Alma Mater Studiorum – Università di Bologna

Department of Pharmacy and Biotechnology FaBiT

International Master in Bioinformatics 28th September 2023

FunCoup-like inference of functional association networks from tissue specific data

Author

Paolo Mastrogiovanni

Supervisor

Prof. Castrense Savojardo

External Supervisors

Prof. Erik Sonnhammer

Davide Buzzao

Emma Persson

Academic Year 2022/2023

Abstract

The FunCoup network is based on a global model of protein functional associations which integrates hundreds of experimental datasets by a data-driven Bayesian framework, without distinction between tissues of origin. Tissue-specific networks can find disease-related relationships and disclose the varying functional roles of genes across tissues. Some attempts to model tissue-specific functional association networks were done by applying filters on genome-scale at pre- or post-training phase: (i) GIANT infers tissue-specific networks from genome-wide generic data using filtered gold standard interactions via tissue-specific annotations; (ii) FunCoup5 implements a filter to visualize pre-computed networks with only genes expressed in specific tissues according to protein abundance in the Human Protein Atlas. In this project, we attempted to infer tissue-specific networks starting from single cell RNA sequencing data, which holds high cell resolution and allows to extract tissue-specific functional associations. Inference was performed on seven human tissues (ovary, heart, pancreas, adrenal gland, muscle, lung, liver) applying the FunCoup's Bayesian framework, which establishes link confidence based on four gold standards. Resulting networks were analysed mainly to assess coverage, statistical significance, and biological specificity. Coverage was addressed by evaluating network features, such as dimension and node connectivity, after filtering for high confidence links. By computing similarity measures between networks for nodes, edges and significantly connected pathways, we highlighted differences in their overall gene interactions, showing the feasibility of this approach. While acknowledging the current limitations, such as the lack of a state-of-the-art approach for scRNA-seq quality control and the overall poor data availability, this research provides a foundation for future investigations.

Table of Contents

1 - Introduction

1.1 - Biological networks

During the past years, system biology has gained much attention. One of the reason behind this interest, is that in the post-genomics era the focus in biology has shifted from gathering information about genes and proteins from databases, to elucidate the functioning of their collective interplay [1]. At a highly abstract level, the various components that take part of most complex systems can be reduced to a series of nodes that are connected to each other by links, representing their interaction. Nodes and links together form a network, formally known in mathematical language as graph [2]. The number of edges incident to each node in the network it's called node degree; their distribution, in most real-world networks, follows a power-law, indicating a heterogeneous topology in which the majority of the nodes have a small degree, but there is a significant fraction of highly connected nodes that play an important role in the overall connectivity of the structure [3], [4]. Networks that follow a power-law distribution are defined as scale-free, and are characterized by high robustness, since the random failure of a node will most likely affect a low degree node, and high efficiency, having a short average path length [5]. Robustness can be measured by the size of the largest connected component (LCC), the highest number of connected nodes in the network [6].

Several biochemical networks have traditionally been studied: metabolic networks, that represent chemical transformation between metabolites; protein networks, representing protein-protein interactions; and gene networks, to show the relationships that can be established between genes, when observing how the expression level of one affects the others [7]. Gene networks can be either causal, if the edge between two nodes has always a biological correspondence that can be found experimentally (molecular interaction), or associative, if this is not necessarily the case[1]. One of the most used strategies to infer these networks is to start from experimental data of mRNA levels, since they can easily be gathered via high-throughput technologies. These act as snapshots of the molecular state of cell populations at the transcript level and are rich in information about gene networks[7]. Multiple sources of data can then be integrated using computational methods, that employ statistical and machine learning techniques, to assign confidence scores or weights to each one of the links. The current attempts to uncover the interactome consists of precise low-coverage experiments and affordable error-prone high-throughput techniques,

that, providing databases with bulk knowledge, may propagate the error. The integration of diverse data sources helps to improve the accuracy and coverage of the network, since it will cancel out potential non-systematic errors deriving from the individual experiments [8], [9].

1.2 - Naïve Bayes

Bayes' Rule, also known as Bayes' Theorem, is a mathematical formula used to determine conditional probability, which is the likelihood of an event occurring, given that another event has occurred. It allows to update probabilities based on new evidence. The formula for Bayes' Rule is as follows:

$$P(A|B) = \frac{P(B|A)*P(A)}{P(B)} \qquad [1.1]$$

- P(A|B) is the posterior probability of event A occurring given that B has occurred.
- P(B|A) is the posterior probability of event B occurring given A
- P(A) and P(B) are respectively the prior probabilities of event A and B occurring.

The Naïve bayes algorithm is a classification technique based on Bayes' Rule, with an independence assumption among predictors. It is obvious that the conditional independence assumption is rarely true, and therefore, the algorithm is found to work poorly for regression problems and produces poor probability estimates. But even if the posteriori probabilities predicted by the naïve bayes lack precision, the model seems to perform quite well on classification problem. The common rule of the Naïve Bayes is in fact to pick the hypothesis that is most probable (known has maximum a posteriori decision rule), which is not particularly affected by imprecision estimate, if they stay above the classification threshold. This means that the naïve Bayes tolerates the estimation error of class probabilities to some extent[10]–[12].

1.3 - Kernel Density Estimation

Kernel Density Estimation (KDE) is a non-parametric method used to estimate the probability density function of a random variable[13]. It is particularly useful when there is a need to model the probabilistic or stochastic structure of a dataset without making assumptions on the underlying distribution[14]. The basic idea of KDE is to place a kernel function, typically a symmetric and smooth function, at each data point and then average these functions to obtain the density estimate. Commonly used functions include Gaussian, Epanechnikov and uniform kernels[15]. The estimation process involves selecting a bandwidth hyperparameter, which controls the smoothness of the density. The selection of this parameter is crucial, as it can significantly affect the quality of the estimation. There are several methods to select the bandwidth, including least squares cross-validation and direct plug-in strategy[16].

The advantage of using KDE instead of similar approaches relies on its clarity. It provides a smooth estimate, which can be more informative and visually appealing[17]. However, the method also has some limitations. It can suffer from boundary bias when the data is non-negative or has a specific domain, as most kernels do not consider the domain of the data. Additionally, KDE can be computationally expensive, especially for large datasets and high dimensional data[18], [19].

1.4 - FunCoup

FunCoup is a comprehensive database of functional association networks that uses a unique Bayesian approach to combine different data types for the inference, to increase coverage. The database applies orthology transfer to share functional association information between 22 species from all domains of life [9], [20]. Base ground of the network generation are its evidence data types, signals that support or contradict the presence of functional coupling between two genes. Funcoup latest version (5.0) integrates 11 different data types:

- Protein interaction (PIN) from iRefIndex [21], assigning an higher score to the links confirmed by multiple publications.
- mRNA co-expression (MEX) from multiple experimental conditions or tissues. Values are obtained by computing the spearman correlation of expression profiles.

- Protein co-expression (PEX) measured from the Human Protein atlas [22].
- Genetic interaction profile similarity (GIN).
- Shared transcription factor binding (TFB), considering TF profile similarity.
- Co-miRNA regulation by shared miRNA targeting (MIR), considering profile similarity.
- Sub-cellular co-localization (SCL), derived from the GO ontology[23]. The more specific is the location the higher the weight assigned.
- Domain interaction (DOM) predicted with UniDomInt[24].
- Phylogenetic profile similarity (PHP), scored across multiple species as a fraction of branch lengths shared by both genes.
- Quantitative mass spectrometry (QMS) obtained via PaxDB[25].
- Gene regulatory (GRG) as directed links inferred from transcription factor-gene bindings. Chip-seq data from Encode is used as evidence[26].

The database differentiates between 5 classes of functional associations: physical protein interaction, sharing the same signalling pathway, participation in the same protein complex, sharing the same metabolic pathway and co-occurrence in the same operon. Evidences are integrated for each one of the classes separately to infer individual networks. These networks can then be merged in a single one, by keeping the max score that was assigned to each link[9], [20].

In the first version of FunCoup the network framework was based on a simple Naïve Bayes approach, as it tolerates well missing values and gives straightforward and interpretable scores. However, the algorithm relied on the independence of the underlying data, assumption that is violated by some of the integrated datasets. For instance, co-expression analysis of two different microarray studies might provide redundant information, based on the focus of the studies. For this reason, on the third release of the database, a redundancy weighted variant of the method was introduced, with the aim to down-weight redundant information in the evidence. The idea behind the weighting scheme is that evidence of the same type will only increase the final score to the extent that they provide novel information to the link[20], [27].

To train the network, a positive and negative gold standard are required for each association class. In biology, a gold standard dataset is usually a collection of

data derived from high-quality verifiable experimental evidence. In our case, they are known set of functional association retrieved as follows:

- Protein-Protein Interactions (PPI) are collected from iRefindex [21].
- Gold standard couplings for protein complexes are collected from iRefindex, ComplexPortal [28] and Corum [29].
- Metabolic and Signaling gold standard couplings are collected from KEGG pathways[30].
- Gold standard couplings from shared Operons were collected from OperonDB (Only for prokaryotic organisms)[31].

For each one of the positive sets, a negative gold standard is required. As there are not methods to prove the absence of association, the negative gold standard is selected as background distribution of random links (absent in the positive counterpart). By increasing the number of negative links, the training becomes less vulnerable to errors. For each one of the evidence datasets, scores are associated to positive and negative gold standards, to then compute two probability density curves via Kernel density estimation. The curves allow the association of each value in the evidence dataset to two probabilities, used to compute the log-likelihood ratio (LLR).

The integration of likelihoods obtained by different evidence datasets is performed as follows:

$$LLR(a,b)_t = \sum_e LLR(a,b)_e \prod_{k<e} d_{ek}, \qquad [1.2]$$

The LLR score for the genes (a and b) on the evidence type (t) is computed by summing the weighted LLR scores derived from each evidence dataset (e). LLRe are ranked by their absolute value in decreasing order and d is the distance between evidence e and evidence k. Each LLRe is weighted by the product of the distances to each evidence on the left side in this ranking. Distance is defined as $d_{ek} = \alpha(1 - max(0, r_{ek}))$, where r is the Spearman correlation between the LLRs for evidence e and evidence k, and is the baseline redundancy. The parameter alpha has the role to correct for redundancy underestimation due to noise and is set to 0.7 [9], [20].

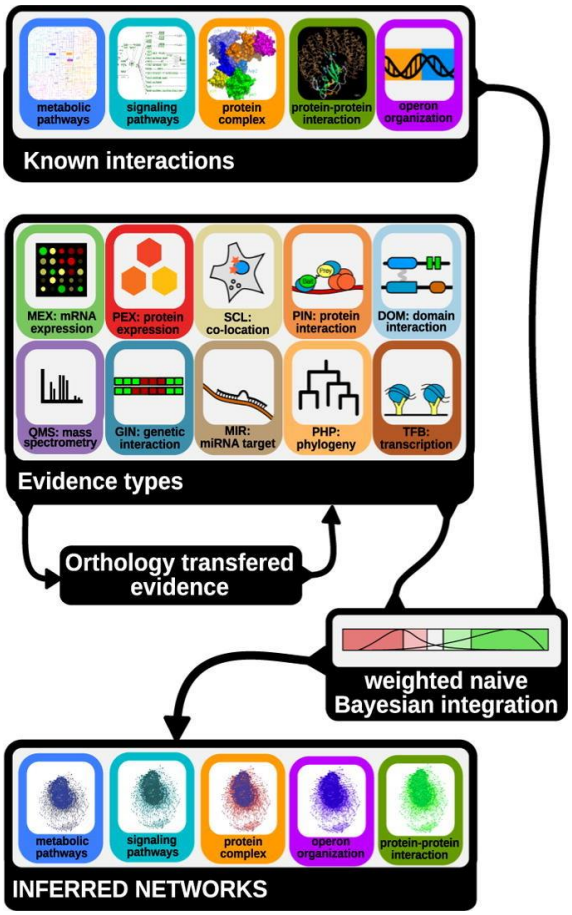

Figure 1 – The FunCoup framework. Likelihood scores are computed for each evidence dataset by comparison with known interactions (gold standards) in the weighted Bayesian framework. For each gold standard a separate network can be generated **[9]**.

In FunCoup 5 (the latest version) a new functionality has been introduced to allow the user to filter genes based on tissue expression. For Homo Sapiens, tissue specificity is based on protein expression values extracted from the Human Protein Atlas. This filter annotates a total of 241,719 unique genes in the FunCoup networks with at least one tissue. To showcase the impact of tissue-specific genes in the interactome, the researchers have computed the mean shortest distance (msd) between all the genes expressed in a specific tissue, and then compared it to randomly selected gene sets of the same size as the tissue-specific gene set. All 49 human tissues in the database showed to have a significantly lower msd between their genes compared to random. Moreover, these subsets of genes also show a significantly larger LCC, suggesting that tissue-specific genes tend to cluster together[9].

1.5 - Giant

GIANT (Genome-wide Analysis of gene Networks in Tissues) serves as a prediction server focused on human tissue-specific gene interactions, featuring genome-wide functional association networks for 144 human tissues and cell types. The process of building functional networks entails performing weighted tissue-specific Bayesian integration based on a genome-scale dataset. For each tissue, a naïve Bayesian classifier is trained, leveraging gene-to-tissue annotations obtained from sources like GTEx [32], FANTOM5 [33], and the Gene Ontology [23] as gold standards. This training allows the assignment of tissue-specific weights to each gene, calculated as the median transcription frequency of the gene across all samples corresponding to the tissue.

When a user initiates a GIANT prediction, they provide a set of genes and one or more tissues of interest (Figure 2A). Subsequently, the server employs probabilistic integration of thousands of genome-scale experiments, tailored to the specified tissues, to predict the likelihood of functional relationships between the user's genes and all other genes within the human genome (Figure 2B). The results are then presented to the user in the form of a gene network specific to each queried tissue, complete with posterior probabilities denoting the likelihood of functional relationships for the genes of interest within that tissue (Figure 2C)[34], [35].

Although GIANT is today one of the mostly used databases for tissue-specific networks, the specificity of its interactions relies on predictions applied during the training phase. This provides us with enough motivation to attempt in a network construction that starts directly from tissue-specific data, with the goal to achieve a more precise snapshot of the cellular state.

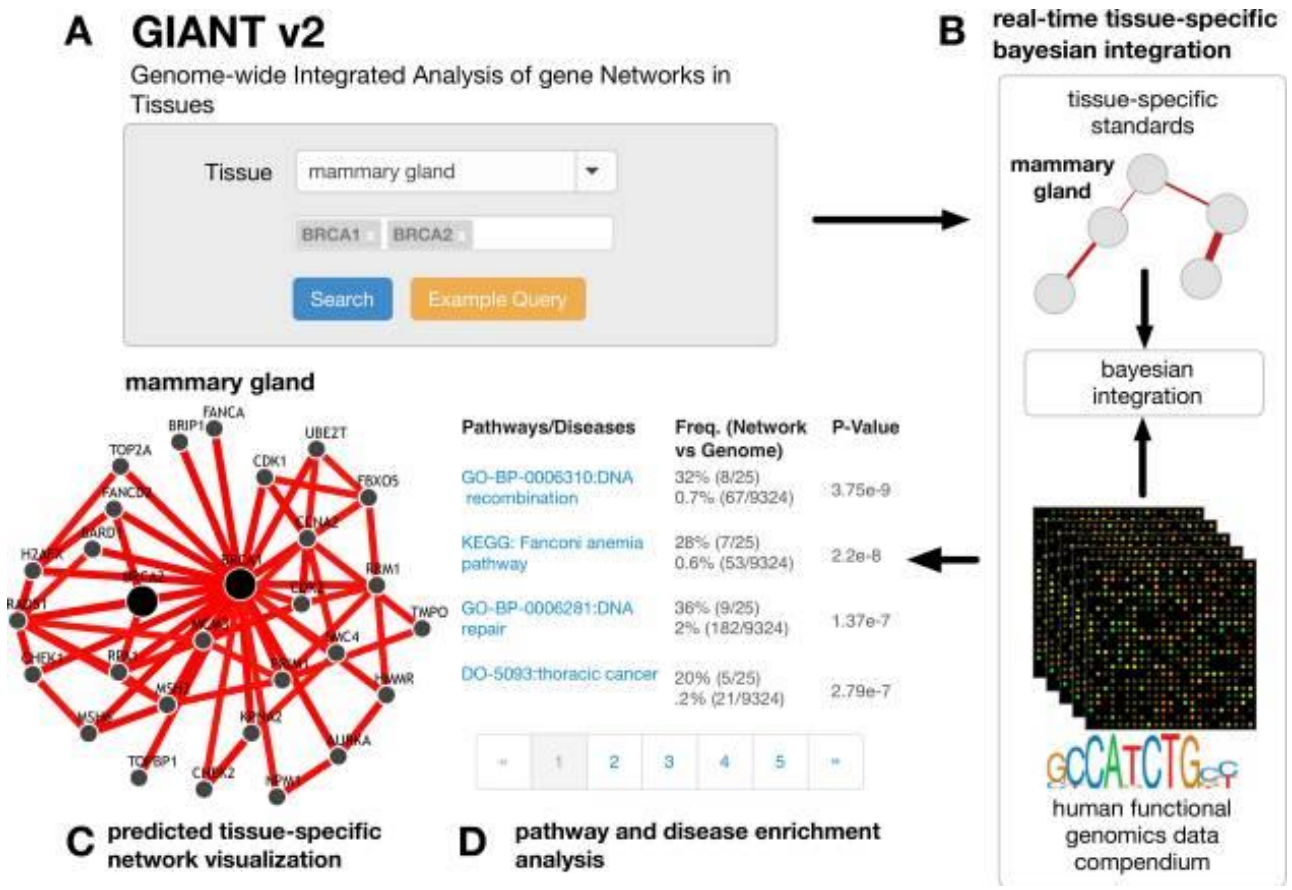

Figure 2 – A schematic of the GIANT tissue-specific interaction prediction server. The user can query genes directly to the tissue of interest (A) . Predictions are made on-the-fly based on pre-computed Bayesian models (B). Additional pathway and disease enrichment analysis of the network is also available (D) **[34]**.

1.6 – Single Cell RNA sequencing

Almost all cells in the human body have the same set of genetic material, but their transcriptome information reflects the unique activity of only a subset of genes. For this reason, profiling the gene expression activity in samples is considered as one of the most efficient approaches to probe cell identity, state, function and response[36]. Single-cell RNA sequencing is a powerful genomic approach that enables the detection and quantitative analysis of messenger RNA molecules within individual cells[37]. Since its discovery, scRNA-seq has revolutionized the medicine field, by offering a higher resolution of cellular differences compared to traditional bulk RNA sequencing methods[36].

The first step to scRNA-seq is isolation of individual cells, although the capture efficiency is a big challenge for the current technologies. To date, several

approaches are available for the isolation, including limiting dilution, fluorescence-activated cell sorting (FACS), magnetic-activated cell sorting, microfluidic system and laser microdissection. The choice of the technology should be done based on the characteristics of the specific organ/tissue of interest. The objective of the capture is to have each single cell in an isolated reaction mixture to allow for reverse transcription (conversion of RNA to cDNA) and barcoding of the resulting transcripts [36],[38]. The small amount of synthesized cDNAs is then further amplified using conventional PCR or in vitro transcription before sequencing, for which the Illumina platform is widely used[39] (Figure 3).

The isolation procedure contains many drawbacks, as cells are heavily influenced by their surrounding environment and cell-to-cell interactions. Samples that have been manipulated and isolated are no longer in their native environment, which could result in an unnatural response. Moreover, the isolation process can induce unintended stress, altering the behaviour, viability and profile of the cell [40]. These technical limitations are to be addressed during the pre-processing step, but, to this date, computational pipelines for handling raw data files remain limited. Some commercial companies provide software tools for the task, but this area remains in its infancy, and gold-standard tools have yet to be developed[39].

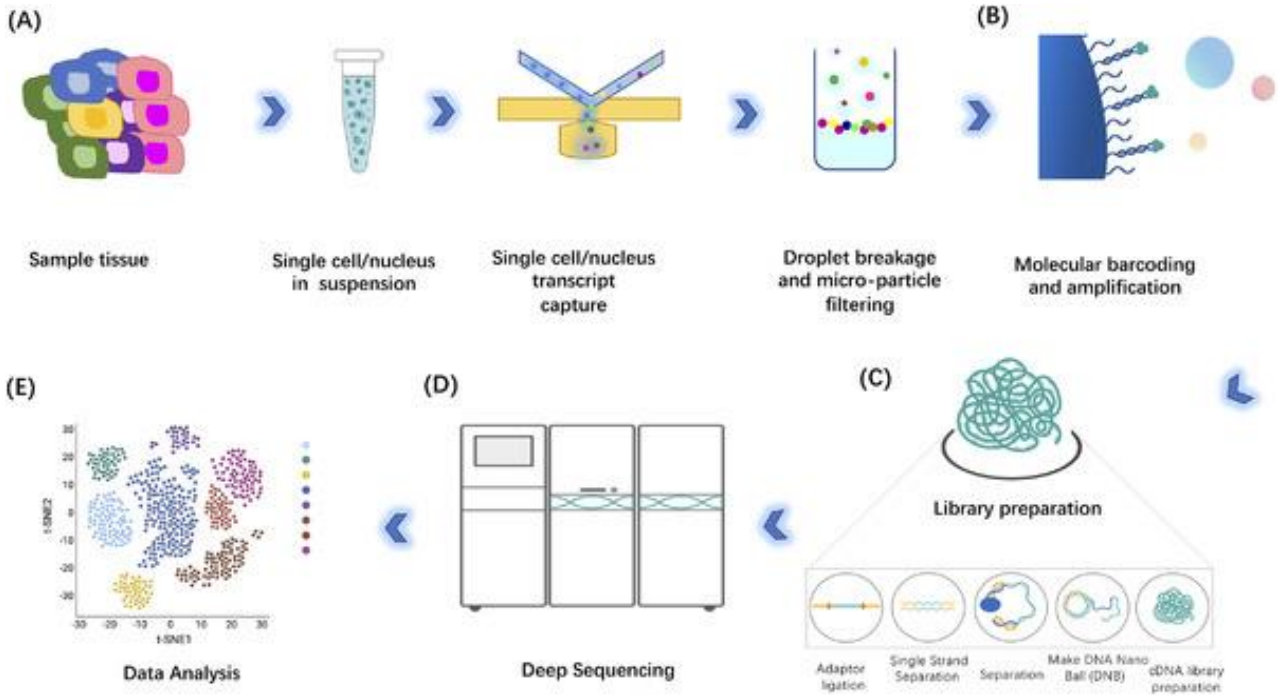

Figure 3 – ScRNA-seq pipeline, from sample isolation up to data analysis ***[36]****.*

ScRNA-seq protocols can generate a portion of low-quality data from cells that are broken, dead or contaminated. These cells hinder the downstream analysis and may lead to misinterpretation of data, hence must be removed during the quality control process. Samples containing only a few reads should be discarded since insufficient sequencing depth may lead to the loss of a large portion of lowly and moderately expressed genes. Cytoplasmic RNAs are usually lost but mitochondrial RNAs are retained for broken cells, hence the ratio of reads mapped to mitochondrial genome can also be used to identify low-quality samples[38]. To date, the standard practice in quality control is to filter out cells by setting arbitrarily defined thresholds on the QC metrics. Main flaw of this approach is the fact that these filters do not account for biological variation between samples. For example, mitochondrial transcript abundance is dependent on cellular physiology, and metabolically active tissues (such as muscle or kidney) have a higher mitochondrial transcript content. For This reason, while it is widely used, it is important to keep in mind that such approach might result in data loss for certain cell types [41].

2 - Materials and Methods

2.1 - Data gathering

Most of our datasets were sourced from the ENCODE Data Coordination Center (DCC) consisting of validated projects submitted by members of the ENCODE Consortium [26]. Tissue selection was based on data availability, excluding tissues with fewer than 4 scRNA-seq datasets in the DCC (Figure 4). Consequently, we chose ovary, muscle, heart, pancreas, liver, adrenal gland, and lung tissues for our study. Although we initially retrieved sufficient datasets for the intestine, this tissue was later discarded after quality control checks. To ensure a minimum of 5 starting datasets per network, we supplemented our data with information from the Tabula Sapiens project [42], a comprehensive human reference atlas comprising nearly 500,000 cells from 24 different tissues and organs, accessible via the Human Cell Atlas database [43]. Our data management was facilitated by the Anndata Python library [44], designed for handling sparsity matrices while preserving raw expression values and metadata within a single data structure. The expression matrix was structured so that each row represented a sample (cell), and each column represented a gene. In most cases, the raw datasets contained approximately 60,000 genes (including isoforms, mitochondrial and ribosomal genes) and a number of samples ranging from 10,000 to 80,000.

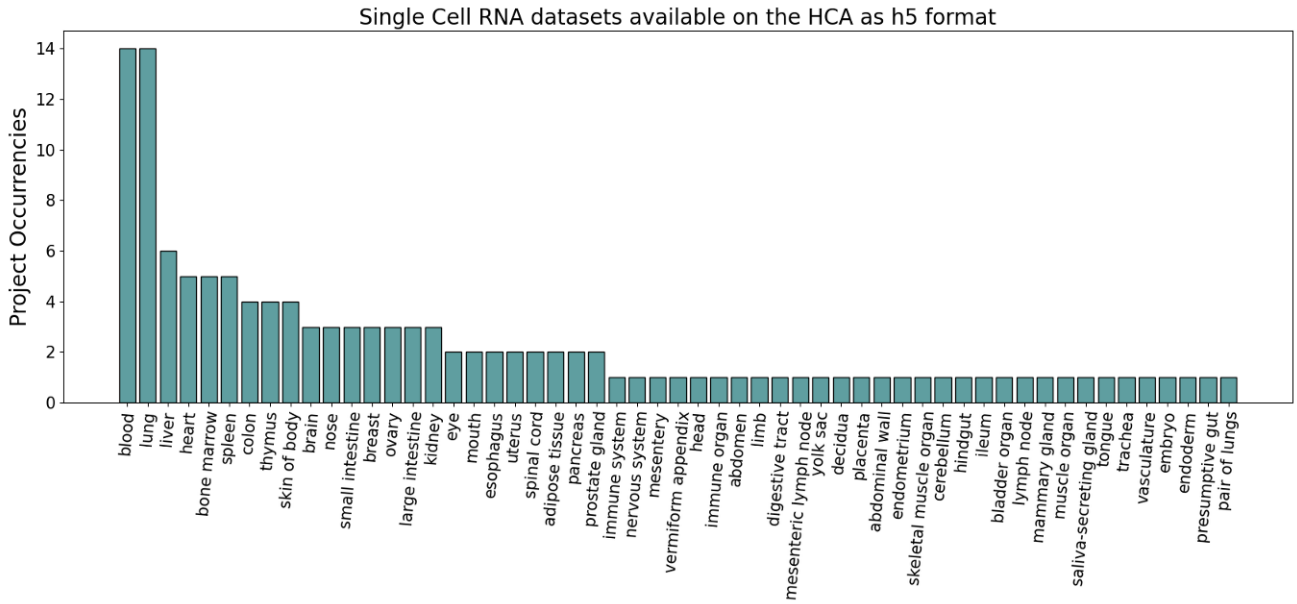

Figure 4 - Tissue distribution of scRNA-seq datasets on the ENCODE project. Only tissues with more than 4 available datasets were selected for the study.

2.2 - Data pre-processing

All potential non-existent or non-coding genes were excluded from the analysis by performing ID mapping of the gene identifiers from each dataset against the UniprotKB [45]. Redundances and isoforms were excluded by keeping always the first mapped entry; this reduced the gene number for each dataset to around 20.000. Additionally, we subjected our samples to rigorous filtering procedures aimed at improving data quality and reducing the likelihood of including potentially damaged or contaminated cells. These filtering criteria were established based on existing literature and involved the removal of cells meeting the following conditions:

- Cells with more than 10% of Mitochondrial or Ribosomal Genes [46], [47].
- Cells with fewer than 500 read counts [41].
- Cells with fewer than 200 detected genes [41].

2.3 - Correlation

To assess the overall co-expression of genes within a specific tissue, we computed the correlation of expression values across samples. Given that the relationship between these values cannot be assumed to be linear, we opted for the Spearman method. Spearman's rank correlation coefficient (ρ) is a non-parametric measure that gauges the strength and direction of association between two ranked variables. This coefficient ranges from -1 (indicating a perfect negative correlation) to +1 (indicating a perfect positive correlation). The computation was performed with the python NumPy library [48], which first converts the expression values into ranked data, and then applies the following formula:

$$\rho = 1 - \frac{6\Sigma\, \mathrm{d}_i^2}{n(n^2 - 1)} \qquad [2.1]$$

Where p is the score, d is the difference between the two ranks of each observation and n is the number of observations.

To reduce the introduction of noise and meaningless information, the score was computed only for gene pairs for which both expression values were observed in at least 100 cells. For each dataset, one co-expression matrix was produced and stored as NumPy array.

2.4 - Gold Standards

Gold standards represent curated collections of known functionally associated gene pairs obtained from established reference databases (see chapter 1.4). In our study, these gold standards served as proxies for authentic interactions, aiding us in assigning confidence scores to each gene pair. We employed four distinct categories of gold standards:

- *Protein-Protein Interactions* (PPI).
- *Metabolic* and *Signaling* Pathways.
- *Protein complexes*.

For each positive gold standard, a corresponding negative gold standard was created, starting from randomized gene pairs that were absent in the positive set. For each dataset, the correlation values are mapped to the correspondent gene pair in the positive and negative gold standard files. Each pair of positive/negative gold standards is used to compute probability density functions (pdf) with Kernel Density Estimation. The resulting correlation value versus PDF curves were visually represented using Matplotlib [49]. It's worth noting that we initiated this process with 50,000 pairs for both the negative and positive sets, as further increasing the dataset size only prolonged computational time without significantly altering the distribution of data points.

2.5 - Log-Likelihood ratio scores

The log-likelihood ratios were calculated for each data point within the KDE curves. LLRs serve as measures that assess the likelihood of two competing hypotheses. In our case, these hypotheses compare the likelihood of a data point belonging to the positive gold standard dataset, indicating a functional association, against the likelihood of it belonging to a null distribution. To achieve this, we used the SciPy interpolation[50], which builds a function

starting from the correlations and associated probabilities derived from the curves. This function can be used to interpolate the pdf values for any given correlation value within the range of the original data. For each correlation value within each dataset, the interpolation function retrieved the probabilities from both the positive and negative gold standard curves. It then computed the LLR as the logarithmic ratio between the probability of the correlation value (representing the gene pair) belonging to the positive set and the probability of it belonging to the negative set.

The LLRs obtained for each individual dataset of the same tissue are then combined with a redundancy weighted integration approach (see chapter 1.4, formula 1.2). This procedure is done for each one of the 4 gold standards, resulting into 4 networks per tissue. These can then be merged by taking the max link strength for each gene pair.

2.5 - Network analysis

Networks were generated and analysed using the Python library NetworkX [51], which allows for the creation of graph objects. In these graphs, genes served as nodes, while the weighted edges conveyed the probability of functional association between them. The analysis focused on the following characteristics:

- Number of nodes and edges.
- Number of nodes and edges in the Largest Connected Component (LCC).
- Mean node degree, computed as the average number of edges connected to each node in the network.
- Median node degree, determined by identifying the middle value when sorting node degrees in ascending order.

To ensure robustness in the analysis, all statistics were calculated on networks after filtering out edges with scores lower than 0.5, effectively excluding low-confidence functional associations. Moreover, we created scale-freeness plots using Matplotlib to visualize the node degree distribution, allowing us to gain a deeper understanding of the network topology.

2.6 - Network similarity

To assess the degree of similarity between different tissue networks, two different measures were computed over nodes and edges:

- Jaccard index: calculated by dividing the number of elements in the intersection of the two sets by the number of elements in the union of the two sets.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad [2.2]$$

- Szymkiewicz–Simpson coefficient (SS): computed by dividing the number of elements in the intersection of the two sets by the number of elements within the smaller of the two sets.

$$S(A,B) = \frac{|A \cap B|}{\min(|A|,|B|)} \quad [2.3]$$

Both measures yield values within the range of 0 to 1, where 0 signifies no similarity, and 1 indicates complete similarity between the two sets being compared.

2.7 - Statistical Significance

To assess the statistical significance of the networks, we utilized the KEGG [30] pathways list as reference. It contains 334 pathways, each one associated to the participating genes. For each network (links with score >0.5), we compared the lengths of the largest connected component within the subgraphs composed by the genes of each pathway to a null model.

To establish a suitable null model for each pathway, we created 10,000 subgraphs with randomly selected nodes from the overall network, with the number of sampled nodes matching the number of genes in the pathway under investigation. For consistency and reproducibility, each sampling process employed a predetermined seed. The LCC lengths of these randomly generated subgraphs were recorded and collectively constituted a background distribution for each pathway.

We computed p-values as the number of random LCCs smaller than the observed LCC for the given pathway, divided by the total number of iterations (10,000); with a precautionary addition of 1 to both the denominator and numerator to prevent any division by zero errors. Subsequently, to account for multiple comparisons, the obtained p-values underwent Benjamini-Hochberg correction. The Benjamini-Hochberg (BH) correction is a statistical method used to control the false discovery rate (FDR) in multiple hypothesis testing. It helps to decrease the number of false positives (Type I errors) that may occur when performing multiple tests simultaneously. The correction was applied using the Statsmodel library for Python[52], using 0.05 as alpha (expected false discovery rate).

Pathways were considered significant when their corrected p-values were less than 0.05. These significant pathways were compared across different networks using heatmaps and histograms to assess similarities. Additionally, we performed comparisons with the Homo Sapiens FunCoup5 database, and a network created by merging the seven individual tissue networks.

3 - Results and discussion

3.1 - Quality control

As single cell data is highly prone to the introduction of error due to cell rupture or contamination, all datasets were subjected to stringent quality controls procedures. While it would be advantageous to adopt a data-driven approach, considering the varying dimensions and content of these datasets, the current state of available methods for adaptive quality control lacks the required precision and may introduce undesirable artifacts into the study. For instance, we explored the potential of the DDQC algorithm [41], which seeks to apply adaptive filtering thresholds based on the clustering of similar groups of cells (aiming to identify cell types). However, our experimentation with this approach revealed unrealistic correlation values, often indicating perfect correlations among gene pairs. Given our aim to maintain the study's linear integrity and avoid introducing excessive artificial variability between networks, we opted for fixed filtering thresholds based on established literature (see chapter 2.2).

Figure 5 provides a visual representation of our filtering process, using a heart dataset as an example. In most instances, for cells of the Homo Sapiens species, a mitochondrial or ribosomal RNA content exceeding 10% serves as an indicator of apoptotic, stressed, or low-quality cells, leading to their exclusion. To avoid data loss from some less organized cell types, the filtering of cells with low number of genes was kept as light as possible, adopting 200 as threshold, as it is the minimum number of genes to keep a cell alive. Furthermore, we excluded cells with less than 500 read counts, as these cells might have experienced genetic material loss due to cell membrane rupture. This threshold aligns with the common practice observed in most single-cell RNA research.

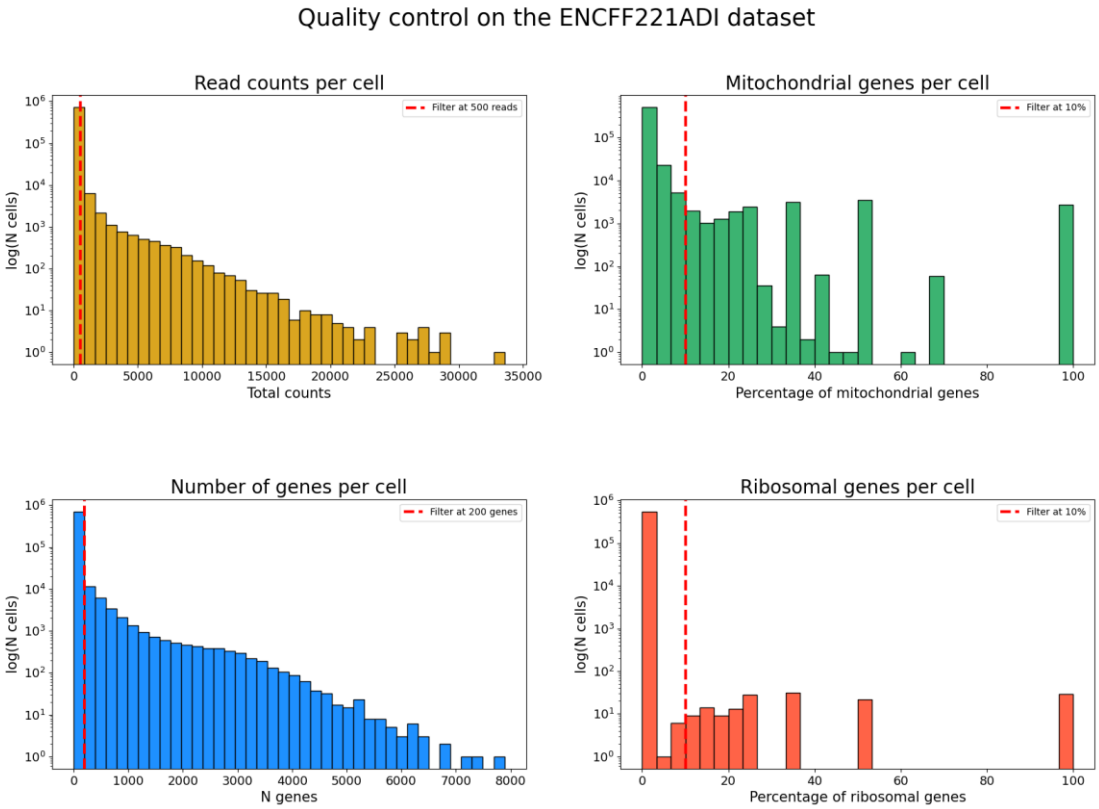

Figure 5 – The data filtering process applied to the ENFC221ADI heart dataset is here illustrated. For read counts and number of genes samples located before the red line in the graph were removed. Conversely, for ribosomal and mitochondrial percentages, the removal was executed for samples positioned after the red line.

3.2 – Gene co-expression

Gene co-expression was computed as Spearman correlation across samples. Fine-tuning the minimum number of observations for these computations emerged as a pivotal factor in obtaining meaningful results. In fact, choosing a low value such as 10 (default in the Numpy library function) resulted in high coverage for what regards gene pairs, but also in highly skewed correlation values, with many perfect positive correlations. Reasonably, the lower is the number of observed expressions for the pair, the higher is the probability of them following an identical pattern, regardless of the accuracy of the data. On the other hand, selecting a high threshold, such as 1000, yielded considerably low coverage (Figure 6), and despite the correlations being supported by a more substantial number of observations, they could not be effectively utilized for

network construction. Following comprehensive parameter testing, detailed in Figure 7, we established a threshold of 100 as the most suitable compromise. This threshold appeared to strike a balance between achieving adequate coverage and maintaining a well-distributed correlation profile, ultimately enhancing the quality and reliability of our network analysis.

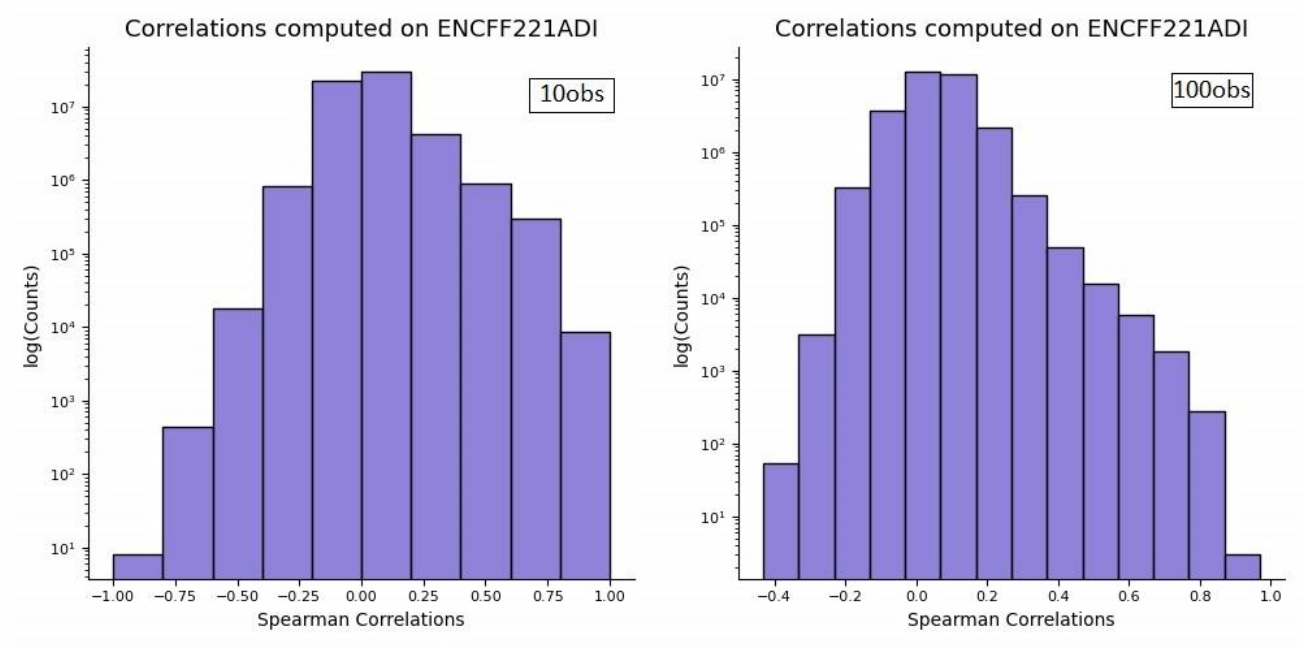

Figure 6 – Comparison between the correlation values computed on the same dataset, but with different n_obs parameters: 10 on the left, and 100 on the right.

N obs	N links
10	5.9E+07
100	3.1E+07
250	1.7E+07
500	8.4E+06
1000	3.3E+06

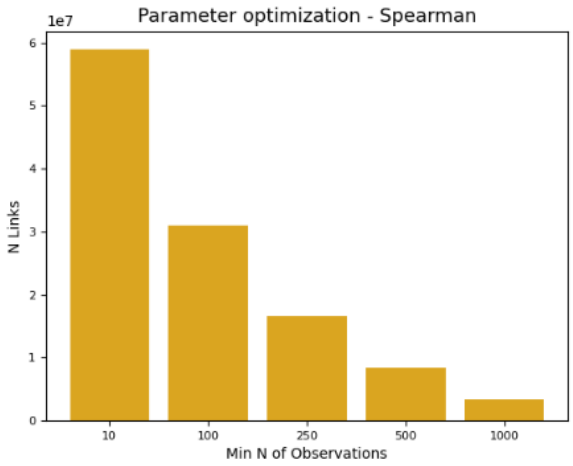

Figure 7 – As highlighted from the table, an increase in the number of minimum observations during spearman correlation computation results in an exponential decrease in link coverage.

3.3 – Probability density curves and LLRs

Using kernel density estimation, we generated two probability density curves for each dataset with respect to each gold standard. Visualizing these curves already provides valuable insights. Ideally, we would expect a clear separation between the probability curves, as one is built starting from true interactions (positive gold standard) and the other from random interactions (negative gold standard). This would lead to higher log-likelihood ratios, but it's not always the case. In fact, differences in cell type distribution across various datasets introduce variations. These discrepancies are primarily due to the fact that, since the four gold standards were generated based on global interactions, they do not necessarily mirror the state of cells in each specific tissue. This limitation is quickly addressed by integrating the LLR obtained from each gold standard with the others, increasing the network coverage. Positive LLRs can be derived from every point in which the positive curve is above the negative curve. This typically occurs at the extremes, as strong correlation values are less common and, consequently, are more likely to represent genuine functional associations.

What is shown in the KDE curves (Figure 8) reflects the derived LLRs distribution. Higher LLRs are typically obtained from strong positive or negative correlations within the data, indicating strong evidence of functional associations. Conversely, low or negative LLRs are associated with correlation values around 0, suggesting weak or non-existent associations. For the build of the Network, only values that get past a certain LLR threshold are used.

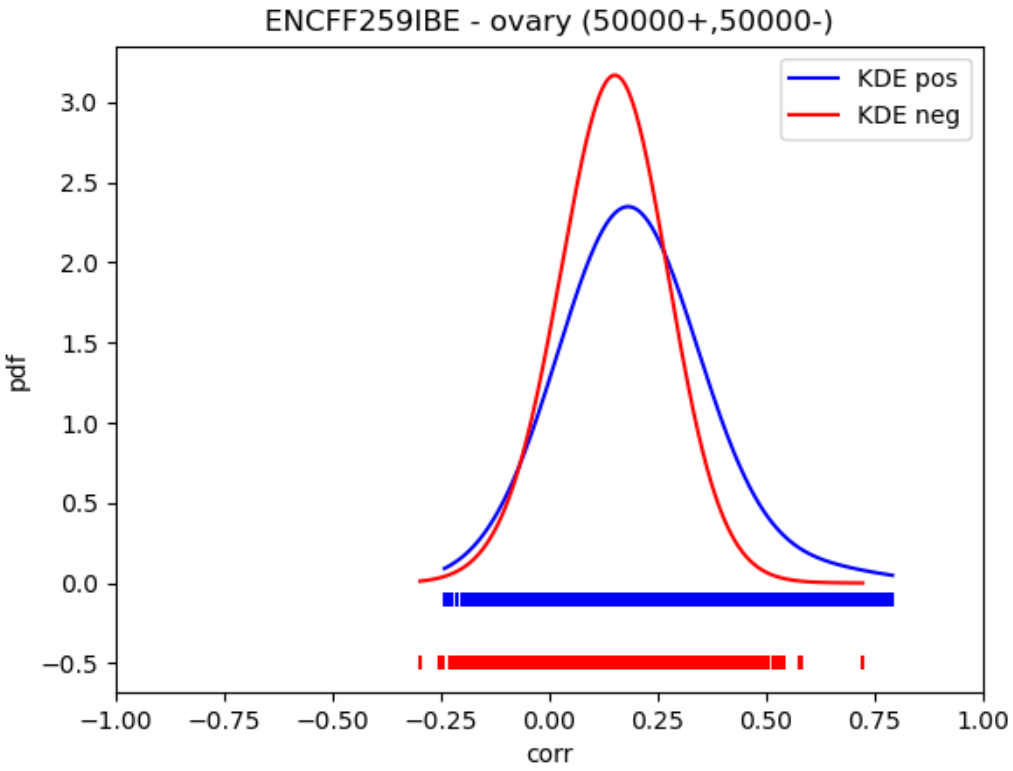

Figure 8 – Kernel density estimation curves generated from the metabolic gold standard on the ENCFF259IBE ovary dataset. Blue curve is derived from the positive GS, red curve is derived from the negative GS.

3.4 – Network analysis

Log-likelihood ratios were computed individually for five datasets per tissue and subsequently integrated to yield final scores, serving as weights for the network links. This produced networks for seven distinct tissues, including: ovary, heart, lungs, adrenal gland, muscle, pancreas, and liver. In addition to these tissue-

specific networks, we also incorporated the human FunCoup5 database, and a network generated through the union of all seven tissue datasets.

Figure 10 and Table 1 illustrates noticeable differences in both network coverage and distribution of log-likelihood ratios (LLRs). In general, most networks exhibit scores ranging from -1 to approximately 4, except for the ovary tissue, which attains a maximum score of 8, and the adrenal gland tissue, where scores barely surpass 2. Given that the network analysis necessitates the retention of only high-confidence scores, it is unsurprising that certain networks may exhibit reduced coverage. For this study, we selected an LLR threshold of 0.5. While this value may not appear particularly stringent, especially when compared to the FunCoup5 LLR distribution, which reaches values as high as 40, it can be justified. In fact, it's important to note that these networks were generated using a single evidence type, in contrast to FunCoup5, which integrates data from 11 distinct sources. Thus, applying a more stringent filtering criterion would not allow for a fair comparison. Even so, substantial coverage is lost following the application of the filtering criteria. Out of the initial 20,000 genes, around 12,000 nodes remain in most of the tissue specific networks (except for the Adrenal Gland tissue, which has less than 3,000).

	FC5	merged	AdrGland	liver	lung	muscle	ovary	pancreas	heart
N nodes	12688	15695	2656	9281	13676	12158	12245	11809	12126
N edges	5.90E+05	1.90E+06	7.00E+03	3.40E+04	8.60E+05	3.10E+05	1.80E+05	1.10E+05	6.40E+05
N nodes in the LCC	12491	15695	1602	8771	13668	12154	12245	11792	12043
N edges in the LCC	5.90E+05	1.90E+06	6.30E+03	3.40E+04	8.60E+05	3.10E+05	1.80E+05	1.10E+05	6.40E+05
Mean node degree	92.85	239.49	5.3	7.37	126.2	51.18	28.81	18.58	104.97
Median node degree	13	106	1	2	51	17	16	7	11

Table 1 – Networks were filtered for (relatively) high confidence links (>0.5) before the computation of the shown statistics.

Interestingly, when taking the union among all the tissue specific network, a total of 15,695 nodes are retained, more than FunCoup5, stopping at 12,688. The number of edges, the median node degree and mean node degree also appear significantly higher compared to the other networks. But before considering this as an improvement, further studies are needed to verify if by integrating other evidence types, and increasing the LLR threshold, the coverage and connectivity are kept.

It's also worth noting that the Largest Connected Component (LCC) of each network includes most of the nodes, reinforcing the network's overall connectivity and coherence. Moreover, as shown in figure 11, all networks can be considered scale-free, as their node degree distribution follows a power law or closely approximates it.

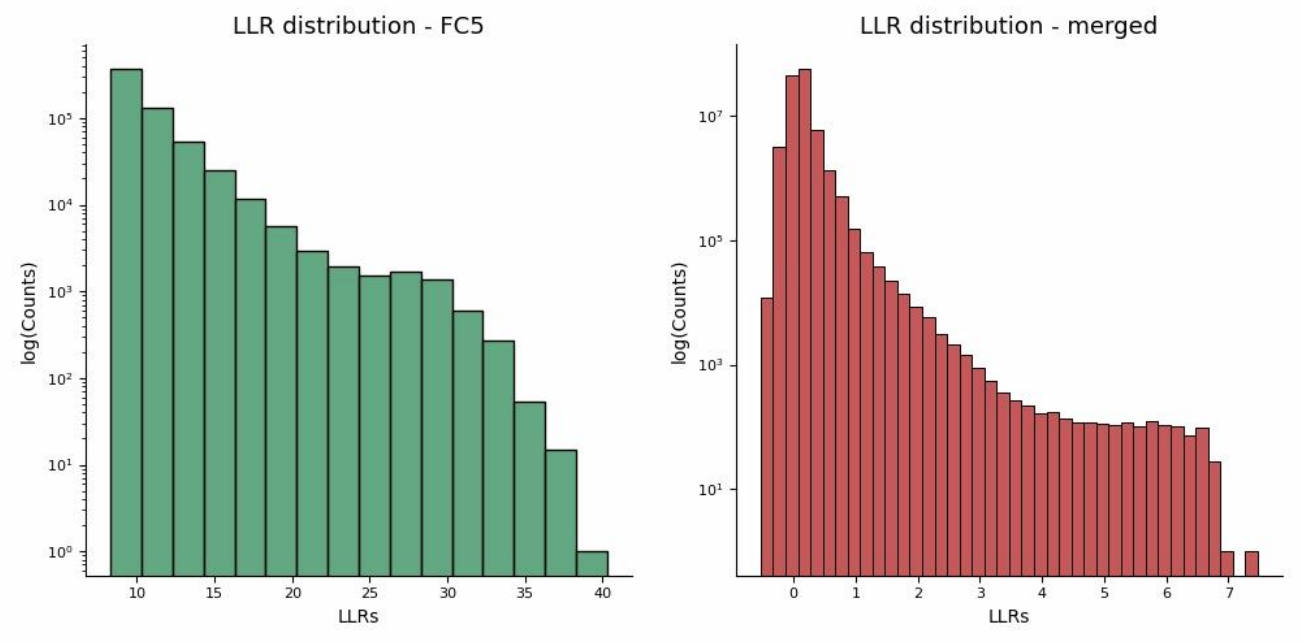

Figure 9 – LLRs distribution. On the left side, the FunCoup5 network; on the right side, the network resulted from the union of the tissue specific networks.

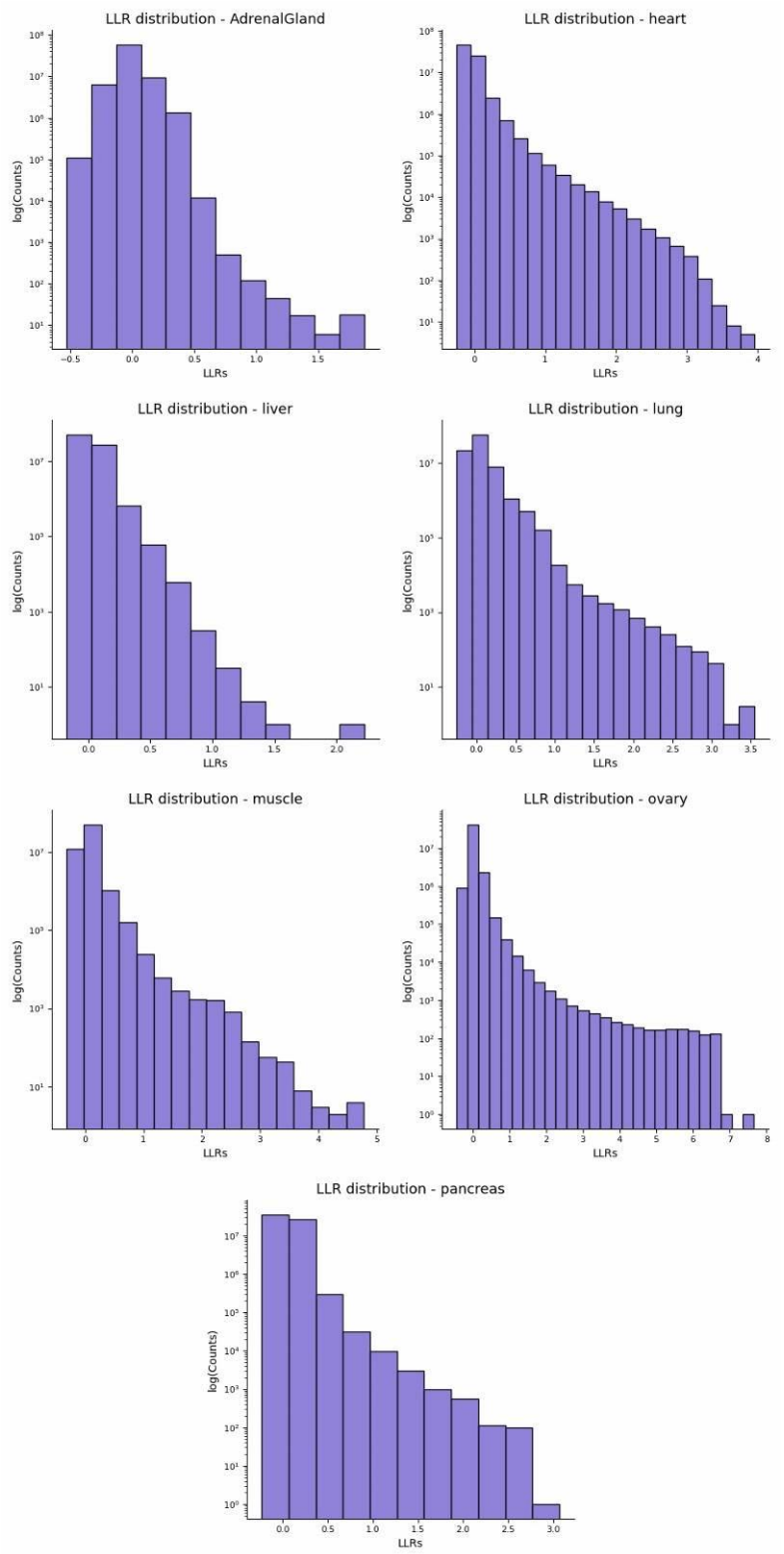

Figure 10 – Final LLR distribution for each one of the tissue specific networks.

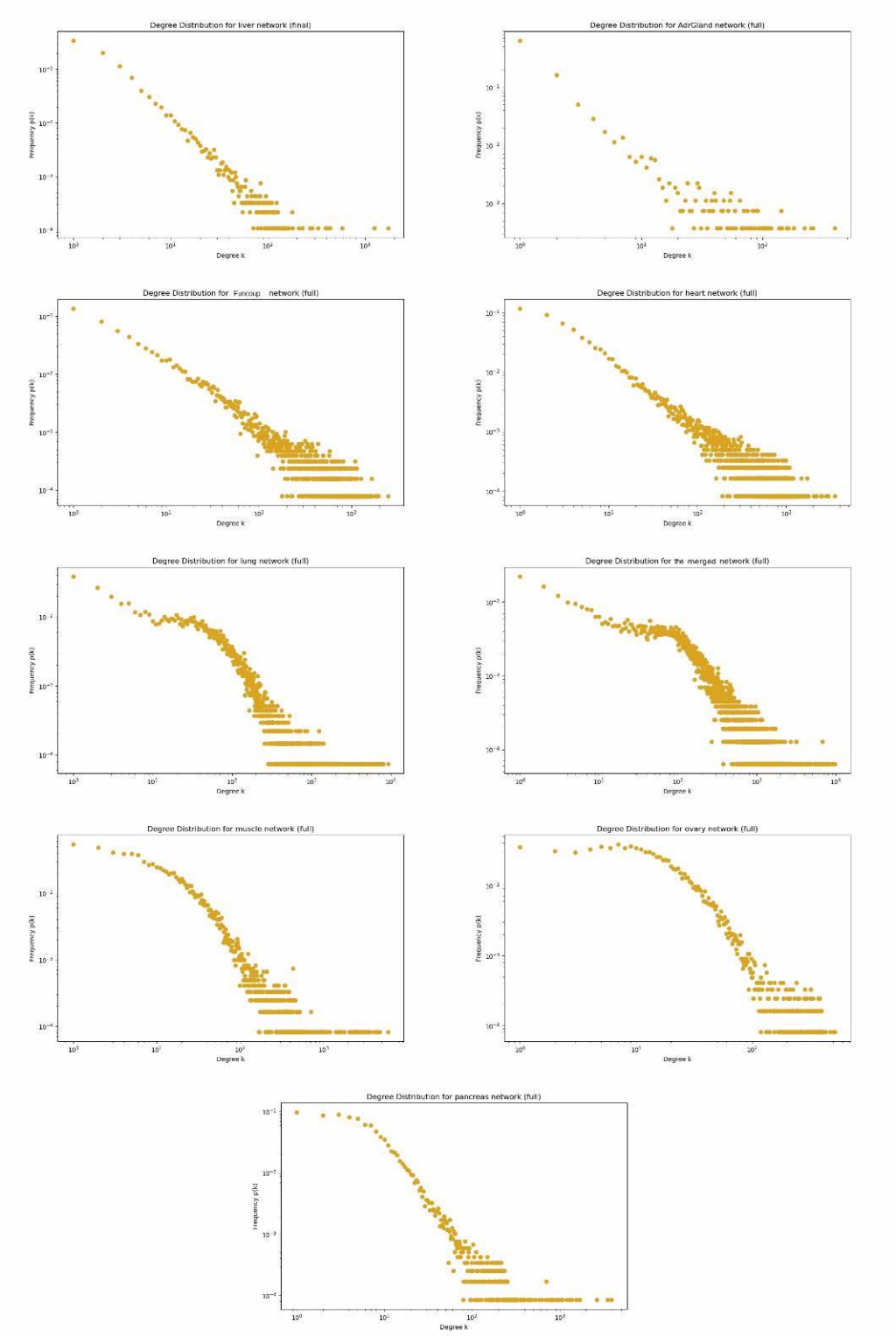

Figure 11 - Network degree distribution. All tissues-specific networks follow a power law distribution, or approximate it, and can be defined as scale free.

3.5 – Network comparison

Following an in-depth analysis of the network's primary features, we proceeded to compute similarity measures between their nodes and edges (Figure 12). The Szymkiewicz–Simpson coefficient (SS) proved particularly valuable for comparing networks of differing dimensions. This is because the denominator in the SS formula corresponds to the number of nodes in the smaller network, making the similarity metric less dependent on dimension and more focused on content. From the SS matrix, the “merged” network exhibits perfect node similarity with all the other networks, as they essentially function as subsets. However, the situation differs when considering the Jaccard similarity matrix, where dimension plays a substantial role. Here, smaller networks like the Adrenal Gland and Liver display no similarity with the others.

The expectation was that these heatmaps would unveil meaningful biological relationships. For instance, one might anticipate that tissues with similar functions or tissues from related organ systems (e.g., heart and muscle; glands like pancreas, ovary, and adrenal) would exhibit comparable functional associations between genes, as they often perform analogous roles. Regrettably, these plots fail to highlight notable relationships, which can be attributed to several factors, including the network's low coverage and the low filtering (LLR > 0.5). Consequently, the heatmaps hardly provide us with biological information about the networks.

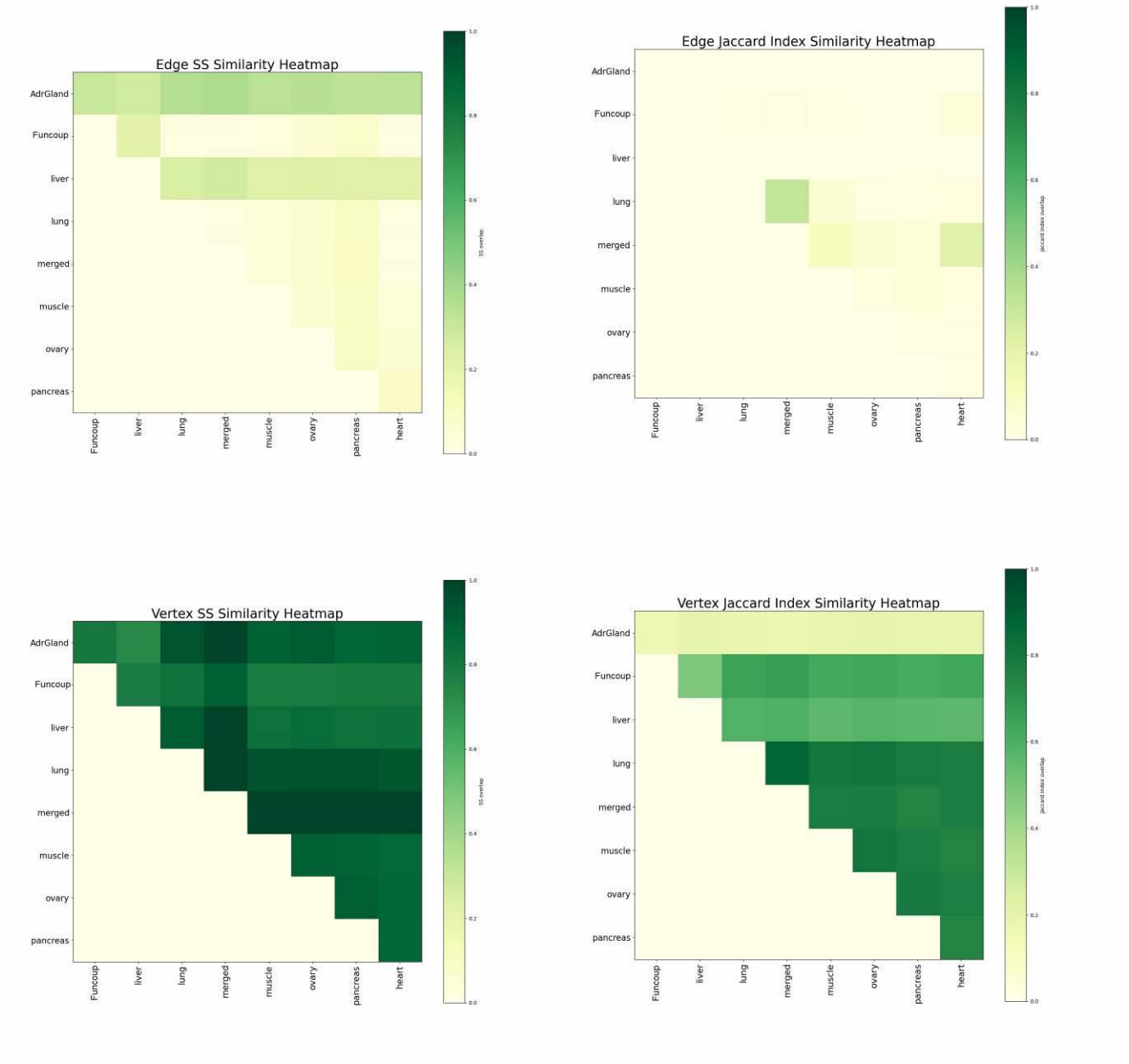

Figure 12 – Similarity comparison between networks with both Jaccard index (on the right) and SS similarity (on the left). The upper pair of heatmaps compares edges, while the lower pair compares nodes.

3.6 – Significant pathways

Up to this point, our analysis has predominantly emphasized network coverage, but a more important aspect lies in its statistical significance. To address this critical dimension, we computed p-values by extracting sub-graphs consisting of 323 known pathways sourced from KEGG within the network. The lengths of the LCC of these sub-graphs were then compared with LCCs of randomized subgraphs of equal dimensions. In an ideal scenario, where the network perfectly mirrors the biological state of a cell, we would anticipate that most of these sub-graphs exhibit high connectivity, resulting in elevated p-values.

As a reference benchmark, we subjected the FunCoup5 database to the same testing procedure, revealing over 200 significant pathways, thereby highlighting the expected statistical robustness of an established network (Figure 13). However, when considering our tissue-specific networks, none of them exceeded 50 significant pathways. The heart network boasted the highest count at 42, while the Adrenal Gland network featured the fewest at just 5. It's worth noting that, in part, these results also align with the relative coverage of these networks.

To assess biological specificity, we identified common pathways among each pair of networks and presented them in a heatmap (Figure 14). This approach provides more informative and realistic insights compared to previous similarity comparisons. For example, the pancreas and ovary tissues, both being glands, exhibit an high similarity in pathway expression, reflecting their shared biological characteristics.

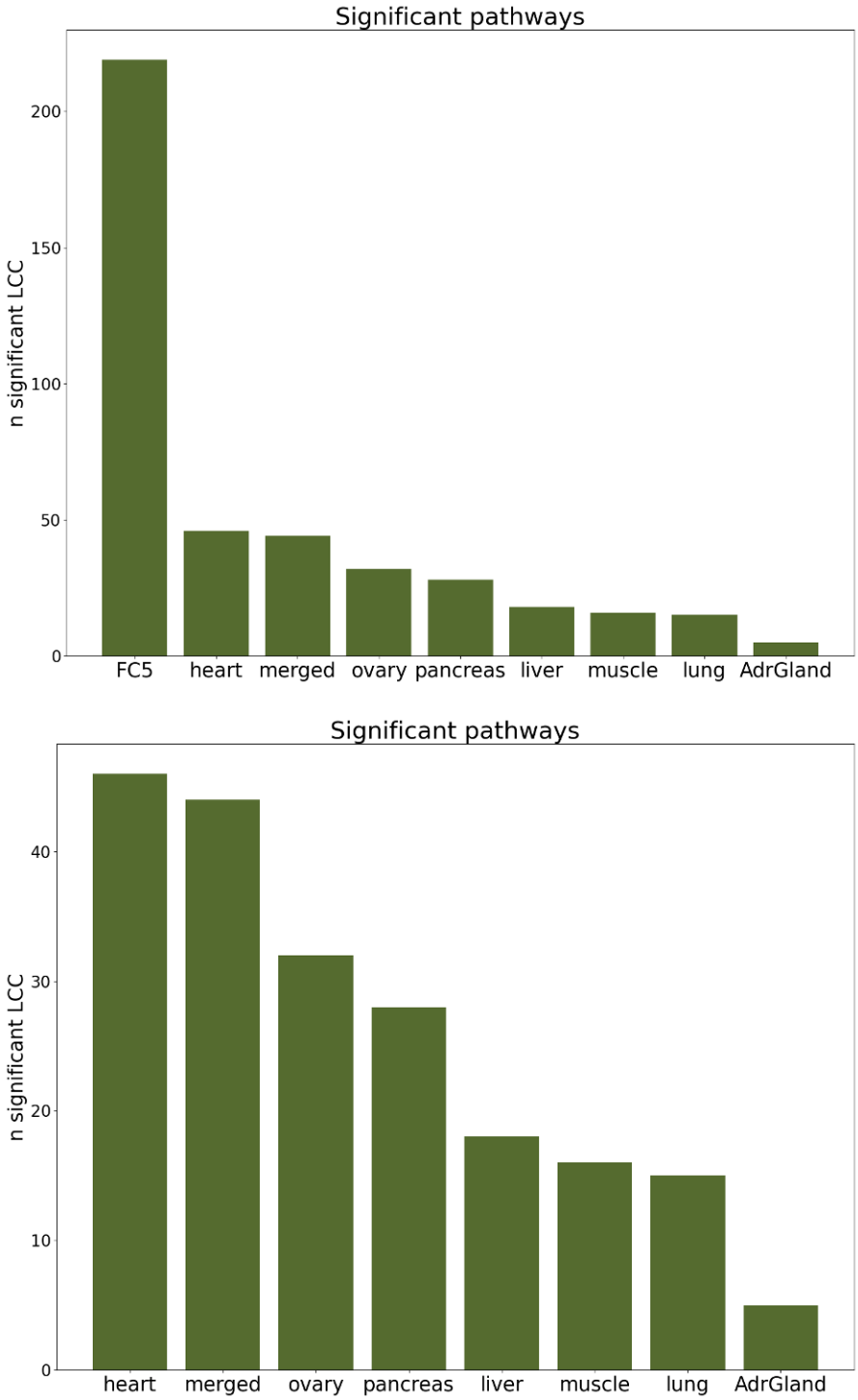

Figure 13 – Number of significant pathways found in each network. First histogram includes FunCoup5; the second one does not for scaling purposes.

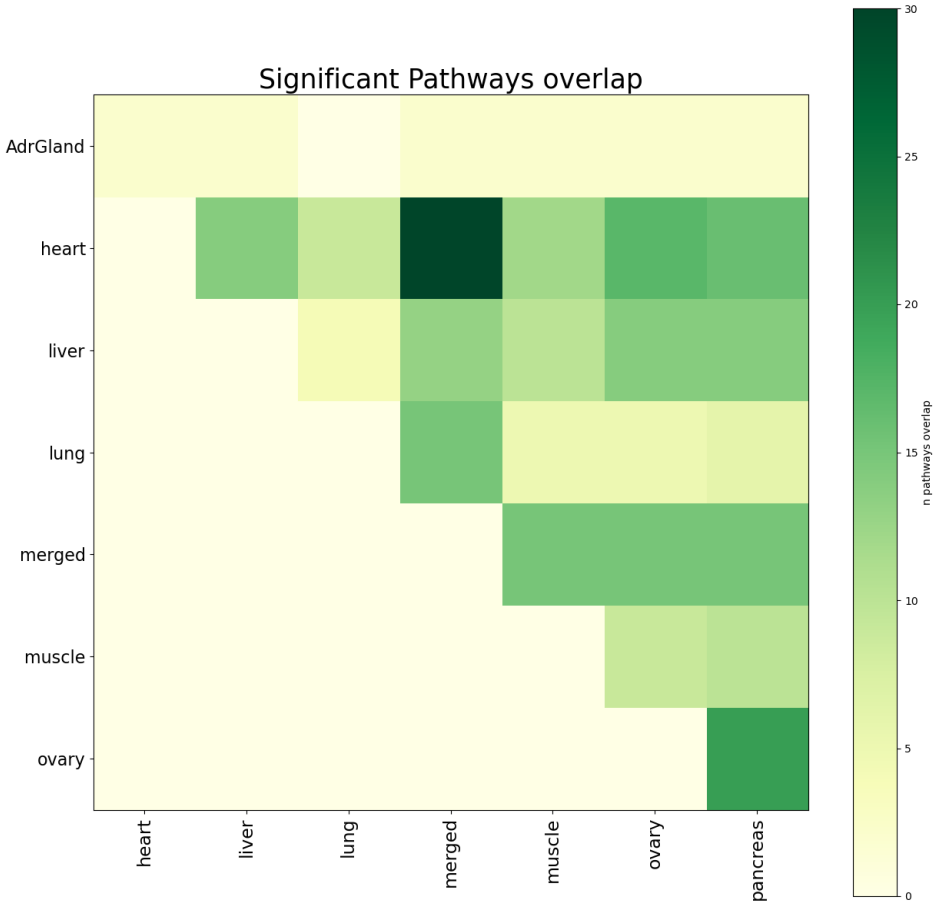

Figure 14 – Common significant pathways between networks. Darker colours indicate higher similarity.

4 - Conclusions

This project serves as an initial exploration for what regards tissue-specific network generation and should primarily be referenced as a foundation for future research, rather than a source of definitive conclusions. Even so, from the results, several key takeaways emerge. Despite the inherent challenges posed by single-cell data, including its high sparsity and generally low quality, we have demonstrated that it is possible to use them to construct networks that retain certain significance and biological differences.

To enhance the efficacy of such procedure, future studies could address the current limitations in data cleaning methods, aiming to minimize data loss without introducing artifacts. In particular, the adoption of data-driven algorithms, applied to the specific use case, holds significant promise. This could help mitigate the flaws of using fixed thresholds, which may inadvertently exclude valuable data from certain cell types. Additionally, handling isoforms differently, by considering their tissue-specific expression patterns rather than relying solely on their arbitrary selection, could enhance the specificity of the resulting networks.

Drawing inspiration from the FunCoup5 database, further explorations might involve the integration of additional evidence types, global genomics, and phylogenetic studies, with the goal to increase link strength and coverage. This could allow us to increase the threshold for high confidence links and perform a fairer comparison with the database. However, even if these improvements would increase the significance of the study, a huge limit in the utilisation would still be the overall low availability of single cell data, which prevents us to generate networks for every tissue of the human body.

References

[1] F. Emmert-Streib and G. V. Glazko, “Network biology: a direct approach to study biological function.,” *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, vol. 3, no. 4, pp. 379–391, Jul. 2011, doi: 10.1002/wsbm.134.

[2] A.-L. Barabási and Z. N. Oltvai, “Network biology: understanding the cell’s functional organization,” *Nature Reviews Genetics*, vol. 5, no. 2, pp. 101–113, Feb. 2004, doi: 10.1038/nrg1272.

[3] Y. Zhang, E. D. Kolaczyk, and B. D. Spencer, “Estimating network degree distributions under sampling: An inverse problem, with applications to monitoring social media networks,” *The Annals of Applied Statistics*, vol. 9, no. 1, pp. 166–199, 2015, doi: 10.1214/14-aoas800.

[4] R. Albert and A.-L. Barabási, “Statistical mechanics of complex networks,” *Reviews of Modern Physics*, vol. 74, no. 1, pp. 47–97, Jan. 2001, doi: 10.1103/revmodphys.74.47.

[5] H. Ohsaki, K. Yagi, and M. Imase, “On the Effect of Scale-Free Structure of Network Topology on End-to-End Performance,” *International Symposium on Applications and the Internet*, pp. 12–12, Jan. 2007, doi: 10.1109/saint.2007.18.

[6] M. Kitsak *et al.*, “Stability of a giant connected component in a complex network.,” *Physical Review E*, vol. 97, no. 1, pp. 012309–012309, Jan. 2018, doi: 10.1103/physreve.97.012309.

[7] P. Brazhnik, Alberto de la Fuente, Alberto de la Fuente, A. de la Fuente, and P. Mendes, “Gene networks: how to put the function in genomics Paul Brazhnik, Alberto de la Fuente and Pedro Mendes,” *Trends in Biotechnology*, Jan. 2002, doi: 10.1016/s0167-7799(02)02053-x.

[8] R. Dannenfelser, N. R. Clark, and A. Ma’ayan, “Genes2FANs: connecting genes through functional association networks.,” *BMC Bioinformatics*, vol. 13, no. 1, pp. 156–156, Jul. 2012, doi: 10.1186/1471-2105-13-156.

[9] E. Persson, M. Castresana-Aguirre, D. Buzzao, D. Guala, and E. L. L. Sonnhammer, “FunCoup 5: Functional Association Networks in All Domains of Life, Supporting Directed Links and Tissue-Specificity.,” *Journal of Molecular Biology*, vol. 433, no. 11, pp. 166835–166835, 2021, doi: 10.1016/j.jmb.2021.166835.

[10] S. Ahmed, Shahjaman, M. Rana, Md. Masud Rana, and N. H. Mollah, “Robustification of Naïve Bayes Classifier and Its Application for Microarray Gene Expression Data

Analysis," *BioMed Research International*, vol. 2017, pp. 3020627–3020627, Aug. 2017, doi: 10.1155/2017/3020627.

[11] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian networks to analyze expression data," *Journal of Computational Biology*, vol. 7, pp. 601–620, Jan. 2000, doi: 10.1089/106652700750050961.

[12] H. Zhang and J. Su, "Naive Bayes for optimal ranking," *Journal of Experimental and Theoretical Artificial Intelligence*, vol. 20, no. 2, pp. 79–93, Jun. 2008, doi: 10.1080/09528130701476391.

[13] Arsalane Chouaib Guidoum and A. C. Guidoum, "Kernel Estimator and Bandwidth Selection for Density and its Derivatives: The kedd Package," *arXiv: Computation*, 2020.

[14] A. Z. Zambom and R. Dias, "A Review of Kernel Density Estimation with Applications to Econometrics," vol. 5, no. 1, pp. 20–42, Apr. 2013.

[15] Y.-C. Chen, "A Tutorial on Kernel Density Estimation and Recent Advances," vol. 1, no. 1, pp. 161–187, Dec. 2017, doi: 10.1080/24709360.2017.1396742.

[16] M. P. Wand, M. P. Wand, M. P. Wand, M. P. Wand, and J. C. F. Yu, "Density Estimation via Bayesian Inference Engines.," *arXiv: Machine Learning*, 2020, doi: 10.1007/s10182-021-00422-8.

[17] H. Hang and Hanyuan Hang, "Histogram Transform Ensembles for Density Estimation," *arXiv: Statistics Theory*, Nov. 2019.

[18] Z. I. Botev, J. F. Grotowski, and D. P. Kroese, "Kernel density estimation via diffusion," *Annals of Statistics*, Nov. 2010, doi: 10.1214/10-aos799.

[19] Joseph A. Gallego-Mejia, J. Osorio, and Fabio A. Gonz'alez, "Fast Kernel Density Estimation with Density Matrices and Random Fourier Features," *Ibero-American Conference on AI*, 2022, doi: 10.48550/arxiv.2208.01206.

[20] T. Schmitt, C. Ogris, and E. L. L. Sonnhammer, "FunCoup 3.0: database of genome-wide functional coupling networks," *Nucleic Acids Research*, vol. 42, pp. 380–388, Jan. 2014, doi: 10.1093/nar/gkt984.

[21] S. Razick, G. Magklaras, and I. M. Donaldson, "iRefIndex: A consolidated protein interaction database with provenance," *BMC Bioinformatics*, vol. 9, no. 1, p. 405, Sep. 2008, doi: 10.1186/1471-2105-9-405.

[22] “Tissue-based map of the human proteome | Science.” https://www.science.org/doi/10.1126/science.1260419 (accessed Sep. 07, 2023).

[23] “Gene Ontology knowledgebase in 2023 | Genetics | Oxford Academic.” https://academic.oup.com/genetics/article/224/1/iyad031/7068118?login=false (accessed Sep. 07, 2023).

[24] P. Björkholm and E. Sonnhammer, “Comparative analysis and unification of domain-domain interaction networks,” *Bioinformatics (Oxford, England)*, vol. 25, pp. 3020–5, Sep. 2009, doi: 10.1093/bioinformatics/btp522.

[25] Q. Huang, D. Szklarczyk, M. Wang, M. Simonovic, and C. von Mering, “PaxDb 5.0: curated protein quantification data suggests adaptive proteome changes in yeasts,” *Molecular & Cellular Proteomics*, vol. 0, no. 0, Aug. 2023, doi: 10.1016/j.mcpro.2023.100640.

[26] Y. Luo *et al.*, “New developments on the Encyclopedia of DNA Elements (ENCODE) data portal,” *Nucleic Acids Res*, vol. 48, no. D1, pp. D882–D889, Jan. 2020, doi: 10.1093/nar/gkz1062.

[27] A. Alexeyenko and E. L. L. Sonnhammer, “Global networks of functional coupling in eukaryotes from comprehensive data integration.,” *Genome Research*, vol. 19, no. 6, pp. 1107–1116, Jun. 2009, doi: 10.1101/gr.087528.108.

[28] B. H. M. Meldal *et al.*, “The complex portal - an encyclopaedia of macromolecular complexes,” *Nucleic Acids Res*, vol. 43, no. Database issue, pp. D479–D484, Jan. 2015, doi: 10.1093/nar/gku975.

[29] A. Ruepp *et al.*, “CORUM: the comprehensive resource of mammalian protein complexes,” *Nucleic Acids Res*, vol. 36, no. Database issue, pp. D646–D650, Jan. 2008, doi: 10.1093/nar/gkm936.

[30] M. Kanehisa and S. Goto, “KEGG: kyoto encyclopedia of genes and genomes,” *Nucleic Acids Res*, vol. 28, no. 1, pp. 27–30, Jan. 2000, doi: 10.1093/nar/28.1.27.

[31] M. Pertea, K. Ayanbule, M. Smedinghoff, and S. L. Salzberg, “OperonDB: a comprehensive database of predicted operons in microbial genomes,” *Nucleic Acids Res*, vol. 37, no. Database issue, pp. D479–D482, Jan. 2009, doi: 10.1093/nar/gkn784.

[32] “The Genotype-Tissue Expression (GTEx) project,” *Nat Genet*, vol. 45, no. 6, pp. 580–585, Jun. 2013, doi: 10.1038/ng.2653.

[33] “FANTOM5 CAGE profiles of human and mouse samples | Scientific Data.” https://www.nature.com/articles/sdata2017112 (accessed Sep. 07, 2023).

[34] A. K. Wong, A. Krishnan, and O. G. Troyanskaya, “GIANT 2.0: genome-scale integrated analysis of gene networks in tissues,” *Nucleic Acids Res*, vol. 46, no. Web Server issue, pp. W65–W70, Jul. 2018, doi: 10.1093/nar/gky408.

[35] C. S. Greene *et al.*, “Understanding multicellular function and disease with human tissue-specific networks,” *Nat Genet*, vol. 47, no. 6, Art. no. 6, Jun. 2015, doi: 10.1038/ng.3259.

[36] D. Jovic, X. Liang, H. Zeng, L. Lin, F. Xu, and Y. Luo, “Single-cell RNA sequencing technologies and applications: A brief overview,” *Clin Transl Med*, vol. 12, no. 3, p. e694, Mar. 2022, doi: 10.1002/ctm2.694.

[37] “A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications | Genome Medicine | Full Text.” https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-017-0467-4 (accessed Sep. 07, 2023).

[38] G. Chen, B. Ning, and T. Shi, “Single-Cell RNA-Seq Technologies and Related Computational Data Analysis,” *Front Genet*, vol. 10, p. 317, Apr. 2019, doi: 10.3389/fgene.2019.00317.

[39] B. Hwang, J. H. Lee, and D. Bang, “Single-cell RNA sequencing technologies and bioinformatics pipelines,” *Exp Mol Med*, vol. 50, no. 8, Art. no. 8, Aug. 2018, doi: 10.1038/s12276-018-0071-8.

[40] “Single Cell Analysis – Advantages, Challenges, and Applications,” *Drug Discovery from Technology Networks*. http://www.technologynetworks.com/drug-discovery/blog/single-cell-analysis-advantages-challenges-and-applications-322768 (accessed Sep. 07, 2023).

[41] A. Subramanian, M. Alperovich, Y. Yang, and B. Li, “Biology-inspired data-driven quality control for scientific discovery in single-cell transcriptomics,” *Genome Biology*, vol. 23, no. 1, p. 267, Dec. 2022, doi: 10.1186/s13059-022-02820-w.

[42] “The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans | Science.” https://www.science.org/stoken/author-tokens/ST-495/full (accessed Sep. 07, 2023).

[43] A. Regev *et al.*, “The Human Cell Atlas,” *Elife*, vol. 6, p. e27041, Dec. 2017, doi: 10.7554/eLife.27041.

[44] I. Virshup, S. Rybakov, F. J. Theis, P. Angerer, and F. A. Wolf, “anndata: Annotated data.” bioRxiv, p. 2021.12.16.473007, Dec. 19, 2021. doi: 10.1101/2021.12.16.473007.

[45] The UniProt Consortium, “UniProt: the Universal Protein Knowledgebase in 2023,” *Nucleic Acids Research*, vol. 51, no. D1, pp. D523–D531, Jan. 2023, doi: 10.1093/nar/gkac1052.

[46] T. R. Mercer *et al.*, “The human mitochondrial transcriptome,” *Cell*, vol. 146, no. 4, pp. 645–658, Aug. 2011, doi: 10.1016/j.cell.2011.06.051.

[47] D. Osorio and J. J. Cai, “Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell RNA-sequencing data quality control,” *Bioinformatics*, vol. 37, no. 7, pp. 963–967, Aug. 2020, doi: 10.1093/bioinformatics/btaa751.

[48] C. R. Harris *et al.*, “Array programming with NumPy,” *Nature*, vol. 585, no. 7825, Art. no. 7825, Sep. 2020, doi: 10.1038/s41586-020-2649-2.

[49] J. D. Hunter, “Matplotlib: A 2D Graphics Environment,” *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, May 2007, doi: 10.1109/MCSE.2007.55.

[50] P. Virtanen *et al.*, “SciPy 1.0: fundamental algorithms for scientific computing in Python,” *Nat Methods*, vol. 17, no. 3, Art. no. 3, Mar. 2020, doi: 10.1038/s41592-019-0686-2.

[51] A. Hagberg, P. J. Swart, and D. A. Schult, “Exploring network structure, dynamics, and function using NetworkX,” Los Alamos National Laboratory (LANL), Los Alamos, NM (United States), LA-UR-08-05495; LA-UR-08-5495, Jan. 2008. Accessed: Sep. 07, 2023. [Online]. Available: https://www.osti.gov/biblio/960616

[52] S. Seabold and J. Perktold, “Statsmodels: Econometric and Statistical Modeling with Python,” presented at the Python in Science Conference, Austin, Texas, 2010, pp. 92–96. doi: 10.25080/Majora-92bf1922-011.