

A Differential Expression Analysis of Ischaemic and Healthy Cardiac Tissue Highlights Associated Genes and Pathways

1 Introduction

Ischaemic heart disease (IHD), also known as coronary artery disease, represents the leading cause of mortality globally, accounting for 182 million disability-adjusted life years, and 9.14 million deaths in 2019 (Roth *et al.*, 2020). Ischaemic heart disease is the result of the chronic atherosclerotic process, whereby lipid deposition upon the arterial walls results in occlusion of the vessels and restricted bloodflow to the heart via

the coronary artery (Malaker *et al.*, 2019). Subsequent sequela of atherosclerosis includes myocardial infarction, ischaemic strokes, and ischaemic heart disease.

Ischaemic heart disease is a complex trait, resulting in part from genetic susceptibility, environmental exposures, and the complex interplay between the two (Musunuru and Kathiresan, 2019). Numerous putative genes and pathways have been associated in the genetic susceptibility to ischaemic heart disease via genome-wide association studies, but the complex mechanistic network that mediate the pathophysiological progression of the disease remain unclear (Nordlie, Wold, and Kloner, 2005). Previous genome-wide association studies have attributed a heritability of between 0.35 and 0.55, suggesting a substantial genetic component to the disorder. However, due to the complex nature of the trait however, the environmental component and its possible effects upon gene expression represent a key target for the management of ischaemic heart disease.

This report employs a bioinformatic analysis of bulk RNA sequencing data from the primary cardiac fibroblasts of 10 healthy, control patients and 13 patients with ischaemic heart disease. The omic-level analyses herein

aim to identify significantly enriched expressed genes in ischemic heart tissue relative to controls to elucidate key genes and pathways involved in ischaemic heart disease.

2 Results

RNA-seq data was firstly pre-processed to generate a master dataframe, upon which most analyses were conducted. Using this master dataframe, preliminary analysis concerning quality control of the input dataset of healthy control ($n = 10$) and ischaemic ($n = 13$) heart tissue was employed (Figure 1).

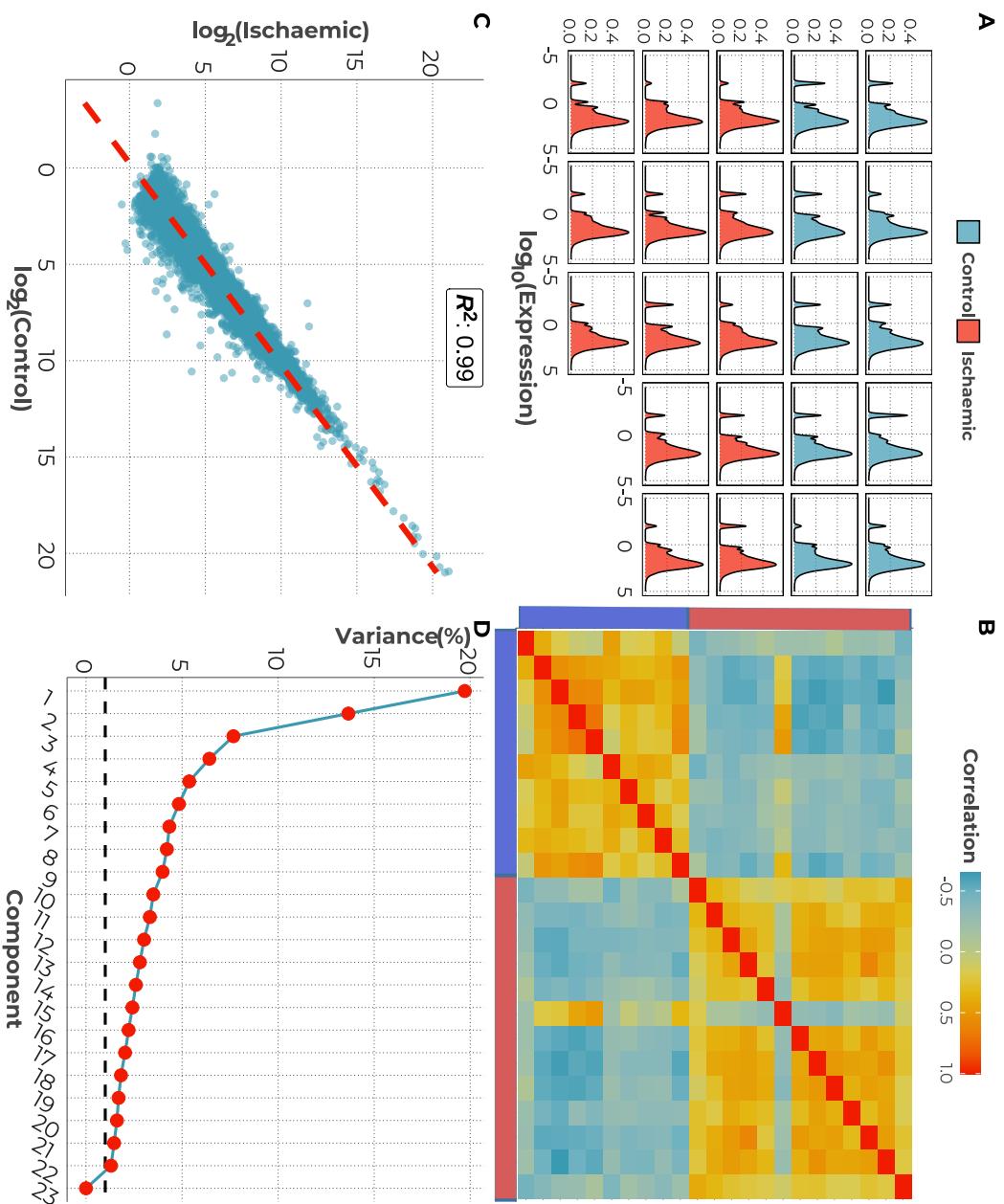


Figure 1: **Quality Control of Dataset ($n_{control} = 10$, $n_{ischaemic} = 13$)**

(A) Expression density of 10 control replicates and 23 ischaemic replicates. **(B)** Sample-by-sample heatmap of the Spearman correlation of each sample pairwise for the significantly differentially expressed genes, with the fill of each tile representing the correlation value (blue tile = control, red tile = ischaemic sample) ($n = 434$). **(C)** Pearson correlation coefficient (R^2) of the mean expression values of all genes, grouped by sample, with the values then being normalised to their logarithm base-2. **(D)** Scree plot illustrating the proportion of variance explained by each of the principal components following principal component analysis.

Quality Control

Initial analysis for sample quality revealed high variability of expression patterns within and across samples (Figure 1A). All samples displayed a bimodal distribution of expression patterns, with a peak representing non-expressed genes ($\log_{10}\text{Expression} \simeq -2.5$) and a larger peak with a small shoulder representing those expressed genes ($0 \leq \log_{10}\text{Expression} \leq 4$). However, gross differences in the density of these expression patterns were noted within the sample subgroups. Analysis of sample-by-sample correlation however revealed that a weak correlation did exist between samples within the subgroups with respect to the subset of significantly-altered genes (Figure 1B). However, a significant positive correlation between the two subgroups was observed by comparison of the mean expression value for all assayed gene per subgroup ($R^2 = 0.99, p < 2.2 \times 10^{-16}$) (Figure 1C). Principal component analysis yielded no principal components that explained a significant proportion of variance (Figure 1D) and subsequent biplots revealed no evident discrimination of the subgroups, corroborating the complexity and overall noise of the sample groups previously observed

(Supplementary Figure 1).

Identification of Differential Gene Expression

Following general analysis of sample structure for quality, analysis of the RNA-seq read count data was employed to illustrate differential gene expression between the control and ischaemic samples (Figure 2).

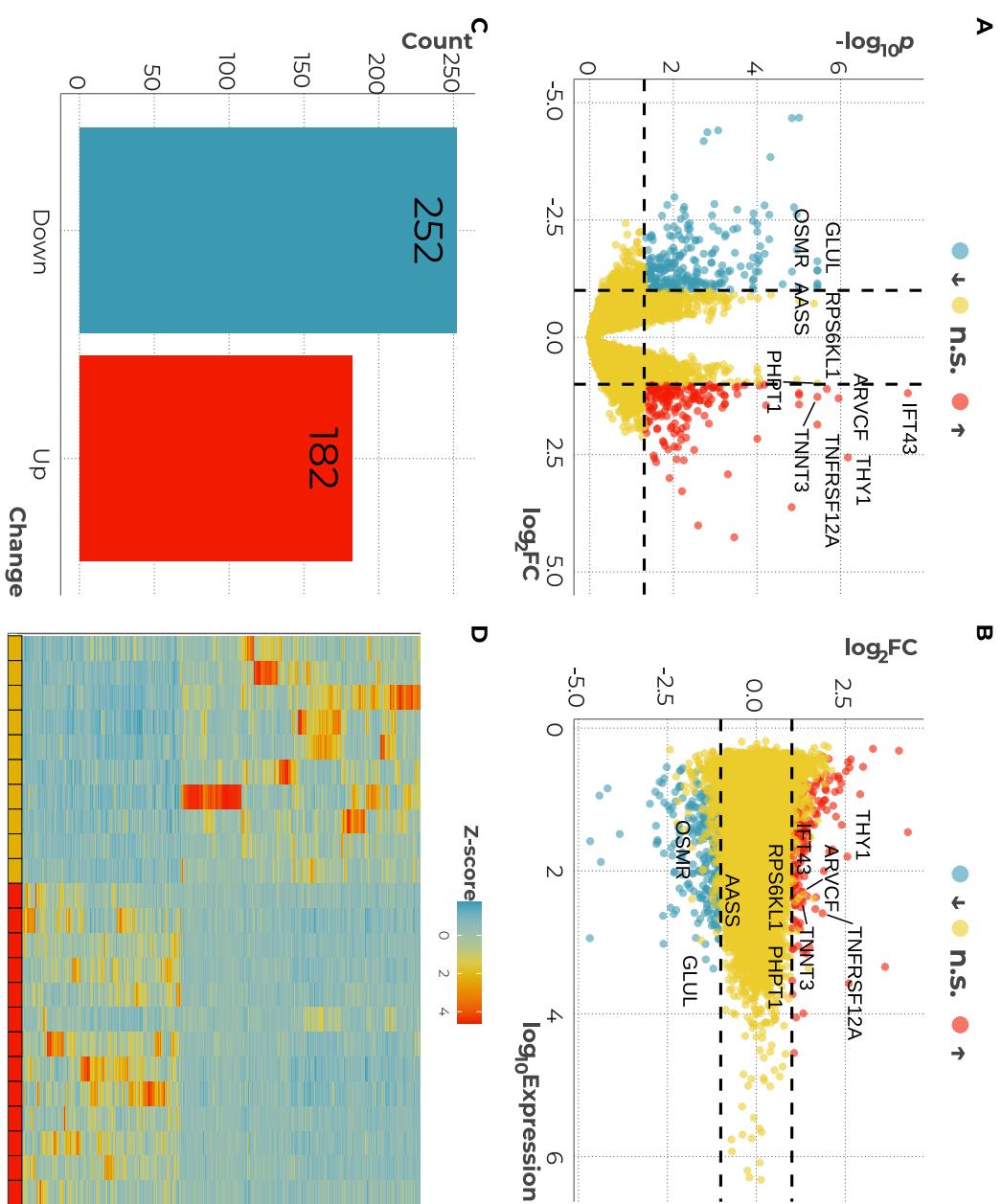


Figure 2: **Visualisation of RNA-seq analysis data**

(A) Volcano plot illustrating the \log_2 of fold change in gene expression between samples and the $-\log_{10}$ -normalised adjusted p -value. The 10 genes with the lowest p -value following differential gene analysis are labelled. **(B)** MA plot displaying $\log_{10}\text{Expression}$ against $\log_2\text{Fold Change}$ for all assayed genes. The same 10 genes as in the 1A are labelled. **(C)** Barchart of the number of genes that are significantly down- and upregulated between the two sample groups. **(D)** Hierarchically-clustered heatmap of the significantly differentially expressed gene subset ($n = 434$) with control and ischaemic tissue samples denoted by the yellow and red tiles displayed at below the heatmap, respectively.
 (Blue = significantly downregulated, yellow = no significant change, red = significantly upregulated)

Transcriptome-wide analyses revealed numerous significantly up- and downregulated genes when comparing control and ischaemic cardiac tissue, with 252 upregulated and 182 downregulated genes, respectively (Figure 2A, C). Visualisation via MA plotting highlighted a low statistical power within the sample as indicated by the symmetric decrease in absolute fold-change with increasing expression, suggesting a greater sample is needed to better quantify differential expression between the two samples (Figure 2C). Moreover, hierarchical clustering of the significant gene subset ($n = 434$) illustrated only a weak clustering effect of gene expression patterns between control and ischaemic samples, in line with observations during the quality control analysis pointing to insufficiently designed sample structure (Figure 2D).

Analysis of Enriched Gene Ontologies

Differential gene analysis highlighted differences in expression patterns, with subsequent gene enrichment analysis then being conducted to identify pathways and gene ontologies that differ between the healthy and ischaemic heart tissue (Figure 3).

Significantly enriched gene ontologies associated with molecular

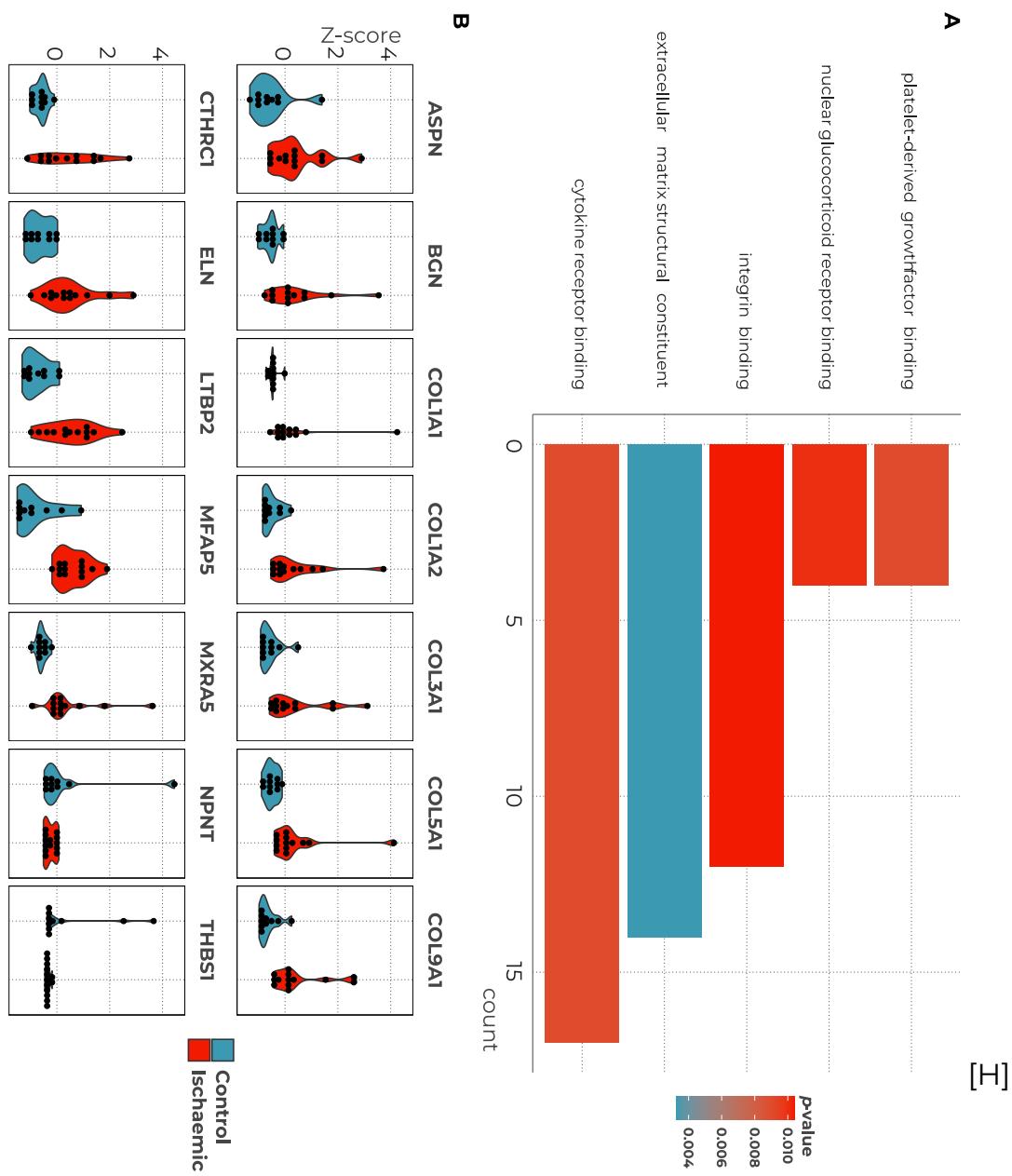


Figure 3: **Gene ontology based upon over-representation analysis:** **(A)** Molecular function-based gene ontology, illustrating the top 5 most significantly enriched gene ontology sets, the gene count per set, and the adjusted p-value. **(B)** Violin plots comparing the normalised (z-score) expression of genes within the extracellular matrix structural constituent gene ontology between control (blue) and ischaemic (red) samples.

function were identified, illustrating a particular bias towards gene products with binding activity and roles in extracellular matrix function as well as inflammatory signalling (Figure 3A). Analysis of the most significantly-enriched gene set, those comprising the extracellular matrix, demonstrated numerous collagen isoforms particularly enriched in ischaemic heart tissue as compared to healthy tissue (Fig 3B).

Employing the previous data from over-representation analysis, a relational network of gene ontologies and the associated genes with the ontology set was created (Figure 4).

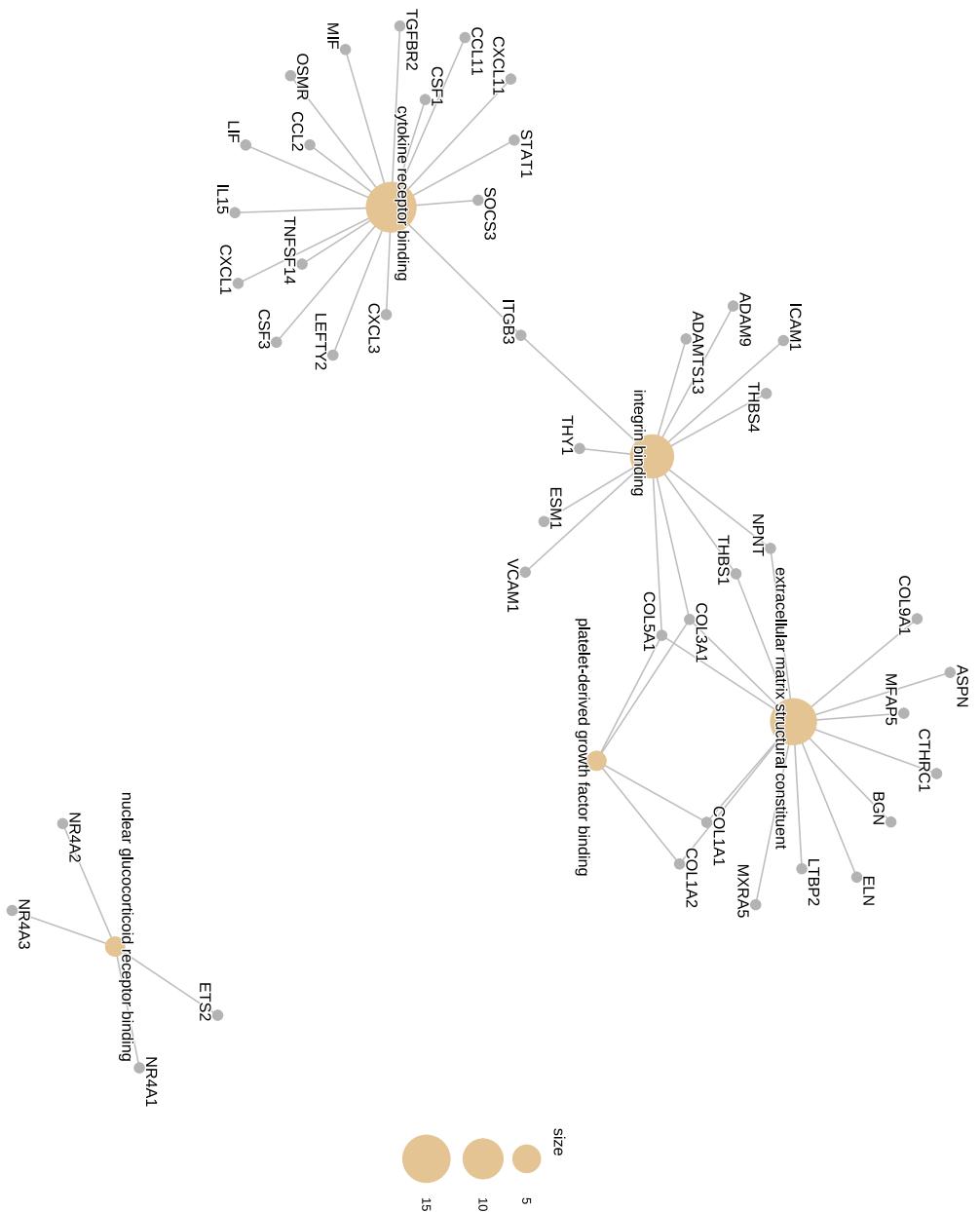


Figure 4: Gene-concept network of the five molecular function gene ontologies identified by over-representation analysis. The network demonstrates the genetic relationships between the gene ontology sets (yellow nodes) and their associated members (grey). The node size represents the number of genes within the gene set while the connecting edges between nodes and genes represent their biological relationship.

As was previously noted, members of the collagen superfamily were noted to be over-represented within the dataset. In addition to this, the same collagen isoforms seem to possess an important role in bridging the ontology nodes, suggesting that collagens and their differential expression are intimately involved with the pathology of ischaemic heart disease via their multifaceted role in numerous associated pathways.

Moreover, the most prolific differentially expressed genes in the enrichment analysis are those centred around cytokine receptor binding, highlighting the key role of inflammatory processes in ischaemic heart disease progression and symptomology. While the cytokine-receptor associated gene set is relatively distinct, it forms a bridge with the integrin binding gene ontology via Integrin beta 3 (IGTB3), a cell surface protein in platelets, highlighting the interplay of inflammation and coagulation in the atherosclerotic process and consequent ischaemia of the heart.

3 Discussion

Findings

Notably, the most significantly enriched gene set, with respect to molecular function, comprised numerous members of the collagen superfamily. This is in line with the current understanding of ischaemic heart disease, whereby atherosclerosis, the deposition of lipids and subsequently, fibrous tissue comprising collagen at the site of the atherosclerotic plaque. Moreover, these data highlight how these collagen members span a relational network in which they both constitute the extracellular matrix, while also physically interacting with platelet-derived growth factor, which has a potent role in regulating angiogenesis in the event of myocardial infarction (Lopez, Gonzales, and Diez, 2020).

In addition to this, genes associated with pro-inflammatory signalling such as those of the tumour necrotic factor alpha ($\text{TNF}\alpha$) and transforming growth factor beta ($\text{TGF}\beta$) signalling networks were observed to be significantly differentially expressed between tissues as well as being enriched within the gene ontology over-representation

analysis. In this instance, differential gene analysis and gene enrichment analysis implicated TGF β Receptor 2 (TGF β R2), TNFSF14 (TNF superfamily member 14), TNFRSF12A (TNF receptor superfamily member 12A), and STAT1 as being differentially expressed between control and ischaemic heart tissue. These observations taken together agree with the understanding that risk of ischaemic heart disease is predicted by numerous comorbid inflammatory conditions such as obesity, Diabetes Mellitus, and atherosclerosis, which is itself an inflammatory pathology and a significant risk factor of ischaemic heart disease (Lyu *et al.*, 2018; Heinein, 2022; Doetschmann *et al.*, 2011; Ng, 2003). However, it should be noted that while ischaemic heart disease is often comorbid with chronic inflammatory pathologies, lifestyle factors such as poor diet, sedentary lifestyles, smoking, and stress are all contributors to systemic inflammation as well as the aforementioned pathologies and thus may be confounders, influencing both risk factors and ischaemic heart disease (Martin, 2024; Roivanem *et al.*, 2000).

Limitations

During processing of the dataset for quality control, rows of the master dataframe containing NA values within any column were omitted from further analysis. Scrutiny of this revealed the omission of 1554 genes on the basis of containing an NA value within the p.adj column, with 122 of these genes having an unadjusted p-value less than 0.05. Thus, the dataset employed here may have some caveats in preprocessing that should be addressed prior to further analysis

Additionally, preliminary analysis of the dataset highlighted no distinct clustering of the subgroups, instead showing strong intragroup variability. Taken together, this is suggestive that data collection may need to be optimised in order to better determine the contribution of disease state to the respective transcriptome profiles. Moreover, no information was available with reference to sex, age, or other possible confounding factors, which again may prove useful in determining if expression profiles are stratified across factors not considered here.

Future Prospects

Ischaemic heart disease has increased in incidence steadily until this present day and it is predicted to continue on this course (Roth *et al.*, 2019). Robust understanding of the involved pathways in addition to candidate genes that may represent therapeutic targets is a definitive goal of large-scale bioinformatic analyses. However, in order to elucidate the underlying mechanisms and networks, large-scale analyses are required, in order to increase the statistical power of any observations. The work here may be replicated in future, employing greater sample sizes and accounting for stratification due to sex, age and other factors, which may strengthen the reliability and reproducibility of future work.

4 Supplementary Figures

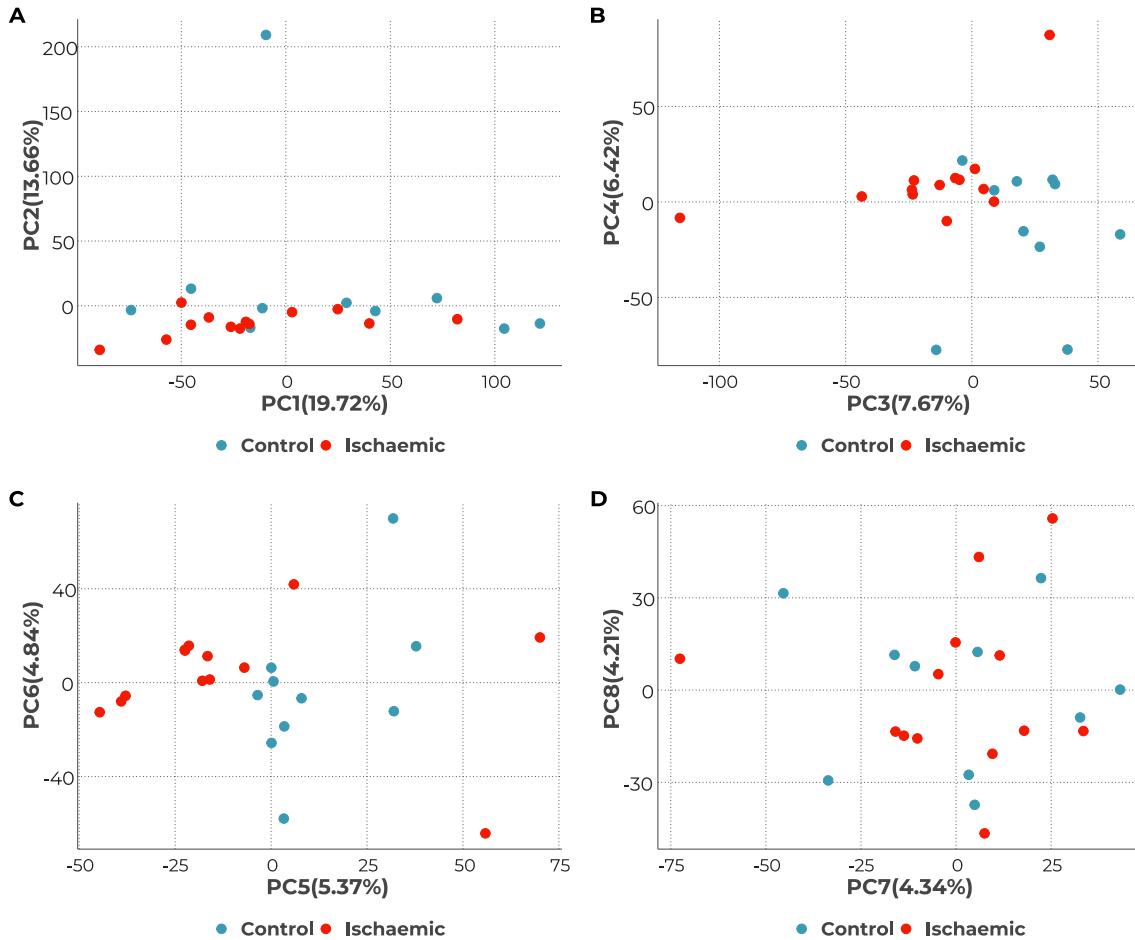


Figure 5: Biplots of principal component 1 to 8 following principal component analysis

(A)

References

- Doetschman, T. et al. (2011) 'Transforming growth factor beta signaling in adult cardiovascular diseases and Repair', *Cell and Tissue Research*, 347(1), pp. 203–223. doi:10.1007/s00441-011-1241-3.
- Henein, M.Y. et al. (2022) 'The role of inflammation in cardiovascular disease', *International Journal of Molecular Sciences*, 23(21), p. 12906. doi:10.3390/ijms232112906.
- Lopez, B., Gonzalez, A. and Diez, J. (2010) 'Circulating biomarkers of collagen metabolism in cardiac diseases', *Circulation*, 121(14), pp. 1645–1654. doi:10.1161/circulationaha.109.912774.
- Lyu, M. et al. (2018) 'TNFRSF12A-mediated atherosclerosis signaling and inflammatory response as a common protection mechanism of shuxuening injection against both myocardial and cerebral ischemia-reperfusion injuries', *Frontiers in Pharmacology*, 9. doi:10.3389/fphar.2018.00312.
- Malakar, A.K. et al. (2019) 'A review on coronary artery disease, its risk factors, and therapeutics', *Journal of Cellular Physiology*, 234(10), pp.

16812–16823. doi:10.1002/jcp.28350.

- Martin, S.S. et al. (2024) '2024 heart disease and stroke statistics: A report of US and Global Data from the American Heart Association', *Circulation*, 149(8). doi:10.1161/cir.0000000000001209.
- Musunuru, K. and Kathiresan, S. (2019) 'Genetics of common, complex coronary artery disease', *Cell*, 177(1), pp. 132–145. doi:10.1016/j.cell.2019.02.015.
- Ng, D. (2003) 'Activation of signal transducer and activator of transcription (STAT) pathways in failing human hearts', *Cardiovascular Research*, 57(2), pp. 333–346. doi:10.1016/s0008-6363(02)00664-8.
- Nordlie, M., Wold, L. and Kloner, R. (2005) 'Genetic contributors toward increased risk for ischemic heart disease', *Journal of Molecular and Cellular Cardiology*, 39(4), pp. 667–679. doi:10.1016/j.jmcc.2005.06.006.
- Roivainen, M. et al. (2000) 'Infections, inflammation, and the risk of coronary heart disease', *Circulation*, 101(3), pp. 252–257. doi:10.1161/01.cir.101.3.252.

- Roth, G.A. et al. (2018) 'Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: A systematic analysis for the global burden of disease study 2017', *The Lancet*, 392(10159), pp. 1736–1788.
doi:10.1016/s0140-6736(18)32203-7.