

https://dspace.org/

Using ChIP-seq and Gene Expression Microarray data to explore transcriptional dysregulation of PXDN and PXDNL in cardiovascular diseases

Naidoo, Shiven

2024

University of the Witwatersrand, Johannesburg

https://hdl.handle.net/10539/42148

Naidoo, Shiven. (2024). Using ChIP-seq and Gene Expression Microarray data to explore transcriptional dysregulation of PXDN and PXDNL in cardiovascular diseases [Master's dissertation, University of the Witwatersrand, Johannesburg]. WireDSpace.

Downloaded from a DSpace repository. https://dspace.org/



UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

Using ChIP-seq and Gene Expression Microarray data to explore transcriptional dysregulation of *PXDN* and *PXDNL* in cardiovascular diseases

by

Shiven Naidoo

(1943318)

Dissertation

Submitted in fulfilment of the requirements for the degree

Master of Science

in

Molecular and Cell Biology

in the Faculty of Science, University of the Witwatersrand, Johannesburg, South Africa

Supervisor: Professor Demetra Mavri-Damelin

Co-supervisor: Doctor Nikki Gentle

June 2024

Declaration

I, Shiven Naidoo (1943318), am a student registered for the degree of Master of Science in the academic year 2024.

I hereby declare the following:

● I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.

● I confirm that the work submitted for assessment for the above degree is my own unaided work except where explicitly indicated otherwise and acknowledged.

● I have followed the required conventions in referencing the thoughts and ideas of others.

● I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.



Shiven Naidoo

6 June 2024

Abstract

Background: Cardiovascular diseases (CVDs) remain one of the leading causes of death globally. The genes *PXDN* and *PXDNL* are both expressed in the cardiovascular system, and their dysregulation has been linked to various disorders, including CVDs, but little is known of their transcriptional regulation in the cardiovascular system or their roles in CVD pathogenesis.

Methods: This study developed two custom bioinformatics pipelines in R to mine and analyse ChIP-seq data from ChIP-Atlas and gene expression microarray data from the Gene Expression Omnibus (GEO). The first pipeline used ChIPseeker to identify regulatory transcription factors (TFs) of *PXDN* and *PXDNL* in cardiovascular cells and tissues. ChIP-seq data from 400 experiments across 63 TFs was filtered to isolate TFs with high confidence binding peaks in the promoter and first intron of *PXDN* and *PXDNL*. The second pipeline used R Bioconductor packages to explore the expression profiles of *PXDN, PXDNL*, and their TFs in seven microarray datasets across three CVD-related contexts: cardiomyopathies, heart failure and TNF-α stimulation.

Results and discussion: This study identified 27 TFs binding to *PXDN* and 18 TFs binding to *PXDNL* in cardiovascular cells. Sixteen of these TFs were shared by both *PXDN* and *PXDNL*, suggesting potential coregulatory mechanisms in cardiovascular cells where they are both expressed. Unique TFs were also identified for *PXDN* (11) and *PXDNL* (2). Differential gene expression analysis revealed no significant change in expression ($log_2FC > 0.5$; $p.adj < 0.05$) for *PXDN, PXDNL* and many of their identified TFs in the CVD-related conditions investigated, suggesting that changes at the transcript level may not contribute to the progression of these conditions.

Conclusions: This study advances our understanding of the transcriptional regulation of *PXDN* and *PXDNL* in healthy cardiovascular cells as well as their expression levels in the investigated CVD-related contexts. This study also contributes a bioinformatics pipeline which can be further developed and applied to analysing data from ChIP-Atlas and GEO. Future research can elucidate the roles of each TF in regulating *PXDN* and *PXDNL* in healthy and diseased cell lines.

Acknowledgements

I would like to thank my supervisors Professor Demetra Mavri-Damelin and Dr Nikki Gentle for their guidance. I value the time and effort you have spent on me.

I would like to acknowledge the National Research Foundation, whose financial support gave me the freedom to pursue my studies.

To my colleagues in the Applied Bioinformatics and Functional Genetics Research laboratories, I am grateful for your support.

To my friends, thank you for dragging me out of my study whenever I needed a break.

And lastly, to my family, thank you for your patience, love, and support.

Table of Contents

1 INTRODUCTION 1

1.1 The cardiovascular system 1

1.1.1 The cardiovascular disease burden 2

1.1.2 Molecular pathways for CVD pathogenesis 4

1.1.3 *PXDN* Background 6

1.1.4 Structure and function 7

1.1.5 Regulation of *PXDN* expression 10

1.1.6 *PXDN*'s role in disease 11

1.2 *PXDNL* 14

1.3 Bioinformatics: Data mining 15

1.4 Introduction to current study 16

1.5 Aims and objectives 17

2 MATERIALS AND METHODS 18

2.1 ChIP-seq data mining pipeline 19

2.1.1 ChIP-Atlas background 20

2.1.2 Data collection 21

2.1.3 Data pre-processing 22

2.1.4 Data analysis and visualisation 22

2.2 Microarray data mining pipeline 23

2.2.1 GEO background 25

2.2.2 Data collection 25

2.2.3 Data preparation 27

2.2.4 Exploratory data analysis and quality control 30

2.2.5 Differential gene expression analysis 32

2.2.6 DGE result visualisation 33

2.3 Validation of DGE results 34

2.4 Programs and R script 34

3 RESULTS 35

3.1 ChIP-seq analysis 35

3.1.1 TF binding visualization using peak browser 35

3.1.2 Identifying regulatory TFs for *PXDN* and *PXDNL* 38

3.2 Gene expression microarray analysis 43

3.2.1 Quality control and preprocessing 43

3.2.2 DGE analysis 51

4 DISCUSSION 56
4.1 Insights from ChIP-seq data mining 56
4.1.1 Exploration of TF binding patterns 57
4.1.2 Shared TFs of *PXDN* and *PXDNL* 57
4.1.3 Unique TFs: *PXDN* 60
4.1.4 Unique TFs: *PXDNL* 63
4.2 Microarray data exploration 64
4.2.1 DGE analysis in CVDs 64
4.3 Theoretical and practical implications 67
4.4 Limitations 68
4.5 Future research 69
4.6 Conclusion 70
REFERENCES 71
APPENDIX 90

List of Figures

Figure 3.1 Distribution of TF binding peaks identified in genomic features within a 10kb range of (A) *PXDN* and (B) *PXDNL*'s TSS. 37

Figure 3.2 High confidence TF binding peaks located within the first introns of (A) *PXDN* and (B) *PXDNL* in different cardiovascular cells/tissues. 39

Figure 3.3 High confidence TF binding peaks located within the promoter of (A) *PXDN* and (B) *PXDNL* in different cardiovascular cells/tissues. 40

Figure 3.4 Number of shared and unique TFs binding to the promoter and first introns of *PXDN* and *PXDNL*. 42

Figure 3.5 Heatmaps and hierarchical clustering. 44

Figure 3.6 Distribution of fluorescence values for raw and preprocessed microarray datasets. .. 46

Figure 3.7 Principal Component Analysis (PCA) plot showing sample clustering along PC1 and PC2 for GSE26887's (A) raw expression data (B) RMA normalized expression data. 48

Figure 3.8 PCA plot showing sample clustering along PC1 and PC2 for GSE2638 (A) raw expression data (B) RMA normalized expression data. 50

Figure 3.9 Volcano plots showing the distribution of significantly differentially expressed genes 52

Figure 3.10 RMA normalized fluorescence values. 54

Figure A.1 Heatmap and hierarchical clustering showing the consistency within and between sample groups 91

Figure A.2 PCA plot showing clustering of sample groups along PC1. 91

Figure A.3 Contribution of each principal component to observed variances. 92

Figure A.4 RNA degradation plot to assess RNA quality of each sample. ..... **Error! Bookmark not defined.**

Figure A.5 MA plots to compare fluorescence values of probes between different samples. . **Error! Bookmark not defined.**

Figure A.Boxplots of fluorescence values for GSE1400.6 ..... **Error! Bookmark not defined.**

Figure A.7 Micrographs of microarray samples generated from the GSE1400 dataset. ..... **Error! Bookmark not defined.**

Figure A.8 Volcano plot showing the number of differentially expressed genes in the membrane fraction of MCF-7s compared to the cytosolic fraction. 95

List of Tables

Table 1.1 Functions of Peroxidasin's structural motifs 7
Table 1.2 Diseases linked to *PXDN* dysregulation 13
Table 2.1 Summary of key selection criteria for the inclusion of identified datasets 26
Table 2.2 Summary of selected microarray datasets 28
Table 3.1 Summary of cell types, treatments, TFs and gene loci for all ChIP-seq experiments in cardiovascular cells 36
Table 3.2 Summary of DGE results for seven selected datasets 55
Table A.1 Summary of function parameters used for Exploratory Data Analysis 95

List of Abbreviations

Antioxidant response element (ARE)

Aortic adventitial fibroblasts (AoAF)

Acute myeloid leukemia (AML)

Anterior segment dysgenesis (ASD)

Cardiovascular diseases (CVD)

CCAAT/enhancer binding protein delta (CEBPD)

CCCTC-binding factor (CTCF)

Coronary artery disease (CAD)

Chromatin Immuno-Precipitation (ChIP)

E1A-associated protein p300 (EP300)

European Bioinformatics Institute (EBI)

Endothelial progenitor cell (EPC)

Endothelial sodium channel (EnNaC)

ETS related gene (ERG)

Extracellular matrix (ECM)

FBJ murine osteosarcoma viral oncogene homolog (FOS)

Friend leukemia integration 1 transcription factor (FLI1)

Focal Adhesion Kinase (FAK)

Forkhead box C2 (FOXC2)

Forkhead box O1 (FOXO1)

Gene Expression Omnibus (GEO)

Genome-wide association studies (GWAS)

High mobility group box 1 (HMGB1)

Human aortic endothelial cell (HAEC)

Human brain microvascular endothelial cell (HBMEC)

Human coronary artery smooth muscle cell (HCASMC)

Human cardiac fibroblast – adult atrial (HCFaa)

Human cardiac myocytes (HCM)

Human pulmonary artery fibroblast (HPAF)

Human embryonic kidney (HEK)

Human umbilical vein endothelial cell (HUVEC)

Human-induced pluripotent stem cell-derived cardiomyocytes (hiPSC-CMs)

Hypobromous acid (HOBr)

Hypochloric acid (HOCl)

Immunoglobulin (Ig)

Interferon regulatory factor 1 (IRF1)

Interferon-gamma-inducible protein (IP-10)

Jun proto-oncogene, activator protein 1 (JUN)

Jun B proto-oncogene, activator protein 1 (JUNB)

Jun D proto-oncogene, activator protein 1 (JUND)

Kilobases (kb)

Kruppel-like factor 4 (KLF4)

Interleukin (IL)

Lamin A/C (LMNA)

Leucine-rich repeats (LRR)

Locally Weighted Scatterplot Smoothing (LOWESS)

Monocyte chemoattractant protein 1 (MCP-1)

MYC associated factor X (MAX)

MYC proto-oncogene, bHLH transcription factor (MYC)

Myocardial Infarction (MI)

Mesenchymal transition (EMT)

National Center for Bioinformatics Information (NCBI)

Nuclear factor erythroid 2-related factor 2 (*Nrf2*)

Oxidized low-density lipoprotein (ox-LDL)

Peripheral Artery Disease (PAD)

Principal Component Analysis (PCA)

Progesterone receptor (PGR)

Rapid Kidney Function Decline (RKFD)

Reverse transcription-quantitative polymerase chain reaction (RT-qPCR)

Robust Multi-array Average (RMA)

Reactive Oxygen Species (ROS)

Sequence Read Archive (SRA)

Serum and glucocorticoid kinase 1 (SGK-1)

Sulforaphane (SFN)

SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily A, member 4 (SMARCA4)

Structural maintenance of chromosomes 1A (SMC1A)

Stromal antigen (STAG)

T-cell acute lymphocytic leukemia 1 (TAL1)

Tert-butylhydroquinone (tBHQ)

Transcription factor (TF)

Transcription factor 21 (TCF21)

Transcription factor binding site (TFBS)

Tumour necrosis factor-α (TNF-α)

Transcription Start Site (TSS)

Urinary bladder cancer (UBC)

von Willebrand factor C module (VWC)

1 INTRODUCTION

Cardiovascular diseases (CVDs) remain a major global health challenge (Wang *et al.*, 2016; WHO, 2021a), compounded by gaps in our understanding of their underlying pathogenic mechanisms. This is particularly evident in our understanding of the roles of genes such as *PXDN* and its paralog *PXDNL*. *PXDN* and *PXDNL* encode transcripts for the peroxidasin and peroxidasin-like proteins, respectively. *PXDN*, which has been shown to be expressed in various cell types in the cardiovascular system, plays key roles in vital cellular processes (Cheng and Shi, 2022; Péterfi *et al.*, 2014). The dysregulation of *PXDN* expression has been correlated with the pathogenesis of CVDs and many other diseases (Dougan *et al.*, 2019; Kurihara-Shimomura *et al.*, 2020; Zheng and Liang, 2018; Costas *et al.*, 2021; Liu *et al.*, 2019). Similarly, *PXDNL*, which has been shown to be expressed in the heart, has been linked to CVD pathogenesis as well (Cheng *et al.*, 2008; Péterfi *et al.*, 2014). However, little is known about the transcriptional regulators of *PXDN* and *PXDNL* in the cardiovascular system and the regulatory pathways linking these genes to CVD pathogenesis remain largely unexplored. This study aims to bridge these knowledge gaps by analysing data mined from publicly available -omics databases for insights into the transcriptional regulation of *PXDN* and *PXDNL* in cardiovascular cells and their expression profiles in different CVD-related contexts.

1.1 The cardiovascular system

The cardiovascular system is a complex network of organs and tissues that play critical roles in maintaining physiological homeostasis. Composed of the heart, blood vessels, and a variety of cell types, this system is responsible for the circulation of blood, thereby regulating temperature and transporting oxygen, nutrients, hormones and metabolic waste throughout the body (Levick, 1991). The heart acts as the primary pump, facilitating blood flow through the pulmonary and systemic circulatory systems (Katz, 2010; Levick, 1991). To perform this role, the heart itself is composed of specialized cells, including cardiac myofibroblasts, cardiomyocytes, and endothelial cells. Cardiac myofibroblasts play an important role in maintaining the extracellular matrix (ECM), which provides structural support to the heart (Porter and Turner, 2009), while cardiomyocytes are essential for the rhythmic contractions that pump blood (Katz, 2010). The blood vessels, consisting of arteries, veins, and capillaries, are lined by endothelial cells and surrounded by smooth muscle cells. These cells not only facilitate the transport of blood but also play a crucial role in regulating blood flow and pressure

(Aird, 2007). Furthermore, endothelial cells, alongside pericytes in capillaries and venules, also function to promote angiogenesis, control the selective permeability and vascular tone of vascular tissues, and maintain haemostasis (Kutcher and Herman, 2009; Sattler and Kennedy-Lydon, 2017). In addition to these structural cell types, free-floating cells such as lymphocytes, macrophages and platelets can be found circulating within the blood, playing key roles in processes such as immune response and tissue repair (Cao, 2018; Sonmez and Sonmez, 2017). Given the diversity of cell types and their interconnected roles in the cardiovascular system, a variety of diseases can be caused from disorders in the functioning of these cells. Due to the cardiovascular system's vital role in maintaining physiological homeostasis, these diseases represent a significant threat to human health.

1.1.1 The cardiovascular disease burden

Cardiovascular diseases (CVDs) are a large collection of complex disorders which affect the cardiovascular system, and include diseases such as Coronary Heart Disease, Stroke, Heart Failure, Congenital Heart Disease, Peripheral Arterial Disease, and Cardiomyopathies. CVDs are a major health concern globally and particularly within Sub-Saharan Africa. In 2005, CVDs contributed to 16 million deaths globally, rising to over 17 million deaths in 2015 (Wang *et al.*, 2016). Furthermore, in a recent report from the World Health Organization (WHO), CVDs accounted for an estimated 17.9 million global deaths in 2019 (WHO, 2021a). In Sub-Saharan Africa, the situation is particularly alarming, as the prevalence of CVDs is rising, compounded by challenges such as limited healthcare resources, the high prevalence of communicable diseases, and rising rates of risk factors like hypertension and obesity (Hamid *et al.*, 2019).

In addition to posing a threat on their own, CVDs are known to interact with a variety of other diseases. For example, Covid-19 was shown to promote CVD development and increase the severity of pre-existing CVDs (Chung *et al.*, 2021; Ogah *et al.*, 2021). Furthermore, treatments for Covid-19 were also shown to have negative effects on patients. For example, Nicin *et al.* (2020) showed that ACE inhibitor therapy for Covid-19 patients disrupted ACE2 expression in some cardiovascular cells, leading to the progression of pre-existing CVDs. CVD patients were also shown to be more vulnerable to Covid-19 infections (Boukhris *et al.*, 2020). Additionally, CVD-free Covid-19 patients were more likely to develop a CVD due to the immune system damaging cardiovascular tissues and/or adverse drug effects from Covid-19 therapies (Boukhris *et al.*, 2020).

CVDs also exhibit complex interactions with non-communicable diseases. Conditions like diabetes and obesity are closely linked with CVDs, where the metabolic imbalances caused by these diseases can accelerate the progression of CVDs (Casula *et al.*, 2017; Poirier and Eckel, 2002; Sowers *et al.*, 2001). For instance, insulin resistance contributes to CVD pathogenesis through mechanisms that impact vascular function and stiffness. A review by Hill *et al.* (2021) noted that elevated levels of aldosterone and insulin reduced nitric oxide production through the activation of specific cellular pathways like the endothelial sodium channel (EnNaC) and serum and glucocorticoid kinase 1 (SGK-1). This cascade diminishes vascular relaxation and leads to increased vascular and cardiac fibrosis and stiffness, promoting the development and progression of CVDs. Given that both communicable and non-communicable diseases can promote the development of CVDs and the progression of pre-existing CVDs, the demand for effective CVD treatments is ever-increasing.

Conventional drug-based therapies that can effectively treat CVDs are limited and come with inherent risks. For example, statin therapies, which act through reducing low-density lipoprotein (LDL) cholesterol biosynthesis (Stancu and Sima, 2001), have both positive and negative outcomes for CVD patients. In their meta-analysis, Mills *et al.* (2011) highlighted that statin therapy is correlated with significant reductions in the relative risk of major cardiac and vascular events brought on by CVDs. However, this meta-analysis also identified several studies which observed significantly increased risk for the development of diabetes in patients taking statins. A more recent review corroborates the correlation between statin therapies and reduced CVD risk, but highlights relatively poor performance in some contexts, identifying several primary and secondary prevention studies that observed reduced risks of CVD events ranging from 13% to 50% for patients treated with statins (Toth and Banach, 2019). This review also attributed poor statin performance in practice to underdosing and high statin discontinuation rates.

However, high dosages of certain statins, such as rosuvastatin, have also been associated with the development of type II diabetes (Ridker *et al.*, 2008). Notably, studies have shown that statins can induce new-onset diabetes in both low-risk (Casula *et al.*, 2017) and high-risk patients (Crandall *et al.*, 2017). Laakso and Kuusisto (2017) argue that the risk of developing type II diabetes is outweighed by the potential benefits statins may have against CVDs. However, it should be noted that diabetes was responsible for 1.5 million deaths in 2019 (WHO, 2021b). Furthermore, diabetes is already a great economic burden, with treatment costs estimated at $1.31 trillion in 2015 (Bommer *et al.*, 2017). Notably, another meta-analysis by

Armitage *et al.* (2019) observed seemingly reduced statin efficacy in older patients, but their findings were not statistically significant, in part due to a small sample size for older patients. Furthermore, their meta-analysis identified statin trials exclusively consisting of patients suffering from heart failure and kidney failure where the risk of major vascular events and vascular death were not significantly reduced. As statin-based therapies continue to be developed and refined, it is important to develop novel therapies targeting alternative molecular pathways driving CVD pathogenesis.

1.1.2 Molecular pathways for CVD pathogenesis

CVDs affect various cell types, including fibroblasts, cardiomyocytes, endothelial cells, macrophages, and vascular smooth muscle cells. In order to perform their functions, the activities of these cells are regulated by various molecular pathways. The dysregulation of these molecular pathways can disrupt vital cellular processes. The main function of cardiac myofibroblasts is the production of connective tissue in the heart's ECM, a dense network of proteoglycans, glycoproteins, and collagens that form a solid structural foundation (van den Borne *et al.*, 2010). Dysregulation of cardiac myofibroblast activity is known to promote the pathogenesis of CVDs such as cardiac fibrosis, which causes stiffening of the heart muscle due to increased ECM deposition and scar formation by overactive myofibroblasts (Frangogiannis, 2019).

On the other hand, cardiomyocytes drive the contraction and expansion of the heart. Their dysregulation has also been linked to the pathogenesis of CVDs. For example, in patients with ischaemic cardiomyopathy (ICM), a disease characterised by the weakening and dilation of the heart's left ventricle, cardiac myocytes may become elongated in response to the disease, diminishing their ability to function effectively and stressing the ventricular walls, thereby weakening the heart muscle further (Gerdes *et al.*, 1992).

Dysregulation of endothelial and vascular smooth muscle cells within blood vessels also contributes to CVD pathogenesis. Under healthy conditions, these cells play roles in distributing, absorbing, and secreting molecules throughout the body. One of the most common CVDs affecting blood vessels is coronary artery disease (CAD), a condition where plaque build-up in the arterial walls creates blockages and prevents blood flow (also known as atherosclerosis) (Mussbacher *et al.,* 2022). The dysregulation of endothelial cells in diseased coronary arteries has been shown to promote plaque formation by disrupting the activity of

smooth muscle cells, immune cells and microtubules (Hirase and Node, 2012). The dysregulation of these components leads to further progression of CAD.

Inflammatory pathways have also been shown to promote the development of CVDs (Henein *et al.,* 2022). Inflammation refers to an immune response to potentially harmful pathogens or cell damage characterised by greatly increased activity from the immune system (Chen *et al.*, 2017). Ideally, inflammation is responsible for removing damaged cells and potential pathogens/toxins/irritants, as well as initiating the repair of damaged tissues. However, dysregulated inflammation can contribute to disease pathogenesis. For example, Bryant *et al.* (1998) demonstrated that the overexpression of a pro-inflammatory factor, Tumour necrosis factor-α (TNF-α), in the cardiac cells of transgenic mice can promote heart failure due to increased collagen deposition, causing biventricular fibrosis. Similar studies, like Kolattukudy et al. (1998) found that other pro-inflammatory factors, such as the chemokine monocyte chemoattractant protein 1 (MCP-1), also promote cardiac fibrosis and myocarditis (inflammation of the heart) by promoting the proliferation of portal fibroblasts and subsequent trans-differentiation into collagen-secreting myofibroblasts (Kruglov *et al.*, 2006). In contrast, the chemokine IP-10 (Interferon-gamma-inducible protein) appears to attenuate fibrotic pathogenesis by inhibiting pathways responsible for promoting fibroblast migration and proliferation (Bujak *et al.*, 2009). This suggests that reducing anti-inflammatory molecules like IP-10 might disrupt fibroblast regulation and contribute to disease progression.

There are still many gaps in our understanding of how inflammatory factors influence CVD pathogenesis. For example, recent studies of inflammatory factors and CVDs have produced conflicting findings regarding interleukins. Interleukins are a group of inflammatory cytokines secreted by immune cells. A study by Bujak *et al.* (2008) found that Interleukin-1 (IL-1) signalling pathways were vital for activating both inflammatory and fibrogenic responses in infarcted hearts, and suppression of these pathways attenuated pathogenic fibrogenesis. This is supported by findings from Lim *et al.* (2002), which suggested that the disruption of IL-1 signalling using an IL-1 receptor antagonist attenuated both inflammation and fibrosis in mice. Conversely, a study by Mezzaroma *et al.* (2015) found that IL-1 knockdown had no effect on the pathogenesis of cardiac fibrosis in transgenic mice. Similarly, there are conflicting studies that support (González *et al.*, 2015) and contest (Lai *et al.*, 2012) the role of Interleukin-6 (*IL6*) signalling in promoting cardiac fibrosis.

While the pathogenesis of CVDs is entangled in a variety of complex systems and molecular pathways, gene dysregulation has been shown to play a critical role in CVDs (Sciarretta et al., 2018; Fujita and Ishikawa, 2011; Popa et al., 2007; Sethi, 2008). Investigating dysregulated genes in CVDs can reveal mechanisms for disease pathogenesis and provide direct targets for effective therapies. For example, disruption of the mTOR pathway has been linked to cardiac hypertrophy and increased heart failure risk (Sciarretta *et al.*, 2018), p53 dysregulation has been linked to cardiac cell apoptosis, also increasing heart failure risk (Fujita and Ishikawa, 2011), and TNF-α dysregulation has been linked to vascular inflammation, promoting plaque development and atherosclerosis (Popa *et al.*, 2007; Sethi, 2008).

By fully elucidating the molecular pathways and genes which are dysregulated in different CVD contexts and cardiovascular cell types, the aforementioned studies revealed potential therapeutic and prognostic targets for CVDs. However, there are still gaps in our understanding of the cell-specific molecular pathways and genes involved in driving the pathogenesis of many CVDs. To narrow this gap, this study investigated two genes *PXDN* and *PXDNL*, the former of which has been linked directly to CVD pathogenesis, while the latter has been hypothesized to regulate *PXDN* (Péterfi *et al*., 2014). The functions, pathways and transcriptional regulation of these genes have yet to be fully described in the context of the cardiovascular system and CVDs

1.1.3 *PXDN* Background

PXDN is the gene encoding the protein peroxidasin. Peroxidasin was first described in Drosophila where it was shown to interact with both immune cells and the ECM (Nelson *et al.*, 1994). Peroxidasin was later identified in humans where it was shown to be expressed in and secreted from various cells and tissues within the cardiovascular system (Cheng *et al.*, 2011; Cheng and Shi, 2022). Peroxidasin is characterised by its conserved haem-peroxidase domain, classing it under the protein superfamily: peroxidase-cyclooxygenases (Soudi *et al.*, 2012). Soudi *et al.* (2012) identified three other key structural motifs, leucine-rich repeats (LRR), immunoglobulin (Ig) domains, and a von Willebrand factor C module (VWC). Currently, peroxidasin is the only animal peroxidase known to contain both a haem-peroxidase domain and an Ig domain (Ero-Tolliver *et al.* 2015). Peroxidasin's structural motifs and expression across various cardiovascular cells and tissue types suggest that it may be a part of the immune response and ECM-based pathways in the cardiovascular system. Thus, the regulation of *PXDN*'s expression may play an important role in maintaining a healthy cardiovascular phenotype.

1.1.4 Structure and function

Peroxidasin's four major structural motifs have been linked to various cellular functions (Table 1.1): a conserved haem-peroxidase domain, as well as three additional structural motifs which are important for protein-protein interactions: a leucine-rich repeat domain; immunoglobulin domain and a von Willebrand factor C domain, (Shi *et al.*, 2011).

Table 1.1 Functions of *PXDN's* structural motifs

Structural Motif	Function

Haem-peroxidase domain	• Supports dityrosine crosslinking (Cheng *et al.*, 2011)
	• Catalyses oxidation of halides to form hypohalous acids
	  ○ Hypohalous acids hypothesised to promote collagen IV crosslinking (Bhave *et al.*, 2012)
	  ○ Hypohalous acids exhibit bactericidal activity (Li *et al.*, 2012)
	  ○ Hypobromous acid (HOBr) brominates Tyrosine residues (Bathish *et al.*, 2020)
	• Essential for cell proliferation in HUVEC and HEK cell lines (Lee *et al.*, 2020)
	• Essential for tissue development and maintenance in murine and fly models (Peebles *et al.*, 2024)

Leucine-rich repeat (LRR) domain	• Recognises and binds Collagen IV substrate sites (Bhave *et al.*, 2012) • Recognises and binds ECM proteins (eg. laminin-111) (Sevcnikar *et al.*, 2020)
Immunoglobulin domain(s)	• Recognises and binds Collagen IV substrate sites (Bhave *et al.*, 2012; Ero-Tolliver *et al.*, 2015) • Essential for cell proliferation in HUVEC and HEK cell lines (Lee *et al.*, 2020)
von Willebrand factor C terminal domain	• Recognises and binds Collagen IV substrate sites (Bhave *et al.*, 2012) • Cleavage determines structural conformation (Paumann-Page *et al.*, 2020; Sevcnikar *et al.*, 2020)
Structural mechanism undefined	• Promotes necrosis in endothelial cells (Zhang *et al.*, 2018) • Promotes angiogenesis (Medfai *et al.*, 2019)

The haem-peroxidase domain is a key structural motif of peroxidasin. This motif catalyses the oxidation of hydrogen peroxide (H_2O_2), producing hypohalous acid as a by-product (Cheng *et al.*, 2008). The source of H_2O_2 is NADPH oxidase, which is co-expressed alongside peroxidasin in many vascular cells (Shi *et al.*, 2011). These hypohalous acids play important roles in the ECM and in immune response (Li *et al.*, 2012). Bhave *et al.* (2012) demonstrated that hypohalous acid intermediates produced from *PXDN* activity catalysed the formation of sulfilimine bonds in the ECM and promoted collagen IV crosslinking. Notably, *PXDN* mutants had disorganized collagen IV networks and far fewer collagen IV crosslinks, suggesting that *PXDN*-mediated hypohalous acid production may be the primary pathway for forming these bonds.

Studies have shown that *PXDN* retains its hypohalous acid production in mammals. Bathish *et al.* (2020) investigated the function of hypobromous acid (HOBr) catalysed by *PXDN* in a PFHR-9 mouse cell line. The *PXDN* induced HOBr brominated the tyrosine residues of ECM proteins, which promoted collagen IV crosslinking in the ECM. Furthermore, Lee *et al.* (2020) showed that *PXDN*-induced collagen IV crosslinking in the ECM of human umbilical vein endothelial cells (HUVECs) promoted cell survival and proliferation. Peebles *et al.* (2024) extend these findings by linking *PXDN* to the development and maintenance of the basement membrane through collagen IV crosslinking. In their study, they observed significantly decreased viability in fly and murine mutants with catalytically inactive *PXDN*. Interestingly, more murine mutants survived compared to fly mutants, suggesting that mammals may have alternative pathways for forming collagen IV crosslinks. Furthermore, *PXDN* activity was required in adult mice to maintain tissue stiffness, suggesting that *PXDN* expression is important for maintaining the ECM even after cells have fully developed.

In addition to its role in collagen IV crosslinking, the haem peroxidase domain performs other functions. Firstly, hypohalous acid intermediates produced from *PXDN* have also been shown to increase cell survival through bactericidal and antioxidant activity. For example, Li *et al.* (2012) have shown that hypochloric acid (HOCl) and HOBr enhance the antimicrobial and antioxidant activity of resveratrols. Secondly, a study by Cheng *et al.* (2011) has shown that *PXDN*'s haem peroxidase domain may also catalyse dityrosine crosslinking between proteins in the ECM of human cell lines. While human studies on the function of tyrosine cross-linkage are limited, it has been suggested that these bonds may function to stabilize proteins (Partlow *et al.*, 2016).

PXDN's Ig domain has been shown to interact with proteins in the ECM. Lee *et al.* (2020) investigated the effects of adding mutant versions of *PXDN* without the Ig domain, as well as adding *PXDN* with Ig domain inhibitors, to human umbilical vein endothelial cells (HUVECs) without *PXDN*. Short interfering RNA (siRNA) was initially used to knock down *PXDN* expression in the HUVECs, resulting in decreased cell viability due to attenuated sulfilimine bonding and collagen IV crosslinking (Lee *et al.,* 2020). Notably, both rescue experiments failed to restore *PXDN* activity, suggesting that the Ig domain is essential for *PXDN*'s ECM functions. Studies have shown that *PXDN*'s other structural motifs may also facilitate interactions with proteins in the ECM. For example, Sevcnikar *et al.* (2020) showed that the von Willebrand factor C (VWC) terminal domain and Leucine-rich repeat (LRR) domain both facilitate strong binding interactions between human laminin and *PXDN* in the ECM.

Studies have also elucidated links between *PXDN*'s structural motifs and its structural conformation. Paumann-Page *et al.* (2020) identified two structural conformations for *PXDN*, a homo-trimer and an oligomer. The protein took on the homo-trimer conformation when the two structural motifs, VWC and LRR domain, were cleaved along with the C-terminal domain during post-transcriptional processing. This structural conformation was shown to catalyse more collagen IV crosslinking (Lázár *et al.*, 2015). If the C-terminal domain was not cleaved, the protein would take on the oligomer conformation. Lázár *et al.* (2015) showed that this conformation was more effective at binding to collagen IV "hot spots", suggesting that this post-transcriptional processing may play a key role in modulating *PXDN* activity. Recent studies have confirmed that proteolytic processing of *PXDN* by furin, a proprotein-convertase, is essential for both *PXDN's* integration into the ECM and its collagen IV crosslinking activity (Colon and Bhave, 2016; Kovács *et al.*, 2021).

Many of *PXDN*'s functions require further investigation to elucidate their mechanism of action. For example, one known function of *PXDN* is its interaction with signalling pathways involving Extracellular signal-Related Kinases (ERK1/2), Protein kinase B (Akt) and Focal Adhesion Kinase (FAK) to promote endothelial cell migration and angiogenesis (Medfai *et al.* 2019). However, it is still unclear how *PXDN* expression influences these pathways, as the study showed that neither the VWC nor the LRR domains were involved in pro-angiogenic or migratory signalling. Similarly, a study by Zhang *et al.* (2018) showed that *PXDN* expression induced necrosis of HUVECs damaged by oxidized low-density lipoprotein (ox-LDL), but the mechanism of action was not identified.

1.1.5 Regulation of *PXDN* expression

PXDN expression is vital for the survival, proliferation, and activity of various cardiovascular cell types. However, further investigations are needed to elucidate the pathways responsible for *PXDN*'s expression in the cardiovascular system, as these are largely unknown. This section will discuss three known regulatory pathways of *PXDN* expression.

Epithelial to Mesenchymal Transition (EMT) is responsible for the dedifferentiation of epithelial cells into a stem cell-like state (Kalluri and Weinberg, 2009). This increases cell survivability and gives cells the ability to migrate and invade (Kalluri and Weinberg, 2009). These characteristics are vital for processes such as regeneration and development in healthy cells but also occurs during the pathogenesis of diseases such as fibrosis and cancer (Das *et al.*, 2019; Thiery *et al.*, 2009). Thus, EMT is a potential target for disease therapies. Snai1 is the master regulator of EMT, and mediates *PXDN* repression in cervical carcinoma cells (Sitole and Mavri-Damelin, 2018). Sitole and Mavri-Damelin (2018) showed that by stimulating EMT using TGF-β, *PXDN*'s expression could be repressed in HeLa and SiHa cells. *PXDN*'s repression was associated with the binding of the Snai1 transcription factor (TF) to *PXDN*'s promoter region.

The TF Nuclear factor erythroid 2-related factor 2 (*Nrf2*), a master regulator of redox response, has been shown to upregulate *PXDN* expression in HeLa and HEK-293 cells (Hanmer and Mavri-Damelin, 2018). An *Nrf2* binding site was also identified in the promoter region of *PXDN*. Interestingly, the upregulation of *PXDN* expression by *Nrf2* was mediated by reactive oxidative species (ROS), namely tert-butylhydroquinone (tBHQ), H_2O_2 and sulforaphane (SFN). Given that the presence of ROSs stimulates *PXDN* expression, and since the *PXDN* protein is known to catalyse H_2O_2, it is possible that *PXDN* may play an important role in managing ROS-induced oxidative stress. Since excessive oxidative stress can disrupt important chemical pathways and is harmful to CVD patients (Senoner and Dichtl, 2019), it will be important for future studies to elucidate the role of *PXDN* expression in managing ROS-induced oxidative stress.

PXDN expression has also been shown to be silenced by folic acid-induced epigenetic silencing (Cui *et al.*, 2018b). Cui *et al.* (2018b) investigated the effects of folic acid supplementation on ox-LDL-treated HUVECs. The study found that folic acid silenced *PXDN* expression through extensive methylation of its promoter, which led to a marked increase in cell survival. Other epigenetic factors have also been shown to silence *PXDN* expression. For example, Kaufman

et al. (2018) found that the methylation of a site within *PXDN* was strongly associated with the risk of developing obesity in children with adverse childhood experiences. *PXDN* has also been suggested to be upregulated by the microRNA (miRNA) hsa-miR-203 in oesophagael cancer (Cai *et al.*, 2018) or even silenced (Desmond *et al.*, 2007) in acute myeloid leukemia (AML).

Thus, the regulation of *PXDN* expression can worsen, or inhibit disease pathogenesis depending on the context. This highlights that the regulation of *PXDN* expression can play a vital role in treating, diagnosing, and predicting the development of diseases, but further research is needed to elucidate the interplay between *PXDN* expression and the pathogenesis of different diseases.

1.1.6 *PXDN*'s role in disease

Due to its wide range of roles, the dysregulation of *PXDN* contributes to several diseases, from congenital defects of the eye to cancers and CVDs. As summarised below, studies have shown that both the upregulation and downregulation of *PXDN* expression can contribute to disease pathogenesis through a variety of pathways.

PXDN mutations have been associated with developmental disorders of the eye. In murine models, *PXDN* knockout mutants have been shown to exhibit impaired eye development, resulting in severe morphological defects (Kim *et al.*, 2019). Similarly, *PXDN* mutations in humans have been linked to developmental disorders of the eye, such as Anterior Segment Dysgenesis (ASD) and microphthalmia (Choi *et al.*, 2015; Zazo-Seco *et al.*, 2020), as well as congenital cataracts and glaucoma (Khan *et al.*, 2011; Micheal *et al.*, 2016). A study by Yang *et al.* (2016) identified at least seven genes which were differentially expressed in mice due to a *PXDN* mutation associated with eye disorders, suggesting that *PXDN*'s expression may regulate downstream targets associated with eye development. It is likely that the role of *PXDN* in developing and maintaining the ECM during cell differentiation is important for healthy eye development.

PXDN is also known to be dysregulated in cancers (Desmond et al., 2007). *PXDN* has many functions that can promote cancer growth and progression, ranging from stimulating angiogenesis and regulating oxidative stress (Dougan et al., 2019; Kurihara-Shimomura et al., 2020; Zheng and Liang, 2018) to resisting apoptotic signals and increasing cell invasion and motility. Thus, a shared pattern between ovarian cancer, prostate cancer and oral squamous cell carcinoma is the increased expression of *PXDN*. (Dougan *et al.*, 2019; Kurihara-Shimomura *et*

al., 2020; Zheng and Liang, 2018). Furthermore, a recent pan-cancer analysis found significant differences in *PXDN* transcript and protein levels for 31 tumour types when comparing healthy and diseased tissues (Zhou *et al.*, 2022). Additionally, Zhou *et al.* (2022) found that *PXDN* hypermethylation in tumour tissues was correlated with longer patient survival times. However, Desmond *et al.* (2007) have linked the epigenetic silencing of *PXDN* to the progression of acute myeloid leukaemia (AML). The silencing of *PXDN* may be beneficial for liquid tumours, like AML, which retain a dedifferentiated phenotype through decreasing activities linked to ECM stabilisation and cell differentiation.

Like solid tumours, fibrotic diseases may also use *PXDN*'s ECM stabilising function to promote their pathogenesis, as they are characterised by hardening of the ECM and the excess deposition of ECM proteins (Wynn, 2008). One study by Péterfi *et al.* (2009) investigated *PXDN*'s activity in cell and animal models for kidney fibrosis. They found that treatment of human dermal and pulmonary fibroblasts with TGF-β1 (a fibrogenic signalling molecule) upregulated the expression of *PXDN*, which then co-localised in structural molecules in the extracellular space. Notably, *PXDN* expression was also increased in a murine unilateral ureteral obstruction (UUO) model of kidney fibrosis. Furthermore, *PXDN* was concentrated in the peritubular space of the kidneys. These findings suggest that *PXDN*'s expression and secretion may promote fibrosis through increasing crosslinking ECM proteins and increasing ECM stability. A recent study in a murine UUO model of kidney fibrosis found that *PXDN* knockout mice exhibited reduced fibrosis, suggesting that *PXDN* may be a therapeutic target for fibrosis (Colon *et al.*, 2019).

Similarly to cancers and fibrosis, CVDs are often associated with increased *PXDN* expression. Cui *et al.* (2018b) have shown that elevated *PXDN* expression can increase oxidative stress. Oxidative stress is well known to contribute to the development of new CVDs and the progression of existing CVDs (Senoner and Dichtl, 2019). For example, Wang *et al.* (2019) showed that mice with pulmonary hypertension suffered degradation of their endothelial progenitor cells due to *PXDN*-induced oxidative stress. A recent study by Costas *et al.* (2021) in patients with Peripheral Artery Disease (PAD) observed Rapid Kidney Function Decline (RKFD) in patients with elevated *PXDN* expression. Elevated *PDXN* expression has also been linked to the progression of cardiac fibrosis post-myocardial infarction (MI), likely due to *PXDN*'s roles in the ECM (Liu *et al.*, 2019).

Table 1.2 Diseases linked to *PXDN* dysregulation

Disease	Aberration of *PXDN*
Anterior Segment Dysgenesis, Congenital Cataracts, Congenital Glaucoma, Eye Malformation and Microphthalmia	• Mutant genotypes (Choi *et al.*, 2015; Khan *et al.*, 2011; Kim *et al.*, 2019; Micheal *et al.*, 2016; Yang *et al.*, 2016; Zazo-Seco *et al.*, 2020)
Prostate Cancer	• Increased *PXDN* expression (Dougan *et al.*, 2019) o Regulates oxidative stress o Prevents apoptosis o Promotes cancer progression
Oral Squamous Cell Carcinoma	• Increased *PXDN* expression (Kurihara-Shimomura *et al.*, 2020) o Regulates oxidative stress o Prevents apoptosis o Promotes cancer progression
Ovarian Cancer	• Increased *PXDN* expression (Zheng and Liang, 2018) o Regulates oxidative stress o Prevents apoptosis o Promotes invasion and proliferation o Promotes cancer progression
Acute Myeloid Leukaemia	• Silenced *PXDN* expression (Desmond *et al.*, 2007) o Promotes cancer progression
Kidney Fibrosis	• Increased *PXDN* expression (Péterfi *et al.*, 2009) o Promotes renal fibrosis o Increased secretion into kidney ECM o *PXDN* knockout mice exhibit reduced renal fibrosis (Colon *et al.*, 2019)
CVD; Pulmonary Hypertension	• Increased *PXDN* expression (Wang *et al.*, 2019) o Increased H_2O_2 and Hypochlorous acid o Increased apoptosis of endothelial progenitor cells (EPCs) o Dysregulation of EPC differentiation o Promotes oxidative damage of EPCs
CVD; Peripheral Artery Disease (PAD)	• Increased *PXDN* expression o Associated with Rapid Kidney Function Decline (RKFD) (Costas *et al.*, 2021) • Mechanism unknown
CVD; Cardiac fibrosis	• Increased *PXDN* expression (Liu *et al.*, 2019) o Promotes pathogenesis

The mechanisms behind many of the pathologies in Table 1.2 remain unresolved, and the growing need for effective therapies calls for the identification of new therapeutic targets. TFs regulating *PXDN* expression in cardiovascular cells may be an ideal target for regulating *PXDN* expression. However, another potential regulator/inhibitor of *PXDN* is its homolog *PXDNL*.

1.2 *PXDNL*

The gene *PXDNL* is a poorly characterized homolog of *PXDN,* which differs in expression, structure, and function (Péterfi *et al.,* 2014). Unlike *PXDN*, which is expressed more widely, *PXDNL* expression is limited to the heart, specifically in cardiomyocytes (Cheng *et al.*, 2008; Péterfi *et al.*, 2014). The structure of *PXDNL*, as described by Péterfi *et al.* (2014) , shares the same domains that are unique to *PXDN* (LRR domain, VWC domain, and Ig domain). With these domains, *PXDNL* is capable of translocating to the ECM like *PXDN*, however, it has lost its catalytic activity due to the amino acid changes in its peroxidase domain. Péterfi *et al.* (2014) have hypothesized that since *PXDNL* is expressed only in the heart, loss of the peroxidase domain may be beneficial because the heart has a high production of ROS, and peroxidase activity may increase oxidative stress.

Like *PXDN*, the dysregulation of *PXDNL* has been linked to disease. For example, Lu *et al.* (2023) found that *PXDNL* was highly expressed in urinary bladder cancer (UBC) and promoted cell motility. Further knockdown and overexpression experiments in UM-UC-3 and TCCSUP UBC cell lines revealed that *PXDNL* may have acted through the Wnt/β-catenin pathway. These findings suggest that *PXDNL* could serve as a potential therapeutic target and prognostic biomarker for UBC. In relation to CVDs, Péterfi *et al.* (2014) have shown that *PXDNL* expression is increased in failing hearts, where it may play an antagonistic role and suppress *PXDN*. In addition to heart failure, variants of *PXDNL* have been linked to a cardiac arrythmia syndrome in human-induced pluripotent stem cell-derived cardiomyocytes (hiPSC-CMs) (Barajas-Martinez *et al.*, 2020), but only in the presence of other genetic variants.

PXDN and *PXDNL* have both been shown to be expressed in cardiovascular tissues (Péterfi et al., 2014) and have been linked to the pathogenesis of CVDs (Costas et al., 2021; Liu et al., 2019; Barajas-Martinez et al., 2020). However, the functions of *PXDN* and *PXDNL* in cardiovascular cells and the molecular mechanisms linking their dysregulation to CVD pathogenesis remain unclear. In order to fill these gaps, we need to identify the transcriptional elements regulating *PXDN* and *PXDNL* expression in cardiovascular cells and explore their expression patterns in different CVD contexts.

1.3 Bioinformatics: Data mining

Bioinformatics tools enable the mining of novel insights from a wide range of -omics data stored in openly accessible databases. This study looks at two different types of data: ChIP-seq data and gene expression microarray data. Each data type reveals insights into molecular pathways involving the regulation and expression of *PXDN* and *PXDNL*. However, the utility of public datasets is often hindered by variability in data quality, lack of standardization, and incomplete metadata (Cahan *et al.*, 2007), thus sufficient quality control and data selection criteria are essential for mining useful insights.

Chromatin Immuno-Precipitation (ChIP) is a technique wherein proteins are bound to DNA to form protein-DNA complexes which are subsequently isolated using antibody-mediated precipitation (Park, 2009). "-seq" refers to the application of DNA to sequencing to determine the sequence of DNA to which the protein (transcription factor/regulator) is bound. A major application of ChIP-seq is identifying transcription factor binding sites (TFBSs) (Spencer *et al.*, 2003). ChIP-seq can also be used to elucidate variations in regulatory transcription factor (TF) binding in different cell lines, offering deeper insights into cell-specific gene regulation. For this study, ChIP-seq data was sourced from the ChIP-Atlas database (https://chip-atlas.org/). ChIP-Atlas contains an extensive library of annotated ChIP-seq datasets as well as built-in visualisation and analysis tools (Oki *et al*., 2018).

The second -omics data type used in this study is from gene expression microarrays. Microarray data reveals insights into the transcriptome by indirectly measuring the expression of RNA transcripts for thousands of genes. Gene expression microarrays function by hybridizing cDNA or RNA samples to thousands of gene-specific probes coated or fixed onto wells in a microarray chip (Lockhart *et al.*, 1996; Schena *et al.*, 1995). Hybridized nucleic acids are subsequently detected and quantified using scanners which measure fluorescent signals released upon hybridization, offering a snapshot of gene expression across the genome (Lockhart *et al.*, 1996; Schena *et al.*, 1995). Microarrays have a variety of applications, notably in elucidating changes in gene expression that contribute to disease pathogenesis and biomarker identification. For example, a study by Seo *et al.* (2004) used gene expression microarrays to investigate genes linked to the pathogenesis of atherosclerosis. Their study identified unique gene expression profiles for predicting atherosclerosis as well as candidate genes that may contribute to atherosclerosis-promoting pathways, making them potential therapeutic targets.

Given the complexity and variety of CVDs, microarrays offer notable advantages in identifying potential gene pathways for further investigation. High-throughput analysis allows for the simultaneous examination of thousands of genes to uncover informative expression patterns and pathways which may be involved in conditions like heart failure and atherosclerosis. Furthermore, the public availability of microarray datasets through databases such as Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/) and ArrayExpress (https://www.ebi.ac.uk/biostudies/arrayexpress) provide a rich resource for validation and comparison, expanding the scope and reliability of findings in CVD research (Barrett and Edgar, 2006; Brazma *et al.*, 2003). For example, in a recent review by Khomtchouk *et al.* (2020), publications on CVD research had produced over 200 microarray datasets, which can be accessed from GEO. Despite these advantages, microarrays also have limitations, such as issues with probe sensitivity, that can impact the accuracy of gene expression profiling, particularly in detecting low-abundance transcripts which may be relevant to cardiovascular diseases (Draghici *et al.*, 2008; Murphy, 2002).

1.4 Introduction to current study

The growing prevalence of CVDs prompts research into the vital molecular pathways of the cardiovascular system, identifying therapeutic targets and developing effective CVD therapies. Two promising targets are *PXDN* and *PXDNL*, two genes expressed in the cardiovascular system and linked to the pathogenesis of CVDs such as cardiac fibrosis (Liu *et al.*, 2019), PAD (Costas *et al.*, 2021), heart failure (Péterfi *et al.*, 2014), and cardiac arrythmias (Barajas-Martinez *et al.*, 2020). However, there is a significant gap in our knowledge of their transcriptional regulation in the cardiovascular system as well as their role in different CVDs. Bioinformatics data-mining tools provide a unique opportunity to repurpose publicly available datasets to investigate these genes. ChIP-seq and gene expression microarray data offer opportunities to identify transcriptional regulators of *PXDN* and *PXDNL* and investigate their gene expression profiles in CVD-related contexts. However, when mining public databases, the availability and quality of datasets require careful consideration. As no datasets were available for CVDs such as cardiac fibrosis and PAD, datasets investigating related diseases and disease-simulating conditions were selected. Furthermore, extra quality control steps were taken to ensure the integrity of selected datasets. Exploring the roles of *PXDN* and *PXDNL* in CVDs may reveal novel pathways that contribute to the pathogenesis of different CVDs and provide valuable insights for developing targeted therapies and biomarkers.

1.5 Aims and objectives

This study aimed to identify the transcriptional regulators of *PXDN* and *PXDNL* in the cardiovascular system and explore their mRNA expression profiles in CVD-related contexts. To accomplish this, the study was guided by the following objectives:

Objectives:

1. To develop a bioinformatics pipeline to mine processed ChIP-seq data from the ChIP-Atlas database
2. To identify transcription factors that potentially regulate the expression of *PXDN* and *PXDNL* in cardiovascular cells
3. To develop a bioinformatics pipeline to mine and process microarray data from the GEO sequence data repository
4. To establish if the expression of *PXDN*, *PXDNL*, and/or their regulatory TFs is altered in the context of CVDs.

2 MATERIALS AND METHODS

In order to complete the objectives of this study, a data mining and analysis workflow was developed (Figure 2.1). This workflow was divided into two pipelines. The first pipeline was used to identify regulatory transcription factors (TFs) of *PXDN* and *PXDNL* in cardiovascular cells, while the second pipeline was used to explore the expression profiles of these genes and their regulatory TFs under different cardiovascular disease (CVD) contexts.

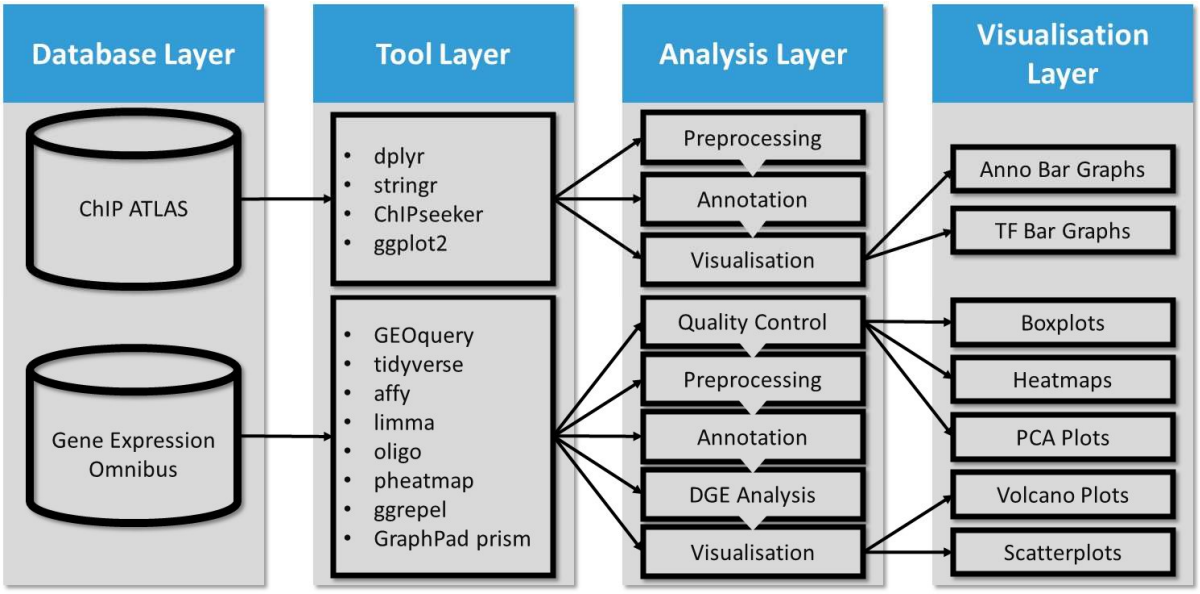


Figure 2.1 Overview of the workflow used to mine and analyse data from the ChIP-Atlas and GEO databases. Four-layer workflow used for ChIP-seq and gene expression microarray data mining. Both data mining pipelines extensively used R-based tools for data handling, analysis and visualization. The microarray pipeline also required the GraphPad Prism tool for further visualization.

This study used datasets from two public databases, namely ChIP-Atlas (Oki *et al.*, 2018) and the Gene Expression Omnibus (GEO) (Barrett and Edgar, 2006; Edgar, 2002). The tool layer showcases the tools used in both pipelines, with the majority of the included R packages being sourced from the Bioconductor repository (Huber *et al.*, 2015). A single standalone tool, GraphPad Prism v10.1.0 was used for additional visualisations. Both pipelines were primarily developed using Rv4.3.0 (R Core Team, 2023). The analysis layer shows how these tools were used to process the datasets, while the final layer, visualization, refers to the various graphical representations and plotting methods used to investigate the raw and processed data.

2.1 ChIP-seq data mining pipeline

The first pipeline, shown in Figure 2.2, was used to identify the potential regulatory TFs of *PXDN* and *PXDNL* in cardiovascular cells using ChIP-seq data. ChIP-seq data is generated by combining chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing (-seq) to identify the binding sites of DNA-associated proteins (Barski *et al.*, 2007; Johnson *et al.*, 2007; Park, 2009). ChIP is a technique used for isolating transcription factors of interest while they are bound to DNA, thereby allowing for characterization of TF binding sites (Weinmann and Farnham, 2002; Wells and Farnham, 2002). As described by Wells and Farnham (2002), ChIP begins with the crosslinking of proteins to DNA using formaldehyde. Following protein-DNA crosslinking, the DNA is fragmented using either sonication or enzymatic digestion. These fragments are then immunoprecipitated using antibodies that target the protein of interest. The antibody-protein-DNA complexes are isolated, and the crosslinks are reversed, freeing the DNA. In ChIP-seq, the previously protein-bound DNA is then purified and sequenced, producing a library of reads that can be mapped back to a reference genome using peak calling software to characterize TF binding sites (Park, 2009).

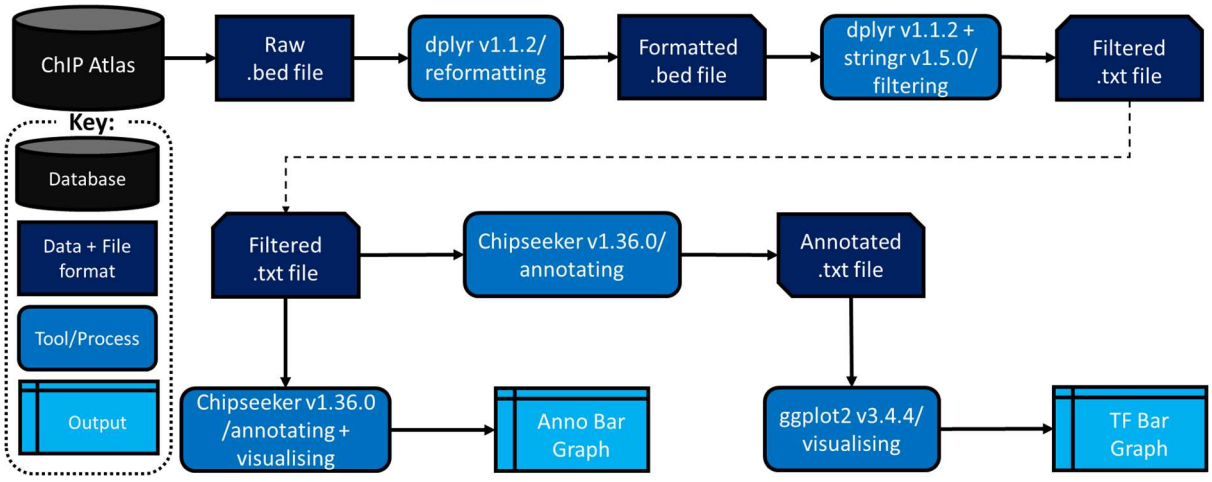


Figure 2.2 ChIP-seq data mining pipeline. This pipeline highlights the key packages, files and processes in identifying regulatory TFs from ChIP-seq data. First, data was collected from ChIP-Atlas, then reformatted and filtered using dplyr v1.1.2 (Wickham *et al.*, 2023) and stringr v1.5.0 (Wickham, 2023). Next, ChIPseeker v1.36.0 (Yu *et al.*, 2015) was used to annotate and visualize the distribution of TF binding peaks within *PXDN* and *PXDNL* on clustered Anno bar graphs. Finally, ggplot2 v3.4.4 (Wickham, 2011) was used to visualize TF binding peaks using the annotated genomic regions of interest.

2.1.1 ChIP-Atlas background

ChIP-Atlas, established in 2015, is an online data mining resource that stores omics data from multiple public databases, covering a wide range of species, cell types, and conditions (Oki *et al.*, 2018). ChIP-Atlas was selected as the ChIP-seq data source in this study for two reasons. Firstly, ChIP-Atlas contains peak binding results from a large number of ChIP-seq studies, with approximately 400 experiments across 63 TFs in cardiovascular cell lines. Secondly, these results are standardised to ensure consistency and comparability between different datasets. This is achieved through a standardised pipeline which processes the raw data from ChIP-seq experiments and analyses them using the MACS2 (Model-based Analysis of ChIP-seq 2) peak calling algorithm (Oki *et al.*, 2018; Zhang *et al.*, 2008).

Developed by Zhang *et al.* (2008), the original MACS peak calling software was used for analysing ChIP-seq data and identifying genomic regions enriched for DNA-binding proteins. The algorithm begins by aligning tags (ChIP-seq reads) to a reference genome. These tags represent the ends of the protein-bound DNA fragments. The estimated distance between the tags and the centre of their DNA fragment, known as the shift size, is then modelled to accurately predict the protein binding site (Zhang *et al.*, 2008). MACS then extends these tags in both directions to the estimated fragment length, allowing for the accumulation of signal at the centre of the fragments, and applies a sliding window across the genome to identify regions with a significantly higher number of overlapping tags than expected by chance. This algorithm applies a dynamic statistical model to accommodate local genomic variations, enhancing the reliability of peak detection (Zhang *et al.*, 2008). MACS2 builds on the original MACS peak caller by introducing an improved model for better fragment size estimation, offering the capability to analyse both narrow peaks typical of transcription factor binding sites and broad peaks associated with histone modifications, and implementing more sophisticated methods for controlling the false discovery rate and handling duplicates (Gaspar, 2018; Mistry, 2017). In ChIP-Atlas, the scores output by MACS2, typically in the form of p-values or q-values (adjusting for false discovery rates), indicate the statistical significance of each peak.

2.1.2 Data collection

All ChIP-seq TF binding data used in this study was retrieved using the Peak Browser tool in ChIP-Atlas' data mining suite, accessible at https://chip-atlas.org/. The Peak Browser tool offers several filters when downloading ChIP-seq results. These filters include the Antigen, Cell class, Cell type and Threshold for significance. For this study, ChIP-seq data was filtered to include only binding results for TFs in cardiovascular cell lines with a threshold value of 200. The threshold for significance is a log-transformed MACS2 q-value, where higher scores imply stricter criteria, excluding more false positive peaks but simultaneously increasing the number of false negative peaks. A threshold of 50 was the lowest selectable value, equivalent to MACS2 q-value $< 1 \times 10^{-5}$. This study adopted a more stringent threshold value of 200 to identify high confidence TFs.

$$Threshold\ for\ Significance = (-10 \times \log_{10}[MACS2\ q\ value])$$

Following the documentation provided by Oki (2020), ChIP-seq peak binding data was downloaded as a .bed file, which contained binding data for TFs aligned to the GRCh38 human reference genome. For every identified peak, this file contained the chromosomal location (chromosome number and start and end positions of the peak), the TF associated with the peak, the statistical significance of the peak (represented as threshold values), and additional metadata regarding the experiment from which the peak results was generated. These metadata allowed for identifying details about each experiment such as whether cells were untreated or treated and the treatment duration, which were necessary to identify TFs bound to *PXDN* and *PXDNL* under various conditions. While the focus of this investigation was the identification of transcriptional regulators under healthy cardiovascular conditions, TFs identified under treated or unspecified conditions were still retained as they may still bind under healthy conditions.

2.1.3 Data pre-processing

Pre-processing (i.e. filtering and reformatting) was essential in preparing the .bed file for further analysis. Two R packages were used for these tasks. First, dplyr v1.1.2 (Wickham *et al.*, 2023) and stringr v1.5.0 (Wickham, 2023) were used to convert the raw .bed file downloaded from ChIP-Atlas in BED9 + GFF3 format into a standard .bed format compatible with ChIPseeker v1.36.0 (Yu *et al.*, 2015). Notably, this .bed file contained peak binding data aligned to the entire genome and needed to be filtered to include only the regions of interest. Thus, dplyr v1.1.2 was used to filter the formatted .bed file for the genomic regions encompassing *PXDN* and *PXDNL*. To remain consistent with the selected reference genome, the genome coordinates for both genes were based on the GRCh38 reference genome. Genomic coordinates for both genes were sourced from Ensembl, and the regions selected for *PXDN* and *PXDNL* were chr2:1629887-1746901 and chr8:51317577-51810445 respectively (Ensembl, 2023a, 2023a). 2 kb regions flanking each gene were also included to capture TFs binding within the distal intergenic region. This preprocessing stage produced two formatted, filtered .bed files, one for TF binding peaks within and around *PXDN*, the other for *PXDNL*. These files would be filtered further after annotation to focus on key regions for transcriptional regulation.

2.1.4 Data analysis and visualisation

ChIPseeker v1.36.0 (Yu *et al.*, 2015) was used in conjunction with the TxDb.Hsapiens.UCSC.hg38.knownGene package to annotate the filtered TF binding peaks. ChIPseeker was designed to annotate ChIP-seq peaks relative to gene locations, categorize peaks into genomic regions, and produce informative visualizations (Yu *et al.*, 2015). In this study ChIPseeker was used to annotate the genomic features of interest within a 10 kb region up- and downstream of *PXDN* and *PXDNL's* transcription start sites (TSS). In addition to annotating genomic regions, ChIPseeker was used to explore and compare TF binding patterns within these genomic regions to identify potential regulatory regions enriched with binding peaks.

This study focused on two genomic regions enriched for binding peaks, the promoter and the first intronic region. The promoter region, typically found upstream of the transcription start site (TSS), plays a crucial role in the initiation of transcription and is a common binding site for TFs that regulate gene expression (Smale and Kadonaga, 2003). TF binding in the promoter region often correlates with gene activation or repression, making it a key target in understanding gene regulatory networks. Similarly, the first intronic region has emerged as an important site for transcriptional regulation. Introns, particularly the first intron, are known to contain regulatory elements such as enhancers, silencers, and insulators, which can significantly modulate gene expression (Andersson *et al.*, 2014; Blackwood and Kadonaga, 1998). These regions can affect the efficiency of transcription, mRNA processing, and export to the cytoplasm (Luco *et al.*, 2011).

The high confidence TFs binding within each gene's promoter and first intron were analysed further to investigate the cell types in which binding peaks were observed and to compare TFs between both genes. Bar graphs were used to investigate the number of studies identifying binding peaks for each high confidence TF in *PXDN* and *PXDNL* and the cell types they were identified in. Venn diagrams were then used to compare the number of TFs binding to *PXDN* and *PXDNL*, as well as TFs binding to their first introns and promoters. The expression profiles of these high confidence TFs were investigated further in CVD-related contexts.

2.2 Microarray data mining pipeline

The second pipeline (Figure 2.3) analysed gene expression microarray data to investigate the expression profiles of *PXDN*, *PXDNL*, and their regulatory TFs in different CVDs. Microarrays are used to compare genome-wide expression levels between treatment groups (Heller *et al.*, 1997; Schena, 1996; Schena *et al.*, 1995). First, mRNA is isolated from sample cells, reverse transcribed into more stable cDNA transcripts, and fluorescently labelled. The cDNA is then applied to a microarray chip containing thousands of wells containing probes corresponding to a specific gene. After the cDNA hybridizes with the probes, a scanner is used to detect the fluorescent intensity of each well, which is used to infer changes in gene expression levels relative to a reference/control (Schena *et al.*, 1995). The current study investigated data from single dye systems and dual dye systems. The single dye microarrays, (Affymetrix platforms), measure gene expression with a single fluorescent label, while dual dye systems, (Unigene 3.1 cDNA platform), utilize two different fluorescent labels for comparative expression analysis (Lockhart *et al.*, 1996; Lockhart and Winzeler, 2000).

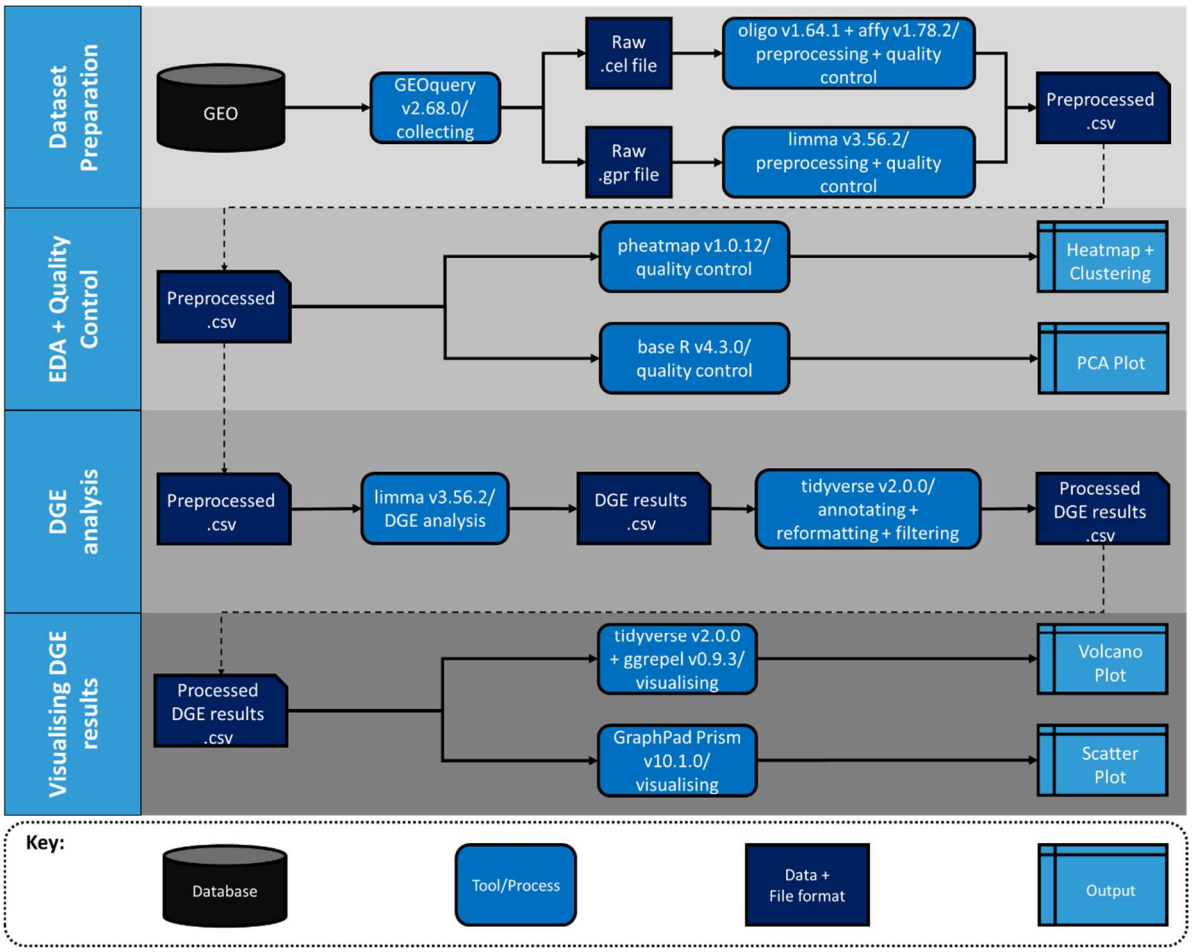


Figure 2.3 Gene Expression Microarray data mining pipeline. This pipeline highlights the key components and steps in mining insights from microarray datasets. First, the raw and supplementary data of each dataset were loaded using GEOquery v2.68.0 (Davis and Meltzer, 2007). Next, limma v3.56.2 (Ritchie *et al.*, 2015) was used to preprocess single-dye microarray datasets while oligo v1.64.1 (Carvalho and Irizarry, 2010) and affy v1.78.2 (Gautier *et al.*, 2004) were used to preprocess dual-dye microarray datasets. The preprocessed data underwent further quality control and exploratory data analysis using base R v4.3.0 (R Core Team, 2023) and pheatmap v1.0.12 (Kolde, 2019). Differential gene expression analysis was then performed using limma v3.56.2. Finally, the Differential gene Expression (DGE) results were mined for insights and visualized using tidyverse v2.0.0 (Wickham *et al.*, 2019), ggrepel v0.9.3 (Slowikowski, 2024) and GraphPad Prism v10.1.0.

2.2.1 GEO background

The Gene Expression Omnibus (GEO) (established in 2002) is public database for that offers an extensive collection of gene expression datasets, including microarrays (Barrett and Edgar, 2006; Edgar, 2002). There were three reasons this study opted for GEO over other microarray repositories, such as ArrayExpress. Firstly, a preliminary investigation identified 112 datasets in GEO compared to only 81 in Expression Atlas when searching for gene expression microarray datasets in cardiovascular cells for *Homo sapiens*. Secondly, GEO has GEO2R, a built-in tool which utilises the GEOquery (Davis and Meltzer, 2007) and limma (Ritchie *et al.*, 2015) R packages for assessing dataset quality, and performing differential gene expression analysis (Davis and Meltzer, 2007; Smyth, 2005, 2004). Lastly, the search function in GEO allowed for more advanced filtering of datasets, which facilitated the identification of CVD-related microarray datasets. It should be noted that while GEO2R was useful for performing rough analyses and quality control on microarray datasets, it did not provide the same flexibility or transparency as a custom pipeline, which could use a wider variety of tools for preprocessing, analysing and visualising data.

2.2.2 Data collection

Sixteen CVD-related datasets were filtered from roughly 112 datasets for human cardiovascular microarrays. In this study, CVD-related refers to studies investigating cardiovascular tissues/cells under CVD conditions or conditions which simulate and/or promote CVD conditions. Table 2.1 summarizes the tissue/cell type, treatment conditions, sample size, and microarray platform for each GEO dataset. Additionally, the table includes important supplementary information, such as whether the microarray results were validated using quantitative polymerase chain reaction (qPCR) (in the study they were sourced from), whether probes for *PXDNL* were present in the dataset, or whether the datasets were excluded from DGE analysis.

Table 2.1 Summary of key selection criteria for the inclusion of identified datasets

GEO ID	Cell type	Treatment conditions	Sample size	*PXDNL* probes	RT-qPCR validation	Dataset Exclusion
GSE26887	Non-Infarcted Left Ventricular Tissue	Diabetic Heart Failure/ Non-diabetic Heart Failure/ Control	24	Present	None	Included
GSE42955	Left Ventricular Tissue	Dilated Cardiomyopathy/ Control	17	Present	Validated (5/7)	Included
GSE3585	Subendocardial Left Ventricular Tissue	Dilated cardiomyopathy/ Non-failing heart	12	Absent	Validated (11/12)	Included
GSE3586	Septal Myocardial Tissue	Dilated cardiomyopathy/ Non-failing heart	28	Absent	Validated (11/12)	Included
GSE1869	Myocardial Tissue	Ischemic Cardiomyopathy/ non-Ischemic Cardiomyopathy/ Control	37	Present	Validated (27/32)	Included
GSE2638	Microvascular endothelial cells	TNF-α/ Control	6	Absent	Validated (19/19)	Included
GSE2639	Human umbilical vein endothelial cells (HUVECs)	TNF-α/ Control	8	Absent	Validated (19/19)	Included
GSE48060	Peripheral Blood	first-time Acute Myocardial Infarction (AMI) with recurrent cardiovascular events at 18 months/first-time AMI without recurrent cardiovascular events at 18 months	31	Present	None	Excluded†
GSE13760	Arterial Tissue	T2 Diabetic Males/Non-diabetic Controls	21	Absent	Validated (3/5)	Excluded†
GSE6257	Lymphatic endothelial cells	TNF-α/ Control	8	Present	None	Excluded†
GSE4567	Human Pulmonary Artery Endothelial Cells (HPAECs)	Ultrafine particle exposure/Control	8	Present	Validated (2/2)	Excluded†
GSE16683	HUVECs	Estradiol/ control	6	Present	None	Excluded*
GSE1145	Left Ventricular Tissue	Idiopathic Dilated Cardiomyopathy/Ischemic Cardiomyopathy/Alcoholic Cardiomyopathy/Valvular Cardiomyopathy/Hypertrophic Cardiomyopathy/Control	107	Present	None	Excluded*
GSE4172	Parvovirus B19 Positive Endomyocardial Biopsies	Inflammatory Dilated Cardiomyopathy/ Control	12	Present	None	Excluded*
GSE9128	Peripheral Blood Mononuclear Cells	Ischemic Heart Failure/ non-Ischemic Heart Failure/ Control	11	Absent	Validated (5/5)	Excluded*
GSE67492	Right Ventricular Tissue	Idiopathic Dilated Cardiomyopathy (IPC)/ Pulmonary Arterial Hypertension with BMPR2 mutation/ Healthy control	6	Present	None	Excluded ‡

* *Missing raw/supplementary data*
† *Treatment/cell line/tissue type invalid for modelling cardiovascular disease*
‡ *Insufficient sample size*

The inclusion criteria for datasets were important for ensuring data integrity, as poor-quality datasets would compromise the validity of the differential gene expression results. The criteria were as follows: datasets had to be in cardiovascular cells or tissues, consist of treatment groups that were CVD-related (as defined earlier), provide the raw datasets and possess an adequate sample size for statistically significant analysis. The latter criterion is especially important as a review by Ioannidis (2010) noted insufficient sample size as a major concern in research. CVD-related datasets were identified and manually reviewed using GEO's profile browsing tool (accessed at: https://www.ncbi.nlm.nih.gov/geoprofiles/). For this study, datasets were only included if they had three or more samples per treatment group. Priority was also given to datasets validated via reverse transcription-quantitative polymerase chain reaction (RT-qPCR), providing an extra dimension of data reliability. However, it should be noted that none of the genes or TFs of interest for this study were included in the qPCR validated genes, and so differential gene expression observed in this analysis would still need to be validated in future studies.

Unfortunately, as shown in Table 2.1, only two datasets were identified that included probes for both *PXDN* and *PXDNL*. Due to the small number of datasets, an additional five datasets were added. These datasets met all other selection criteria but lacked probes for *PXDNL*. The seven selected datasets spanned five studies and three CVD contexts, namely cardiomyopathies, heart failure and TNF-α stimulation. While not being directly CVD-related, the TNF-α datasets were chosen because TNF-α stimulation in HUVEC cell lines may mimic the inflammatory response seen in certain CVDs. Studies have suggested that TNF-α induced inflammation may promote vascular inflammation, plaque development and endothelial dysfunction, leading to the development of CVDs such as atherosclerosis and heart failure (Popa *et al.*, 2007; Sarziputtini *et al.*, 2005; Sethi, 2008; Zhang *et al.*, 2009), however, the precise molecular mechanisms have not been fully elucidated.

2.2.3 Data preparation

The first step in data preparation was the collection of microarray expression data from the relevant GEO datasets using GEOquery (Davis and Meltzer, 2007) an R-based tool used for retrieving raw datasets and supplementary materials from GEO. The GEO IDs, microarray platforms and study citations are included in Table 2.2.

Table 2.2 Summary of selected microarray datasets

GEO ID	Microarray Platform	Study Citation
GSE42955	Affymetrix Human Gene 1.0 ST Array	(Molina-Navarro *et al.*, 2013)
GSE3585	Affymetrix Human Genome U133A Array	(Barth *et al.*, 2006)
GSE3586	Unigene3.1 cDNA Array	(Barth *et al.*, 2006)
GSE26887	Affymetrix Human Gene 1.0 ST Array	(Greco *et al.*, 2012)
GSE76701	Affymetrix Human Genome U133 Plus 2.0 Array	(Kim *et al.*, 2016)
GSE2638	Affymetrix Human Genome U133A Array	(Viemann *et al.*, 2006)
GSE2639	Affymetrix Human Genome U133A Array	(Viemann *et al.*, 2006)

In addition to consisting of cell line and tissue data, the selected datasets also spanned across three distinct microarray platforms: Affymetrix Human Genome U133A Array, Affymetrix Human Genome U133 Plus 2.0 Array, Unigene3.1 cDNA Array, and Affymetrix Human Gene 1.0 ST Array. A consequence of incorporating datasets from multiple platforms was that each used a unique file format. The Affymetrix datasets consisted of .cel files, which contained the fluorescence data of probes for a single-dye microarray. In contrast, the Unigene3.1 dataset contained the fluorescence data of probes for a dual-dye microarray, where each microarray chip produced two files (one for each dye). Additionally, the newer Affymetrix array (Affymetrix Human Gene 1.0 ST Array) used a different file compression format, .gz or gzipped, which required a separate R package and the use of external tools such as 7-zip (https://7-zip.org/).

Different microarray platforms with different file formats required different tools and packages for preprocessing. Thus, the pipeline incorporates the affy v1.78.2 (Gautier *et al.*, 2004) and oligo v1.64.1 (Carvalho and Irizarry, 2010) packages to preprocess Affymetrix datasets. The affy v1.78.2 package was developed for older Affymetrix array types and is known for its robust methods for data preprocessing and normalization (Gautier *et al.*, 2004). In this study, affy v1.78.2 was used for preprocessing datasets from the Affymetrix U133 Arrays. The oligo package was developed for handling newer Affymetrix arrays and was used to preprocess datasets from Gene 1.0 ST Arrays. Both packages were used to load raw .cel files into R as AffyBatch ExpressionSet objects, providing a unified structure for subsequent data analysis. This step was essential as the raw .cel files store the experiment metadata, as well as the fluorescent intensity and xy coordinates for each probe well, in a binary format. The AffyBatch ExpressionSet structure contains both raw (binary) and processed (human readable) data along with the relevant metadata.

Limma v3.56.2 was developed primarily for the analysis of gene expression data, and has in-built tools for preprocessing datasets from Unigene cDNA arrays (Ritchie *et al.*, 2015), which consist of gpr files. Like .cel files, .gpr files contain raw intensity data. However, this data originates from dual-dye microarray scans, and includes spot fluorescent intensity values, local background estimates, and quality measures. In this study, limma was used to load raw .gpr files (also in a binary format) into R as ExpressionSet objects. This data structure resembles that of the AffyBatch ExpressionSet objects and allowed for the convergence of preprocessing outputs in the microarray pipeline for further analysis.

Once the datasets were loaded into ExpressionSet objects, they needed to be processed further. The fluorescence values at this stage were tied to xy coordinates, reflecting each chip's multiple wells for the same probe and multiple probes for the same gene, which is a method used to reduce variance and enhance reliability. Furthermore, these fluorescence values varied across different samples within the same dataset, thus requiring normalization. Robust Multi-array Average (RMA) normalization is included in both the affy v1.78.2 (Gautier *et al.*, 2004) and oligo v1.64.1 (Carvalho and Irizarry, 2010) packages as it is specifically designed for preprocessing Affymetrix microarrays. RMA normalization involves several steps: first, background correction is applied to adjust probe intensities, minimizing background noise. This is followed by quantile normalization to ensure consistent distribution of probe intensities across all arrays. The data is then log-transformed to stabilize variance. This is followed by probe summarization, where multiple probes mapping to the same gene are combined into a single expression value. This process effectively standardizes the data, making it suitable for comparative analysis across different samples and conditions (Irizarry, 2003).

The preprocessing of expression sets derived from .gpr files followed a similar approach to that used for the Affymetrix datasets. After converting the .gpr raw files into ExpressionSet objects, several key preprocessing steps were taken to ensure data integrity. For the Unigene 3.1 two-dye microarray dataset, normalization was performed using the Locally Weighted Scatterplot Smoothing (LOWESS) method instead of RMA. LOWESS normalization was designed to correct for dye-bias and intensity-dependent variations, common issues in dual-dye arrays (Smyth and Speed, 2003; Yang, 2002). This method involves a local regression approach to adjust the log ratios based on the overall intensity. Essentially, it adjusts the balance between two dyes used in the experiment by looking at the overall brightness of each spot on the microarray. This helps to correct any inconsistencies or biases that might appear across the entire microarray, ensuring that the data is more accurate and reliable (Yang, 2002). The

application of RMA and LOWESS normalization in their respective datasets were crucial for preparing the data for further analysis, however, they do prevent making comparisons across different datasets.

2.2.4 Exploratory data analysis and quality control

In this study, heatmaps and Principal Component Analysis (PCA) were used as tools for exploratory data analysis (EDA) and quality control (QC). Correlation heatmaps were generated using the pheatmap function in R (Figure A.1). This method involved first condensing the RMA/LOWESS-normalized expression data into a correlation matrix (corMatrix). The cor() function was used to calculate pairwise Pearson correlation coefficients between gene expression levels across samples. The resultant correlation matrix was then visualized using pheatmap (Kolde, 2019). The heatmaps were used to identify patterns and relationships in the gene expression data, providing a visual representation of the correlations across treatment groups and samples. These heatmaps included clustering as part of their visualization. The pheatmap package uses hierarchical clustering for grouping similar patterns (Eisen *et al.*, 1998). This type of clustering works by iteratively merging samples based on their expression similarity, using Pearson correlation. By linking samples within groups according to similarities in their expression profiles, hierarchical clustering in heatmaps provides insights into the relationships between samples, highlighting outliers or well separated treatment groups.

PCA was used both to support and extend on the insights generated from the correlation heatmaps. PCA reduces the dimensionality of the data while preserving as much variance as possible. This method can be used for the visualization of relationships within and between data sub-structures, allowing for the identification of patterns and outliers (Raychaudhuri *et al.*, 1999). In the context of genome wide expression data especially, PCA is a powerful tool for reducing the complexity of thousands of genes into an easy to interpret visualisation. For the creation of PCA plots (Figure A.2), the 'prcomp' function in R v4.3.0 was used, which supplemented the heatmap findings and assisted in the identification of outliers (R Core Team, 2023). 'prcomp' was applied to both the raw and normalised expression set objects, creating a prcomp object containing both the results of the PCA (singlular value decomposition applied to a centred matrix of expression values) as well as the proportion of variance explained by each principal component (PC). These were used to generate PCA plots comparing major PCs as well as bar graphs showing the percentage variance explained by each PC. Both PCA and

heatmaps were integral in assessing the quality of the datasets, ensuring their suitability for further analysis. These methods aided in detecting batch effects, technical artifacts, and outliers.

Several other QC tools were used to ensure data integrity. These included RNA degradation plots (Figure A.4), MA plots (Figure A.5), boxplots of fluorescent intensity (Figure A.6), and image plots (Figure A.7) for each microarray chip. RNA degradation plots were used to assess the quality of RNA samples, highlighting samples where RNA integrity may have been compromised (Fleige *et al.*, 2006). These plots were only available for U133 arrays as they make use of the AffyRNAdeg() function from affy v1.78.2. MA plots were used to visualize the relationship between the intensity (M) and average expression levels (A) for every pairwise comparison of samples. This approach was effective in identifying intensity-dependent biases and artifacts in the data (Dudoit *et al.*, 2002). To draw these plots, the MAplot() function from limma v3.56.2 was applied to the ExpressionSet objects.

Boxplots of fluorescent intensity were then used to investigate the distribution of signal intensities between samples in each dataset before and after normalization. This was used to evaluate the overall quality of microarray datasets, as it highlighted the presence of any skewness or anomalous samples in the intensity distribution (Bolstad *et al.*, 2003). For this study, the boxplot() function from base R v4.3.0 was applied to the raw and normalized fluorescent values in the ExpressionSet objects. Additionally, image plots for each microarray chip were used to visually inspect the arrays for spatial artifacts such as scratches, bubbles, or uneven hybridization (Figure A.7), which are common issues in microarray experiments (Quackenbush, 2002). These image plots for single dye and dual dye microarrays were generated using the image() function from affy v1.78.2, which uses the fluorescence values and well coordinates to reconstruct images of the microarray chips for each sample. A summary of all parameters used for each EDA plot is provided in Table A.1.

2.2.5 Differential gene expression analysis

Differential Gene Expression (DGE) analysis was conducted using limma v3.56.2 (Ritchie *et al.*, 2015), a tool designed for the analysis of gene expression microarray data. Initially, it fits a linear model for each gene, using the experimental design of each dataset to contrast between two groups. For example, GSE1869 had three treatment conditions, ischemic cardiomyopathy, non-ischemic cardiomyopathy, and the control. When performing DGE analysis on this dataset, two contrasts were used, 'ischemic cardiomyopathy – control' and 'non-ischemic cardiomyopathy – control'. Thus, the expression levels of genes in each treated group was contrasted with the control group. An important aspect of limma is its ability to adjust the influence of outliers identified during QC in the linear modelling phase. For this study, limma's arrayWeights tools was used to penalize outlier samples. These weights/penalties (w_i) were determined by calculating the median absolute deviation (MAD_i) for each array and identifying arrays with greater spread. During this step, residuals were calculated from an initial fit of a linear model comparing the treatment groups. Residuals represent the difference between observed and fitted values, capturing the variability unexplained by the model. By examining these residuals, arrays with consistently large residuals were identified as outliers. These arrays, that have higher variability, are considered poorer quality and are assigned lower weights. Thus, when limma fits linear models between contrasts during DGE analysis, outlier samples have proportionally less impact on the model, in line with their variability (Ritchie *et al.*, 2015).

$$w_i = \frac{1}{MAD_i}$$

Post model fitting, limma employs empirical Bayes methods to moderate the standard errors of the estimated log-fold changes. This moderation, crucial for genes with limited sample sizes, stabilizes variance estimates by borrowing strength across all genes. Following this, limma computes moderated t-statistics and associated p-values for each gene. These p-values are then adjusted for multiple testing, typically using the Benjamini-Hochberg method, to control the false discovery rate, ensuring more reliable and statistically sound identification of differentially expressed genes (Ritchie *et al.*, 2015; Smyth, 2004).

Genes were considered significantly differentially expressed if their adjusted p-values fell below 0.05 and their $\log_2$ fold change in expression was greater than 0.5. This low fold change corresponds to an increase or decrease in gene expression of roughly 41%. While many DGE studies typically use changes of 1 or greater, this study is exploratory in nature, so by using a smaller fold change threshold, a wider range of potentially differentially expressed genes could be identified, albeit at the cost of more false positives. Concluding the DGE analysis, the tidyverse v2.0.0 collection of R packages was used to further process the results, preparing the data for visualization (Wickham *et al.*, 2019).

2.2.6 DGE result visualisation

Two distinct software tools were used to generate visualizations from the processed DGE results. Firstly, 'ggplot2', included as part of the tidyverse v2.0.0 collection, was used for creating volcano plots (Figure A.8). Volcano plots are scatter plots used to analyse DGE results by visualizing the number of significantly up- and downregulated genes on a single set of axes, where the negative log-transformed adjusted p-value is plotted on the y-axis, and $\log_2$ fold change is plotted on the x-axis. The volcano plots were used to investigate general trends in gene expression changes in each of the seven datasets. These plots were valuable in uncovering insights into the number of genes that were up- vs downregulated in each CVD-related dataset. Furthermore, the control genes and genes of interest were colour-coded and labelled using the ggrepel v0.9.3 package, allowing for result validation and target gene profiling.

The control genes used in this study included *GAPDH*, *NPPA* and *IL6*. *GAPDH* was selected as a negative control due to its known role as a housekeeper gene and evidence showing *GAPDH* expression levels to remain consistent in CVD contexts such as heart failure (Li *et al.*, 2017; Yang *et al.*, 2021). *NPPA* was selected as a positive control due to evidence showing it is upregulated in two CVD contexts, dilated cardiomyopathies and cardiac stress (Feng *et al.*, 2022; Man *et al.*, 2018). *IL6* was selected as a positive control for the HUVEC and MVEC cell line studies for its known associations with inflammatory pathways, where it is known to be upregulated in response to *IL6* treatment (Viemann *et al.*, 2006).

In contrast to the volcano scatter plot, which provided a visualisation focusing on general trends in differential gene expression, GraphPad Prism v10.1.0 was used to draw gene-specific scatter plots showing the fluorescence values for each sample in their respective treatment groups. These visualisations provided a more detailed look at genes that were significantly differentially expressed and allowed for the identification of outliers. By visualising each sample individually, clear distinctions could be made between high confidence and low confidence differentially expressed genes.

2.3 Validation of DGE results

To validate the pipeline's DGE results an independent dataset, GSE1400 (Stitziel *et al.*, 2004), was selected from GEO. GSE1400 investigated the differences in gene expression between the cytosolic and membrane fractions of MCF-7 cells. Given that *PXDN* is known to be expressed in MCF-7 cells, and plays its primary roles in the ECM, it was hypothesized that *PXDN* expression would be greater in the membrane fraction. The DGE results produced from the pipeline were compared to the results for the same dataset's original study.

2.4 Programs and R script

The script and `sessionInfo()` (includes all packages and version numbers) for the ChIP-seq pipeline and microarray pipeline are available at GitHub:

(https://github.com/Shiven-Naidoo/ShivenNaidoo_MScProject)

All quality control results for each dataset are also available in this repository.

3 RESULTS

3.1 ChIP-seq analysis

3.1.1 TF binding visualization using peak browser

Table 3.1 summarises metadata from a subset of ChIP-seq experiments which contained binding peaks within 10 kb of *PXDN* and/or *PXDNL's* TSS. The ChIP-seq experiments span 15 cardiovascular cell types, ranging from fibroblasts and endothelial cells to tissue samples from the ventricles and aorta. TFs were identified for most TFs in untreated cell lines, with the exception of *KLF4*, *SMARCA4*, *BRD4*, *IRF1*, and *HMGB1*.

BRD4 was observed binding near *PXDN*'s TSS only in TNF-α stimulated HUVECs. Further investigation of binding peaks within *PXDN* revealed *BRD4* binding sites for untreated HUVECs were present just beyond the 10 kb range downstream of the TSS. It should also be noted that the ChIP-Atlas database only contained the results of 5 ChIPseq experiments on *BRD4* in HUVEC cells, and only one in untreated HUVEC cells (although no binding was observed near the TSS of either gene). Similarly, the only ChIP-seq experiments for *HMGB1* were in proliferating HUVEC cells and *IRF1* was only observed under IL-1β and TNF-α stimulated conditions. ChIP-seq experiments for *SMARCA4* and *KLF4* were treated with either constitutively active form of MEK5 or GFP (Green Fluorescent Protein). In some cases, there were ChIP-seq experiments in both treated and untreated conditions. Treated cell lines reveal some unique binding behaviours for TFs of both *PXDN* and *PXDNL*. For example, *RELA* was observed binding only to *PXDN* in HAECs in IL-1β and TNF-α stimulated cell lines, despite binding *RELA* peaks being observed for both *PXDN* and *PXDNL* in untreated HAECs.

Table 3.1 Summary of all ChIP-seq experiments in cardiovascular cells

Cardiovascular Cell Type	Cell Treatment	TF	Gene Loci
Aorta and ventricles	None	*CTCF*	Both
Endothelial cells (ECs)	None	*TAL1*	*PXDN*
	100 ng/ml TSA	*TAL1*	*PXDN*
	None	*EP300*	*PXDN*
Cardiac Progenitor Cells (CPCs)	None	*EP300*	*PXDN*
Coronary Artery Endothelial cells (CAECs)	None	*STAG1*	Both
	None	*STAG2*	*PXDN*
	None	*SMC1A*	*PXDN*
	None	*CTCF*	Both
Human Aortic Endothelial cells (HAECs)	IL-1 β	*RELA*	Both
	TNF-α	*RELA*	Both
	None	*RELA*	*PXDN*
	None	*EP300*	*PXDN*
	None	*ERG*	*PXDN*
	IL-1 β	*IRF1*	*PXDNL*
	TNF-α	*IRF1*	*PXDNL*
	None	*NFE2L2*	*Both*
	None	*JUNB*	*PXDN*
Human Umbilical Vein Endothelial Cells (HUVECs)	None	*RELA*	Both
	Scr siRNA	*RELA*	*PXDN*
	Scr siRNA and IL-1 β	*RELA*	*PXDN*
	Notch 1 siRNA and IL-I β	*RELA*	*PXDN*
	None	*EP300*	Both
	VEGF	*EP300*	*PXDNL*
	None	*CTCF*	Both
	None	*RAD21*	Both
	Proliferating	*HMGB1*	*PXDN*
	None	*NOTCH1*	*PXDN*
	None	*TEAD1*	*PXDN*
	None	*FLI1*	*PXDN*
	None	*RBPJ*	*PXDN*
	None	*ERG*	*PXDN*
	None	*MAX*	*PXDN*
	TNF-α	*BRD4*	*PXDN*
	None	*FOS*	*PXDN*
	Progesterone	*PGR*	Both
Primary Human Cardiac Myocytes (HCMs)	None	*CTCF*	Both
	ad.caMEK5 transduced	*SMARCA4*	Both
	ad.GFP transduced	*SMARCA4*	*PXDN*
Pulmonary Artery Endothelial Cells (PAECs)	ad.caMEK5 transduced	*KLF4*	*PXDN*
Human Aortic Adventitial Fibroblasts (HAoAFs)	None	*CTCF*	Both
Human Brain Microvascular Endothelial cells (HBMECs)	None	*CTCF*	Both
Human Cardiac Fibroblasts (HCFaas)	None	*CTCF*	Both
Human Coronary Artery Smooth Muscle cells (HCASMCs)	None	*JUND*	*PXDN*
	None	*JUN*	*PXDN*
	None	*TCF21*	*PXDN*
Human Pulmonary Artery Fibroblasts (HPAF)	None	*CTCF*	Both
Tibial Arteries	None	*CTCF*	Both

PXDN and *PXDNL* shared many TFs, including *EP300*, *STAG1*, *RELA*, *CTCF* and *RAD21*. However, for this 10 kb range surrounding their TSSs *PXDN* appeared to have a wider variety of TFs, and significantly more binding peaks. *PXDN* also had binding peaks in a wider variety of cell types compared to *PXDNL*. Notably, most of these TF binding peaks were produced under either untreated or inflammatory conditions. Next, the distributions of binding peaks within genomic features of *PXDN* and *PXDNL* were visualised using bar plots (Figure 3.1).

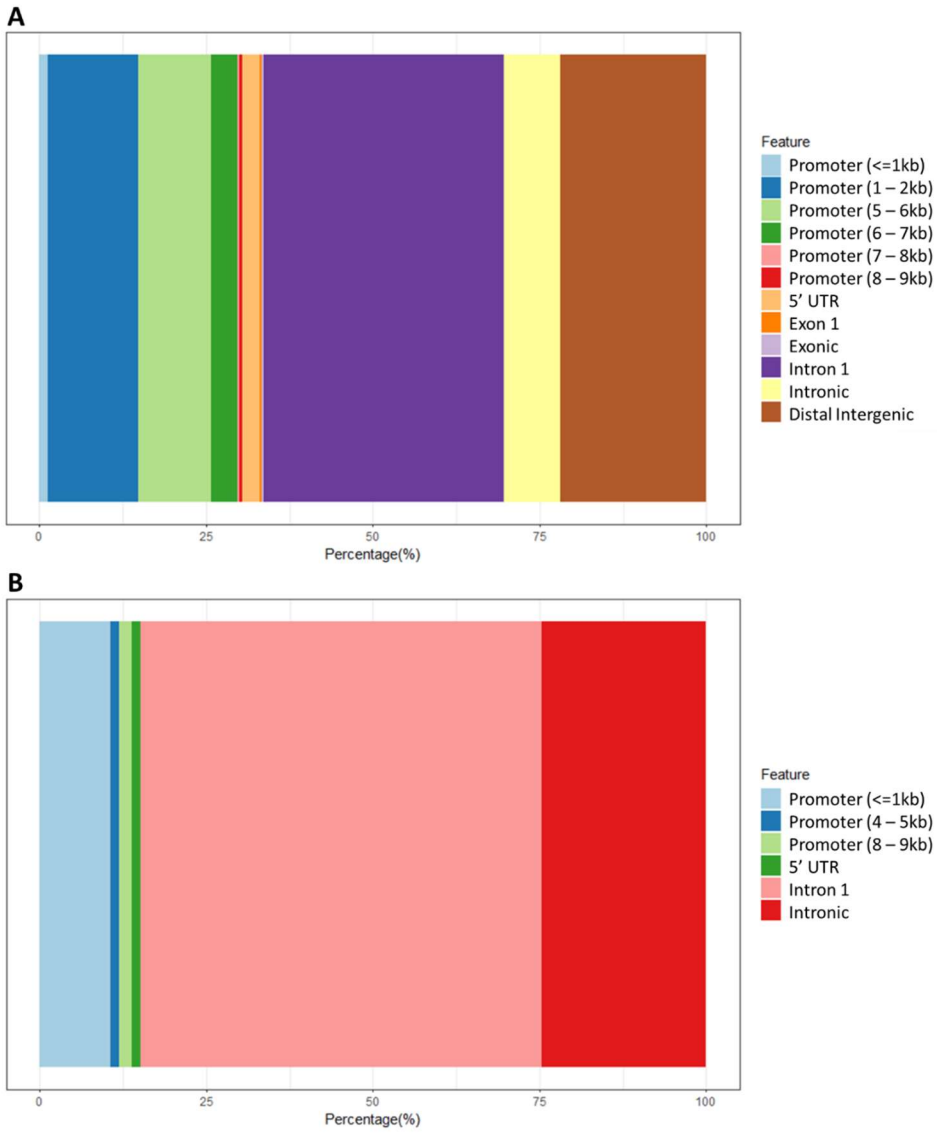


Figure 3.1 Distribution of TF binding peaks identified in genomic features within a 10kb range of (A) *PXDN* and (B) *PXDNL*'s TSS. TF binding peak data was filtered to within 10 kb of the target gene's TSS using dplyr v1.1.2 (Wickham *et al.*, 2023) and annotated in relation to genomic elements of the target genes using the GRCh38 reference genome and ChIPseeker v1.36.0. These images were generated using ChIPseeker v1.36.0 (Yu *et al.*, 2015).

The majority of TF binding peaks fall within the intronic regions, with the first introns accounting for over 30% of the observed binding peaks in *PXDN* and *PXDNL*. A large proportion of TFs also seem to bind to the core promoter region of both genes, accounting for over 10% of binding peaks. *PXDN* has a much higher percentage of TF binding peaks within the distal intergenic region, likely due to *PXDNL* having a much larger core promoter region and, thus, a smaller distal intergenic region within 10kb of the TSS when compared to *PXDN*. Furthermore, *PXDN* has a larger portion of TF binding within its first exon and promoter outer boundaries (beyond 2 kb) compared to *PXDNL*.

3.1.2 Identifying regulatory TFs for *PXDN* and *PXDNL*

High confidence TFs observed binding to *PXDN* and *PXDNL's* core promoters and first introns were identified across a wide range of cell/tissue types. As shown in Figure 3.2, *PXDN*'s first intron had binding peaks for 22 TFs while *PXDNL* had 17 TFs. Despite having more TFs, binding peaks were identified across fewer cell types in *PXDN* (12 cell/tissue types) compared to *PXDNL* (19 cell/tissue types). Both genes shared many TFs in similar cell/tissue types. For example, several studies contained high confidence *RELA* binding sites in human aortic endothelial cells (HAECs) and human umbilical vein endothelial cells (HUVECs) for both genes. Interestingly, there appear to be exceptions such as *CTCF*, which had far fewer binding peaks across fewer cell/tissue types in *PXDN* compared to *PXDNL*.

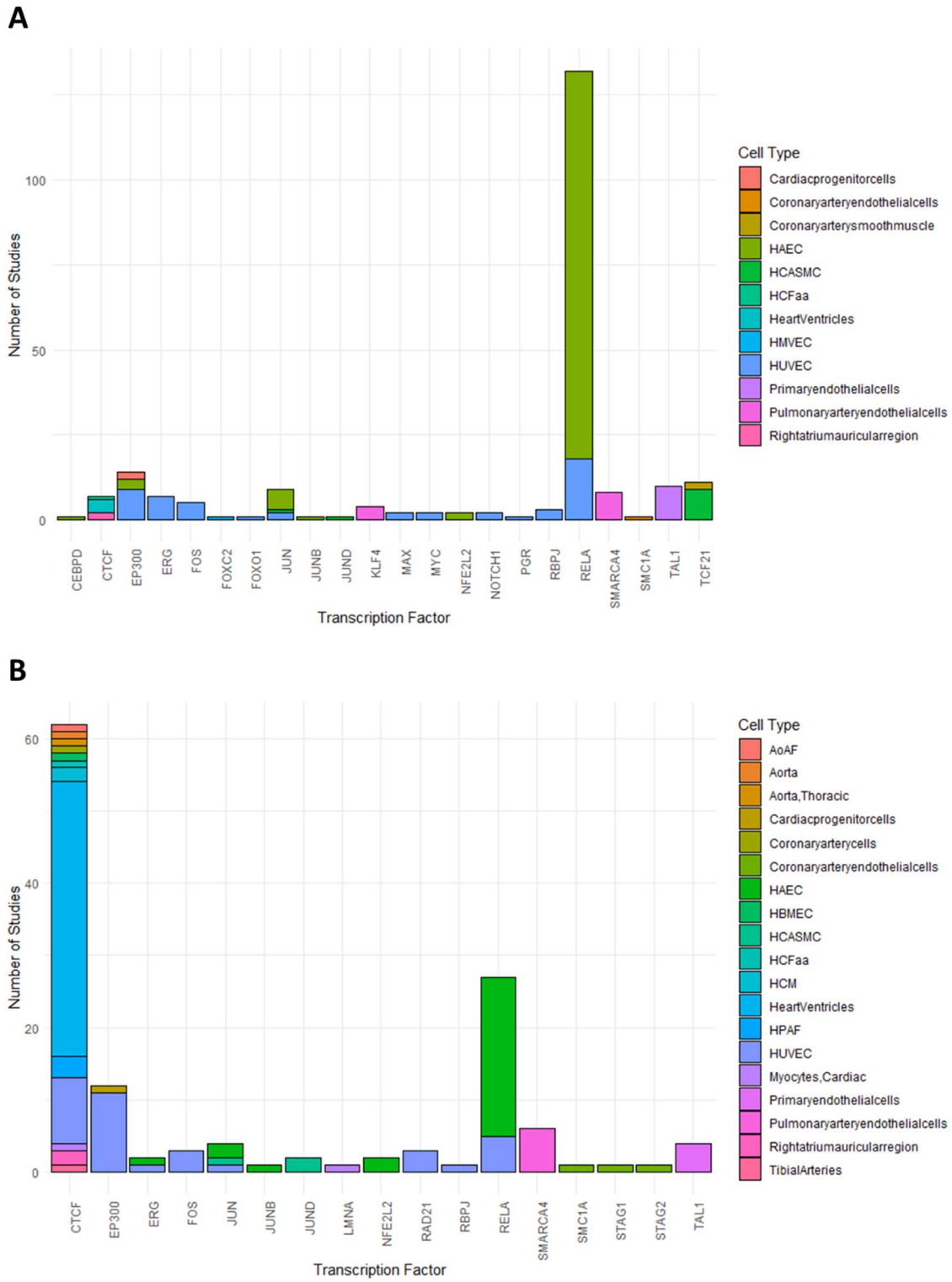


Figure 3.2 High confidence TF binding peaks located within the first introns of (A) *PXDN* and (B) *PXDNL* in different cardiovascular cells/tissues. tidyverse v2.0.0 was used to filter annotated .bed files for both genes to include only TF binding peaks within the first intron. Each bar on the x axis represents a unique TF while the y axis represents the number of unique studies where a TF binding peak was observed. The stacks represent the cell type for each study. This image was generated using ggplot v3.4.4 (Wickham, 2011).

Figure 3.3 shows high confidence binding peaks within the promoters of *PXDN* (Figure 3.3 A) and *PXDNL* (Figure 3.3 B) respectively. Once again, *PXDN* (15 TFs) is seen to have a much wider variety of TFs compared to *PXDNL* (4 TFs). Many of the TFs identified in the intronic regions recur in the promoter for *PXDN* and *PXDNL*. However, unlike the first intron, *PXDN*'s binding peaks were also identified in a wider diversity of cell/tissue types compared to *PXDNL*. Interestingly, both the cell/tissue types and TFs identified in *PXDN*'s promoter align very closely with those seen in *PXDNL*'s first intron. In contrast, *PXDNL*'s promoter had a much smaller range of cell/tissue types (11).

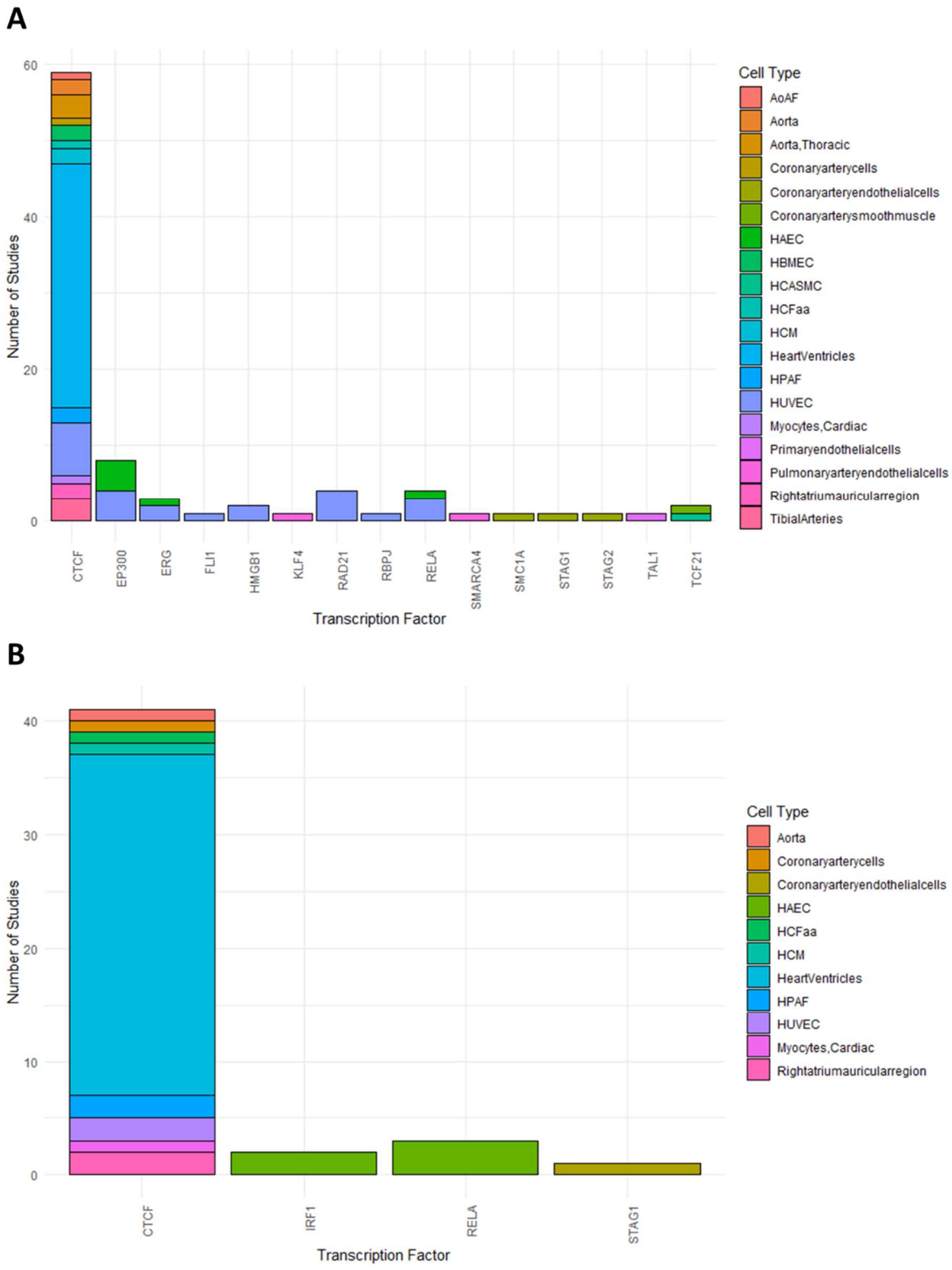


Figure 3.3 High confidence TF binding peaks located within the promoter of (A) *PXDN* and (B) *PXDNL* in different cardiovascular cells/tissues. Replicates the analysis presented in Figure 3.2, but .bed files were filtered to include TF binding peaks in the core promoter regions of each gene. This image was generated using ggplot v3.4.4 (Wickham, 2011).

In total, 29 high confidence TFs were observed binding to *PXDN* and/or *PXDNL*. *PXDN* and *PXDNL* were found to share 16 TFs binding either the promoter or first intron (Figure 3.4 A). These TFs include *RELA*, *SMC1A*, *SMARCA4*, *EP300*, *TAL1*, *JUN*, *ERG*, *RBPJ*, *CTCF*, *FOS*, *NFE2L2*, *JUNB*, *JUND*, *STAG2*, *STAG1*, and *RAD21*. While *PXDN* and *PXDNL* share many TFs, they also have unique TFs. As shown in Figure 3.4 A, *PXDN* has 11 unique TFs, including *TCF21*, *KLF4*, *PGR*, *NOTCH1*, *MAX*, *MYC*, *FOXO1*, *CEBPD*, *FOXC2*, *FLI1*, and *HMGB1* while *PXDNL* only has two unique TFs, *IRF1* and *LMNA*. Notably, the first intron has a larger variety of TFs in both *PXDN* (Figure 3.4 B) and *PXDNL* (Figure 3.4 C) when compared to the promoter.

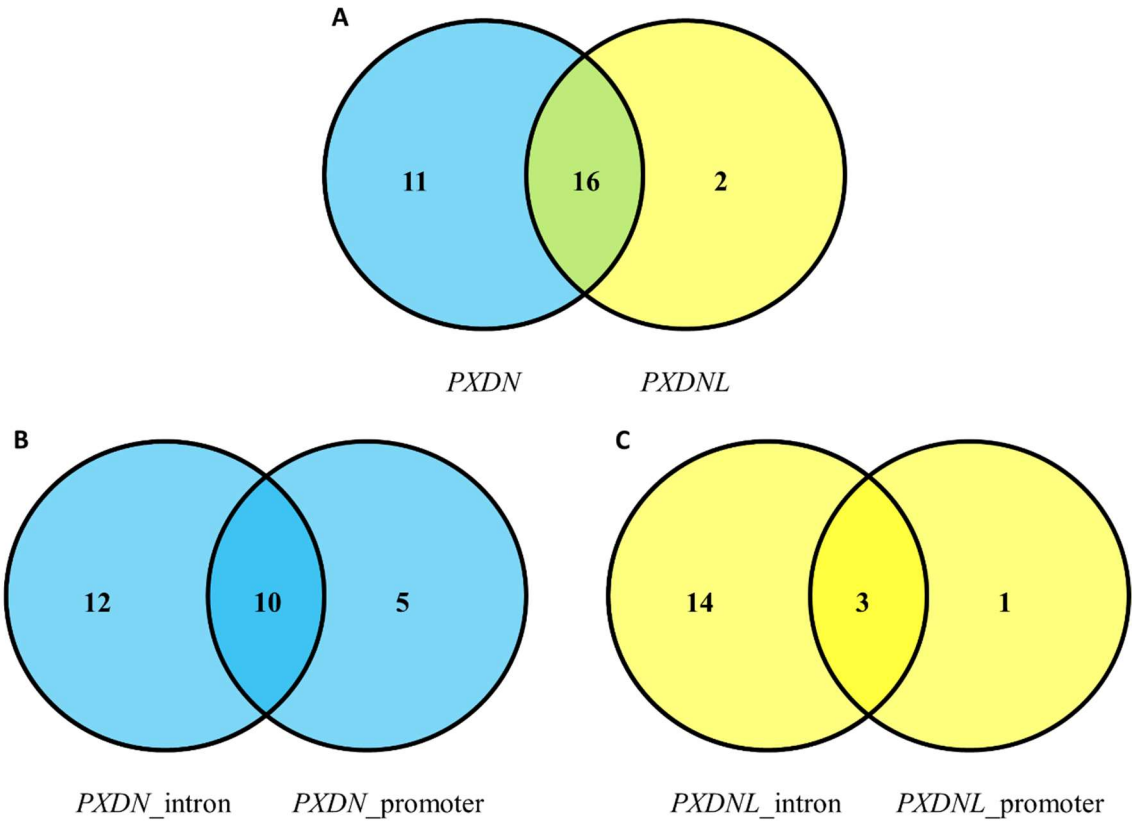


Figure 3.4 Number of shared and unique TFs binding to the promoter and first introns of *PXDN* and *PXDNL*. Unique TFs identified for each gene and genomic region were compiled into lists and compared using base R v4.3.0 (R Core Team, 2023) and tidyverse v2.0.0 (Wickham *et al.*, 2019) (A) number of unique TFs identified in *PXDN* vs *PXDNL*, (B) *PXDN*'s first intron vs promoter and, (C) *PXDNL*'s first intron vs promoter.

3.2 Gene expression microarray analysis

3.2.1 Quality control and preprocessing

Before analysing the expression profiles of *PXDN*, *PXDNL* and their 29 high confidence TFs, quality control (QC) tools were used to evaluate the integrity and reliability of the seven selected microarray datasets. For this section, only the QC results from two datasets will be shown. First is the tissue dataset GSE26887 (Affymetrix Human Gene 1.0 ST Array) and second is the HUVEC cell line dataset GSE2638 (Affymetrix Human Genome U133 Plus 2.0 Array). These datasets were chosen as they showcase the key QC tools used, allow for the visual detection of outliers, and highlight the challenges unique to tissue vs cell line datasets. The full QC results for all datasets, including additional visualisations such as RNA degradation plots, MVA plots and microarray images, are available on github (refer to methods). First, heatmaps and hierarchical clustering were used to investigate the relationships between samples and treatment groups for each dataset (Figure 3.5).

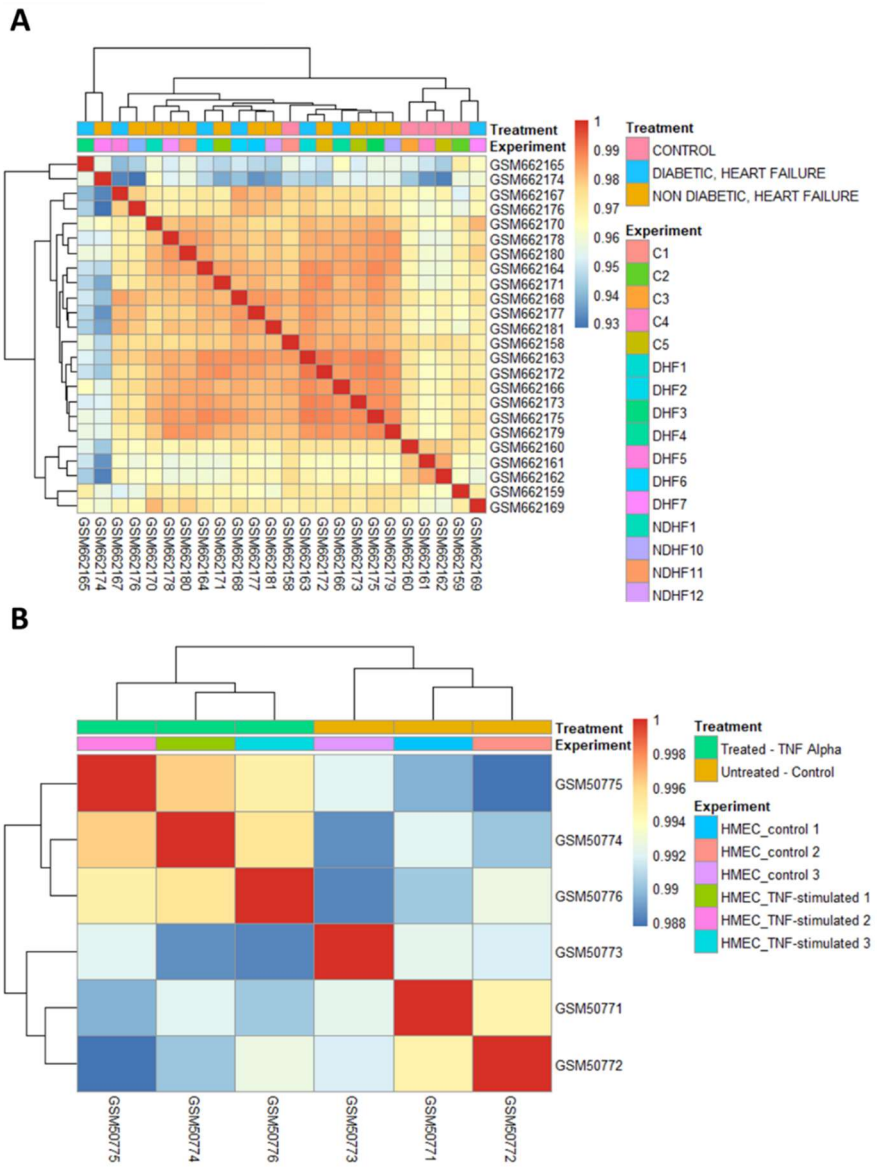


Figure 3.5 Heatmaps and hierarchical clustering. Show pairwise correlations between samples based on expression data from all probes in each sample. (A) highlights relationships between 24 samples across three treatment groups for the GSE26887 tissue dataset. (B) highlights relationships between six samples across two treatment groups for the GSE2638 cell line dataset. These images were generated using ggplot v3.4.4 (Wickham, 2011).

As shown in Figure 3.5 A, the majority of control samples cluster separately from both treatment groups, with the exception of C1. However, both treatment groups, diabetic and non-diabetic heart failure (HF), tend to cluster closely together. Figure 3.5 B shows that GSE2638's samples clustered according to their treatment group, with no apparent outliers. The cell line dataset here exhibits a clearer separation of treatment groups, while treatment groups are harder to distinguish in the tissue dataset. Additionally, the tissue dataset shows potential outliers

whereas no clear outliers are seen in the cell line dataset. When applied to all datasets, similar patterns were seen when comparing cell line and tissue data, with clearer treatment separation seen in the cell line data and some potential outliers detected in the tissue data. The datasets were then preprocessed using Robust Multi-array Average (RMA) and Locally Weighted Scatterplot Smoothing (LOWESS) for single-dye and dual-dye microarrays, respectively. Boxplots of fluorescence value distributions were used to compare the raw and preprocessed datasets (Figure 3.6).

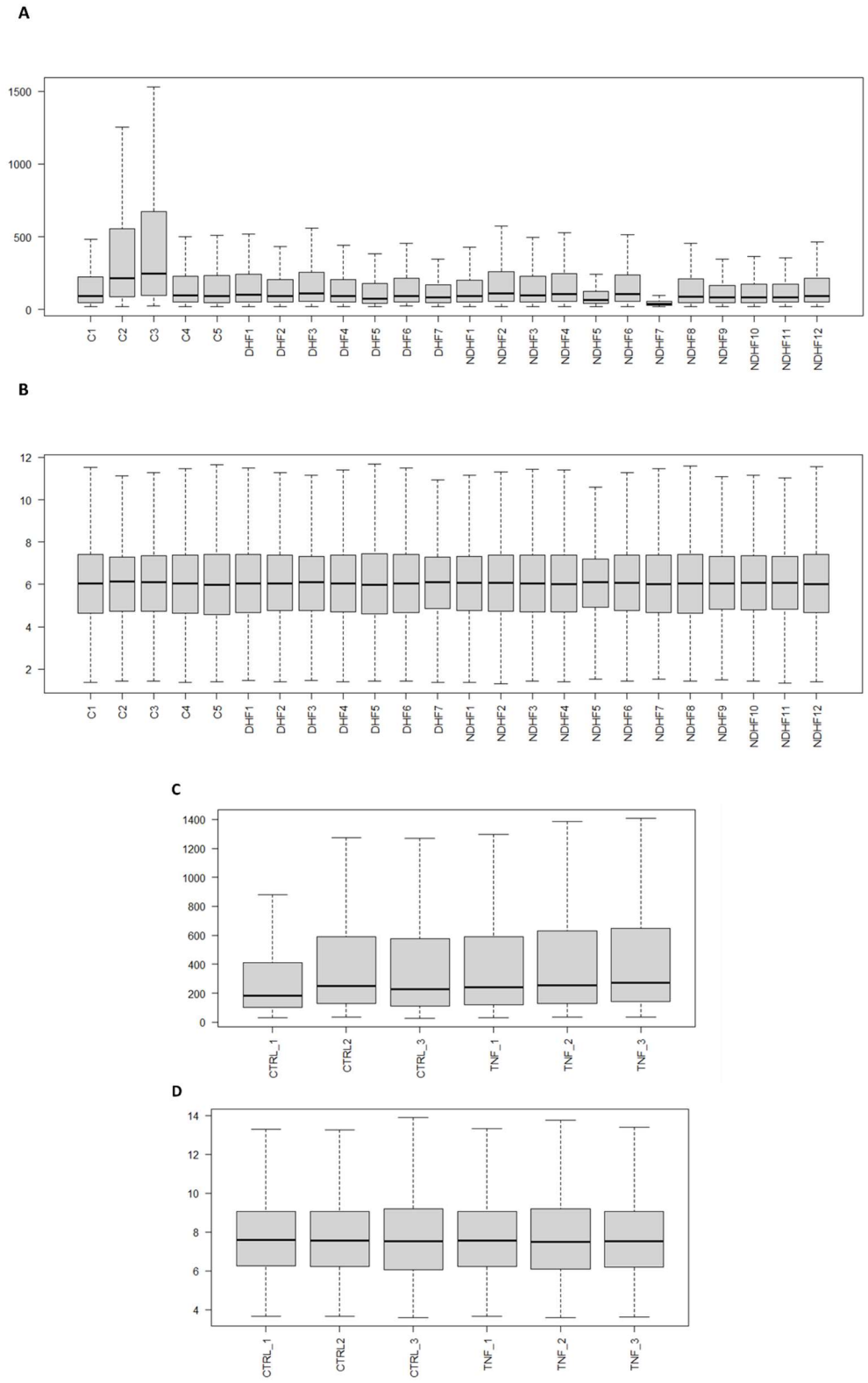


Figure 3.6 Distribution of fluorescence values for raw and preprocessed microarray datasets. Panels A and B represent the raw and normalized fluorescence values for all samples in GSE26887, respectively. Panels C and D represent the raw and RMA normalized fluorescence values for all samples in GSE2638 respectively. These images were generated using ggplot v3.4.4 (Wickham, 2011).

Before preprocessing, datasets were screened for samples with abnormally high or low distributions of fluorescence values. Figure 3.6 A shows four abnormal distributions in GSE26887, wherein two control samples exhibited higher than average fluorescence values (C2 and C3), and two non-diabetic heart failure samples exhibited lower than average fluorescence values (NDHF5 and NDHF7). These abnormal distributions may arise from technical variations between samples. For samples to be compared fairly in later analyses, they needed to be appropriately scaled. Figure 3.6 B shows that RMA preprocessing successfully normalized the fluorescence value distributions across all samples in GSE26887.

Similarly, abnormalities were seen in GSE2638. Figure 3.6 C shows that the first control sample (CTRL_1) had an abnormally low distribution of fluorescence values. However, all fluorescence values were normally distributed after RMA preprocessing (Figure 3.6 D), indicating that the samples were successfully scaled. Notably, there was much more variation between tissue sample fluorescence distributions compared to the cell line datasets, suggesting that tissue sampling differences and biological variation may also be affecting fluorescence distributions. As similar results were obtained for all datasets, preprocessing was considered successful in normalizing fluorescence values. Next, Principal Component Analysis (PCA) was used to investigate the impact of normalization on samples within treatment groups and identify potential outliers. Using ggplot v3.4.4 (Wickham, 2011), PCA plots were drawn for the raw and normalized data (Figures 3.7 and 3.8).

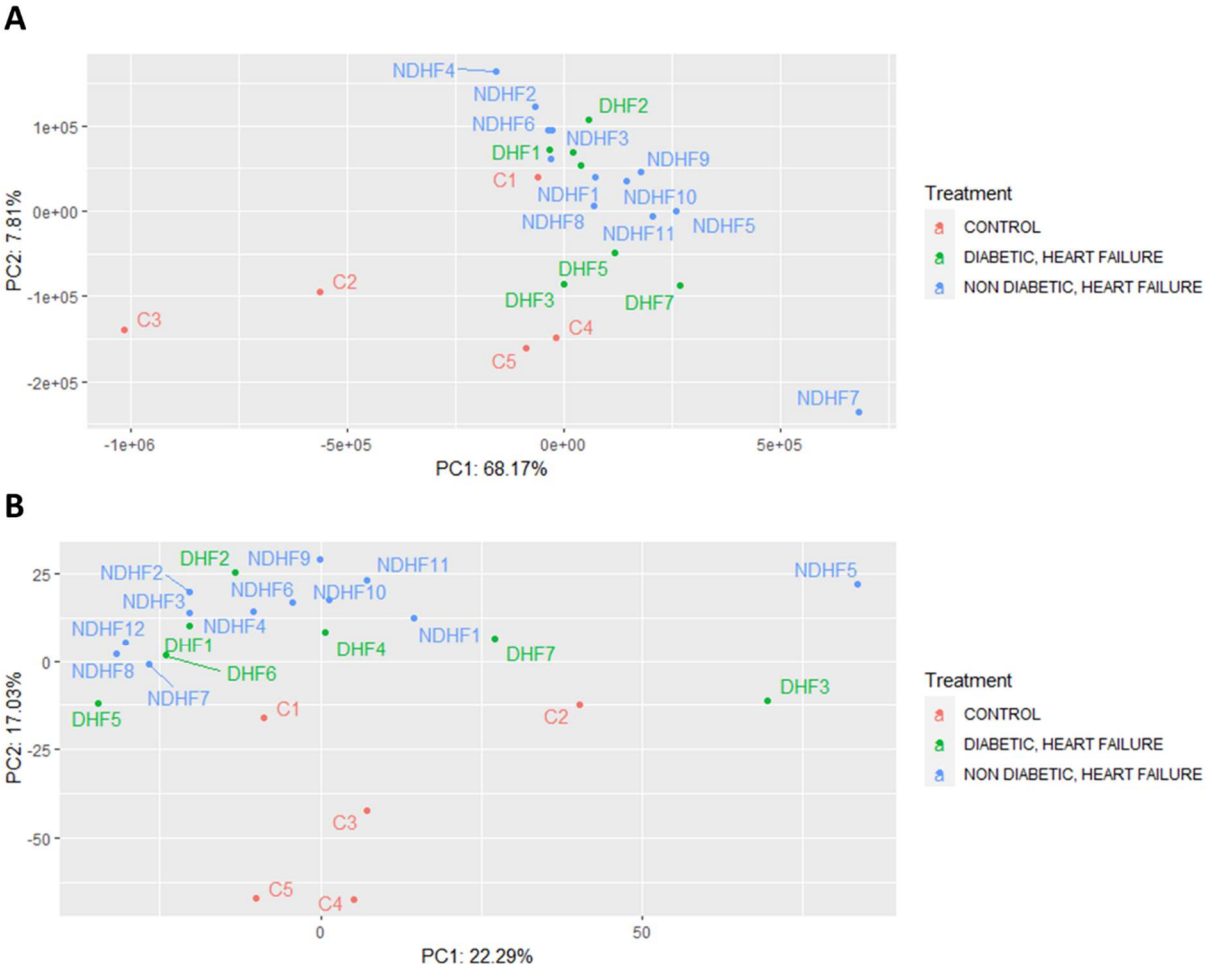


Figure 3.7 Principal Component Analysis (PCA) plot showing sample clustering along PC1 and PC2 for GSE26887's (A) raw expression data (B) RMA normalized expression data. Each dot represents a sample, and each color represents a treatment group. These images were generated using ggplot v3.4.4 (Wickham, 2011).

PCA results using raw expression values from GSE26887 (Figure 3.7 A) show no clear clustering of treatment groups along PC1. Furthermore, previously identified outliers (C2, C3, NDHF5, and NDHF7) are spread uniformly across PC1, suggesting that they may account for most of the variance observed. After RMA normalization (Figure 3.7 B), there appears to be clearer separation of the control group from both disease groups, with four of five controls clustering away from disease groups along PC2, with the exception of C1. Notably, there was still no clear separation between disease groups. PC1 only captured 22% of the variance from this dataset after normalization, suggesting significant heterogeneity within treatment groups. While this could be due to poor data quality, it could also reflect biological variation inherent in the disease state and/or the patients from which the samples were derived. Given that the

control group did show separation from both treatment groups along PC2, which captured 17% of the variance, the dataset was retained for further investigation. PCA of GSE2868 showed similar results, with outliers from the raw data clustering away from their treatment groups (Figure 3.8 A) while RMA normalization improved the clustering of treatment groups (Figure 3.8 B). GSE2638 and the other cell line datasets showed significantly clearer separation of treatment groups compared to the tissue datasets.

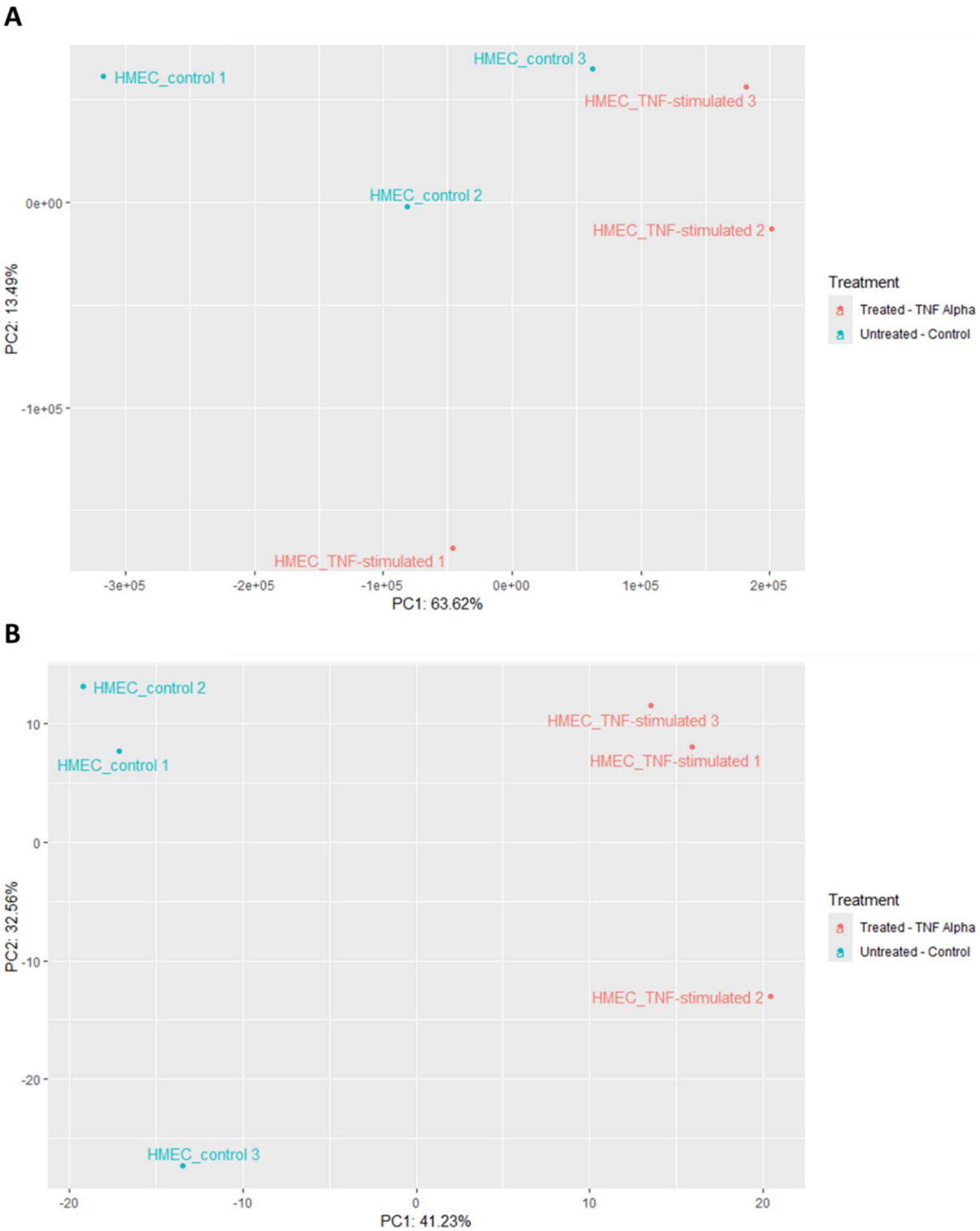


Figure 3.8 PCA plot showing sample clustering along PC1 and PC2 for GSE2638 (A) raw expression data and (B) RMA normalized expression data. Replicates the analysis shown in Figure 3.7 for GSE2638.

All tissue datasets followed similar trends where the variance explained by PC1 and PC2 were more dispersed, and the separation of treatment and control groups was less clear compared to the cell line datasets. Additionally, outliers (such as C2, C3, DHF3, and NDHF5) were identified and subsequently penalized using arrayWeights. Conversely, the cell line datasets were consistently clearly separated according to their treatment groups with few outliers. All datasets showed significant enough separation between treatment and control groups to warrant further investigation. Additionally, while outliers were identified in most of the tissue datasets, the arrayWeights tool from limma v3.56.2 was used to penalize outlier samples instead of removing them altogether, thereby retaining as much information as possible without compromising the integrity of these datasets. It should be noted that GSE42955 did show an overlap of treatment groups, but its outlier samples were either removed or heavily penalized, and it was retained for further analysis.

3.2.2 DGE analysis

Volcano plots were used to visualize significantly differentially expressed genes for each dataset. In addition to clearly showing the number of significantly upregulated and downregulated genes, each probe for the genes of interest was labelled. Furthermore, each plot was labelled with a positive and negative control. *GAPDH* was used as a negative control for all datasets, while the positive controls for tissues and cell lines were *NPPA* and *IL6* respectively. The figures below contain the volcano plots for the GSE26887 (Figure 3.9 A and 3.9 B) and GSE2638 (Figure 3.9 C) datasets.

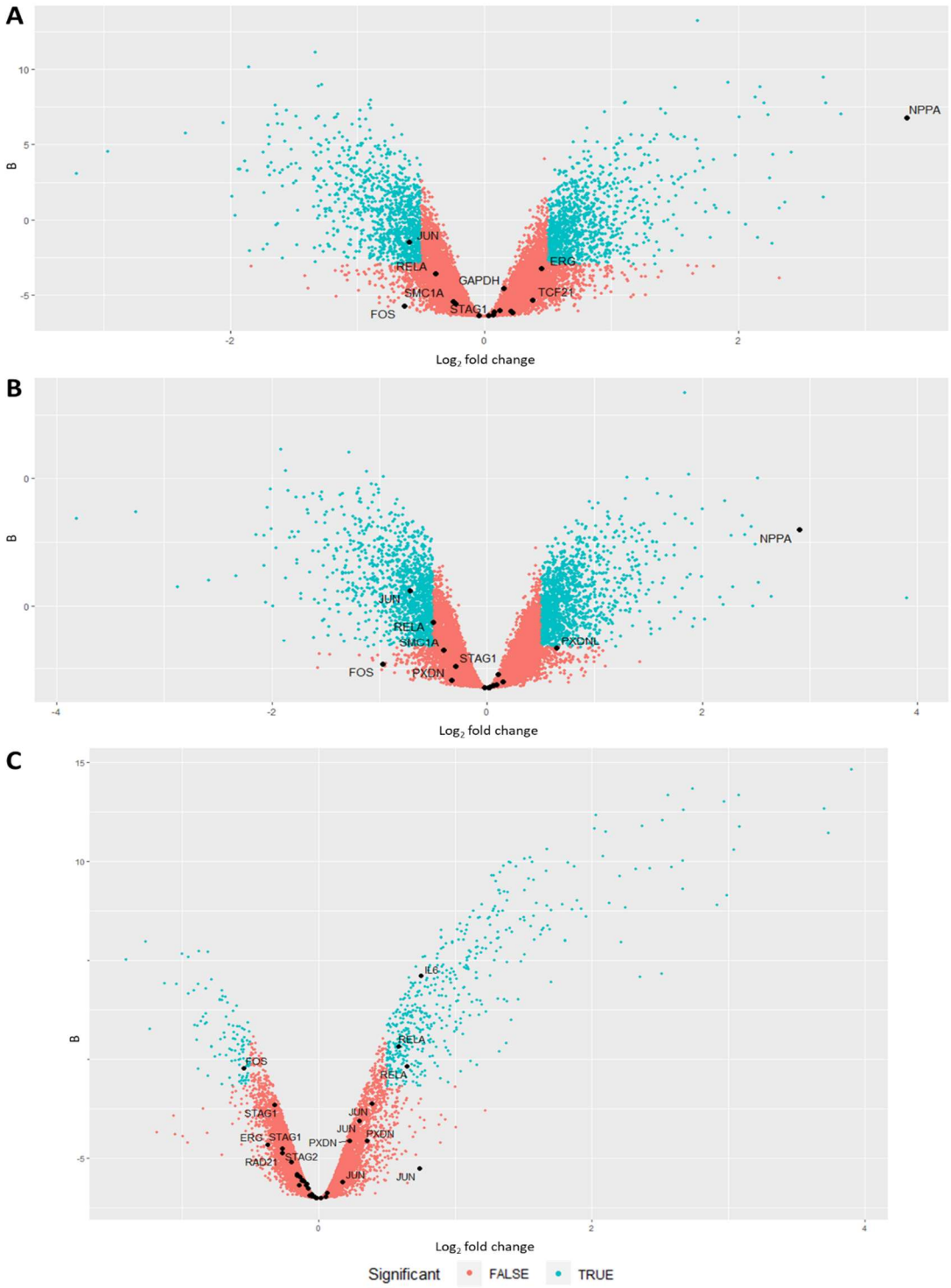
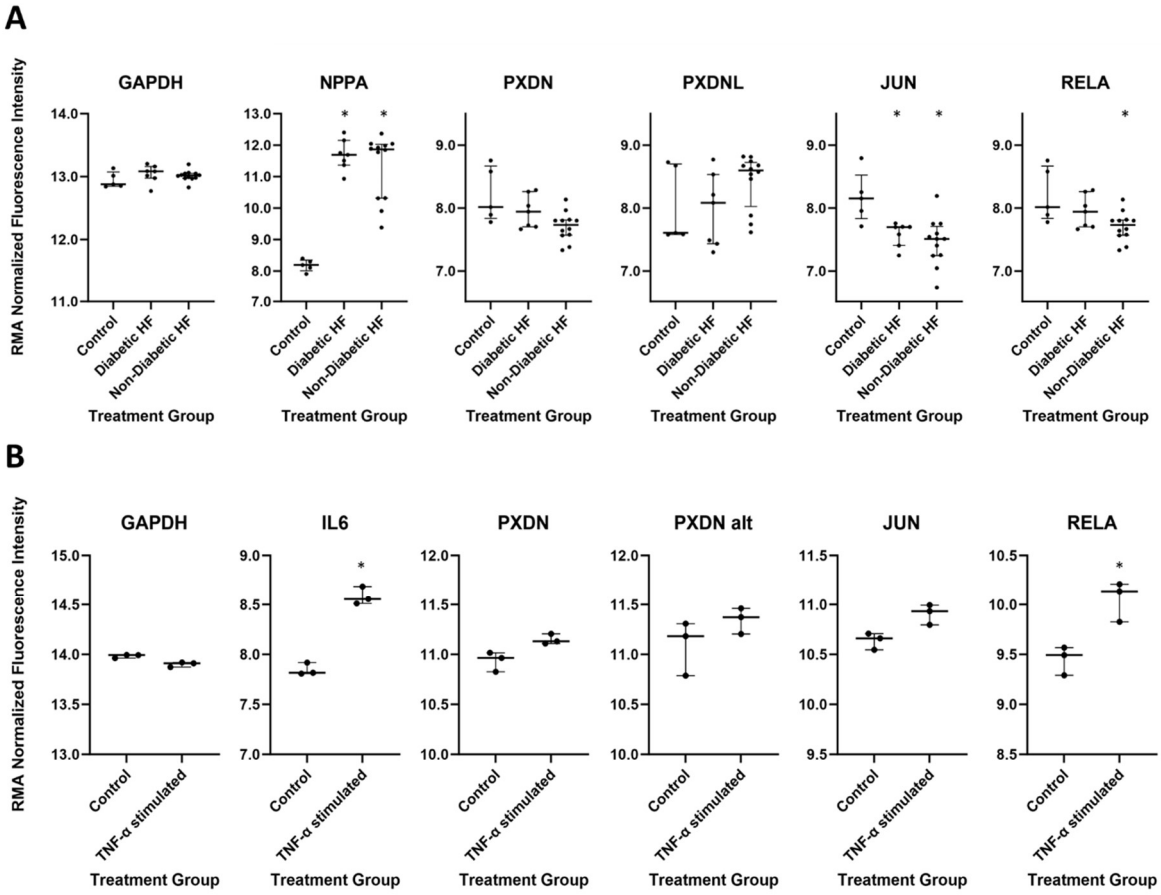


Figure 3.9 Volcano plots showing the distribution of significantly differentially expressed genes based on $\log_2$ fold changes in expression (x-axis) and the negative $\log_{10}$(adjusted p-value) (y-axis). Distributions of differentially expressed genes in GSE26887 when comparing the control samples to either (A) diabetic or (B) non-diabetic heart failure conditions. (C) differentially expressed genes in GSE2638 when comparing the control samples to TNF-α stimulated cells.

The volcano plots in Figure 3.9 revealed useful insights into general trends in differential gene expression between sample groups. Firstly, both positive and negative control genes behaved as expected for both datasets, with *NPPA* and *IL6* being upregulated in tissue and cell lines, respectively, and *GAPDH* expression remaining constant across all conditions. Secondly, a large number of genes were both significantly upregulated and downregulated under both diabetic and non-diabetic heart failure conditions in GSE26887. In contrast, fewer genes were significantly differentially expressed in GSE2638, with a larger proportion of genes being upregulated. Furthermore, the tissue datasets showed a wider range of fold changes in gene expression compared to the cell line dataset. Notably, the only dataset where the positive control, *NPPA*, was not differentially expressed was GSE42955.

The differential gene expression of *PXDN*, *PXDNL* and their TFs were also explored in Figure 3.10. For both diabetic (Figure 3.9 A) and non-diabetic (Figure 3.9 B) heart failure in GSE26887, neither *PXDN* nor *PXDNL* were significantly differentially expressed. The TFs *JUN* and *RELA* were exceptional cases, with both TFs being downregulated in non-diabetic heart failure, and only *JUN* being downregulated in diabetic heart failure. Similarly, *PXDN* was not differentially expressed in GSE2638 (Figure 3.9 C) (the dataset lacked probes for *PXDNL*). Furthermore, two TFs were shown to be differentially expressed under TNF-α stimulation, two probes for *RELA* were significantly upregulated and one probe for *FOS* was significantly downregulated. Notably, genes of interest that did show differential changes in gene expression had relatively low $\log_2$ fold changes (< 1). Next, scatter plots were used to investigate gene expression changes more closely (Figure 3.10).



** Significantly differentially expressed compared to control group, p-adj < 0.05*

Figure 3.10 RMA normalized fluorescence values. (A) GSE26887, control and gene of interest fluorescence values for 24 samples across three treatments: Control, Diabetic and Non-diabetic (B) GSE2638, control and gene of interest fluorescence values for six samples across two treatments: Control and TNF-α treated. Generated using GraphPad Prism v10.1.0.

The scatter plots in Figure 3.10 highlight the differences in expression values for samples within the same treatment group for several genes of interest. Notably, the differences between same-treatment samples were generally much higher in the tissue datasets compared to the cell lines. Outlier samples were also easier to identify in tissue datasets. For example, in Figure 3.10 A two outlier samples can be seen giving generally higher expression values in *PXDN*, *PXDNL*, *RELA* and the control *GAPDH*. The final DGE results for all seven analysed datasets are shown above in Table 3.2 below. For the three datasets that included probes for *PXDNL*, no differential gene expression was detected, while all seven datasets showed no significant change in *PXDN* expression. However, several high confidence TFs were significantly differentially expressed in a variety of CVDs.

Table 3.2 Summary of DGE results for seven selected datasets

GEO ID	Cell/Tissue Type	Samples	Treatment/ Disease	Significant change in gene expression
GSE3585*	Heart: left ventricle	12	Dilated cardiomyopathy	None
GSE3586*	Heart: septum	28	Dilated cardiomyopathy	*HMGB1* ↓
GSE42955	Heart: left ventricle apex	17	Dilated cardiomyopathy	None
			Ischemic cardiomyopathy	None
GSE26887	Heart: left ventricle	24	Diabetic heart failure	*JUN* ↓ *JUNB* ↓ *CEBPD* ↓ *MYC* ↓
			Non-Diabetic heart failure	*JUN* ↓ *JUNB* ↓ *CEBPD* ↓ *MYC* ↓ *RELA* ↓
GSE76701	Heart: left ventricle	8	End-stage heart failure	None
GSE2638*	Microvascular endothelial cells	6	2ng/ml TNF-α for 5 hours	*FOXC2* ↓ *FOS* ↓ *IRF1* ↑ *JUNB* ↑ *RELA* ↑
GSE2639*	Human umbilical vein endothelial cells	8	2ng/ml TNF-α for 5 hours	*IRF1* ↑ *JUNB* ↑

**Microarray platform lacked probes for PXDNL*

Note that datasets GSE2638, GSE2639, GSE3585 and GSE3586 did not have probes for *PXDNL*. Neither *PXDN* nor *PXDNL* were differentially expressed in all datasets. Similarly, the majority of their TFs were not differentially expressed across all CVD-related conditions. None of the genes and TFs of interest were differentially expressed in the cardiomyopathy datasets, except for *HMGB1*, which was downregulated. Both diabetic and non-diabetic heart failure datasets shared similar patterns, with *JUN*, *JUNB*, *CEBPD*, and *MYC* all being significantly downregulated. Notably, *RELA* was upregulated in only the non-diabetic heart failure group. Similar patterns were also shared between the TNF-α stimulated endothelial cell lines, where *IRF1* and *JUNB* were upregulated in both datasets. Notably, the dataset in microvascular cells also saw significant changes in the expression of *FOXC2*, *FOS* and *RELA*.

4 DISCUSSION

PXDN and *PXDNL* have been shown to be expressed in the cardiovascular system and the heart, respectively (Cheng and Shi, 2022; Péterfi *et al.*, 2014). Notably, the increased expression of *PXDN* has been correlated to the progression of cardiovascular diseases (CVDs) such as peripheral artery disease (PAD) (Costas *et al.*, 2021) and cardiac fibrosis (Liu *et al.*, 2019). Additionally, Péterfi *et al.* (2014) have observed increased expression of *PXDNL* in failing hearts, where it has been shown to act as a competitive inhibitor of *PXDN*. These CVDs can all be influenced by similar pathways, with inflammation playing a major role in CVD pathogenesis (Poirier and Eckel, 2002; Yndestad *et al.*, 2006). Furthermore, the progression of these CVDs is intertwined, with conditions like cardiac fibrosis being observed in late-stage heart failure or promoting the pathogenesis of certain cardiomyopathies (Frangogiannis, 2021, 2019). Thus, the dysregulation of *PXDN* and *PXDNL* expression in cardiovascular cells may contribute to the pathogenesis of different CVDs as well. However, to date, few studies have explored the transcriptional regulation of these genes in the cardiovascular system, or changes in their expression profiles in CVD-related contexts. This study offers novel insights into the transcriptional pathways regulating *PXDN* and *PXDNL*, as well as their expression patterns in CVD-related conditions.

4.1 Insights from ChIP-seq data mining

The first aim of this study was to investigate the transcriptional regulation of *PXDN* and *PXDNL* in the cardiovascular system by identifying transcription factors (TFs) which may regulate their expression in cardiovascular cells. To bridge this gap, this study analysed roughly 400 ChIP-seq datasets in cardiovascular cells and tissues from the ChIP-Atlas database (Oki *et al.*, 2018) using ChIPseeker (Yu *et al.*, 2015). The molecular pathways of these TFs were then explored to elucidate the potential roles of *PXDN* and *PXDNL* in the cardiovascular system. It should be noted that not all ChIP-seq datasets were in untreated cells and tissues, and TFs which were only identified under treated conditions will be specified.

4.1.1 Exploration of TF binding patterns

In this study we observed large proportions of TF binding peaks enriched in the promoter and first introns of *PXDN* and *PXDNL*, suggesting that they may be key sites of transcriptional regulation. The promoter, located upstream of a gene's TSS, is a primary binding site for TFs that initiate transcription (Levine and Tjian, 2003). TFs binding to this region may regulate gene expression by recruiting RNA polymerase II and other essential transcription machinery (Smale and Kadonaga, 2003). The first intron also plays a key role in regulating gene expression. TF binding in this region can influence alternative splicing, enhancer activity, and the efficiency of transcriptional elongation (Blackwood and Kadonaga, 1998; Luco *et al.*, 2011). Intronic binding is particularly relevant in complex genomes where regulatory elements are often dispersed and not confined to the immediate promoter region (Andersson *et al.*, 2014).

TFs which were more likely to regulate the transcription of *PXDN* and/or *PXDNL*, termed high confidence TFs, were selected based on the genomic regions and threshold values of their binding peaks (refer to methods). In total, 27 high confidence TFs were identified for *PXDN*, and 18 high confidence TFs were identified for *PXDNL*. Of these high confidence TFs, *PXDN* and *PXDNL* were found to share 16 TFs. For *PXDN*, the intron and promoter shared 10 TFs, with 12 TFs unique to the first intron and 5 TFs unique to the promoter. *PXDNL* had 14 high confidence TFs in its first intron, but significantly fewer TFs were identified binding to its promoter, with three TFs binding in both the promoter and first intron and only one TF unique to the promoter. This was surprising given that *PXDNL*'s core promoter region (2200 bp) is much larger compared to *PXDN*'s (1401 bp) (Ensembl, 2023b, 2023a; Martin *et al.*, 2023). However, it is possible that *PXDNL*'s heart-specific expression may not require as many TF binding sites within the promoter (Levine and Tjian, 2003; Smale and Kadonaga, 2003). Alternatively, given that there are estimated to be over 1600 transcription factors that interact with the human genome (Lambert *et al.*, 2018), it is possible that *PXDNL* may also be regulated by other TFs not investigated in this study.

4.1.2 Shared TFs of *PXDN* and *PXDNL*

PXDN and *PXDNL* shared 16 high confidence TFs including: *NFE2L2*, *RELA*, *FOS*, *JUN*, *JUNB*, *JUND*, *EP300*, *CTCF*, *STAG1*, *STAG2*, *SMC1A*, *RAD21*, *SMARCA4*, *TAL1*, *ERG*, *RBPJ*. Given that *PXDNL* is a paralogue of *PXDN*, it was expected to share some regulatory

TFs, especially given their high sequence similarity (73%) (Péterfi *et al.*, 2014). Furthermore, since both *PXDN* and *PXDNL* are expressed in the heart, some of these shared TFs may play similar roles in regulating their expression. These TFs are linked to a variety of molecular pathways and cardiovascular functions, from oxidative stress management to cell differentiation.

NFE2L2, *RELA*, and the AP-1 components (*FOS*, *JUN*, *JUNB*, *JUND*) were all observed binding to both *PXDN* and *PXDNL* in endothelial cell lines. These TFs play crucial roles in the regulation of oxidative stress and inflammation. *NFE2L2* (nuclear factor erythroid-derived 2-like transcription factor 2), also known as *Nrf2*, regulates the expression of antioxidant response element (ARE)-driven genes (Nguyen *et al.*, 2000), and has been shown to play a vital role in mitigating oxidative damage and inflammation in cardiovascular cell lines (Huang *et al.*, 2022). Notably, Hanmer and Mavri-Damelin (2018) have demonstrated that *NFE2L2/Nrf2* binds to *PXDN*'s promoter, inducing expression in HeLa and HEK293 cell lines. The next TF, *RELA*, is a subunit of the NF-κB complex, which plays a central in the transcriptional response to inflammation and stress, inducing the expression of genes involved in immune responses and cell survival (Hayden and Ghosh, 2012; Tak and Firestein, 2001). Furthermore, Hodgkinson *et al.* (2018) showed that the expression of *RELA* and interactions with the TLR3-NF-κB pathway is essential in cardiomyocyte maturation. The AP-1 complex, consisting of *FOS*, *JUN*, *JUNB*, and *JUND*, regulates gene expression in response to various stimuli, including oxidative stress, playing significant roles in cell proliferation, differentiation, and apoptosis (Shaulian and Karin, 2002).

PXDN has been shown to play a key role in managing oxidative stress through catalysing the oxidation of halide ions by hydrogen peroxide, which in turn modulates inflammation through the production of hypohalous acids (Li *et al.*, 2012). *PXDNL*, on the other hand, may compete with *PXDN* in the heart where it acts as an antagonistic inhibitor (Péterfi *et al.*, 2014). Given that *PXDNL* is not known to be expressed in endothelial cells, it is possible that the aforementioned TFs may only modulate the expression of *PXDN* in response to oxidative stress in these cell lines. These transcription factors could influence *PXDN* expression directly through activating its transcription or indirectly by modulating signalling pathways, such as NF-κB, that may affect *PXDN* during and after transcription in response to oxidative stress.

Key regulators of chromatin accessibility and transcription were also observed binding to both *PXDN* and *PXDNL*. These consisted of two master regulators as well as constituents of the cohesin complex. The first master regulator was the histone acetyltransferase *EP300* (E1A binding protein p300). *EP300* contributes to cardiac cell development through various mechanisms, from post-translational interactions and chromatin remodelling to directly binding gene promoters to activate transcription (Cui *et al.*, 2018a). Bamforth *et al.* (2001) demonstrated that mutations impairing *EP300* and its paralog *CREBBP* adversely affect cardiovascular cell differentiation, cardiac development, and angiogenesis. *CTCF* (CCCTC-binding factor) is another master regulator that intersects various molecular pathways. The roles of *CTCF* include interacting with gene enhancer and promoter regions, thereby modulating, repressing and/or activating gene expression, as well as organising and maintaining the 3D chromatin genome structure (Braccioli and de Wit, 2019). In addition to regulating gene expression, *CTCF* also plays a role in transcript splicing, acting through direct and indirect pathways to generate alternative splice variants (Alharbi *et al.*, 2021).

STAG1, *STAG2*, *SMC1A*, *RAD21* and *SMARCA4* were all observed binding to *PXDN* and *PXDNL* in endothelial cell lines. The first four TFs interact to form the cohesin complex, a crucial structure that is both ubiquitously expressed and conserved across a wide range of eukaryotes (Peters *et al.*, 2008). Studies have suggested that these TFs contribute to chromatin organisation, cell development and DNA repair (Nasmyth and Haering, 2009; Peters *et al.*, 2008). Notably, the cohesin complex interacts with both *CTCF* (Davidson *et al.*, 2023) and *EP300* (DeMare *et al.*, 2013) to regulate chromatin organisation. Given that these TFs regulate large-scale gene accessibility, investigating their effects on chromatin accessibility in different cell types may reveal insights into how *PXDNL* is expressed in only heart tissues while *PXDN* is expressed across a wider range of cell types.

The latter TF, *SMARCA4* (SWI/SNF related matrix associated actin regulator protein 4) is a key component of SWI/SNF, another chromatin remodelling complex. Interestingly, endothelial conditional *SMARCA4* knockout was shown to attenuate liver fibrosis in mice and decrease ROS production (Li *et al.*, 2019). Thus, *SMARCA4*-modulated expression of *PXDN* and *PXDNL* may contribute to the management of ROS in cardiovascular endothelial cell lines. However, it should be noted that in the current study *SMARCA4* was only observed binding in experiments where samples were treated with constitutively active forms of MEK5 or GFP.

ERG was also observed binding to *PXDN* and *PXDNL* in endothelial cells. Notably, its expression has been linked to positive cardiovascular outcomes. Zhang *et al.* (2021) demonstrated that *ERG* knockdown in murine cardiac fibroblasts and HUVECs aggravated cardiac fibrosis and stimulated the release of pro-fibrotic markers. Conversely, *ERG* overexpression appeared to restore normal cellular function, suggesting a protective role in cardiac health. The dysregulation observed in the heart's ECM in the absence of *ERG* might be partially attributed to increased *PXDN* expression, considering its role in facilitating collagen IV and dityrosine crosslinking in the ECM (Cheng *et al.*, 2011; Ero-Tolliver *et al.*, 2015).

The last shared TF, *TAL1*, was identified binding in endothelial cell lines. While there is little research on *TAL1*'s role in endothelial cells, Bussmann *et al.* (2007) demonstrated that *TAL1* mutations in Zebrafish embryos led to malformations in the endocardium, highlighting its role in cardiac development. Furthermore, research by Schumacher *et al.* (2013) established *TAL1*'s importance in endocardial morphogenesis, particularly in maintaining endocardial cell identity and the formation of intercellular junctions. Thus, *TAL1* may regulate the expression of genes which play a vital role in endothelial cell development, which may intersect with *PXDN's* role in the development of the ECM.

The shared TFs of *PXDN* and *PXDNL* suggest that these genes may be co-regulated by a complex network of TFs. However, further research is needed to elucidate the pathways jointly regulating these genes and their roles in the cardiovascular system. Additionally, *PXDNL's* exclusive expression in cardiomyocytes (Péterfi *et al.*, 2014) may be attributed to many of the binding peaks for its regulatory elements being exclusive to *PXDNL's* first intron (Figure 3.2B), while the same TFs are located in *PXDN's* promoter (Figure 3.3A) and first intron (Figure 3.2A). Future research could investigate whether the regulatory elements in *PXDNL's* first intron are still capable of inducing transcription.

4.1.3 Unique TFs: *PXDN*

PXDN's 11 unique TFs include *TCF21*, *KLF4*, *PGR*, *NOTCH1*, *MAX*, *MYC*, *FOXO1*, *CEBPD*, *FOXC2*, *FLI1*, and *HMGB1*. *PXDN* is known to participate in several vital processes, from managing oxidative stress, inflammation and immune response (Cui *et al.*, 2018b), to strengthening the ECM (Bhave *et al.*, 2012; Cheng *et al.*, 2011). Some of the TFs identified in this study can elucidate *PXDN's* roles and transcriptional regulation in the cardiovascular system.

FOXO1 was observed binding to *PXDN* in endothelial cell lines. *FOXO1*, a member of the forkhead box (FOX) family of TFs, has been shown to promote both protective and dysfunctional redox activities. For example, *FOXO1* promotes cardiomyopathy in diabetic mice by inducing the expression of *KLF5*, which can increase oxidative stress by inducing NADPH oxidase (NOX) 4 expression (Kyriazis *et al.*, 2021). However, *FOXO1* has also been observed inducing the expression of antioxidant enzymes such as manganese superoxide dismutase and catalase (Yun *et al.*, 2014). Interestingly, a recent study by Li *et al.* (2021) has shown that *PXDN* modulates the expression and activation of *FOXO1* under conditions of oxidative stress. *FOXO1*'s ability to activate genes that protect against oxidative stress suggests that it may also regulate PXDN, suggesting that they may form part of a feedback loop that dynamically responds to oxidative stress.

HMGB1, also observed binding *PXDN* in endothelial cells, is a member of the high mobility group box protein family and is known for its role as a pro-inflammatory cytokine. *HMGB1* can stimulate antibacterial activity and activate the immune system when present in low concentrations, but causes acute damage when present in high concentrations (Lotze and Tracey, 2005). Notably, oxidative stress-induced activation of the Pi3Kγ/Akt pathway has been shown to promote *HMGB1* expression in cardiomyocytes (Song *et al.*, 2016). While this study did not examine *HMGB1* binding sites in cardiomyocytes, it can be assumed that given the binding site for *HMGB1* in *PXDN*, *HMGB1* may regulate *PXDN's* expression in cardiovascular contexts as well. *HMGB1's* ability to stimulate antibacterial activity and respond to oxidative stress may be linked to *PXDN's* catalytic activity which can modulate oxidative stress and produce hypohalous acids with bactericidal functions (Dougan *et al.*, 2019; Li *et al.*, 2012).

In addition to TFs linked to oxidative stress and immune response, many identified TFs play key roles in cell development. For example, *MAX* and *MYC* are known for their ability to dimerize with other TFs to form complexes which play a critical role in cell proliferation and differentiation (Amati and Land, 1994). Meanwhile, *FOXC2* and *FLI1* both play important roles in vascular development and remodelling (Abedin *et al.*, 2014; Kume, 2008). *FOXC2* has also been shown to promote angiogenesis through activating the *CXCR4-CXCL12* signalling pathway (Hayashi and Kume, 2008). Interestingly, *FOXC2* is also linked to other diseases and processes alongside *PXDN*, namely anterior segment dysgenesis (ASD) and the epithelial-to-mesenchymal transition (EMT) (Hollier *et al.*, 2013; Kuang *et al.*, 2023). *TEAD1* is another TF involved in cardiac development, particularly in reprogramming cardiomyocytes through the

BRD4/*Wnt4* pathway (Song *et al.*, 2024). Furthermore, *TEAD1* and *TCF21* have both been shown to promote the development of cardiac fibroblasts, the former through the *Hippo* pathway, the latter through influencing the ability of progenitor cells to undergo EMT (Acharya *et al.*, 2012; Singh *et al.*, 2021). Similarly, *NOTCH1* has been linked to the development of various cardiovascular cell types and tissues, where it has been shown to play a key role in EMT (High and Epstein, 2008; Niessen and Karsan, 2008). Given that *PXDN's* expression is known to be repressed by the master regulator of EMT, *Snai1* (Sitole and Mavri-Damelin, 2018), *TCF21*, *FOXC2*, and *NOTCH1* may play a complimentary role in mediating *PXDN* expression in cardiac progenitor cells to modulate *PXDN*-driven stabilisation of the ECM.

The relationship between *PXDN* and the TFs *CEBPD*, *KLF4* and *BRD4* remains unclear. *CEBPD*, CCAAT/enhancer-binding protein Delta, has been shown to promote the expression of connective tissue growth factor (*CTGF*) and ECM production in cardiomyocytes (Zhang *et al.*, 2011). However, it should be noted that *CEBPD* ChIP-seq data was only available in 10 ng/mL IL-1B treated endothelial cells (after 4 hours), thus further research is needed to determine if *CEBPD* binds to *PXDN* under non-inflammatory conditions. ChIP-seq data for Kruppel-like factor 4 (KLF4) was also only available in MEK5-treated cells. *KLF4*, while lacking direct links to known functions of *PXDN*, has been identified as a key regulator in previously mentioned pathways. For example, *KLF4* has been shown to inhibit NF-κB activity under inflammatory conditions in endothelial cell lines (Yoshida *et al.*, 2014). Thus, *KLF4* could play an important role in balancing the activation of *PXDN* in inflammatory conditions by inhibiting the activity of NF-κB and other proinflammatory genes like *RELA*.

As with the aforementioned TFs, only treated datasets were available for *BRD4* (Bromodomain containing protein 4). In this case, *BRD4* was observed binding to *PXDN* in TNF-α stimulated endothelial cells. Notably, *BRD4* is a TF with established profibrotic properties. Stratton *et al.* (2019) highlighted *BRD4*'s role in fibrosis, particularly through the transcriptional upregulation of genes that contribute to excessive ECM deposition. This study proposes that these TFs may induce *PXDN* expression, enhancing fibroblast activity and ECM deposition in cardiomyocytes, potentially through *PXDN* overexpression. This may contribute to PXDN-driven cardiac fibrosis observed post-myocardial infarction (Liu *et al.*, 2019).

4.1.4 Unique TFs: *PXDNL*

PXDNL had only two unique TFs, *IRF1* in endothelial cells and *LMNA* in cardiomyocytes. *IRF1* (Interferon Regulatory Factor 1) is known for its role in immune response regulation. In the cardiovascular system, *IRF1* has been shown to mediate inflammatory responses in endothelial cell lines (Neish *et al.*, 1995) and has been implicated in the pathogenesis of atherosclerosis due to its role in modulating endothelial cell function and macrophage activation (Du *et al.*, 2019). On the other hand, *LMNA*, encoding Lamin A/C, is involved in maintaining nuclear structure and integrity (Butin-Israeli *et al.*, 2012; Lin and Worman, 1993). Mutations in *LMNA* have been linked to a range of cardiac abnormalities, including cardiomyopathies (Lu *et al.*, 2011). These conditions highlight *LMNA*'s importance in cardiac structural integrity and electrical signalling. *PXDNL*, known for its expression in cardiomyocytes and elevated levels in failing hearts (Péterfi *et al.*, 2014), may interact with *LMNA* and *IRF1* pathways. The exploration of these TFs and *PXDNL* in cardiovascular cells and CVDs could potentially reveal novel regulatory mechanisms for cardiovascular cell health and new roles for *PXDNL*.

PXDN and *PXDNL*'s TFs share many common molecular pathways linked to inflammation, oxidative stress, vascular remodelling, angiogenesis and apoptosis, as well as gene regulatory pathways linked to chromatin organisation, gene regulation and alternative splicing. However, the additional regulatory TFs unique to *PXDN* suggest a more complex regulatory system involving pathways such as cell proliferation, differentiation, and development. Additionally, many of *PXDN*'s TFs have been noted to have cardioprotective properties or contribute to CVD development, while *PXDNL*'s unique TFs are constrained to inflammation, endothelial cell function and electrical signalling. Despite these findings, the regulatory roles of these TFs with both *PXDN* and *PXDNL* in the cardiovascular system and the impact of CVDs on their expression remain unknown.

4.2 Microarray data exploration

The second aim of this study was to explore the expression profiles of high confidence TFs alongside *PXDN* and *PXDNL* in cardiovascular disease (CVD)-related contexts. *PXDN* has been linked to CVDs such as cardiac fibrosis and peripheral artery disease (PAD) (Costas *et al.*, 2021; Liu *et al.*, 2019). Furthermore, *PXDNL* has been suggested to competitively inhibit *PXDN*'s activity, and could thereby modulate its activity in the heart and in CVDs. However, few studies have investigated the expression patterns of these *PXDN*, *PXDNL* and their regulatory TFs in CVDs. This study bridges this gap by performing differential gene expression analysis on CVD-related gene expression microarray datasets from a public data repository, the Gene Expression Omnibus (GEO).

4.2.1 DGE analysis in CVDs

This study investigated seven datasets across three conditions: cardiomyopathies, heart failure and TNF-α treatment (Table 3.2). These conditions were selected as they provide insights into gene expression changes in CVD conditions (cardiomyopathy and heart failure) and CVD-related conditions (TNF-α). The first condition, cardiomyopathies, represents a heterogenous group of diseases which primarily affect the myocardium and impair cardiac function. Both the underlying mechanisms and symptoms of these diseases are varied. For example, dilated cardiomyopathies (DCM), characterised by enlargement of the ventricular chambers, can be caused by both genetic and environmental factors, with genetic variants playing a significant role (Rosenbaum *et al.*, 2020). The second condition, heart failure, refers to a clinical syndrome where the heart cannot pump sufficiently to meet the body's needs for blood and oxygen. The underlying structural and functional mechanisms leading to heart failure can vary. Non-diabetic heart failure can be caused by several conditions, including diseases such as cardiomyopathies, hypertension and ischemic heart disease (Ceriello *et al.*, 2021; Rosenbaum *et al*., 2020). Diabetic heart failure, however, is synonymous with diabetic cardiomyopathies, and is exclusive to people suffering from diabetes (Ceriello *et al.*, 2021). Given that the previous conditions are subject to various molecular mechanisms, the final condition was selected to specifically investigate the effect of TNF-α induced inflammation in cardiovascular cells, which is known to induce and progress CVDs (Yndestad *et al.*, 2006).

Three datasets were used to explore whether the genes and TFs of interest were differentially expressed in cardiomyopathies: GSE3585, GSE3586, and GSE42955. GSE3585 and GSE3586 were analysed to determine gene expression changes in DCM relative to healthy controls in the left ventricle of the heart and the septum, respectively. Differential gene expression analysis of both datasets detected no significant changes in the expression of *PXDN* or its high confidence TFs. It should be noted that neither of these datasets had probes for *PXDNL*. Similar results were obtained from the analysis of GSE42955, which compared healthy controls to both DCM and ischemic cardiomyopathy (ICM) tissue samples and contained probes for *PXDNL*. Notably, the positive control gene, *NPPA*, did not behave as expected in GSE42955. Furthermore, very few differentially expressed genes were observed in this dataset, alongside a large number of outliers, despite applying limma's arrayWeights. This suggests that the dataset's results may be unreliable. Thus, it appears that only *HMGB1* may be dysregulated in cardiomyopathies.

Many of the previously identified high confidence TFs have been linked to cardiomyopathies through direct and indirect mechanisms. For example, studies have shown that pathogenic genetic variants of *LMNA* can induce and contribute to DCM (Rosenbaum *et al.*, 2020). However, *LMNA*-DCM is rare, with its prevalence estimated to be less than 5% of DCM cases (Pugh *et al.*, 2014). Cell survival/apoptosis pathways such as PI3K/Akt have also been linked to cardiomyopathies. Inhibition of the PI3K/Akt, in conjunction with *FOXO1*, has been shown to attenuate apoptosis, oxidative stress, myocardial fibrosis and cardiac dysfunction in murine diabetic cardiomyopathies (DiCM) (Ren *et al.*, 2020; Yang *et al.*, 2019). The lack of differential expression for *PXDN*, *PXDNL* and their high confidence TFs, including *FOXO1* and PI3K/Akt targets, suggests that their expression may not play a key role in the cardiomyopathy cases represented in these samples. Given the wide variety of molecular mechanisms which can induce cardiomyopathies, it is possible that only certain variants of cardiomyopathy are driven by the dysregulation of *PXDN*, such as *LMNA*-DCM (Rosenbaum *et al.*, 2020). Alternatively, this suggests that if *PXDN* does contribute to the pathogenesis of cardiomyopathies, it may occur through post-translational modifications, protein-protein interactions, or other mechanisms not captured by gene expression changes. Additionally, the changes in gene expression of *PXDN* and *PXDNL* may still be disease specific as well as time-specific and/or too small for microarrays to detect, highlighting the need for more detailed investigations.

The heart failure datasets GSE26887 and GSE76701 revealed similar patterns. Neither *PXDN* nor *PXDNL* exhibited significantly altered gene expression patterns. Furthermore, analysis of GSE76701 revealed no changes for any of the high confidence TFs. In contrast, there were some observable changes for a few TFs in GSE26887. It should be noted that for GSE76701, end-stage heart failure samples were compared with healthy controls, while for GSE26887, healthy controls were compared to both diabetic and non-diabetic heart failure. Furthermore, GSE76701 had a much smaller sample size (n = 8), and quality control detected significant variation between samples. When analysing GSE26887's results, similar gene expression profiles were observed when comparing diabetic and non-diabetic heart failure to healthy controls. *JUN*, *JUNB*, *CEBPD* and *MYC* were all downregulated in both diabetic and non-diabetic conditions. Curiously, *RELA* was only downregulated in non-diabetic heart failure, which may represent a regulatory attempt to dampen the pro-inflammatory state often observed in failing hearts (Yndestad *et al.*, 2006). The observed maintenance of *PXDN* and *PXDNL*'s expression, however, suggests that the expression of their transcripts may not contribute to both diabetic and non-diabetic heart failure. It also suggests that the downregulation of these TFs may not influence their expression under the conditions investigated.

The datasets GSE2638 and GSE2639 were chosen for their relevance in studying the effects of TNF-α, a pro-inflammatory cytokine implicated in cardiovascular diseases (CVDs) (Popa *et al.*, 2007; Sarziputtini *et al.*, 2005), on the expression of *PXDN, PXDNL*, and their regulatory transcription factors. Both datasets compared untreated endothelial cells to endothelial cells stimulated with TNF-α for 5 hours. Once again, no changes were observed in the expression of *PXDN*, *PXDNL* or most of their high confidence TFs. The only exceptions observed in GSE2638 were *FOXC2*, *JUNB*, *RELA* and *IRF1*, which were upregulated, and *FOS*, which was downregulated. *JUNB* and *IRF1* were also upregulated in GSE2638. Unfortunately, these datasets lacked probes for *PXDNL*, so its regulatory relationship with *IRF1* remains unexplored. Furthermore, since *PXDN*'s expression remained the same, it implies that the differentially expressed TFs may not regulate *PXDN*'s expression under the investigated inflammatory conditions. However, these datasets only looked at expression levels after 5 hours of TNF-α. Thus, it may be possible to observe changes that influence PXDN's expression at different time points, especially if *PXDN* is regulated by networks that act through complex signalling cascades that take longer to affect their downstream targets.

To summarise, DGE analysis detected no significant changes in *PXDN* and *PXDNL* expression in DCM, ICM and heart failure. No significant changes were observed for *PXDN* expression in TNF-α treated cell lines as well. Notably, significant expression changes were observed in a handful of high confidence TFs. However, the role of these TFs in the regulation of *PXDN* and *PXDNL* expression and activity remains unclear. These findings highlight the need for further investigation into the molecular pathways linking *PXDN* and *PXDNL* to CVDs.

4.3 Theoretical and practical implications

The findings of this study have significant theoretical implications for our understanding of the cardiovascular system and CVDs. By identifying potential cardiovascular-specific TFs for *PXDN* and *PXDNL* and exploring their expression patterns in CVD-related conditions, this research has extended our understanding of the transcriptional mechanisms regulating these genes in the cardiovascular system. Furthermore, the lack of differential gene expression of *PXDN* and *PXDNL* in cardiomyopathies, heart failure, and TNF-α treatment suggests that their dysregulation may not contribute to the pathogenesis of these CVDs.

Practically, the bioinformatics pipeline developed during this study will be valuable for future research. The custom pipeline developed for TF identification and microarray data analysis is a valuable toolkit that can mine public data repositories for ChIP-seq and microarray data, perform quality control to ensure adequate data integrity, and preprocess datasets for further analyses. As noted by Khomtchouk *et al.* (2020), the complexity of CVDs prompts the development of existing computational methods to analyse a wider variety of data types. By incorporating both gene expression and DNA-protein interaction data, these pipelines bridge these gaps and provide a solid foundation that can be extended to include a wider array of tools and analyse more diverse CVD data types. More broadly, they also offer new avenues for future research into -omics data mining and pipeline development.

Additionally, by demonstrating the value of publicly available bioinformatics data to extract novel insights, this study serves as a model for future investigations in the field. The practical applications of these findings and methodologies could significantly influence the development of targeted and effective therapies. The study's contributions, therefore, not only advance our understanding of the regulation of *PXDN* and *PXDNL* expression in CVDs but also highlight the potential of bioinformatics in mining insights into disease pathogenesis.

4.4 Limitations

This study is subject to several limitations that warrant consideration. Firstly, the reliance on publicly available data, such as those from GEO, inherently carries the risk of compromised data integrity. Despite rigorous quality control measures, the initial quality and consistency of these datasets were largely dependent on the original submitters, affecting the robustness of subsequent analyses. Secondly, the methods used to generate these datasets, including ChIP-seq and gene expression microarrays, have inherent limitations. For example, ChIP-seq can sometimes yield false-positive or false-negative results due to antibody specificity or chromatin accessibility. Similarly, microarrays may not detect subtle gene expression changes or transcripts without representative probes on the array. Thirdly, the analytical methods employed, specifically limma for differential expression analysis and RMA for normalization, while robust and widely used, have their constraints. limma might not fully account for complex biological variability, and RMA normalization, though effective in reducing technical noise, might over-correct or miss subtle biological changes.

Lastly, the molecular mechanisms driving CVD pathogenesis extend beyond gene dysregulation. As this study focused on gene expression microarray data, it cannot comment on post-transcriptional and post-translational modifications that can affect the alternative splicing, localisation and activity of the proteins and transcripts of interest. There were also limitations in the availability of ideal datasets, as some microarray datasets lacked probes for *PXDNL* and ChIP-seq datasets for TFs like *CEBPD*, *SMARCA4* and *BRD4* lacked data for untreated conditions. Additionally, there was a lack of diversity in the types of CVDs examined using microarrays. Each of these limitations has implications for the interpretation of the study's findings and highlights the need for cautious inference and validation in future research.

4.5 Future research

Building on the recognized limitations of this study, several avenues for future research emerge, particularly focusing on the newly identified transcription factors (TFs) for *PXDN* and *PXDNL*. To further validate and elucidate the functional implications of these TFs, comprehensive *in vitro* and *in vivo* studies are recommended. These could include experiments to confirm the physical interactions between these TFs and their target genes, and to elucidate the biological outcomes of these interactions in various cardiovascular disease (CVD) contexts. Additionally, while no significant changes were observed in the expression of *PXDN* and *PXDNL*, future research is needed to determine the role of post-transcriptional and post-translational modifications in regulating their activity, as well as the regulation of their TFs.

Moreover, given that this was primarily an exploratory study leveraging publicly available data, there is a substantial opportunity for more focused investigations using freshly collected, high-quality samples, or adopting more rigorous experimental designs. This approach would allow for the assessment of the roles of *PXDN* and *PXDNL* and their regulatory TFs across a broader spectrum of CVD conditions, employing a combination of *in silico* and *in vivo* methods to provide a more comprehensive understanding of their roles. Furthermore, given that CVDs can exhibit altered expression profiles at different stages in their development, future studies employ longitudinal research to investigate changes in these expression profiles over time.

Additionally, the bioinformatics pipeline developed in this study presents a valuable tool for future research. It is advisable for subsequent studies to not only utilize this pipeline but also to continue its development. Enhancements could include integrating additional normalization and analysis methods to address the limitations identified with RMA and limma or adapting the pipeline to accommodate emerging technologies like next-generation sequencing. Furthermore, incorporating machine learning approaches could refine the identification and prediction of gene-TF interactions, offering deeper insights into the complex regulatory networks at play in CVDs. Incorporating multi-omics data and analysis tools can also enable the investigation of the complex interplay of genomic, transcriptomic, and proteomic factors in cardiovascular diseases in more detail.

By addressing the current limitations and building on the bioinformatics contributions, future work can advance our understanding of cardiovascular biology and contribute to the development of more targeted and effective therapies. This exploratory study opens new paths for further *in silico* and *in vivo* investigations into the roles of genes like *PXDN* and *PXDNL* in different cardiovascular diseases.

4.6 Conclusion

In conclusion, this study has made three significant contributions. Firstly, the identification of potential cardiovascular-specific TFs for *PXDN* and *PXDNL* extends our understanding of the transcriptional factors that regulate their expression and may guide future investigations into their regulation and functions in the cardiovascular system. Secondly, despite expectations, this study found that *PXDN* and *PXDNL* were not differentially expressed in CVD conditions such as cardiomyopathies, heart failure, and TNF-α treatment. These findings challenge previous assumptions but also suggest that future research could elucidate the post-transcriptional and post-translational changes of *PXDN* and *PXDNL* in CVDs. Lastly, the in-house pipelines for TF identification and microarray data analysis assembled during this study can form the foundation for future data-mining and research into the molecular mechanisms driving disease pathogenesis. Furthermore, these pipelines can be further developed to incorporate a wider range of tools and data types, offering new avenues for future bioinformatics-focused research. These contributions demonstrate the value of mining public datasets using bioinformatics tools and advance our understanding of *PXDN* and *PXDNL*'s transcriptional regulation in the cardiovascular system and their expression profiles in CVD-related contexts.

REFERENCES

Abedin, M.J., Nguyen, A., Jiang, N., Perry, C.E., Shelton, J.M., Watson, D.K., Ferdous, A., 2014. Fli1 Acts Downstream of Etv2 to Govern Cell Survival and Vascular Homeostasis via Positive Autoregulation. Circulation Research 114, 1690–1699. https://doi.org/10.1161/CIRCRESAHA.1134303145

Acharya, A., Baek, S.T., Huang, G., Eskiocak, B., Goetsch, S., Sung, C.Y., Banfi, S., Sauer, M.F., Olsen, G.S., Duffield, J.S., Olson, E.N., Tallquist, M.D., 2012. The bHLH transcription factor Tcf21 is required for lineage-specific EMT of cardiac fibroblast progenitors. Development 139, 2139–2149. https://doi.org/10.1242/dev.079970

Aird, W.C., 2007. Phenotypic Heterogeneity of the Endothelium: I. Structure, Function, and Mechanisms. Circulation Research 100, 158–173. https://doi.org/10.1161/01.RES.0000255691.76142.4a

Alharbi, A.B., Schmitz, U., Bailey, C.G., Rasko, J.E.J., 2021. CTCF as a regulator of alternative splicing: new tricks for an old player. Nucleic Acids Research 49, 7825–7838. https://doi.org/10.1093/nar/gkab520

Amati, B., Land, H., 1994. Myc—Max—Mad: a transcription factor network controlling cell cycle progression, differentiation and death. Current Opinion in Genetics & Development 4, 102–108. https://doi.org/10.1016/0959-437X(94)90098-1

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., Ntini, E., Arner, E., Valen, E., Li, K., Schwarzfischer, L., Glatz, D., Raithel, J., Lilje, B., Rapin, N., Bagger, F.O., *et al.*, 2014. An atlas of active enhancers across human cell types and tissues. Nature 507, 455–461. https://doi.org/10.1038/nature12787

Armitage, J., Baigent, C., Barnes, E., Betteridge, D.J., Blackwell, L., Blazing, M., Bowman, L., Braunwald, E., Byington, R., Cannon, C., Clearfield, M., Colhoun, H., Collins, R., Dahlöf, B., Davies, K., Davis, B., De Lemos, J., Downs, J.R., Durrington, P., Emberson, J., *et al.*, 2019. Efficacy and safety of statin therapy in older people: a meta-analysis of individual participant data from 28 randomised controlled trials. The Lancet 393, 407–415. https://doi.org/10.1016/S0140-6736(18)31942-1

Bamforth, S.D., Bragança, J., Eloranta, J.J., Murdoch, J.N., Marques, F.I.R., Kranc, K.R., Farza, H., Henderson, D.J., Hurst, H.C., Bhattacharya, S., 2001. Cardiac malformations, adrenal agenesis, neural crest defects and exencephaly in mice lacking Cited2, a new Tfap2 co-activator. Nature Genetics 29, 469–474. https://doi.org/10.1038/ng768

Barajas-Martinez, H., Smith, M., Hu, D., Goodrow, R.J., Puleo, C., Hasdemir, C., Antzelevitch, C., Pfeiffer, R., Treat, J.A., Cordeiro, J.M., 2020. Susceptibility to Ventricular Arrhythmias Resulting from Mutations in *FKBP1B*, *PXDNL*, and *SCN9A* Evaluated in hiPSC Cardiomyocytes. Stem Cells International 2020, 1–16. https://doi.org/10.1155/2020/8842398

Barrett, T., Edgar, R., 2006. Mining Microarray Data at NCBI\'s Gene Expression Omnibus (GEO)*, in: Gene Mapping, Discovery, and Expression. Humana Press, New Jersey, pp. 175–190. https://doi.org/10.1385/1-59745-097-9:175

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., Zhao, K., 2007. High-Resolution Profiling of Histone Methylations in the Human Genome. Cell 129, 823–837. https://doi.org/10.1016/j.cell.2007.05.009

Barth, A.S., Kuner, R., Buness, A., Ruschhaupt, M., Merk, S., Zwermann, L., Kääb, S., Kreuzer, E., Steinbeck, G., Mansmann, U., Poustka, A., Nabauer, M., Sültmann, H., 2006. Identification of a Common Gene Expression Signature in Dilated Cardiomyopathy Across Independent Microarray Studies. Journal of the American College of Cardiology 48, 1610–1617. https://doi.org/10.1016/j.jacc.2006.07.026

Bathish, B., Paumann-Page, M., Paton, L.N., Kettle, A.J., Winterbourn, C.C., 2020. Peroxidasin mediates bromination of tyrosine residues in the extracellular matrix. Journal of Biological Chemistry 295, 12697–12705. https://doi.org/10.1074/jbc.RA120.014504

Bhave, G., Cummings, C.F., Vanacore, R.M., Kumagai-Cresse, C., Ero-Tolliver, I.A., Rafi, M., Kang, J.-S., Pedchenko, V.P., Fessler, L.I., Fessler, J.H., Hudson, B.G., 2012. Peroxidasin Forms Sulfilimine Chemical Bonds Using Hypohalous Acids In Tissue Genesis. Nature chemical biology 8, 784–790. https://doi.org/10.1038/nchembio.1038

Blackwood, E.M., Kadonaga, J.T., 1998. Going the Distance: A Current View of Enhancer Action. Science 281, 60–63. https://doi.org/10.1126/science.281.5373.60

Bolstad, B.M., Irizarry, R.A., Åstrand, M., Speed, T.P., 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19, 185–193. https://doi.org/10.1093/bioinformatics/19.2.185

Bommer, C., Heesemann, E., Sagalova, V., Manne-Goehler, J., Atun, R., Bärnighausen, T., Vollmer, S., 2017. The global economic burden of diabetes in adults aged 20–79 years: a cost-of-illness study. The Lancet Diabetes & Endocrinology 5, 423–430. https://doi.org/10.1016/S2213-8587(17)30097-9

Boukhris, M., Hillani, A., Moroni, F., Annabi, M.S., Addad, F., Ribeiro, M.H., Mansour, S., Zhao, X., Ybarra, L.F., Abbate, A., Vilca, L.M., Azzalini, L., 2020. Cardiovascular Implications of the COVID-19 Pandemic: A Global Perspective. Canadian Journal of Cardiology 36, 1068–1080. https://doi.org/10.1016/j.cjca.2020.05.018

Braccioli, L., de Wit, E., 2019. CTCF: a Swiss-army knife for genome organization and transcription regulation. Essays in Biochemistry 63, 157–165. https://doi.org/10.1042/EBC20180069

Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G.G., Oezcimen, 2003. ArrayExpress--a public repository for microarray gene expression data at the EBI. Nucleic Acids Research 31, 68–71. https://doi.org/10.1093/nar/gkg091

Bryant, D., Becker, L., Richardson, J., Shelton, J., Franco, F., Peshock, R., Thompson, M., Giroir, B., 1998. Cardiac Failure in Transgenic Mice With Myocardial Expression of Tumor Necrosis Factor-α. Circulation 97, 1375–1381. https://doi.org/10.1161/01.CIR.97.14.1375

Bujak, M., Dobaczewski, M., Chatila, K., Mendoza, L.H., Li, N., Reddy, A., Frangogiannis, N.G., 2008. Interleukin-1 Receptor Type I Signaling Critically Regulates Infarct Healing and Cardiac Remodeling. The American Journal of Pathology 173, 57–67. https://doi.org/10.2353/ajpath.2008.070974

Bujak, M., Dobaczewski, M., Gonzalez-Quesada, C., Xia, Y., Leucker, T., Zymek, P., Veeranna, V., Tager, A.M., Luster, A.D., Frangogiannis, N.G., 2009. Induction of the CXC Chemokine Interferon-γ-Inducible Protein 10 Regulates the Reparative Response Following Myocardial Infarction. Circulation Research 105, 973–983. https://doi.org/10.1161/CIRCRESAHA.109.199471

Bussmann, J., Bakkers, J., Schulte-Merker, S., 2007. Early Endocardial Morphogenesis Requires Scl/Tal1. PLOS Genetics 3, e140. https://doi.org/10.1371/journal.pgen.0030140

Butin-Israeli, V., Adam, S.A., Goldman, A.E., Goldman, R.D., 2012. Nuclear lamin functions and disease. Trends in genetics 28, 464–471. https://doi.org/10.1016/j.tig.2012.06.001

Cahan, P., Rovegno, F., Mooney, D., Newman, J.C., St. Laurent, G., McCaffrey, T.A., 2007. Meta-analysis of microarray results: challenges, opportunities, and recommendations for standardization. Gene 401, 12–18. https://doi.org/10.1016/j.gene.2007.06.016

Cai, X., Yang, Xiwen, Jin, C., Li, L., Cui, Q., Guo, Y., Dong, Y., Yang, Xiaohua, Guo, L., Zhang, M., 2018. Identification and verification of differentially expressed microRNAs and their target genes for the diagnosis of esophageal cancer. Oncology Letters https://doi.org/10.3892/ol.2018.9066

Cao, D.J., 2018. Macrophages in Cardiovascular Homeostasis and Disease. Circulation 138, 2452–2455. https://doi.org/10.1161/CIRCULATIONAHA.118.035736

Carvalho, B.S., Irizarry, R.A., 2010. A framework for oligonucleotide microarray preprocessing. Bioinformatics 26, 2363–2367. https://doi.org/10.1093/bioinformatics/btq431

Casula, M., Mozzanica, F., Scotti, L., Tragni, E., Pirillo, A., Corrao, G., Catapano, A.L., 2017. Statin use and risk of new-onset diabetes: A meta-analysis of observational studies. Nutrition, Metabolism and Cardiovascular Diseases 27, 396–406. https://doi.org/10.1016/j.numecd.2017.03.001

Ceriello, A., Catrinoiu, D., Chandramouli, C., Cosentino, F., Dombrowsky, A.C., Itzhak, B., Lalic, N.M., Prattichizzo, F., Schnell, O., Seferović, P.M., Valensi, P., 2021. Heart failure in type 2 diabetes: current perspectives on screening, diagnosis and management. Cardiovascular Diabetology 20, 218. https://doi.org/10.1186/s12933-021-01408-1

Chen L., Deng H., Cui H., Fang J., Zuo Z., Deng J., Li Y., Wang X., Zhao L., 2017. Inflammatory responses and inflammation-associated diseases in organs. Oncotarget 9, 7204-7218. https://doi.org/10.18632/oncotarget.23208

Cheng, G., Li, H., Cao, Z., Qiu, X., McCormick, S., Thannickal, V.J., Nauseef, W.M., 2011. Vascular peroxidase-1 is rapidly secreted, circulates in plasma, and supports dityrosine cross-linking reactions. Free Radical Biology and Medicine 51, 1445–1453. https://doi.org/10.1016/j.freeradbiomed.2011.07.002

Cheng, G., Salerno, J.C., Cao, Z., Pagano, P.J., Lambeth, J.D., 2008. Identification and characterization of VPO1, a new animal heme-containing peroxidase. Free Radical Biology and Medicine 45, 1682–1694. https://doi.org/10.1016/j.freeradbiomed.2008.09.009

Cheng, G., Shi, R., 2022. Mammalian peroxidasin (PXDN): From physiology to pathology. Free Radical Biology and Medicine 182, 100–107. https://doi.org/10.1016/j.freeradbiomed.2022.02.026

Choi, A., Lao, R., Ling-Fung Tang, P., Wan, E., Mayer, W., Bardakjian, T., Shaw, G.M., Kwok, P., Schneider, A., Slavotinek, A., 2015. Novel mutations in PXDN cause microphthalmia and anterior segment dysgenesis. European Journal of Human Genetics 23, 337–341. https://doi.org/10.1038/ejhg.2014.119

Chung, M.K., Zidar, D.A., Bristow, M.R., Cameron, S.J., Chan, T., Harding, C.V., Kwon, D.H., Singh, T., Tilton, J.C., Tsai, E.J., Tucker, N.R., Barnard, J., Loscalzo, J., 2021. COVID-19 and Cardiovascular Disease. Circulation Research 128, 1214–1236. https://doi.org/10.1161/CIRCRESAHA.121.317997

Colon, S., Bhave, G., 2016. Proprotein Convertase Processing Enhances Peroxidasin Activity to Reinforce Collagen IV. Journal of Biological Chemistry 291, 24009–24016. https://doi.org/10.1074/jbc.M116.745935

Colon, S., Luan, H., Liu, Y., Meyer, C., Gewin, L., Bhave, G., 2019. Peroxidasin and eosinophil peroxidase, but not myeloperoxidase, contribute to renal fibrosis in the murine unilateral ureteral obstruction model. American Journal of Physiology-Renal Physiology 316, F360–F371. https://doi.org/10.1152/ajprenal.00291.2018

Costas, L., Herz, C.T., Höbaus, C., Koppensteiner, R., Schernthaner, G.-H., 2021. Vascular peroxidase 1 is independently associated with worse kidney function in patients with peripheral artery disease. Journal of Nephrology 34, 165–172. https://doi.org/10.1007/s40620-020-00818-w

Crandall, J.P., Mather, K., Rajpathak, S.N., Goldberg, R.B., Watson, K., Foo, S., Ratner, R., Barrett-Connor, E., Temprosa, M., 2017. Statin use and risk of developing diabetes: results from the Diabetes Prevention Program. BMJ Open Diabetes Research and Care 5, e000438. https://doi.org/10.1136/bmjdrc-2017-000438

Cui, M., Wang, Z., Bassel-Duby, R., Olson, E.N., 2018a. Genetic and epigenetic regulation of cardiomyocytes in development, regeneration and disease. Development 145, dev171983. https://doi.org/10.1242/dev.171983

Cui, S., Lv, X., Li, W., Li, Z., Liu, H., Gao, Y., Huang, G., 2018b. Folic acid modulates VPO1 DNA methylation levels and alleviates oxidative stress-induced apoptosis in vivo and in vitro. Redox Biology 19, 81–91. https://doi.org/10.1016/j.redox.2018.08.005

Das, V., Bhattacharya, S., Chikkaputtaiah, C., Hazra, S., Pal, M., 2019. The basics of epithelial–mesenchymal transition (EMT): A study from a structure, dynamics, and functional perspective. Journal of Cellular Physiology 234, 14535–14555. https://doi.org/10.1002/jcp.28160

Davidson, I.F., Barth, R., Zaczek, M., Van Der Torre, J., Tang, W., Nagasaka, K., Janissen, R., Kerssemakers, J., Wutz, G., Dekker, C., Peters, J.-M., 2023. CTCF is a DNA-tension-dependent barrier to cohesin-mediated loop extrusion. Nature 616, 822–827. https://doi.org/10.1038/s41586-023-05961-5

Davis, S., Meltzer, P.S., 2007. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics 23, 1846–1847. https://doi.org/10.1093/bioinformatics/btm254

DeMare, L.E., Leng, J., Cotney, J., Reilly, S.K., Yin, J., Sarro, R., Noonan, J.P., 2013. The genomic landscape of cohesin-associated chromatin interactions. Genome Research 23, 1224–1234. https://doi.org/10.1101/gr.156570.113

Desmond, J.C., Raynaud, S., Tung, E., Hofmann, W.-K., Haferlach, T., Koeffler, H.P., 2007. Discovery of epigenetically silenced genes in acute myeloid leukemias. Leukemia 21, 1026–1034. https://doi.org/10.1038/sj.leu.2404611

Dougan, Hawsawi, Burton, Edwards, Jones, Zou, Nagappan, Wang, Zhang, Danaher, Bowen, Hinton, Odero-Marah, 2019. Proteomics-Metabolomics Combined Approach Identifies Peroxidasin as a Protector against Metabolic and Oxidative Stress in Prostate Cancer. IJMS 20, 3046. https://doi.org/10.3390/ijms20123046

Draghici, S., Khatri, P., Eklund, A.C., Szallasi, Z., 2008. Reliability and reproducibility issues in DNA microarray measurements. TRENDS in Genetics 22, 101–109

Du, M., Wang, X., Mao, X., Yang, L., Huang, Kun, Zhang, F., Wang, Y., Luo, X., Wang, C., Peng, J., Liang, M., Huang, D., Huang, Kai, 2019. Absence of Interferon Regulatory Factor 1 Protects Against Atherosclerosis in Apolipoprotein E-Deficient Mice. Theranostics 9, 4688–4703. https://doi.org/10.7150/thno.36862

Dudoit, S., Yang, Y.H., Callow, M.J., Speed, T.P., 2002. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica sinica 12, 111–139.

Edgar, R., 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research 30, 207–210. https://doi.org/10.1093/nar/30.1.207

Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D., 1998. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences U.S.A. 95, 14863–14868. https://doi.org/10.1073/pnas.95.25.14863

Ensembl, 2023a. Gene: PXDNL ENSG00000147485 [WWW Document]. EMBL-EBI. URL https://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000147485;r=8:51319577-51809445 (accessed 12.7.2023).

Ensembl, 2023b. Gene: PXDN ENSG00000130508 [WWW Document]. EMBL-EBI. URL https://www.ensembl.org/Homo_sapiens/Gene/Summary?g=ENSG00000130508;r=2:163188 7-1744852 (accessed 12.7.2023).

Ero-Tolliver, I.A., Hudson, B.G., Bhave, G., 2015. The Ancient Immunoglobulin Domains of Peroxidasin Are Required to Form Sulfilimine Cross-links in Collagen IV. Journal of Biological Chemistry 290, 21741–21748. https://doi.org/10.1074/jbc.M115.673996

Feng, Y., Cai, L., Hong, W., Zhang, C., Tan, N., Wang, M., Wang, C., Liu, F., Wang, X., Ma, J., Gao, C., Kumar, M., Mo, Y., Geng, Q., Luo, C., Lin, Y., Chen, H., Wang, S.-Y., Watson, M.J., Jegga, A.G., *et al.*, 2022. Rewiring of 3D Chromatin Topology Orchestrates Transcriptional Reprogramming and the Development of Human Dilated Cardiomyopathy. Circulation 145, 1663–1683. https://doi.org/10.1161/CIRCULATIONAHA.121.055781

Fleige, S., Walf, V., Huch, S., Prgomet, C., Sehm, J., Pfaffl, M.W., 2006. Comparison of relative mRNA quantification models and the impact of RNA integrity in quantitative real-time RT-PCR. Biotechnology letters 28, 1601–1613. https://doi.org/10.1007/s10529-006-9127-2

Frangogiannis, N.G., 2021. Cardiac fibrosis. Cardiovascular Research 117, 1450–1488. https://doi.org/10.1093/cvr/cvaa324

Frangogiannis, N.G., 2019. Cardiac fibrosis: Cell biological mechanisms, molecular pathways and therapeutic opportunities. Molecular Aspects of Medicine 65, 70–99. https://doi.org/10.1016/j.mam.2018.07.001

Fujita, T., Ishikawa, Y., 2011. Apoptosis in Heart Failure - The Role of the β-Adrenergic Receptor-Mediated Signaling Pathway and p53-Mediated Signaling Pathway in the Apoptosis of Cardiomyocytes -: – The Role of the β-Adrenergic Receptor-Mediated Signaling Pathway and p53-Mediated Signaling Pathway in the Apoptosis of Cardiomyocytes –. Circulation Journal 75, 1811–1818. https://doi.org/10.1253/circj.CJ-11-0025

Gaspar, J.M., 2018. Improved peak-calling with MACS2. bioRxiv 496521. https://doi.org/10.1101/496521

Gautier, L., Cope, L., Bolstad, B.M., Irizarry, R.A., 2004. affy—analysis of *Affymetrix GeneChip* data at the probe level. Bioinformatics 20, 307–315. https://doi.org/10.1093/bioinformatics/btg405

Gerdes, A.M., Kellerman, S.E., Moore, J.A., Muffly, K.E., Clark, L.C., Reaves, P.Y., Malec, K.B., McKeown, P.P., Schocken, D.D., 1992. Structural remodeling of cardiac myocytes in patients with ischemic cardiomyopathy. Circulation 86, 426–430. https://doi.org/10.1161/01.CIR.86.2.426

González, G.E., Rhaleb, N.-E., D'Ambrosio, M.A., Nakagawa, P., Liu, Y., Leung, P., Dai, X., Yang, X.-P., Peterson, E.L., Carretero, O.A., 2015. Deletion of interleukin-6 prevents cardiac inflammation, fibrosis and dysfunction without affecting blood pressure in angiotensin II-high salt-induced hypertension. Journal of Hypertension 33, 144–152. https://doi.org/10.1097/HJH.0000000000000358

Greco, S., Fasanaro, P., Castelvecchio, S., D'Alessandra, Y., Arcelli, D., Di Donato, M., Malavazos, A., Capogrossi, M.C., Menicanti, L., Martelli, F., 2012. MicroRNA Dysregulation in Diabetic Ischemic Heart Failure Patients. Diabetes 61, 1633–1641. https://doi.org/10.2337/db11-0952

Hamid, S., Groot, W., Pavlova, M., 2019. Trends in cardiovascular diseases and associated risks in sub-Saharan Africa: a review of the evidence for Ghana, Nigeria, South Africa, Sudan and Tanzania. The Aging Male 22, 169–176. https://doi.org/10.1080/13685538.2019.1582621

Hanmer, K.L., Mavri-Damelin, D., 2018. Peroxidasin is a novel target of the redox-sensitive transcription factor Nrf2. Gene 674, 104–114. https://doi.org/10.1016/j.gene.2018.06.076

Hayashi, H., Kume, T., 2008. Forkhead transcription factors regulate expression of the chemokine receptor CXCR4 in endothelial cells and CXCL12-induced cell migration. Biochemical and Biophysical Research Communications 367, 584–589. https://doi.org/10.1016/j.bbrc.2007.12.183

Hayden, M.S., Ghosh, S., 2012. NF-κB, the first quarter-century: remarkable progress and outstanding questions. Genes & development 26, 203–234. https://doi.org/10.1101/gad.183434.111

Heller, R.A., Schena, M., Chai, A., Shalon, D., Bedilion, T., Gilmore, J., Woolley, D.E., Davis, R.W., 1997. Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. Proceedings of the National Academy of Sciences 94, 2150–2155.

Henein, M. Y., Vancheri, S., Longo, G., Vancheri, F., 2022. The role of inflammation in cardiovascular disease. International Journal of Molecular Sciences 23, 12906. https://doi.org/10.3390/ijms232112906

High, F.A., Epstein, J.A., 2008. The multifaceted role of Notch in cardiac development and disease. Nature Reviews Genetics 9, 49–61. https://doi.org/10.1038/nrg2279

Hill, M.A., Yang, Y., Zhang, L., Sun, Z., Jia, G., Parrish, A.R., Sowers, J.R., 2021. Insulin resistance, cardiovascular stiffening and cardiovascular disease. Metabolism 119, 154766. https://doi.org/10.1016/j.metabol.2021.154766

Hirase, T., Node, K., 2012. Endothelial dysfunction as a cellular mechanism for vascular failure. American Journal of Physiology-Heart and Circulatory Physiology 302, H499–H505. https://doi.org/10.1152/ajpheart.00325.2011

Hodgkinson, C.P., Pratt, R.E., Kirste, I., Dal-Pra, S., Cooke, J.P., Dzau, V.J., 2018. Cardiomyocyte Maturation Requires TLR3 Activated Nuclear Factor Kappa B. Stem Cells 36, 1198–1209. https://doi.org/10.1002/stem.2833

Hollier, B.G., Tinnirello, A.A., Werden, S.J., Evans, K.W., Taube, J.H., Sarkar, T.R., Sphyris, N., Shariati, M., Kumar, S.V., Battula, V.L., Herschkowitz, J.I., Guerra, R., Chang, J.T., Miura, N., Rosen, J.M., Mani, S.A., 2013. FOXC2 Expression Links Epithelial–Mesenchymal Transition and Stem Cell Properties in Breast Cancer. Cancer Research 73, 1981–1992. https://doi.org/10.1158/0008-5472.CAN-12-2962

Huang, Z., Wu, M., Zeng, L., Wang, D., 2022. The Beneficial Role of Nrf2 in the Endothelial Dysfunction of Atherosclerosis. Cardiology Research and Practice 2022, 4287711. https://doi.org/10.1155/2022/4287711

Huber, W., Carey, V.J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B.S., Bravo, H.C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K.D., Irizarry, R.A., Lawrence, M., Love, M.I., MacDonald, J., Obenchain, V., Oleś, A.K., Pagès, H., *et al.*, 2015. Orchestrating high-throughput genomic analysis with Bioconductor. Nature Methods 12, 115–121. https://doi.org/10.1038/nmeth.3252

Ioannidis, J.P.A., 2010. Meta-research: The art of getting it wrong. Research Synthesis Methods 1, 169–184. https://doi.org/10.1002/jrsm.19

Irizarry, R.A., 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4, 249–264. https://doi.org/10.1093/biostatistics/4.2.249

Johnson, D.S., Mortazavi, A., Myers, R.M., Wold, B., 2007. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. Science 316, 1497–1502. https://doi.org/10.1126/science.1141319

Kalluri, R., Weinberg, R.A., 2009. The basics of epithelial-mesenchymal transition. The Journal of clinical investigation 119, 1420-1428. https://doi.org/10.1172/JCI39104

Katz, A.M., 2010. Physiology of the Heart. Lippincott Williams & Wilkins.

Kaufman, J., Montalvo-Ortiz, J.L., Holbrook, H., O'Loughlin, K., Orr, C., Kearney, C., Yang, B.-Z., Wang, T., Zhao, H., Althoff, R., Garavan, H., Gelernter, J., Hudziak, J., 2018. Adverse Childhood Experiences, Epigenetic Measures, and Obesity in Youth. The Journal of Pediatrics 202, 150-156.e3. https://doi.org/10.1016/j.jpeds.2018.06.051

Khan, K., Rudkin, A., Parry, D.A., Burdon, K.P., McKibbin, M., Logan, C.V., Abdelhamed, Z.I.A., Muecke, J.S., Fernandez-Fuentes, N., Laurie, K.J., Shires, M., Fogarty, R., Carr, I.M., Poulter, J.A., Morgan, J.E., Mohamed, M.D., Jafri, H., Raashid, Y., Meng, N., Piseth, H., *et al.*, 2011. Homozygous Mutations in PXDN Cause Congenital Cataract, Corneal Opacity, and Developmental Glaucoma. The American Journal of Human Genetics 89, 464–473. https://doi.org/10.1016/j.ajhg.2011.08.005

Khomtchouk, B.B., Tran, D.-T., Vand, K.A., Might, M., Gozani, O., Assimes, T.L., 2020. Cardioinformatics: the nexus of bioinformatics and precision cardiology. Briefings in Bioinformatics 21, 2031–2051. https://doi.org/10.1093/bib/bbz119

Kim, E.H., Galchev, V.I., Kim, J.Y., Misek, S.A., Stevenson, T.K., Campbell, M.D., Pagani, F.D., Day, S.M., Johnson, T.C., Washburn, J.G., Vikstrom, K.L., Michele, D.E., Misek, D.E., Westfall, M.V., 2016. Differential protein expression and basal lamina remodeling in human heart failure. PROTEOMICS–Clinical Applications 10, 585–596. https://doi.org/10.1002/prca.201500099

Kim, H.-K., Ham, K.A., Lee, S.-W., Choi, H.S., Kim, H.-S., Kim, H.K., Shin, H.-S., Seo, K.Y., Cho, Y., Nam, K.T., Kim, I.-B., Joe, Y.A., 2019. Biallelic Deletion of Pxdn in Mice Leads to

Anophthalmia and Severe Eye Malformation. IJMS 20, 6144. https://doi.org/10.3390/ijms20246144

Kolattukudy, P.E., Bergese, S., Breckenridge, S., Hensley, J., Klenotic, S., Orosz, C., Parker-Thornburg, J., 1998. Myocarditis Induced by Targeted Expression of the MCP-1 Gene in Murine Cardiac Muscle. The American Journal of Pathology 152, 11.

Kolde, R., 2019. pheatmap: Pretty Heatmaps. R package version 1.0.12. https://CRAN.R-project.org/package=pheatmap

Kovács, H.A., Lázár, E., Várady, G., Sirokmány, G., Geiszt, M., 2021. Characterization of the Proprotein Convertase-Mediated Processing of Peroxidasin and Peroxidasin-like Protein. Antioxidants 10, 1565. https://doi.org/10.3390/antiox10101565

Kruglov, E.A., Nathanson, R.A., Nguyen, T., Dranoff, J.A., 2006. Secretion of MCP-1/CCL2 by bile duct epithelia induces myofibroblastic transdifferentiation of portal fibroblasts. American Journal of Physiology-Gastrointestinal and Liver Physiology 290, G765–G771. https://doi.org/10.1152/ajpgi.00308.2005

Kuang, L., Zhang, M., Wang, T., Huang, T., Li, J., Gan, R., Yu, M., Cao, W., Yan, X., 2023. The molecular genetics of anterior segment dysgenesis. Experimental Eye Research 234, 109603. https://doi.org/10.1016/j.exer.2023.109603

Kume, T., 2008. Foxc2 Transcription Factor: A Newly Described Regulator of Angiogenesis. Trends in Cardiovascular Medicine 18, 224–228. https://doi.org/10.1016/j.tcm.2008.11.003

Kurihara-Shimomura, M., Sasahira, T., Shimomura, H., Kirita, T., 2020. Peroxidan Plays a Tumor-Promoting Role in Oral Squamous Cell Carcinoma. IJMS 21, 5416. https://doi.org/10.3390/ijms21155416

Kutcher, M.E., Herman, I.M., 2009. The pericyte: Cellular regulator of microvascular blood flow. Microvascular Research 77, 235–246. https://doi.org/10.1016/j.mvr.2009.01.007

Kyriazis, I.D., Hoffman, M., Gaignebet, L., Lucchese, A.M., Markopoulou, E., Palioura, D., Wang, C., Bannister, T.D., Christofidou-Solomidou, M., Oka, S., Sadoshima, J., Koch, W.J., Goldberg, I.J., Yang, V.W., Bialkowska, A.B., Kararigas, G., Drosatos, K., 2021. KLF5 Is Induced by FOXO1 and Causes Oxidative Stress and Diabetic Cardiomyopathy. Circulation research 128, 335–357. https://doi.org/10.1161/CIRCRESAHA.120.316738

Laakso, M., Kuusisto, J., 2017. Diabetes Secondary to Treatment with Statins. Current diabetes reports 17, 10. https://doi.org/10.1007/s11892-017-0837-8

Lai, N.C., Gao, M.H., Tang, E., Tang, R., Guo, T., Dalton, N.D., Deng, A., Tang, T., 2012. Pressure overload-induced cardiac remodeling and dysfunction in the absence of interleukin 6 in mice. Laboratory Investigation 92, 1518–1526. https://doi.org/10.1038/labinvest.2012.97

Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R., Weirauch, M.T., 2018. The Human Transcription Factors. Cell 172, 650–665. https://doi.org/10.1016/j.cell.2018.01.029

Lázár, E., Péterfi, Z., Sirokmány, G., Kovács, H.A., Klement, E., Medzihradszky, K.F., Geiszt, M., 2015. Structure–function analysis of peroxidasin provides insight into the mechanism of collagen IV crosslinking. Free Radical Biology and Medicine 83, 273–282. https://doi.org/10.1016/j.freeradbiomed.2015.02.015

Lee, S., Kim, H., Naidansuren, P., Ham, K.A., Choi, H.S., Ahn, H., Kim, M., Kang, D.H., Kang, S.W., Joe, Y.A., 2020. Peroxidasin is essential for endothelial cell survival and growth signaling by sulfilimine crosslink-dependent matrix assembly. The FASEB journal 34, 10228–10241. https://doi.org/10.1096/fj.201902899R

Levick, J.R., 1991. Overview of the cardiovascular system, in: An Introduction to Cardiovascular Physiology. Butterworth-Heinemann, pp. 1–12. https://doi.org/10.1016/B978-0-7506-1028-5.50004-4

Levine, M., Tjian, R., 2003. Transcription regulation and animal diversity. Nature 424, 147–151. https://doi.org/10.1038/nature01763

Li, C., Liu, Z., Xu, Q., Peng, H., Cao, J., Zhou, H., Zhang, G., Cheng, G., Shi, R., 2021. PXDN reduces autophagic flux in insulin-resistant cardiomyocytes via modulating FoxO1. Cell death & disease 12, 418. https://doi.org/10.1038/s41419-021-03699-4

Li, M., Rao, M., Chen, K., Zhou, J., Song, J., 2017. Selection of reference genes for gene expression studies in heart failure for left and right ventricles. Gene 620, 30–35. https://doi.org/10.1016/j.gene.2017.04.006

Li, X.-Z., Wei, X., Zhang, C.-J., Jin, X.-L., Tang, J.-J., Fan, G.-J., Zhou, B., 2012. Hypohalous acid-mediated halogenation of resveratrol and its role in antioxidant and antimicrobial activities. Food Chemistry 135, 1239–1244. https://doi.org/10.1016/j.foodchem.2012.05.043

Li, Z., Chen, B., Dong, W., Kong, M., Shao, Y., Fan, Z., Yu, L., Wu, D., Lu, J., Guo, J., Xu, Y., 2019. The Chromatin Remodeler Brg1 Integrates ROS Production and Endothelial-Mesenchymal Transition to Promote Liver Fibrosis in Mice. Frontiers in cell and developmental biology 7, 245. https://doi.org/10.3389/fcell.2019.00245

Lim, B.-K., Choe, S.-C., Shin, J.-O., Ho, S.-H., Kim, J.-M., Yu, S.-S., Kim, S., Jeon, E.-S., 2002. Local Expression of Interleukin-1 Receptor Antagonist by Plasmid DNA Improves Mortality and Decreases Myocardial Inflammation in Experimental Coxsackieviral Myocarditis. Circulation 105, 1278–1281. https://doi.org/10.1161/circ.105.11.1278

Lin, F., Worman, H.J., 1993. Structural organization of the human gene encoding nuclear lamin A and nuclear lamin C. Journal of Biological Chemistry 268, 16321–16326.

Liu, Z., Xu, Q., Yang, Q., Cao, J., Wu, C., Peng, H., Zhang, X., Chen, J., Cheng, G., Wu, Y., Shi, R., Zhang, G., 2019. Vascular peroxidase 1 is a novel regulator of cardiac fibrosis after myocardial infarction. Redox Biology 22, 101151. https://doi.org/10.1016/j.redox.2019.101151

Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Michiko, K., Norton, H., Brown, E.L., 1996. Expression monitoring by

hybridization to high-density oligonucleotide arrays. Nature Biotechnology 14, 1675–1680. https://doi.org/10.1038/nbt1296-1675

Lockhart, D.J., Winzeler, E.A., 2000. Genomics, gene expression and DNA arrays. Nature 405, 827–836. https://doi.org/10.1038/35015701

Lotze, M.T., Tracey, K.J., 2005. High-mobility group box 1 protein (HMGB1): nuclear weapon in the immune arsenal. Nature Reviews Immunology 5, 331–342. https://doi.org/10.1038/nri1594

Lu, J.T., Muchir, A., Nagy, P.L., Worman, H.J., 2011. *LMNA* cardiomyopathy: cell biology and genetics meet clinical medicine. Disease Models & Mechanisms 4, 562–568. https://doi.org/10.1242/dmm.006346

Lu, M., Liu, B., Li, D., Gao, Z., Li, W., Zhou, X., Zhan, H., 2023. PXDNL activates the motility of urothelial bladder carcinoma cells through the Wnt/β-catenin pathway and has a prognostic value. Life Sciences 312, 121270. https://doi.org/10.1016/j.lfs.2022.121270

Luco, R.F., Allo, M., Schor, I.E., Kornblihtt, A.R., Misteli, T., 2011. Epigenetics in Alternative Pre-mRNA Splicing. Cell 144, 16–26. https://doi.org/10.1016/j.cell.2010.11.056

Man, J., Barnett, P., Christoffels, V.M., 2018. Structure and function of the Nppa–Nppb cluster locus during heart development and disease. Cellular and Molecular Life Sciences 75, 1435–1444. https://doi.org/10.1007/s00018-017-2737-0

Martin, F.J., Amode, M.R., Aneja, A., Austine-Orimoloye, O., Azov, A.G., Barnes, I., Becker, A., Bennett, R., Berry, A., Bhai, J., Bhurji, S.K., Bignell, A., Boddu, S., Branco Lins, P.R., Brooks, L., Ramaraju, S.B., Charkhchi, M., Cockburn, A., Da Rin Fiorretto, L., Davidson, C., *et al.*, 2023. Ensembl 2023. Nucleic Acids Research 51, D933–D941. https://doi.org/10.1093/nar/gkac958

Medfai, H., Khalil, A., Rousseau, A., Nuyens, V., Paumann-Page, M., Sevcnikar, B., Furtmüller, P.G., Obinger, C., Moguilevsky, N., Peulen, O., Herfs, M., Castronovo, V., Amri, M., Van Antwerpen, P., Vanhamme, L., Zouaoui Boudjeltia, K., 2019. Human peroxidasin 1 promotes angiogenesis through ERK1/2, Akt, and FAK pathways. Cardiovascular Research 115, 463–475. https://doi.org/10.1093/cvr/cvy179

Mezzaroma, E., Mikkelsen, R.B., Toldo, S., Mauro, A.G., Sharma, K., Marchetti, C., Alam, A., Van Tassell, B.W., Gewirtz, D.A., Abbate, A., 2015. Role of Interleukin-1 in Radiation-Induced Cardiomyopathy. Molecular Medicine 21, 210–218. https://doi.org/10.2119/molmed.2014.00243

Micheal, S., Siddiqui, S.N., Zafar, S.N., Iqbal, A., Khan, M.I., den Hollander, A.I., 2016. Identification of Novel Variants in LTBP2 and PXDN Using Whole-Exome Sequencing in Developmental and Congenital Glaucoma. PLoS ONE 11, e0159259. https://doi.org/10.1371/journal.pone.0159259

Mills, E.J., Wu, P., Chong, G., Ghement, I., Singh, S., Akl, E.A., Eyawo, O., Guyatt, G., Berwanger, O., Briel, M., 2011. Efficacy and safety of statin treatment for cardiovascular

disease: a network meta-analysis of 170 255 patients from 76 randomized trials. QJM 104, 109–124. https://doi.org/10.1093/qjmed/hcq165

Mistry, M., 2017. Peak calling with MACS2 [WWW Document]. Introduction to ChIP-Seq using high-performance computing. URL https://hbctraining.github.io/Intro-to-ChIPseq/lessons/05_peak_calling_macs.html (accessed 7.23.21).

Molina-Navarro, M.M., Roselló-Lletí, E., Ortega, A., Tarazón, E., Otero, M., Martínez-Dolz, L., Lago, F., González-Juanatey, J.R., España, F., García-Pavía, P., Montero, J.A., Portolés, M., Rivera, M., 2013. Differential Gene Expression of Cardiac Ion Channels in Human Dilated Cardiomyopathy. PLoS ONE 8, e79792. https://doi.org/10.1371/journal.pone.0079792

Murphy, D., 2002. Gene expression studies using microarrays: principles, problems, and prospects. Advances in Physiology Education 26, 256–270. https://doi.org/10.1152/advan.00043.2002

Mussbacher, M., Schossleitner, K., Kral-Pointner, J.B., Salzmann, M., Schrammel, A., Schmid, J.A., 2022. More than just a monolayer: The multifaceted role of endothelial cells in the pathophysiology of atherosclerosis. Current Atherosclerosis Reports 24, 483–492. https://doi.org/10.1007/s11883-022-01023-9

Nasmyth, K., Haering, C.H., 2009. Cohesin: Its Roles and Mechanisms. Annual Review of Genetics 43, 525–558. https://doi.org/10.1146/annurev-genet-102108-134233

Neish, A.S., Read, M.A., Thanos, D., Pine, R., Maniatis, T., Collins, T., 1995. Endothelial Interferon Regulatory Factor 1 Cooperates with NF-κB as a Transcriptional Activator of Vascular Cell Adhesion Molecule. Molecular and cellular biology. 15, 5. https://doi.org/10.1128/MCB.15.5.2558

Nelson, R. E., Fessler, L. I., Takagi, Y., Blumberg, B., Keene, D. R., Olson, P. F., Parker, C. G., Fessler, J. H., 1994. Peroxidasin: a novel enzyme-matrix protein of Drosophila development. The EMBO Journal 13, 3438–3447. https://doi.org/10.1002/j.1460-2075.1994.tb06649.x

Nguyen, T., Huang, H.C., Pickett, C.B., 2000. Transcriptional Regulation of the Antioxidant Response Element. Journal of Biological Chemistry 275, 15466–15473. https://doi.org/10.1074/jbc.M000361200

Nicin, L., Abplanalp, W.T., Mellentin, H., Kattih, B., Tombor, L., John, D., Schmitto, J.D., Heineke, J., Emrich, F., Arsalan, M., Holubec, T., Walther, T., Zeiher, A.M., Dimmeler, S., 2020. Cell type-specific expression of the putative SARS-CoV-2 receptor ACE2 in human hearts. European Heart Journal 41, 1804–1806. https://doi.org/10.1093/eurheartj/ehaa311

Niessen, K., Karsan, A., 2008. Notch Signaling in Cardiac Development. Circulation Research 102, 1169–1181. https://doi.org/10.1161/CIRCRESAHA.108.174318

Ogah, O.S., Umuerri, E.M., Adebiyi, A., Orimolade, O.A., Sani, M.U., Ojji, D.B., Mbakwem, A.C., Stewart, S., Sliwa, K., 2021. SARS-CoV 2 Infection (Covid-19) and Cardiovascular Disease in Africa: Health Care and Socio-Economic Implications. Global Heart 16, 18. https://doi.org/10.5334/gh.829

Oki, S., 2020. ChIP-Atlas / Documents [WWW Document]. GitHub. URL https://github.com/inutano/chip-atlas (accessed 10.17.21).

Oki, S., Ohta, T., Shioi, G., Hatanaka, H., Ogasawara, O., Okuda, Y., Kawaji, H., Nakaki, R., Sese, J., Meno, C., 2018. ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. EMBO reports 19, e46255. https://doi.org/10.15252/embr.201846255

Park, P.J., 2009. ChIP–seq: advantages and challenges of a maturing technology. Nature Reviews Genetics 10, 669–680. https://doi.org/10.1038/nrg2641

Partlow, B.P., Applegate, M.B., Omenetto, F.G., Kaplan, D.L., 2016. Dityrosine Cross-Linking in Designing Biomaterials. ACS Biomaterials Science & Engineering 2, 2108–2121. https://doi.org/10.1021/acsbiomaterials.6b00454

Paumann-Page, M., Tscheliessnig, R., Sevcnikar, B., Katz, R.-S., Schwartz, I., Hofbauer, S., Pfanzagl, V., Furtmüller, P.G., Obinger, C., 2020. Monomeric and homotrimeric solution structures of truncated human peroxidasin 1 variants. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics 1868, 140249. https://doi.org/10.1016/j.bbapap.2019.07.002

Peebles, K.E., LaFever, K.S., Page-McCaw, P.S., Colon, S., Wang, D., Stricker, A.M., Ferrell, N., Bhave, G., Page-McCaw, A., 2024. Peroxidasin is required for full viability in development and for maintenance of tissue mechanics in adults. Matrix Biology 125, 1–11. https://doi.org/10.1016/j.matbio.2023.11.005

Péterfi, Z., Donkó, Á., Orient, A., Sum, A., Prókai, Á., Molnár, B., Veréb, Z., Rajnavölgyi, É., Kovács, K.J., Müller, V., Szabó, A.J., Geiszt, M., 2009. Peroxidasin Is Secreted and Incorporated into the Extracellular Matrix of Myofibroblasts and Fibrotic Kidney. The American Journal of Pathology 175, 725–735. https://doi.org/10.2353/ajpath.2009.080693

Péterfi, Z., Tóth, Z.E., Kovács, H.A., Lázár, E., Sum, A., Donkó, Á., Sirokmány, G., Shah, A.M., Geiszt, M., 2014. Peroxidasin-like protein: a novel peroxidase homologue in the human heart. Cardiovascular Research 101, 393–399. https://doi.org/10.1093/cvr/cvt256

Peters, J.-M., Tedeschi, A., Schmitz, J., 2008. The cohesin complex and its roles in chromosome biology. Genes and Development 22, 3089–3114. https://doi.org/10.1101/gad.1724308

Poirier, P., Eckel, R.H., 2002. Obesity and cardiovascular disease. Current atherosclerosis reports 4, pp.448-453.0020 https://doi.org/10.1007/s11883-002-0049-8

Popa, C., Netea, M.G., Van Riel, P.L.C.M., Van Der Meer, J.W.M., Stalenhoef, A.F.H., 2007. The role of TNF-α in chronic inflammatory conditions, intermediary metabolism, and cardiovascular risk. Journal of Lipid Research 48, 751–762. https://doi.org/10.1194/jlr.R600021-JLR200

Porter, K.E., Turner, N.A., 2009. Cardiac fibroblasts: At the heart of myocardial remodeling. Pharmacology & Therapeutics 123, 255–278. https://doi.org/10.1016/j.pharmthera.2009.05.002

Pugh, T.J., Kelly, M.A., Gowrisankar, S., Hynes, E., Seidman, M.A., Baxter, S.M., Bowser, M., Harrison, B., Aaron, D., Mahanta, L.M., Lakdawala, N.K., McDermott, G., White, E.T., Rehm, H.L., Lebo, M., Funke, B.H., 2014. The landscape of genetic variation in dilated cardiomyopathy as surveyed by clinical DNA sequencing. Genetics in Medicine 16, 601–608. https://doi.org/10.1038/gim.2013.204

Quackenbush, J., 2002. Microarray data normalization and transformation. Nature Genetics 32, 496–501. https://doi.org/10.1038/ng1032

R Core Team, 2023. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Raychaudhuri, S., Stuart, J.M., Altman, R.B., 1999. Principal components analysis to summarize microarray experiments: application to sporulation time series, in: Biocomputing 2000. Presented at the Proceedings of the Pacific Symposium, World scientific, Honolulu, Hawaii, USA, pp. 455–466. https://doi.org/10.1142/9789814447331_0043

Ren, B., Zhang, Y., Liu, S., Cheng, X., Yang, X., Cui, X., Zhao, X., Zhao, H., Hao, M., Li, M., Tie, Y., Qu, L., Li, X., 2020. Curcumin alleviates oxidative stress and inhibits apoptosis in diabetic cardiomyopathy via Sirt1-Foxo1 and PI3K-Akt signalling pathways. Journal of Cellular and Molecular Medicine 24, 12355–12367. https://doi.org/10.1111/jcmm.15725

Ridker, P.M., Danielson, E., Fonseca, F.A.H., Genest, J., Gotto, A.M., Kastelein, J.J.P., Koenig, W., Libby, P., Lorenzatti, A.J., MacFadyen, J.G., Nordestgaard, B.G., Shepherd, J., Willerson, J.T., Glynn, R.J., 2008. Rosuvastatin to Prevent Vascular Events in Men and Women with Elevated C-Reactive Protein. New England journal of medicine 359, 2195–2207. https://doi.org/10.1056/NEJMoa0807646

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., Smyth, G.K., 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 43, e47–e47. https://doi.org/10.1093/nar/gkv007

Rosenbaum, A.N., Agre, K.E., Pereira, N.L., 2020. Genetics of dilated cardiomyopathy: practical implications for heart failure management. Nature Reviews Cardiology 17, 286–297. https://doi.org/10.1038/s41569-019-0284-0

Sarziputtini, P., Atzeni, F., Shoenfeld, Y., Ferraccioli, G., 2005. TNF-alpha, rheumatoid arthritis, and heart failure: a rheumatological dilemma. Autoimmunity Reviews 4, 153–161. https://doi.org/10.1016/j.autrev.2004.09.004

Sattler, S., Kennedy-Lydon, T. (Eds.), 2017. The Immunology of Cardiovascular Homeostasis and Pathology, Advances in Experimental Medicine and Biology. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-57613-8

Schena, M., 1996. Genome analysis with gene expression microarrays. BioEssays 18, 427–431. https://doi.org/10.1002/bies.950180513

Schena, M., Shalon, D., Davis, R.W., Brown, P.O., 1995. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. Science 270, 467–470.

Schumacher, J.A., Bloomekatz, J., Garavito-Aguilar, Z.V., Yelon, D., 2013. tal1 regulates the formation of intercellular junctions and the maintenance of identity in the endocardium. Developmental Biology 383, 214–226. https://doi.org/10.1016/j.ydbio.2013.09.019

Sciarretta, S., Forte, M., Frati, G., Sadoshima, J., 2018. New Insights Into the Role of mTOR Signaling in the Cardiovascular System. Circulation research 122, 489–505. https://doi.org/10.1161/CIRCRESAHA.117.311147

Senoner, T., Dichtl, W., 2019. Oxidative stress in cardiovascular diseases: Still a therapeutic target? Nutrients 11. https://doi.org/10.3390/nu11092090

Seo, D., Wang, T., Dressman, H., Herderick, E.E., Iversen, E.S., Dong, C., Vata, K., Milano, C.A., Rigat, F., Pittman, J., Nevins, J.R., West, M., Goldschmidt-Clermont, P.J., 2004. Gene Expression Phenotypes of Atherosclerosis. Arteriosclerosis, thrombosis, and vascular biology 24, 1922–1927. https://doi.org/10.1161/01.ATV.0000141358.65242.1f

Sethi, G., 2008. TNF: A master switch for inflammation to cancer. Front Biosci 13, 5094–5107. https://doi.org/10.2741/3066

Sevcnikar, B., Schaffner, I., Chuang, C.Y., Gamon, L., Paumann-Page, M., Hofbauer, S., Davies, M.J., Furtmüller, P.G., Obinger, C., 2020. The leucine-rich repeat domain of human peroxidasin 1 promotes binding to laminin in basement membranes. Archives of Biochemistry and Biophysics 689, 108443. https://doi.org/10.1016/j.abb.2020.108443Shaulian, E., Karin, M., 2002. AP-1 as a regulator of cell life and death. Nature cell biology 4, E131–E136. https://doi.org/10.1038/ncb0502-e131

Shi, R., Hu, C., Yuan, Q., Yang, T., Peng, J., Li, Y., Bai, Y., Cao, Z., Cheng, G., Zhang, G., 2011. Involvement of vascular peroxidase 1 in angiotensin II-induced vascular smooth muscle cell proliferation. Cardiovascular Research 91, 27–36. https://doi.org/10.1093/cvr/cvr042

Singh, V.P., Pinnamaneni, J.P., Pugazenthi, A., Sanagasetti, D., Mathison, M., Martin, J.F., Yang, J., Rosengart, T.K., 2021. Hippo Pathway Effector Tead1 Induces Cardiac Fibroblast to Cardiomyocyte Reprogramming. Journal of the American Heart Association10, e022659. https://doi.org/10.1161/JAHA.121.022659

Sitole, B.N., Mavri-Damelin, D., 2018. Peroxidasin is regulated by the epithelial-mesenchymal transition master transcription factor Snai1. Gene 646, 195–202. https://doi.org/10.1016/j.gene.2018.01.011

Slowikowski, 2024. ggrepel: Automatically Position Non-Overlapping Text Labels with “ggplot2”. https://ggrepel.slowkow.com/, https://github.com/slowkow/ggrepel (accessed 10.6.2023).

Smale, S.T., Kadonaga, J.T., 2003. The RNA Polymerase II Core Promoter. Annual review of biochemistry 72, 449–479. https://doi.org/10.1146/annurev.biochem.72.121801.161520

Smyth, G.K., 2005. Linear Models for Microarray Data User’s Guide, in: Ritchie, M., Thorne, N., Wettenhall, J. (Eds.), Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer, New York, pp. 397–420.

Smyth, G.K., 2004. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. Statistical Applications in Genetics and Molecular Biology 3, 1–25. https://doi.org/10.2202/1544-6115.1027

Smyth, G.K., Speed, T., 2003. Normalization of cDNA microarray data. Methods 31, 265–273. https://doi.org/10.1016/S1046-2023(03)00155-5

Song, J., Liu, Q., Tang, H., Tao, A., Wang, H., Kao, R., Rui, T., 2016. Activation of PI3Kγ/Akt pathway increases cardiomyocyte HMGB1 expression in diabetic environment. Oncotarget 7, 80803–80810. https://doi.org/10.18632/oncotarget.13096

Song, S., Zhang, Xiaokai, Huang, Z., Zhao, Y., Lu, S., Zeng, L., Cai, F., Wang, T., Pei, Z., Weng, X., Luo, W., Lu, H., Wei, Z., Wu, J., Yu, P., Shen, L., Zhang, Xiaochun, Sun, A., Ge, J., 2024. TEA domain transcription factor 1(TEAD1) induces cardiac fibroblasts cells remodeling through BRD4/Wnt4 pathway. Signal Transduction and Targeted Therapy 9, 45. https://doi.org/10.1038/s41392-023-01732-w

Sonmez, O., Sonmez, M., 2017. Role of platelets in immune system and inflammation. Porto Biomedical Journal 2, 311–314. https://doi.org/10.1016/j.pbj.2017.05.005

Soudi, M., Zamocky, M., Jakopitsch, C., Furtmüller, P.G., Obinger, C., 2012. Molecular Evolution, Structure, and Function of Peroxidasins. Chemistry & Biodiversity 9, 1776–1793. https://doi.org/10.1002/cbdv.201100438

Sowers, J.R., Epstein, M., Frohlich, E.D., 2001. Diabetes, Hypertension, and Cardiovascular Disease: An Update. Hypertension 37, 1053–1059. https://doi.org/10.1161/01.HYP.37.4.1053

Spencer, V.A., Sun, J.-M., Li, L., Davie, J.R., 2003. Chromatin immunoprecipitation: a tool for studying histone acetylation and transcription factor binding. Methods, Histone Modifications 31, 67–75. https://doi.org/10.1016/S1046-2023(03)00089-6

Stancu, C., Sima, A., 2001. Statins: mechanism of action and effects. Journal of cellular and molecular medicine 5, 378–387. https://doi.org/10.1111/j.1582-4934.2001.tb00172.x

Stitziel, N.O., Mar, B.G., Liang, J., Westbrook, C.A., 2004. Membrane-Associated and Secreted Genes in Breast Cancer. Cancer Research 64, 8682–8687. https://doi.org/10.1158/0008-5472.CAN-04-1729

Stratton, M.S., Bagchi, R.A., Felisbino, M.B., Hirsch, R.A., Smith, H.E., Riching, A.S., Enyart, B.Y., Koch, K.A., Cavasin, M.A., Alexanian, M., Song, K., Qi, J., Lemieux, M.E., Srivastava, D., Lam, M.P.Y., Haldar, S.M., Lin, C.Y., McKinsey, T.A., 2019. Dynamic Chromatin Targeting of BRD4 Stimulates Cardiac Fibroblast Activation. Circulation Research 125, 662–677. https://doi.org/10.1161/CIRCRESAHA.119.315125

Tak, P.P., Firestein, G.S., 2001. NF-κB: a key role in inflammatory diseases. The Journal of clinical investigation 107, 7–11. https://doi.org/10.1172/JCI11830

Thiery, J.P., Acloque, H., Huang, R.Y.J., Nieto, M.A., 2009. Epithelial-Mesenchymal Transitions in Development and Disease. Cell 139, 871–890. https://doi.org/10.1016/j.cell.2009.11.007

Toth, P.P., Banach, M., 2019. Statins: Then and Now. Methodist DeBakey Cardiovascular Journal 15, 23. https://doi.org/10.14797/mdcj-15-1-23

van den Borne, S.W.M., Diez, J., Blankesteijn, W.M., Verjans, J., Hofstra, L., Narula, J., 2010. Myocardial remodeling after infarction: the role of myofibroblasts. Nat Rev Cardiol 7, 30–37. https://doi.org/10.1038/nrcardio.2009.199

Viemann, D., Goebeler, M., Schmid, S., Nordhues, U., Klimmek, K., Sorg, C., Roth, J., 2006. TNF induces distinct gene expression programs in microvascular and macrovascular human endothelial cells. Journal of Leukocyte Biology 80, 174–185. https://doi.org/10.1189/jlb.0905530

Wang, E.-L., Jia, M.-M., Luo, F.-M., Li, T., Peng, J.-J., Luo, X.-J., Song, F.-L., Yang, J.-F., Peng, J., Liu, B., 2019. Coordination between NADPH oxidase and vascular peroxidase 1 promotes dysfunctions of endothelial progenitor cells in hypoxia-induced pulmonary hypertensive rats. European Journal of Pharmacology 857, 172459. https://doi.org/10.1016/j.ejphar.2019.172459

Wang, H., Naghavi, M., Allen, C., Barber, R.M., Bhutta, Z.A., Carter, A., Casey, D.C., Charlson, F.J., Chen, A.Z., Coates, M.M., Coggeshall, M., Dandona, L., Dicker, D.J., Erskine, H.E., Ferrari, A.J., Fitzmaurice, C., Foreman, K., Forouzanfar, M.H., Fraser, M.S., Fullman, N., *et al.*, 2016. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015. The Lancet 388, 1459–1544. https://doi.org/10.1016/S0140-6736(16)31012-1

Weinmann, A.S., Farnham, P.J., 2002. Identification of unknown target genes of human transcription factors using chromatin immunoprecipitation. Methods 26, 37–47. https://doi.org/10.1016/S1046-2023(02)00006-3

Wells, J., Farnham, P.J., 2002. Characterizing transcription factor binding sites using formaldehyde crosslinking and immunoprecipitation. Methods 26, 48–56. https://doi.org/10.1016/S1046-2023(02)00007-5

WHO, 2021a. Cardiovascular diseases (CVDs) [WWW Document]. URL https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (accessed 6.13.21).

WHO, 2021b. Diabetes [WWW Document]. URL https://www.who.int/news-room/fact-sheets/detail/diabetes (accessed 6.14.21).

Wickham, H., 2023. stringr: Simple, Consistent Wrappers for Common String Operations. https://github.com/tidyverse/stringr (accessed 3.1.2023).

Wickham, H., 2011. ggplot2. WIREs Computational Stats 3, 2. 180–185. https://doi.org/10.1002/wics.147

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms,

J., Robinson, D., Seidel, D., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H., 2019. Welcome to the Tidyverse. JOSS 4, 1686. https://doi.org/10.21105/joss.01686

Wickham, H., François, R., Henry, L., Müller, K., Vaughan, D., 2023. dplyr: A Grammar of Data Manipulation. R package version 1.1.2. https://github.com/tidyverse/dplyr (accessed 9.5.23).

Wynn, T., 2008. Cellular and molecular mechanisms of fibrosis. The Journal of Pathology 214, 199–210. https://doi.org/10.1002/path.2277

Yang, M., Yan, J., Wu, A., Zhao, W., Qin, J., Pogwizd, S.M., Wu, X., Yuan, S., Ai, X., 2021. Alterations of housekeeping proteins in human aged and diseased hearts. Pflügers Archiv-European Journal of Physiology 473, 351–362. https://doi.org/10.1007/s00424-021-02538-x

Yang, X., Li, X., Lin, Q., Xu, Q., 2019. Up-regulation of microRNA-203 inhibits myocardial fibrosis and oxidative stress in mice with diabetic cardiomyopathy through the inhibition of PI3K/Akt signaling pathway via PIK3CA. Gene 715, 143995. https://doi.org/10.1016/j.gene.2019.143995

Yang, Y., Xing, Y., Liang, C., Hu, L., Xu, F., Mei, Q., 2016. An examination of the regulatory mechanism of Pxdn mutation-induced eye disorders using microarray analysis. International Journal of Molecular Medicine 37, 1449–1456. https://doi.org/10.3892/ijmm.2016.2572

Yang, Y.H., 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Research 30, 15e–115. https://doi.org/10.1093/nar/30.4.e15

Yndestad, A., Kristian Damås, J., Øie, E., Ueland, T., Gullestad, L., Aukrust, P., 2006. Systemic inflammation in heart failure – The whys and wherefores. Heart Fail Rev 11, 83–92. https://doi.org/10.1007/s10741-006-9196-2

Yoshida, T., Yamashita, M., Horimai, C., Hayashi, M., 2014. Deletion of Krüppel-Like Factor 4 in Endothelial and Hematopoietic Cells Enhances Neointimal Formation Following Vascular Injury. JAHA 3, e000622. https://doi.org/10.1161/JAHA.113.000622

Yu, G., Wang, L.-G., He, Q.-Y., 2015. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics 31, 2382–2383. https://doi.org/10.1093/bioinformatics/btv145

Yun, H., Park, S., Kim, M., Yang, W.K., Im, D.U., Yang, K.R., Hong, J., Choe, W., Kang, I., Kim, S.S., Ha, J., 2014. AMP -activated protein kinase mediates the antioxidant effects of resveratrol through regulation of the transcription factor FoxO1. The FEBS Journal 281, 4421–4438. https://doi.org/10.1111/febs.12949

Zazo-Seco, C., Plaisancié, J., Bitoun, P., Corton, M., Arteche, A., Ayuso, C., Schneider, A., Zafeiropoulou, D., Gilissen, C., Roche, O., Frémont, F., Calvas, P., Slavotinek, A., Ragge, N., Chassaing, N., 2020. Novel PXDN biallelic variants in patients with microphthalmia and anterior segment dysgenesis. Journal of Human Genetics 65, 487–491. https://doi.org/10.1038/s10038-020-0726-x

Zhang, H., Park, Y., Wu, J., Chen, X. ping, Lee, S., Yang, J., Dellsperger, K.C., Zhang, C., 2009. Role of TNF-α in vascular dysfunction. Clinical Science 116, 219–230. https://doi.org/10.1042/CS20080196

Zhang, Ji, Chang, L., Chen, C., Zhang, M., Luo, Y., Hamblin, M., Villacorta, L., Xiong, J.-W., Chen, Y.E., Zhang, Jifeng, Zhu, X., 2011. Rad GTPase inhibits cardiac fibrosis through connective tissue growth factor. Cardiovascular Research 91, 90–98. https://doi.org/10.1093/cvr/cvr068

Zhang, X., Hu, C., Yuan, Y.-P., Song, P., Kong, C.-Y., Wu, H.-M., Xu, S.-C., Ma, Z.-G., Tang, Q.-Z., 2021. Endothelial ERG alleviates cardiac fibrosis via blocking endothelin-1-dependent paracrine mechanism. Cell Biol Toxicol. https://doi.org/10.1007/s10565-021-09581-5

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W., Liu, X.S., 2008. Model-based Analysis of ChIP-Seq (MACS). Genome Biol 9, R137. https://doi.org/10.1186/gb-2008-9-9-r137

Zhang, Y.-Z., Wang, L., Zhang, J.-J., Xiong, X.-M., Zhang, D., Tang, X.-M., Luo, X.-J., Ma, Q.-L., Peng, J., 2018. Vascular peroxide 1 promotes ox-LDL-induced programmed necrosis in endothelial cells through a mechanism involving β-catenin signaling. Atherosclerosis 274, 128–138. https://doi.org/10.1016/j.atherosclerosis.2018.04.031

Zheng, Y.-Z., Liang, L., 2018. High expression of PXDN is associated with poor prognosis and promotes proliferation, invasion as well as migration in ovarian cancer. Annals of Diagnostic Pathology 34, 161–165. https://doi.org/10.1016/j.anndiagpath.2018.03.002

Zhou, X., Sun, Q., Xu, C., Zhou, Z., Chen, X., Zhu, X., Huang, Z., Wang, W., Shi, Y., 2022. A systematic pan-cancer analysis of PXDN as a potential target for clinical diagnosis and treatment. Front. Oncol. 12, 952849. https://doi.org/10.3389/fonc.2022.952849

APPENDIX

Validation dataset for DGE analysis

This appendix includes the QC and DGE results for the GSE1400 dataset, as well as a table summarising the function parameters used for each QC tool. This dataset investigated differences in transcript concentrations between the cytosol and membrane of MCF-7 cells. *PXDN* is known to be expressed in MCF-7 cells and was expected to localise to the membrane.

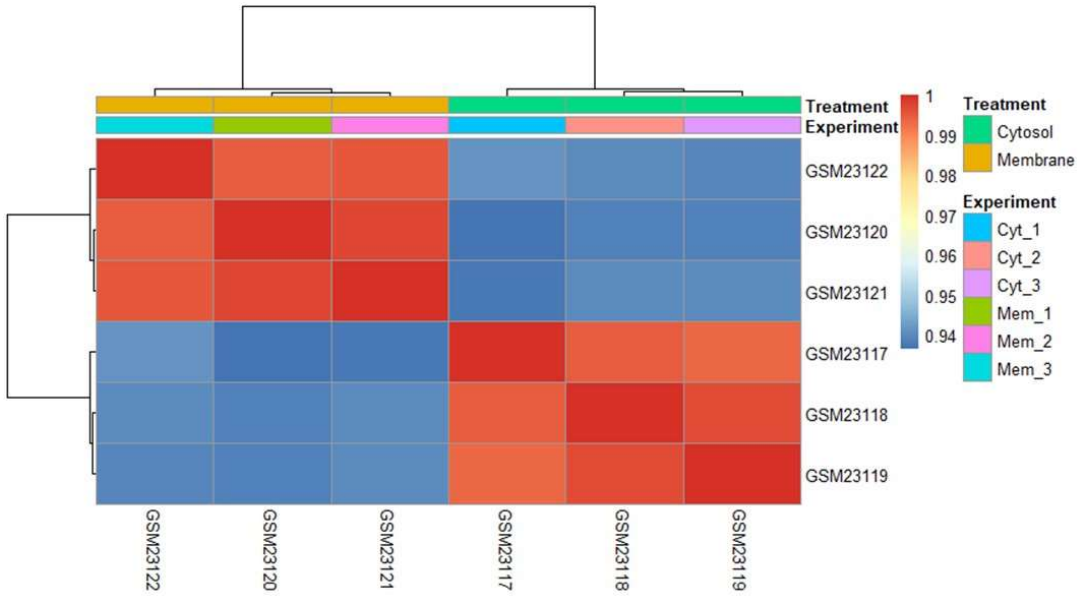


Figure A.1 Heatmap and hierarchical clustering showing the consistency within and between sample groups. Clear separation of treatment groups with no obvious outliers.

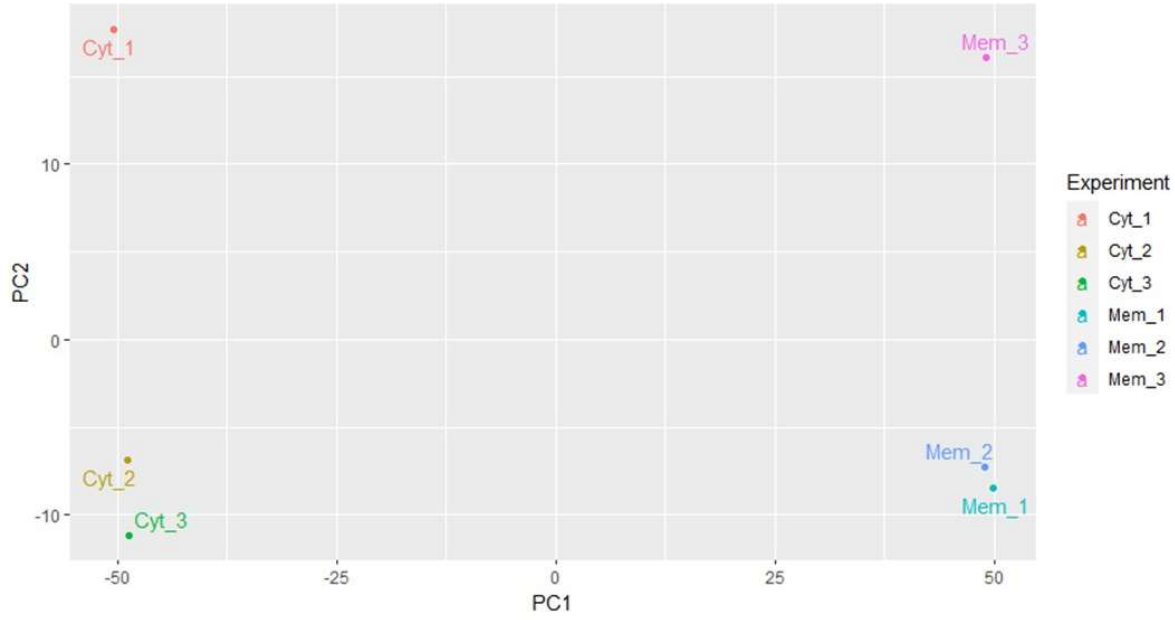


Figure A.2 PCA plot showing clustering of sample groups along PC1. Cyt_1 and Mem_3 were noted as potential outliers, but given the low contribution of PC2 were retained

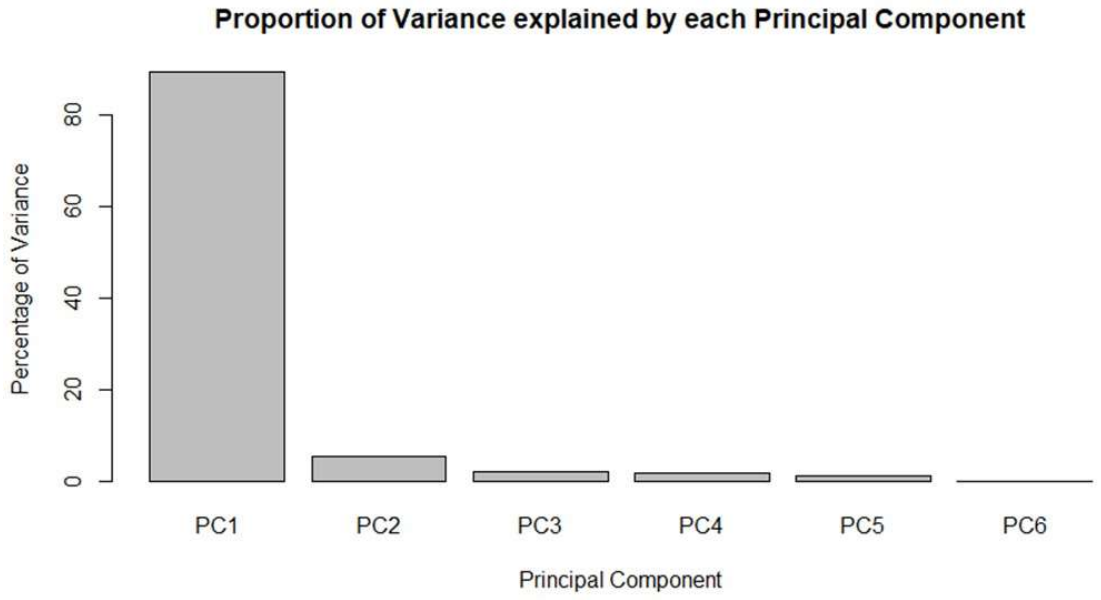


Figure A.3 Contribution of each principal component to observed variances. Notably, PC1 had the largest contribution, over 80%. Thus outliers in the PCA were only given minor penalties for separating along PC2.

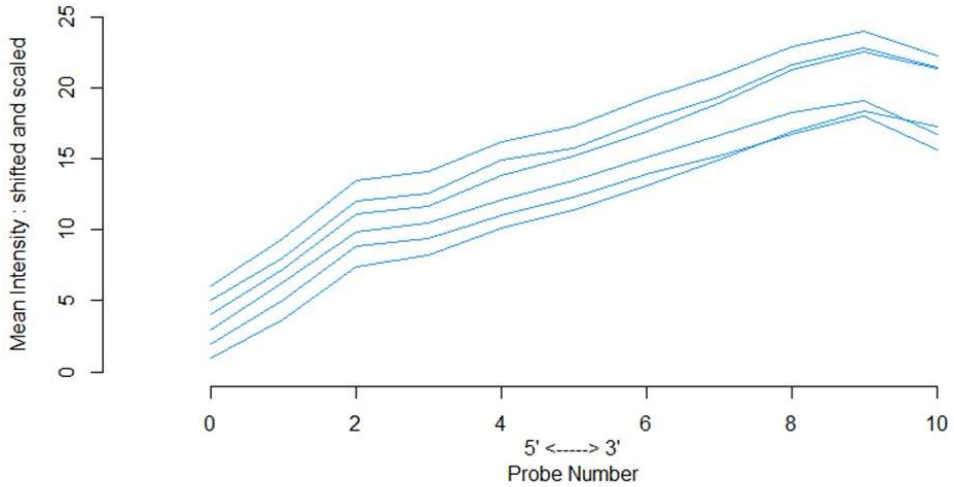


Figure A.4 RNA degradation plot to assess RNA quality of each sample.

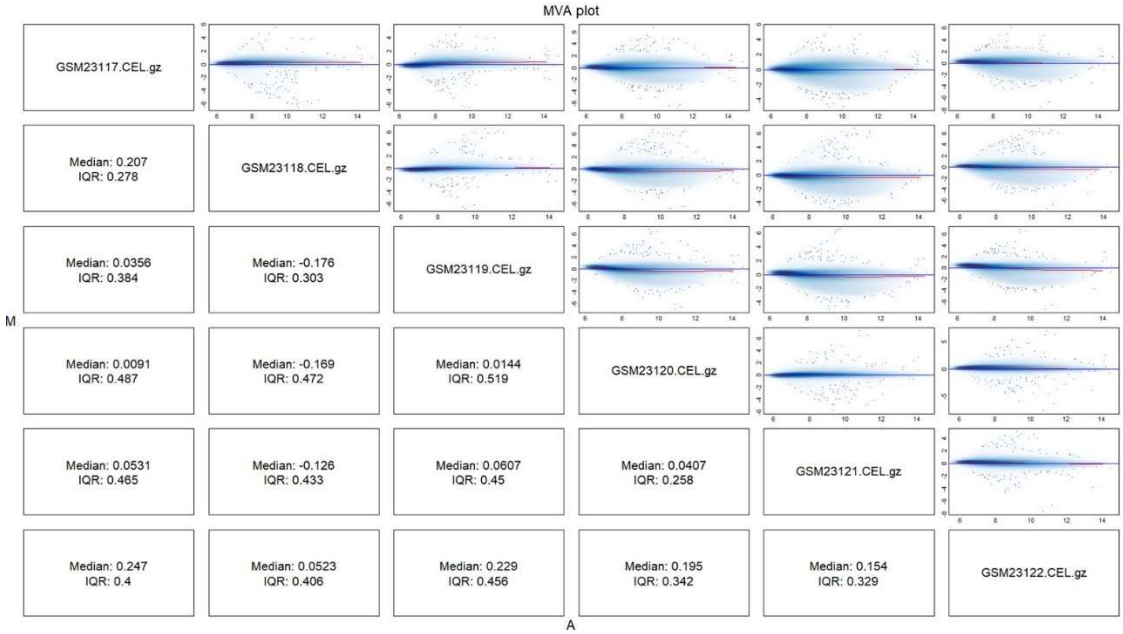


Figure A.5 MA plots to compare fluorescence values of probes between different samples.

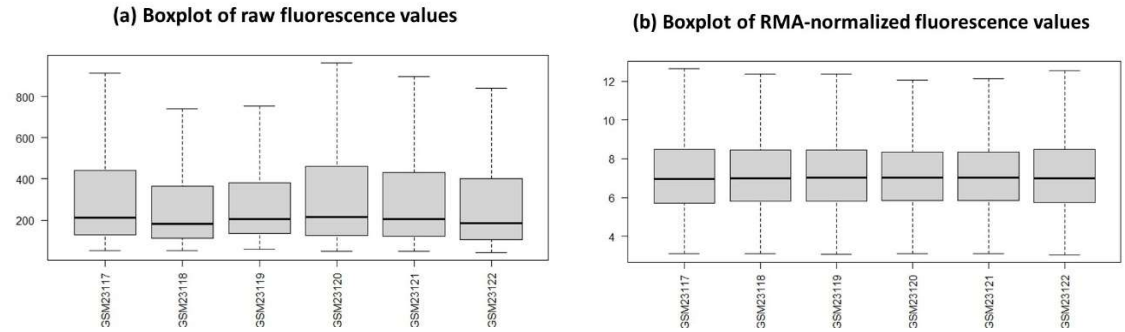


Figure A.6 Boxplots of fluorescence values for GSE1400. (a) raw fluorescence values for all probes. (b) RMA-normalized fluorescence values. Cell line is shown to produce stable fluorescence values across samples even before normalization. However, RMA did show a marked improvement in consistency across samples.

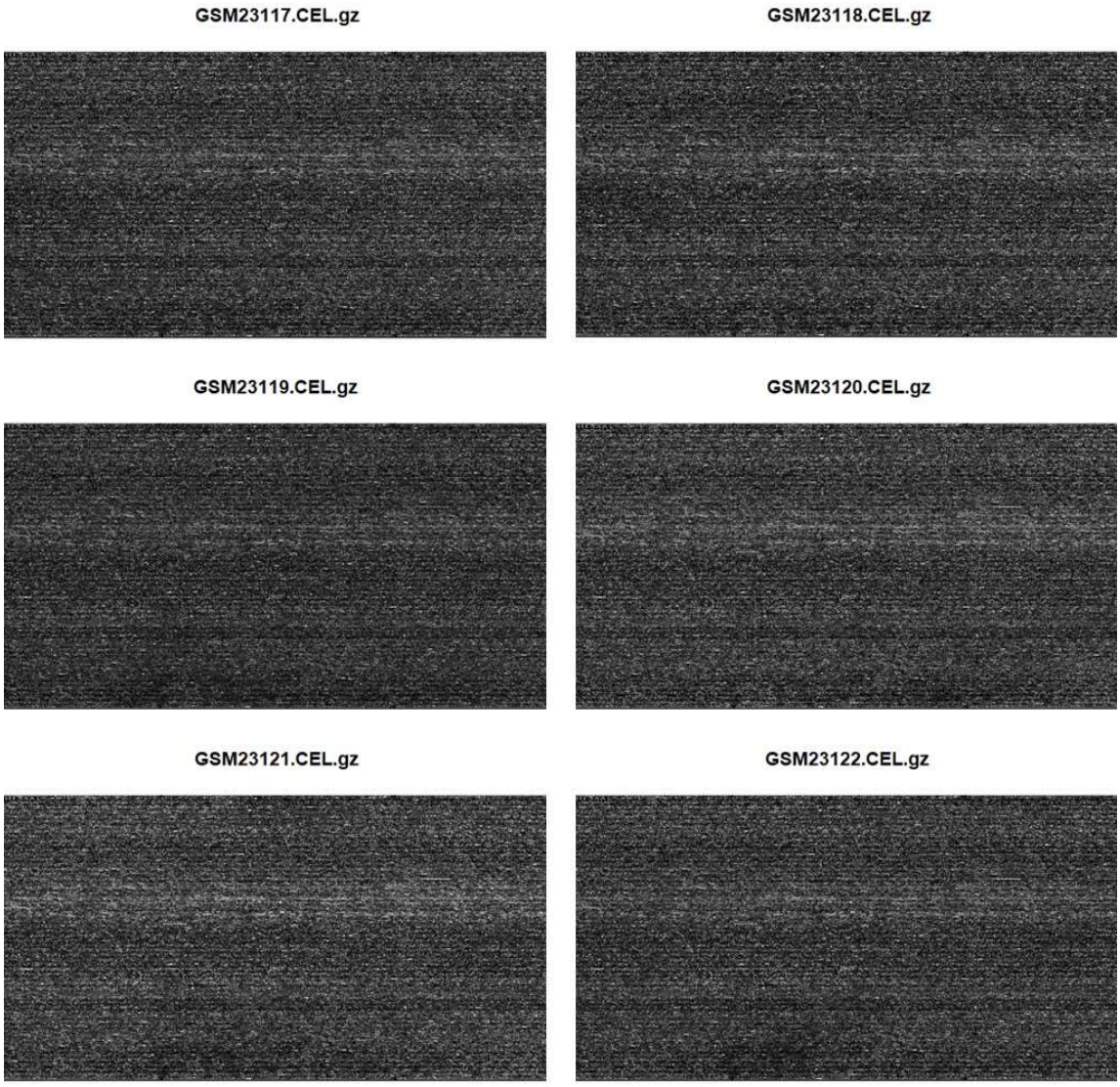


Figure A.7 Micrographs of microarray samples generated from the GSE1400 dataset.

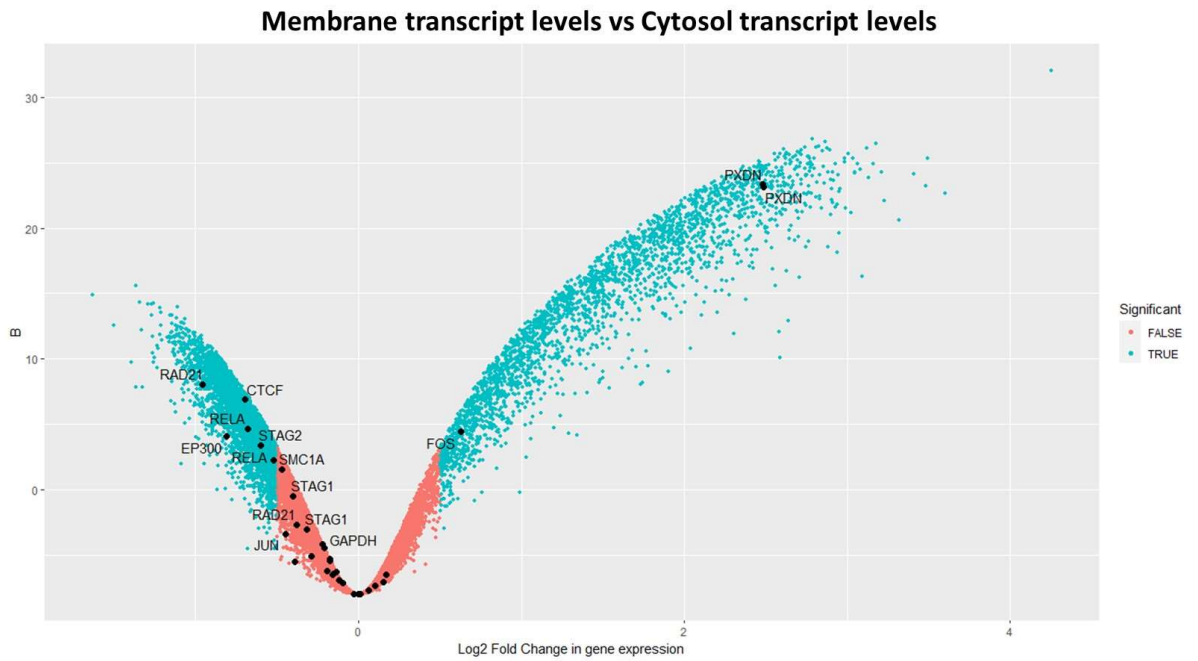


Figure A.8 Volcano plot showing the number of differentially expressed genes in the membrane fraction of MCF-7s compared to the cytosolic fraction. As expected, *PXDN* was present in higher concentrations in the membrane, while many of its regulatory TFs were unchanged or localized more to the cytosolic fraction.

Table A.1 Summary of function parameters used for Exploratory Data Analysis

EDA visualisation	Parameters
Heatmaps	pheatmap(corMatrix, annotation_col = sample_info), clustering_distance_rows = "euclidean", clustering_distance_cols = "euclidean", clustering_method = "complete"
PCA Plots	ggplot(aes(x = PC1, y = PC2, col = Treatment, label = paste("", Experiment))) + geom_point() + geom_text_repel(), labs(title = ""), theme(plot.title = element_text(hjust = 0.5))
PCA Bar Plots	barplot(pca_proportions_percent, main = "Proportion of Variance explained by each Principal Component", xlab = "Principal Component", ylab = "Percentage of Variance")
RNA Degradation Plots	plotAffyRNAdeg(deg)
MA Plots	MAplot(raw_data, pairs = TRUE, plot.method = "smoothScatter", names = plotting_col_names)
Expression Boxplots	boxplot(exprs(raw_data), outline = F, las = 2, names = sample_info$Experiment), boxplot(rma_normalised_expr, outline = F, las = 2, names = sample_info$Experiment)
Image Plots	image(raw_data)