already known to be associated, this would support the overall trustworthiness of the coexpression results. It should be possible to demonstrate that strong associations in the calculations point to interesting gene relationships instead of random unrelated gene pairs. To investigate this, the top 50 coexpressed gene pairs of several genes in the calculations were searched in literature papers.

A systematic literature search was done with the help of the *Pubtator* tool [126]. This resource has an R package bundled with records of scientific publications mentioning genes with ENTREZ identifiers [118]. This tool was chosen because it builds a local database that is perfect for quickly dealing with numerous queries. Figure 4.16 presents the results of this analysis. The used publications database focuses only on genes with ENTREZ identifiers, so only a subset of 17,584 genes involved in the coexpression estimations was analyzed. These genes had ENTREZ identifiers, were mentioned in at least one of the publications compiled by the *pubtatordb* package, and had at least one gene pair in their top 50 coexpressed partners with a mention in at least one publication. Details of the distributions observed in the particular case of Chi-square coexpression are shown in Figure 4.17.

Statistical significance between each coexpression metric and its corresponding random run was assessed with the Wilcoxon signed-rank test in Table 4.1 (2.9.2.3). The differences between occurrences of top coexpressed gene pairs in literature observed with each metric were also addressed in Table 4.2. All metrics as employed in the calculations of this work prove to be significantly more useful at finding interesting gene associations compared to what is possible by random chance. Interestingly, all metrics also perform significantly differently between them. The PCC is the metric that calls more top coexpressed pairs that also appear together in the literature publications.

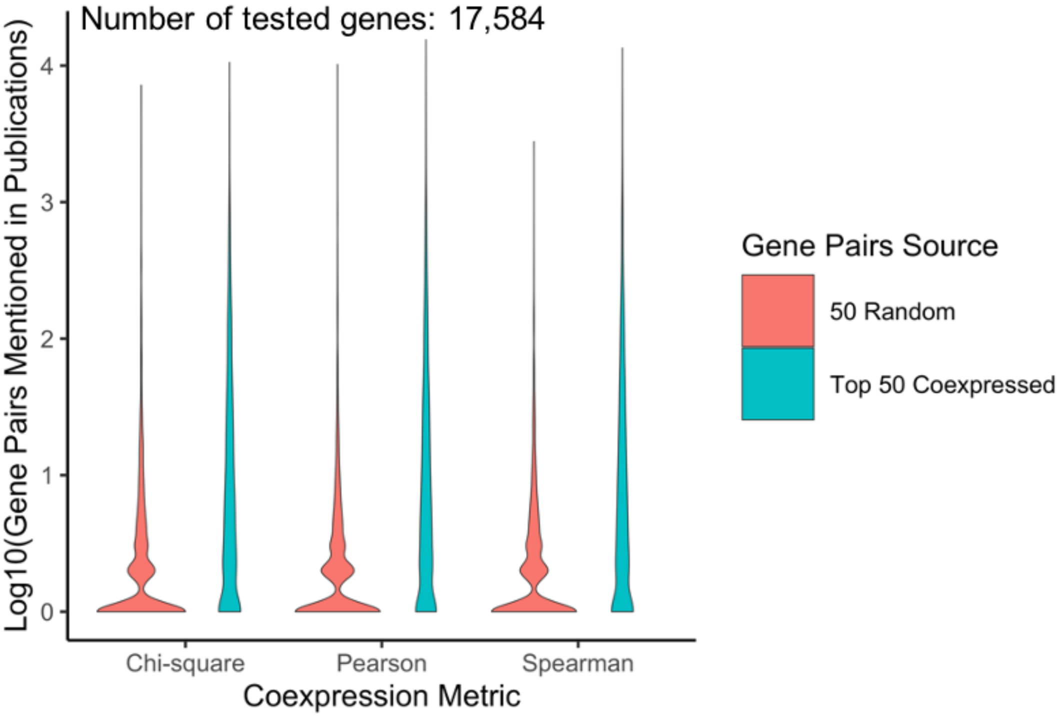


Figure 4.16: Violin plots of the occurrences distributions of gene pairs appearing together in scientific publications. Occurrences have been log-scaled to improve visualization. Distributions in blue correspond to occurrences when querying every gene in the analysis paired with each of its top 50 coexpressed partners (i.e. 17,584 times 50 queries were issued). Distributions in red are similar, but the 50 pairs for each gene are picked randomly instead of by coexpression strength. Distributions arising from random queries have a higher density around smaller occurrence values, indicating that gene pairs mentioned together in literature are hard to find randomly. Distributions in blue show higher overall density in higher occurrence values, indicating that strongly coexpressed pairs correspond better to pairs mentioned in literature than random pairs.

Table 4.1: Literature mentions of top coexpressed gene pairs against mentions of randomly chosen gene pairs.

Coexpression Metric	Sum of gene pairs appearing in publications		Wilcoxon test
	Top 50 Coexpressed	Random 50	p-value
Chi-square	552,222	43,321	$< 2.2 \times 10^{-16}$
Pearson	655,260	44,564	$< 2.2 \times 10^{-16}$
Spearman	592,335	27,846	$< 2.2 \times 10^{-16}$

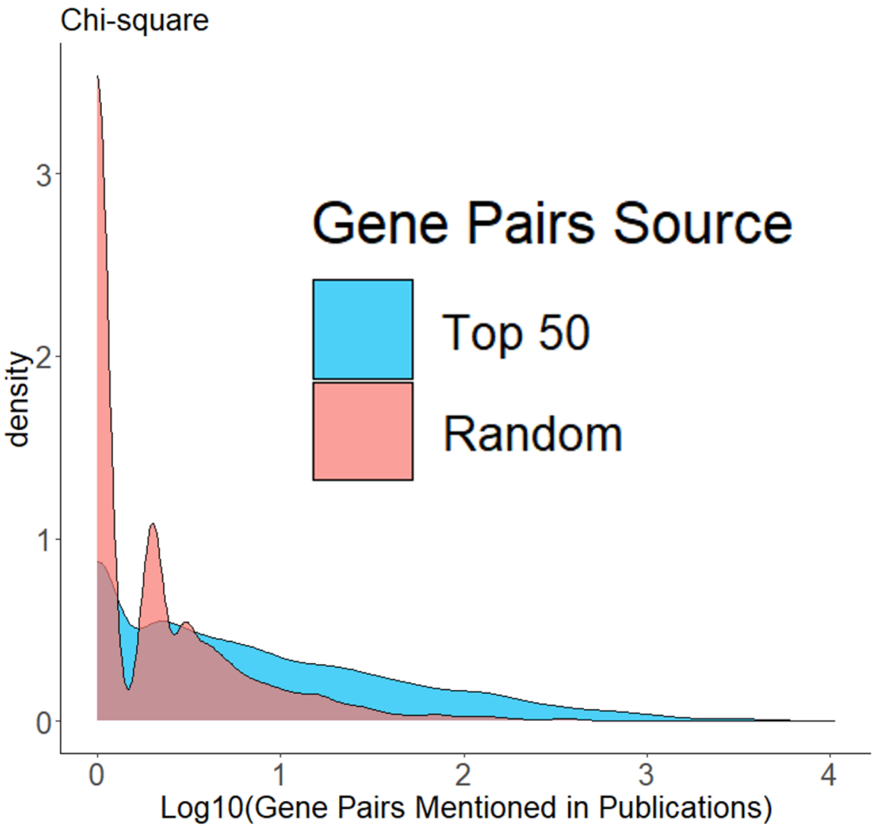


Figure 4.17: Distributions of observed and at random Chi-square coexpressed gene pairs in literature publications. Mentions of pairs of genes are more frequent when the pairs are picked according to coexpression strength (top 50 coexpressed with each gene of the 17,584 tested).

Table 4.2: Pairwise Wilcoxon signed-rank test between coexpression metrics. Occurrences in the literature of the top 50 coexpressed gene pairs with each of 17,584 tested genes are the input for the test.

p-values	Chi-square	Pearson	Spearman
Chi-square			
Pearson	8.48×10^{-10}		
Spearman	0.0007	0.005	

4.4.4 Number of associations per gene

Another interesting comparison between metrics is to look at the number of associations that each gene has with other genes in the project (i.e. Gene pairs with strong coexpression measurements). One way to analyze this is to establish a threshold for each metric above which an association between genes will be considered. To find a reasonable threshold, the cumulative distributions of all coexpressions for the 3 computed metrics were visualized as shown in Figure 4.18. The value for each of the metrics at which 95% of all coexpressions are accumulated was found. All associations with greater magnitude than these threshold values are therefore within the top 5% of all results independently of the implicated genes. With these thresholds defined, it is now possible to check the number of associations per gene. The thresholds were rounded to 300, 0.35, and 0.3 for Chi-square, PCC, and SRCC respectively. This is done for convenience as the number of associations at different predefined thresholds was calculated a priori before defining these specific thresholds during the parsing of coexpression results to a database format that is presented later (4.7.1)

A comparative histogram is presented in Figure 4.19 where the distribution of the number of associations per gene for each metric can be seen. In this figure there are also noted some interesting genes which had the maximum number of associations for the different coexpression metrics. The maximum for Chi-square was the nuclear body protein SP110 with a baffling number of 11,546 associations. While many of these associations could be attributed directly or indirectly to the biological activity of SP110, it is unfeasible to think that all of them are in fact relevant from a biological point of view. Other genes have a reasonable number of associations by means of the 300 Chi-square threshold. This suggests that depending on the gene, the number of associations might need to be determined with dynamic thresholds. However, it is still useful to investigate the number of associations with a set threshold as it provides insight into what kind of genes are frequently associated with others and can help validate the calculations.

In the case of SP110, it makes sense that this gene is associated with many other genes as it is a chromatin regulator that directly interacts with DNA, histones, and protein complexes [127]. The importance of this gene in the regulation of expression of many other genes is highlighted by the fact that mutations in SP110 and other members of the Speckled Protein family of genes are associated with multi-organ diseases such as Multiple Sclerosis and Crohn's Disease [127]. The gene with the most associations by the PCC was the DEAD-box helicase 25 (DDX25), a gene for which a great number of associations is also biologically justified. DEAD-box helicases are required for the process of transcription perse, for processing precursor messenger RNA (mRNA), and for ribosome biosynthesis among other functions that relate them with many genes [128]. Neugrin (NRGN) was the gene with the most associations by the SRCC. Once again, it makes sense that a gene such as this has a high number of associations because NRGN is directly implicated in a complex that mediates ribosomal biosynthesis in the mitochondria [129]. Ribosomes are the molecular structures in charge of translating mRNAs to proteins, so genes related to ribosomes are frequently coexpressed with many genes [16]. The fact that these kinds of genes are popping up with a high number of associations in the analyses carried out of this work is a sign of the correctness of the calculations. Upon inspecting genes that had zero associations for all metrics it was observed that the majority of these comprise unannotated genes of the RP11 or RP13 type.

As part of the exploration of the number of associations per gene deriving from the coexpression calculation carried out in this work, 21 interesting genes were inspected (Table 4.3). These genes are known genes associated with cancer. Hence they are expected to be associated with other genes due to the great impact they have over normal human biology when they become disturbed by the disease process. As expected, all of these genes have associations with others by all metrics.

Table 4.3: List of example cancer associated genes and their number of associations by metric. Order of appearance is given by the number of associations by Chi-square statistic.

Gene	Chi-Square	PCC	SRCC
FAM135B	6,378	4,916	2,938
PARK7	5,642	3,365	4,199
SOD1	5,443	2,793	3,201
PCLO	5,043	3,478	2,593
TP53	4,691	2,256	3,162
ZNF569	3,615	4,187	3,598
BMPR2	3,249	2,359	3,390
BRCA2	2,165	1,915	2,782
IGF1	2,061	1,489	2,838
ADAMTS7	1,815	1,722	2,735
KDM6A	1,174	1,013	1,311
ACTB	1,120	813	1,417
IDH1	540	751	1,395
BRCA1	408	563	310
MUC4	370	508	284
MCU	366	415	856
DNMT3A	295	632	943
CUBN	217	244	1,047
MUC17	137	175	89
KRAS	115	525	531
ESR1	95	345	1,890

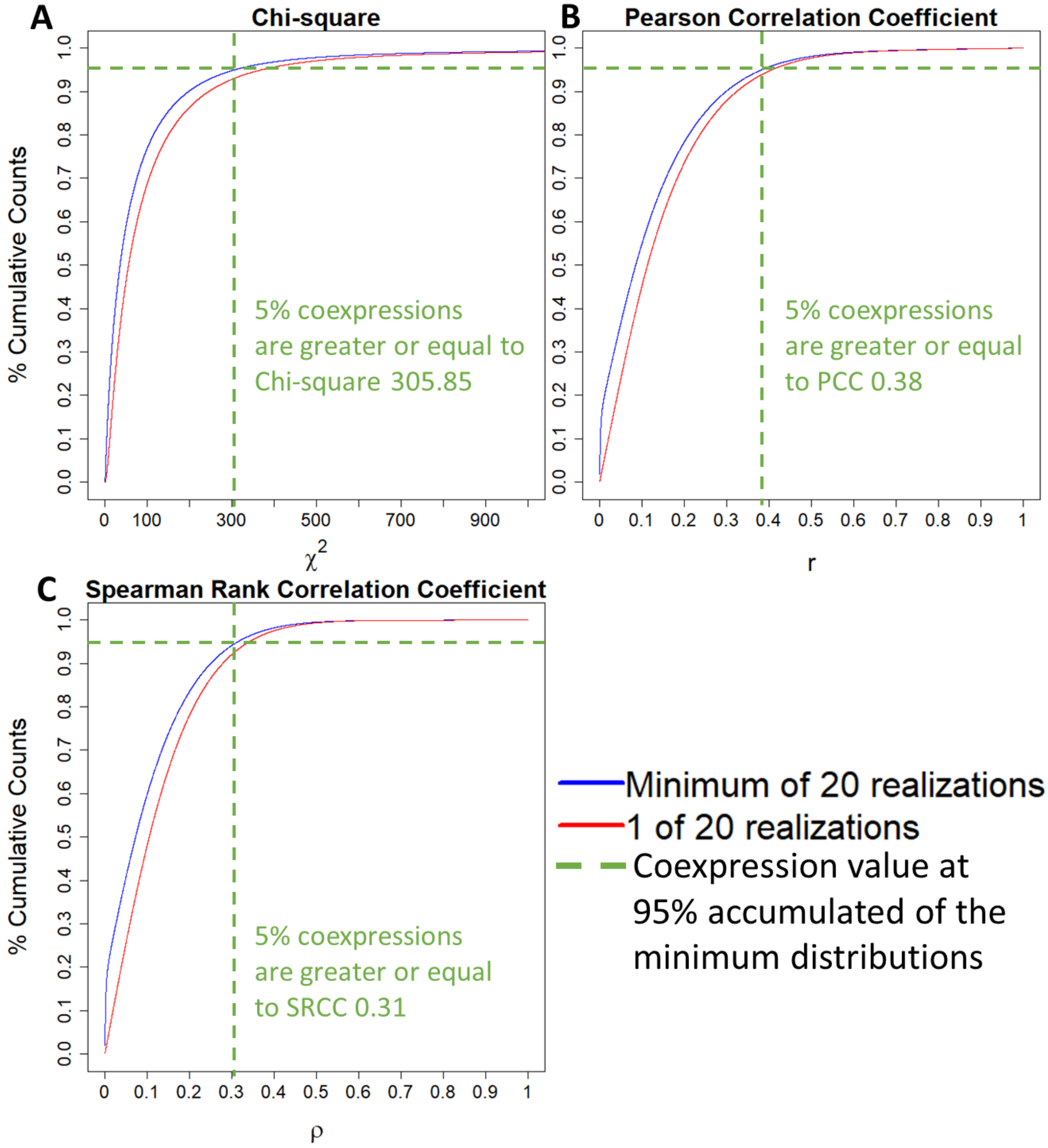
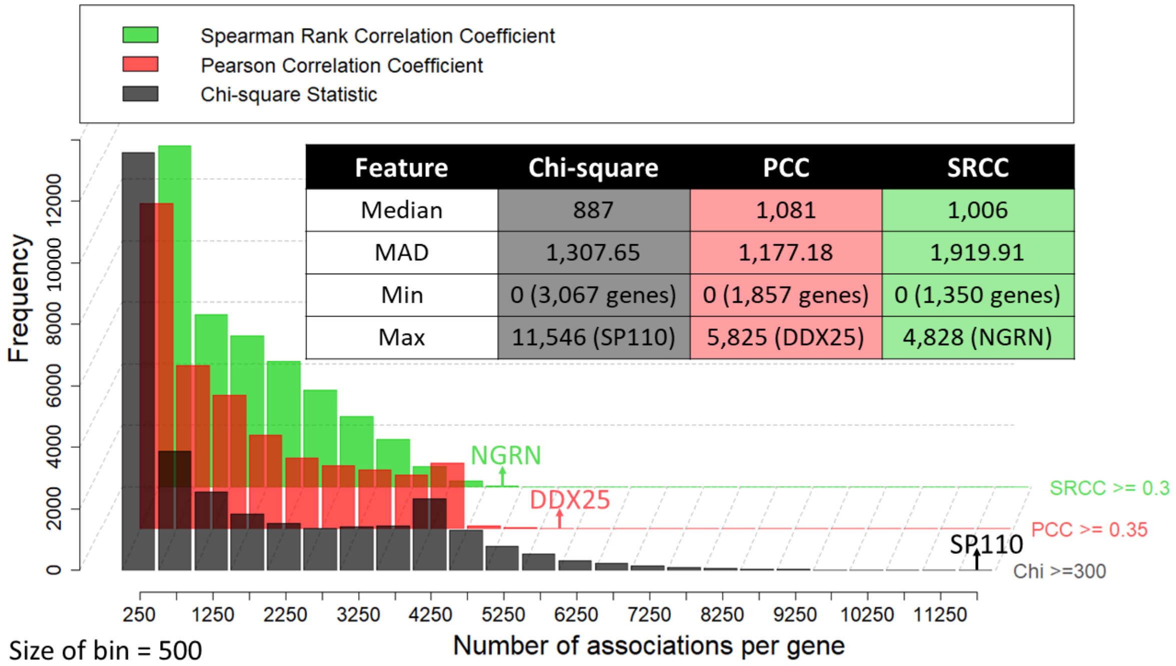


Figure 4.18: Cumulative distributions of observed coexpression values for all metrics. For each metric, the value accumulating 95% of the calculated coexpressions in the minimum of 20 distribution is indicated. A: Chi-square statistic. B: PCC. C: SRCC.



Feature	Chi-square	PCC	SRCC
Median	887	1,081	1,006
MAD	1,307.65	1,177.18	1,919.91
Min	0 (3,067 genes)	0 (1,857 genes)	0 (1,350 genes)
Max	11,546 (SP110)	5,825 (DDX25)	4,828 (NGRN)

Figure 4.19: Distribution of the number of associations per gene per coexpression metric. Overall, the Chi-square statistic seems to have the most genes with a high number of associations as this distribution has a more dense tail compared to the correlation coefficients. All metrics have cases of genes for which no associations are found with any other gene in the data according to the selected thresholds. For descriptive statistics, the median and MAD are chosen over mean and variance as the distributions are skewed. Interestingly, Chi-square has the smallest median and MAD number of associations per gene despite exhibiting several genes with more associations than the maximum observed for the correlation coefficients.

4.5 Gene expression states contribution to system level coexpression

In the system-level coexpression estimation, a balanced collection of samples from different gene expression states is used as input for the calculations. The observed coexpression measures resulting from this estimation reflect what is discovered globally as a combination of trends across all gene expression states. A natural question to ask is: in what proportion do different gene expression states support specific coexpression measurements? To be able to answer this question, Algorithm 3 was devised as a heuristic to investigate this specifically for the Chi-square measure.

The idea is to repeat the estimation of coexpression but using a ”permute one cluster out” strategy where it is possible to compare the observed statistic without any permutations to each one of the permuted realizations. This allows for estimating the importance of the corresponding clusters. If the difference between the observed statistic and the permuted one is small, this implies that the permuted samples were not very relevant to the calculation. On the other hand, if a great difference ensues, then it means that the permuted samples were key for the coexpression.

Algorithm 3: Estimate the contribution to Chi-square coexpression of a particular class in a partition of samples

```
input  : An expression matrix $E$ with $n$ genes in the rows and $m$ samples in the columns, the column
         indices $p$ of the samples in a particular cluster
output: A coexpression matrix of chi-square measures

Function partContrib($E, p$):
    // Obtain the limits between gene expression categories as given by
       k-means in the non-permuted data
    $L \leftarrow$ initializeMatrix($n, 2$);
    $c \leftarrow$ initializeVector($n$);
    for $i \leftarrow 1$ to $n$ do
        $c \leftarrow$ kmeans($E[i,]$);
        // Get the gene expression value that separates low expression
           from medium expression, and medium expression from high
           expression
        $L[i,] \leftarrow$ findLimits($c, E[i,]$);
    end
    // Permute each gene vector in $E$
    $M \leftarrow$ initializeMatrix($n, m$);
    for $i \leftarrow 1$ to $n$ do
        $M[i,] \leftarrow$ sample($E[i,], n$);
    end
    // Replace the currently processed cluster in $E$ with the permuted
       data at the same position in $M$
    for $i \leftarrow 1$ to $n$ do
        $E[i,p] \leftarrow M[i,p]$
    end
    // Discretize each gene vector in $E$ with the expression limits found
       when no cluster is permuted
    for $i \leftarrow 1$ to $n$ do
        for $j \leftarrow 1$ to $m$ do
            if $E[i,j] < L[i,1]$ then
                // Low expression
                $E[i,j] \leftarrow 1$ ;
            else if $E[i,j] > L[i,2]$ then
                // High expression
                $E[i,j] \leftarrow 3$ ;
            else
                // Medium expression
                $E[i,j] \leftarrow 2$ ;
            end
        end
    end
    $Y \leftarrow$ coexp($X$);
    return $Y$;
```

Algorithm 3 was run 34 times (one for each cluster) to obtain the distributions shown in Figure 4.20. For this experiment, one realization of the 20 in the global estimation was used as a reference (the same 1 of 20 shown in Figure 4.1). The latter means that the expression matrix used as an input for said realization was the one undergoing each of the permutations (i.e. It was kept the same for all the ”permute one cluster out” runs) and subsequent coexpression calculation. This reference realization is labeled as ”not permuted” in Figure 4.20. The distribution that was more affected by the permutation procedure was by far the one corresponding to the Testis cluster. This implies that this cluster has a very marked contribution in more coexpressions than any other gene expression state. This clicks with the fact that many of the genes in the input data that exhibit some degree of tissue-specificity are Testis-specific (3.1). Expression is an obvious prerequisite for coexpression, so the associations of these numerous testis-specific genes are frequently driven only by samples of this cluster.

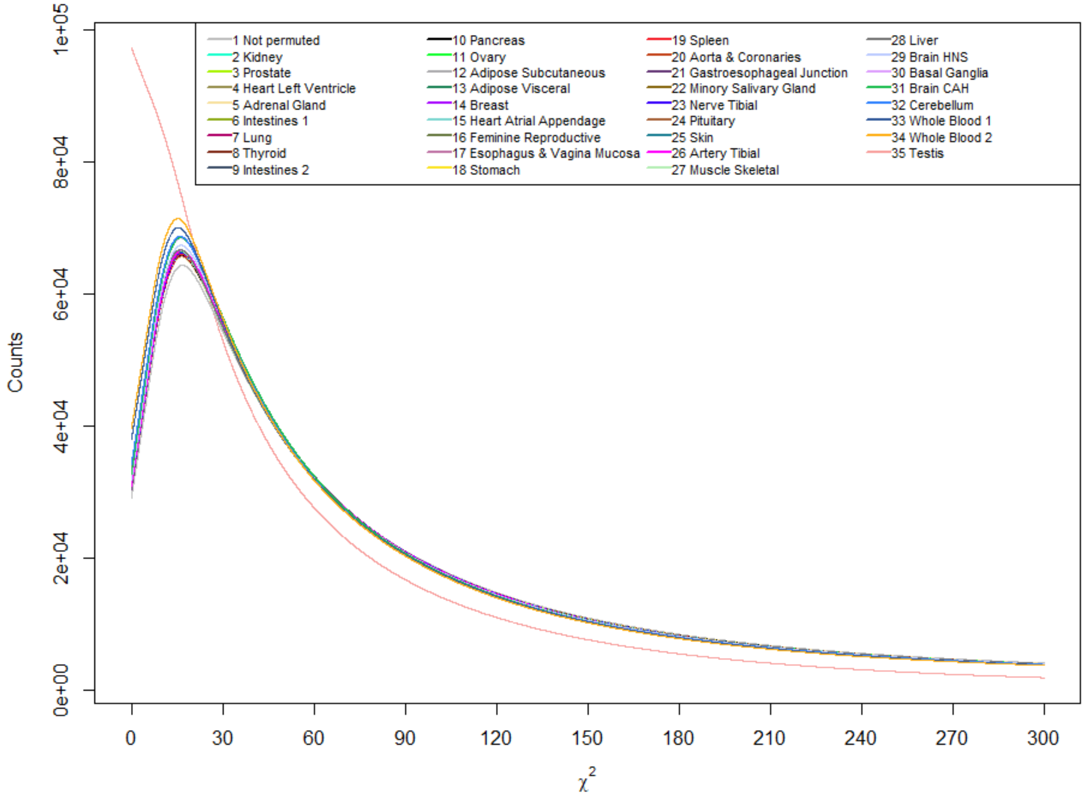


Figure 4.20: Chi-square distributions found by permuting one cluster at a time in the input matrix of 1 of 20 realizations of the BRS procedure (curve labeled as ”not permuted” derives from this original matrix). Lower ranks (e.i. Bigger numbers) assigned in the legend signify a lower amount of coexpressions that remained approximately the same after permuting compared to before permuting. This histogram is plotted in lineplot style to allow for the visualization of multiple distributions at once. Curves have been smoothed to gain resolution when lines become crowded. The smoothing does not change the rank of the clusters nor causes the counts’ values to change significantly.

Figure 4.21 illustrates the concept of the ”permute one out” technique more graphically. The example was chosen for validation by searching an external tissue-specific genes database where the UMOD gene was found as Kidney-specific [130]. The coexpressions for such gene were retrieved from the calculated data.

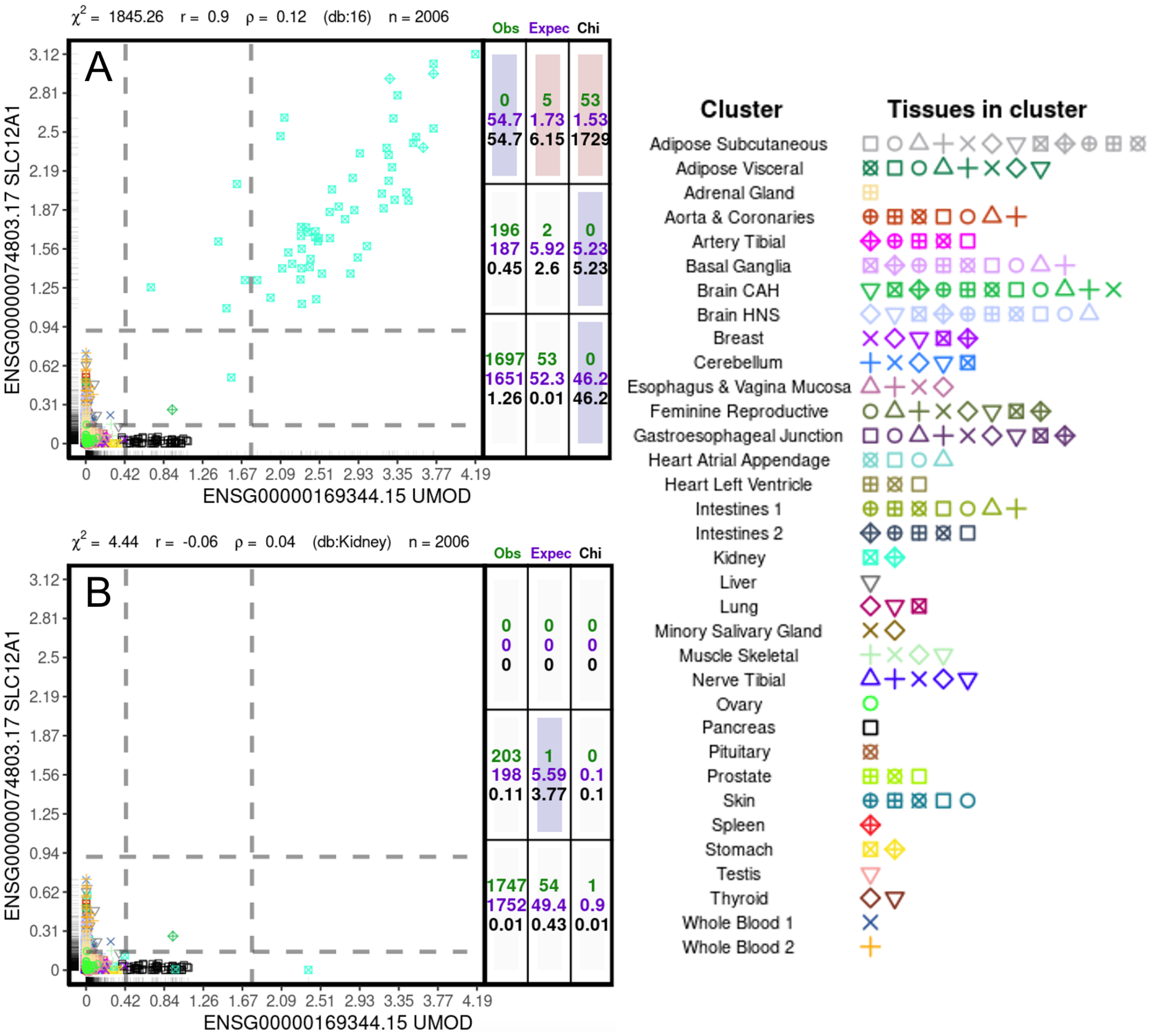


Figure 4.21: Example of a Kidney-driven coexpression between the uromodulin (UMOD) gene and the solute carrier family 12 member 1 gene (SLC12A1). A: observed coexpression in the reference realization of BRS (labeled *db:16*). B: Kidney realization of the ”permute one cluster out” experiment. The Chi-square statistic after permuting the Kidney samples is less than 1% the original calculated statistic.

The biggest changes between the permuted Kidney realization and the observed reference Chi-square were checked. It is very clear that before the cluster is permuted, the coexpression is driven solely by Kidney samples. Moreover, in the rest of the clusters, there is even an obvious relationship of no coexistence between genes. When one is expressed, the other one is basically at 0 expression. Upon permuting the Kidney cluster, the association is destroyed. As of now, a very simple metric $permScore$ is used for measuring these effects for all gene expression states across the hundreds of millions of coexpressions estimated in this project:

$$permScore_{ic} = 100 - [(\chi^2_{ic} * 100)/\chi^2_{ir}] \tag{4.1}$$

Where χ^2_{ic} is the Chi-square statistic obtained for coexpression i after permuting cluster c in the input data. χ^2_{ir} is the analogous statistic calculated from the reference data r in which no cluster is permuted. For the example in Figure 4.21, the $permScore$ is 99.75%. This indicates that when Kidney samples are permuted, 99.75% of the initial association observed between the genes disappears. The $permScore$ metric comprises a way for quickly knowing if any particular coexpression is strongly supported by some specific gene expression states.

One important annotation to make is that using $permScore$ does not always lead to a positive or 0 value. Sometimes the Chi-square obtained after permuting one cluster out can increase instead of decreasing or staying the same when compared to the reference. An example of this situation is shown in Figure 4.22 where it can be seen that the increased statistic after the permutation is a product of chance. This can be said because the sample causing the increase in the Chi-square statistic is of the same cluster that is being permuted. Many other studied examples like this exhibited similar patterns. Negative $permScore$ values can be interpreted in a similar fashion to 0 values because permuting the cluster in question does not decrease the real observed coexpression.

4.6 Estimation for comparison of coexpression by covariates

An additional experiment inspired by a series of recent 2020 publications using the GTEx data to investigate the effect of different covariates such as gender in gene expression [131] was performed (such papers are discussed further in the last chapter of this work). One realization of gene coexpression data (Chi-squared) was calculated on some selected subsets of samples representing different covariate levels across gene expression states. The covariates used are the ones that are freely accessible as GTEx metadata. The idea is to allow for exploration of similar concepts to those presented in the aforementioned papers, but for coexpression relationships instead of single gene expression. Balance of gene expression states was also sought for these calculations, but it proved to be harder due to limited numbers of samples representing certain covariable levels in some gene expression states. The exact composition of samples used is documented in this section for each covariable.

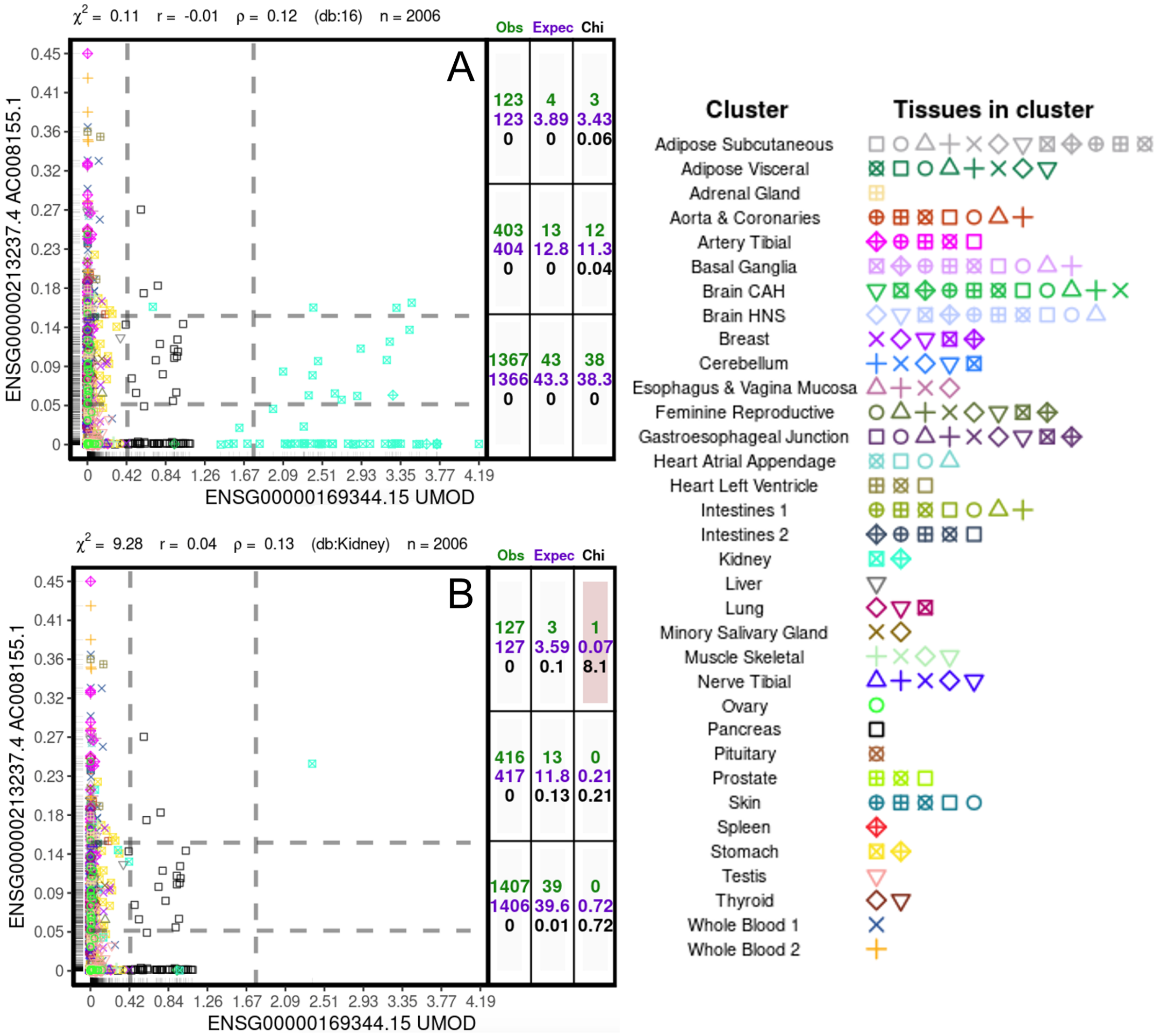


Figure 4.22: Example of a coexpression that strengthens upon permuting a cluster (in this case Kidney). A: observed coexpression in the reference realization of BRS (labeled *db:16*). B: Kidney realization of the ”permute one cluster out” experiment.

4.6.1 Gender

Gender of the GTEx donors is distributed across the gene expression states estimated in this work as shown in Figure 4.23. The two levels of this covariable are simply described as:

- *Female*: samples from female donors (5,562 samples total).
- *Male*: samples from male donors (11,142 samples total).

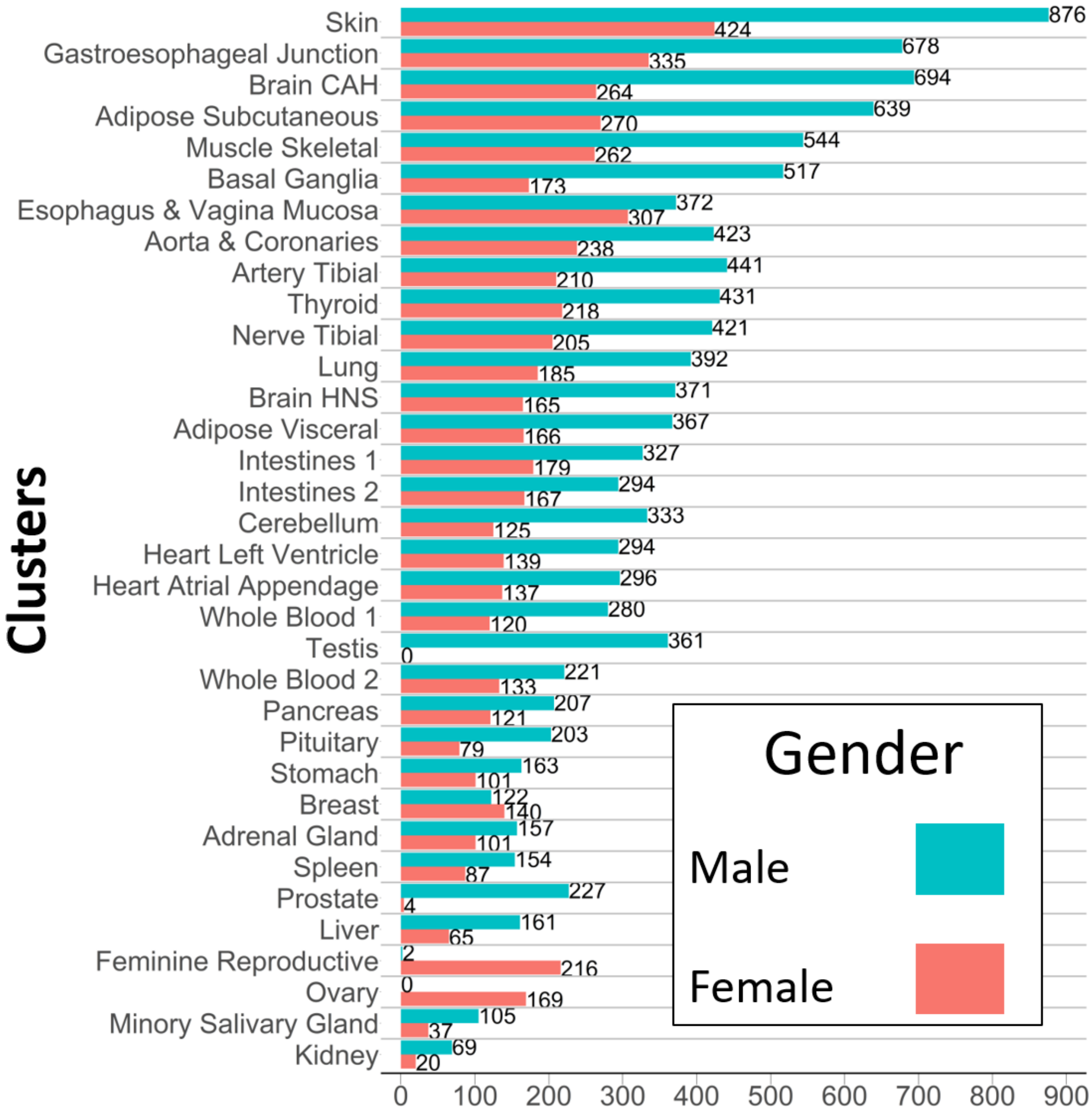


Figure 4.23: Distribution of GTEx donors gender across the gene expression states that were characterized in this work (3.2).

Gene expression matrices were assembled that only considered samples from female or male donors respectively. Logically, and as seen in Figure 4.23, some gene expression states are exclusive (or nearly exclusive) of a certain gender. These need not be considered for this experiment as they will not be comparable. Such clusters are Prostate, Feminine Reproductive, Testis, and Ovary. 34 samples from each of the other clusters were randomly chosen to assemble the input gene expression matrices. One exception was made with the Kidney cluster, which only counted with 20 samples for females. In the case of the latter cluster, only 20 samples were taken for both genders. After calculating coexpression with these gender-specific inputs, resulting distributions were plotted in Figure 4.24.

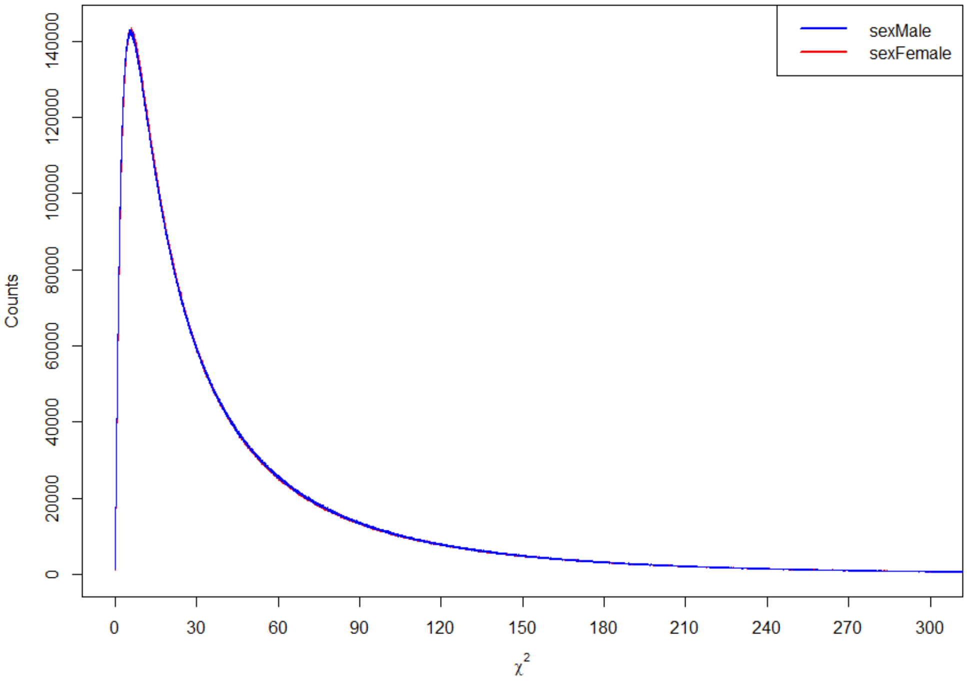


Figure 4.24: Coexpression Chi-squared distributions resulting from gender-separated gene expression. As expected, distributions are extremely similar since it is a known result that the effect of gender on individual gene expression is overall small [131]. This histogram is plotted in lineplot style to allow for a clearer visualization of both distributions at once.

After examining some coexpressions that had big differences between genders, most of them involve chromosome Y genes such as the ones exemplified in Figure 4.25. Even if this is expected, interesting additional information may still be obtained from these comparisons. For example, in Figure 4.25, the coexpression is not supported globally by all clusters in males as one may initially assume. Only brain gene expression states support the observed statistic. The coexpression is not only gender-specific but also encountered exclusively in the brain of males. Another example shown in Figure 4.26 presents a case of a female-specific coexpression. This example was actually found through a literature revision which yielded sex-biased genes in their expression [132]. It was hypothesized that these genes may have sex-biased coexpressed pairs due to their selective expression. Some of the biggest differences in coexpressions involving these genes between females and males were checked.

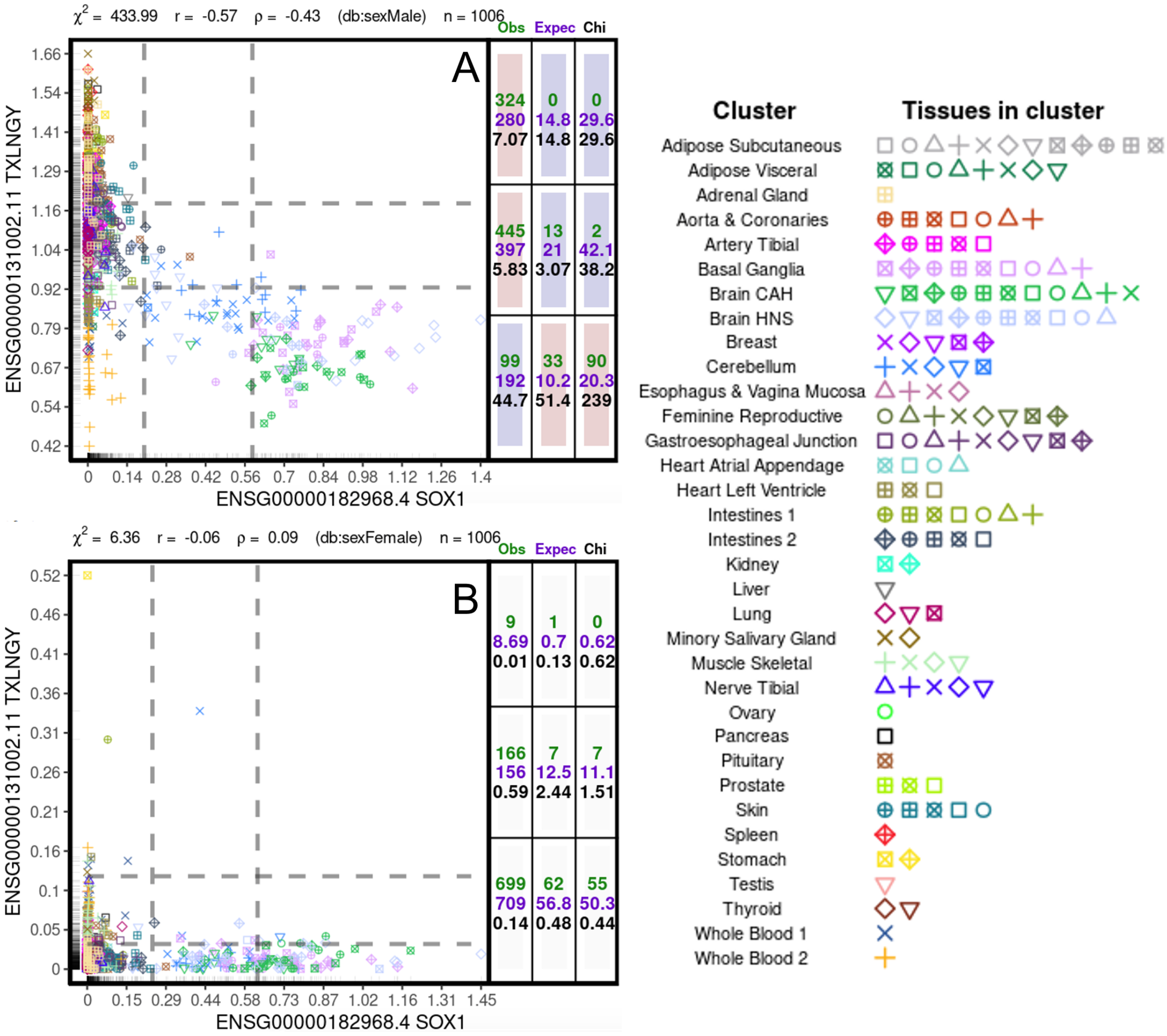


Figure 4.25: Example of a coexpression with a big difference (as quantified by Chi-square) between genders. SRY-box transcription factor gene (SOX1) and taxilin gamma pseudogene (TXLNGY). A: coexpression as observed males. B: coexpression as observed in females.

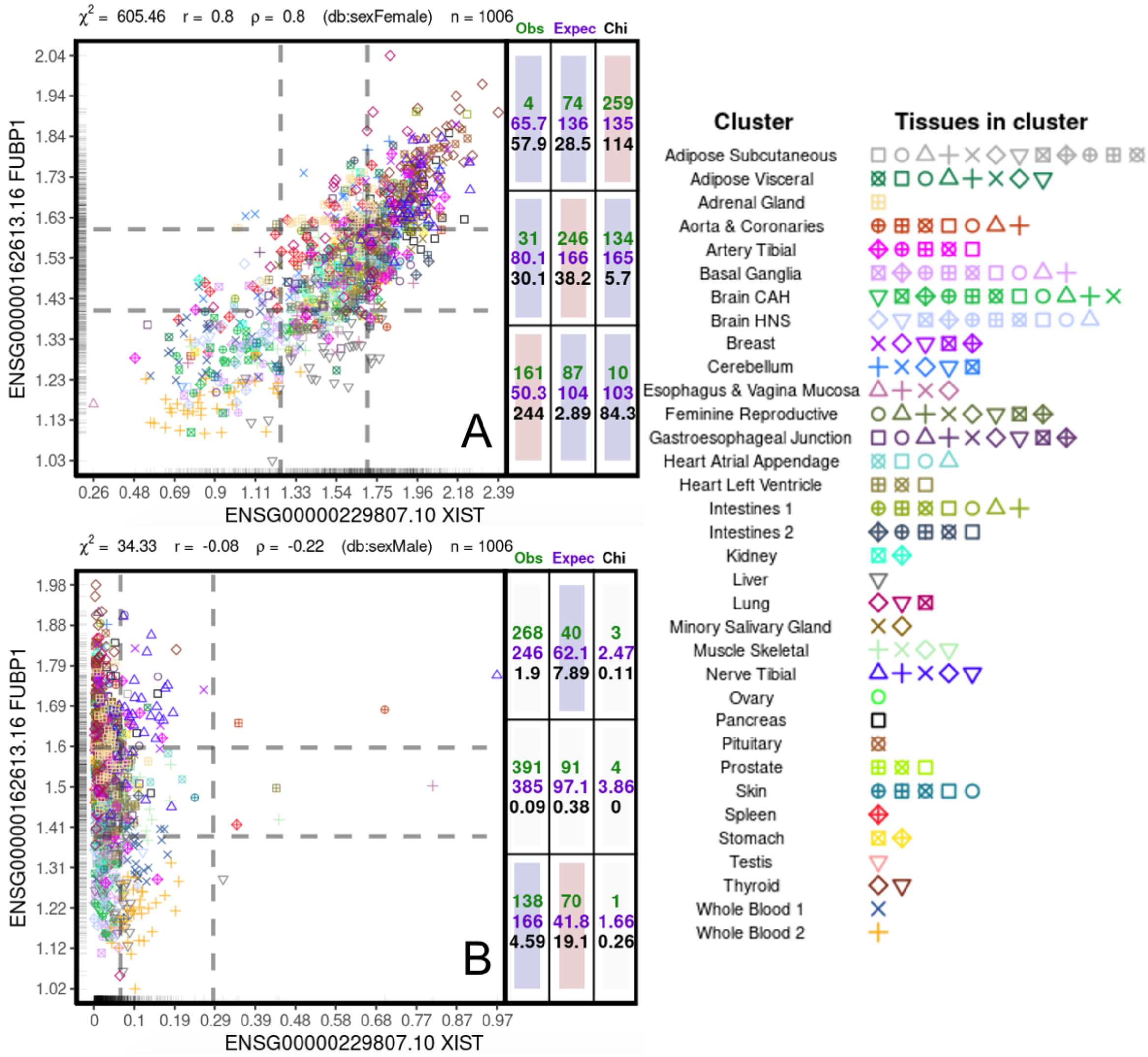


Figure 4.26: Example of a coexpression with a big difference (as quantified by Chi-square) between genders. X-inactive specific transcript gene (XIST) and far upstream element binding protein (FUBP1). A: coexpression as observed males. B: coexpression as observed in females. Coexpression in females is given by all gene expression states.

4.6.2 Age

This covariable is freely provided by GTEx in a binned format (i.e. Predefined categories such as 40-49, 50-59, etc.). As some of these bins contained very few samples from certain clusters, these categories were rebinned into 3 major groups as shown in Figure 4.27:

- $<$ *50*: donors of less than 50 years of age (5,102 samples total).
- *50-59*: donors that have between (inclusive) 50 and 59 years of age (5,392 samples total).
- $>$ *59*: donors that have more than 59 years of age (6,210 samples total).

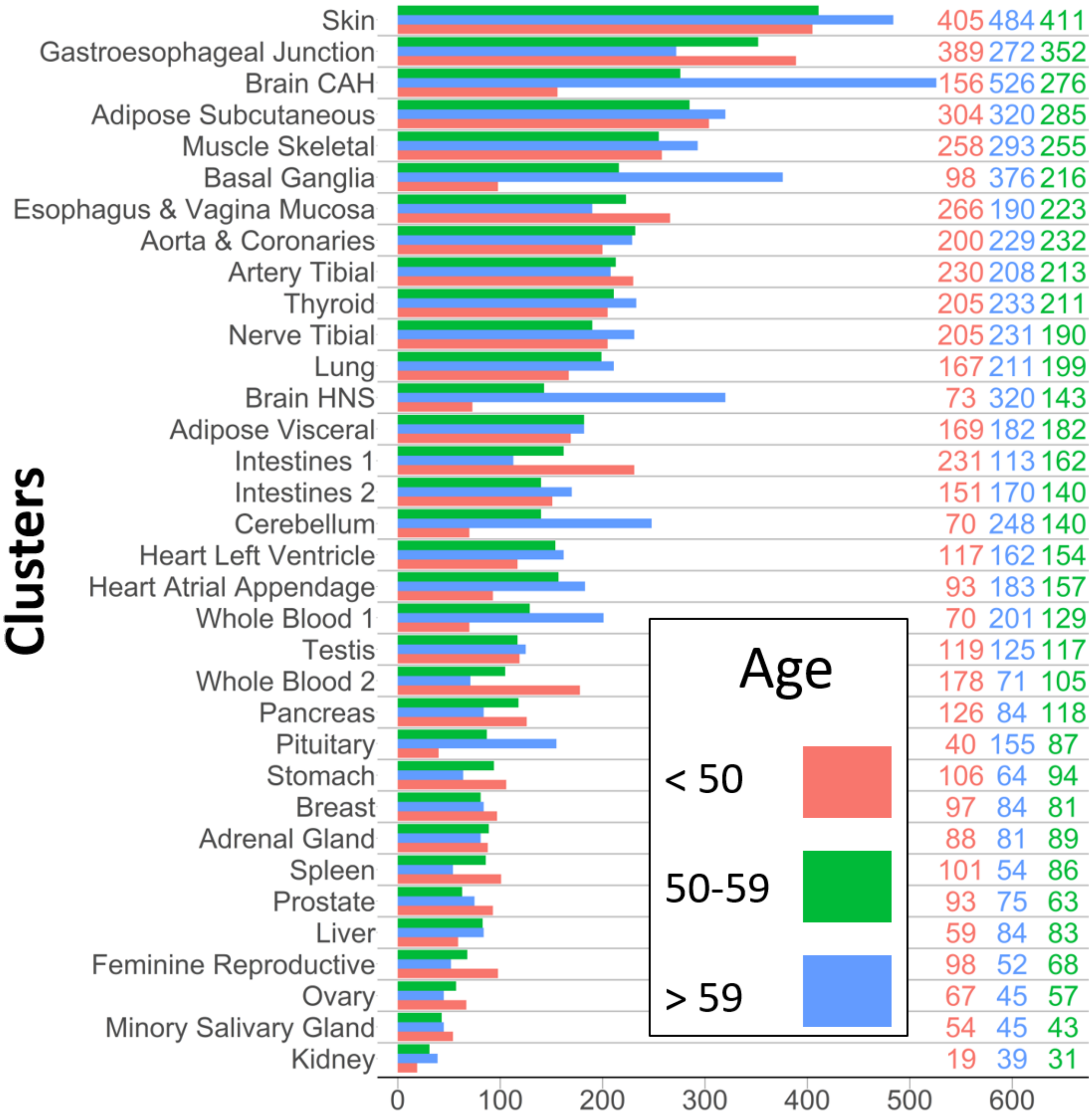


Figure 4.27: Distribution of GTEx donors age across the gene expression states that were characterized in this work (3.2).

For each age group, a gene expression matrix was assembled that exclusively considers samples from donors of each group. 40 samples from each cluster were randomly chosen to assemble these matrices. One exception was made with the Kidney cluster, which only counted with 19 samples for one of the age groups defined. In the case of the latter cluster, only 19 samples were taken for all age groups. Upon calculating coexpression with these age-specific input gene expression matrices, the distributions shown in Figure 4.28 summarize the results.

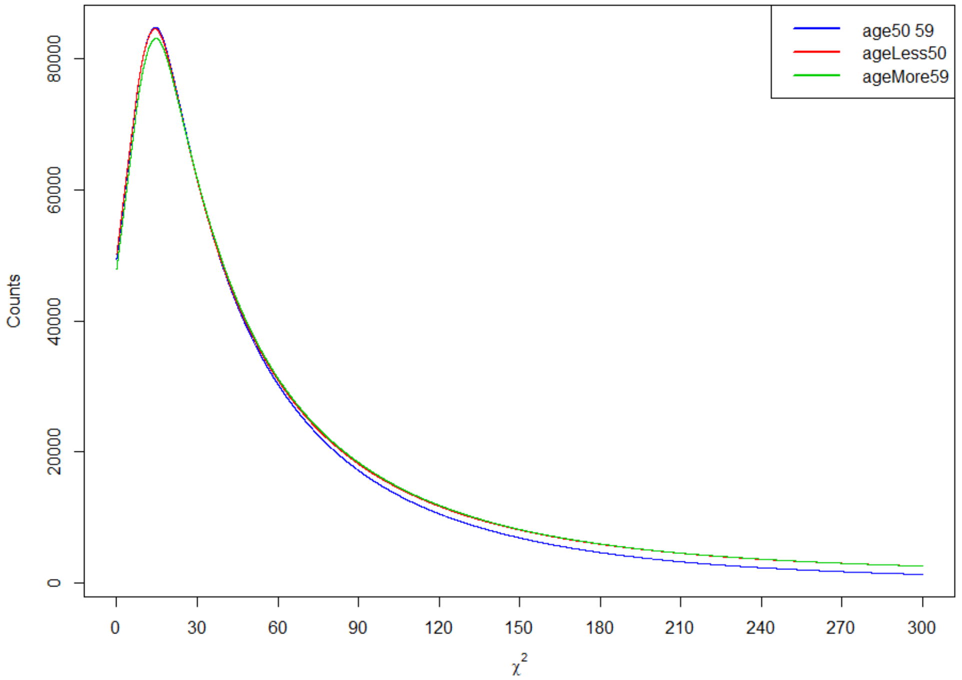


Figure 4.28: Coexpression Chi-squared distributions resulting from age-separated gene expression. As expected, distributions are similar since it is a known result that the effect of age on individual gene expression is small [133]. However, there are still subtle changes that can be observed between the age variable levels. Age $>$ 59 has more density in weaker Chi-square values as apposed to the other two categories. The 50-59 group seems to have less density for stronger values. This histogram is plotted in lineplot style to allow for the visualization of multiple distributions at once. Curves have been smoothed to gain resolution when lines become crowded. The smoothing does not cause the counts' values to change significantly.

After examining some coexpressions that had big differences between age groups, even the ones with the greatest differences derive from subtle changes in the gene expression of the involved pairs. The Chi-square test can detect these subtle changes during coexpression calculation. One of these cases is presented in Figure 4.29 where the gene expression of the pair shown diminishes on the more aged group of donors specifically for the Intestines 1 cluster. This is what mainly drives the change in the observed statistics. The latter is concordant with literature as age only explains a small fraction of the variance in gene expression and this variance is on its majority tissue-specific [133].

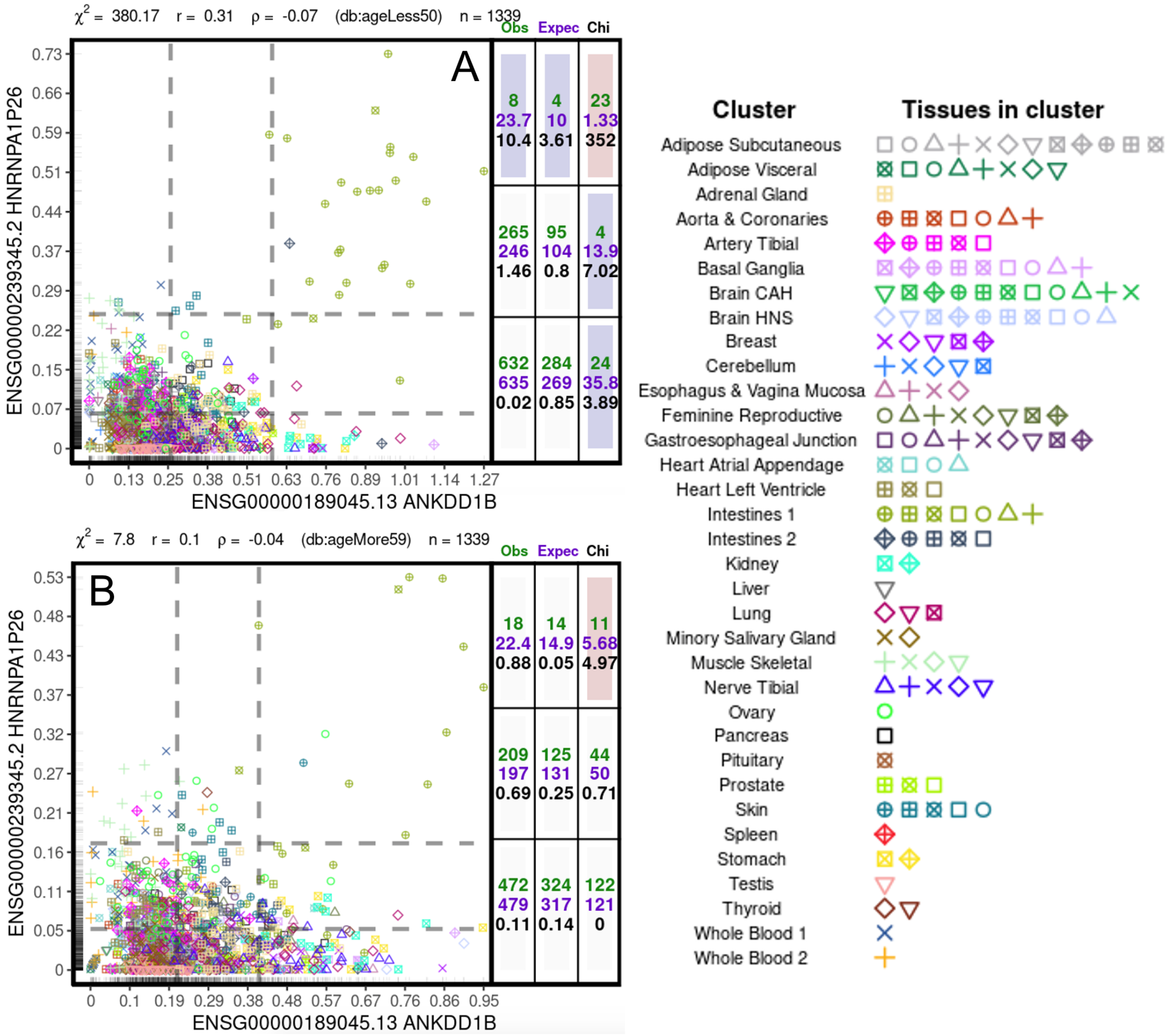


Figure 4.29: Example of a coexpression with a big difference (as quantified by Chi-square) between age groups. Ankyrin repeat and death domain containing 1B gene (ANKDD1B) and heterogeneous nuclear ribonucleoprotein 26 A1 pseudogene (HNRNPA1P26). A: coexpression as observed in the $<$ 50 group. B: coexpression as observed in the $>$ 59 group.

4.6.3 Post mortem interval

Ischemia time or Post Mortem Interval (PMI) of the GTEx samples is the time in minutes that passes from the death of the donor to the obtention of the sample. It is freely given in the GTEx metadata as integer values. These values are discretized into two categories representing low and high ischemia times respectively using a similar strategy to what is done for discretizing gene expression (2.5.1.1). The two obtained categories are distributed across the gene expression states estimated in this work as shown in Figure 4.30. These levels of this covariate are described as:

- *0-648*: samples with a PMI between 0 and 168 (inclusive) minutes (9,313 samples total).
- $\geq$ *648*: samples with a PMI greater than or equal to 649 minutes (7,053 samples total).

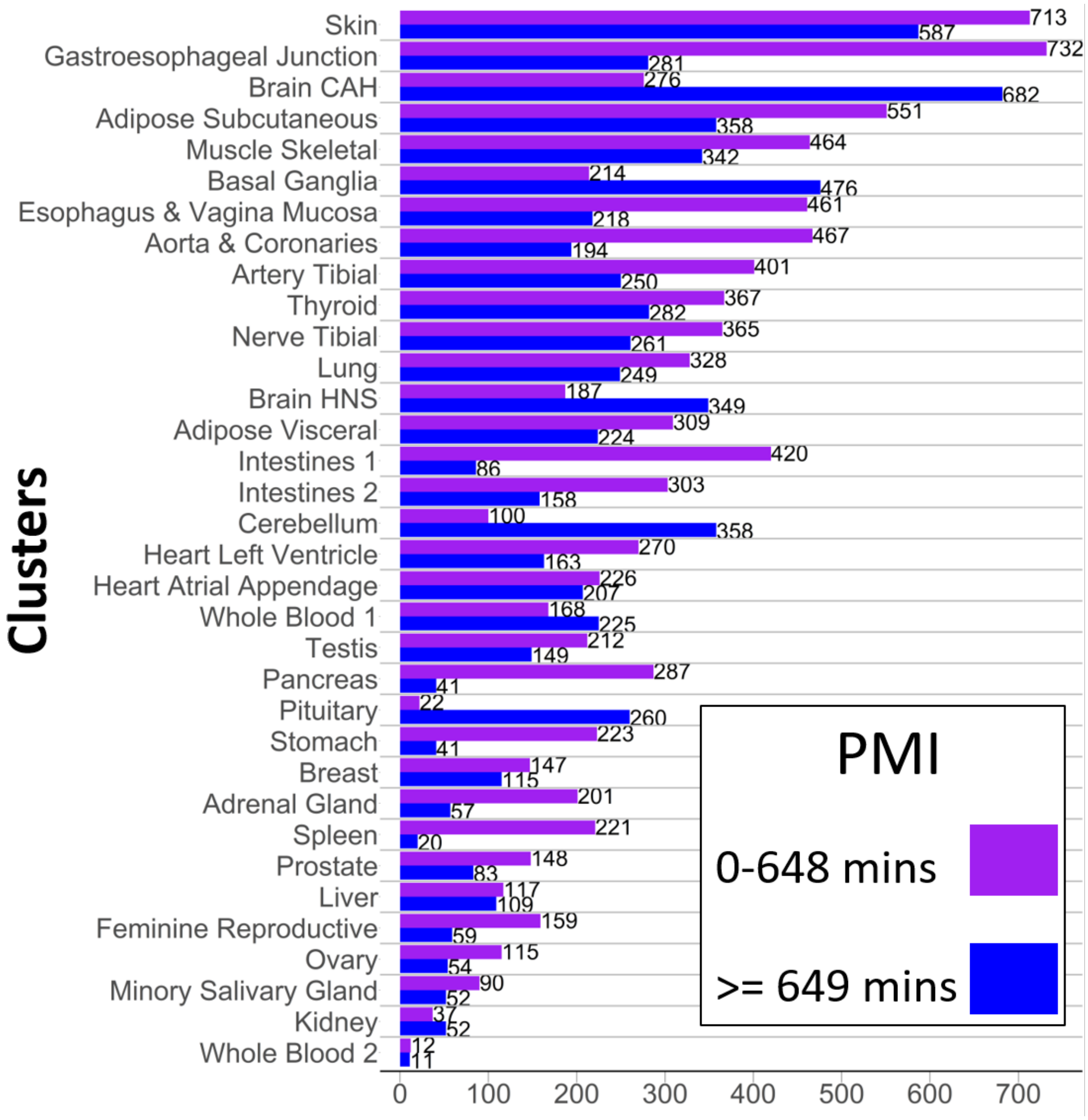


Figure 4.30: Distribution of GTEx samples PMI categories across the gene expression states that were characterized in this work (3.2).

Gene expression matrices were assembled that only consider samples from the low or high ischemia groups respectively. As seen in Figure 4.30, one of the Whole Blood gene expression states has almost no samples in any of the 2 PMI groups. The reason for this is that most of the samples in that cluster have negative PMI values (e.i. They were taken before the death of the donor). Considering these samples would make groups less comparable as there are almost no samples from that cluster with higher PMIs. That Whole Blood cluster is removed from the experiment and 22 samples from each of the other clusters were randomly chosen to assemble the input gene expression matrices. Two exceptions were made with the Pituitary and Spleen clusters, which did not have 22 samples for both PMI categories. In these cases, only 1 sample and 20 samples respectively were taken for both groups. After calculating coexpression with these PMI-specific inputs, Chi-square distributions were obtained as shown in Figure 4.31.

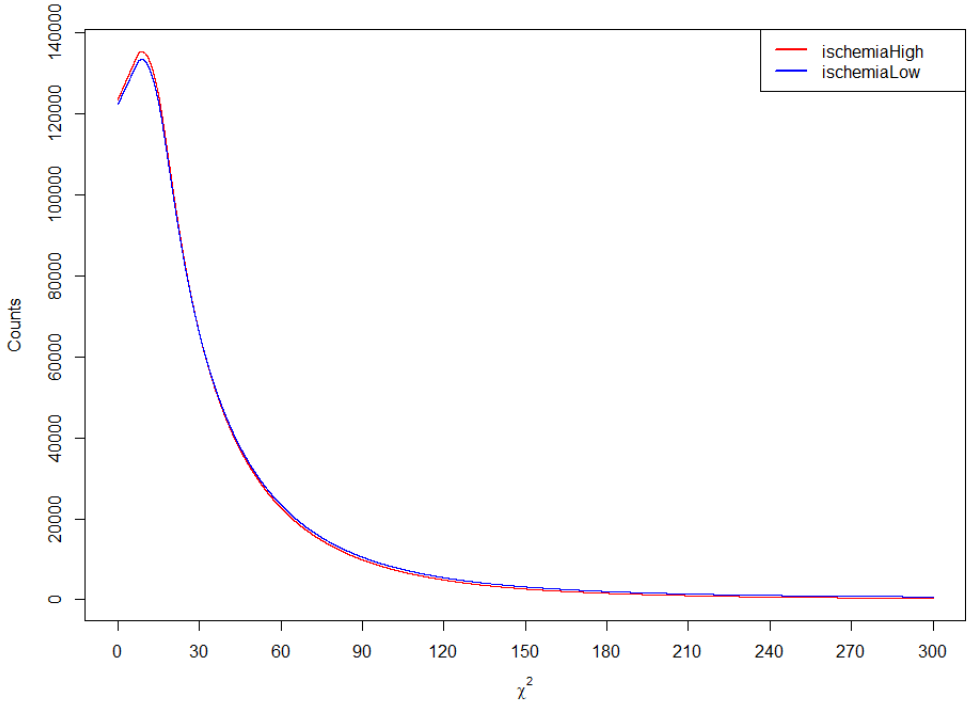


Figure 4.31: Coexpression Chi-squared distributions resulting from PMI-separated gene expression. Distributions are similar since it is a known result that the effect of cold ischemia on individual gene expression is overall small [117]. Nevertheless, a small gain in the density for stronger Chi-square values in the low ischemia group is notable. This histogram is plotted in lineplot style to allow for the visualization of multiple distributions at once. Curves have been smoothed to gain resolution when lines become crowded. The smoothing does not cause the counts' values to change significantly.

After examining some coexpressions that had big differences between PMI categories, most of them derive from subtle changes in the gene expression of the involved pairs. The Chi-square test can detect these subtle changes during coexpression calculation. The example shown in Figure 4.32 depicts one such case which involves an intensification of the expression of two hemoglobin related genes in the high ischemia group when compared to the low ischemia group. The upregulation in the expression of these kinds of genes due to hypoxia in tissues other than Whole Blood has been shown for individual genes [117]. It can be explored through the PMI-based estimation that this is also reflected at the coexpression level.

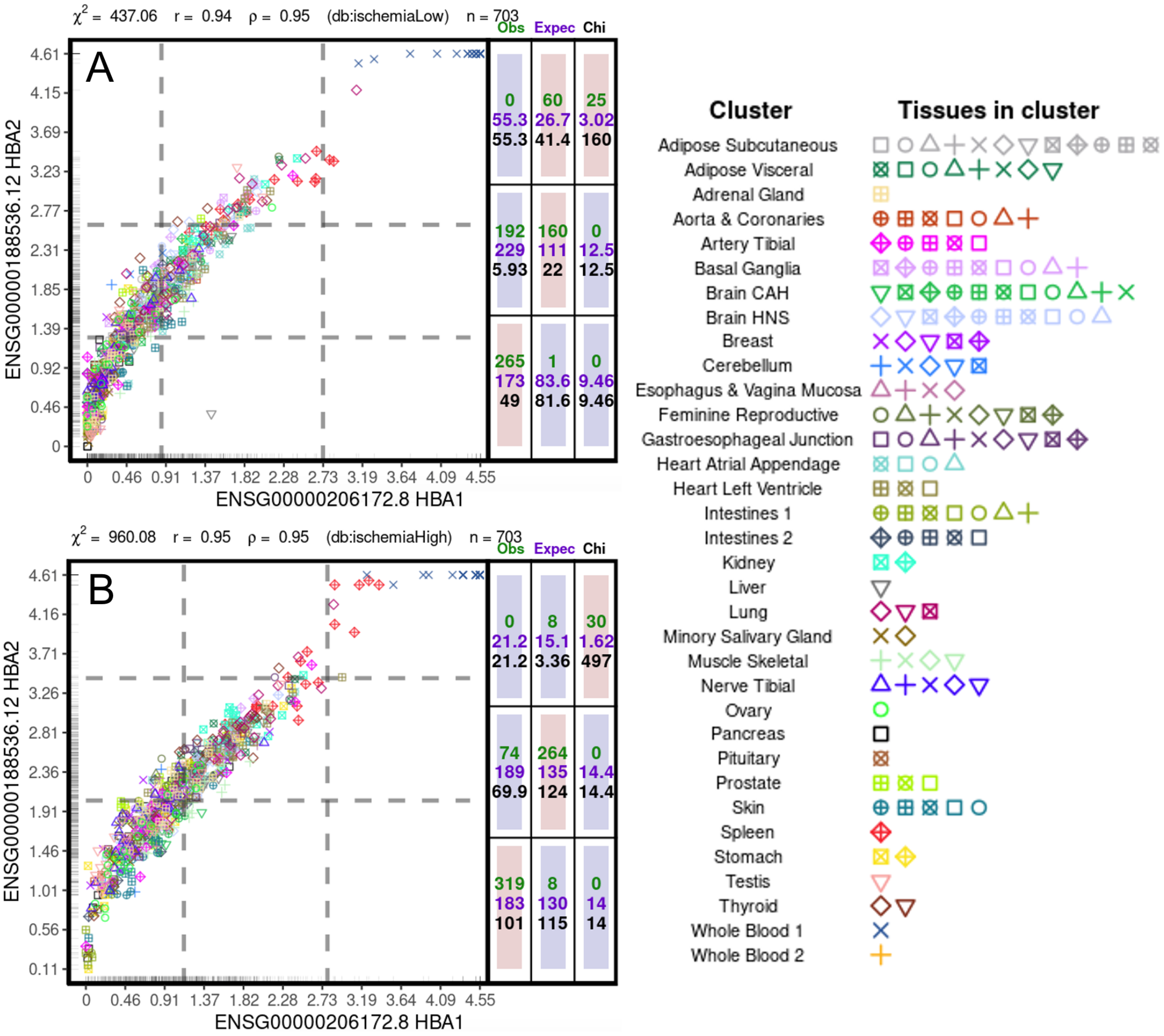


Figure 4.32: Example of a coexpression with a big difference (as quantified by Chi-square) between PMI categories. Hemoglobin subunit alpha 1 gene (HBA1) and hemoglobin subunit alpha 2 gene (HBA2). A: coexpression as observed in the 0-648 PMI group. B: coexpression as observed in the $\geq$ 648 PMI group.

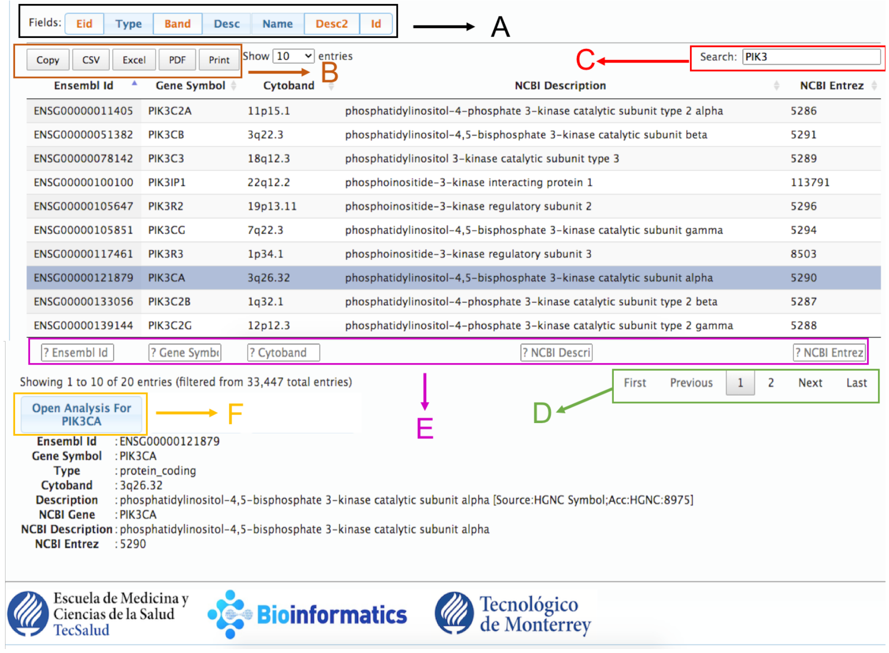
4.7 Results accessibility

As stated originally in the objectives of this work, one of the aspirations of the project was to share the results with the scientific community. This prevents others from having to go through the long process that has been documented throughout this master thesis to use the data. The intent is to provide other researchers with the possibility of using the data for their own analyses. A grand total of 57,611,420,643 individual calculations were made during this project after putting together the robust system-level coexpression estimation for Chi-Square, PCC, and SRCC across 20 realizations of the BRS algorithm, the estimation of the contribution of the gene expression states to individual coexpressions, and the coexpressions by covariate levels.

To make all the aforementioned data available, a Java Server Pages (JSP) architecture created by Dr. Victor Treviño (advisor of this master thesis) was used. The data was parsed and organized into a database queryable format by the author of this thesis. The R programs that can be called from JSP for in-web visualizations of coexpression data were also created by the author of this thesis. A description of the features of this web tool is provided in this section.

4.7.1 Web tool navigation

The web tool, which is available at `http://bioinformatica.mty.itesm.mx:8080/Chi2Expression/index.jsp`, hosts an easy to understand interface that is based on a gene search page (see Figure 4.33). This page acts as the home of the database and allows users to search for any gene of interest. Searches can be made by gene symbol, ENTREZ identifier [118], ENSEMBL identifier [119], cytoband [134] or gene description. Once a gene of interest has been selected, an option to retrieve its list of coexpressed genes will appear.



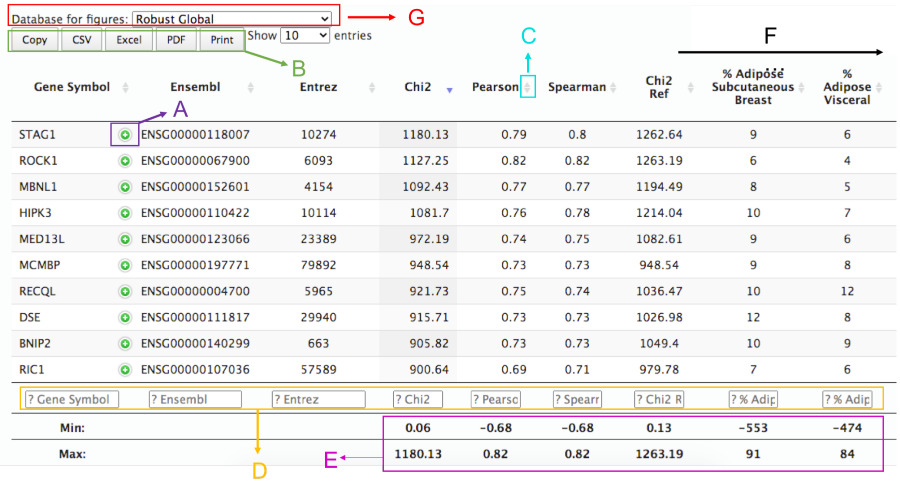
Ensembl Id	Gene Symbol	Cytoband	NCBI Description	NCBI Entrez
ENSG00000011405	PIK3C2A	11p15.1	phosphatidylinositol-4-phosphate 3-kinase catalytic subunit type 2 alpha	5286
ENSG00000051382	PIK3CB	3q22.3	phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit beta	5291
ENSG00000078142	PIK3C3	18q12.3	phosphatidylinositol 3-kinase catalytic subunit type 3	5289
ENSG00000100100	PIK3IP1	22q12.2	phosphoinositide-3-kinase interacting protein 1	113791
ENSG00000105647	PIK3R2	19p13.11	phosphoinositide-3-kinase regulatory subunit 2	5296
ENSG00000105851	PIK3CG	7q22.3	phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit gamma	5294
ENSG00000117461	PIK3R3	1p34.1	phosphoinositide-3-kinase regulatory subunit 3	8503
ENSG00000121879	PIK3CA	3q26.32	phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha	5290
ENSG00000133056	PIK3C2B	1q32.1	phosphatidylinositol-4-phosphate 3-kinase catalytic subunit type 2 beta	5287
ENSG00000139144	PIK3C2G	12p12.3	phosphatidylinositol-4-phosphate 3-kinase catalytic subunit type 2 gamma	5288

Showing 1 to 10 of 20 entries (filtered from 33,447 total entries)

Ensembl Id : ENSG00000121879
Gene Symbol : PIK3CA
Type : protein_coding
Cytoband : 3q26.32
Description : phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha [Source:HGNC Symbol;Acc:HGNC:8975]
NCBI Gene : PIK3CA
NCBI Description : phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha
NCBI Entrez : 5290

Figure 4.33: Web tool gene search page. A: choosing columns to be shown in the gene search table. B: download options. C: general bar searching every column. D: tabs to navigate the genes returned by the search. E: column-specific search bars. F: button to open the gene list of the gene of interest once highlighted in the search table. In this example the highlighted row is the phosphatidylinositol-4,5-biphosphate 3-kinase catalytic subunit alpha gene (PIK3CA).

In the gene list page (see Figure 4.34), users will be presented with all the coexpression results for the selected gene. This means that there is access to all estimations in which the selected gene participates (33,446 for each gene) regardless of if they comprise strong associations or not. Gene lists are sorted decreasingly by default according to the Chi-square value obtained in the robust minimum estimation (2.6), but they may be re-sorted decreasingly or increasingly by any of the columns in the gene list.



Database for figures: Robust Global

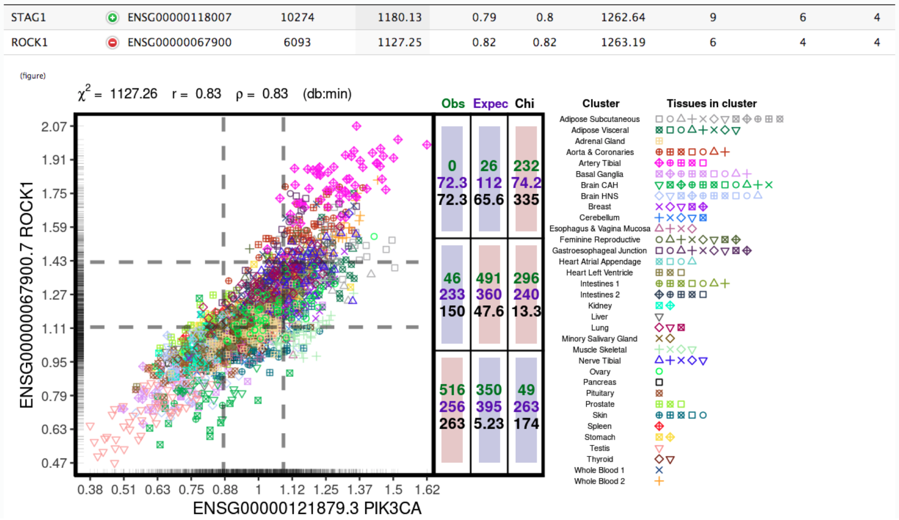
Show 10 entries

Gene Symbol		Ensembl	Entrez	Chi2	Pearson	Spearman	Chi2 Ref	% Adipose Subcutaneous Breast	% Adipose Visceral
STAG1	⊕	ENSG00000118007	10274	1180.13	0.79	0.8	1262.64	9	6
ROCK1	⊕	ENSG00000067900	6093	1127.25	0.82	0.82	1263.19	6	4
MBNL1	⊕	ENSG00000152601	4154	1092.43	0.77	0.77	1194.49	8	5
HIPK3	⊕	ENSG00000110422	10114	1081.7	0.76	0.78	1214.04	10	7
MED13L	⊕	ENSG00000123066	23389	972.19	0.74	0.75	1082.61	9	6
MCMBP	⊕	ENSG00000197771	79892	948.54	0.73	0.73	948.54	9	8
RECQL	⊕	ENSG00000004700	5965	921.73	0.75	0.74	1036.47	10	12
DSE	⊕	ENSG00000111817	29940	915.71	0.73	0.73	1026.98	12	8
BNIP2	⊕	ENSG00000140299	663	905.82	0.73	0.73	1049.4	10	9
RIC1	⊕	ENSG00000107036	57589	900.64	0.69	0.71	979.78	7	6
? Gene Symbol		? Ensembl	? Entrez	? Chi2	? Pearsc	? Spearr	? Chi2 R	? % Adip	? % Adip
Min:				0.06	-0.68	-0.68	0.13	-553	-474
Max:				1180.13	0.82	0.82	1263.19	91	84

Figure 4.34: Web tool gene list page (PIK3CA example). A: inspect coexpression button. It plots a contingency scatter plot (4.3). B: download options. C: buttons to sort by desired column. D: column-specific search bars. These support expressions such as greater than or equal, lesser, etc. Multiple filters can be applied at once. E: summary statistics. F: the gene list has many more columns additional to the ones shown here. They contain the information of the contributions to the coexpressions by the rest of the clusters, the estimation by covariates, and a column of differences between covariate levels for gender, age, and PMI. G: dropdown menu to switch the source of the contingency scatter plots. All estimations are available for plotting including the robust minimum, permuted clusters, and covariate estimations. Each one of the realizations of BRS is also available for visualization (4.1).

4.7.2 Built-in functions

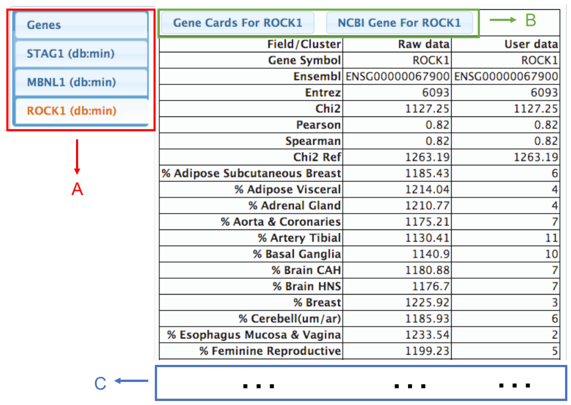
The main goal of the web tool is to provide the coexpression data to interested users by allowing them to download individual gene lists or the entirety of data in bulk depending on their needs. However, a couple of handy functions that are nice to have inside the webpage to quickly explore the data were also added to the web-tool.



STAG1	⊕	ENSG00000118007	10274	1180.13	0.79	0.8	1262.64	9	6	4
ROCK1	⊖	ENSG00000067900	6093	1127.25	0.82	0.82	1263.19	6	4	4

Figure 4.35: Example of in-web tool contingency scatter plot visualization of PIKCA3 and one of its top coexpressed gene pairs, the rho associated coiled-coil containing protein kinase 1 (ROCK1).

The first handy feature consists in being to visualize any coexpression with the help of the contingency scatter plots designed in this work (4.3). A program was designed specifically for this that can compute the plots on the fly in an efficient manner upon user request from any data source shown in the columns of the gene table. An example of how it looks embedded in the gene list page can be seen in Figure 4.35. The second handy feature consists of a queue list which is automatically built with the genes that the user has clicked in the gene list page (see Figure 4.36). The list allows for quickly navigating the plots produced so far and also features external links to gene information databases to check the full description of the genes, functions, genomic context, etc. A summary table for the pair of genes in question is also provided.



Field/Cluster	Raw data	User data
Gene Symbol	ROCK1	ROCK1
Ensembl	ENSG00000067900	ENSG00000067900
Entrez	6093	6093
Chi2	1127.25	1127.25
Pearson	0.82	0.82
Spearman	0.82	0.82
Chi2 Ref	1263.19	1263.19
% Adipose Subcutaneous Breast	1185.43	6
% Adipose Visceral	1214.04	4
% Adrenal Gland	1210.77	4
% Aorta & Coronaries	1175.21	7
% Artery Tibial	1130.41	11
% Basal Ganglia	1140.9	10
% Brain CAH	1180.88	7
% Brain HNS	1176.7	7
% Breast	1225.92	3
% Cerebell(um/ar)	1185.93	6
% Esophagus Mucosa & Vagina	1233.54	2
% Feminine Reproductive	1199.23	5

Figure 4.36: Example queue list in gene list pages of the web tool. A: clickable list for quick navigation of the genes selected so far in the gene list as shown in Figure 4.34. B: external links to detailed information of the gene pair of interest (in this case a top coexpressed pair of PIKC3A called ROCK1). C: continuation of the summary table shown holds the data for the remaining clusters and covariates. Note that both the $permScore$ metric defined in Equation 4.1 (User data) can be seen as well as the ”raw” Chi-square. These raw values are obtained from the cluster permutation experiments to estimate their contributions to the reference statistic (Chi2 Ref).

Chapter 5

Discussion and conclusions

In this final chapter, important results and observations regarding the work done are discussed. Thoughts on future work related to this research are also addressed as many potential interesting investigations can derive from the work done so far. Final remarks on the project are also given.

5.1 Discussion

In this work, a robust process for estimating large-scale system-level human coexpression using computational and statistical methods was performed and described. The project was mainly motivated by areas of opportunity in the domain of large-scale coexpression projects. No resources with results focused on normal human coexpression that used RNA-seq as the source of their input data existed before the realization of this work. The discussion presented here is thought under the light of the results produced for each objective introduced in Chapter 1.

The first results obtained were related to properly characterizing the input gene expression data intended to enter the coexpression calculation process. It was observed that GTEx tissues initially featured highly similar samples that were labeled differently (e.g. Adipose Subcutaneous and Brest samples) as well as dissimilar samples that were labeled the same (e.g. Whole Blood). The differences and similarities between samples were better described by a clustering partition found in this work that allowed to conceptualize the GTEx samples in terms of less redundant gene expression states or clusters (3.2). The partition of samples found was more suited for the downstream analyses when compared to the original GTEx samples tissue-based partition. However, it was noticeable at least in t-SNE space that there were still some sample clustering assignments that could potentially be improved. This could be attempted with advanced clustering validation techniques such as co-association matrices [135] or the Validity Index using supervised Classifiers (VIC) [136]. It is not expected, however, that downstream analyses vary much with slight improvements of the clustering partition.

Finding well-defined and non-redundant gene expression states was important in this work because the goal was to produce a coexpression estimation which was representative of humans as a whole. By identifying groups of highly similar samples, it was made sure that no particular sample type was overrepresented during coexpression calculation. Without defining the groups, doing this was imprecise and difficult. Not doing it would bias the results towards what is observed only in the overrepresented subset of samples. Several strategies were appropriate for addressing this problem once the groups of samples were defined. To decide which one allowed for a more adequate discovery of coexpressions, experiments based on biological pathways were carried out (3.3). Adequate was defined as a balance between a method's ability to identify coexpressions that are highly likely to be true based on documented biological pathways/gene sets and the method's total discovered coexpressions. As some gene expression data transformations were also considered as variants of the methods, these were also tested taking into account if they disrupted the initial gene expression states defined before coexpression calculation.

Once the coexpression estimation was completed, one of the main findings was that the Chi-square statistic is reliable as a metric to estimate gene coexpression in a large-scale setting. This observation is supported by the fact that the statistic is correlated with the PCC (4.4.1.1), a metric which is the standard measure used in many works including large-scale and smaller coexpression analyses. When looking at the most strongly coexpressed pairs for the majority of genes, which are usually the pairs of interest in coexpression analysis, there is an agreement of about 65% between Chi-square and PCC when comparing the top 100 coexpressed pairs found by each metric (4.4.2). Even in the cases where the percentages of agreement were smaller, they were still statistically significant by an FDR-corrected hypergeometric test for 33,445 genes out of the 33,447 that were considered in this work. Since the Chi-sqaure statistic is reliable, one can enjoy the additional benefits it offers in terms of interpretability of individual coexpression relationships 4.3.

Despite identifying that the Chi-square and PCC yielded similar results on average, it was also shown that there exist certain genes for which the correlation between the metrics is noticeably smaller (4.4.1.1). These genes seemed to exhibit some degree of tissue or cluster-specificity, thus making associations in which they are involved more complex to describe. Since these kinds of genes only get a chance to be coexpressed with other genes in the tissues or clusters in which they are expressed in the first place, it is frequent to encounter patterns in the coexpression scatter plot that are difficult to characterize only with linear functions. In such cases, the Chi-square was much more sensitive to detecting associations driven by these tissue or cluster-specific genes.

Results also showed that in terms of top coexpressed genes with each gene, the SRCC was the metric that produced the most different results when compared to the other two metrics. It was particularly interesting to find that the pair Chi-square/PCC agreed more on what were the top coexpressed pairs of each gene compared to the pair PCC/SRCC. Broadly speaking, the PCC and SRCC usually produce very similar results in the literature outside coexpression [125]. The latter applies to input data with certain distributional assumptions. In practice, while dealing with the complexity of real biological data, there were several genes for which the SRCC and PCC did not manage to agree on with statistical significance (4.4.2). It was also observed that these genes with different top coexpressed pairs depending on if they were measured with PCC and SRCC seemed to be tissue or cluster-specific. The agreement between Chi-square/SRCC was even less than that of PCC/SRCC (4.4.2).

In Chapter 2 it was mentioned that the significance of the Chi-square statistic can be assessed by looking at the random Chi-square distribution with the appropriate parameters (2.5.1). Initially, it was anticipated that it would be possible to do this with the calculated coexpressions to assess how many genes were truly coexpressed with each gene. However, once the calculations were complete, it was discovered that the resulting Chi-square distributions of observed values were highly skewed. The robust minimum distribution featured a whooping percentage of 80% coexpressions (447,467,345 out of 559,334,181) that would be interesting by the statistical significance criteria. This was mentioned at the moment to be unrealistic (4.2). Small additional experiments were made to investigate this further and two sources impacting the shape of the calculated distributions were identified:

- Tissue-specific genes as seen in Figure 5.1
- Sample size as seen in Figure 5.2

The tissue-specific genes observation is obviously a particularity of coexpression analysis, but the sample size effect is domain-independent. The issue is well-characterized not just for the Chi-square, but for many statistical tests [137]. In instances where the sample size is large, the tests still behave as they are supposed. It can therefore be said that the found associations do exist however subtle. Thanks to the sample size, the tests will be powerful enough that they can detect very mild associations. This does not mean that all these statistically significant associations are useful to know or are practical in some way (in fact the great majority of them will not be). Practical significance must be considered along with statistical significance in these cases.

Regarding the estimation of the contribution of gene expression states to individual coexpressions (4.5), Algorithm 3 was a nice first approximation for solving the problem since it allowed for easily identifying coexpressions driven by a strong cluster-specific component even inside the web-tool created. However, applying the algorithm to some instances in which the permuted cluster was not important for the particular coexpression in question would result in an increase of the Chi-square statistic when compared to the reference (a realization without any permuted clusters). This was explainable due to the way expected counts changed upon permutation and could be interpreted as the cluster in question not being important for the coexpression. Nevertheless, it is thought that it can be confusing to other researchers consulting the data. It is believed that there should be a metric that is more robust for these instances.

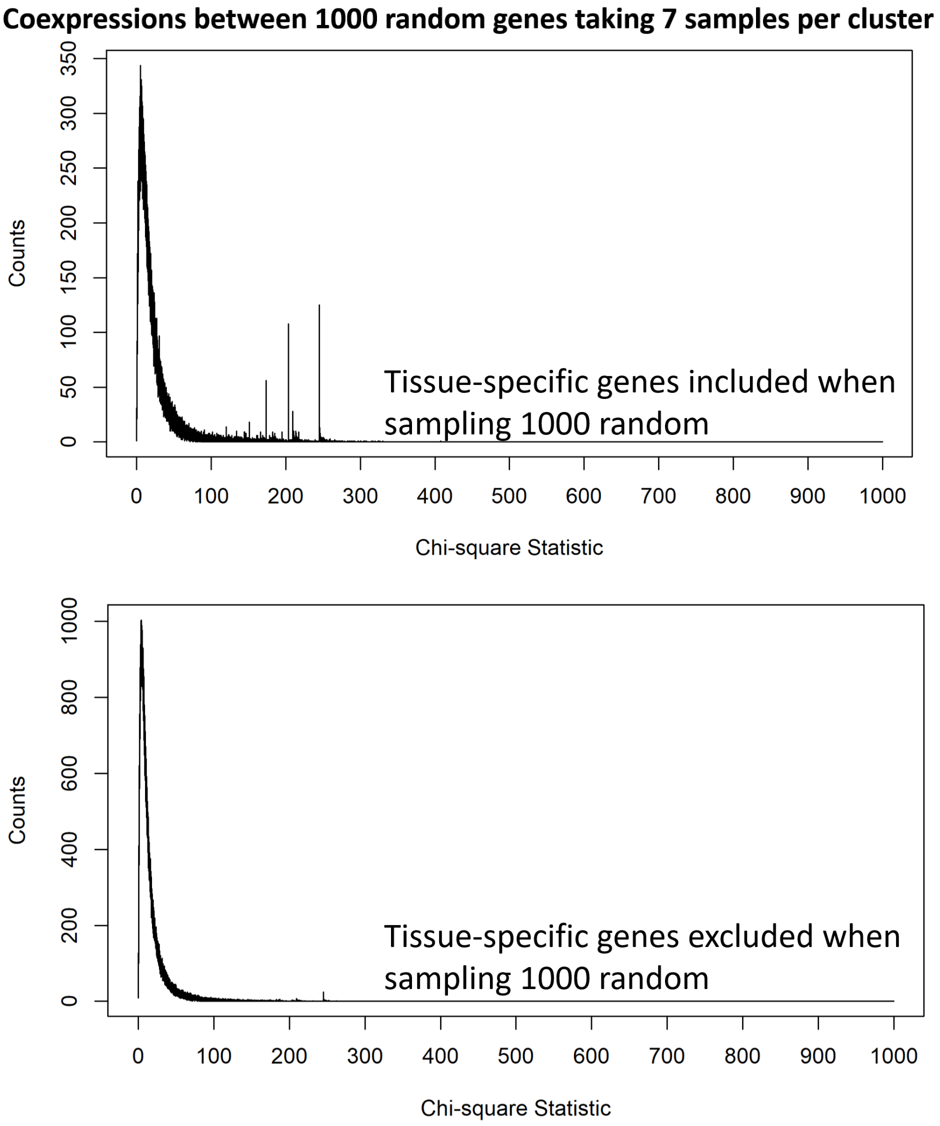


Figure 5.1: Removing tissue-specific genes from a quick estimation using a representative random sample of 1000 genes and 7 samples per gene expression state made the resulting distribution less dense for strong Chi-square values (2.2).

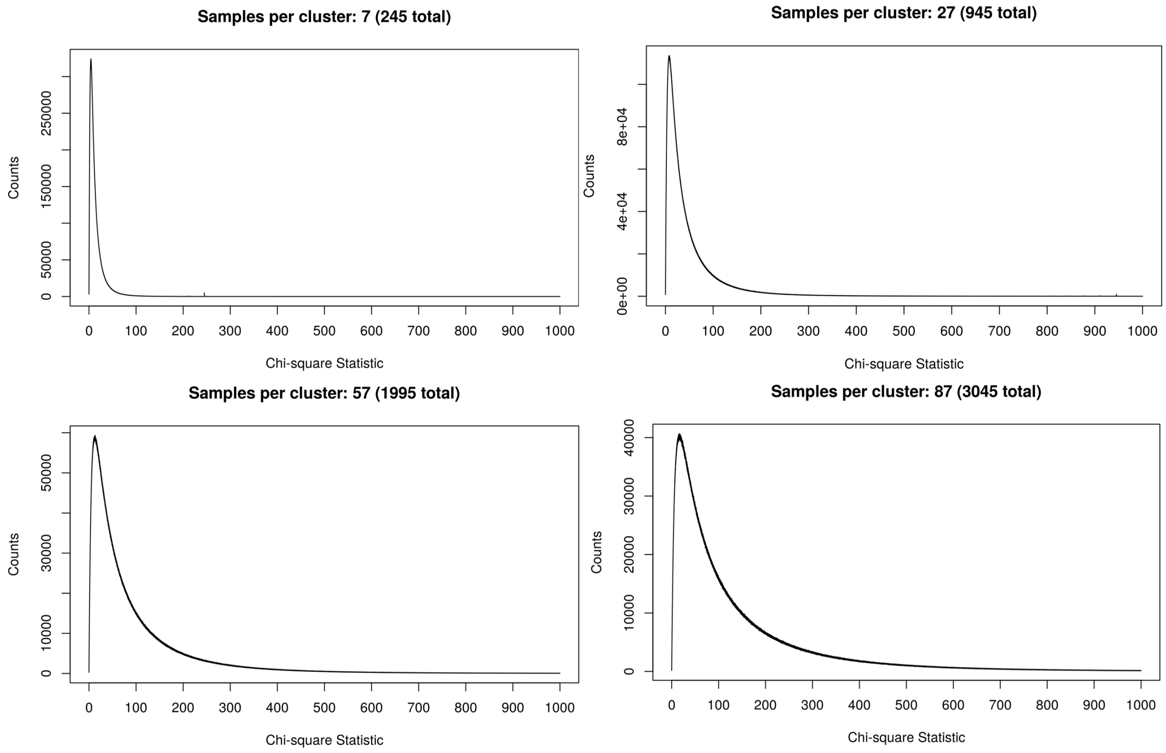


Figure 5.2: Increasing the number of samples considered for a quick estimation on a constant representative random sample of 1000 genes causes the Chi-square distribution to gain density for stronger Chi-square values.

5.2 Future work

Due to the nature of the work carried out in this master thesis, there are many research questions and potential analyses that could stem from the data generated and the observations made. For instance, during the year this work was completed, a series of publications addressing specifically the GTEx data were published in the *Science* journal. Important topics such as the role of gender in gene expression across tissues were investigated in these papers [131]. In another work, phenome-wide and genome-wide association studies combined with gene expression data were used to predict the outcome of clinical trials for new drugs. This represents an important step towards reducing unnecessary time and financial expenses caused by failed trials [138]. The idea is that all these kinds of analyses that are possible using gene expression data can be extended to their gene coexpression versions. These analyses would be able to take advantage of the extra information that coexpression yields regarding interactions between genes. The complexity of the analyses will go up for sure, but it may be possible to come to conclusions that otherwise would remain unexplored by only using gene expression data.

Some promising ideas that have been on the table as a continuation of the work done here comprise the creation of a genomic coexpression map where the data can be studied in the context of neighboring genes in the human genome at a DNA sequence level. In general, it is thought that genes with similar expression profiles and that need to be expressed at the same time tend to be clustered together in the human genome [139]. The data produced in this work can help investigate this further with good coverage across the human genome. Another interesting possibility lies in classifying coexpression types in terms of the scatter plot patterns describing them. There are many examples of linear coexpression relationships between genes, but what about other patterns such as exponential, crossed lines, or even non-functional associations such as no-coexistence? [23]. It would be very interesting to see what patterns may be uncovered from a vast collection of coexpressions present in normal human biology such as the one created here.

Future work could also be sought from a wet lab validation point of view. This means looking for some promising and previously undescribed gene pairs that could be interesting for a better understanding of some biological function. The prioritization of these gene pairs would be guided by the data estimated in this work. Therefore this would be a way to start using the knowledge gained from this computational project and apply it to a molecular biology experiment. Dr. Victor Treviño and the author of this thesis continue to work on the web tool presented in this project to facilitate the exploration of the coexpression data by other researchers and favor the translation of the knowledge produced to practical applications (4.7).

Regarding the future of large-scale coexpression computational projects, a natural continuation of the same line of thought presented here would be to try to compute the coexpression data using MIC [23]. This is the metric that originally inspired the use of a grid-based strategy such as the Chi-square statistic in this master thesis work. There is the concern regarding how feasible this would be due to the more computationally intensive process that needs to be followed for MIC computation, but it is definitely worth the try as it could represent a more robust metric for coexpression while keeping some of the interpretability features that the Chi-square test has in the context of coexpression.

5.3 Conclusions

The main contributions of this work can be summarized as follows:

- Characterization of non-redundant gene expression states in an important dataset of normal human tissues such as the GTEx project (3.2)

- Implementation of a strategy (BRS) for large-scale coexpression which prevents sample type biases in downstream calculations and that performs better than the traditionally used method in literature (weighted coexpression) for this purpose in a comparison based on biological pathways and reference gene sets (2.6.4, 3.3)

- Estimation of large-scale human coexpression data based exclusively on normal samples, based on data originating from RNA-seq, and calculated for three different coexpression metrics (Chapter 4). All of this combined for the first time in the literature.

- Usage of the Chi-square statistic in the context of large-scale coexpression for the first time in the literature demonstrating its reliability and usefulness at the moment of interpreting individual coexpression relationships between genes (4.3, 4.4)

- Description and application of a permutation-based algorithm to estimate the contribution of specific types of samples to individual coexpressions calculated with the Chi-square test of independence (4.5)

- Sharing of all computed data with the scientific community through a web tool that incorporates highly comprehensive individual coexpression visualizations as well as download options for further analysis by interested researchers (4.7)

As a final remark of this master thesis project, the author wishes to express the successful fulfillment of the initially established project goals. This includes the assembly of a gene expression dataset suitable for downstream system-level coexpression calculations using different metrics and a robust method for ensuring that the results are representative of normal human coexpression associations.

Insight was provided into how different coexpression metrics compare to each other. The Chi-square statistic was used for the first time in the context of large-scale coexpression proving to be a valuable tool when interpreting individual coexpression associations. Other features related to coexpression relationships between genes were characterized including the importance of different gene expression states at the moment of calculating the observed coexpression statistics, as well as the potential effect that some covariates have on observed coexpressions.

Work culminated with the depositing of the results in a web tool that makes all estimations obtained publicly available and that is already available online. The advisor and the author of this work think that this tool will be very valuable to other researchers since it is the first resource of its kind to host results derived from normal human gene expression quantified with sequencing technologies. The work presented throughout this master thesis project comprises an important step towards a better understanding of interactions between human genes in normality from a systems biology point of view.

Bibliography

[1] Sam Behjati and Patrick S. Tarpey. What is next generation sequencing? *Archives of disease in childhood. Education and practice edition*, 98(6):236–238, Dec 2013.

[2] Zichen Wang, Alexander Lachmann, and Avi Ma'ayan. Mining data and metadata from the gene expression omnibus. *Biophysical Reviews*, 11, 12 2018.

[3] Juliana Costa-Silva, Douglas Domingues, and Fabricio Martins Lopes. Rna-seq differential expression analysis: An extended review and a software tool. *PLOS ONE*, 12(12):1–18, 12 2017.

[4] Jun Han, Meijun Chen, Yihan Wang, Boxuan Gong, Tianwei Zhuang, Lingyu Liang, and Hong Qiao. Identification of biomarkers based on differentially expressed genes in papillary thyroid carcinoma. *Scientific Reports*, 8, 12 2018.

[5] En-Guo Chen, Pin Wang, Haizhou Lou, Yunshan Wang, Hong Yan, Lei Bi, Liang Liu, Bin Li, Antoine Snijders, Jian-Hua Mao, and Bo Hang. A robust gene expression-based prognostic risk score predicts overall survival of lung adenocarcinoma patients. *Oncotarget*, 9, 12 2017.

[6] Bao-Hong Liu. *Differential Coexpression Network Analysis for Gene Expression Data*, pages 155–165. Springer New York, New York, NY, 2018.

[7] Joshua Stuart, Eran Segal, Daphne Koller, and Stuart Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science (New York, N.Y.)*, 302:249–55, 11 2003.

[8] Raina Robeva. Systems biology – old concepts, new science, new challenges. *Frontiers in psychiatry / Frontiers Research Foundation*, 1:1, 01 2010.

[9] Sipko van Dam, Urmo Võsa, Adriaan van der Graaf, Lude Franke, and João Pedro de Magalhães. Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in Bioinformatics*, 19(4):575–592, 01 2017.

[10] Y. Feng, J. Hurst, M. Almeida-De-Macedo, X. Chen, L. Li, N. Ransom, and E. S. Wurtele. Massive human co-expression network and its medical applications. *Chem Biodivers*, 9(5):868–887, May 2012.

[11] R. R. Nayak, M. Kearns, R. S. Spielman, and V. G. Cheung. Coexpression network based on natural variation in human gene expression reveals gene interactions and functions. *Genome Res*, 19(11):1953–1962, Nov 2009.

[12] H. K. Lee, A. K. Hsu, J. Sajdak, J. Qin, and P. Pavlidis. Coexpression analysis of human genes across many microarray data sets. *Genome Res*, 14(6):1085–1094, Jun 2004.

[13] Z. F. Gerring, E. R. Gamazon, and E. M. Derks. A gene co-expression network-based analysis of multiple brain tissues reveals novel genes and molecular pathways underlying major depression. *PLoS Genet*, 15(7):e1008245, 07 2019.

[14] Y. Yang, L. Han, Y. Yuan, J. Li, N. Hei, and H. Liang. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nat Commun*, 5:3231, 2014.

[15] Takeshi Obayashi, Yuki Kagaya, Yuichi Aoki, Shu Tadaka, and Kengo Kinoshita. Coxpresdb v7: a gene coexpression database for 11 animal species supported by 23 coexpression platforms for technical evaluation and evolutionary inference. *Nucleic acids research*, 47, 11 2018.

[16] Sipko Dam, Thomas Craig, and Joao Pedro de Magalhaes. Genefriends: A human rna-seq-based gene and transcript co-expression database. *Nucleic acids research*, 43, 10 2014.

[17] Ioannis Michalopoulos, Georgios Pavlopoulos, Apostolos Malatras, Alexandros Karelas, Myrto Kostadima, Reinhard Schneider, and Sophia Kossida. Human gene correlation analysis (hgca): A tool for the identification of transcriptionally co-expressed genes. *BMC research notes*, 5:265, 06 2012.

[18] Sunmo Yang, Chan Kim, Sohyun Hwang, Eiru Kim, Hyojin Kim, Hongseok Shim, and Insuk Lee. Coexpedia: Exploring biomedical hypotheses via co-expressions associated with medical subject headings (mesh). *Nucleic Acids Research*, 45, 09 2016.

[19] Daniel Jupiter, Hailin Chen, and Vincent VanBuren. Starnet 2: A web-based tool for accelerating discovery of gene regulatory networks using microarray co-expression data. *BMC bioinformatics*, 10:332, 10 2009.

[20] Tomas Hruz, Oliver Laule, Gábor Szabó, Frans Wessendorp, Stefan Bleuler, Lukas Oertle, Peter Widmayer, Wilhelm Gruissem, and Philip Zimmermann. Genevestigator v3: a reference expression database for the meta-analysis of transcriptomes. *Advances in bioinformatics*, 2008:420747, 07 2008.

[21] Rajeshwar Govindarajan, Jeyapradha Duraiyan, Karunakaran Kaliyappan, and Murugesan Palanisamy. Microarray and its applications. *Journal of pharmacy & bioallied sciences*, 4:S310–2, 08 2012.

[22] Joseph Lee Rodgers and W. Alan Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, 1988.

[23] David Reshef, Yakir Reshef, Hilary Finucane, Sharon Grossman, Gilean McVean, Peter Turnbaugh, Eric Lander, Michael Mitzenmacher, and Pardis Sabeti. Detecting novel associations in large data sets. *Science (New York, N.Y.)*, 334:1518–24, 12 2011.

[24] S. Roy, D. K. Bhattacharyya, and J. K. Kalita. Reconstruction of gene co-expression network from microarray data using local expression patterns. *BMC Bioinformatics*, 15 Suppl 7:S10, 2014.

[25] Dominic Allocco, Isaac Kohane, and Atul Butte. Quantifying the relationship between co-expression, co-regulation and gene function. *BMC bioinformatics*, 5:18, 03 2004.

[26] Esra Gov (Korurer) and Kazim Arga. Differential co-expression analysis reveals a novel prognostic gene module in ovarian cancer. *Scientific Reports*, 7, 07 2017.

[27] Sipko van Dam, Rui Cordeiro, Thomas Craig, Jesse van Dam, Shona H. Wood, and João Pedro de Magalhães. Genefriends: An online co-expression analysis tool to identify novel gene targets for aging and complex diseases. *BMC Genomics*, 13(1):535, 10 2012.

[28] Zhenhong Jiang, Xiaobao Dong, Zhigang Li, Fei He, and Ziding Zhang. Differential coexpression analysis reveals extensive rewiring of arabidopsis gene coexpression in response to pseudomonas syringae infection. *Scientific Reports*, 6:35064, 10 2016.

[29] Sara Ballouz, Melanie Weber, Paul Pavlidis, and Jesse Gillis. Egad: Ultra-fast functional analysis of gene networks. *Bioinformatics (Oxford, England)*, 33, 12 2016.

[30] Alexander Platzer, Thomas Nussbaumer, Thomas Karonitsch, Josef S. Smolen, and Daniel Aletaha. Analysis of gene expression in rheumatoid arthritis and related conditions offers insights into sex-bias, gene biotypes and co-expression patterns. *PLOS ONE*, 14(7):1–23, 07 2019.

[31] Zhizhu su, Paweł Łabaj, Sheng Li, Jean Thierry-Mieg, Danielle Thierry-Mieg, Wei Shi, Chun Wang, Gary Schroth, Robert Setterquist, John Thompson, Wendell Jones, Wenzhong Xiao, Weihong Xu, Roderick Jensen, Reagan Kelly, Joshua Xu, Ana Conesa, Cesare Furlanello, Hanlin Gao, and Leming Shi. A comprehensive assessment of rna-seq accuracy, reproducibility and information content by the sequencing quality control consortium. *Nature Biotechnology*, 32, 08 2014.

[32] A Ugoni and Bruce Walker. The chi square test: an introduction. *COMSIG review / COMSIG, Chiropractors and Osteopaths Musculo-Skeletal Interest Group*, 4:61–4, 12 1995.

[33] C Spearman. The proof and measurement of association between two things. *International Journal of Epidemiology*, 39(5):1137–1150, 10 2010.

[34] Donald Sharpe. Your chi-square test is statistically significant: Now what? *Practical Assessment, Research and Evaluation*, 20:1–10, 01 2015.

[35] Guogen Shan and Shawn Gerstenberger. Fisher's exact approach for post hoc analysis of a chi-squared test. *PLOS ONE*, 12(12):1–12, 12 2017.

[36] Robert Yang, Jie Quan, Reza Sodaei, Francois Aguet, Ayellet Segre, John Allen, Thomas Lanz, Veronica Reinhart, Matthew Crawford, Samuel Hasson, Kristin Ardlie, Roderic Guigo, and Hualin Xi. A systematic survey of human tissue-specific gene expression and splicing reveals new opportunities for therapeutic target identification and evaluation. *bioRxiv*, 04 2018.

[37] Takeshi Obayashi, Yuki Kagaya, Yuichi Aoki, Shu Tadaka, and Kengo Kinoshita. Coexviewer. `https://coxpresdb.jp/coexview/?geneID1=7535&geneID2=919`, 2020. Accessed: 2020-19-02.

[38] J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia, N. Young, B. Foster, M. Moser, E. Karasik, B. Gillard, K. Ramsey, S. Sullivan, J. Bridge, H. Magazine, J. Syron, J. Fleming, et al. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, 45(6):580–585, 7 2013.

[39] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

[40] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[41] Michel Morange. The central dogma of molecular biology. *Resonance*, 14:236–247, 03 2009.

[42] Adam Frankish, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M Mudge, Cristina Sisu, James Wright, Joel Armstrong, If Barnes, Andrew Berry, Alexandra Bignell, Silvia Carbonell Sala, Jacqueline Chrast, Fiona Cunningham, Tomás Di Domenico, Sarah Donaldson, Ian T Fiddes, Carlos García Girón, et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*, 47(D1):D766–D773, 10 2018.

[43] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: A revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10:57–63, 12 2008.

[44] Rajandeep Sekhon, Roman Briskine, Candice Hirsch, Chad Myers, Nathan Springer, C Buell, Natalia Leon, and Shawn Kaeppler. Maize gene atlas developed by rna sequencing and comparative evaluation of transcriptomes based on rna sequencing and microarrays. *PloS one*, 8:e61005, 04 2013.

[45] Shanrong Zhao, Wai-Ping Fung-Leung, Anton Bittner, Karen Ngo, and Xuejun Liu. Comparison of rna-seq and microarray in transcriptome profiling of activated t cells. *PLOS ONE*, 9(1):1–13, 01 2014.

[46] Alicia Oshlack and Matthew Wakefield. Transcript length bias in rna-seq data confounds systems biology. *Biology direct*, 4:14, 05 2009.

[47] Gunter Wagner, Koryu Kin, and Vincent Lynch. Measurement of mrna abundance using rna-seq data: Rpkm measure is inconsistent among samples. *Theory in biosciences*, 131, 08 2012.

[48] Zachary Abrams, Travis Johnson, Kun Huang, Philip Payne, and Kevin Coombes. A protocol to evaluate rna sequencing normalization methods. *BMC Bioinformatics*, 20, 12 2019.

[49] Isabella Zwiener, Barbara Frisch, and Harald Binder. Transforming rna-seq data to improve the performance of prognostic gene signatures. *PLOS ONE*, 9(1):1–13, 01 2014.

[50] Stephanie Hicks and Rafael Irizarry. quantro: A data-driven approach to guide the choice of an appropriate normalization method. *Genome biology*, 16:117, 06 2015.

[51] B Bolstad, RA Irizarry, Magnus Åstrand, and T Speed. A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Bioinformatics (Oxford, England)*, 19:185–93, 02 2003.

[52] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, 41(Database issue):D991–995, 1 2013.

[53] M. Suntsova, N. Gaifullin, D. Allina, A. Reshetun, X. Li, L. Mendeleeva, V. Surin, A. Sergeeva, P. Spirin, V. Prassolov, A. Morgan, A. Garazha, M. Sorokin, and A. Buzdin. Atlas of RNA sequencing profiles for normal human tissues. *Sci Data*, 6(1):36, 04 2019.

[54] Yi Wang, Stephanie C. Hicks, and Kasper D. Hansen. Co-expression analysis is biased by a mean-correlation relationship. *bioRxiv*, 2020.

[55] P. Langfelder and Steve Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC Bioinformatics*, 9:1–13, 01 2008.

[56] Charu Aggarwal and Haixun Wang. *A Survey of Clustering Algorithms for Graph Data*, volume 40, pages 275–301. Springer, 02 2010.

[57] A.e.a Subramanian. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. proc natl acad sci u s a. *Proceedings of the National Academy of Sciences*, 102:15545–15550, 10 2005.

[58] Elizabeth Boyle, Shuai Weng, Jeremy Gollub, Heng Jin, David Botstein, J Cherry, and Gavin Sherlock. Go::termfinder - open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics (Oxford, England)*, 20:3710–5, 01 2005.

[59] Xavier Teitsma, Johannes Jacobs, Michal Mokry, Michelle Borm, Attila Pethö-Schramm, Jacob Laar, Johannes Bijlsma, and Floris Lafeber. Identification of differential co-expressed gene networks in early rheumatoid arthritis achieving sustained drug-free remission after treatment with a tocilizumab-based or methotrexate-based strategy. *Arthritis Research & Therapy*, 19, 12 2017.

[60] Elissa Cosgrove, Timothy Gardner, and Eric Kolaczyk. On the choice and number of microarrays for transcriptional regulatory network inference. *BMC bioinformatics*, 11:454, 09 2010.

[61] Samuel A. Lambert, Arttu Jolma, Laura F. Campitelli, Pratyush K. Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R. Hughes, and Matthew T. Weirauch. The human transcription factors. *Cell*, 172(4):650–665, 2 2018.

[62] Takeshi Obayashi, Shinpei Hayashi, Masayuki Shibaoka, Motoshi Saeki, Hiroyuki Ohta, and Kengo Kinoshita. Coxpresdb: A database of coexpressed gene networks in mammals. *Nucleic acids research*, 36:D77–82, 02 2008.

[63] Takeshi Obayashi and Kengo Kinoshita. Coxpresdb: a database to compare gene coexpression in seven model animals. *Nucleic acids research*, 39(Database issue):D1016–D1022, 1 2011.

[64] Takeshi Obayashi, Yasunobu Okamura, Satoshi Ito, Shu Tadaka, Ikuko Motoike, and Kengo Kinoshita. Coxpresdb: A database of comparative gene coexpression networks of eleven species for mammals. *Nucleic acids research*, 41, 11 2012.

[65] Yasunobu Okamura, Yuichi Aoki, Takeshi Obayashi, Shu Tadaka, Satoshi Ito, Takafumi Narise, and Kengo Kinoshita. Coxpresdb in 2015: Coexpression database for animal species by dna-microarray and rnaseq-based expression data with multiple quality assessment systems. *Nucleic acids research*, 43, 11 2014.

[66] M. Kanehisa and S. Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 1 2000.

[67] M. Ashburner, C.A. Ball, Judith Blake, David Botstein, Heather Butler, and J. Cherry. Gene ontology: Tool for the unification of biology. *The Gene Ontology Consortium. Nat Genet*, 25:25–29, 01 2000.

[68] Takeshi Obayashi, Yuki Kagaya, Yuichi Aoki, Shu Tadaka, and Kengo Kinoshita. Expression viewer. `https://coxpresdb.jp/exview/?geneID=919`, 2020. Accessed: 2020-09-02.

[69] Takeshi Obayashi and Kengo Kinoshita. Rank of correlation coefficient as a comparable measure for biological significance of gene coexpression. *DNA research : an international journal for rapid publication of reports on genes and genomes*, 16:249–60, 09 2009.

[70] Ian Jolliffe and Jorge Cadima. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374:20150202, 04 2016.

[71] Rasko Leinonen, Hideaki Sugawara, and Martin Shumway. The sequence read archive. *Nucleic acids research*, 39:D19–21, 11 2010.

[72] Sean Simmons, Jian Peng, Jadwiga Bienkowska, and Bonnie Berger. Discovering what dimensionality reduction really tells us about rna-seq data. *Journal of computational biology : a journal of computational molecular cell biology*, 22, 06 2015.

[73] Da Wei (David) Huang, Brad Sherman, and Richard Lempicki. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature Protocols*, 4:44–57, 12 2008.

[74] Luman Wang, Qiaochu Mo, and Jianxin Wang. Mirexpress: A database for gene coexpression correlation in immune cells based on mutual information and pearson correlation. *Journal of immunology research*, 2015:140819–140819, 2015.

[75] Jesse Hoey. The two-way likelihood ratio (g) test and comparison to two-way chi squared test. *arXiv*, 06 2012.

[76] A. Battle, C. D. Brown, B. E. Engelhardt, S. B. Montgomery, F. Aguet, K. G. Ardlie, B. B. Cummings, E. T. Gelfand, G. Getz, K. Hadley, R. E. Handsaker, K. H. Huang, S. Kashin, K. J. Karczewski, M. Lek, X. Li, D. G. MacArthur, J. L. Nedzel, D. T. Nguyen, M. S. Noble, et al. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213, 10 2017.

[77] GTEx Consortium. The gtex project documentation webpage. `https://www.gtexportal.org/home/documentationPage`, 2019. Accessed: 2020-09-14.

[78] GTEx Consortium. The gtex project frequently asked questions. `https://www.gtexportal.org/home/documentationPage`, 2019. Accessed: 2020-09-14.

[79] Y. Sha, J. H. Phan, and M. D. Wang. Effect of low-expression gene filtering on detection of differentially expressed genes in RNA-seq data. *Conf Proc IEEE Eng Med Biol Soc*, 2015:6461–6464, 2015.

[80] The GTEx Consortium. The gtex consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330, 2020.

[81] Cian Glenfield and Aoife McLysaght. Pseudogenes provide evolutionary evidence for the competitive endogenous rna hypothesis. *Molecular biology and evolution*, 35, 09 2018.

[82] C. Park, Tanima de, Y. Xu, Y. Zhong, Erin Smithberger, C. Alarcon, Eric Gamazon, and Minoli Perera. Hepatocyte gene expression and dna methylation as ancestry-dependent mechanisms in african americans. *npj Genomic Medicine*, 4, 12 2019.

[83] Pradipta Ray, Jawad Khan, Andi Wangzhou, Diana Tavares-Ferreira, Armen Akopian, Gregory Dussor, and Theodore Price. Transcriptome analysis of the human tibial nerve identifies sexually dimorphic expression of genes involved in pain, inflammation, and neuro-immunity. *Frontiers in Molecular Neuroscience*, 12:37, 03 2019.

[84] Xiaoguang Xu, James M. Eales, Artur Akbarov, Hui Guo, Lorenz Becker, David Talavera, Fehzan Ashraf, Jabran Nawaz, Sanjeev Pramanik, John Bowes, Xiao Jiang, John Dormer, Matthew Denniff, Andrzej Antczak, Monika Szulinska, Ingrid Wise, Priscilla R. Prestes, Maciej Glyda, Pawel Bogdanski, Ewa Zukowska-Szczechowska, Carlo Berzuini, Adrian S. Woolf, Nilesh J. Samani, Fadi J. Charchar, and Maciej Tomaszewski. Molecular insights into genome-wide association studies of chronic kidney disease-defining traits. *Nature Communications*, 9(1):4800, Nov 2018.

[85] G. Box and David Cox. An analysis of transformations (with discussion). *Journal of the Royal Statistical Society, Series B*, 26:211–252, 01 1964.

[86] Laurens van der Maaten and Geoffrey Hinton. Viualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 11 2008.

[87] Erich Schubert and Michael Gertz. Intrinsic t-stochastic neighbor embedding for visualization and outlier detection. In *Similarity Search and Applications*, pages 188–203, 10 2017.

[88] Amit Saxena, Mukesh Prasad, Akshansh Gupta, Neha Bharill, op Patel, Aruna Tiwari, Meng Er, Weiping Ding, and Chin-Teng Lin. A review of clustering techniques and developments. *Neurocomputing*, 267, 07 2017.

[89] Joe H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

[90] Vincent Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics Theory and Experiment*, 2008, 04 2008.

[91] Karthik Shekhar, Sylvain W Lapan, Irene E Whitney, Nicholas M Tran, Evan Macosko, Monika Kowalczyk, Xian Adiconis, Joshua Z Levin, James Nemesh, Melissa Goldman, Steven Mccarroll, Constance L Cepko, Aviv Regev, and Joshua R Sanes. Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell*, 166:1308–1323.e30, 08 2016.

[92] Mark Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 103:8577–82, 07 2006.

[93] Peter Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. comput. appl. math. 20, 53-65. *Journal of Computational and Applied Mathematics*, 20:53–65, 11 1987.

[94] J.C. Dunn. Well-separated clusters and optimal fuzzy partitions. *Cybernetics and Systems*, 4:95–104, 04 1974.

[95] Mary McHugh. The chi-square test of independence. *Biochemia medica*, 23:143–9, 06 2013.

[96] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[97] H. Ahn, S. Son, and S. Kim. Deepfunnet: Deep learning for gene functional similarity network construction. In *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 1–8, 2019.

[98] Chris Cheadle, Marquis Vawter, William Freed, and Kevin Becker. Analysis of microarray data using z score transformation. *The Journal of molecular diagnostics : JMD*, 5:73–81, 06 2003.

[99] Princy Parsana, Claire Ruberman, Andrew Jaffe, Michael Schatz, Alexis Battle, and Jeffrey Leek. Addressing confounding artifacts in reconstruction of gene co-expression networks. *Genome Biology*, 20, 12 2019.

[100] Astrid Schneider, Gerhard Hommel, and Maria Blettner. Linear regression analysis part 14 of a series on evaluation of scientific publications. *Deutsches Ärzteblatt international*, 107:776–82, 11 2010.

[101] Andreas Buja and Nermin Eyuboglu. Remarks on parallel analysis. *Multivariate Behavioral Research*, 27, 03 2000.

[102] William S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 1979.

[103] Alaa Tharwat. Classification assessment methods. *Applied Computing and Informatics*, 2018.

[104] J. C. Rodriguez, G. A. Merino, A. S. Llera, and E. A. Fern?ndez. Massive integrative gene set analysis enables functional characterization of breast cancer subtypes. *J Biomed Inform*, 93:103157, 05 2019.

[105] Maxim Kuleshov, Matthew Jones, Andrew Rouillard, Nicolas Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry Jenkins, Kathleen Jagodnik, Alexander Lachmann, Michael McDermott, Caroline Monteiro, Gregory Gundersen, and Avi Ma'ayan. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, 44:gkw377, 05 2016.

[106] Darryl Nishimura. Biocarta. *Biotech Software & Internet Report*, 2(3):117–120, 2001.

[107] Andreas Ruepp, Brigitte Waegele, Martin Lechner, Barbara Brauner, Irmtraud Dunger, Gisela Fobo, Goar Frishman, Corinna Montrone, and Hans-Werner Mewes. Corum: The comprehensive resource of mammalian protein complexes-2009. *Nucleic acids research*, 38:D497–501, 11 2009.

[108] Huaiyu Mi and Paul Thomas. Panther pathway: An ontology-based pathway database coupled with data analysis tools. *Methods in molecular biology (Clifton, N.J.)*, 563:123–40, 02 2009.

[109] Lisa Matthews, Gopal Gopinath, Marc Gillespie, Michael Caudy, David Croft, Bernard de Bono, Phani Garapati, Jill Hemish, Henning Hermjakob, Bijay Jassal, Alex Kanapin, Suzanna Lewis, Shahana Mahajan, Bruce May, Esther Schmidt, Imre Vastrik, Guanming Wu, Ewan Birney, Lincoln Stein, and Peter D'Eustachio. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Research*, 37(suppl1):D619–D622, 11 2008.

[110] Denise Slenter, Martina Kutmon, Kristina Hanspers, Anders Riutta, Jacob Windsor, Nuno Nunes, Jonathan Mélius, Elisa Cirillo, Susan Coort, Daniela Digles, Friederike Ehrhart, Pieter Giesbertz, Marianthi Kalafati, Marvin Martens, Ryan Miller, Kozo Nishida, Linda Rieswijk, Andra Waagmeester, Lars Eijssen, and Egon Willighagen. Wikipathways: A multifaceted pathway database bridging metabolomics to other omics research. *Nucleic acids research*, 46, 11 2017.

[111] T. A. Knijnenburg, L. F. Wessels, M. J. Reinders, and I. Shmulevich. Fewer permutations, more accurate P-values. *Bioinformatics*, 25(12):i161–168, Jun 2009.

[112] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J. Royal Statist. Soc., Series B*, 57:289 – 300, 11 1995.

[113] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[114] J. Ahn, Y. J. Park, P. Chen, T. J. Lee, Y. J. Jeon, C. M. Croce, Y. Suh, S. Hwang, W. S. Kwon, M. G. Pang, C. H. Kim, S. S. Lee, and K. Lee. Comparative expression profiling of testis-enriched genes regulated during the development of spermatogonial cells. *PLoS ONE*, 12(4):e0175787, 2017.

[115] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.

[116] L.J.P. van der Maaten. Accelerating t-sne using tree-based algorithms. *Journal of Machine Learning Research*, 15:3221–3245, 2014.

[117] Pedro Ferreira, Manuel Muñoz-Aguirre, Ferran Reverter, Caio Godinho, Abel Sousa, Alicia Amadoz, Reza Sodaei, Marta Hidalgo, Dmitri Pervouchine, Jose Carbonell-Caballero, Ramil Nurtdinov, Alessandra Breschi, Raziel Amador, Patrícia Oliveira, Cankut Cubuk, João Curado, François Aguet, Carla Oliveira, Joaquin Dopazo, and Roderic Guigó. The effects of death and post-mortem cold ischemia on human tissue transcriptomes. *Nature Communications*, 9, 12 2018.

[118] Donna Maglott, Jim Ostell, Kim Pruitt, and Tatiana Tatusova. Entrez gene: Gene-centered information at ncbi. *Nucleic acids research*, 39:D52–7, 01 2011.

[119] EMBL-EBI. Ensembl stable ids. `https://www.ensembl.org/info/genome/stable_ids/index.html`, 2020. Accessed: 2020-09-16.

[120] Steffen Durinck, Paul Spellman, Ewan Birney, and Wolfgang Huber. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nature protocols*, 4:1184–91, 02 2009.

[121] Takeshi Obayashi, Yuki Kagaya, Yuichi Aoki, Shu Tadaka, and Kengo Kinoshita. Calculation of gene coexpression. `https://coxpresdb.jp/static/help/coex_cal.shtml`, 2018. Accessed: 2020-20-02.

[122] Frank Feltus, Stephen Ficklin, Scott Gibson, and Melissa Smith. Maximizing capture of gene co-expression relationships through pre-clustering of input expression samples: An arabidopsis case study. *BMC systems biology*, 7:44, 06 2013.

[123] Dirk Eddelbuettel and Romain François. Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18, 2011.

[124] Laura Comley, Heidi Fuller, Thomas Wishart, Chantal Mutsaers, Derek Thomson, Ann Wright, Richard Ribchester, Glenn Morris, Simon Parson, Karen Horsburgh, and Thomas Gillingwater. Apoe isoform-specific regulation of regeneration in the peripheral nervous system. *Human molecular genetics*, 20:2406–21, 06 2011.

[125] Joost de Winter, Samuel Gosling, and Jeff Potter. Comparing the pearson and spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. *Psychological Methods*, 21:273–290, 09 2016.

[126] C. H. Wei, H. Y. Kao, and Z. Lu. PubTator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Res*, 41(Web Server issue):W518–522, Jul 2013.

[127] Isabella Fraschilla and Kate Jeffrey. The speckled protein (sp) family: Immunity's chromatin readers. *Trends in Immunology*, 41, 05 2020.

[128] Sanda Rocak and Patrick Linder. Dead-box proteins: The driving forces behind rna metabolism. *Nature reviews. Molecular cell biology*, 5:232–41, 04 2004.

[129] J. D. Arroyo, A. A. Jourdain, S. E. Calvo, C. A. Ballarano, J. G. Doench, D. E. Root, and V. K. Mootha. A Genome-wide CRISPR Death Screen Identifies Genes Essential for Oxidative Phosphorylation. *Cell Metab*, 24(6):875–885, 12 2016.

[130] Xiong Liu, Xueping Yu, Don Zack, Heng Zhu, and Jiang Qian. Tiger: A database for tissue-specific gene expression and regulation. *BMC bioinformatics*, 9:271, 02 2008.

[131] Meritxell Oliva, Manuel Muñoz-Aguirre, Sarah Kim-Hellmuth, Valentin Wucher, Ariel D. H. Gewirtz, Daniel J. Cotter, Princy Parsana, Silva Kasela, Brunilda Balliu, Ana Viñuela, Stephane E. Castel, Pejman Mohammadi, François Aguet, Yuxin Zou, Ekaterina A. Khramtsova, Andrew D. Skol, Diego Garrido-Martín, Ferran Reverter, Andrew Brown, Patrick Evans, Eric R. Gamazon, Anthony Payne, Rodrigo Bonazzola, Alvaro N. Barbeira, Andrew R. Hamel, Angel Martinez-Perez, José Manuel Soria, Brandon L. Pierce, Matthew Stephens, Eleazar Eskin, Emmanouil T. Dermitzakis, Ayellet V.

Segrè, Hae Kyung Im, Barbara E. Engelhardt, Kristin G. Ardlie, Stephen B. Montgomery, Alexis J. Battle, Tuuli Lappalainen, Roderic Guigó, and Barbara E. Stranger. The impact of sex on gene expression across human tissues. *Science*, 369(6509), 2020.

[132] Marquis Vawter, Simon Evans, Prabhakara Choudary, Hiroaki Tomita, Jim Meador-Woodruff, Margherita Molnar, Jun Li, Juan Lopez, Rick Myers, David Cox, Stanley Watson, Huda Akil, Edward Jones, and William Bunney. Gender-specific gene expression in post-mortem human brain: Localization to sex chromosomes. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology*, 29:373–84, 03 2004.

[133] A. Vinuela, A. A. Brown, A. Buil, P. C. Tsai, M. N. Davies, J. T. Bell, E. T. Dermitzakis, T. D. Spector, and K. S. Small. Age-dependent changes in mean and variance of gene expression across tissues in a twin cohort. *Hum Mol Genet*, 27(4):732–741, 02 2018.

[134] Broad Institute. Cytoband. `https://software.broadinstitute.org/software/igv/Cytoband#:~:text=The%20Cytoband%20file%20format%20is,column%20tab%2Ddelimited%20text%20file.`, 2018. Accessed: 2020-10-28.

[135] Caiming Zhong, Ting Luo, and Xiaodong Yue. Cluster ensemble based on iteratively refined co-association matrix. *IEEE Access*, PP:1–1, 11 2018.

[136] Jorge Rodríguez, Miguel Medina-Pérez, Andrés Gutiérrez-Rodríguez, Raúl Monroy, and Hugo Terashima-Marín. Cluster validation using an ensemble of supervised classifiers. *Knowledge-Based Systems*, 145:134–144, 01 2018.

[137] Mingfeng Lin, Henry Lucas, and Galit Shmueli. Too big to fail: Large samples and the p-value problem. *Information Systems Research*, 24:906–917, 12 2013.

[138] Áine Duffy, Marie Verbanck, Amanda Dobbyn, Hong-Hee Won, Joshua L. Rein, Iain S. Forrest, Girish Nadkarni, Ghislain Rocheleau, and Ron Do. Tissue-specific genetic features inform prediction of drug side effects in clinical trials. *Science Advances*, 6(37), 2020.

[139] Pawel Michalak. Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. *Genomics*, 91:243–8, 04 2008.

Curriculum Vitae

Miguel Ángel Cortés Guzmán was born in 1995 in Mexico City. He achieved his bachelor in Biosciences with honorable mention from Tecnológico de Monterrey, Mexico in 2018. That same year, he underwent an internship at the sequencing and bioinformatics division of Servicio Nacional de Sanidad, Inocuidad y Calidad Agroalimentaria (SENASICA), a government organism in Mexico. There he participated analysing data from food related bacterial genomes. Due to interest in bioinformatics and computational biology, he enrolled in the Master in Computer Science program of that same university in 2019. He started research activities at the Bioinformatics for Clinical Diagnosis research group under the supervision of Dr. Victor Treviño where he has been focused on computational projects dealing with cancer and coexpression analysis.

This document was typed in using LaTeX 2_ε[1] by Miguel Ángel Cortés Guzmán.

[1] The template `MCCi-DCC-Thesis.cls` used to set up this document was prepared by the Research Group with Strategic Focus in Intelligent Systems of Tecnológico de Monterrey, Monterrey Campus.