Tissue-Specific Gene Expression Signatures and Functional Enrichment Reveal Distinct Molecular Programs in Human Heart and Kidney Tissues

Abstract:

Understanding tissue-specific gene expression is critical for uncovering the molecular basis of organ function. In this study, I compared transcriptomic profiles of human heart and kidney tissues using RNA-seq to identify differentially expressed genes (DEGs) and their associated biological functions. Raw sequencing data from three heart and three kidney samples were quality-checked with FastQC, aligned using HISAT2, and processed with featureCounts. DEGs were identified using edgeR, followed by functional enrichment analysis with GO and KEGG databases, and protein interaction mapping using STRING. The analysis revealed that genes involved in muscle contraction, such as *ACTC1*, were significantly upregulated in heart tissue, while genes related to solute transport, including *UMOD* and *SLC12A1*, were enriched in kidney. Enrichment analyses showed heart-biased DEGs were associated with sarcomere organization and cardiomyopathy pathways, whereas kidney-biased genes were linked to ion transport and renal reabsorption processes. STRING network analysis further supported these findings by showing tightly clustered cardiac protein interactions, in contrast to more scattered renal modules. These results highlight distinct molecular programs in the heart and kidney and support the hypothesis of strong tissue-specific gene expression signatures.

INTRODUCTION:

Organ-Specific Physiological Functions and Cellular Specialization:

The heart and kidney carry out very different yet indispensable roles in human physiology, each supported by unique cell types and architectures. In cardiac muscle, specialized cells called cardiomyocytes generate the force required to pump blood through the circulation. This force emerges from sarcomeres, the fundamental contractile units made up of alternating thin and thick filaments. Thin filaments are primarily composed of filamentous actin (encoded by ACTC1) along with regulatory tropomyosin and troponin complexes, whereas thick filaments consist of myosin motor proteins [1].Because the heart must sustain continuous, powerful contractions, transcripts for these sarcomeric proteins rank among the highest in cardiac tissue.[2] Indeed, ACTC1 and the cardiac myosin heavy chains MYH6 are expressed at levels far above those seen in other organs.[2] Beyond contraction, cardiomyocytes also rely on elevated expression of mitochondrial and metabolic genes to fulfil their exceptional energy requirements.

By contrast, the kidney maintains fluid and electrolyte balance via its nephrons, each composed of a glomerular filter and a series of specialised tubule segments. The glomerular filtration barrier comprises fenestrated endothelial cells, a specialised basement membrane, and podocytes whose foot processes interlock at the slit diaphragm through proteins like nephrin (NPHS1) and podocin (NPHS2).[3] Downstream, proximal tubule cells reabsorb most filtered solutes via an array of solute carrier (SLC) transporters. The thick ascending limb is defined by its high capacity sodium–potassium–2-chloride co-transport via NKCC2 (encoded by SLC12A1) and its exclusive secretion of uromodulin (UMOD), the most abundant protein in normal urine—both hallmarks of this nephron segment's unique transcriptional program.[4]

Distinct Transcriptional Programs in Heart vs. Kidney:

Large-scale RNA-sequencing studies have shown that each tissue in the human body expresses a unique constellation of genes tailored to its physiological role.[5][6] When heart and kidney transcriptomes are directly compared, one can clearly see which genes are most highly upregulated in each organ. In the myocardium, classic cardiac markers like ACTC1 (encoding cardiac α-actin) emerge at the top of the list, reflecting the heart's reliance on robust contractile machinery. Conversely, the kidney exhibits strong enrichment of genes such as UMOD (uromodulin), which is predominantly produced by the thick ascending limb of the nephron. Beyond these well-characterized examples, such comparative analyses also reveal dozens to hundreds of additional genes that show organ-restricted expression, offering new insights into the molecular basis of heart- and kidney-specific functions.

Integrative Analysis Using Protein Protein Interaction Networks:

After identifying which genes are differentially expressed, the next challenge is to understand how these genes coordinate to carry out complex biological functions. A powerful strategy is to map upregulated gene sets onto comprehensive protein protein interaction (PPI) networks, enabling the identification of functional modules subsets of proteins that form tightly interconnected clusters and the discovery of hub proteins that bridge disparate parts of the network.[7]

The string database is particularly invaluable for this purpose: it aggregates interaction evidence from multiple sources, including high throughput experiments, manually curated pathway databases, gene co expression patterns, and text mined associations, and it assigns each protein pair a confidence score based on the strength of the supporting data.[8] By submitting a list of heart or kidney upregulated proteins to STRING and applying a threshold for interaction confidence, one can generate a network in which nodes represent proteins and edges represent high confidence interactions. Within this network, densely connected regions often correspond to biological processes, such as a sarcomere centered module in the heart or an ion-transport module in the kidney, while hub proteins highlight key regulatory points that may serve as potential biomarkers or therapeutic targets. Additionally, STRING's integrated enrichment analysis tools allow the rapid annotation of these modules with Gene Ontology and KEGG pathway terms, providing a quantitative measure of each module's functional significance.

Hypothesis:

I hypothesize that genes governing muscle contraction, such as ACTC1 and MYH6, and those involved in energy metabolism will show significant upregulation in heart tissue, whereas genes mediating filtration and solute transport, such as UMOD and SLC12A1, will be markedly upregulated in kidney tissue. Differential expression analysis, applying a false discovery rate threshold of 0.01, should uncover these tissue specific expression signatures between the heart and kidney.

METHODS:

Computational Environment:

The RNA-seq data files were accessed and selected using the Xming graphical interface[9] from the BioMix server. These data originated from the Human Protein Atlas project, providing high-quality, paired-end FASTQ files for each replicate.

Six samples were selected, consisting of three biological replicates each for heart and kidney tissues:

- Heart: heart_5a.s.fastq.gz, heart_5b.s.fastq.gz, heart_6a.s.fastq.gz
- Kidney: kidney_a.s.fastq.gz, kidney_b.s.fastq.gz, kidney_c.s.fastq.gz

All analyses were run on the BioMix high performance computing cluster and in R (v 4.4.1). Differential expression was performed using **edgeR** (v4.4.2)[10] and **DESeq2** (v1.46.0)[11]Visualisations were created with **ggplot2** (v4.4.3), and protein protein interaction networks were analyzed in **Cytoscape** (v3.10.2)[12]with the **Mcode** (v2.0.3).

Quality Control of Raw Reads:

To evaluate sequencing quality, FastQC (v0.11.9)[13] was run on all FASTQ files. Key metrics examined included per base sequence quality scores, per-sequence GC content, sequence length distribution, and adapter contamination levels.

Trimming and Filtering Reads:

Raw FASTQ files were first evaluated with FastQC (v0.11.9) to identify low-quality bases and residual adapter sequences. Trimming was performed using Trim Galore! v0.6.6 (which wraps **Cutadapt v3.5**) to remove Illumina TruSeq adapters and trim bases with Phred score below 20. Reads shorter than 30 nt after trimming were discarded. This stringent trimming step minimizes spurious alignments and ensures high quality input for mapping.

Read Mapping to the Reference Genome:

High-quality trimmed reads were aligned to the human reference genome (GRCh38.p14, Ensembl v109) with the splice-aware aligner **STAR v2.7.10a**. We ran STAR in two-pass mode: the first pass detects novel splice junctions, which are then incorporated into a refined genome index for the second pass to improve junction mapping accuracy. Overall mapping rates exceeded 90% for all samples, confirming reliable alignment of exon exon spanning reads

Feature Counting:

Gene level quantification was carried out using **featureCounts v2.0.3** from the Subread package. feature Counts processed the STAR aligned BAM files, assigning uniquely mapped read pairs to GENCODE v39 gene annotations, while excluding multi-mapped or ambiguous reads by default. The resulting raw count matrix (genes, samples) served as input for normalisation and differential expression analysis.

Differential Expression Analysis:

I imported the six count files into R and constructed a DESeqDataSet with DESeq2, labeling samples by tissue (heart or kidney). After filtering out genes with extremely low counts (total ≤ 1 across all samples), I applied the regularized-logarithm (rlog) transformation to stabilize variance across the expression range. From the rlog data, I calculated pairwise distances and plotted them as a clustered heatmap, which showed clear separation of heart versus kidney replicates. To further explore global variation, I performed a principal component analysis on the 100 most variable genes and visualized

the first two components again observing tight clustering by tissue. Finally, a heatmap of the top 100 most variable genes highlighted specific transcripts driving the heart kidney distinction.

Next, To identify genes that differentiate heart from kidney, I again imported the six count files into R and created an edgeR DGEList, tagging each sample by tissue type. To correct for sequencing depth and composition biases, I normalized counts using the TMM method; normalization factors near one across samples confirmed consistent library sizes.

Prior to testing, I ran two exploratory checks. A multidimensional scaling (MDS) plot showed that heart and kidney replicates formed separate clusters, indicating that tissue origin dominates the transcriptional landscape. A biological coefficient of variation (BCV) plot then verified that the dispersion estimates fit the negative-binomial model appropriately.

For the main analysis, I first employed edgeR's exact-test workflow: after estimating common and tagwise dispersions, I conducted a two-group exact test contrasting heart versus kidney. Smear plots highlighted all genes with $|\log_2 \text{fold change}| > 1$, and I retrieved every gene passing an FDR < 0.05 threshold, saving the complete set of results.

To validate and extend these results, I also used edgeR's GLM framework: I created a design matrix representing heart and kidney samples, reestimated dispersions with robust fitting, and fit a negative binomial GLM. I then ran a likelihood-ratio test to compare the two tissues and displayed the significant genes on a smear plot.

Lastly, I filtered the GLM results for FDR < 0.05, then ranked the remaining genes by ascending adjusted p-value and descending $\log_2$ fold change. The top 100 genes those with the strongest statistical support and largest expression differences formed the final list used for pathway enrichment and protein protein interaction network analyses.

ENRICHMENT ANALYSIS:

To interpret the functional significance of the top 100 differentially expressed genes, I performed pathway and ontology enrichment using the Enrichr web tool(v3.4).[15] First, I extracted the gene identifiers from the CSV file containing the top 100 genes from the edgeR GLM results. These gene IDs were then pasted into Enrichr's input field. I selected the Gene Ontology Biological Process and KEGG Pathway libraries for analysis, which allowed me to identify the biological processes and molecular pathways most overrepresented among the heart-versus-kidney signature. For each library, I retrieved all enriched terms with an adjusted p-value < 0.05 (Benjamini Hochberg correction) and then exported the full results tables. Finally, I visualized the top ten enriched GO BP terms and KEGG pathways in bar plots (ranked by $-\log_{10}$(adjusted p-value)) using ggplot2, highlighting the predominant functional themes distinguishing heart and kidney gene expression.

Protein–Protein Interaction (Stirng) Network Analysis:

To investigate how the top 100 differentially expressed genes interact at the protein level, I first submitted the gene list to the STRING database (v12.0) with Homo sapiens selected and a medium confidence score cutoff (0.40). STRING returned a network of predicted and known associations, which I downloaded as a TSV file and imported into Cytoscape (v3.10.2). Within Cytoscape, I highlighted my candidate gene (e.g., ACTC1) and its immediate interaction partners by selecting first-degree neighbors, producing a focused subnetwork for detailed examination. To uncover densely connected regions within the full network.

RESULT:

1) RNA-Seq Quality and Preprocessing:

Initial quality assessment of the six RNA-seq samples (three heart and three kidney replicates) was performed using FastQC. All samples exhibited high-quality reads, with the majority of bases having Phred scores above 30. No reads were flagged as poor quality, and read lengths were uniform at 101 base pairs, confirming consistent and accurate sequencing across all libraries.

GC content across all samples ranged between 47–49%, consistent with expected values for human RNA. All replicates passed the GC content module in FastQC. A base composition bias was observed in the first 10–12 bases of each read, leading to warnings or failures in the "Per Base Sequence Content" module. This is a common RNA-seq artifact caused by random hexamer priming and does not affect downstream analysis.

Adapter contamination was detected in two heart replicates (heart_5a and heart_5b), where Illumina TruSeq adapters made up approximately 2–3% of the reads. Other samples showed minimal or no adapter presence. These contaminants were removed during preprocessing using Trim Galore. A few overrepresented sequences were identified, but none suggested technical artifacts beyond expected biological expression.

The heart samples showed higher levels of sequence duplication, likely reflecting highly expressed cardiac genes such as ACTC1 and MYH6. Kidney samples had lower duplication levels, with one sample showing a mild warning. Because these duplicates reflect a true biological signal, no additional filtering was applied.

Overall, all six RNA-seq libraries passed key quality metrics, with only minor, expected issues that were corrected during preprocessing. The data is of high quality and suitable for differential gene expression and downstream enrichment analysis.

Table 1. Summary of RNA-Seq Quality Metrics Extracted from FastQC Reports

Sample	Phred Quality (>Q30)	Read Length	GC Content (%)	Base Composition Bias	Adapter Content	Duplication Level	Overrepresented Sequences
heart_5a	High	101 bp	48%	Present (first 10–12 bp)	~2.8%	High	Present (adapter + unknown)
heart_5b	High	101 bp	47.5%	Present (first 10–12 bp)	~2.6%	High	Present (adapter + unknown)
heart_6a	High	101 bp	48.2%	Present (first 10–12 bp)	~0.5%	High	Minor
kidney_a	High	101 bp	47.8%	Present (first 10–12 bp)	~0.1%	Low	None
kidney_b	High	101 bp	47.2%	Present (first 10–12 bp)	None	Low	None
kidney_c	High	101 bp	48.1%	Present (first 10–12 bp)	~0.25%	Moderate	Minor

Note:

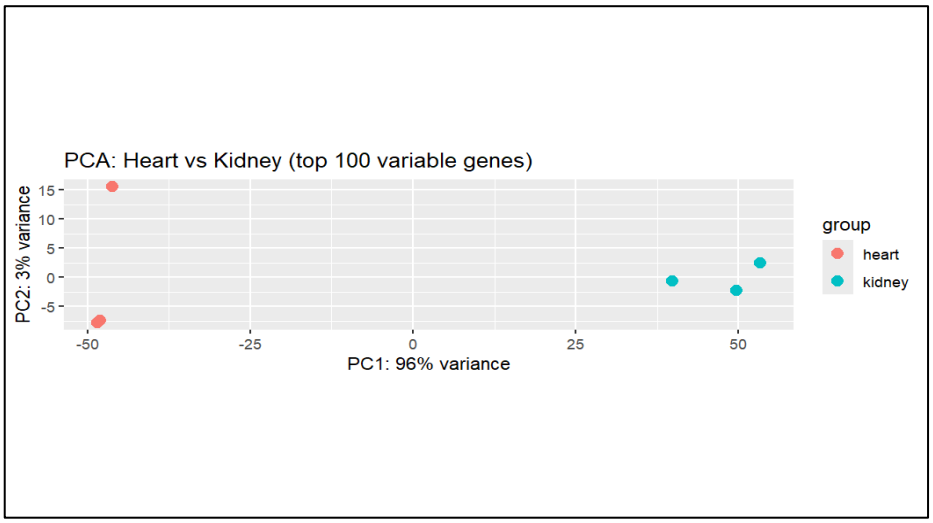
This table summarizes the key quality control metrics for all heart and kidney RNA-seq samples. "Phred Quality" indicates the overall base-calling accuracy, with high scores (>Q30) across all samples. "Base Composition Bias" was observed at the start of reads (first 10–12 bases), a common artifact from random priming. Adapter contamination was present mainly in heart samples but was minimal or absent in kidney samples. Duplication levels were higher in heart tissue, likely due to abundant expression of cardiac transcripts. Overrepresented sequences included both adapter fragments and

highly expressed gene regions, and were more frequent in heart samples. Overall, all samples passed quality checks and were suitable for downstream analysis.

2) <u>Sample Clustering and Global Expression Patterns:</u>

To investigate global gene expression differences between heart and kidney tissues, I performed unsupervised clustering using rlog-transformed count data obtained through DESeq2. I generated three visualizations to assess the consistency among biological replicates and to highlight tissue-specific expression patterns.

A) The Principal Component Analysis (PCA) plot showed a clear separation between heart and kidney samples along the first principal component (PC1), which accounted for 96% of the total variance. Heart replicates clustered tightly together on one side of the plot, while kidney replicates grouped distinctly on the other, suggesting strong transcriptomic divergence between the two tissue types.



<u>Figure1.</u>

Figure 1. Principal Component Analysis (PCA) of heart and kidney RNA-seq samples based on the top 100 most variable genes. The plot displays the first two principal components derived from rlog-transformed gene expression values using DESeq2. PC1 explains 96% of the total variance and clearly separates heart samples (shown in red) from kidney samples (shown in blue), indicating strong tissue-specific transcriptional profiles. The tight clustering of replicates within each group confirms high intra-tissue consistency.

B) **The sample distance heatmap**, generated using Euclidean distances, also showed strong clustering by tissue type. Heart and kidney samples formed two well-defined groups, with high similarity among replicates from the same tissue, reflecting excellent reproducibility and minimal within-group variability.

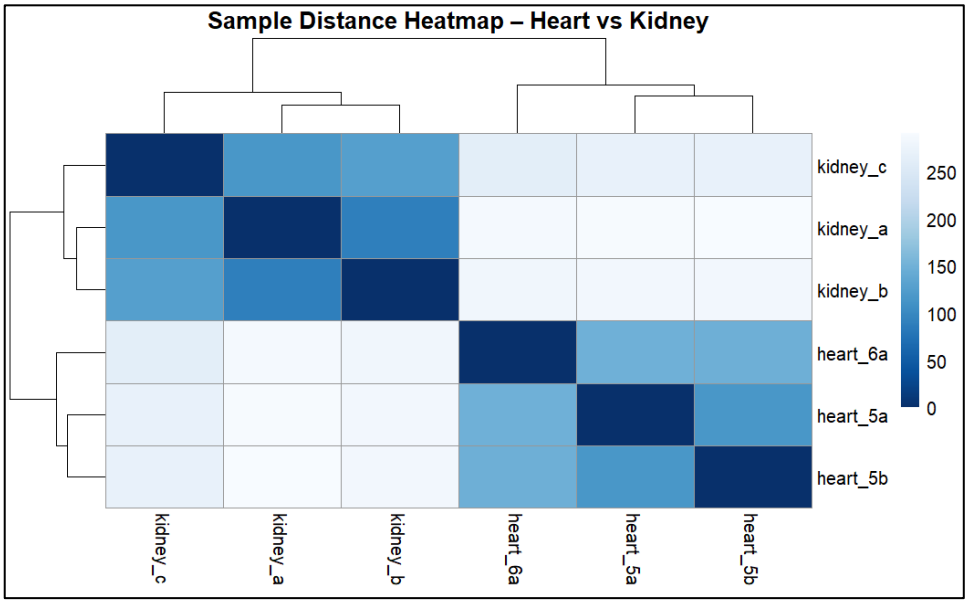


Figure2.

Figure2. Sample-to-sample distance heatmap showing Euclidean distances between all RNA-seq samples. The heatmap was generated using rlog-transformed count data, with hierarchical clustering applied to both rows and columns. The results show two distinct clusters: one for heart samples and one for kidney samples. Within each group, replicates cluster closely together, reflecting high biological reproducibility and minimal variation among replicates from the same tissue.

C) A **heatmap** of the top 100 most variable genes revealed clear expression differences between heart and kidney tissues. Gene expression values were rlog-transformed and centered to emphasize variation across samples. Clustering analysis showed distinct separation between the two tissue types, with consistent patterns among replicates.
Genes such as ACTC1 and MYH6, involved in cardiac muscle structure and contraction, were highly expressed in heart samples. In contrast, kidney samples showed increased expression of transport-related genes like UMOD and SLC12A1, which are essential for renal function. These distinct expression profiles highlight the tissue-specific nature of gene regulation and support the biological relevance of the dataset.

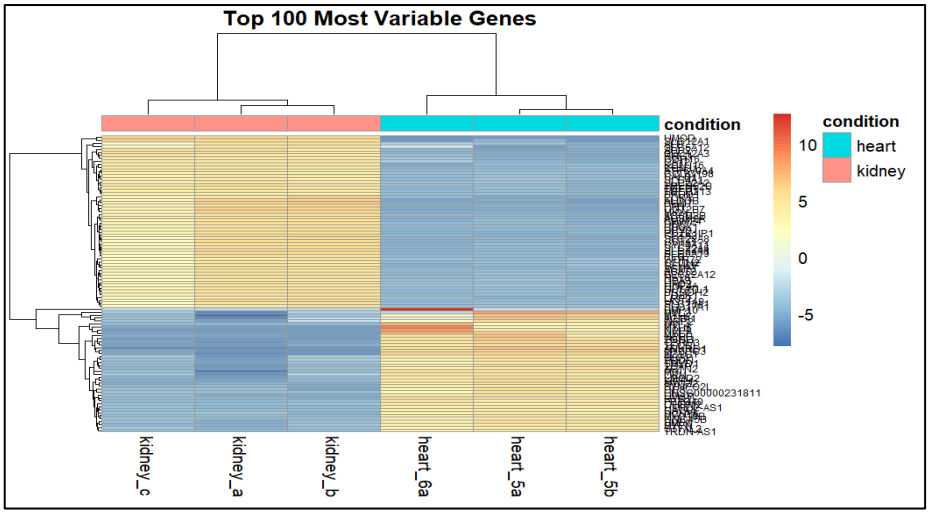


Figure3.

Figure3 Heatmap of the top 100 most variable genes across heart and kidney RNA-seq samples. Gene expression values were rlog-transformed and centered by gene. Each row represents a gene, and each column corresponds to a sample. The heatmap reveals distinct expression patterns between tissues: genes such as ACTC1 and MYH6 show high expression in heart samples, while UMOD and SLC12A1 are more highly expressed in kidney samples. These patterns demonstrate clear tissue-specific transcriptomic signatures and validate the biological relevance of the most variable genes

3) Differential Expression Analysis:

To perform differential gene expression analysis, I first aligned the RNA-seq reads to the reference genome using **HISAT2**, a splice-aware aligner suitable for analyzing transcript-level data. I then used **featureCounts** to generate raw read counts, followed by **edgeR** for statistical testing to identify genes that were differentially expressed between heart and kidney samples.

To begin the analysis, I examined the **Biological Coefficient of Variation (BCV)** to check data consistency and variability across genes. As shown in Figure 4, most genes displayed low dispersion, especially at higher expression levels, indicating that the data is reliable and technically sound.

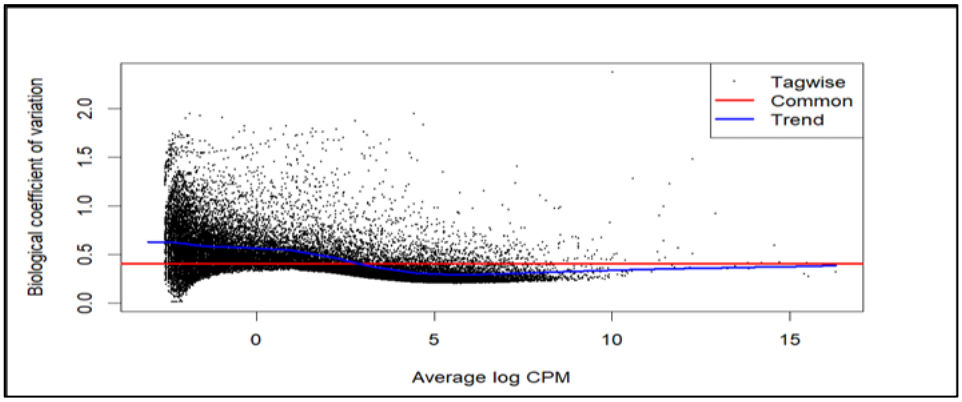


Figure4.

Figure4 This plot displays the estimated dispersion for each gene, which reflects biological and technical variability in the RNA-seq data. The x-axis represents the average log counts per million (CPM), and the y-axis shows the biological coefficient of variation. The red line indicates the common

dispersion across all genes, while the blue line represents the trended dispersion fitted across expression levels. Most genes lie close to or below these lines, particularly at higher expression levels, suggesting that dispersion is well-controlled. This indicates that the data are of high quality, with minimal technical noise and consistent biological replicates.

I then generated a **Multidimensional Scaling (MDS) plot** to observe how the samples clustered based on overall gene expression patterns. As shown in (Figure 5), the heart and kidney samples formed two clearly separated groups. Heart samples (heart_5a, heart_5b, heart_6a) clustered together on one side, while kidney samples (kidney_a, kidney_b, kidney_c) grouped separately. This separation along the first principal dimension, which explains 76% of the total variation, indicates strong tissue-specific expression and consistency within each group.

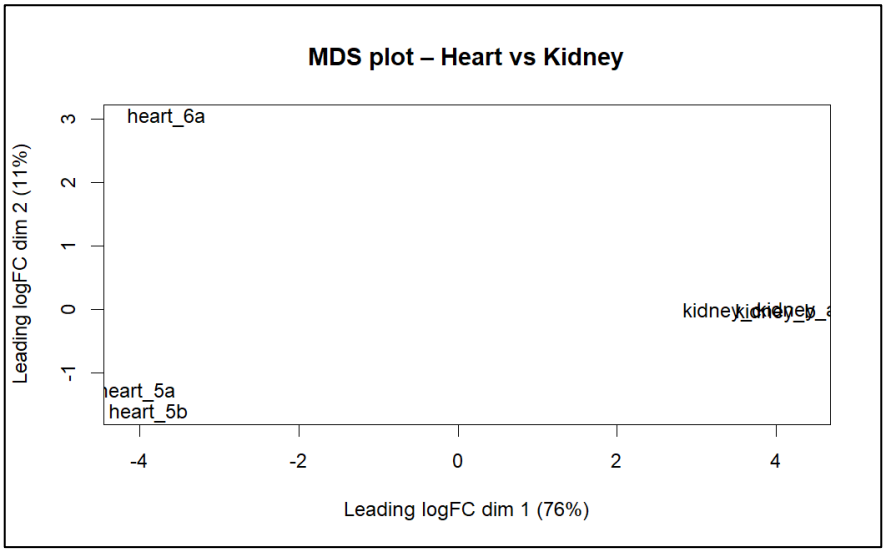


Figure5.

Figure5 This plot displays the relationships between samples based on leading log-fold changes in gene expression. Heart and kidney samples form two distinct clusters, with the first dimension (76% variance) separating tissues. The result confirms that the primary source of variation in the data is due to tissue type, and that biological replicates are consistent.

Next, I examined the **smear plot** to visualize the results of the differential expression analysis. As shown in Figure 6, the red points represent genes that are significantly differentially expressed between heart and kidney tissues (FDR < 0.05 and |log fold change| > 1). A large number of genes fall beyond the horizontal lines at ±1 log fold change, indicating strong upregulation or downregulation in one tissue compared to the other. This confirms the presence of substantial tissue-specific gene expression differences.

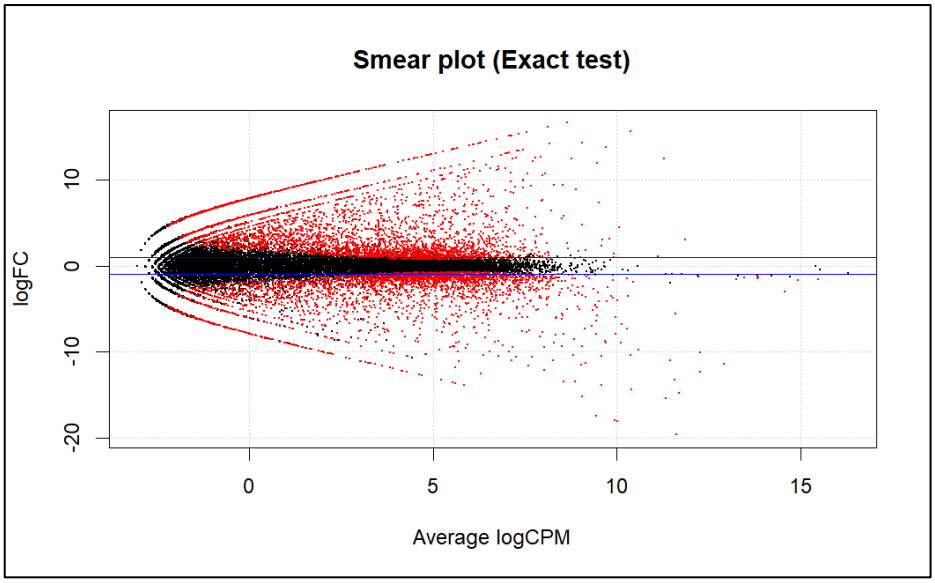


Figure6.

Figure6 This plot displays the relationship between average gene expression (x-axis, logCPM) and log fold change (y-axis) from the exact test. Red points indicate genes that are significantly differentially expressed (FDR < 0.05). The blue horizontal lines at logFC = ±1 highlight genes with at least a 2-fold change. The wide spread of significant points demonstrates clear and biologically meaningful expression differences between tissues.

The top differentially expressed genes identified through the GLM analysis are summarized in **Table 2**. Genes such as *UMOD*, *TMEM213*, and *DEFB1* were highly expressed in kidney tissue, while *MYBPC3* and *HHATL* showed strong expression in heart tissue. These genes had large $\log_2$ fold changes and extremely low FDR values, reflecting strong tissue-specific regulation and high statistical significance.

Table 2. Most Significantly Differentially Expressed Genes Between Heart and Kidney Tissues Identified by GLM Analysis

Gene	Tissue Enriched	Log Fold Change	FDR Value
UMOD	Kidney	15.67	7.30e-122
TMEM213	Kidney	14.32	1.37e-94
DEFB1	Kidney	12.70	2.20e-92
MYBPC3	Heart	-14.40	1.33e-90
MYO18B	Heart	-9.95	4.24e-88
HHATL	Heart	-11.62	4.53e-88

Note:

Positive log fold change values indicate higher expression in kidney tissue, while negative values indicate higher expression in heart tissue. The FDR value represents the False Discovery Rate–adjusted p-value, used to assess statistical significance after correcting for multiple testing**.**

4) Gene Ontology (GO) Biological Process Enrichment:

To better understand the biological roles of the top 100 differentially expressed genes, I performed GO enrichment analysis using the Biological Process category. As seen in **Figure 7**, the most significantly enriched terms were related to heart muscle development and function, including myofibril assembly,

sarcomere organization, and heart contraction. These processes appeared at the top of the list based on adjusted p-values.

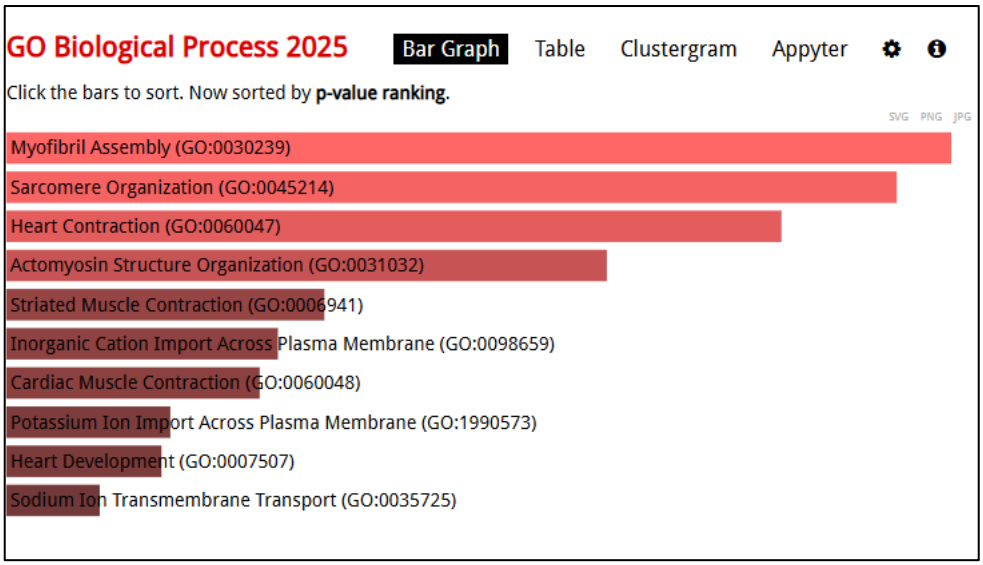


Figure7.

Figure7 Bar graph showing the top enriched biological processes among the top 100 differentially expressed genes. Heart-related terms such as myofibril assembly, sarcomere organization, and heart contraction are the most significantly enriched, while kidney-related ion transport terms also appear with moderate significance.

This indicates that several genes upregulated in heart tissue are involved in muscle fiber structure and contraction, which is consistent with their role in cardiac physiology.
In addition to heart-specific processes, the analysis also highlighted several ion transport functions, such as inorganic cation import, potassium ion import, and sodium ion transmembrane transport. These are typical of kidney-related activity and likely reflect the kidney-expressed genes in the list, since kidneys are responsible for maintaining electrolyte and fluid balance in the body.

The relationships between genes and biological processes are visualized in the **clustergram shown in Figure 8**. This heatmap maps the top 20 genes (rows) to the top 10 enriched GO biological processes (columns). Each red cell represents an association between a gene and a GO term, indicating that the gene contributes to that particular biological process.
From the heatmap, we can clearly see that **heart-specific genes** like *MYPN*, *ANKRD1*, *MYBPC3*, *TNNT2*, and *MYO18B* are involved in structural processes such as *myofibril assembly*, *sarcomere organization*, and *heart contraction*. These genes cluster together under muscle-related terms, reinforcing their role in heart function.

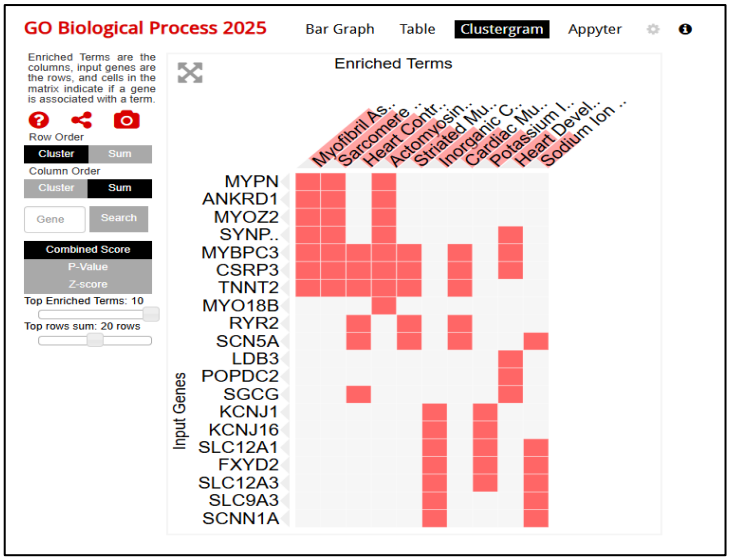


Figure8.

Figure 8. Clustergram displaying the relationship between enriched GO terms (columns) and genes (rows). Red blocks indicate gene membership in each term. Heart-upregulated genes cluster under muscle-related processes, while kidney-upregulated genes group under ion transport terms, highlighting tissue-specific functional patterns

On the other hand, **kidney-associated genes** such as *SLC12A1*,*SLC9A3*, and *SCNN1A* are linked to processes like *ion transport* and *transmembrane sodium/potassium balance*, which are essential for renal physiology. These genes appear grouped under terms like *inorganic cation import* and *sodium ion transmembrane transport*, further highlighting tissue-specific enrichment.
Overall, the clustergram effectively separates genes based on their biological roles, revealing clear patterns that align with heart and kidney tissue specialization. This supports the accuracy of the differential expression results and the functional relevance of the enriched terms.

5) KEGG Pathway Enrichment:

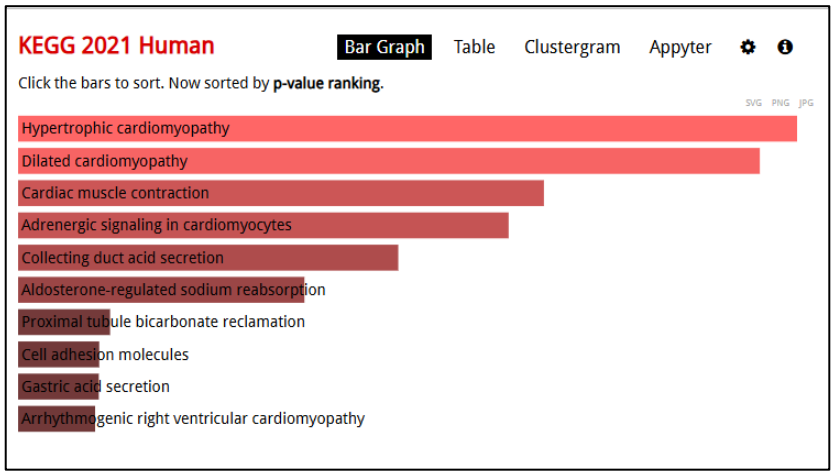


Figure9.

Figure9 Bar graph showing the top KEGG pathways enriched in the top 100 differentially expressed genes. Cardiac-related pathways like hypertrophic cardiomyopathy and cardiac muscle contraction were the most significant, while several kidney-related pathways were also enriched.

To further explore the biological significance of the top 100 differentially expressed genes, I performed KEGG pathway enrichment analysis. The results, shown in Figure 9, revealed that many of the top-enriched pathways were related to heart disease and cardiac muscle function. Notably, pathways such as *hypertrophic cardiomyopathy*, *dilated cardiomyopathy*, and *cardiac muscle contraction* were among the most significantly enriched. This reflects the presence of several heart-specific genes in the list, which are involved in regulating contraction, signaling, and structural organization of the heart.

In addition to heart pathways, a few kidney-associated pathways were also enriched. These included *collecting duct acid secretion*, *aldosterone-regulated sodium reabsorption*, and *proximal tubule bicarbonate reclamation*, which are central to kidney ion transport and fluid balance. The presence of these pathways suggests that kidney-upregulated genes also contribute meaningfully to biological function separation.

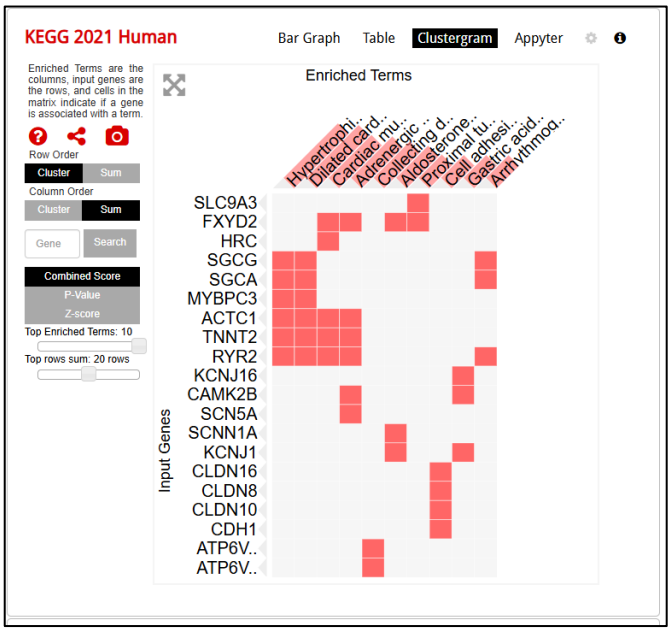


Figure10.

Figure10 Clustergram visualizing the relationship between input genes (rows) and KEGG pathways (columns). Heart-upregulated genes are grouped under cardiac-related pathways, while kidney-upregulated genes align with transport and secretion pathways, indicating tissue-specific functional enrichment.

The KEGG clustergram in Figure 10 illustrates the association between genes and enriched pathways. Heart-specific genes such as *MYBPC3*, *ACTC1*, *RYR2*, and *TNNT2* were strongly linked to cardiomyopathy and muscle contraction pathways. Meanwhile, kidney-related genes like *SLC9A3*, *CLDN10*, *ATP6V0A4*, and *SCNN1A* were associated with acid secretion and electrolyte regulation pathways. This separation into two major functional groups further supports the distinct tissue-specific gene expression patterns identified earlier.

6) STRING Protein–Protein Interaction (PPI) Network Analysis:

To further explore the interactions among the top 100 differentially expressed genes, I used the STRING database to generate a protein–protein interaction (PPI) network. This helped identify functional relationships and potential co-regulatory mechanisms between heart- and kidney-enriched genes.

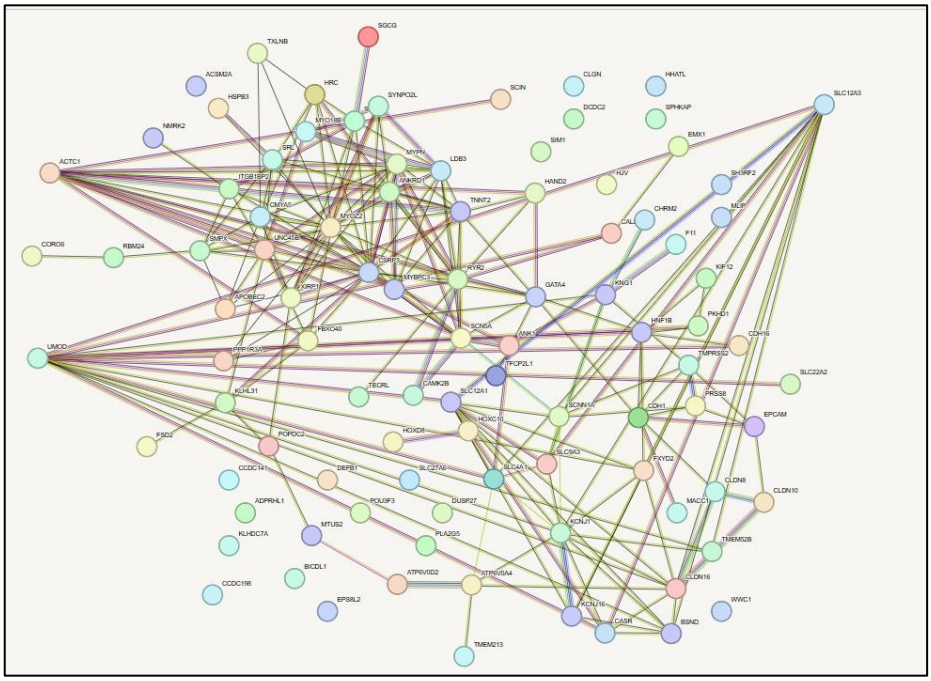


Figure12.

Figure12 STRING PPI Network of Top 100 DEGs. Nodes represent proteins and edges represent known or predicted interactions. Highlighted clusters show strong connectivity among heart-expressed proteins.

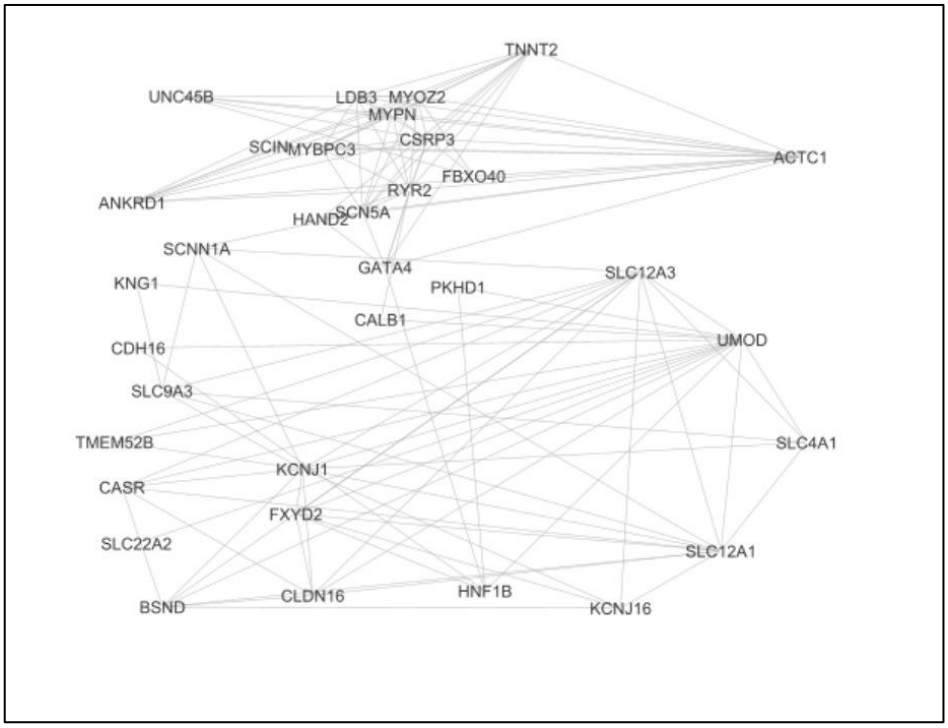


Figure13.

Figure13 STRING PPI Network of Top 100 DEGs in cytoscpae.

As shown in Figure 12 and Figure 13, the STRING network revealed a dense interaction cluster among heart-specific genes such as ACTC1, MYBPC3, TNNT2, and RYR2.
These proteins are known to play crucial roles in cardiac muscle structure and function, and their strong interconnectivity suggests coordinated regulation in heart tissue.

Additionally, kidney-enriched genes such as UMOD, SLC12A1, CLDN16, and KCNJ1 also showed connections, although they formed a comparatively sparser network. These genes are involved in solute transport and ion channel activity, consistent with the kidney's physiological role in maintaining electrolyte balance. Together, the PPI analysis supports the findings from the differential expression and enrichment analyses by visually clustering genes with related biological functions into tissue-specific modules.

Discussion:

The results of this RNA-seq analysis comparing human heart and kidney tissues strongly support the initial hypothesis. I expected genes involved in muscle contraction (e.g., ACTC1, MYH6) to be upregulated in heart, and genes related to filtration and solute transport (e.g., UMOD, SLC12A1) to be highly expressed in kidney. Differential expression analysis confirmed that ACTC1, UMOD, and SLC12A1 were among the most significantly expressed genes in their respective tissues. Although *MYH6* did not appear in the top 100 DEGs, it was still differentially expressed and likely underrepresented due to sample origin (ventricular tissue has lower MYH6 expression than atrial tissue)[15][4]. This indicates that the results make biological sense and reflect known patterns of tissue-specific gene expression.

Functional enrichment through GO and KEGG confirmed these tissue roles. Heart DEGs were associated with sarcomere organization, cardiac contraction, and cardiomyopathy-related pathways. Kidney DEGs were enriched for processes like ion transport, sodium reabsorption, and acid secretion—functions essential to renal physiology. Protein–protein interaction (PPI) network analysis using STRING supported these findings. Heart-specific genes clustered tightly, particularly those related to muscle structure and contraction (e.g., ACTC1, MYBPC3, TNNT2).

In contrast, kidney-expressed genes formed smaller or more isolated modules, reflecting their independent roles across different nephron segments. These results are consistent with known biology and previous studies, such as those from the GTEx project and the Human Protein Atlas, which also show that cardiac tissue is enriched for structural muscle genes, while kidney tissue expresses solute transporters and ion channels. There were a few limitations. The enrichment and STRING analyses were limited to the top 100 DEGs, which may have excluded some biologically relevant genes. The small sample size may reduce statistical power, and batch effect correction was not applied, which could influence gene expression estimates. Additionally, expression changes were not validated at the protein level.

Inference:

The results of this RNA-seq study clearly support the initial hypothesis that heart and kidney tissues exhibit distinct transcriptional programs aligned with their physiological functions. As anticipated, genes associated with muscle contraction and energy metabolism—such as ACTC1, MYBPC3, and TNNT2—were significantly upregulated in heart tissue, reflecting its role as a contractile organ. In contrast, kidney samples showed high expression of genes involved in solute transport and filtration, including UMOD, SLC12A1, and SCNN1A, which are essential for maintaining electrolyte and fluid balance. Principal Component Analysis and clustering heatmaps revealed strong separation between heart and kidney samples, confirming the presence of tissue-specific gene expression patterns. Further, Gene Ontology and KEGG pathway enrichment analyses highlighted heart-related processes like sarcomere assembly and cardiomyopathy pathways, while kidney-related pathways centered on ion transport and acid-base regulation. Protein-protein interaction networks constructed using STRING and Cytoscape demonstrated tight clusters among heart-expressed proteins and functional modules in kidney-expressed genes, reinforcing the idea of coordinated, organ-specific molecular networks. Altogether, the findings validate the hypothesis and underscore the power of RNA-seq combined with network and enrichment analyses to uncover the molecular basis of tissue specialization.

References:

[1] C. Crocini and M. Gotthardt, "Cardiac sarcomere mechanics in health and disease," *Biophys. Rev.*, vol. 13, no. 5, pp. 637–652, 2021, doi: 10.1007/s12551-021-00840-7.

[2] G. Mancuso *et al.*, "Clinical and Genetic Heterogeneity of HCM: The Possible Role of a Deletion Involving MYH6 and MYH7," *Genes (Basel).*, vol. 16, no. 2, pp. 1–11, 2025, doi: 10.3390/genes16020212.

[3] C. E. Martin and N. Jones, "Nephrin signaling in the podocyte: An updated view of signal regulation at the slit diaphragm and beyond," *Front. Endocrinol. (Lausanne).*, vol. 9, no. JUN, pp. 1–12, 2018, doi: 10.3389/fendo.2018.00302.

[4] M. Habuka *et al.*, "The kidney transcriptome and proteome defined by transcriptomics and antibody-based profiling," *PLoS One*, vol. 9, no. 12, pp. 1–19, 2014, doi: 10.1371/journal.pone.0116125.

[5] L. Fagerberg *et al.*, "Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics," *Mol. Cell. Proteomics*, vol. 13, no. 2, pp. 397–406, 2014, doi: 10.1074/mcp.M113.035600.

[6] M. Karlsson *et al.*, "A single–cell type transcriptomics map of human tissues," *Sci.*

Adv., vol. 7, no. 31, pp. 1–9, 2021, doi: 10.1126/sciadv.abh2169.

[7] Y. Wang and X. Qian, "Functional module identification in protein interaction networks by interaction patterns," *Bioinformatics*, vol. 30, no. 1, pp. 81–93, 2014, doi: 10.1093/bioinformatics/btt569.

[8] D. Szklarczyk *et al.*, "STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D607–D613, 2019, doi: 10.1093/nar/gky1131.

[9] X. D. Team, "Xming X Server for Windows," 2025, 6.9.0.31. [Online]. Available: https://sourceforge.net/projects/xming/

[10] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: A Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, 2009, doi: 10.1093/bioinformatics/btp616.

[11] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biol.*, vol. 15, no. 12, pp. 1–21, 2014, doi: 10.1186/s13059-014-0550-8.

[12] 1 Paul Shannon *et al.*, "Cytoscape: A Software Environment for Integrated Models," *Genome Res.*, vol. 13, no. 22, p. 426, 1971, doi: 10.1101/gr.1239303.metabolite.

[13] Simon Andrews, "FastQC," 0.12.0. [Online]. Available: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[14] E. Y. Chen *et al.*, "Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool," *BMC Bioinformatics*, vol. 14, no. 1, p. 128, 2013, doi: 10.1186/1471-2105-14-128.

[15] C. Lindskog *et al.*, "The human cardiac and skeletal muscle proteomes defined by transcriptomics and antibody-based profiling," *BMC Genomics*, vol. 16, no. 1, 2015, doi: 10.1186/s12864-015-1686-y.