RNA Sequencing of Non-coding RNAs in Ischaemic Heart Disease

A thesis submitted for the Degree of

Doctor of Philosophy

Zoe Ward

At University of Otago, Christchurch,

New Zealand

2020

Abstract

Ischaemic heart disease is a major cause of death worldwide and a leading cause of mortality and morbidity in New Zealand. Older adults and those of Māori and Pacific ancestry are particularly affected. Ischaemic heart disease accounts for over half of all cardiovascular disease mortality and, again, rates are more than twice as high among Māori than non-Māori. Ischaemic heart disease can lead to myocardial infarction (heart attack) which, if not fatal, can then lead to heart failure, a complex, multifactorial disease characterised by neurohormonal signalling and remodelling of the heart. Currently the natriuretic peptides are the international gold standard for diganosing heart failure and are also excellent prognostic markers in patients with heart failure. However, there is still a clinical need for early biomarkers of myocardial ischaemia (to identify people at risk of myocardial infarction) and to identify patients at risk of developing heart failure before detrimental remodelling has occurred.

As sequencing technologies have evolved there has been intense research in the fields of circulating cell free DNA and RNA, especially non-coding RNA. As RNA is actively transcribed, it has the advantage of providing a 'real time' insight into the disease status of an individual. Recent discoveries have highlighted the regulatory roles and diseases associated with non-coding RNAs, including long non-coding RNAs (lncRNAs) and circular RNAs (circRNAs). LncRNAs have been demonstrated to have mutliple functional roles both within the nucleus and cytoplasm such as chromatin remodelling, histone modification, transcription factor recruitment, formation of subnuclear structures and control of mRNA translation and decay. CircRNA, a relative newcomer, has also been demonstrated to have functional roles such as sequestering miRNAs, binding proteins and even coding for peptides. There is great excitment for the potential utility of circRNAs as biomarkers as, due to their circular structure, they are more resistant to degradation in the circulation than their linear RNA counterparts.

The overall aim of this thesis was to identify non-coding RNAs associated with ischaemic heart disease. To address this aim, a bioinformatics pipeline was developed to identify mRNAs, lncRNAs including putative novel lncRNAs, and circRNAs using short-read RNA Sequencing (RNA-Seq) data. This pipeline was tested and validated with publicly available data and used to screen for candidate mRNA and lncRNA biomarkers associated with ischaemic heart disease in human heart tissue. A whole genome network correlation approach identified several promising candidate biomarkers for myocardial ischaemia including several

novel lncRNAs, which were validated with long-read Nanopore sequencing in independent samples. The sub-cellular localisation of three promising lncRNAs candidates (two annotated lncRNAs, one novel lncRNA) was identified using the in-situ hybridisation assay, RNAscope®. Next, an RNA-Seq protocol was developed to detect mRNAs, lncRNAs and circRNAs in human plasma. This protocol was applied to plasma from patients with ischaemic heart disease and healthy controls to screen for candidate mRNA, lncRNA and circRNA biomarkers for progression from ischaemic heart disease to heart failure. Although candidate biomarkers for disease progression could not be detected in these patients several additional lncRNA candidates for the presence of ischaemic heart disease were identified.

In summary, this study has established a bioinformatics pipeline and methodology for identifying and validating putative novel lncRNAs and circRNAs in human tissue and plasma. This work has identified several promising candidate lncRNA biomarkers for ischaemic heart disease, which, if validated, may provide early diagnostic information in high-risk patients.

The pipeline is freely available to download at https://github.com/zoeward-nz/PhD

Acknowledgements

This thesis would have not been possible without the expertise and generosity of many people whom I would like to take this opportunity to thank.

First and foremost I would like to thank my primary supervisory Dr Anna Pilbrow. You saw potential in me that I didn't see in myself and convinced me that I had the qualities to make a good PhD candidate. You have never once turned me away when I have approached you with a question, for a brainstorming session or for advice - be that PhD advice or life advice. Not only did you encourage me to embark on this crazy journey but you have been there throughout, giving your guidance and I will be forever grateful to you that you gave me the belief that I could complete the journey.

I would also like to thank Professor Vicky Cameron who was also extremely generous with her time of which she has precious little of as she is in such high demand. I have really appreciated your open door policy and again, always made me feel welcome to ask a question or seek advice at any time. I have also enjoyed our chats about politics, films and books over tea.

Many thanks to my other supervisors also - Associate Professor John Pearson for your help when Janice (the computer) was being difficult and for your statistics expertise and Dr Sebastian Schmeier who helped enormously with the pipeline implementation.

Thanks to Dr Jochen D. Muehlschlegel and Dr Simon Body for the generous sharing of your rare dataset which makes up a whole chapter of this thesis.

A big thankyou to Heart Foundation of New Zealand for the scholarship that enabled me to carry out this work.

Thankyou to Dr Allamanda Faatoese for many a word of encouragment in the office in the difficult times and apologies for any expletives I may have shouted at my computer.

To all the members of the CHI especially the CHI study coordinators for collecting and meticulously recording the plasma samples. Thanks also goes to Dr Arthur Morley-Bunker for his help with RNAscope, Allison Miller for her Nanopore advice and Dr Aaron Jeffs at Otago Genomics for making probably the trickiest RNA Sequencing library ever made.

To all of the cohort patients around New Zealand who have very generously donated their time and plasma so that this, as well as many other studies, can go ahead. Without you the Scientists have no data to work on.

To my parents, Gloria and John who have supported me my whole life even though your are not quite sure what it is I do. Thanks to the rest of my family for your support and maybe by some miracle we can all be at the graduation.

Finally, to my amazing partner George and Shezzie dog melter of hearts who have been on this crazy journey every second of every day for the past 3.5 years. You have picked me up when I needed it and gave me the encouragement and strength to carry on. This PhD is as much yours George as it is mine and it goes without saying I could not have done it without you. Now you know all about non-coding RNA.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AAV9	Adenovirus9
ACC	American College of Cardiology
ACE	Angiotensin-converting enzyme
ACS	Acute coronary syndrome
ADAR	Adenosine deaminases acting on RNA
AGO2	Argonaute 2
AHA	American Heart Association
ANOVA	Analysis of variance
ANP	Atrial natriuretic peptide
ARBs	Angiotensin-receptor blockers
AS	Atherosclerosis
ASO	Antisense oligonucleotides
AVP	Arginine vasopressin
BACE	β-site APP-cleaving enzyme 1
BMI	Body mass index
BNP	Brain natriuretic peptide
BSJ	Back-splice junction
CAD	Coronary Artery Disease
CaMKIIδ	Ca/calmodulin dependent protein kinase Iiδ
CDCS	Coronary Heart Disease Cohort Study
cDNA	Complementary DNA
CDS	Coding regions
ceRNA	Competitive endogenous RNA
CFs	Cardiac fibroblasts
CHI	Christchurch Heart Institute
ChIP	Chromatin immunoprecipitation
ChIRP	Chromatin Isolation by RNA Purification
circRNA	Circular RNA
ciRNAs	Intronic circRNAs
CLIP	Crosslinking-immunoprecipitation
Col1a2	Collagen alpha 2
CPAT	Coding Potential Assessment Tool
CRISPR	Clustered regularly interspaced short palindromic repeats
CRP	C-reactive protein
CTCF	CCCTC-binding transcription factor
CTGF	Connective tissue growth factor
cTnI	Cardiac specific troponin I
cTnT	Cardiac specific troponin T
DAMPs	Damage-associated molecular patterns
DNA	Deoxyribonucleic acid
DOX	Doxorubicin
ECM	Extra cellular matrix
Ecs	Endothelial cells
EF	Ejection fraction
EIcircRNAs	Exon and intron containing circRNAs
eIF3	Eukaryotic initiation factor 3

ENCODE	Encyclopedia of DNA Elements
eQTLs	Expression quantitative trait locus
eRNA	Enhancer RNA
ESTs	Expressed sequence tags
FANTOM	Functional Annotation of Mammals
FFPE	Formalin-fixed paraffin-embedded
FISH	Fluorescence in situ hybridization
FPKM	Fragments Per Kilobase of transcripts per Million mapped reads
GFP	Green fluorescent protein
GTF	Gene transfer format
GWAS	Genome-wide association studies
HAECs	Human aorta endothelial cells
HF	Heart Failure
HFpEF	HF with preserved ejection fraction
HFrEF	HF with reduced ejection fraction
HLA	Human leukocyte antigen
HsCRP	High-sensitivity C-reactive protein
HUVECs	Human Umbilical Vein Endothelial Cells
HVOLs	Healthy Volunteers
ICON	International Collaborative of NT-proBNP
IGF2BP3	Insulin-like-growth factor 2 mRNA-binding protein 3
IHD	Ischaemic Heart Disease
IPA	Ingenuity Pathway Analysis
IQR	Inter-Quartile Range
IR	Ischemia reperfusion
IRESs	Internal ribosome entry sites
LDLs	Low-density lipoproteins
lincRNAs	Long intergenic non-coding RNAs
lncRNA	Long non-coding RNA
LV	Left ventricle
m^6A	N6-methyladenosine
MEFs	Mouse embryonic fibroblasts
MF	Myocardial fibrosis
MI	Myocardial Infarction
miRNA	Micro RNA
mRNA	Messenger RNA
mTOR	Mechanistic target of rapamycin (a serine/threonine protein kinase)
NATs	Natural antisense transcripts
NMVCs	Neonatal mouse ventricular cardiomyocytes
NSTEMI	Non-ST-segment elevation MI
NTC	No template control
NT-proBNP	N-terminal proBNP
OGD	Oxygen-glucose-deprivation
OGF	Otago Genomics facility
ONT	Oxford Nanopore Technologies
ORF	Open reading frame
oxLDLs	Oxidized low-density lipoprotein
PBMCs	Peripheral blood mononuclear cell
PCA	Principal Component Analysis
PCC	Pearson Correlation coefficient

piRNAs	Piwi-interacting RNAs
polyA	Polyadenylated
PRC	Polycomb repressor complex
PROMPTs	Promoter upstream transcripts
QKI	Quaking
RAAS	Renin-angiotensin-aldosterone system
RAP	RNA Antisense Purification
RBPs	RNA-binding proteins
RBR	RNA-binding region
RIN	RNA Integrity Number
RIP	RNA immunoprecipitation
RISC	RNA-induced silencing complex
RMRP	RNA processing endoribonuclease
RNA	Ribonucleic acid
RNAPII	RNA polymerase II
RNP	Ribonucleoprotein
ROS	Reactive oxygen species
RPF	Ribosome Protected Fragments
rRNA	Ribosomal RNA
RT-PCR	Reverse transcription-polymerase chain reaction
Seq	Sequencing
shRNAs	Small hairpin RNAs
siRNAs	Short interfering RNAs
SMARTer	Stranded Total RNA-Seq Kit, Pico Input Mammalian
snoRNAs	Small nucleolar RNAs
SNPs	Single nucleotide polymorphisms
SNS	Sympathetic nervous system
sORFs	Short open reading frames
ST2	Soluble Interleukin 1 receptor-like 1
STEMI	ST-segment elevation MI
SUMO	Small Ubiquitin-like Modifier
TADs	Topologically associated domains
tiRNAs	Transcription initiation RNAs
TPM	Transcripts per million
tRNAs	Transfer RNAs
TSS	Transcription start sites
TUNEL	Terminal deoxynucleotidyl transferase-mediated dUTP nick end labelling
UTR	Untranslated region
VSMC	Vascular smooth muscle cells
WGCNA	Whole Genome Correlation Network Analysis

List of Gene Names

αHIF 1A AS RNA 2	Hypoxia inducible factor 1a antisense RNA 2
α-SMA	α-smooth muscle actin
ADCY6	Adenylate cyclase type 6
ALPK3	α-protein kinase 3
ANKRD1	Ankyrin repeat domain 1
ANRIL	Antisense non-coding RNA in the INK4 locus
AQP1	Aquaporin-1
ARC	Activity-regulated cytoskeleton-associated protein
Arl2	ADP ribosylation factor like 2
ATG7	Autophagy related protein 7
BMP10	Bone morphogenetic protein 10
CARL	Cardiac apoptosis-related lncRNA
CDIP1	Cell death-inducing protein
Chaer	Cardiac-hypertrophy-associated epigenetic regulator
CHRF	Cardiac hypertrophy related factor
circNFIB	circ Nuclear factor 1 B-type
COA6	Cytochrome c oxidase assembly factor 6
DANCR	Differentiation Antagonizing Non-Protein Coding RNA
DDX6	DEAD-Box Helicase 6
Dnmt3B	DNA (cytosine-5-)-methyltransferase 3 beta
EMP1	Epithelial membrane protein 1
FEM1A	Fem-1 homolog A
FIRRE	Functional intergenic repeating RNA element
FUS	Fused in Sarcoma
GAS5	Growth arrest-specific 5
HAS2	Hyaluronan Synthase 2
hnRNPU	Heterogeneous nuclear ribonucleoprotein U
HOTAIR	HOX transcript antisense RNA
HOTTIP	HOXA transcript at the distal tip
HOXA	Homeobox A
HRCR	Heart-related circRNA
IGF2BP3	Insulin-like-growth factor 2 mRNA-binding protein 3
KCNQ1OT1	KQT-like subfamily, member 1 opposite strand/antisense transcript 1
KIF1C	Kinesin-like protein
KLK3	Kallikrein-related peptidase 3
LIPCAR	Long intergenic non-coding RNA predicting cardiac remodelling
LSD1	Lysine (K)-specific demethylase 1A
MALAT1	Metastasis Associated Lung Adenocarcinoma Transcript 1
MBNL1	Muscleblind-like splicing regulator
MFACR	Mitchondrial fission and apoptosis-related circRNA
MFN2	Mitofusion 2
MHRT	Myosin heavy-chain-associated RNA transcripts
MIAT	Myocardial infarction associated transcript
MICRA	Myocardial Infarction-Associated CircRNA
MLL1	Mixed-lineage leukemia 1
MMP9	Matrix metallopeptidase 9

Myd88	Myeloid differentiation primary response gene 88
MYH6	Myosin heavy chain, α isoform
MYH7	Myosin heavy chain beta
MyoCD	Myocardin
NDUFAF2	NADH:ubiquinone oxidoreductase complex assembly factor 2
NEAT1	Nuclear enriched abundant transcript 1
NELF	Negative elongation factor
NFIB	Nuclear factor 1 B-type
NFkB	Nuclear factor kappa
Nkx2.5	Homeobox protein Nkx
NORAD	Non-Coding RNA Activated By DNA Damage
NPPB	Brain natriuretic peptide
NRIP1	Nuclear Receptor Interacting Protein 1
NRON	Non-coding repressor of NFAT
OGDH	Oxoglutarate Dehydrogenase
p53	Tumour protein p53
PES1	Pescadillo homologue 1
PINK1	PTEN-induced kinase 1
PLOD2	Procollagen Lysine,2 Oxoglutarate 5 Dioxygenase 2
PTEN	Phosphatase and tensin homologue
RDH13	Retinol dehydrogenase 13
RIPK1/3	Receptor interacting serine/threonine protein kinase 1/3
RWDD3	RWD Domain-Containing Sumoylation Enhancer
SAFA	Scaffold attachment factor A
SDC4	Syndecan-4
SFRP5	Secreted frizzled-related protein 5
SMD	STAU1-mediated mRNA decay
SMMILR	Smooth Muscle Enriched LncRNA
STIM1	Stromal Interaction Molecule 1
TGFB1	Transforming growth factor beta 1
TIA1	TIA1 Cytotoxic Granule Associated RNA Binding Protein
TNF	Tumour necrosis factor
UBC	Ubiquitin C
UCA1	Urothelial Cancer Associated 1
Uchl1	Ubiquitin carboxy-terminal hydrolase L1
VASH1	Vasohibin 1
VEGF-A	Vascular endothelial growth factor A
VTCN1	V-Set Domain Containing T Cell Activation Inhibitor 1
WDR5	WD Repeat Domain 5
WISPER	Wisp2 super-enhancer–associated RNA
XIST	X-inactive specific transcript
Ybx1	Y-Box Binding Protein 1
ZNF584	Zinc finger protein 84

Chapter 1
Introduction

Despite the discovery of the cardiac troponins and the natriuretic peptides as powerful diagnostic and prognostics biomarkers for myocardial infarction (MI) and heart failure (HF), it remains difficult to identify high-risk individuals early in the disease course. Traditionally, clinical biomarkers have taken the form of proteins, peptides or neurohormones, but more recently DNA and RNA have been detected in biofluids which has opened up a whole new field of biomarker discovery. DNA and RNA biomarkers have already begun to appear clinically for cancer and are starting to gain attention in the cardiovascular field. RNA-Seq technology has evolved at a rapid pace since its inception in the genomic world roughly 15 years ago and is starting to shed light on the relatively new non-coding RNA players, long non-coding RNAs (lncRNAs) and circular RNAs (circRNAs). These represent an exciting new source of biomarkers as they have been associated with a wide range of cardiovascular diseases and are detectable in various body fluids. In parallel, bioinformatics software involved in RNA detection and analysis has been developed to allow detection of mRNAs, lncRNAs and circRNAs at a greater precision.

1.1 Aims of the thesis

The overall aim of this thesis is to identify new candidate RNA markers to improve detection of myocardial ischaemia and progression from ischaemic heart disease to HF. Specific aims are:

1. To develop a bioinformatic pipeline that identifies annotated mRNAs, lncRNAs and circRNAs, and putative novel lncRNAs from RNA-Seq data.

2. To identify coding and lncRNAs (including novel lncRNAs) that are associated with myocardial ischaemia in human heart tissue (data provided in collaboration with Harvard Medical School).

3. To establish an RNA-Seq method to detect mRNAs, lncRNAs and circRNAs in human plasma.

4. To identify coding and lncRNAs in human plasma (including novel lncRNAs and circRNAs) that are associated with the presence of ischaemic heart disease or progression from ischaemic heart disease to heart failure.

1.2 Hypotheses:

This thesis is founded on the hypothesis that novel RNA biomarkers will aid cardiovascular diagnosis and prognosis. Specific hypotheses are:

- That putative novel RNA transcripts can be identified from RNA-Seq data using a bioinformatics approach.

- That expression of genes and regulatory RNA transcripts (mRNAs, lncRNAs and circRNAs) coordinate the cascade of cellular, inflammatory, and biochemical events triggered in response to myocardial ischaemia. Some of these transcripts (or their translated protein products) will be secreted from cardiomyocytes into the circulation in detectable concentrations and have clinical utility as biomarkers.

- That circulating RNA transcripts (mRNAs, lncRNAs and circRNAs) reflect the disease status of an individual and can be detected by RNA-Seq in human plasma. Some transcripts will have utility as biomarkers for detecting the presence of ischaemic heart disease or the progression of coronary heart disease to heart failure.

1.3 Thesis outline

The first part of Chapter 2 of this thesis includes an introduction to the physiology of the ischaemic heart and heart failure, the current clinical biomarkers for diagnosis and prognosis and their limitations. The second part of Chapter 2 reviews our current understanding of lncRNAs and circRNAs including biogenesis and functions and their roles in cardiovascular disease. Experimental methods are outlined on Chapter 3. Chapter 4 explains the development of the bioinformatic pipeline after which its validation is presented and discussed. This pipeline is the used to analyse ischaemic heart tissue data in Chapter 5 with further analysis using a second data set from human plasma is presented in Chapters 6 and 7. The results are discussed in Chapter 8. Chapter 9 presents the conclusions and future directions arising from this work.

Chapter 2
Literature Review

2.1 Heart Disease in New Zealand

Ischaemic heart disease (IHD) occurs when the major blood vessels in the heart become narrow and stiff due to the build-up of fatty, atherosclerotic plaques and is a major cause of death in both developed and developing countries [1]. Similarly, IHD is a leading cause of mortality and morbidity in New Zealand, accounting for 14 deaths per day [2] and loss of >90,000 disability-adjusted life years each year [3]. One in twenty adults in NZ live with IHD [4] with older New Zealanders and those of Māori and Pacific ancestry, particularly affected [5, 6]. Ischaemic heart disease accounts for over half of all cardiovascular disease mortality and the mortality rate among Māori was more than twice as high as that among non-Māori[7]

The chronic result of IHD is angina (pain due to inadequate blood supply to the heart), or myocardial infarction (MI, a blockage of blood flow to the heart or prolonged ischaemia usually caused by blood clot). MI remains the most common cause of heart failure (HF) worldwide [8].

In lay terms, HF is where the heart can no longer pump enough blood to meet the physiological demands of the body and is associated with compensatory and adverse structural remodelling of the heart. Clinically, HF is defined as a syndrome characterised by typical symptoms (including breathlessness, ankle swelling and fatigue) that may be accompanied by signs (including elevated jugular venous pressure, pulmonary crackles and peripheral oedema) caused by a structural and/or functional cardiac abnormality, resulting in a reduced cardiac output and/or elevated intracardiac pressures at rest or during stress [9].

HF is a complex syndrome that can be caused by multiple factors that lead to disorders of the pericardium (a thin sac that surrounds the heart), myocardium (the muscle tissue of the heart),

endocardium (the inner most layer of the heart), heart valves, the great vessels (the superior and inferior vena cava, the pulmonary artery and vein, and the aorta). Risk factors for HF include IHD, hypertension, metabolic abnormalities, congenital heart defects, cardiomyopathy, inflammation, and toxins. [10].

The good news is that the rates of death from acute MI have reduced with improved coronary care –a 60% reduction in mortality during the first 30 days post MI in the last 30 years [11]. However, the consequences of patients living longer is an increased incidence of post-infarction HF [11]. HF is now a global pandemic estimated to affect 26 million people worldwide [12] and 80,000 New Zealanders [13], and this is set to increase with an aging population. Again, Māori fare worse than non- Māori: HF mortality rate among Māori is more than twice as high as that of non-Māori and Māori are about 4 times as likely as non-Māori to be hospitalised for HF [7]

Coronary heart disease and heart failure involve many different genes, pathways, and tissues. There have been exciting discoveries in recent years with regards to the potential of long non-coding RNAs and circular RNAs as biomarkers and regulators of cardiovascular disease. Also, as Because these relatively newly discovered classes of non-coding RNAs are less studied than mRNAs and microRNAs this thesis will focus on long non-coding RNAs involved in myocardial ischaemia, MI and ischaemic HF.

2.1.1 IHD, Atherosclerosis, and plaque formation

To understand myocardial ischaemia, it is important to understand the underlying pathophysiology of CAD which begins with atherosclerosis. Atherosclerosis is a silent, chronic, inflammatory, vascular pathology [14]. Over the decades, our understanding of atherosclerosis has evolved from a model of passive cholesterol deposition to a dynamic, complex interplay of different cell types and cytokines (cell signalling molecules) resulting in a chronic inflammatory disease [15]. Atherosclerosis can develop silently over decades and is

driven by excessive exposure to ‘bad’ cholesterol - low density lipoproteins (LDLs) along with other pathogenic factors such as infectious disease, hypertension, diabetes and smoking which cause injury to endothelial cells (the cells that line our blood vessels). This sets off a chain reaction of inflammatory responses which begins with normal LDLs becoming internalised inside the artery wall and damaged by free radicals through oxidation (Figure 2-1A-C). This causes endothelial cells to secrete adhesion proteins which encourage capture and activation of monocytes, a type of white blood cell capable of differentiating into a macrophage that can phagocytose (engulf) various substances such as cell debris, microbes and any other foreign substances. A second consequence is recruitment of platelets to the injury site, which release additional inflammatory cytokines to encourage further aggregation of leukocytes (a type of white blood cell involved with counteracting foreign substances). The monocytes penetrate the compromised endothelial wall and differentiate into macrophages [16]. The process continues with further recruitment and ingestion of lipoproteins and cholesterols by the macrophages. The macrophages phagocytose the oxidised cholesterol and become ‘foam cells’. The foam cells, along with cell debris and further inflammatory cells, form a fatty core [17], over which a fibrous cap of collagen, smooth muscle cells and elastin is formed, resulting in an atherosclerotic plaque [14] (Figure 2-1D). Accumulation of these plaques over time can narrow the lumen of the artery and, if severe enough, will restrict the flow of blood. When this occurs within a coronary artery and oxygen demand outstrips supply, for example during physical activity, ischaemia results, leading to angina pectoris (otherwise known as angina) which causes chest pain but is not fatal [18]. If the plaque cap ruptures, blood clots can quickly form at the site of rupture, which can either completely block blood flow or can break away to cause a blockage elsewhere (Figure 2-1E). If this blockage is formed in one of the coronary arteries, then heart muscle (myocardium) downstream of the blockage can die as it is starved of oxygen, resulting in myocardial infarction (Figure 2-1F).

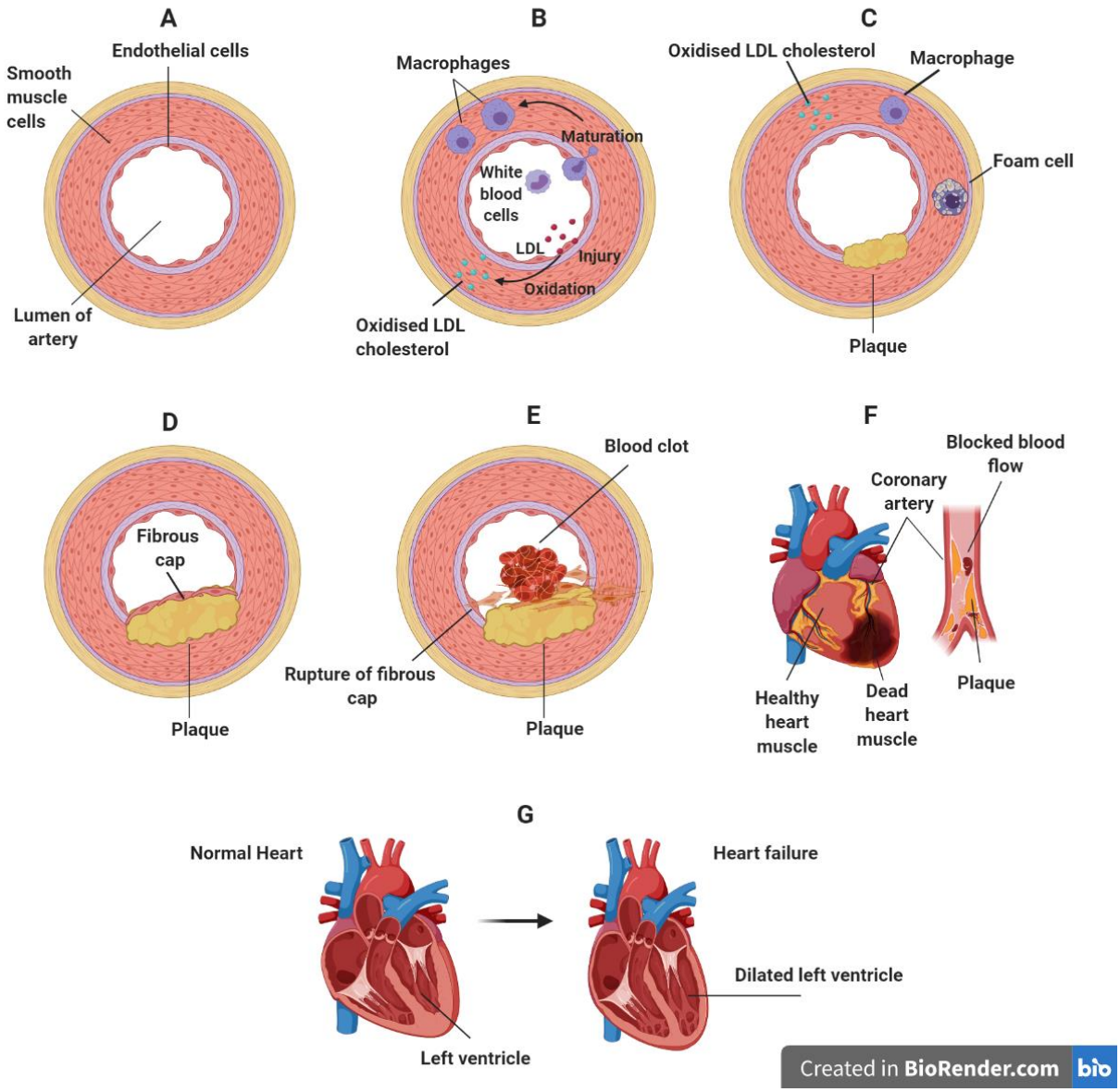


Figure 2-1 Development of atherosclerosis, leading to myocardial infarction and heart failure

A) *A healthy coronary artery: the lumen is lined with endothelial cells which are surrounded by smooth muscle cells.* ***B)*** *Damage to the lining of the lumen results in inflammation. The lining becomes permeable allowing LDLs inside the wall of the artery. Once inside, LDLs become oxidised which attracts white blood cells to the area. These penetrate the endothelial cells and mature into macrophages which can engulf the LDLs.* ***C)*** *Macrophages full of LDL cholesterol become foam cells which eventually burst to deposit cholesterol. Accumulation of cholesterol results in plaque formation.* ***D)*** *This process continues, narrowing the lumen of the artery, reducing the blood flow and causing angina. Proteins and smooth muscle cells form a thin, fibrous cap over the plaque.* ***E)*** *If the cap ruptures blood clots can form that can block the artery or break away and block an artery elsewhere* ***F)*** *If blood flow in the coronary artery which supplies the heart muscle becomes blocked the heart muscle is starved of oxygen and can die causing myocardial infarction.* ***G)*** *An example of a normal, normal heart (left), a heart in failure with altered morphology and dilated left ventricle (right). Figure adapted from [19]and created in Biorender.com*

2.1.2 Myocardial Infarction, left ventricular remodelling and heart failure

Under normal, aerobic conditions the heart generates most of its energy (90-95%) from fatty acid and carbohydrate (glucose and lactate) breakdown via mitochondrial oxidative phosphorylation, with the remainder supplied by glycolysis (the process of breaking down glucose without the need of oxygen). In the presence of oxygen, cardiomyocytes produce energy in the form of ATP using the aerobic oxidative phosphorylation pathway in mitochondria to produce 32 molecules of ATP per molecule of glucose. In contrast, under ischaemic conditions, mitochondrial oxidative phosphorylation decreases, and energy production is immediately shifted to anaerobic glycolysis which only produces 2 molecules of ATP per glucose molecule [20]. As there are no energy reserves in the heart, in severely ischaemic hearts a depletion of mitochondrial ATP occurs rapidly [21] The reduction in ATP leads to cell swelling caused by an increase in intracellular calcium and chloride ions (approximately a third of energy consumption under normal conditions is used to drive ion pumps). Eventually the osmotic overload becomes too much and the myocyte is irreversibly damaged leading to cell death [20].

Human cardiomyocytes have limited capacity to regenerate and cannot regenerate after MI [22]. MI causes both apoptosis (a controlled, programmed cell death) and necrosis (a less tightly controlled, unregulated cell death) of cardiomyocytes due to reduced oxygen supply [23]. After MI, a two-phase process is initiated in an attempt to repair the heart. An initial inflammatory response is triggered with leukocytes flooding into the infarcted area as they try to clear necrotic tissue and extracellular matrix debris [24]. This is followed by a proliferative phase defined by activation of fibroblasts to myofibroblasts. Myofibroblasts, usually absent in healthy myocardium, are phenotypically between smooth muscle cells and fibroblasts, and secrete large amounts of extra cellular matrix (ECM) proteins such as collagen into the interstitial space. This leads to expansion of tissue that has reduced contractile ability at the site of infarction [25]. The scar elongates and thins as the surrounding cardiomyocytes

undergo hypertrophy (an increase in size, rather than number, of cardiomyocytes). This increases the volume of the left ventricle (LV), an initial compensatory process that improves stroke volume and cardiac output but over time further overloads the heart [26, 27].

Reperfusion (reoxygenation) after MI can save further loss of cardiomyocytes but can also cause more injury (up to 50% of MI injury) by damaging the microvasculature via oxidative stress [28].

Together, these processes influence the size of the infarct, the degree of cardiac function and the amount of compensatory cardiac remodelling. After the initial compensatory response, ongoing cardiomyocyte hypertrophy in the non-infarcted border zone increases ventricular wall thickness and further enlarges the ventricular chamber, and there is a shift in the geometry of the ventricle from a normal elliptical shape to a spherical one [29].

The remodelling process continues in an effort to compensate for reduced cardiac output, but eventually becomes detrimental. Ultimately, hypertrophy and scarring decrease cardiac contractility to the point where the heart cannot pump the blood sufficiently to meet the body's demands and enters HF. [30]. The transition from compensatory to adverse remodelling is poorly understood, as is the biological reason why some people regain ventricular function whilst others do not [31].

HF can be broadly classified into two groups depending on the volume of blood pumped from the left ventricle with each contraction (left ventricular ejection fraction, LVEF); HF with reduced ejection fraction (HFrEF, LVEF< 40%) or HF with preserved ejection fraction (HFpEF, LVEF ≥ 40%). LVEF is calculated by the:

$$\frac{\text{volume of blood pumped from the left ventricle per beat (stroke volume)}}{\text{volume of blood collected in the left ventricle at the end of diastolic filling (end-diastolic volume)}} \quad [9]$$

Although HFrEF and HFpEF share the same clinical classification, they should be considered as two distinct subtypes of HF as they differ in terms of pathophysiology and aetiology, as well as treatment responses [32]. Because this thesis focuses on the progression of ischaemic heart disease and IHD and MI are more strongly associated with HFrEF than HFpEF, the following section on neurohormonal compensatory mechanisms will focus on HFrEF alone [33, 34].

2.1.3 Neurohormonal compensatory mechanisms of HFrEF

Cardiac output is a combination of heart rate and stroke volume (the volume of blood pumped at each contraction). In a healthy heart, if left ventricular contraction decreases there will be an incomplete emptying of the chamber. This leads to an increase in end diastolic volume (the volume of the blood in the ventricles just prior to contraction), an increased stretch of myocardial fibres and consequently an increase in force on the subsequent contraction (part of the mechanism known as the Frank-Starling law) [20]. This is not the case in the HFrEF heart, where there is ventricular impairment and a diminished contraction capability. In HFrEF, the heart cannot achieve the required stroke volume and cardiac output is reduced. To increase stroke volume, the ventricle responds by remodelling resulting in ventricular dilation. The consequences of these factors are an increased end diastolic volume which leads to increased left atrial pressure and pulmonary venous pressure. These can cause a build-up of pressure behind the heart, which can force fluid into the lungs and result in shortness of breath and other congestive symptoms [35].

Structural remodelling, a decrease in cardiac output and falling blood pressure all trigger compensatory neurohormonal mechanisms. These include the sympathetic nervous system (SNS), renin-angiotensin-aldosterone system (RAAS), natriuretic peptides and arginine vasopressin (AVP) [34, 36]. Activation of the SNS increases blood pressure by vasoconstriction (narrowing the blood vessels) and increases contractility and heart rate by releasing epinephrine (adrenaline) and norepinephrine (noradrenaline) [36]. In response to

SNS activation, the kidney releases the hormone renin into the bloodstream to activate the renin-angiotensin-aldosterone system (RAAS) in an attempt to raise blood pressure and blood volume [36]. Renin converts angiotensinogen to angiotensin I which is then processed to angiotensin II by the angiotensin-converting enzyme (ACE). Angiotensin II induces vasoconstriction and activates the antidiuretic hormones AVP and aldosterone. Both AVP and aldosterone regulate blood volume: AVP is both a potent vasoconstrictor and antidiuretic hormone (acts on the kidneys to decrease urine formation), whereas aldosterone causes an increase in reabsorption of sodium and water from the kidneys [36].

Again, there is a temporary reprieve from the declining cardiac output however, prolonged exposure to these hormones leads to ever-increasing fluid and salt in the body and eventually into HF. Chronic SNS activity can lead to desensitising of cardiomyocytes to catecholamines (the hormones adrenaline and noradrenaline) leading to reduced contractility as well as cardiac arrhythmias [34]. Chronic exposure to the RAAS can lead to further myocyte hypertrophy and interstitial fibrosis thereby increasing detrimental remodelling [34]. To counteract the renin and aldosterone secretion, the heart releases the natriuretic peptides, atrial natriuretic peptide (ANP, released by the atria) and B-type natriuretic peptide (BNP, released primarily by the cardiac ventricles). These natriuretic peptides promote diuresis (increased production of urine), natriuresis (excretion of sodium in the urine) and vasodilation. Unfortunately, their effects soon become diminished as their target organs become less responsive through receptor desensitisation and/or further release of counter-regulatory hormones from the RAAS and SNS [37].

The two cardiac natriuretic peptide hormones are worth introducing here as they are of considerable importance in HF. ANP is encoded by the gene *NPPA*, which is translated into a 151 amino acid pre-prohormone (pre-proANP), and subsequently cleaved to produce the prohormone $proANP_{1-126}$. A second cleavage of this prohormone produces the biologically active ANP_{1-28} and a 98 amino-acid NT-proANP fragment. The half-life of ANP in the

circulation is approximately two minutes although proANP is more stable. Assays have been developed to detect the mid region of proANP (MR-proANP) which show it provides similar diagnostic (and prognostic) information as NT-proBNP (described in the following section, 1.1.5) in patients with acute HF [38].

The gene encoding BNP, *NPPB* is transcribed and translated in response to cardiac ventricular myocyte stretch in the form of pre-proBNP, a 134 amino acid peptide [39]. This peptide is cleaved to remove a 26 amino acid signalling peptide leaving the prohormone $proBNP_{1\text{-}108.}$ A second cleavage of this prohormone produces the biologically active C-terminal peptide $BNP_{1\text{-}32}$ and the inactive (N-terminal) NT-proBNP which are both released into circulation. Whereas BNP is cleared from plasma by the natriuretic peptide clearance receptor type-C and proteolysis, NT-proBNP is thought to be cleared from plasma by renal excretion. BNP has a half-life of 20 minutes in circulation whereas NT-proBNP has a half-life of 120 minutes which has considerable advantages for use as a biomarker [40-42].

2.1.4 Biomarkers for heart failure

For patients presenting to the emergency department with breathlessness, there needs to be a fast and accurate test to distinguish between HF or pulmonary disease. Traditionally HF diagnosis consisted of taking a clinical history, along with a chest X-ray or echocardiogram. However, X-rays have low sensitivity and specificity for making a clinical diagnosis of HF [40] and both X-rays and echocardiograms may be difficult to obtain out of the normal working hours in the acute setting. In the last 20 years, the natriuretic peptides have emerged as the international gold standard for HF diagnosis [9]. In the early 2000's, the Breathing Not Properly Study (B.N.P. Study) was the first major study (1586 patients) to look at circulating BNP levels in patients presenting to the emergency department with acute breathlessness [43] (the very first study to measure ANP and BNP to assess their effectiveness at diagnosing HF was from the Christchurch Heart Institute [44]). Not only were higher BNP levels seen in plasma from patients with clinically diagnosed HF compared to those without HF (mean 675

± 450 pg/mL versus 110 ± 225 pg/mL, respectively; p=0.001) but the levels of BNP correlated with increasing severity of HF. Using a BNP cut-off of 100 pg/mL, BNP concentration had a 90% sensitivity and 76% specificity for HF and using a cut off of 50 pg/mL, BNP had a negative predictive value of 96%.

Building on this work, the ICON (International Collaborative of NT-proBNP) clinical trial investigated NT-proBNP (the amino terminal fragment of BNP) in 1,256 patients presenting with new-onset shortness of breath at emergency departments in New Zealand, the United States, Spain, and the Netherlands [45] The trial reported much higher circulating levels of NT-proBNP in patients with HF compared to those without (4,639 pg/ml versus 108 pg/ml, respectively; p<0.001) and again, plasma concentrations increased with HF severity. Using a cut-off of 300 pg/mL of NT-proBNP was sufficient to rule out acute HF with a sensitivity of 99% and a specificity of 60%. Because circulating concentrations of the natriuretic peptides increase with age, the investigators also developed age stratified cut-off points to improve specificity of a rule-in test (Table 2-1) [45]. These cut-off levels are now used in international guidelines for the diagnosis of HF [9].

Table 2-1 Adapted from Januzzi et al 2006 [45]. Rule-out and age stratified 'rule-in' cut off levels of NT-proBNP for diagnosis of acute heart failure.

Category	Optimal cut off (pg/mL)	Sensitivity (%)	Specificity (%)
Rule out (n=1256)	300	99	60
Rule in			
<50 years old (n = 184)	450	97	93
50-75 years old (n = 537)	900	90	82
> 75 years old (n=535)	1800	85	73

pg/mL = picogram per millilitre

While the ICON study primarily looked at diagnostic capabilities, they also demonstrated levels of NT-proBNP as a prognostic biomarker for survival in patients with acute HF during the first 76 days following presentation. Other trials have also demonstrated both BNP and

NT-proBNP as independent predictors of all-cause death and of readmission for chronic HFrEF [46-48] Again, of note, both BNP and NT-pro BNP are also strong predictors of adverse events (death, HF, and new myocardial infarction) after MI. [49, 50]

2.1.5 Limitations of natriuretic peptides

Unfortunately, circulating concentrations of the natriuretic peptides can vary in response to a range of physiological factors, in addition to HF. These include age, gender, obesity, atrial fibrillation, renal failure, anaemia, or inflammation of the heart muscle (myocarditis) [9]. Further, most studies have been carried out on patients of European ancestry and do not account for differences in natriuretic peptide levels across ethnic groups. A meta-analysis looking at BNP levels in 92,072 HF patients from white, black, Hispanic and Asian populations showed that despite having similar severity of disease Asian and black patients had higher BNP levels at admission compared with white and Hispanic patients, although BNP provided prognostic value regardless of ethnicity [51] Another complication is that the natriuretic peptides are excreted by the kidneys and so concentrations can potentially be elevated in patients with renal failure without HF [52].

These potential confounding factors are compounded by the considerable variation can be apparent between patients in the extent of the initial myocardial insult, prior medical history and lifestyle factors. Variability of BNP levels are due to several factors including age, gender, kidney function, atrial fibrillation and BMI. Certain confounders also seem to cluster in certain ethnic groups with Maori and Pasifika having higher BMI than Pakahe. A recent paper suggests that the pro-BNP precursor is glycosylated (affecting both cleavage to form BNP / NT-proBNP and antibody binding in the laboratory assays). Also, there seems to be higher glycosylation with a higher BMI. It may be the case that a different class of biomarker (RNA and not protein) would avoid this issue of glycosylation but there could also be an underlying genetic component even after adjusting for confounders. If this is the case then this genetic component could equally affect the lncRNAs and circRNAs.

Thus, despite the utility of the natriuretic peptides in diagnosis and prognosis in HF and in prognosis post MI, it remains difficult to identify patients at risk of progression from myocardial ischaemia/infarction to ischaemic HF.

2.1.6 Other cardiac biomarkers and their limitations

Troponin is a protein made up of three subunits (troponin C, troponin I and troponin T) and is found in both cardiac and skeletal muscle. Because the I and T subunits are structurally different in the heart compared to skeletal muscle, this makes them specific to cardiac myocyte injury [53]. Cardiac specific troponin T (cTnT) and troponin I (cTnI) are released into the circulation by cardiomyocytes after cardiac injury and remain elevated for multiple days afterwards. High sensitivity assays for cTnT and cTnI have significantly improved the diagnostic accuracy of MI as they detect troponin at much lower concentrations [54]. The European Society and American College of Cardiology define acute MI as an increase in serum troponin greater than the 99 percentile of a healthy reference population (with other signs of cardiac ischaemia [53].

Troponin levels are also elevated in patients with HF [55]; however, they are a marker for myocardial injury rather than a specific marker for HF and are useful in risk stratification [9].

C-reactive protein (CRP) is synthesised by hepatocytes and is part of the immune response. Early studies demonstrated increased CRP levels in individuals with ongoing ischaemia, unstable angina and chronic atherosclerotic disease [56]. The Physicians Health Study (PHS) demonstrated that high-sensitivity C-reactive protein (hsCRP) is elevated decades before the first acute ischaemic event and is a strong predictor of risk of acute MI [57, 58]. HsCRP is a clinically useful biomarker however, there is doubt as to whether CRP is itself a target for intervention and upstream cytokines in the inflammatory response such as interleukin-6and interleukin-1 are potential targets[59]. It should be noted that CRP is a general inflammation

marker and could be elevated by a number of diseases or injuries and needs to be interpreted in a clinical context.

Elevated hsCRP levels are also associated with HF [60]. However, a meta-analysis showed that there were discrepancies of hsCRP cut-off values when used for stratifying patients for HF development between studies; most studies included only subjects older than 65 years and sex differences were apparent [61]. In addition to the natriuretic peptides, Galectin-3 (expressed during tissue inflammation, repair and fibrosis) [62] and ST2 (Soluble Interleukin 1 receptor-like 1,expressed by cardiomyocytes in response to mechanical stress) have also been proposed for HF [63]. However, because they lack specificity for HF the American College of Cardiology (ACC) and American Heart Association (AHA) currently recommend they be for additive risk stratification to the natriuretic peptides [64].

2.1.7 Future Biomarkers?

Until recently biomarkers have generally been proteins or peptide neurohormones. The cardiac troponins and the natriuretic peptides are currently the gold standard markers for diagnosis of MI and HF respectively, but they have their limitations discussed above. Cardiovascular disease is complex; many factors influence the speed and route from sub-clinical atherosclerosis to MI to ischaemic HF and how severely an individual will be affected. The aim of this thesis is to identify new candidate markers to improve detection of ischaemia, MI, and prediction of progression from ischaemic heart disease to ischaemic HF. This thesis investigates potential RNA biomarkers an alternative molecular class to proteins, neurohormones and DNA.

There is relatively new field of genomics which focusses on the non-coding genes of the genome. Only 2-3% of the human genome codes for proteins and around 80% is transcribed into non-coding RNA [65]. While a considerable proportion of non-coding RNAs represent the ‘work horses’ of fundamental cellular processes such as transfer RNAs (tRNAs) and

ribosomal RNAs (rRNAS), over the last 20 years discoveries in the fields of lncRNAs, circRNAs and microRNAs (miRNAs) suggest important roles for these newer classes of non-coding of RNAs in almost all cellular processes. Non-coding RNAs of all three classes have been implicated in many disease states and their roles as biomarkers are beginning to be established [66-68]. Although miRNAs are a group of non-coding RNAs they are beyond the remit of this thesis. miRNAs are well studies as biomarkers in comparison to lncRNAs and circRNAs. This thesis focusses on the more understudied class of noncoding RNAs. The following sections discuss lncRNAs and circRNAs in detail, their biogenesis, functions, and finally their role in cardiovascular disease.

2.2 The Long Non-Coding Genome

2.2.1 LncRNAs – An overview

The rather protein-centric view of molecular biology, with its central dogma of DNA being transcribed to RNA and then translated into protein, is having to be rethought. One of the biggest misnomers in genetics has been the term 'junk DNA' [69]. A belief in a vast transcriptional wasteland residing in our genome was encouraged by the C-paradox (Swift, 1950). This term was coined in the 1950s to describe the puzzling disparity between the size of an organism's genome and its complexity. It was assumed that the complexity of an organism would be correlated with the amount of cellular DNA content, but this was not the case - the single-celled amoebae has a genome up to 100-fold larger than the human genome [70]. The Human Genome Project, completed in 2004, added a second paradox - the G-value paradox which assumes a relationship between an organism's complexity and its number of protein-coding genes [71]. Instead of the original prediction of around 100,000 protein-coding genes, humans had a mere 31,000 (which was later re-estimated to be around 21,000). More than 97% of our genome does not code for protein [72] thus, rather than the size of the genome or the number of protein-coding genes, it is the *relative* abundance of non-*protein-coding* genes or non-coding RNA within total genomic DNA that is most strongly correlated

with an organism's complexity [73]. Many of these non-coding regions are conserved across mammalian species suggesting their biological importance [74].

In the mid 2000's, two large scale projects from the FANTOM (Functional Annotation of Mammals) and ENCODE (Encyclopedia of DNA Elements) consortia set out to provide a comprehensive catalogue of mammalian transcription. In 2005, the FANTOM Consortium concluded that there were more non-coding genes than coding genes in the mouse genome [75] and, in 2007, the ENCODE Consortium began a pilot project to delineate all functional elements encoded in the human genome [76]. Their seminal paper which comprised 32 institutions studying 147 cells types was published five years later [77]. From this study they assigned 80% of the genome to be biochemically active (in other words transcribed or potentially functional), most of which is not translated into protein and is termed non-coding RNA. The rapid pace of discovery in this non-coding world can be attributed to the development of RNA-Seq technologies that have allowed the transcriptome to be sequenced to greater depths with greater precision enabling the discovery of rare transcripts. This is especially pertinent to certain classes of non-coding RNAs, such as lncRNAs, as these transcripts are typically less abundant than protein-coding transcripts [78].

Non-coding RNAs are rather arbitrarily separated into two groups depending on their length. Any transcripts shorter than 200 nucleotides are categorised as 'small non-coding RNA'. This group is heterogeneous and contains the well-known RNAs such as rRNAs and transfer RNAs (tRNAs), both essential for messenger RNA (mRNA) translation, as well as small nuclear RNAs (snRNAs), involved with RNA splicing; small nucleolar RNAs (snoRNAs), which guide chemical modifications of other RNAs; miRNAs and short interfering RNAs (siRNAs), both involved with transcriptional and post-transcriptional silencing; piwi-interacting RNAs (piRNAs), which act to silence transposons in germ cells, and more recently; transcription initiation RNAs (tiRNAs) and splice site RNAs (spliRNAs) which are thought to guide nucleosome positioning [79].

LncRNAs are transcripts longer than 200 nucleotides that lack protein-coding potential [78]. Over the last decade there has been an explosion of novel lncRNAs being discovered due to the rapid development of accurate, high-throughput sequencing technologies, as well as decreasing costs and enhancements in bioinformatic analysis tools [80, 81]. The arrival of third generation sequencing technologies, bypassing the need for PCR amplification steps, generate much longer read lengths. With these advantages in mind, it is tempting to think that the list will continue to grow or, at the very least increase the accuracy and completeness of lncRNA annotations [82]. Although the functional roles of many lncRNAs have yet to be determined, numerous lncRNAs have already been implicated in diseases ranging from cancer [83], neurological disorders [84], diabetes [85] and DNA imprinting [86]. In the coming years, many more lncRNAs may be implicated in disease as a result of large-scale searches for genetic variants that differ in frequency between healthy and diseased individuals (genome-wide association studies GWAS), which have concluded that >90% of our inherited susceptibility for disease comes from these non-coding regions [87].

The following sections of the literature review discusses lncRNA annotation which leads on to lncRNA classification. The next part discusses their functions within the nucleus and then the cytoplasm. The final part discusses our current understanding of their role as biomarkers in cardiovascular disease and their roles in atherosclerosis, MI and cardiac remodelling/HF.

2.2.2 Annotation – the state of play

Perhaps the largest challenge facing the non-coding RNA field at present is to collate a thorough and comprehensive ‘gold standard’ annotation resource. Numerous lncRNA annotation resources exist [88-96]; however, with differing techniques and cell/tissue types used for discovery of lncRNAs, there are inconsistencies and variation amongst them. The predicted number of human lncRNA genes ranges from 15,787 to 144,134 (from GENCODE version 26 [88] and NONCODE 2016 [91] respectively) with human lncRNA transcripts ranging from 27,720 to 233,696 [88] and [91] respectively). Xu *et al*. compared 24 current

lncRNA annotation resources and found that 88% of lncRNA transcripts were unique to one resource with only 8 transcripts showing the same exon structure in 5 resources. This figure is even more disappointing when looking at the subset of long intergenic non-coding RNAs (lincRNAs) with 91% of transcripts being unique to one resource and the remaining 9% occurring in only 2 resources [97]. The task of mapping millions of short reads correctly onto the current reference genome annotation and then assembling these into novel transcripts is not a trivial one, and bioinformatic software has not yet achieved 100% specificity and sensitivity. Added to this, potential technical issues such as the methods of preparation of the RNA before sequencing may introduce bias in coverage of the transcriptome [98]. This could potentially introduce errors in recognising all constituent exons, which may lead to assembling incomplete isoforms. As mentioned earlier, this problem should be improved with the advent of much longer reads generated with third generation sequencing technologies. The GENCODE consortium within the ENCODE project has for several years been manually annotating a comprehensive set of human lncRNAs [78]. This constitutes the largest manually curated catalogue of human lncRNAs, making it a good reference point. There are also several databases attempting to provide accurate, manually curated, online repositories from various sources of annotations [90, 97, 99, 100]. A recent database has been established which integrated several previous databases to provide a repository of *experimentally validated* lncRNAs [101]. This is an important next step in accurately classifying this relatively new class of non-coding RNA. Indeed, current annotation may overestimate the numbers of lncRNAs. A growing number of transcripts that have been mistakenly classified as lncRNAs have been shown to have short open reading frames (sORFs) that encode functional peptides known as micro peptides [102-105]. As bioinformatic and functional techniques evolve to identify these micro peptides and potentially other forms of coding RNAs, it will be a challenge for the annotations to remain abreast of these advances.

2.2.3 Characteristics of lncRNAs

On first glance it appears that lncRNAs are mRNAs that lack an open reading frame (ORF); they are both transcribed by RNA polymerase II (RNAPII) from loci with similar epigenetic marks at their promoter regions [74]. LncRNAs are present in the nucleus and/or the cytoplasm and are mostly thought to be 5'-capped (7-Methyl guanosine (m7G) capping at the 5′ end occurs during the initiation phase of Pol II transcription), spliced and polyadenylated (poly-A) [106]. However, there are also distinctions: lncRNAs generally have fewer exons (on average 2.8 exons in lncRNAs compared to 11 exons in protein-coding genes [107], with nearly half having just two exons (42% of lncRNAs have two exons compared to 6% of protein-coding genes [78]). LncRNAs typically have lower expression levels (10-fold lower) and appear to show a high degree of cell type-, tissue-, developmental stage or disease state-specificity [107, 108]. They have fewer conserved primary sequences, [106], although the level of conservation is higher at lncRNA exon splice sites and much higher at their promoters [109]. This reduced conservation at primary sequence level compared to mRNAs is perhaps unsurprising given the fact that lncRNAs do not have to maintain stringent amino acid coding sequences. LncRNAs can form complex secondary and tertiary structures [110] and the functional domains formed from this 3-dimensional folding are thought to be more strictly conserved than their primary sequence. Reports have shown cross-species, tissue-specific conservation of expression in equivalent genomic loci despite no nucleotide sequence conservation [111]. Early predictions were that 40% of lncRNAs are polyadenylated [112], but as the majority of studies may have been biased towards detecting poly-A transcripts (by using oligo dT primers) for cDNA synthesis in library preparation for transcriptome analysis), this figure may be an overestimate. A study by Zhang *et al* (2014) demonstrated that lncRNAs can be processed into mature transcripts and stabilised through non-canonical pathways such as RNase P cleavage (a ribozyme which is known for processing tRNA) to generate a mature

3’end, or capped by snoRNP complexes at both ends [113]. Consequently, this group of lncRNAs do not have poly-A tails.

In general, lncRNAs are sub-divided by their location with respect to protein-coding genes, [97, 114, 115]. To date the subclasses are promoter upstream transcripts (PROMPTs), enhancer RNAs (eRNAs, described in detail below), natural antisense transcripts (NATs), intronic lncRNAs and long intervening/intergenic ncRNAs (lincRNAs), (Figure 2-2).

The definition of promoter-upstream transcripts (PROMPTs) is somewhat arbitrary and depends on the distance between the transcription start sites of a protein-coding gene and the lncRNA. If they share the same start site, they are technically bidirectional, but this term has been used for lncRNA and protein-coding gene pairs with up to 100 base pairs separating their divergent start sites [116]. Wu *et al,* 2017 define this subclass as lncRNAs transcribed in the antisense orientation, approximately 0.5-2.5kb upstream of the active transcription start sites (TSSs) of protein-coding genes [117].

Enhancer RNAs (eRNAs) are bidirectionally transcribed from enhancers with the two directions producing equivalent levels of RNA [118]. Along with PROMPT lncRNAs this subclass has rapid turnover rates as they are targeted by nuclear RNA exosomes (a multi-protein complex capable of degrading various types of RNA) [117].

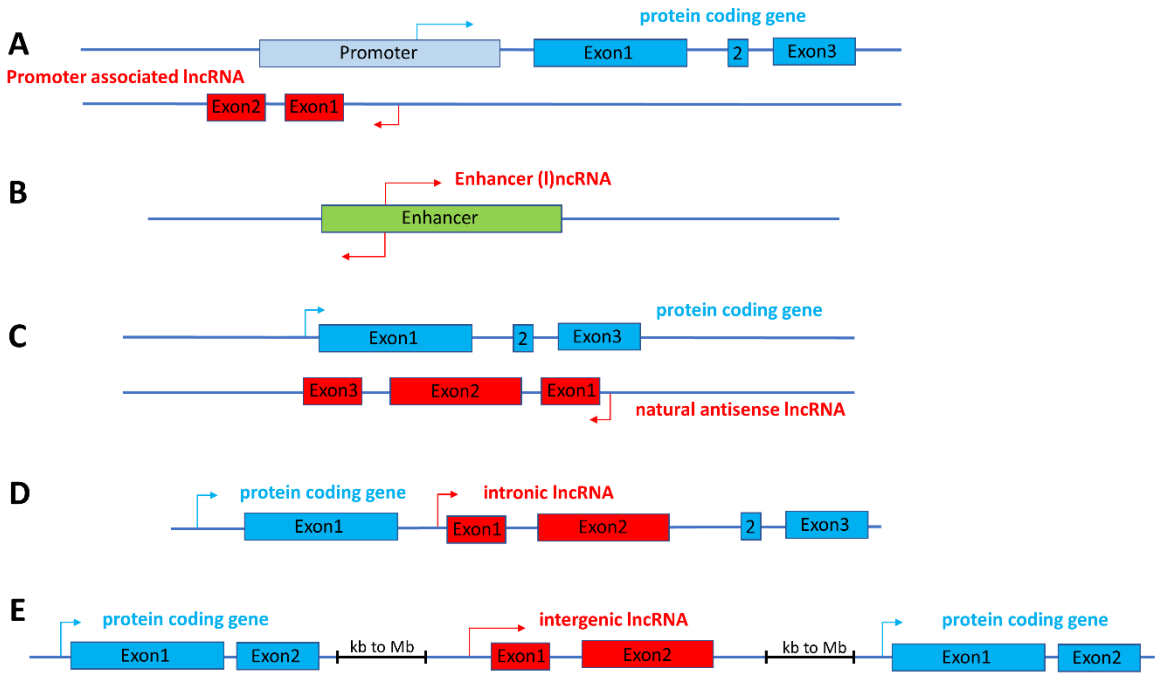


Figure 2-2 Classification of lncRNAs based on their genomic location.
A) *promoter associated lncRNAs (PROMPTS) are transcribed from the sense (not shown) or antisense strand from sequence within the promoter sequence* ***B)*** *Enhancer lncRNA from direct, bi-directional transcription of enhancer elements* ***C)*** *Natural antisense (NATs) lncRNAs are transcribed from the antisense strand and may share complementary sequence with the protein coding gene* ***D)*** *Intronic lncRNAs are located within intron(s) of a protein coding gene* ***E)*** *Long intergenic (Linc)RNAs are transcribed between genes. Blue lines represent the RNA sequence, blue boxes represent protein coding exons, red boxes represent lncRNA exons, red and blue arrows show the direction of transcription start sites, black lines represent gaps in the DNA sequence up to mega bases long.*

Natural antisense lncRNAs overlap any exon of a protein-coding gene but on the opposite strand and contain complementary sequences to the mature mRNA. It is thought that up to 63% of transcripts have an associated antisense transcript, many of which are not protein-coding [119]. The FANTOM consortium suggest that these sense/antisense pairs can either directly interact or that they are separately regulated, the difference suggesting whether the antisense transcript is acting in *cis* or *trans* [120]

Intronic/Overlapping lncRNAs are encoded within an intron of a protein-coding gene and can be transcribed from the same strand or the antisense strand. This class of lncRNA was first

thought to represent transcriptional noise (i.e. unprocessed pre-mRNAs) but has now been shown to be transcribed independently of exonic mRNAs, suggesting they are functionally independent [121]. This transcriptional autonomy is also strengthened by studies that show they can be transcribed from either the sense or antisense strand [122, 123].Long Intergenic RNAs (lincRNAs) do not overlap with protein-coding genes and are distant from any promoter regions or from neighbouring genes.

There are conflicting reports as to which group of lncRNAs is most actively transcribed. The GENCODE study [78] suggested that the majority of transcripts (64%) are long intergenic RNAs (lincRNAs) whereas a later study by St Laurent *et al*. (2012) suggested that intronic RNAs form the largest single class of lncRNAs, making up 35% of all lncRNAs [121]. Again, this discrepancy highlights the need for a more complete and robust annotation for lncRNA.

2.2.4 Functions of lncRNAs

It is now accepted that lncRNAs regulate gene expression in the nucleus and the cytoplasm [115], where they operate at both transcriptional and translational levels [124]. Their modes of operation are diverse, which reflects the versatility of the RNA molecule itself – it can form numerous secondary structures and bind specifically to RNA, DNA and proteins [125]. Because lncRNAs are not translated as their mRNA counterparts are, they can function as soon as they are transcribed within the nucleus and so act upon their targets much quicker. They can also be rapidly up- or down-regulated making them effective regulators of gene expression [126].

2.2.5 Functional mechanisms of lncRNA within the nucleus

Within the nucleus, lncRNAs can act in *cis* or *trans*. The *cis* transcripts act upon a nearby gene on the same allele (typically within 1Mb of the gene TSS); the *trans* transcripts interact with genes located at distant loci (> 1Mb) or on other chromosomes [127].

A growing evidence-base from functional studies suggest mechanisms as to how lncRNAs within the nucleus regulate gene expression. These include identifying and targeting specific sites on DNA to influence gene expression, tethering themselves to chromatin and acting as scaffolds for regulatory protein complexes and regulating compartmentalisation of the nucleus to sequester regulatory elements.

2.2.5.1 Localisation to DNA

Some lncRNAs exert their function by identifying and targeting specific sites on DNA. Engreitz *et al*, 2016 suggest two main strategies: firstly, via high affinity interactions with chromatin by binding to chromatin directly or via chromatin bound proteins, and secondly by three-dimensional (3D) proximity from their site of transcription (chromosomal 3D organisation is evolutionary conserved and looping of the DNA brings into proximity distantly related regulatory elements [128]). Examples of lncRNAs binding DNA using these strategies include MALAT1, HOTTIP and XIST.

MALAT1 (Metastasis Associated Lung Adenocarcinoma Transcript 1), an abundant and stable lncRNA, solely uses affinity interactions to localise to its targets at distant sites. MALAT1 is a highly conserved lncRNA and a regulator of mRNA splicing, and appears to be recruited to sites of actively transcribed mRNA, many of which are at sites distant to its own transcription site [129, 130].

In contrast, chromosomal looping brings the lncRNA HOTTIP (HOXA transcript at the distal tip) into proximity to its target homeobox genes. The homeobox (HOX) gene cluster are a group of related genes that control the body plan of the embryo along the anterior-posterior axis. HOTTIP is a lncRNA expressed at the 5’ end of the HOXA group on chromosome 7 and regulates activation of several of the downstream HOXA genes. HOTTIP directly binds the chromatin modifying WD Repeat Domain 5 (WDR5) and mixed-lineage leukemia 1 (MLL1) proteins, known regulators of gene transcription, and in doing so coordinates histone

modifications for gene transcription [131]. When Wang *et al* knocked out HOTTIP, the WDR5 and MLL proteins were not observed at their usual locations at the transcription start sites. Expressing HOTTIP from another region of the genome was not able to rescue these effects indicating that HOTTIP must work from the chromosome it is transcribed from. Indeed, by using chromosome conformation capture techniques (a method to analyse the 3D spatial organisation of chromatin within the nucleus) the authors were able to show chromatin looping, demonstrating HOTTIP in physical proximity to its target genes (Figure 2-3).

Lastly, X-inactive specific transcript (XIST) is an example of a lncRNA that utilises both strategies to exert its function of silencing one of the X chromosomes in female somatic cells. XIST uses a high affinity mechanism to bind to chromatin via scaffold attachment factor A (SAFA). Knockdown of SAFA leads to dispersed localisation of XIST and the loss of X chromosome inactivation [132-134]. However, this affinity interaction cannot fully explain the mechanism by which XIST exerts its function, as SAFA is apparent across the autosomes as well as the other X chromosome.

High resolution studies show that XIST localises to DNA regions that are in close three dimensional proximity to its transcription site; by moving XIST to another part of the X chromosome (using a cell line that expresses XIST from a transgene at a locus ~ 50 Mb proximal to the endogenous XIST locus) a new localisation pattern occurred which reflected its new proximity contacts [135, 136].

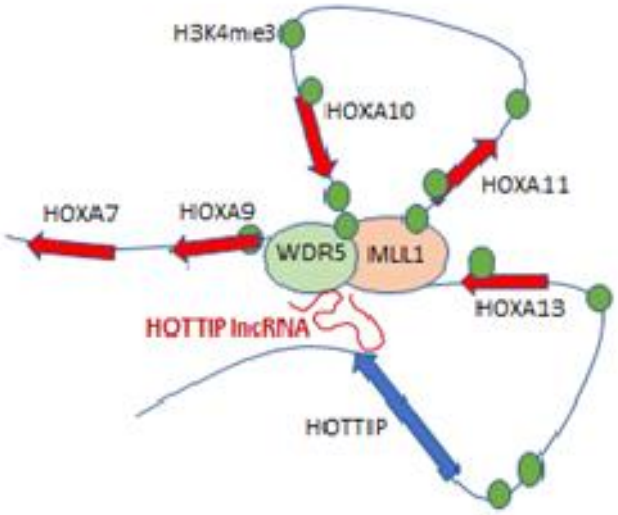


Figure 2-3 An example of a lncRNA using a proximity mechanism to exert its function of recruiting mediator proteins to influence gene transcription.

HOTTIP lncRNA stimulates the transcription of HOXA genes by enforcing H3K4me3 chromatin modifications via the chromatin modifying proteins WDR5 and MLL1. This function is enabled by the DNA looping bringing the HOXA genes into the vicinity of HOTTIP lncRNA Red boxes indicate the HOX genes, the blue box represents the HOTTIP locus, the red line represents HOTTIP lncRNA and green circles show H3K4me3 chromatin modifications.(Modified from [137])

It seems that highly abundant and stable lncRNAs such as MALAT1 can use affinity mechanisms alone, whereas low abundance lncRNAs such as HOTTIP, which is present at <1 copy per cell, regulates genes in very close proximity [138]. XIST sits somewhere in between and uses a combination of both mechanisms.

2.2.5.2 LncRNAs as scaffolds

LncRNAs have discrete domains within their secondary structure that can interact specifically with different proteins as well as DNA and RNA [127, 138]. These functional domains appear to have a high level of conservation and act to recruit various regulatory factors in a co-ordinated way to regulate transcription [139]. A good example of this is XIST, which physically interacts with protein and DNA at the same time (described above). Once tethered, discrete regions of the XIST transcript can bind specific protein complexes independently of one another. In doing so, XIST co-ordinates chromatin modification and chromatin compaction to silence gene expression [138, 140, 141]. Another example of this scaffolding

action involves a different homeobox (HOX) gene cluster to that described above, on chromosome 12 (Figure 2-4). The timing of expression of the HOX genes is precisely controlled by the lncRNA HOX transcript antisense RNA (HOTAIR) through simultaneous binding of regulatory proteins. HOTAIR acts as a scaffold to the histone modifying complexes polycomb repressive complex 2 (PRC2) at its 5' domain and lysine (K)-specific demethylase 1A (LSD1) complex at its 3' domain. Both proteins are involved with epigenetic DNA methylation to repress gene expression. In acting as a scaffold to these two repressive protein complexes and guiding them to their target genes the lncRNA HOTAIR co-ordinates and maintains epigenetic repression over a 40-kb region of the chromosome [125, 142, 143].

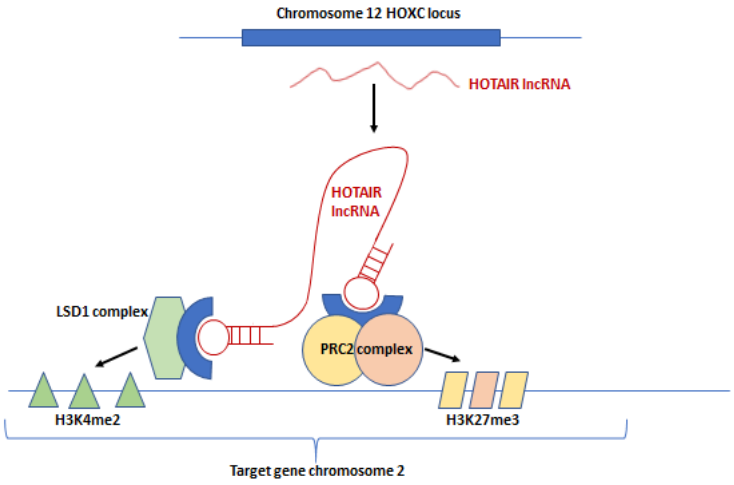


Figure 2-4 LncRNAs can act as scaffolds

HOTAIR is transcribed from the HOXC cluster on chromosome 12. It acts as a scaffold to histone modification complexes PCR2 and LSD1 to regulate gene expression in a trans acting manner 40kb away on chromosome 2 (modified from [144]).

2.2.5.3 LncRNAs and the 3D structure of the nucleus

LncRNAs are also involved with organising the 3D conformation of the nucleus. The lncRNA FIRRE (functional intergenic repeating RNA element), which is transcribed from the X chromosome, can influence the nuclear architecture across chromosomes by interacting with Heterogeneous nuclear ribonucleoprotein U (hnRNPU) and DNA on chromosomes 2,9,15 and 17. In doing so, FIRRE forms its own subcellular compartment [145] (Figure 2-5).

When FIRRE is knocked out, this trans-chromosomal co-localisation is lost and there are transcriptional changes in the genes within these DNA regions. These genes are involved in energy metabolism/adipogenesis [146], suggesting that FIRRE coordinates expression of energy metabolism/adipogenesis by orchestrating this compartmentalisation of the genes involved in this biological process.

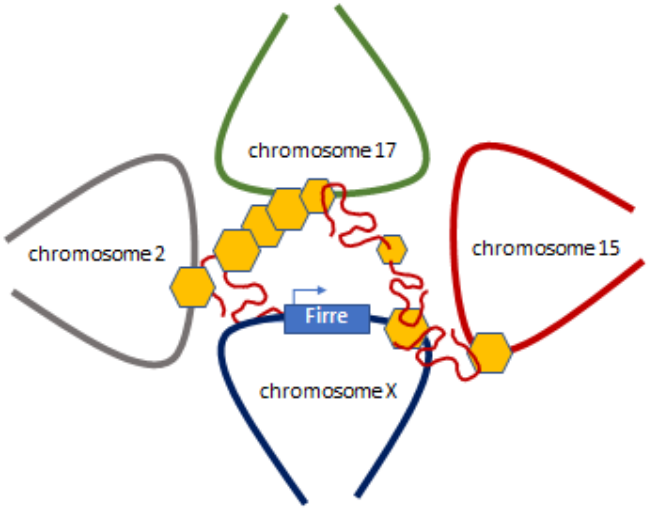


Figure 2-5 LncRNAs can modulate the 3D chromatin structure
The lncRNA FIRRE is transcribed form chromosome X and binds hnRNPU to interact with and regulate genes on other chromosomes forming its own subcellular compartment. Red lines represent FIRRE lncRNA, yellow hexagons represent hnRNPU (Modified from [145]

Another notable example of a lncRNA modulating the 3D structure of the nucleus is NEAT1 (nuclear enriched abundant transcript 1), which regulates the formation and maintenance of paraspeckles within the nucleus [147]. Paraspeckles are compartments within the nucleus that are thought to sequester mRNA and RNA-binding proteins (RBPs) to limit their function [148]. Knockdown of NEAT1 results in the loss of the paraspeckles and expressing it elsewhere in the genome results in ectopic paraspeckles formation [147, 149].

2.2.6 Functional mechanisms of lncRNA within the cytoplasm

The subcellular location of a lncRNA is key to determining its function and there are conflicting thoughts as to the amount of lncRNA localisation in the nucleus versus the

cytoplasm [78, 150, 151]. Quinn and Chang make the interesting case that rather than debate their predominant location, it should be accepted that lncRNAs are ubiquitous throughout the cell. [106].

Following export to the nucleus, mRNAs can be regulated by several post-transcriptional mechanisms. Implicated in these different mechanisms are lncRNAs, which influence mRNA stability, mRNA translation rates and levels of miRNAs within the cytoplasm that in turn have an effect on post-transcriptional regulation.

2.2.6.1 The stability of mRNAs

LncRNAs have been shown to both increase and decrease the stability of mRNAs. One mechanism by which lncRNAs decrease mRNA stability is by STAU1-mediated mRNA decay (SMD). STAU1 is a double-stranded RNA binding protein that binds to transcriptionally active mRNA at its 3' untranslated region (3'UTR) [152] (Figure 2-6A).

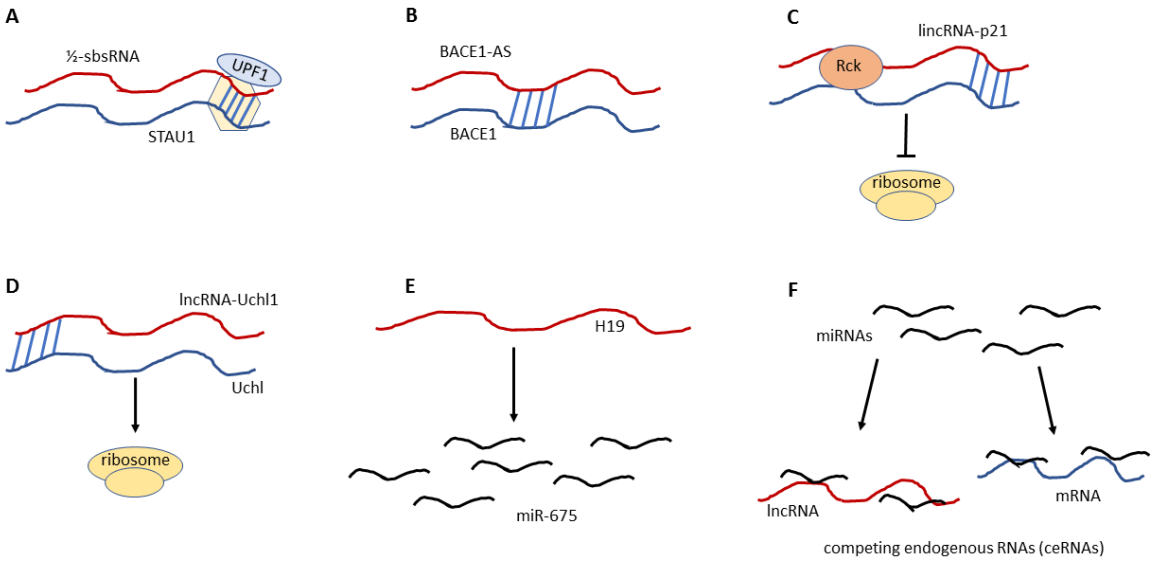


Figure 2-6 Mechanisms of lncRNAs in the cytoplasm

In the cytoplasm lncRNAs can influence mRNA stability, mRNA translation and levels of miRNAs ***A)*** *1/2-sbsRNA is an example of a lncRNAs that can decrease stability of its target mRNA* ***B)*** *In contrast, BACe1-AS increases stability of mRNA* ***C)*** *LncRNAs can modulate mRNA translation by either inhibiting translation e.g. lincRNA-p21 which can interact with translational repressor Rck or* ***D)*** *an example of lncRNA-Uchl1 promoting translation which*

*targets Uchl mRNA to active polysomes for translation **E)** Some lncRNAs such as H19 can give rise to miRNAs, in this case miR-675 **F)** LncRNAs can act as competing endogenous RNAs (ceRNAs) and in this way modulate gene expression. LncRNAs bind miRNAs which sequester them and prevent them from degrading their target mRNA.*

The authors showed binding of STAU1 to mRNA is regulated by a group of lncRNAs which the authors name ½- Staufen 1-binding site lncRNAs (½-sbsRNAs). Binding occurs through Alu elements (a repetitive sequence common throughout the human genome) at the 3' UTR of the targeted mRNA and another Alu element on the lncRNA. STAU1 then recruits the RNA helicase UPF1 which promotes decay. Conversely, other lncRNAs increase mRNA stability, through sense/antisense binding with their complementary mRNA. For example, base pairing between human β-site APP-cleaving enzyme 1 (BACE1) and its antisense lncRNA BACE1-AS increases BACE mRNA stability and consequently BACE1 protein abundance [153]. The formation of this duplex masks a binding site for miRNA miR-485-5p and consequently prevents its degradation (Figure 2-6B).

2.2.6.2 The translation of mRNAs

In addition to altering mRNA stability, lncRNAs can alter protein expression by controlling the rate at which their target mRNAs are translated into protein. LincRNA-p21 inhibits translation by base pairing with complementary mRNA transcripts and enhancing the interaction between the mRNA and translational repressors RCK/DDX6 (DEAD-Box Helicase 6). RCK/DDX6 interacts with argonaute proteins which are a class of proteins that act as repressors of translation [154] (Figure 2-6C). In contrast, the antisense Uchl1 (ubiquitin carboxy-terminal hydrolase L1) lncRNA forms a duplex with the 5' end of Uchl mRNA and targets the mRNA to polysomes for translation (Figure 2-6D) [155].

2.2.6.3 LncRNAs and miRNAs

Lastly, some lncRNAs, such as H19, modulate gene expression through miRNA mechanisms. H19 RNA primarily exists in the cytoplasm and functions as RNA regulators [156], its first exon harbours a template for two conserved miRNAs, *miR-675-3p* and *miR-675-5p*, which are

induced during skeletal muscle differentiation. In myoblasts, knockdown of H19 reduced skeletal muscle differentiation, a phenotype that could be rescued by expression of *miR-675-3p* and *miR-675-5p* [157] (Figure 2-6E). Other lncRNAs act as miRNA 'sponges', to sequester them and prevent their action on mRNAs. Within the cytoplasm a complex regulatory circuitry exists that is mediated by crosstalk among miRNAs, mRNAs and lncRNAs. Downregulation of miRNA interacting lncRNAs leads to an increase in the availability of miRNAs and a decrease in mRNA translation; conversely overexpression of lncRNAs leads to fewer miRNAs and an increase in mRNA translation (Figure 2-6F).

2.2.7 Enhancer RNAs (eRNAs)

There is another class of non-coding RNA within the nucleus called enhancer RNAs (eRNAs), although the classification of eRNAs and lncRNAs is a little blurry [158]

Enhancers are regulatory DNA elements that bind proteins to promote gene expression [159]. Two studies in 2010 demonstrated the expression of stimulus-dependent RNAPII transcription of bi-directional, non-polyadenylated non-coding RNAs at enhancer sites which were termed enhancer-derived ncRNAs or eRNAs [109, 160]. These studies showed that the level of eRNA expression at enhancers positively correlated with the level of mRNA synthesis at nearby genes. This suggested that this stimulus-dependent eRNA transcription was specifically occurring at enhancers that were actively involved in downstream mRNA synthesis [160]. Since 2010 eRNAs transcription has been confirmed in many cell types and tissues by FANTOM and ENCODE [118, 161] and estimations of 40,000 to 65,000 different eRNAs throughout the genome demonstrates their biological importance [118, 162].

Two potential functional mechanisms of eRNAs involve acting as decoys to RNA Pol II repressors and regulating conformational change in DNA. eRNAs can regulate transcription by acting as decoys to complexes that are involved in RNAPII activity. The negative elongation factor (NELF) complex induces RNAPII pausing - a regulatory mechanism for

stimulus-responsive genes in higher eukaryotes [163]. The authors demonstrated activity-regulated cytoskeletal protein (Arc) eRNA acts as an RNA decoy to bind NELF which facilitates NELF mediated RNAPII pausing back to RNAPII elongation (Figure 2-7A).

In addition, eRNAs can facilitate looping of DNA enabling enhancer-bound regulatory proteins to interact with RNA Pol II transcriptional machinery and activate gene transcription. This is demonstrated by the Kallikrein-related peptidase 3 (KLK3) and the Nuclear Receptor Interacting Protein 1 (NRIP1) genes. These were shown to interact with the Mediator complex (a multiprotein complex required for regulating transcription [164]) and Cohesin to promote DNA looping. Knockdown of NRIP1 eRNA showed reduced promoter: enhancer interaction [165] (Figure 2-7B).

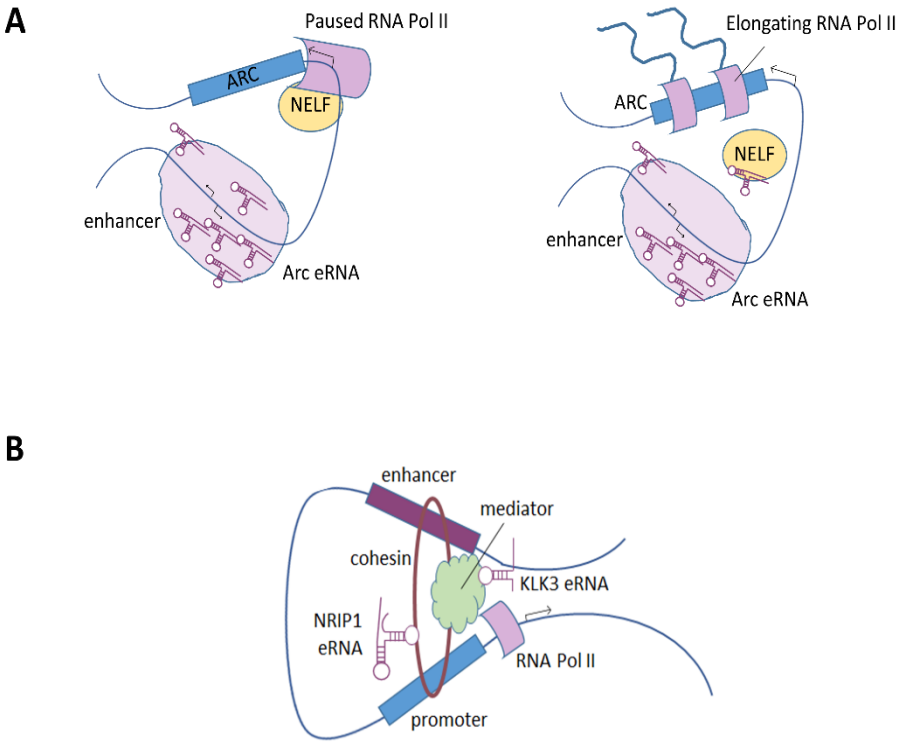


Figure 2-7 Example functions of Enhancer RNAs (eRNAs)

A) *Due to chromatin looping the ARC gene and enhancer are in proximity. Upon stimulation, transcription from both the ARC gene and enhancer element is initiated. ARC Enhancer RNAs (eRNAs) can then mediate the exit from RNA polymerase II (RNA Pol II) pausing by acting as a decoy to the negative elongation factor (NELF) complex.* ***B)*** *The Kallikrein-related peptidase 3 (KLK3) eRNA interacts with the multi-protein complex Mediator which*

facilitates the interaction enhancer and promoter regions to activate transcription. eRNAs like NRIP1 can also interact with Cohesin which acts as a ring to connect two regions of DNA

2.2.8 LncRNAs, Ischaemic Heart Disease (IHD) and Heart Failure (HF)

There is a growing literature supporting the role of lncRNAs in a wide spectrum of biological processes and diseases. In 2013, Cheng *et al* created the 'LncRNA and disease database' (http://www.cuilab.cn/lncrnadisease), which at the last update linked over 1,000 lncRNAs with 221 diseases. The following sections will focus on the roles of lncRNAs involved with cardiovascular disease and, in particular, ischaemic heart disease.

In 2007 the Welcome Trust Case Control Consortium carried out GWAS on 14,000 cases of seven common diseases (with 3000 controls) [166]. This and other studies identified a powerful association between CAD and MI and the chromosome 9p21.3 locus [167-169]. Chromosome 9p21.3 spans a non-coding region of 50 kb that includes several single nucleotide polymorphisms (SNPs) associated with risk of CAD and a lncRNA known as ANRIL (antisense non-coding RNA in the INK4 locus, also known as CDKN2B Antisense RNA). ANRIL has subsequently been shown to be associated with coronary atherosclerosis [170], carotid arteriosclerosis [171], peripheral artery disease [172] and other vascular diseases [173]. There are several isoforms of ANRIL each with tissue-specific expression patterns in endothelial, smooth muscle and inflammatory cells [170]. Carriers of risk alleles at 9p21.3 had higher levels of the two isoforms of ANRIL, EU741058 and NR_003529, but not the isoform, *DQ485454*, in whole blood and atherosclerotic plaque tissue, compared with other patients. Moreover, these two variants were correlated with the severity of atherosclerosis [170]. Holdt *et al*, [170] generated ANRIL over-expressing cell lines which led to significant changes of expression in genes enriched for cell adhesion, proliferation and apoptosis which are all central mechanisms of atherogenesis. These gene expression networks could all be reversed by RNA interference knock down of ANRIL [174]. ANRIL is thought to

regulate genes in *trans* by acting as a scaffold and binding the PRC1 and PRC2 complexes (Polycomb repressive complexes 1 & 2 – protein complexes that can remodel chromatin). An ALU repeat element expressed in both ANRIL and the promoters of target genes is thought to regulate this ANRIL-mediated *trans* regulation [174].

In the literature, much of the lncRNA data in heart comes from mice due to the lack of availability of human heart tissue. However, in the last five years there have been accumulating transcriptome studies using human blood and, to a lesser extent heart tissue, providing insight into the lncRNAs associated with vascular and cardiac disease.

Deep sequencing of the transcriptome of ischaemic and non-ischaemic human hearts by Yang *et al* showed that the expression profiles of lncRNAs differ between failing hearts before and after implantation of a left ventricular assist device. These findings suggest a role of lncRNAs in either the pathogenesis of HF or in the response to the restoration of cardiac function [95], but the functions of these lncRNAs were not investigated. Another study by Saddic *et al* investigated expression profiles of lncRNAs with RNA-Seq before and after ischaemic insult of cardiopulmonary bypass in 85 patients [175]. For the differentially expressed lncRNAs the authors looked at gene ontology of neighbouring genes and found significant enrichment for pathways involved in hydrogen peroxide metabolic processes, response to stress, response to stimulus and immune system processes. Four of the 15 most abundantly expressed lncRNAs in the heart (independent of ischaemia) were the well-known lncRNAs H19, MALAT1, NEAT1 and DANCR. H19 is upregulated in CAD patients [176] and is a regulator of cardiomyocyte hypertrophy [177]. MALAT1 regulates endothelial cell function and angiogenesis, with its inhibition reducing endothelial cell proliferation and vascular inflammation [178-180]. NEAT1 is involved with paraspeckle formation within the nucleus (described above) [181]. The role of Differentiation Antagonizing Non-Protein Coding RNA (DANCR) and cardiac disease is less clear but it has been implicated in reduced cell proliferation and cell cycle arrest in osteosarcoma cells [182].

A study by Zangrando *et al* demonstrated that expression levels of lncRNAs are regulated in cardiac tissue after MI in mice, [183] and a related group went on to look at a panel of five lncRNAs that were suspected to be involved in cardiac pathology in humans. The group looked at ANRIL, hypoxia inducible factor 1a antisense RNA 2 (aHIF 1A antisense RNA 2), potassium voltage-gated channel, KQT-like subfamily, member 1 opposite strand/antisense transcript 1 (KCNQ1OT1), MI-associated transcript 1 (MIAT) MALAT1 in peripheral blood by quantitative polymerase chain reaction [184]. Levels of aHIF 1A antisense RNA 2, KCNQ1OT1, and MALAT1 were higher in patients after MI than in healthy volunteers, while levels of ANRIL were lower in patients after MI. The study demonstrates altered expression profiles of several lncRNAs in patients with acute MI and suggests that these may be useful for prognosis. Interestingly, other studies which have used cardiac tissue, link MIAT upregulation to a higher risk of MI, microvascular dysfunction, cardiac hypertrophy and cardiac fibrosis [185-188], while KCNQ1OT1 may protect against ischemia/reperfusion (I/R) injury following acute MI [189].

These studies provide a list of potential candidates of differentially expressed lncRNAs involved in cardiac disease, but until we understand the functional relevance of these lncRNAs, the story will not be complete. Various strategies to up- or down-regulate expression of candidate lncRNAs are now being applied to tease out the functional effects of these lncRNAs on cardiac disease processes. These include RNA interference (RNAi), small hairpin RNAs (shRNAs), modified antisense oligonucleotides (ASOs) and gapmers (a chimeric antisense oligonucleotide that contains a central stretch of DNA monomers that induces RNase H cleavage), aptamers (single stranded oligonucleotides that target the lncRNA) small-molecule drugs (chemical compounds that block the activity of the target lncRNA [190, 191] and clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR associated protein 9 (Cas9) technology [192].

Techniques of up- or down-regulating expression levels have provided clues as to which mRNAs, proteins and pathways lncRNAs regulate. As shown in Table 2-2, this regulation occurs at all stages of disease, spanning at-risk individuals with atherosclerosis and vascular dysfunction, to before, during and after MI, to cardiac remodelling and HF, thus providing potential biomarkers and drug targets for each stage along the spectrum of heart disease. The rodent lncRNAs didn't appear to have a human equivalent which could be a naming issue. To interrogate this further the gene would have to be looked at in the gene annotation to see if surrounding genes are equivalent. Given that rodent cardiovascular models poorly replicate in the human condition then if these genes are in fact rodent specific then they do not give us a clear insight into human disease.

2.2.9 LncRNA biomarkers

Early diagnosis of cardiovascular disease can improve clinical management and reduce morbidity and mortality rates. Ideal biomarkers are low cost, non-invasive, accurate, stable and strongly predictive. However, none of the cardiovascular biomarkers have 100% sensitivity or specificity (described above) and there remains a critical need for new markers to further improve accuracy in diagnosis and prognosis in cardiovascular disease.

It is clear that lncRNAs play an integral part in various cardiac and vascular diseases and have the potential to become powerful biomarkers. There are technical challenges to their utility as biomarkers, since lncRNAs are expressed at lower levels than mRNAs and miRNAs in the circulation and are prone to degradation by ribonucleases. However, within the circulation, they can be bound by extracellular vesicles known as exosomes which are thought to increase their stability [193, 194]. Although the tissue of origin of the circulating lncRNA is not known, measuring lncRNAs in peripheral blood provides a non-invasive, readily accessible and potentially relatively cheap source of novel biomarkers. Their potential is already being realised in the cancer field where a lncRNA with high specificity and sensitivity for bladder cancer has been developed at an estimated $8 USD per test. [195]. Recent studies suggest that

this success may be possible for cardiac disease. In 2014, Kumarswamy *et al* investigated lncRNAs as potential biomarkers for HF. They used human plasma-derived RNA for microarray-based global transcriptome analysis to look at lncRNAs in 246 patients with or without left ventricular remodelling after MI. -

Table 2-2 LncRNAs involved in Atherosclerosis, Myocardial Infarction and Cardiac Remodelling/Heart Failure

Atherosclerosis				
LncRNA	**Location**	**Phenotype**	**Mechanism/Target Genes**	**Reference**
ANRIL (antisense non-coding RNA in the INK4 locus)	Human	Increased atherosclerosis risk	Impairs ribosome biogenesis, activates p53 promotes apoptosis	[170, 174, 196]
ANRIL (antisense non-coding RNA in the INK4 locus)	Human	Inflammatory factors linked to endothelial activation and atherosclerosis	Downstream of NF-κB pathway	[197]
SMILR (Smooth Muscle Enriched LncRNA)	Human	Atherosclerotic plaques	HAS2 (a critical component of the extracellular matrix that accumulates in human atherosclerotic lesions)	[198]
Biomarker				
COROMARKER	Human	Upregulated in CAD patients		[199]
H19/LIPCAR (long intergenic non-coding RNA predicting cardiac remodelling)	Human	Increased risk of CAD		[176]
novlnc6	Mouse/Human	Downregulated in AMI	Bmp10/Nkx2.5 (Cardiac transcription factors)	[200]
NRON (ncRNA repressor of the nuclear factor of activated T cells)	Human	Plasma levels upregulated in HF		[201]
MHRT (Myosin Heavy Chain Associated RNA Transcripts)	Human	Plasma levels upregulated in HF		[201]
UCA1 (Urothelial Cancer Associated 1)	Human	Lower plasma levels 12 hours post MI, elevated after 72 hours		[202]
Myocardial Infarction				
LncRNA	**Location**	**Phenotype**	**Mechanism/Target Genes**	**Reference**
APF (autophagy promoting factor)	Mouse	Cardiac autophagy and MI	APF↑ → miR-188-3p↓ → ATG7↑ → cardiac autophagy↑	[203]

CARL (cardiac apoptosis-related lncRNA)	Mouse	inhibits anoxia-induced mitochondrial fission and apoptosis in cardiomyocytes	Sequesters miR-539 maintaining mitochondrial homeostasis	[204]
MIAT (myocardial infarction associated transcript)	Human	SNPs linked to increased risk of MI		[185]
MIRT1/MIRT2 (myocardial infarction associated transcript)	Mouse	(MIRT1) Decreasing cardiomyocytes apoptosis, reducing inflammatory cell infiltration	Inhibition of NF-kB pathway	[205]
NRF (necrosis-related factor)	Mouse	Necrosis during MI	NRF↑ → miR-873↓ → RIPK1/3 ↑ → cardiac necrosis	[206]
ANRIL (antisense non-coding RNA in the INK4 locus)/HIF-1α-AS2 (hypoxia inducible factor 1A antisense RNA) 2 ()/KCNQ1OT1 (KQT-like subfamily, member 1 opposite strand/antisense transcript 1)/ MALAT1 (metastasis-associated lung adenocarcinoma transcription 1)	Human	May improve prediction of left ventricular dysfunction	↑ HIF-1α-AS2/KCNQ1OT1/ MALAT1 levels, ↓ ANRIL levels in MI patients compared to healthy volunteers	[184]
Cardiac Remodelling				
LncRNA	**Location**	**Phenotype**	**Mechanism/Target Genes**	**Reference**
Chaer (cardiac-hypertrophy-associated epigenetic regulator)	Mouse/Human	Cardiac Hypertrophy	PRC2/ mTORC1	[207]
Chast (cardiac hypertrophy-associated transcript)	Mouse/Human	Cardiac Remodelling	Pleckstrin - autophagic inhibition and cardiomyocyte hypertrophy	[208]
CHRF (cardiac hypertrophy related factor)	Mouse	Cardiac Hypertrophy	CGRF↑ → miR-489↓ → Myd88↑ → cardiac hypertrophy↑	[209]
GAS5 (Growth arrest-specific 5)	Rat	Cardiac Fibrosis	GAS5↑ → miR-21↓ → PTEN ↑ → cardiac fibrosis↓	[210]

H19	Mouse	Upregulated in pathological cardiac hypertrophy	miR-675/CaMKIIδ pathway (CaMKIIδ is a multifunctional serine/threonine protein kinase)	[177]
H19	Human/Mouse	Upregulated in human failing and mouse hypertrophic hearts	Cathepsin D	[211]
MHRT (myosin heavy chain associated RNA transcript)	Mouse	Protects heart from pathological hypertrophy	MYH6/MYH7 via Brahma Related Gene 1 (BRG1)	[212]
MIAT (myocardial infarction associated transcript)	Mouse	Cardiac Hypertrophy	competing endogenous RNA for miR-150	[188]
MIAT (myocardial infarction associated transcript)	Mouse	Cardiac fibrosis	competing endogenous RNA for miR-24	[187]
MIRT1/MIRT2 (myocardial infarction associated transcript)	Mouse	Left Ventricular remodelling after MI	List of upregulated genes including Nppb, TNF, MMP9, TGFB1, lectin galactoside-binding soluble 3 (lgal3), p53	[183]
ROR (regulator of reprogramming)	Mouse	Cardiac Hypertrophy	ROR↑ → miR-133↓ → ANP/BNP↑ → cardiac hypertrophy↑	[213]
WISPER (Wisp2 super-enhancer–associated RNA)	Mouse/Human	Cardiac Fibrosis	TIA1/PLOD2	[214]
Vascular				
MALAT1 (metastasis-associated lung adenocarcinoma transcription 1)	Human	Endothelial cell function and vessel growth	Cell cycle S-phase cyclins CCNA2, CCNB1, and CCNB2	[178]
MALAT1 (metastasis-associated lung adenocarcinoma transcription 1)	Human	Protects against endothelial dysfunction & upregulated in unstable angina	competing endogenous RNA for miR-22-3p/CXCR2(Interleukin-8 receptor)	[179]
MIAT (myocardial infarction associated transcript)	Human	Microvascular dysfunction	competing endogenous RNA for miR-150-5p/VEGF	[186]
SENCR (Smooth Muscle And Endothelial Cell Enriched Migration/Differentiation-Associated)	Human	Stabilizes smooth muscle cell contractility	MyoCD (a key transcriptional regulator of the smooth muscle cell contractile gene expression)	[96]

One lncRNA, LIPCAR (long intergenic non coding RNA predicting cardiac remodelling) increased significantly with post-MI left ventricular remodelling during chronic HF and could potentially be used to predict future deaths in patients with HF [215].

Further evidence for association of circulating lncRNAs with CAD, MI and HF comes from two recent studies. Higher levels of two lncRNAs, non-coding repressor of NFAT (NRON) and myosin heavy-chain-associated RNA transcripts (MHRT), were detected in HF versus non HF patients using quantitative reverse transcription-polymerase chain reaction (RT-qPCR) [201].

Additionally, levels of a lncRNA named urothelial carcinoma-associated 1 (UCA1) were reduced in early state acute MI patients and then increased at day three after acute MI [202].

Finally, Yang *et al* (2014) used microarray analysis to screen for differentially expressed plasma lncRNAs in CAD patients compared to controls. They found 265 differentially expressed lncRNAs and filtered these down to four candidates based on levels of expression, fold-change, and p values. Of the four candidates, the best candidate was the lncRNA AC100865.1 (Coromarker), which they suggested could be used as a biomarker for CAD [199].

2.2.10 Summary

These studies demonstrate that lncRNAs can be detected in human plasma. RNA-Sequencing (RNA-Seq) is a superior technology for biomarker discovery compared to microarrays in that it does not need pre-determined probes and therefore has an unbiased detection of transcripts. RNA-Seq has a broader dynamic range (not being hampered by background hybridisation and signal saturation), it has increased specificity and, arguably most important for lncRNAs, has increased sensitivity and superior detection of low abundance transcripts. As both studies by Kumarswamy *et al* and Yang *et al* used microarray analysis, it is tempting to think there may

be many more lncRNAs involved in CAD still to be discovered with RNA-Seq that could potentially be used as biomarkers.

2.3 Insights into circRNAs: their biogenesis, detection, and emerging role in cardiovascular disease

The material in the following sections formed the basis of a review article which has been submitted for peer review in RNA Biology.

2.3.1 Circular RNAs – an Overview

CircRNAs are an evolutionarily conserved form of non-coding RNA with covalently closed loop structures. The first studies to establish a functional role for circRNAs showed they can act as potent miRNA sponges and many other studies have focussed solely on this role. However, the biological functions of most circRNAs are still undetermined and other functional roles are gaining traction, including as protein sponges and regulators, and coding for proteins with an alternative mechanism of translation potentially opening up a whole new transcriptome. The first step to gaining insight into circRNA function is accurate identification and various software platforms have been developed for this purpose. What started out as specialised detection software has now evolved into whole bioinformatics pipelines that can be used for detection, *de novo* identification, functional prediction, and validation of circRNAs. However, few cardiovascular circRNA studies have utilised these tools. This section of the literature review summarises current knowledge of circRNA biogenesis, bioinformatic detection tools and the emerging role of circRNAs in cardiovascular disease.

The first endogenous circRNAs in humans were reported in the early 1990's as non-polyadenylated transcripts with a 'scrambled' exon structure (i.e. exons joined at consensus splice sites but in a different order to the primary pre-mRNA transcript) [216, 217]. As most early transcriptomic studies used isolation methods that enriched for poly-A transcripts, circRNAs, with their lack of poly-A tails, went undetected. However, with the evolution of

next generation sequencing kits that interrogated ribosomal depleted total RNA (rather than poly-adenylated RNA), along with RNAse R digestion to enrich for circRNAs, studies identifying and functionally characterising circRNAs quickly proliferated. Fast forward thirty years and we now know that circRNAs can contain a single exon or multiple exons, can contain exonic or intronic sequences or a combination of both (although most are thought to originate from middle exons and most commonly contain two to three exons [113, 218]), are conserved across species and are associated with many different disease states, including cardiovascular disease [219-222]. Despite these advances, our understanding of their functional roles is still in its infancy and for many circRNAs the functional mechanism has not been assigned.

CircRNA transcripts may be located in either the nucleus or the cytoplasm. Whereas exon-intron and intronic circRNAs remain in the nucleus and often appear to regulate the expression of the gene encoding them (parental gene) [223, 224], exonic circRNAs are more commonly exported to the cytoplasm where they exert multiple functions [218]. CircRNAs are highly stable due to a lack of free ends that are vulnerable to exonuclease activity, and can accumulate in non-proliferating tissues [225] and biofluids including plasma, saliva and urine [226-228]

This part of the literature review is divided into four sections. The first section summarises current knowledge of circRNA biogenesis, focussing on the most recent developments in the field. The second section provides a comprehensive list of current bioinformatic detection and downstream *in silico* functional annotation tools. The third section describes circRNA functions, including roles as miRNA sponges, protein sponges, protein scaffolds and protein-coding transcripts. The final section discusses the emerging roles of circRNAs in cardiovascular disease, including as biomarkers and regulatory molecules in atherosclerosis, myocardial infarction/ischaemia-reperfusion injury, HF, cardiac hypertrophy, and cardiac fibrosis.

2.3.2 Biogenesis

A single gene locus can produce multiple circularised transcripts, a process termed alternative circularisation [113, 218]. For example, Titin, the longest gene in the human genome, gives rise to 415 different exonic circRNA isoforms [229], compared with a median of three circRNAs per gene in human brain samples [222].

Canonical splicing of pre-mRNA to linear mRNA involves removing intervening sequences (introns) to join exons in a 5'-3' direction, creating a mature mRNA transcript ready for translation to a protein. This editing is carried out by a complex molecular machine termed the spliceosome which recognises *cis*-regulatory elements such as a GT/U dinucleotide (the donor site at the 5' end of the intron), an AG dinucleotide (the acceptor site at the 3' end of the intron) and a branch point near the 3' end of the intron. Exon-containing circRNAs are formed from non-canonical splicing of linear pre-mRNA where a downstream 5' splice site is joined to an upstream 3' splice site in what is known as 'back-splicing'. This results in a 3', 5' phosphodiester bond at the back-splicing junction forming a covalently closed circular transcript and a linear transcript with a skipped exon [230] (Figure 2-8A). Mutation of the canonical splice sites or inhibition of splicing by isoginkgetin (a general inhibitor of both the major and minor spliceosomes) diminishes circRNA production [231, 232], suggesting that back-splicing of circRNAs utilises the canonical spliceosome and the same splice sites as linear splicing.

A number of factors govern the formation of circRNAs and certain genomic features appear to favour circRNA biogenesis. First, for single exon circRNAs, the exon is, on average, 3-fold longer than other expressed exons [233]. This is in contrast with multi-exon circRNAs, where the exons are of more usual length. Second, exceptionally long introns that contain inverted repeat elements, such as inverted Alu repeat elements are significantly enriched in circRNA loci compared to linear controls, with nearly 90% of circRNAs in humans having ALU repeats in their flanking introns [234]. Third, conserved binding motifs for *trans*-acting RNA-

binding proteins (RBPs) in flanking introns also promote circularisation by allowing for RBP-associated intronic base pairing. Notable examples of RBPs include the protein Quaking (QKI), which regulates the human epithelial–mesenchymal transition [235], the fused in Sarcoma (FUS) protein in motor neurons [236] and Muscleblind-like splicing regulator 1 (MBNL1) in neuronal tissues [232]. Conversely, the RNA-editing enzyme, adenosine deaminases acting on RNA (ADAR) appears to inhibit circularisation and promote formation of linear transcripts. ADARs convert adenosine-to-inosine in double-stranded RNA which disrupts intronic base pairing thus 'melting' the secondary structures within the introns that facilitate circRNA biogenesis (Figure 2-8A). Accordingly, knockdown of ADAR1 significantly upregulates circRNA expression [234]. Fourth, the complementary binding of inverted repeat sequences or the binding of RBP proteins bring the back-splice junctions into close proximity. This allows circularisation of either the exon(s) only or, to a lesser extent, the exon with a retained intron (termed circRNAs containing exonic and intronic sequences, EIciRNAs, Figure 2-8A). Along with EIciRNAs, ciRNAs (circRNAs containing exclusively intronic sequences) are thought to reside in the nucleus and are involved with regulation of their parental gene [223, 224].

In contrast to exon-containing circRNAs, ciRNAs are generated during linear splicing when the 5' splice site of the excised intron joins to an adenosine at the branch point, forming a lariat structure (Figure 2-8B). Lariat structures are usually removed by a debranching enzyme that hydrolyses the 2'-5' phosphodiester bond at the branch point to produce linear molecules. ciRNAs are lariats that escape the debranching process. They contain consensus RNA motifs – a GU rich sequence near the 5' splice site and a C rich sequence near the branchpoint (Figure 2-8B), although it is not yet known how these consensus sequences work or what proteins are involved in escape from the debranching process [224].

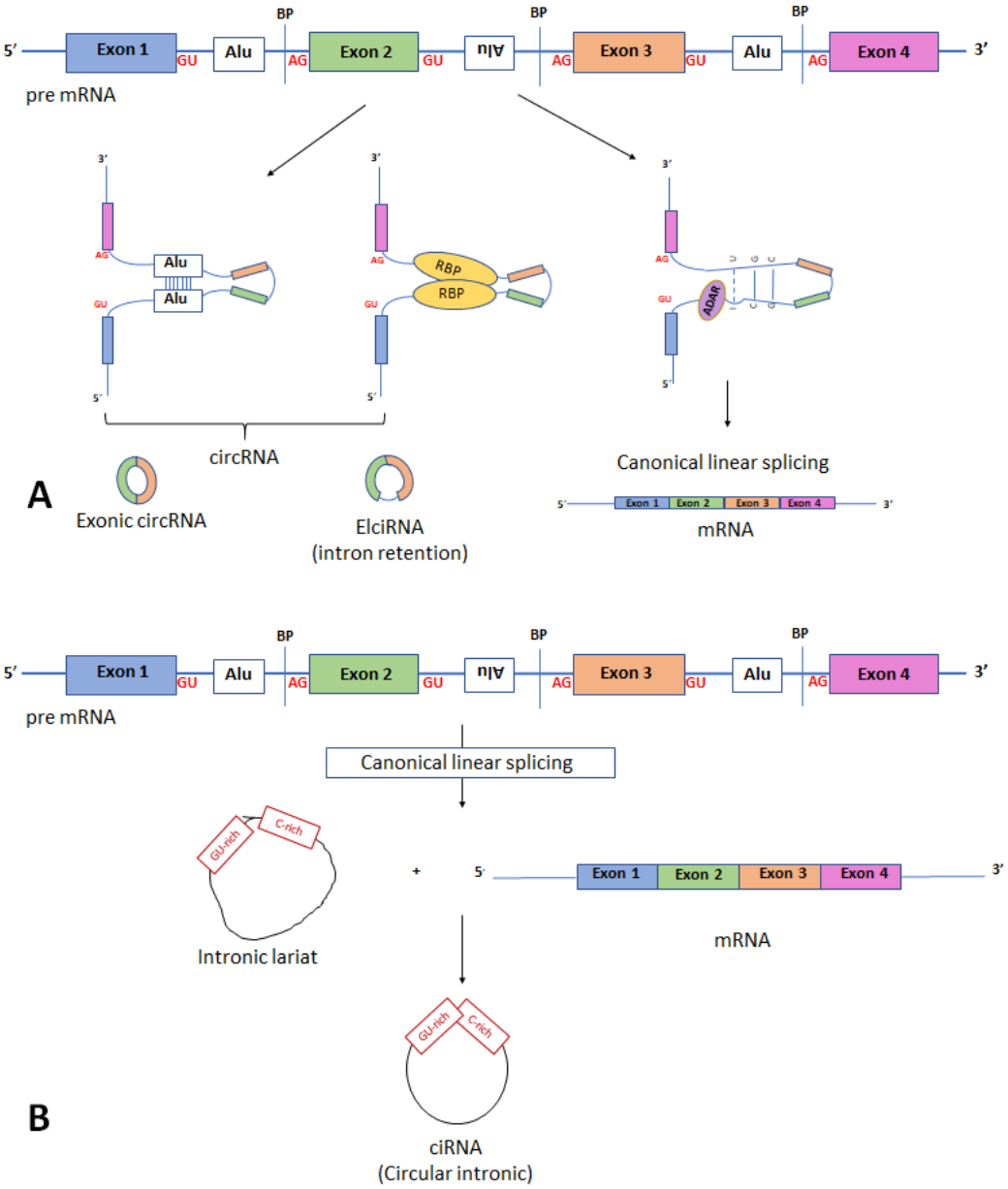


Figure 2-8 Components involved in linear mRNA or circRNA biogenesis

A) *The top panel shows a pre-mRNA transcript with 4 exons. Long flanking introns containing inverted repeat elements such as Alu elements or trans- acting RBP proteins bring the downstream splice site into close proximity with the upstream splice site to favour back splicing and thus circRNA production. Conversely, introns bound by ADAR1 'melt' RNA secondary structures by disrupting intronic base disrupting circularisation.* ***B)*** *Intronic circRNA biogenesis: Linear splicing produces lariat structures that are usually debranched and hydrolysed. However, GU rich and C rich intronic motifs escape debranching to form ciRNAs (intronic circRNAs). ADAR adenosine deaminases acting on RNA, BP: Branch points. RBP: RNA binding proteins. ALU: Alu repeat elements, mRNA messenger RNA*

Most circRNAs are generated from pre-mRNA transcripts that also produce linear mRNA although it is still not clear how or why circRNAs are produced in favour of their linear

counterparts. Given that the same spliceosome and splice sites are used and that many circRNAs consist of exons, it may be tempting to speculate that circRNA biogenesis could be a regulator of mRNA production. However, accumulating evidence show that circRNAs have their own function independent to that of the linear transcript.

In summary, the current model of circRNA biogenesis is that back splicing is promoted by a looping of the intron(s) bringing the (back) splice sites into close proximity. This 'looping' appears to be facilitated by longer introns, inverted repeat sequences and RNA binding proteins [113, 218, 235].

2.3.3 CircRNA detection.

Due to exclusion of non-polyadenylated transcripts in standard RNA-Seq protocols, circRNAs went undetected for years. As libraries evolved from poly-A enriched to (rRNA depleted) total RNA, circRNA detection became possible. Furthermore, addition of 3′-5′ exonuclease Ribonuclease R (RNAse R), which degrades linear RNA, enabled enrichment of circRNAs, although certain circRNAs are sensitive to RNAse R treatment including the well-known circRNA, CDR1as [237, 238]. However, as RNAse R treatment doesn't remove RNA with highly structured 3' ends such as snRNAs, histone mRNAs, or RNAs with G-rich G-quadruplex (G4) secondary structures, circRNA library preparation methods were further refined, using RNAse R treatment followed by polyadenylation of transcripts and then poly(A)+ RNA depletion [239-241].

Following RNA sequencing, bioinformatic detection of circRNAs remains challenging. For a list of current software for circRNA detection and potential functional prediction see Table 2-3. While numerous programmes are freely available, there appears to be limited overlap in the circRNAs identified between the different programs. This suggests that all programmes suffer from a proportion of false positives and that comparing results from several algorithms

and taking the intersection of these combined results will give more reliable outputs than the use of a single algorithm.

Table 2-3 Software for CircRNA detection and downstream applications

Software for CircRNA detection				
Tool	Category	Aligner	Link	Reference
CIRCexplorer CIRCexplorer2 CIRCexplorer3 /CLEAR	Split alignment based	TopHat2, Bowtie, (STAR)	https://pypi.org/project/CIRCexplorer/ https://circexplorer2.readthedocs.io/en/latest/ https://github.com/YangLab/CLEAR	[224] [242] [243]
CIRI CIRI2	Split alignment based	BWA	https://sourceforge.net/projects/ciri/ https://sourceforge.net/projects/ciri/files/CIRI2/	[244] [245]
DCC	Split alignment based	STAR	https://github.com/dieterich-lab/DCC	[246]
find_circ	Split alignment based	Bowtie2	https://github.com/marvin-jens/find_circ	[247]
MapSplice MapSplice2	Split alignment based	Bowtie	http://www.netlab.uky.edu/p/bioinfo/MapSplice2	[248]
segemehl	Split alignment based	segemehl	https://www.bioinf.uni-leipzig.de/Software/segemehl/	[249]
UROBORUS	Split alignment based	TopHat2, Bowtie	https://github.com/WGLab/UROBORUS	[250]
ACFS	Pseudoreference based	BWA-MEM	https://github.com/arthuryxt/acfs	[237]
CircRNAFisher	Pseudoreference based	Bowtie2	https://github.com/duolinwang/CircRNAFisher	[251]
KNIFE	Pseudoreference based	Bowtie, Bowtie2	https://github.com/lindaszabo/KNIFE	[252]
BIQ	No alignment	kmers	https://github.com/pmenzel/biq	[253]
CircDBG	No alignment	kmers	https://github.com/lxwgcool/CircDBG	[254]
CircMarker	No alignment	kmers	https://github.com/lxwgcool/CircMarker	[255]
Software for downstream of CircRNA detection				
ACValidator	*in silico* validation of circRNAs		https://github.com/tgen/ACValidator	[256]
CircCode	identifying circRNA Coding Ability		https://github.com/PSSUN/CircCode	[257]
CirComPara	detect, quantify, and correlate expression of linear and circRNAs from RNA-Seq data		https://github.com/egaffo/CirComPara	[258]
CircInteractome	explore potential interactions of circRNAs with RBPs, design specific divergent primers to detect circRNAs, study tissue- and cell-specific circRNAs, identify gene-specific circRNAs, explore potential miRNAs interacting with circRNAs, and design specific siRNAs to silence circRNAs		https://circinteractome.irp.nia.nih.gov/	[241]

CircPrimer	a software for annotating circRNAs and determining specificity of circRNA primers	http://www.bioinf.com.cn/	[259]
CircPro	identification of circRNAs with protein-coding potential	http://bis.zju.edu.cn/CircPro/	[260]
CircRNAwrap	circRNA identification, transcript prediction, and abundance estimation	https://github.com/liaoscience/circRNAwrap	[261]
CircTools	circRNA identification, RBP enrichment screenings, circRNA primer design, miRNA seed analysis and differential exon usage	https://github.com/dieterich-lab/circtools	[262]
miARma	miRNA, mRNA and circRNA analysis	http://miarmaseq.idoproteins.com/	[263]
NCLcomparator	post-screening non-co-linear transcripts (circular, trans-spliced, or fusion RNAs) for high confidence selection	https://github.com/TreesLab/NCLcomparator	[264]
ReCirc	prediction of circRNA expression and function through probe reannotation of non-circRNA microarrays	http://licpathway.net:8080/ReCirc/	[265]
Ularcirc	circRNA detection independent of gene annotation, visualisation of forward AND backsplice junctions recover predicted circRNA sequence, recover sequence of backsplice junctions and forward splice junctions, detect miRNA binding sites, detect putative open reading frame of circRNA	https://github.com/VCCRI/Ularcirc	[266]

The reader is directed to the following review for a detailed comparison between circRNA software programmes [267], although the field is rapidly evolving and several programmes have been updated and new ones have been developed since these articles were published.

CircRNA detection software needs to distinguish between circRNAs and their linear counterparts as most circRNAs will contain the same exon(s) as the linear mRNA. To do this the software exploits the fact that circRNAs contain back-spliced junctions. The detection strategies can be split into three groups – those that use a multi-stage mapping approach, those that directly detect the back-spliced junction reads using split or chimeric reads and, lastly, those that forgo the alignment step and assemble kmers (short sub sequences). In the multi-

stage mapping approach, there is an initial mapping step where reads that align continuously to the reference genome are filtered out and only unmapped reads are taken for further analysis. These 'unmapped' reads are then aligned to pseudo-sequences which are built around putative back spliced junctions (BSJs) [267] (Figure 2-9A) The newer aligners are 'splice aware' (software that split reads when aligning back to the reference to account for intronic sequences) such as STAR2, HISAT or TopHat2 [268-270], allowing for chimeric reads to be directly aligned to the human reference genome (Figure 2-9B). Care must be taken here as chimeric reads that are in the opposite order can also be generated from other instances in the genome such as intergenic or intragenic *trans* splicing to produce tandem duplications of exons [271]. To account for these potential false positives, one can select only those that occur in the same gene (to eradicate *intergenic trans* splicing) and use of biological replicates is recommended (to minimise false positives due to *intragenic* trans splicing). Lastly, the more recently developed software packages appear to be opting to forgo the mapping stage altogether and instead use the reference and annotation to assemble and store all kmers that are located near exon boundaries (Figure 2-9C). Each sequencing read is then examined for kmers and these are matched to the stored kmers. (Circ DBG builds a De Bruijn graph from the kmers). When two kmers from a single read are out of order to the reference then this suggests the presence of a circRNA. As this strategy assembles all possible kmers from the reference and annotation files only circs from annotated exons will be predicted, no *de novo* circRNAs will be detected.

Although circRNAs may contain multiple exons that align with multiple reads in an RNA-Seq run, it is only the reads that align across this back splice that can be counted as unambiguously originating from the circRNA. The other reads may have originated from either the linear or circular forms. For this reason, the read counts for circRNAs are relatively low. Caution must be followed when inferring the internal sequence of circRNAs for *in silico*

functional predictions (such as miRNA and RBP binding sites) and this is only possible for single exon circRNAs.

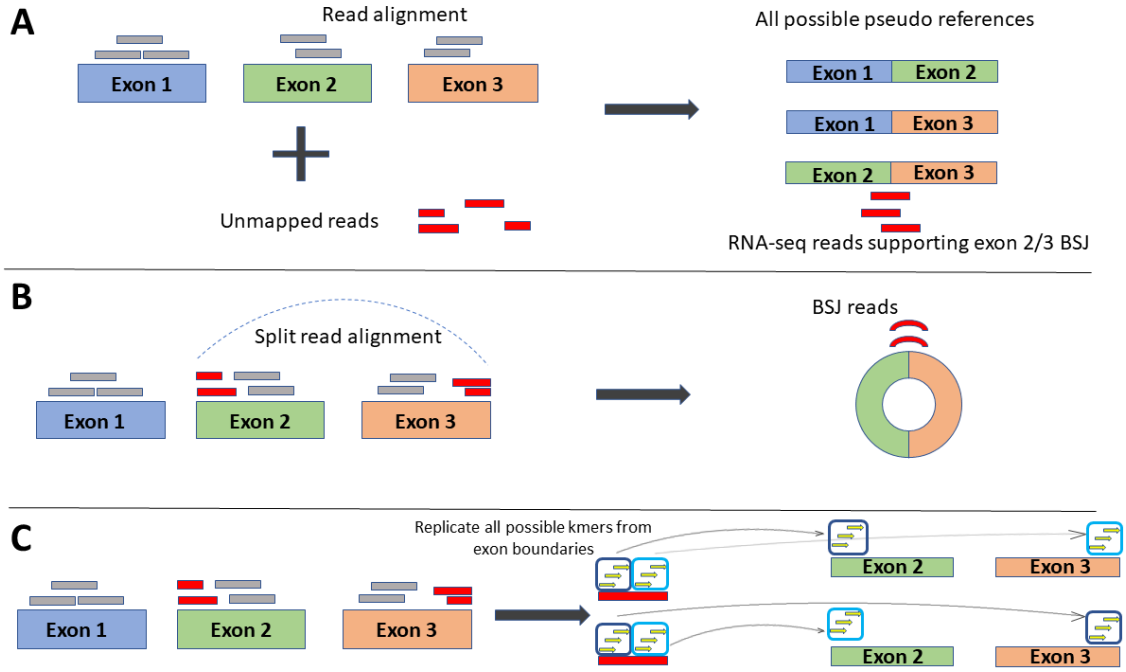


Figure 2-9 Three different strategies of identifying back-splice junction (BSJ) reads employed by circRNA detection software.

A) *Pseudo reference strategy: Reads are first aligned to the reference (grey boxes) any unmapped reads (red boxes) are taken forward and aligned against pseudo references which include all possible combinations of exon junctions. In this way, any reads that align identify the exons involved in the back-spliced junctions.* ***B)*** *Split alignment strategy: Splice-aware aligners split the sequencing reads which allows direct alignment to the reference. Reads can be selected to choose the BSJ reads* ***C)*** *Kmer strategy: The actual step of read alignment is forgone. Instead, all possible kmers (short stretches of sequence) are created from the exon boundaries. Kmers are then matched to the reads. Matched kmers that are in sequence with the reference are considered linear spliced reads (top exons). Matched kmers that are out of sequence to the reference are considered circRNA back spliced junction reads (bottom exons).*

Because only reads that align across the BSJ can be confidently assigned, it is hazardous to make assumptions as to which exons are, or are not, included in multi-exonic circRNAs. Unfortunately, this does not appear to be the case for several circRNA databases, which include many multi-exonic transcripts that have not been experimentally validated [272]. Lastly, as the number of studies using the various algorithms continues to grow so does the list of identified circRNAs and circRNA databases (currently 20 databases; 14 noncurated; six

curated based on literature searches for empirically validated circRNAs). To date there are between 1953 - 1,223,114 noncurated or 249 -3181 curated circRNAs in databases, and there appears to be little overlap between them. For an excellent review of the current state of circRNA databases see Vromman *et al. [272]*

2.3.4 CircRNA functions.

2.3.4.1 CircRNAs as miRNA sponges

The first papers to demonstrate functional roles for circRNAs found that the circRNAs miR-7 *(ciRS-7)* and sex-determining region Y (circSRY) bound and sequestered miRNAs miR-7 and miR-138, via 70 and 16 conserved binding sites, respectively [247, 273]. Both circRNAs strongly supressed the ability of the miRNAs to bind to its target mRNA leading to increased expression of the mRNA, suggesting a competing endogenous RNA regulation of the mRNA.

CircRNAs may bind one or more miRNAs and depending on whether there is partial or complete binding of the miRNA may result in different mechanisms of inhibition or degradation [274]. Piwecka *et al [274]* demonstrated that the circRNA Cdr1as has > 70 biding sites for miRNA-7 that partially binds at the seed region, which they suggest alters the availability of this miRNA. Cdr1as also has a binding site for miRNA-671 that has almost full complementarity and which could direct Argonaute protein mediated cleavage of the circRNA, as has been previously demonstrated [275].

This sponge like mechanism caused great excitement and anticipation of a general functional mechanism for circRNAs. However, bioinformatic analyses suggests that circRNA miRNA-binding sites are no more enriched than would be expected by chance [276] and sequestering miRNAs appears to be just one of several emerging functional roles for circRNAs (Figure 2-10).

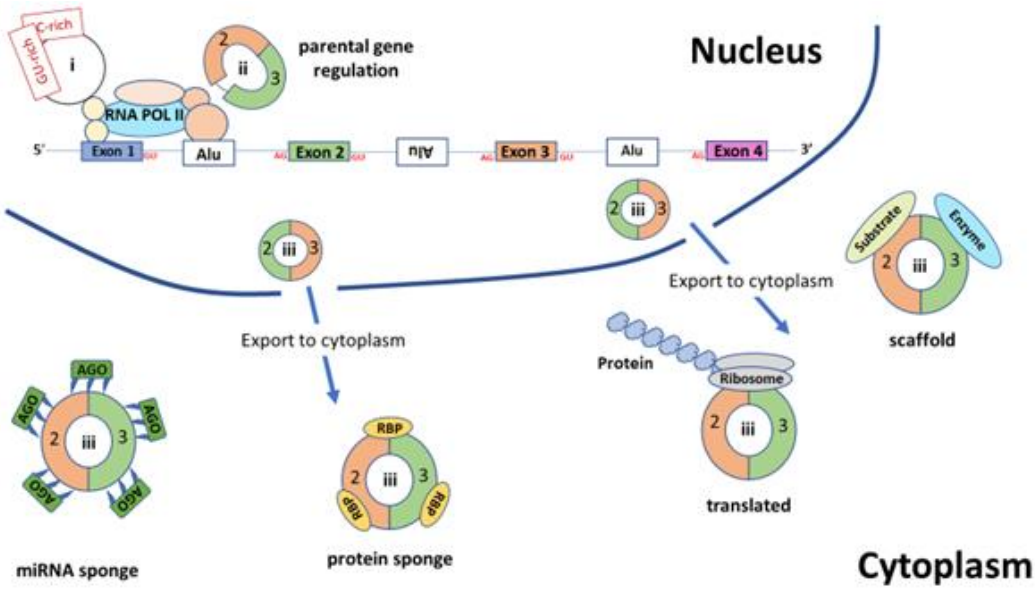


Figure 2-10 An overview of the various functions for circRNAs.

Within the nucleus, intronic circRNAs (ciRNAs (i)) and exon and intron containing circRNAs (EIciRNAs (ii)) are involved in regulation of their parental gene. Exonic circRNAs(iii) are exported to the cytoplasm where they can either act as miRNA sponges (which also bind the Argonaute protein – an essential component of the RNA-induced silencing complex (RISC) which acts upon the targeted mRNA) , RBP sponges, protein scaffolds or are translated. RBP: RNA binding protein, Ago: Argonaute protein, Alu: Alu elements

2.3.4.2 CircRNAs as protein sponges, regulators, and scaffolds

As well as being involved in circRNA biogenesis, emerging evidence suggests that RBPs interact with circRNAs both within the nucleus and cytoplasm. CircRNAs can act as (i) protein sponges to block protein activity [196, 277], (ii) positive regulators of polymerase II to regulate the transcription of their parental genes in the nucleus (examples include circEIF3J, circPAIP2 and ci-ankrd52 [224, 278]), and (iii) protein scaffolds to colocalise enzymes and their substrates in the cytoplasm thereby facilitating recruitment [279] and enzyme activity [280, 281].

2.3.4.3 CircRNAs as protein-coding transcripts

The notion of an alternative circRNA transcriptome is appealing given that circRNAs are predominantly located in the cytoplasm, contain exonic (therefore potential coding) sequences, sometimes contain the canonical AUG initiator codon of the associated mRNA [282] and, could potentially have an infinite open reading frame (ORF) because of their

circular structure (if a stop codon is not encountered beyond the translational start site) [283]. Early studies used ribosome profiling which identifies RNA fragments that are protected by the ribosome and escape ribonucleases. Ribosome protected fragments indicate translation sites so if circRNA back spliced reads overlapped these then this would suggest that the circRNAs are being translated. However, the studies found no evidence of translation [218, 276].

For circRNA translation to occur, it would have to be via an alternative mechanism to linear translation. In eukaryotic translation, the 5'cap of the mRNA transcript is recognised by the eukaryotic initiation factor 4E (eIF4E) complex, which recruits ribosomes to translate the transcript into protein. While circRNA transcripts lack a 5' cap and cannot be translated into protein via this mechanism, there is growing evidence for their translation by alternative mechanisms. Work on viruses and cellular physiological stress responses identified an alternative, cap-independent translation mechanism involving RNA structures known as internal ribosome entry sites (IRESs) which recruit ribosomes to an internal start codon to initiate translation [284]. It has been shown that circRNAs synthesised *in vitro* carrying engineered IRESs can be translated in this manner [285, 286] but endogenous IRESs are rare. Recent bioinformatic analysis identified IRES-like AU-rich hexamers that bind *trans*-acting factors to drive circRNA cap-independent translation in a green fluorescent protein (GFP) cell-based reporter system [287]. The authors predicted that a nucleotide sequence longer than 50 nucleotides would contain an IRES-like hexamer by chance. More than 99% of circRNAs are longer than 100 nucleotides, suggesting that most circRNAs could potentially be translated by this mechanism [287].

In addition to IRESs, methylated adenosine residues (N^6-methyladenosines (m^6A), the most abundant base modification of RNA) in the 5' UTR region can directly recruit the eukaryotic initiation factor 3 (eIF3) and initiate translation in the absence of the cap-binding factor eIF4E [288]. N6-methyladenosine (m^6A) consensus motifs are enriched in circRNAs and a single

m^6A site was found to be sufficient to drive translation in human cells [289]. For example, endogenous circ-ZNF609 is expressed in human and mouse myoblasts and contains a 753-nt open reading frame. Bozzoni *et al* used a CRISPR/Cas9 system in mouse embryonic stem cells to produce constructs expressing circular tagged transcripts which were translated into protein. The authors did not determine the molecular activity of the proteins, but Circ-ZNF609 was found to control myogenic proliferation [290]. Other studies have demonstrated circRNA encoded proteins that suppressed the cell cycle and glioma cell proliferation [291-293], promoting the growth and metastasis of colon cancer [294] and liver cancer cell growth [295].

From the literature it would appear that the predominant functional mechanism of circRNAs is to sequester cytoplasmic miRNAs and, in doing so, participate in regulation of their downstream mRNA targets. However, this focus on miRNAs may reflect the finding that the first circRNAs to be discovered were potent sponges of miRNAs [247, 273] and the relative technical ease of demonstrating this *in vitro*.

From the literature there are three *in vitro* methods to demonstrate circRNA/miRNA interaction. The first approach is a luciferase reporter assay which co-transfects the circRNA sequence along with the miRNA of interest downstream of the luciferase gene and measure luciferase activity. If the miRNA binds to the circRNA sequence, then luciferase activity is inhibited [296]. This demonstrates that the miRNA can bind to this circRNA mimic *in vitro* but not that actual circRNA *in vivo*. The second approach uses a biotin labelled probe to co-precipitate or 'pull down' the circRNA/miRNA. This is more compelling than the previous method in that the actual interaction of the circRNA/miRNA can be demonstrated *in vivo* [297] The third method is RNA immunoprecipitation (RIP) or crosslinking-immunoprecipitation (CLIP) of circRNA by an argonaute 2 protein (AGO2) antibody. The mechanism by which miRNA carries out target degradation is via the RNA-induced silencing

complex (RISC) complex which contains the AGO2 protein. This method demonstrates both circRNA/miRNA interaction and the RNA silencing mechanism.

Robust experimental validation should be the priority as non-specific binding can occur. This was illustrated in a study by Lim *et al. [298],* investigating circSlc8a1. Biotin pull down of circSlc8a1-bound miRNAs resulted in 14 miRNAs being detected which, after Benjamini–Hochberg correction and filtering for a magnitude fold-change >3, led to prioritisation of five candidate miRNAs. Three of these showed significant and consistent inhibition of luciferase activity, with two having *in silico* prediction for several binding sites. The authors repeated the biotin-based pull-down assay using isolated mouse cardiomyocytes for these two miRNAs and saw a 180- and 18-fold enrichment, respectively. It was not until they performed an inverse pull-down assay to test for endogenous circSlc8a1 binding by using biotinylated miRNA mimics that they demonstrated only one of these miRNAs had a robust endogenous interaction between the circRNA and the miRNA.

While circRNA-miRNA sponge mechanism studies are exposing important candidate targets for circRNAs, other gene regulatory mechanisms are now gaining traction and should be explored in parallel, when characterising circRNA function.

2.3.5 CircRNAs in Cardiovascular Disease

There is growing evidence that circRNAs have potential functional roles in cardiac pathologies including atherosclerosis, myocardial infarction, cardiac fibrosis, cardiac hypertrophy, and HF (as demonstrated by the tables below). As we unravel the role of circRNAs in the development and progression of CVD, their potential as therapeutic targets to reverse pathophysiological remodelling or enhance protective mechanisms will become apparent [299].

2.3.5.1 CircRNAs as novel biomarkers

CircRNAs have been reported in plasma, serum, saliva and urine [227, 300-302] and are enriched in biofluids compared to tissues [303]. In addition to their roles in the nucleus and cytoplasm, circRNAs are released from healthy, apoptotic and necrotic cells into biofluids and due to their closed structure are considerably more stable than linear mRNAs, making them excellent candidate biomarkers for the presence and progression of a range of cardiovascular diseases.

Presently only a few studies have demonstrated the use of circRNAs as biomarkers for CAD and MI (summarised in Table 2-4). Although these studies highlight the potential of circRNAs as biomarkers, we should proceed with caution. Studies suffer from small cohort sizes, a lack of independent replication and lack of consideration for confounding medications commonly used for treatment of IHD and HF. Circulating circRNAs levels may change rapidly through disease development and progression, and so timing of collection needs to be considered. As yet, there is little consensus between the studies as to which circRNAs are the lead candidates and more work is needed to identify these. Moreover, these studies have been carried out using targeted technologies of microarray or RT-qPCR highlighting the potential for more discoveries to come with the use of more sensitive RNA-Seq coupled with specialised library preparation protocols for circRNA enrichment.

2.3.5.2 CircRNAs- miRNAs in Atherosclerosis (AS)

Table 2-5 summarises current knowledge of the involvement of circRNAs in atherosclerosis. Of the 10 papers published to date, eight studies have identified the mechanism underlying the association between the circRNA and atherosclerosis, with seven studies identifying *in vitro* circRNA/miRNA associations and one showing a circRNA/protein interaction. Notably, for three circRNAs (circANRIL, circCHFR and circ_0003204), associations with Atherosclerosis have been reported by two independent studies.

Table 2-4 CircRNAs as potential biomarkers

CircRNA	CVD category	Findings of the study	Methods	Reference
hsa_circ_0001445	Coronary Atherosclerosis	hsa_circ_0001445 was remarkably stable in plasma. Plasma levels of hsa_circ_0001445 were proportional to coronary atherosclerotic burden and improved detection of coronary artery atherosclerosis.	RT-qPCR in plasma from 200 patients with suspected stable CAD.	Vilades., *et al* 2020 [226]
Hsa_circ_0001879 and Hsa_circ_0004104	CAD	Both significantly upregulated in CAD patients. Overexpression of hsa_circ_0004104 *in vitro* resulted in dysregulation of atherosclerosis-related genes	Microarray and RT-qPCR in whole blood from 24 CAD patients and 7 healthy controls	Wang *et al* 2019 [304]
Hsa_circ_0124644	CAD	First study to investigate the circRNA profile in the peripheral blood of CAD patients. Circulating levels of hsa_circ_0124644 improved detection of CAD.	Microarray in peripheral blood from 12 CAD patients and 12 healthy controls. Validation in 137 CAD patients and 115 healthy controls	Zhao et a 2017 l [305]
Myocardial Infarction-Associated CircRNA (MICRA)	Myocardial Infarction	Circulating levels of MICRA were lower in MI patients. MICRA was a strong predictor of left ventricle (LV) dysfunction; patients with lower levels of MICRA were at higher risk of LV dysfunction	RT-qPCR in whole blood from 642 acute MI patients, 86 heathy controls	Vausort *et al* 2016 [306]
Myocardial Infarction-Associated CircRNA (MICRA)	Myocardial Infarction	MICRA classified patients into ejection fraction (EF) groups. Improved risk classification after MI	RT-qPCR in whole blood from472 acute MI patients classified into 3 groups according to ejection fraction	Salgado-Somoza *et al* 2017 [307]

Of the seven circRNA/miRNA studies, only two used the more compelling RNA pulldown method to validate the circRNA/miRNA interaction, with the remaining studies using the less compelling luciferase method. In fact, two studies to use the luciferase method suggest two different miRNAs are involved in the functional mechanisms for the same circCHFR.

Table 2-5 CircRNAs associated with atherosclerosis

CircRNA/miRNA interaction					
CircRNA	**miRNA target**	**Protein target**	**Findings of the study**	**Methods**	**Reference**
CircRNA-0044073	miR-107	JAK/STAT	circRNA-0044073 is upregulated in atherosclerosis and promotes the proliferation and invasion of cells by targeting miR-107 and activating the JAK/STAT signalling pathway	RT-qPCR in blood cells from patients with atherosclerosis and healthy controls (n=20 total). RNA pulldown for miRNA interaction.	Shen et al 2019 [308]
Circ_CHFR	miR-370	FOXO1	Knock down of circ_CHFR inhibits the proliferation and migration of VSMCs *in vitro*	Microarray on human vascular smooth muscle cells (VSMCs). Luciferase reporter assay for miRNA interaction	Yang et al 2019 [309]
Circ_CHFR	miR-214-3p	Wnt3/β-catenin pathway	Circ_CHFR was up-regulated in atherosclerotic serum and ox-LDL-stimulated VSMC. Circ_CHFR regulated cell growth, migration, and inflammation via regulating the expression of Wnt3 as a competitive endogenous RNA (ceRNA) of miR-214 in ox-LDL-treated VSMCs	RT-qPCR in atherosclerotic (n=32) and control patient's serum (n=32). Cell model of atherosclerosis in (VSMCs) with oxidized low-density lipoprotein (ox-LDL). Luciferase reporter assay for miRNA interaction	Zhuang et al 2020 [310]
Circ-SATB2	miR-939	Stromal Interaction Molecule 1 (STIM1)	circ-SATB2 and STIM1 up-regulated in proliferative VSMCs, miR-939 down-regulated. circ-SATB2 can regulate VSMC phenotypic differentiation, proliferation, apoptosis and migration by promoting the expression of STIM1	RT-qPCR in VSMCs. Dual luciferase assay for miRNA interaction	Mao et al 2018 [311]
circ-Sirt1	miR-132/212	nuclear factor kappa-light-chain-enhancer of activated B cells (NF-κB)	circ-Sirt1 inhibits inflammatory phenotypic switching of VSMCs. circ-Sirt1 controls NF-κB activation via enhancement of SIRT1 expression by binding to miR-132/212 in vascular smooth muscle cells	RT-qPCR and microarray in human arterial tissues (11 atherosclerotic, 14 control), human plasma (20 CAD, 20 control) and rat VSMCs. RNA pulldown for miRNA interaction	Kong et al 2019 [312]
circ_0003204	miR-370-3p	TGFβR2/phosph-SMAD3	CircRNA circ_0003204 inhibits proliferation, migration and tube formation of endothelial cell in atherosclerosis via miR-370-3p/TGFβR2/phosph-SMAD3 axis	RT-qPCR in human aorta endothelial cells (HAECs). Luciferase activity assay	Zhang et al 2019 [313]
Circ_RUSC2	miR-661	Tyrosine-protein kinase (SYK)	Circ_RUSC2 can promote the expression of SYK, a target gene of miR-661, and regulates VSMC proliferation, apoptosis, phenotypic modulation, and migration	RT-qPCR in human coronary artery smooth muscle cells. Dual luciferase assay	Sun et al 2019 [314]

circRNA/Protein interaction				
CircRNA	Protein target	Findings of the study	Methods	Reference
circANRIL	Pescadillo Homologue 1 (PES1)	binds pescadillo homologue 1 (PES1) to prevent rRNA maturation. This induces nucleolar stress and p53 activation which leads to apoptosis and proliferation inhibition. Atheroprotection is achieved by culling over-proliferating cell types in atherosclerotic plaques.	RT-qPCR for circRNA expression levels from peripheral blood mononuclear cell (PBMC), human primary tissues and cells. CAD patients (n=2,280) undergoing coronary angiography. Also, human endarterectomy specimens (n=218) collected in a cohort of patients undergoing vascular surgery. Genome-wide expression arrays for cells overexpressing circANRIL. λN-Peptide-mediated pull-down of circANRIL-bound proteins	Holdt et al 2016 [196]
Mechanism unknown				
CircRNA	Findings of the study		Methods	Reference
circANRIL	Reduced circANRIL expression could prevent coronary AS by reducing vascular EC apoptosis and inflammatory factor expression		Atherosclerosis rat model (n=60) 5 groups: control, model, empty vector group, over-expressed circANRIL group and low-expressed circANRIL group. serum lipids, apoptosis, protein and mRNA levels analysed in the 5 groups	Song et al 2017 [315]
has_circ_0003204	Oxidized low-density lipoprotein (oxLDL) plays a vital part in the initiation of AS. hsa_circ_0003204 was notably upregulated in HUVECs treated with oxLDL indicating it may be involved in the pathogenesis of atherosclerosis. Knockdown boosted the proliferation, migration, and invasion of oxLDL-induced HUVECs, suggesting that hsa_circ_0003204 may slow down the repair of vascular endothelial cell injury.		RT-PCR from (OxLDL) Human umbilical vein endothelial cells (HUVECs)	Liu et al 2020 [316]

CircANRIL investigated by Holdt *et al* [196] and Song *et al* [315] is interesting as it originates from chromosome 9p21 which houses a lncRNA called ANRIL (described above). Increased linear ANRIL is associated with increased atherosclerosis [170] and multiple linear isoforms as well as circular isoforms have been identified [196, 317]. circANRIL is expressed in human vascular tissue, smooth muscle cells and monocyte/macrophages, all of which play an important role in atherogenesis. The study by Holdt *et al* [196] suggests circANRIL confers atheroprotection. However, another study looking at the effect of circANRIL on the inflammatory response of vascular endothelial cells (ECs) in a rat model of coronary atherosclerosis suggests that expression of circANRIL was correlated with the expression of inflammatory factors in vascular ECs, and the over-expression of circANRIL could exacerbate vascular EC inflammation and promote atherosclerosis. Further studies are needed to confirm whether circANRIL has a protective or antagonistic role in atherosclerosis. [315].

Together these studies provide evidence to suggest circRNAs play a regulatory role in the development and progression of atherosclerosis. To date four studies [196, 226, 308, 310] have reported altered levels of circRNAs in biofluids in patients with atherosclerosis highlighting a potential role for these circRNAs (*circRNA_0001445, circRNA-0044073*, circCHFR, circANRIL) as potential biomarkers.

2.3.5.3 Myocardial infarction / Ischaemia Reperfusion injury

Accumulation of atherosclerotic plaque can lead to myocardial ischaemia, associated with a cascade of cellular, inflammatory and biochemical events [318] subclinical myocardial dysfunction and ultimately MI [319] Cardiomyocytes have limited regenerative capacity and restoration of blood supply or reperfusion is the current clinical procedure to limit death of cardiomyocytes but reperfusion itself can cause injury to the site [320]. There is growing evidence that circRNAs are involved cellular responses to MI and ischaemia reperfusion (IR), either through miRNA regulation or interaction with proteins in the cell cytoplasm.

Of 18 circRNAs associated with MI/IR to date, 13 circRNAs have been suggested to regulate MI/IR induced apoptosis in cardiomyocytes via interaction with miRNAs in animal models or human cells *in vitro* (Table 2-6) and represent potential therapeutic targets in this setting. Geng *et al* [296]were the first study to suggest this mechanism, suggesting CDR1as, acted as a sponge of miR-7 and led to an increase in apoptosis in cultured mouse cardiomyocytes. However, this study demonstrated no *direct* evidence for this interaction. Instead, overexpression of CDR1as through administration of an expression plasmid via intracardiac injection in a mouse model of MI led to an increase in infarct size which could be reversed by overexpression of miR-7a. A further three studies [321-323] provide no direct evidence for circRNA/miRNA interaction and instead provide evidence with overexpression of one affecting expression levels on the other.

Conversely, the study by Wang *et al* [324] was the first to use the more compelling method of AGO2 pulldown to not only demonstrate direct evidence of circMFACR interaction with miR-652-3p but also the AGO2 protein *in vivo*. A further three studies also provide evidence of circRNA/miRNA interaction with AGO2 pulldown [325-327]. In addition, four studies have demonstrated circRNA/protein binding mechanisms involved in MI/IR injury [328-331].

Interestingly and perhaps surprisingly, it was not until the Huang *et al* [330] study in 2019 that RNA-Seq was used to identify CircNfix (albeit using RNA-Seq data provided from another study). As yet, there is no confirmation of the same circRNA between studies with no consensus as to which are the lead circRNA candidates involved with MI or ischaemia reperfusion injury.

Table 2-6 CircRNAs associated with myocardial infarction or ischaemia reperfusion injury

CircRNA/miRNA interaction					
CircRNA	**miRNA target**	**Protein target**	**Findings of the study**	**Methods**	**Reference**
CDR1as	miR-7a	poly(ADP-ribose) polymerase (PRP) and SP/KLF family of transcription factors (SP1)	Overexpression of CDR1 in vivo promoted apoptosis and increased cardiac infarct size; overexpression of miR-7a had the opposite effect	RT-qPCR from mouse cardiomyocytes. No direct evidence of circRNA/miRNA interaction. CDR1as induced apoptosis could be reverse by miR-7 overexpression	Geng et al 2016 [296]
Mitochondrial fission and apoptosis-related circRNA (MFACR)	miR-652-3p	MTP18	Regulated mitochondrial fission and apoptosis in the heart	Mouse cardiomyocytes. AGO2 immunoprecipitation	Wang et al 2017 [324]
ncx1	miR-133a-3p	pro-apoptotic gene cell death-inducing protein (CDIP1)	Increased in response to reactive oxygen species (ROS) and promotes cardiomyocyte apoptosis	H9c2 cells and neonatal rat cardiomyocytes. Biotinylated pulldown of miRNA-133 and circMFACR. AGO2 immunoprecipitation	Li et al 2018 [325]
CircHIPK3	CircHIPK3 by binding to miRNA-124-3p	overexpression upregulated pro-apoptotic Bax, and downregulated anti-apoptotic Bcl-2	Inhibited proliferative ability and induced apoptosis of cardiomyocytes after myocardial IR injury	Mouse cardiomyocytes. Dual-Luciferase Reporter Gene Assay	Bai et al 2019 [332]
Ttc3	miR-15b	ADP ribosylation factor like 2 (Arl2)	Played a cardioprotective role after myocardial infarction. Overexpression counteracted hypoxia-induced ATP depletion and apoptotic death.	Rat cardiomyocytes and cardiac fibroblasts. Dual-luciferase reporter assay	Cai et al 2019 [333]
circ_0010729	miR-145-5p	mTOR and MEK/ERK pathways	Silencing circRNA circ_0010729 protected human cardiomyocytes from oxygen-glucose deprivation-induced injury	4h oxygen-glucose-deprivation (OGD) human cardiomyocytes. No direct evidence for circRNA/miRNA interaction. Expression studies after Overexpression of one against the other.	Jin et al 2019 [321]
CircMACF1	miR-500b-5p	epithelial membrane protein 1 (EMP1)	Attenuated Acute Myocardial Infarction	Mouse cardiomyocytes. Dual Luciferase Reporter Gene Assay	Zhao et al 2020 [334]

CircRNA_101237	let-7a-5p	insulin-like-growth factor 2 mRNA-binding protein 3 (IGF2BP3)	Mediated anoxia/reoxygenation injury	Mouse cardiomyocytes Biotinylated miRNA pull-down. AGO2 immunoprecipitation	Gan et al 2020 [326]
Circ_LAS1L	miR-125b	secreted frizzled-related protein 5 (SFRP5)	Down-regulated in acute myocardial infarction (AMI). regulates cardiac fibroblast activation, growth, and migration and apoptosis	Mouse cardiomyocytes. Luciferase reporter gene assay. Biotinylated RNA pull down. AGO2 immunoprecipitation	Sun et al 2020 [327]
circDLPAG4/HECTD1	miR-143	HECT Domain E3 Ubiquitin Protein Ligase 1 (HECTD1)	circDLPAG4/HECTD1 played a vital role in apoptosis and cell migration in endothelial cells through ER stress in response to ischaemia/reperfusion injury	Primary Human Umbilical Vein Endothelial Cells (HUVEC)/mice. No direct evidence for circRNA/miRNA interaction. Expression studies after Overexpression of one against the other.	Chen et al 2020 [322]
circRNA Pan3	miR-31-5p	quaking (QKI)	Silencing of miR-31-5p significantly alleviated the myocardial apoptosis induced by Doxorubicin (DOX) treatment. MiR-31-5p acts as a negative regulator of circPan3 by directly suppressing QKI *both in vivo* and *in vitro*	Mouse cardiomyocytes. No direct evidence for circRNA/miRNA interaction. Dual-luciferase reporter gene assay of QKI/ miR-31-5p interaction. RNA immunoprecipitation of QKI/circPan3 interaction. Expression levels of circPan3 when miR-31-5p was inhibited	Ji et al 2020 [323]
circ_0007623	miR-297	Vascular Endothelial Growth Factor A (VEGFA)	hsa_circ_0007623 promoted cardiac repair after acute myocardial ischemia, and protect cardiac function	hypoxia-induced human umbilical vein endothelial cells (HUVECs). Dual luciferase reporter gene assay of circ_0007623/miR-297	Zhang et al 2020 [335]
circDENND2A	miR-34a	may work via β-catenin and Ras/Raf/MEK/ERK pathways	Overexpression of circDENND2A enhanced cell viability and migration but declined apoptosis under oxygen glucose deprivation (OGD)	Rat H9c2 cells. Luciferase activity assay For interaction of circDENND2A/ miR-34a	Shao et al 2020 [336]
circRNA/Protein interaction					
CircRNA	**Protein target**		**Findings of the study**	**Methods**	**Reference**
Autophagy-related circRNA (ACR)	DNA (cytosine-5-)-methyltransferase 3 beta (Dnmt3B)		Inhibition of autophagy – a type of cell death protects cardiomyocytes during ischemia/reperfusion. circRNA ACR was able to inhibit autophagy and cell death in	Microarray on Mouse cardiomyocytes to look at differential expression in autophagy induced mouse hearts by I/R injury. Transcriptome microarray after	Zhou et al 2019 [331]

		cardiomyocytes to protect the heart from reducing I/R induced infarct sizes. ARC activated PTEN-induced kinase 1 (PINK1) expression (an autophagy associated gene) by directly binding to and blocking Dnmt3B-mediated methylation of the Pink1 promoter.	ACR knockdown. RNA-binding protein immunoprecipitation (RIP) of ACR/Dnmt3B association. Chromatin immunoprecipitation (ChIP) assay showing Dnmt3B/Pink1 association	
CircNfix	Y-Box Binding Protein 1 (Ybx1)	Ybx1 is a transcription factor that is involved with cardiomyocyte differentiation. Interaction between circNfix and Ybx1 prevented nuclear translocation of Ybx1, causing cytoplasmic retention and degradation. Nfix also bound to miR-214, which has been demonstrated to be essential for cardiomyocyte proliferation. Loss of circNfix induced cardiac regeneration and angiogenesis and inhibited cardiomyocyte apoptosis after MI, which significantly restored cardiac function and improved the prognosis	Used RNA-Seq data from another study [337] to filter for circRNAs differentially expressed in human, mouse and rat MI vs control/sham. RT-qPCR of circNfix. RNA pulldown assays and mass spectrometry analysis for Nfix/Ybx1 association. Luciferase, RNA pulldown for Nfix/mIR-214 association	Huang et al 2019 [330]
circFndc3b	Fused in Sarcoma (FUS)	circFndc3b significantly downregulated in post MI mouse hearts and in human cardiac tissues of ischemic cardiomyopathy patients. Overexpression in cardiac endothelial cells increases vascular endothelial growth factor-A (VEGF-A) expression which enhances angiogenic activity and reduces cardiomyocytes and endothelial cell apoptosis. circFndc3b interacts with the RNA binding protein Fused in Sarcoma to regulate VEGF expression and signalling	Microarrays on sham or MI mouse hearts (n = 2) and ischemic cardiomyopathy human heart tissue (n=7), normal heart tissue (n=4). RNA Immunoprecipitation (RIP)	Garikipati et al 2019 [329]
circFoxo3	p21-CDK2	circ-Foxo3 highly expressed in the tissues of aged mice and patients. Ectopic expression of circFoxo3 induced cellular senescence of mouse embryonic fibroblasts (MEFs), where it interacted with the anti-senescence proteins ID1 and E2F1, and anti-stress proteins FAK and HIF1α. These interactions prevented nuclear translocation of these transcription factors, causing cytoplasmic retention and reduced function	Primary cardiomyocytes isolated from neonatal and 12-week heart mouse tissues (n=20) RT-PCR for expression levels. Antibody pulldown for protein/circRNA association	Du et al 2017 [328]

Mechanism unknown				
CircRNA	**Protein target**	**Findings of the study**	**Methods**	**Reference**
CircRNA 010567	May be related to the inhibition on the TGF-β1 signalling pathway	Cardiac function and myocardial infarction (MI)-induced myocardial fibrosis (MF) improved, myocardial apoptosis was ameliorated in circRNA 010567 siRNA group compared with that in Model group. Regulatory mechanism may be related to the inhibition on the TGF-β1 signalling pathway	Rat MI model by ligation of the left anterior descending coronary artery. Model rats were randomly divided into circRNA 010567 siRNA group and Model group, with sham operation group as Control group (n=30). The effects of circRNA 010567 on myocardial infarction (MI)-induced myocardial fibrosis (MF), myocardial apoptosis, mRNA, and protein expression levels of TGF-β1 and Smad3 in heart tissues of MI rats were detected using the small animal ultrasound system, Masson staining, terminal deoxynucleotidyl transferase-mediated dUTP nick end labelling (TUNEL) staining, reverse transcription-polymerase chain reaction (RT-qPCR), and Western blotting	Bai et al 2020 [338]

2.3.5.4 CircRNAs- miRNAs in Heart Failure/Hypertrophy/Cardiac fibrosis

Once damage has occurred through MI or chronic hypertension, the heart undergoes macro- and microscopic remodelling [339, 340] An early infiltrative inflammatory response is followed by replacement of infarcted myocardium by non-elastic fibrotic tissue. Diffuse interstitial fibrosis may also affect areas of the heart remote from the initial injury. The architecture of the left ventricle exhibits mechanically disadvantageous changes, including overall ventricular dilatation and alteration from an efficient elliptical shape to a spherical chamber morphology. In vulnerable patients, these processes can lead to subclinical myocardial dysfunction and subsequent symptomatic HF [339-341]. Cardiac fibrosis can be caused by abnormal deposition of extra cellular matrix in the myocardium. It can be the result of scarring after MI but can be more widespread and common in HF. The scarring is detrimental by either stiffening the myocardium thereby reducing the pumping ability of the heart by impairing electrical conductance [342].

Table 2-7 summarises nine studies on circRNAs that have been demonstrated regulate gene expression in HF, hypertrophy, and cardiac fibrosis. All 9 studies suggest a circRNA-miRNA-protein axis. Only one study Han *et al* [343] was carried out in human with the remaining studies being carried out in mice. Two studies suggest a role of the circHIPK3 which has been demonstrated to be a key circRNA in a variety of cancers [344] and has been associated with several miRNAs. However, Ni *et al* [345] did not show direct evidence for a circHIPK3/miRNA interaction and only demonstrated cytoplasmic colocalisation by RNA Fluorescence in situ hybridization (FISH).

Table 2-7 CircRNAs associated with Heart Failure/Hypertrophy and Cardiac Fibrosis

CircRNA	miRNA target	Protein target	Action	Cell/Tissue type	Reference
Heart-related circRNA (HRCR)	miR-223	Apoptosis repressor with CARD domain (ARC)	miR-223 acted as a positive regulator of cardiac hypertrophy. Increased expression of HRCR sequesters free miR-223	RT-PCR of heart related circRNAs responsive to saline- vs. isoproterenol-infused mice. (n=7) and transverse aortic constriction or to sham treatment (n=8) Biotinylated RNA pulldown. AGO2 immunoprecipitation	Wang et al 2016 [346]
CircRNA_000203	miR-26b-5p	Collagen alpha 2(I) (Col1a2) and connective tissue growth factor (CTGF)	over-expression of circRNA_000203 could eliminate the anti-fibrosis effect of miR-26b-5p in cardiac fibroblasts	CircRNA expression analysis with microarray from mouse myocardium (n=8 model, n=8 control Biotinylated RNA pull-down of circRNA/miRNA. Luciferase assay for miRNA/mRNA targets	Tang et al 2017 [347]
circRNA_010567	miR-141	TGF-β1	circRNA_010567 promoted myocardial fibrosis via suppressing miR-141 by targeting TGF-β1	Mouse cardiac fibroblasts. CircRNA expression with microarray confirmed with RT-PCR. Luciferase assay for circRNA/miRNA association.	Zhou et al 2017 [348]
Circ-HIPK3	miR-17-3p	Adenylyl cyclase type 6 (ADCY6)	Demonstrated an increase of Adenylate cyclase type 6 (ADCY6) caused by circ-HIPK3 which was ameliorated by miR-17-3p overexpression and vice versa, implicates a circ-HIPK3 - miR-17-3p - ADCY6 axis. Downregulation of circ-HIPK3 can alleviate fibrosis and maintain cardiac function post MI in mice	24 mice divided into 4 groups - normal group (without surgery), control group (without ligation), NC (negative control) group and experiment group. Dual luciferase expression vectors with circ-HIPK3 co-transfected with miR-17-3p	Deng et al 2019 [349]
circSlc8a1	miR-133a		Adenovirus (AAV9)-mediated RNAi knockdown of circSlc8a1 attenuates cardiac hypertrophy from pressure-overload, whereas forced cardiomyocyte specific overexpression of circSlc8a1 resulted in heart failure	Mouse cardiomyocytes. Biotin-based pull-down of circRNA/miRNA association then qPCR of miRNA and luciferase assay	Lim et al 2019 [298]
circNFIB (nuclear factor 1 B-type)	miR-433	AZIN1 and JNK1	circNFIB overexpression reduced cardiac fibroblast proliferation (a process involved with fibrosis) based on TGF-β stimulation), while inhibition of circNFIB promoted fibroblast proliferation. Action via miR-433.	RT-PCR of candidate circRNA expression levels in MI fibrosis mice (n=5 control, n=6 MI) and cardiac fibroblasts. miRNA	Zhu et al 2019 [350]

				association with dual luciferase reporter assay system.	
circHIPK3	miR-29b-3p	α-smooth muscle actin(α-SMA), Collagen type I alpha I (COL1A1, COL3A1	circHIPK3 silencing reduced cardiac fibroblast proliferation, migration and the upregulation of a-SMA expression levels induced by Ang II in vitro	Mouse cardiac fibroblasts (CFs).No direct evidence for circnRNA/miRNA interaction RT-PCR for circRNA expression level. RNA FISH to determine whether circHIPK3 and miR-29b-3p colocalise in CFs.	Ni et al 2019 [345]
circRNA_000203	miR26b-5p and miR-140-3p	Gata4	CircRNA_000203 was found to be upregulated in the myocardium of cardiac hypertrophy induced mice. Demonstrated that circRNA_000203 sponged miR-26b-5p, -140-3p, abolished the suppression of Gata4 by miR-26b-5p, -140-3p, resulting in the increase of GATA4 in NMVCs	Neonatal mouse ventricular cardiomyocytes (NMVCs). Dual-luciferase assays of circRNA_000203/miR-26b-5p, and miR-140-3p Biotinylated RNA pull-down assay between circRNA_000203/ miR-26b-5p,and miR-140-3pincubated	Li et al 2020 [351]
Hsa_circ_0097435	hsa_miR_6799_5P, hsa_miR_5000_5P, hsa_miR_609 and hsa_miR_1294	unknown	Overexpression of circ_0097435 promoted cardiomyocyte apoptosis, Silencing hsa_circ_0097435 inhibited apoptosis	RNA-Seq heart failure patients (n=5) vs control (n=4) Dysregulated expression confirmed with qPCR 40 patients with heart failure. RNA-pulldown experiments and AGO2-immunoprecipitation experiments revealed that hsa_circ_0097435 sponged multiple miRNAs	Han et al 2020 [343]

2.3.5.5 Summary

Growing evidence supports an important role of circRNAs in normal physiology and a range of disease states, including cardiovascular disease. Most studies have focussed on the circRNA/miRNA/mRNA axis which may be due to the first circRNAs being demonstrated as potent miRNA sponges and the relative ease of the *in vitro* methods. However, caution must be applied as several studies suffer from low cohort numbers and some show no direct evidence of the circRNA/miRNA interaction and many other use luciferase assays which demonstrate an interaction with circRNA/miRNA mimics *in vitro* but these should be validated with further, more robust experimental methods such as circRNA/AGO2 immunoprecipitation. Several other mechanisms are now emerging and whilst translated circRNAs have not been associated with cardiovascular disease, evidence for this mechanism in cancer biology [294, 295] suggests it is only be a matter of time before protein-coding cardiovascular circRNAs are identified. As yet, there are few independent replications involving the same circRNA and consequently there is no overlap of circRNAs within or between cardiovascular aetiologies.

It is surprising that so few studies have employed RNA-Seq for circRNA detection with most cardiovascular studies using microarray or RT-qPCR technologies, limiting detection of novel circRNAs. With the development of methods for circRNA enrichment in biofluids and more user-friendly software and bioinformatics pipelines future up take of RNA-Seq technologies for circRNA detection may be encouraged.

2.4 RNA sequencing and bioinformatic analysis pipeline development

As a large part of this thesis focuses on a bioinformatic pipeline developed to analyse these lncRNAs and circRNAs with RNA sequencing, the next section gives a brief introduction to RNA Seq and how it has developed over the past few decades. A later section (2-6) will guide the reader through the bioinformatics pipeline developed and the software chosen.

2.4.1 First and second-generation sequencing

Reducing costs have led to RNA Seq becoming the technique of choice for transcriptomics. RNA Seq is commonly used for quantifying gene expression and performing differential expression analysis of protein-coding genes [352]. However, with the development of new kits, we can now use RNA Seq to perform such techniques as targeted sequencing (to analyse a subset of genes of interest or specific regions of the genome) [353], RNA Seq from total RNA (including coding and non-coding RNA) [354], small RNA (small non-coding RNAs such as miRNAs) [355], ribosomal profiling (ribosome protected mRNAs to analyse which genes are being expressed at a specific time in a cell) [356], single cell RNA-Seq (analysing total RNA from a single cell) [357] and CLIP-Seq (allowing the generation of genome-wide maps of RNA binding protein – RNA interaction sites) [358].

The first manuscripts using RNA Seq started to appear in 2008 [359-362]. Prior to this, gene expression was analysed with probe-based microarrays and, before that, as expressed sequence tags (ESTs, ~200-500 nucleotide 'tags' of a transcript that were used for gene mapping) with Sanger sequencing. To date there have been three 'generations' of sequencing technology. The very first DNA sequencing method to be widely adopted was the Maxam and Gilbert chemical cleavage method, developed in 1976 [363] but this was superseded by Sanger Sequencing which, due to its very high accuracy rate and ease of use, has been the mainstay technique for low throughput sequencing since the late 1970s and is still used today. Collectively these methods are referred to as first generation sequencing.

Second generation sequencing refers to the development of high throughput, short read or massively parallel sequencing, so named because the RNA/DNA is fragmented and amplified, and millions of fragments are sequenced in parallel. The second-generation market has been dominated by the Illumina 'sequence by synthesis' method used in this thesis for RNA-Seq in both human heart tissue and human plasma. This method involves fragmenting the RNA to around 300 bases and reverse transcribing to make complementary DNA (cDNA), ligating

sequencing adapters to each end of each fragment, which are then PCR amplified and then passed over the flow cell – a glass slide with several channels or lanes in which millions of primers which are complementary to the adapters on either end of the fragments are bound. Next, there is an amplification step, which happens millions of times in parallel, termed bridge amplification for Illumina technology. In this step, the fragment is clonally amplified in a cluster on the flow cell so that image detection can take place. During a sequencing run there will be several million clusters in each lane of the flow cell.

Illumina's proprietary 'sequencing by synthesis' method uses these clonally amplified fragments as templates for sequencing. Fluorescently labelled nucleotides are added to extend the template one base at a time and, after each base addition, the imager captures the fluorescent signal. After addition of the last base, the synthesised strand is denatured and the fragment is 'turned' so that the fragment can be sequenced from the opposite direction – this is termed paired end sequencing [364]. This process generates millions of 'reads' of sequence, which then need to be aligned to the reference genome and quantified with bioinformatics the role of the pipeline developed in this project (discussed below).

It is worth mentioning here that a big advance in the capabilities of RNA-Seq is due to the evolution of different library preparation methods. RNA-Seq initially interrogated mRNA expression by selecting transcripts which have poly-A tails from total RNA. The purpose of this was to exclude rRNAs, which make up over 80% of the RNA in human cells [365], and avoid generating reads from transcripts not intended for analysis. Library preparation kits have since evolved to enrich for all RNAs from total RNA, not just those with poly-A tails, such as lncRNAs, circRNAs and miRNAs, while still excluding rRNA (by ribosomal depletion methods). Also, there are now stranded library kits which maintain information as to which DNA strand the RNA was transcribed from. In un-stranded libraries, strand of origin information is lost. In regions of the genome where genes overlap on sense and antisense strands, it is impossible to know which of the two strands is the correct strand with which to

align each fragment, and so which read belongs to which gene cannot be determined. As stranded libraries maintain strand of origin information, gene expression from overlapping genes can be quantified more accurately. This is especially important for lncRNAs as many overlap mRNAs on the opposite DNA strand and hence has particular relevance for the current project.

2.4.2 Third generation sequencing

The throughput capabilities of sequencing increased again with the evolution of third generation sequencing, which improves on the read length capability up to >100kb [366] compared to the second generation technologies. The two main players in the market currently are Pac Biosciences and Oxford Nanopore Technologies, with the latter used in this thesis and discussed below. For DNA applications, this has enabled sequencing of repetitive regions of the genome and has increased the accuracy of genome assembly. For RNA applications, it has allowed sequencing of whole transcripts from start to end and, with this, the ability to distinguish alternative isoforms. Long-read sequencing is achievable because the DNA or RNA does not have to be fragmented prior to sequencing and there is no need to amplify the transcript, as each transcript is read directly on the instrument (even direct RNA-Seq without the need to convert to cDNA is possible). The trade-off is that the error rate is higher than that of Illumina's <1% for short read sequencing (around 13% for both Pac Biosciences and Oxford Nanopore Technologies [367]), along with the need for a higher input of starting material.

Briefly, the Nanopore library preparation takes the input RNA and reverse transcribes it to make cDNA. At the end of this reverse transcribed strand, an additional 3 non-templated cytosine bases are added, and a PCR primer annealed to the non-templated cytosine bases. In this way, the PCR priming sequence is added to the end of full-length cDNA transcripts. After PCR, sequencing adapters are added. Nanopore sequencing works by feeding a single DNA (or RNA) molecule through a protein nanopore (a nano-scale hole) based on an electrical

resistant membrane on the flow cell. A voltage is passed across this membrane which sets an ionic current across the nanopore. When the DNA/RNA molecule passes through the pore each base or combination of several bases disrupts the current in a signature manner. Measurement of this signature allows identification of the bases and so the sequence of the whole fragment can be read [368].

2.5 Rational for Research

Accumulation of atherosclerotic plaque in the coronary arteries, the major blood vessels to the heart muscle, can restrict blood supply to the heart. When the blood supply does not meet the demands of the myocardium, myocardial ischaemia occurs, associated with a cascade of cellular, inflammatory, and biochemical events. Over time, these processes can lead to subclinical myocardial dysfunction and ultimately a MI, causing the death of heart muscle cells. Once damage has occurred, the heart undergoes macro- and microscopic remodelling [331, 332] that can lead to subclinical myocardial dysfunction and subsequent symptomatic HF in vulnerable patients [331-333]. While the cardiac troponins and natriuretic peptides have emerged as vital biomarkers in the clinical diagnosis of MI and HF, we lack specific markers for early myocardial ischaemia that could help identify cell dysfunction before cell damage has become irreversible. We also lack markers to reliably predict progression to ischaemic HF before detrimental remodelling has occurred. Such biomarkers would add substantially to current clinical tools and may help improve assessment of cardiovascular risk early in the disease trajectory, facilitating better monitoring and use of preventative strategies.

This thesis is founded on the idea that novel non-coding RNA biomarkers may aid in cardiovascular diagnosis and prognosis. The following chapters describe the development of a bioinformatics pipeline and a series of RNA-Seq studies focussed on lncRNAs and circRNAs, aimed at identifying new candidate markers to improve detection of myocardial ischaemia and progression from ischaemic heart disease to HF.

Chapter 3
Materials and Methods

3.1 Introduction

This chapter describes the patient cohorts for the RNA-Seq and protocols for RNA extraction from plasma and tissue, Illumina short read RNA-Seq and Nanopore long read RNA- Seq. These studies formed the basis of three core bioinformatics analyses of:

1. Illumina RNA-Seq data provided by Harvard Medical School from patients undergoing valve replacement to look ischaemia in heart tissue
2. Nanopore RNA-Seq generated from Cleveland donor heart tissue to validate results from (1)
3. Illumina RNA-Seq data generated from CDCS and HVOL cohort plasma to look at ischaemic heart disease and heart failure.

As a number of methods were performed by colleagues, I would like to thank the following:

- Dr Anna Pilbrow - RNA extraction from donor human heart to validate putative novel transcripts identified from the Harvard heart tissue
- Dr Christine Moravec and Ms Wendy Sweet - providing the Formalin-Fixed Paraffin-Embedded (FFPE) left ventricle tissue glass slides
- Ms Allison Miller – help with the Nanopore library and loading onto the Nanopore flow cell
- Dr Arthur Morley-Bunker – for RNAscope expertise
- Dr Aaron Jeffs – for the Illumina total RNA library preparation
- Christchurch Heart Institute study coordinators – for plasma sample collection, processing and bio banking

3.2 Clinical Cohorts

3.2.1 Human Heart Tissue Samples

3.2.1.1 Ischaemic Heart Tissue cohort, Brigham and Women's and Hospital, Harvard Medical School

To identify mRNA, lncRNAs and novel lncRNAs involved in myocardial ischaemia, RNA-Seq data from left ventricular tissue from 85 patients sampled before and after cardiopulmonary bypass for aortic valve replacement surgery [175] was analysed using the pipeline developed here. RNA-Seq data were generously provided by Dr Danny Muelschlegeal and Prof Simon Body, Brigham and Women's and Hospital, Harvard Medical School, Massachusetts, USA [175]. The findings from the analysis of these samples are presented in Chapter 5 along with validation using a second cohort (presented in the next Section) using Nanopore long read sequencing.

Patients undergoing nonemergent aortic valve replacement surgery with cardiopulmonary bypass (n=85) were prospectively enrolled at Brigham and Women's and Hospital, Massachusetts, USA. Procedures were performed in accordance with the ethical standards of The Partners HealthCare Institutional Review Board which approved this study, and written informed consent was obtained from each patient. Punch biopsies (≈3-5 µg total RNA content) were taken from the left ventricle apex immediately after the initiation of cardiopulmonary bypass at the time of routine placement of a surgical vent (pre-ischaemia) and after a median of 74 minutes (interquartile range 61–93 minutes; post-ischaemia), during which time the heart was arrested with cold blood cardioplegia for myocardial protection (summarised in Table 3-1). Because the heart muscle is without oxygen for this time, these samples were deemed to be a good model for mild cardiac ischaemia. Tissue samples were immediately placed in RNAlater (Ambion, Life Technologies) at +4°C for 48 hours and then frozen at -80°C until RNA extraction.

Figure 3-1 Patient demographics and clinical characteristics

Demographic	Statistic

Age (years)	71 (64–81)*
Male gender	51 (60)†
Caucasian Descent	84 (99)†
Body mass index (kg/m^2)	30 (26–34)*
Diabetes	37 (44)†
Coronary artery disease	40 (47)†
Aortic cross-clamp (minutes)	74 (61–93)*
Left ventricle ejection fraction (%)	60 (55–65)*

Patients: n=85 *Median (inter-quartile range) †Number of patients (percent of patients). Table modified from Saddic et al [175]

3.2.1.2 Preparation of Harvard heart tissue samples for RNA-Seq

The following steps were performed at the Harvard Medical School, and the RNA-Seq raw reads sent to our laboratory for analysis. Total RNA was isolated with Trizol, and RNA quality was assessed using the Agilent Bioanalyzer 2100 (Agilent, Santa Clara, CA). Ribosomal RNA was removed by performing 1 to 2 washings of RNA annealed to poly-T oligo beads (Invitrogen, Life Technologies, Grand Island, NY). RNA was fragmented and reverse transcribed using random hexamers (Invitrogen). Double stranded cDNA synthesis was performed using Pol I and RNAse H. Short fragments were purified with QiaQuick PCR extraction kit (Qiagen, Hilden, Germany) and resolved with elution buffer for end reparation and poly(A) addition followed by ligation with sequencing adaptors for cluster generation and sequenced on the Illumina HiSeq 2000 (Illumina, San Diego, CA). Read length was 100 base pairs.

3.2.2 Nanopore Validation using RNA from Cleveland Clinic Kaufman Centre Tissue bank Donor Heart Tissue

To validate putative novel transcripts identified from the Harvard heart tissue samples I used RNA that had been extracted from donor human heart by my Primary Supervisor, Dr Anna Pilbrow. The results are presented in Chapter 5.

Heart tissue from the left ventricular free wall of organ donors (n=108) was collected by the Human Tissue Core Facility at the Kaufman Centre for Heart Failure, Cleveland Clinic, between August 1993 - May 2005. The Human Tissue Core Facility holds explanted human hearts from heart transplant recipients and healthy heart tissue from unmatched organ donors (Cleveland Clinic IRB ethics approval: IRB 2378). Left ventricle tissue (< 0.5 g) was provided to the Christchurch Heart Institute (CHI) for gene expression and genotyping studies in collaboration with Professor Christine Moravec, Professor Wilson Tang and Ms Wendy Sweet (New Zealand Health and Disability Ethics Committee approval CTR/03/11/199/AM03).

From these 108 samples, I performed approximate matching on age and gender against the Harvard Heart samples by choosing Cleveland samples within the age range of the Harvard patients (37-90 years) and randomly selecting the same percentage of males/females. From these, I tested the RNA integrity of 44 samples selected at random on an Agilent TapeStation and chose those samples with an RNA Integrity Number (RIN) of ≥8, which yielded 8 samples. The RNA from these 8 samples was then pooled for use in Nanopore library preparation (Section 2.4).

Table 3-1 Samples chosen for Nanopore sequencing.

sample ID	Age	Gender	Ethnicity	Cause of death donor	RIN
41	55	M	Caucasian	CVA	8.5
50	58	M	Caucasian	CVA	8
118	53	F	Caucasian	CVA	8.3
126	47	F	Caucasian	CVA	8.4
131	41	F	Caucasian	GSW HEAD	8
156	71	M	Caucasian	CVA	8.1

165	56	F	Caucasian	CVA	8.1
167	57	M	Caucasian	CVA	8

3.2.3 Human Plasma Samples

3.2.3.1 Healthy Volunteer and Coronary Heart Disease Cohorts

To identify mRNAs, lncRNAs, novel lncRNAs and circRNAs associated with ischaemic heart disease and progression to HF, RNA was extracted from plasma from two cohorts– a healthy volunteer cohort (HVOLs) and the Coronary Disease Cohort Study (CDCS). The CDCS cohort was split into two - those patients who did not develop HF (CDCS HF-) and those that did develop HF (CDCS HF+). The HVOLs and CDCS cohorts are described in the next Sections and a summary of cohort clinical characteristics is presented in Table 3-3. The results of this analysis are presented in Chapters 6 and 7.

Table 3-2 The HVOL and CDCS cohorts clinical characteristics

	HVOL (n=31)	CDCS HF – (n=31)	CDCS HF + (n=30)
Age, years*	70(61-76)	69(64-76)	72(61-77)
Male Gender†	22(71)	23(74)	19(63)
European†	25(81)	26(84)	21(70)
Diabetes†	2(6)	12(39)	12(40)
Smoker†	0	0	1(3)
Diagnosis†		UA:7(22) STEMI:8(26) NSTEMI: 16(52)	UA:7(23) STEMI:7(23) NSTEMI: 16(54)
Hypertensive†	9(29)	23(74)	23(77)
BMI, kg/m²*	29(25-33)	29(26-33)	28(25-33)
eGFR at baseline*	68(63-74)	64(52-75)	64(52-75)
LVEF %*	67(64-70)	52(47-66)	52(39-60)
Systolic BP at baseline*	140(130-147)	135(113-147)	130(117-140)
Creatinine at baseline (µg/L)*	92(79-98)	96(85-107)	100(85-123)

Atrial Fibrillation†		6(19)	6(19)
Cholesterol mmol/L*	5.6(4.9-6.0)	4.6(4.4-4.7)	4.1(3.7-4.7)
NT proBNP at baseline*	14(11-40)	7.6(6.4-9.8)	8.7(7.1-10.0)
hsTnI at baseline*	2.7(1.6-4.9)	2.8(2.4-3.7)	3.2(2.5-4.8)
BB1 at baseline	3(10)	30(97)	26(87)
ACE_or_ARB inhibitors†	6(19)	30(97)	28(93)
Statins at baseline†	9(29)	28(90)	28(93)
Diuretics at baseline†	3(10)	8(26)	9(30)

**Median (inter-quartile range) †Number of patients (percent of patients). BMI, Body Mass Index, eGFR, estimated Glomerular Filtration Rate, LVEF, Left Ventricular Ejection Volume, BP, Blood Pressure, NT-BNP, N-Terminal Brain Natriuretic Peptide, hsTni, High sensitivity Troponin I, BB1, Beta Blocker 1, ACE, angiotensin converting enzyme, ARB, angiotensin-receptor blockers*

3.2.3.1.1 Christchurch Healthy Volunteers for Heart Disease Research Cohort (HVOLs)

Volunteers randomly selected from the Canterbury electoral rolls and age- and sex-matched to existing Christchurch Heart Institute (CHI) acute coronary syndromes, MI and HF patient cohorts were recruited into the Canterbury Healthy Volunteers study between 2003-2013 (HVOLs, n=3,358) Participants were aged 18 to 100 years and were screened before recruitment using hospital Patient Management Systems databases to confirm they had no documented personal history of overt cardiovascular disease, including IHD and MI. Participants attended a research clinic where they completed a study questionnaire on their medical history, smoking status, alcohol consumption, and self-reported physical activity. Height, weight, waist, and hip measurements were documented, blood pressure was recorded (seated, with duplicate readings at least 10 mins apart), and a blood sample was taken for neurohormone and genetic analyses. Subsequent cardiovascular events during follow-up were identified through the hospital Patient Management Systems and New Zealand Health Information Services (NZHIS) databases, with a median follow-up of 9 years. The study was approved by the Upper South A Ethics Committee (Reference No. CTY/01/05/062), and each participant provided written, informed consent.

3.2.3.1.2 Coronary Heart Disease Cohort Study (CDCS)

From July 2002, patients (n=2140) admitted to either Christchurch Hospital or Auckland City Hospital, New Zealand, were recruited into the Coronary Disease Cohort Study (CDCS). Inclusion criteria were ischaemic discomfort plus one or more of the following: ECG changes (ST-segment depression or elevation of at least 0.5 mm, T-wave inversion of at least 3 mm in at least 3 leads, or left bundle branch block), elevated levels of cardiac markers, a history of coronary disease, or 64 years of age in patients with diabetes mellitus or vascular disease. Patients were excluded from the study if they had a severe comorbidity that limited their life expectancy to 3 years. Within the CDCS cohort, unstable angina accounted for 26.1% of all diagnoses at discharge, non-ST-segment elevation MI (NSTEMI) for 51.2%, and ST-segment elevation MI (STEMI) for 22.7%. Anthropometric and clinical characteristics were recorded at planned follow-up clinic visits at baseline, 4 months, and 12 months after admission. Clinical events were recorded from questionnaires, patient notes, and NZHIS and hospital PMS databases. Median follow-up was 3.7 years (range, 0.1–7.9 years). The study conformed to the principles outlined in the Declaration of Helsinki and Title 45, US Code of Federal Regulations, Part 46, was approved by the New Zealand Multi-region Ethics Committee (Reference No. CTY/ 02/02/018) and registered with the Australian New Zealand Clinical Trials Registry (ACTRN12605000431628). Each participating patient provided written, informed consent.

3.2.3.2 Plasma collection

All plasma samples required for this study were collected, processed and bio banked at -80 C by CHI study coordinators. Bloods were taken from an indwelling intravenous cannula placed 30 min prior to sampling, with the patient remaining semi recumbent. Whole peripheral blood samples were drawn into 9mL EDTA tubes. Centrifugation occurred within 30 minutes of collection at 4C, 300rpm for 10 minutes. The separated plasma was carefully pipetted off with care taken not to disturb the red blood cells pelleted at the bottom of the collection tube. Tubes containing the plasma were placed immediately into a -80°C freezer.

3.2.4 Plasma sample selection

Plasma samples, collected ~4 months post-index coronary event once the patients were stable and early left ventricular remodelling was underway, were selected from CDCS patients and healthy control samples from the HVOLs (see Figure 3-1 for a schematic of the procedure). The CDCS group was split into two groups – those that went on to develop heart failure within 3 years (HF+) and those that did not (HF-). Patients who developed HF after admission were filtered to exclude patients that had HF prior to baseline, had a diagnosis of HF beyond 3 years of sampling and lastly, those samples that had less than 5mL of plasma in storage. This left 30 samples in the HF+ group. The CDCS patient samples that had no HF prior to, or after baseline (HF-), and the HVOL control samples were to exclude samples that had less than 5mL of plasma in storage.

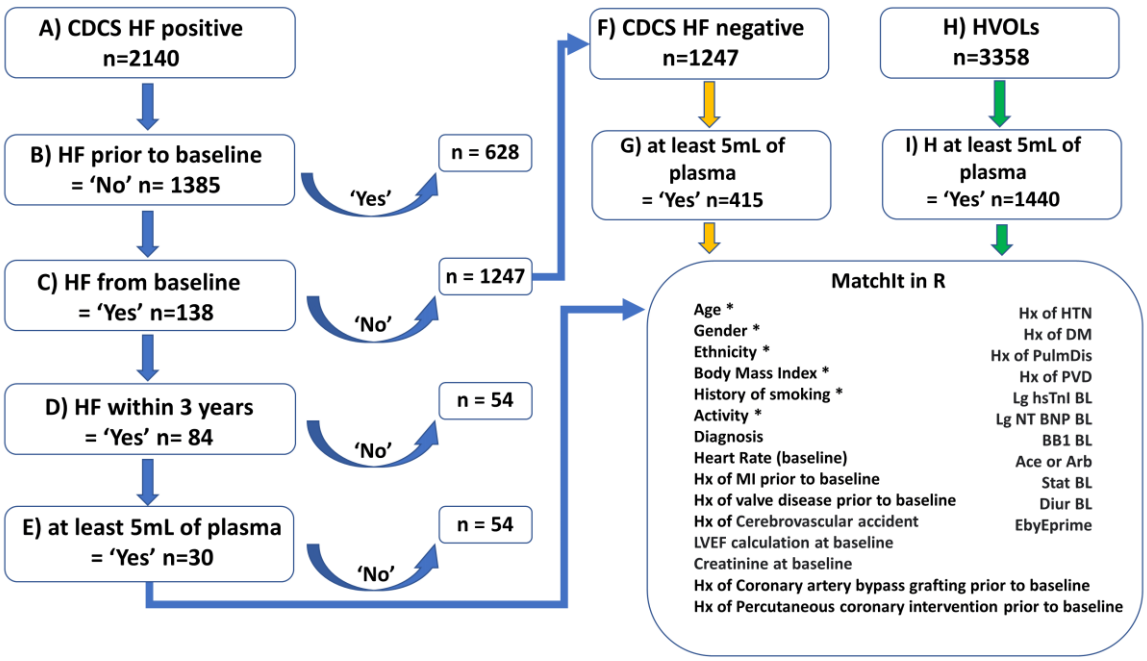


Figure 3-1 The selection procedure for plasma samples from the HVOLs and CDCS cohort. *After filtering the CDCS HF positive group had 30 samples available. The CDCS HF negative and HVOLs samples were then chosen to match the 30 CDCS HF positive samples as closely as possible using the R package MatchIt. All categories were matched between CDCS HF positive and CDCS HF negative, * indicates the categories that were matched with the HVOL groups also. See text for abbreviations explanation.*

The CDCS HF negative and HVOLs samples were then exactly matched to the CDCS HF positive samples for age, gender and ethnicity and nearest matched for previous medical history and severity of disease where possible, using the MatchIt package in R software (https://github.com/kosukeimai/MatchIt). The following variables (collected at the baseline clinic) were selected to be nearest matched: Body mass index, history of smoking, activity, heart rate, history of MI, history of valve disease, history of cerebrovascular accident, LVEF, E/e' (an echocardiogram index of diastolic dysfunction), creatinine, history of coronary artery bypass grafting, history of percutaneous coronary intervention, history of hypertension, history of diabetes, history of pulmonary disorder, history of peripheral vascular disease, levels of NT-proBNP, levels of High-Sensitivity Troponin I, and medications - Beta Blocker I, angiotensin converting enzyme inhibitors (ACE) and angiotensin-receptor blockers (ARBs), statins and diuretics. A total of 31 CDCS HF - samples and 31 HVOL samples were selected to achieve a total sample pool of 95 samples (including CDCS HF + and positive controls) for the Nova-Seq sequencing run, enabling each sample to be sequenced to a depth of ~100M reads. Confirmation that there were no statistically significant differences between the groups for any of the variables that had been selected using chi square and anova tests was carried out (data shown in Appendix A).

3.3 Laboratory Methods

As several of the laboratory methods were commercial protocols, for clarity of the chapter, where possible these have been abbreviated. The next section has been abbreviated and the full-length versions are shown in Appendix B-1.

3.3.1 RNA extraction from tissue

For the validation experiment using Nanopore sequencing (Section 3.4), RNA from the Cleveland donor heart samples was used (previously extracted by Dr Anna Pilbrow).
Left ventricular tissue from donor hearts was broken into small blocks (100-150 mg) in liquid nitrogen using a pestle and mortar and placed into pre-chilled tubes on dry ice. Automated

grinding was performed for 10 minutes in 800 µL pre-chilled TRIzol® (Invitrogen, Carlsbad, CA). Samples were mixed with 160 µL chloroform for 15 seconds, incubated at room temperature for 2-3 minutes and then centrifuged at 12,000g for 15 minutes. The supernatant was transferred to a fresh 1.5mL Eppendorf tube.

RNA clean-up was performed with the Norgen Biotek CleanAll Kit according to manufacturer's instructions. Briefly, RNase-free 70% ethanol was added to the supernatant and vortexed. This was applied to a spin column with collection tube and centrifuged. Flow-through was discarded and the column reassembled. This was repeated. Wash solution was added and centrifuged. This was repeated twice. The column was transferred to a fresh elution tube, elution buffer was added and centrifuged. RNA was quantified using the Nanodrop 8000 (ThermoFisher Scientific Inc, Waltham, USA).

3.3.2 RNA extraction from plasma

Following selection of plasma samples from CDCS and HVOLs cohorts (HVOL n=31, CDCS HF – n=31, CDCS HF + n=30), I performed RNA extraction and clean-up from plasma using Norgen Plasma/Serum RNA Purification kits (cat #56200, Norgen Biotek Corporation, Thorold, Canada). This kit purifies RNA from up to 5 mL of fresh or frozen serum/plasma and concentrates high purity, cell-free circulating and exosomal RNA using a two-column method. The first column processes the large volume of serum/plasma fluid which is followed by a concentration of the RNA on a second mini column. Frozen plasma was thawed on ice and centrifuged at 400 g for 2 minutes to remove cell debris. Up to 5 mL of plasma was transferred to a 50 mL tube (if the sample volume was less than 5 mL it was topped up to 5mL using nuclease free water). Lysis Buffer A (15 mL) along with 150 µl of β-mercaptoethanol was added to each plasma sample and vortexed for 10 seconds. Isopropanol (10 mL) was added and vortexed for 10 seconds. Then 15mL of this solution was transferred to a Maxi Spin column with collection tube (lids were left loose) and centrifuged for 3 minutes at 1000 g, to bind the RNA. The flow-through was discarded, and column and tube

reassembled. This was repeated until all of the mixture had been added to the column. Wash Solution A (5 mL) was added and centrifuged for 3 minutes at 1,000 g. The flow-through was discarded, and column and tube reassembled, and this wash was repeated a second time. The column was spun empty at 2000 g for 3 minutes to dry the column. The Maxi Spin column was transferred to a fresh 50 mL tube, 800 µl of Elution Buffer F was added to the column, incubated at room temperature for 2 minutes and then centrifuged for 2 minutes at 500 g. To maximise RNA recovery the eluted RNA was reloaded onto the same Maxi Spin column, incubated at room temperature for 2 minutes and then centrifuged for 2 minutes at 500 g. To concentrate the RNA, 600 µl of Lysis Buffer A was added to the eluate and vortexed for 10 seconds. This was followed by adding 800 µl of 96-100% Ethanol and vortexing for 10 seconds, and then 750 µl of the ethanol-RNA eluate mix was transferred to a Mini Spin column and centrifuged for 2 minutes at 3,300 g. The flow-through was discarded and this step was repeated until all of the mixture had been transferred. For every sample, a mix of 15 µL of DNase I and 100 µL of Enzyme Incubation Buffer was prepared using Norgen's RNase-Free DNase I Kit (Product # 25710). Prior to DNase I treatment, 400 µl of Wash Solution A was added to the column, centrifuged for 3 minutes at 3,300 g and the flow-through discarded. After adding 100 µl of the prepared RNase-free DNase I solution to the column and centrifuging at 8,000 g for 1 minute, the flow-through was pipetted back to the column and incubated at 25-30°C for 15 minutes, to ensure maximal degradation of DNA. To wash the RNA bound to the column, 400 µl of Wash Solution A was added to the column, centrifuged for 3 minutes at 3,300 g and the flow-through discarded. This step was repeated a second time to ensure removal of digested DNA. The column was spun empty for 2 minutes at 13,000 g to dry it completely before transferring it to a fresh 1.7 mL Elution tube and adding 50 µL of Elution Solution A to the column, incubating at room temperature for 2 minutes and centrifuging for 1 minute at 400 g followed by 2 minutes at 5,800 g. To maximise RNA recovery the eluted buffer was pipetted back to the column, incubated at room

temperature for a further 2 minutes and centrifuged for 1 minute at 400g followed by 2 minutes at 5,800 g. RNA was quantified using the Qubit™ adapted protocol (Section 3.3.3.1.3). This final 50 µl elution containing the RNA was stored at -80 C prior to sequencing.

3.3.3 Assessing RNA Quantity and Integrity

3.3.3.1 RNA Quantification

3.3.3.1.1 Nanodrop Spectrophotometry

The amount of RNA extracted from each Cleveland human heart tissue sample was estimated by ultraviolet spectrophotometry using NanoDrop. RNA (and DNA) absorb light with a characteristic peak at 260 nm. Elution buffer was used as a blank and 1.5 µl of sample was loaded onto the cleaned Nanodrop pedestal to obtain an approximate estimate of the sample concentration prior to quantitation on the Qubit Fluorometer and assessment of RNA integrity on the TapeStation 2200 system.

3.3.3.1.2 Fluorometry (Qubit™) Standard protocol for RNA quantification of heart tissue

RNA extracted from human heart tissue was quantified on the fluorescence-based Qubit 2.0 Fluorometer using the Qubit™ RNA HS Assay Kit (ThermoFisher Scientific Inc., Waltham, USA). The Qubit™ working solution was prepared at room temperature by mixing Qubit RNA HS Reagent to Qubit RNA HS Buffer in a ratio of 1:200 (~200 µL of working solution per sample/standard). Qubit™ working solution (190 µL) was added to 10 µL of the Qubit™ standards (Standard #1 served as a blank, Standard #2 contained RNA at 500 ng/µL) and vortexed for 3 seconds. 199 µL of Qubit™ working solution was added to 1 µL of each sample and vortexed for 3 seconds. All tubes were incubated at room temperature for 2 minutes. The two kit standards were used to create a standard curve that the Qubit™ uses to generate a curve fitting algorithm that is uses to calculate concentration. Each sample measurement was performed three times, averaged and multiplied by 200 (the dilution factor) to calculate the final concentration of RNA in ng/µL.

3.3.3.1.3 Fluorometry (Qubit™) Adapted protocol of RNA quantification for plasma

Because the concentration of RNA extracted from human plasma was below the detection limit of the Qubit™ high sensitivity RNA kit described above, a modified protocol that could detect RNA concentrations between 250 pg/µL and 55.6 pg/µL [369] was used.

The Qubit™ working solution and standards were prepared as described above. A 2.5 ng/µL RNA 'spike in' master mix was prepared by diluting the Qubit™ RNA Standard #2 4 -fold with RNase free water. The 'spike in' master mix (182 µL) was mixed with 18µL of RNAse free water ('spike in' alone tube) or 17 µL of RNAse free water and 1µL of sample. All tubes were vortexed for 3 seconds, centrifuged for ~ 5 seconds (to collect the solution at the bottom of the tube) and incubated at room temperature for 2 minutes. After calibrating the Qubit™ fluorometer with both standards (described above) the RNA 'spike in' alone tube was measured (Read 1) followed by the RNA sample tubes (Read 2). RNA sample concentration was calculated as:

Concentration (pg/uL = (Read2 – Read1) (pg/µL) × 200 (µL) ÷ volume of sample added (µL)

3.3.3.2 Agilent TapeStation RNA ScreenTape System for assessing RNA integrity

RNA from the Cleveland Heart tissue was quantified and assessed for integrity using the Agilent TapeStation 2200 System (Agilent, Santa Clara, CA). This is an automated electrophoresis system which uses 1-2 µl of RNA sample to determine the RNA Integrity Number (RIN) based on the relative intensity of 28S:18S rRNA bands. Ratios >2.0 indicate intact, high-quality RNA and equate to RIN values which can be between 10 (intact) to 1 (totally degraded). [370].

RNA was thawed on ice and all reagents were allowed to equilibrate to room temperature for 30 minutes. 1 µl of RNA ladder or sample was added to 5µl RNA sample buffer and vortexed briefly to mix. Tubes were heated to 72 °C for 3 minutes, placed on ice for 2 minutes and spun briefly to collect the liquid at the bottom of the tube. RNA ScreenTape, tips, and tubes

were loaded onto the instrument and automated electrophoresis performed according to manufacturer's instructions.

3.3.4 Sequins – synthetic internal controls.

The Garvan Institute (Sydney, Australia) have developed RNA standards for use in RNA-Seq experiments called 'sequins' (synthetic spike-ins to act as internal controls) [371]. Sequins are provided with a range of isoforms and at varying concentrations, covering 78 artificial genes ranging from single exons to large, multi-exonic transcripts (up to 36 exons) and including 164 alternative isoforms ranging in length. The sequins are derived from an artificial 11 Mb chromosome designed to reflect the GC composition and repeat density of the human genome. Any sequence similarity of the artificial chromosome to the human genome was avoided by sequence inversion and shuffling. By spiking in these synthetic RNAs of known amounts it is possible to check that the library preparation and sequencing has worked successfully. It is also possible to detect the limit of detection (e.g. what is the lowest abundant sequin transcript we can detect) and extrapolate that to the transcripts we are analysing. Sequins were spiked (1%) into Cleveland Heart tissue RNA when I validated novel heart tissue transcripts from the Harvard data. Also, sequins were tested to see if they could be successfully added to the plasma samples.

3.4 Methods for Nanopore long read sequencing

3.4.1 Reverse transcription and strand-switching

The integrity of RNA previously extracted from Cleveland Heart Donor Patients (section 3.2.2) was assessed on the TapeStation 2200 as described in Section 3.3.3.2. Of 44 RNA samples tested, 8 samples had RIN scores ≥8 indicating good quality RNA and were selected for analysis by Nanopore long read sequencing. These were pooled for the validation experiment, which aimed to confirm the presence of putative novel lncRNAs identified in the Havard samples in independent samples, rather than determine their abundance.

The Nanopore cDNA-PCR Sequencing protocol (SQK-PCS109) uses a 'strand switching' method which gives higher yields of cDNA and enriches for full length cDNAs. When the first strand synthesis reaches the end of the fragment, the reverse transcriptase adds several non-templated Cs to the end of the cDNA. A strand switching primer in the reaction binds to these Cs and the reverse transcriptase switches template from the RNA to the strand switching primer. The second strand cDNA is then synthesised (Figure 3-2).

Briefly, the strand switching method produces full length cDNAs from total RNA (or polyA RNA). These were then amplified by PCR which at the same time adds 'rapid attachment' primers, sequencing primers were then attached, and the library was loaded onto the flow cell.

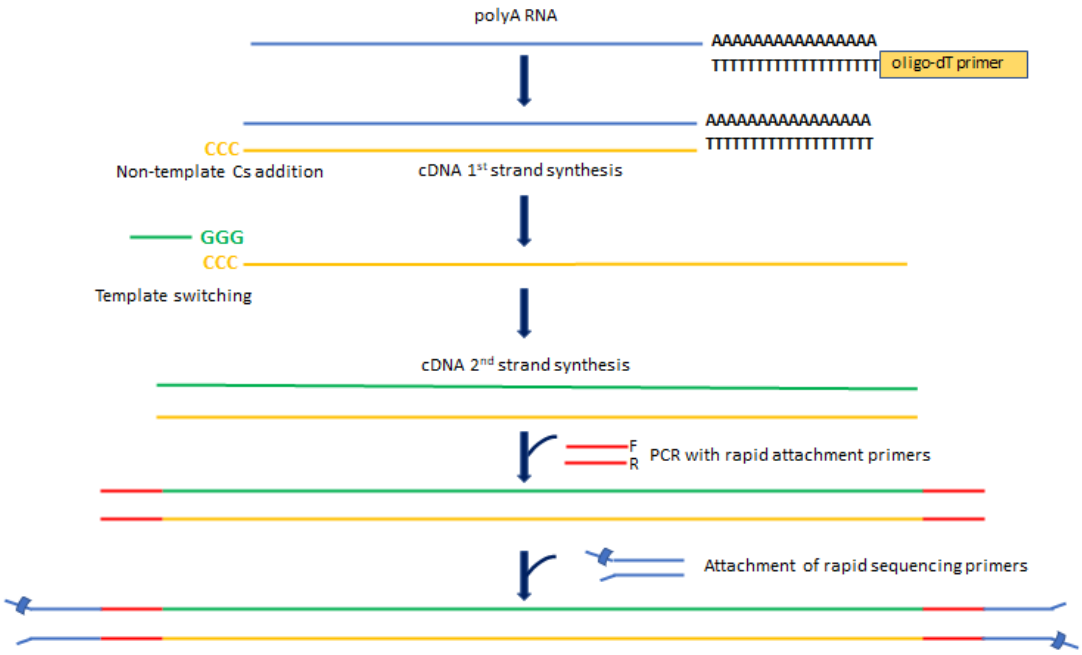


Figure 3-2 A schematic of Nanopore 'strand switching' cDNA library preparation protocol. *An oligo dT primer is annealed to the polyA tail of the RNA and reverse transcription takes place. Several non-templated Cs are then added and a strand switching primer anneals. A second cDNA strand is then generated producing a full-length cDNA. Amplification happens by PCR and sequencing primers are attached. Figure adapted from Nanopore website (https://nanoporetech.com/sites/default/files/s3/literature/Nanopore-cDNA-Guide.pdf)*

Total RNA (50 ng) was combined with 1% RNA sequins (0.5ng) [372], 1 µL of polyT-VN RT primer (VNP, ONT SQK-PCS109) and 1 µL 10 mM dNTP solution (N0447, NEB) and

the final volume was adjusted to 9 µL with RNase-free water. Samples were mixed by tapping, briefly spun by microfuge, incubated at 65°C for 5 minutes and immediately snap cooled on a pre-chilled freezer block. Next, 4 µL of 5x RT Buffer (ThermoFisher, EP0751), 1 µL of RNaseOUT (40 U/µL Life Technologies, 10777019), 1 µL of nuclease free water and 2 µL Strand-Switching Primer (SSP, ONT SQK-PCS109) were added to the samples, which were mixed by tapping, briefly spun, and incubated at 42 °C for 2 minutes. Finally, 1µL of Maxima H Minus Reverse Transcriptase (ThermoFisher, EP0751) was added to each of the 19 µL reaction mixes. Samples were mixed by tapping, briefly spun by microfuge and incubated in a thermocycler (Eppendorf® Mastercycler®, Hamburg, Germany) under the following conditions: 42 °C for 90 minutes (RT reaction and strand-switching), 85°C for 5 minutes (heat inactivation); and then held indefinitely at 4 °C.

3.4.2 Selecting for full-length transcripts by PCR

The reverse-transcribed sample (20µL) was split into four 5µl aliquots and mixed with 25 µL LongAmp Taq 2x Master Mix (NEB M0287), 1.5µL cDNA Primer (cPRM, ONT SQK-PCS109) and 18.5µL of Nuclease-free water. Reactions were incubated in a thermocycler under the following conditions: 95 °C for 30 seconds (initial denaturation) followed by 14 cycles of 95 °C for 15 seconds (denaturation), 62 °C for 15 seconds (annealing), 65 °C for 50 seconds per kilobase (extension), and a final extension of 65 °C for 6 minutes. After thermal cycling, samples were held at 4 °C. Exonuclease (NEB, M0293) (1µL) was added to each tube and each reaction was incubated at 37 °C for 15 minutes followed by 80 °C for 15 minutes. The four sample aliquots were pooled, and the combined cDNA sample was purified by a 5-minute rotating mixer incubation with 160µL (1.8X) resuspended Agencourt AMPure XP magnetic beads (A63880, Beckman Coulter) at room temperature. While bound to the magnet, two 200 µL washes with 70% ethanol were performed, with the ethanol then discarded. Tubes were air dried for 30 seconds and eluted in 12 µL Elution Buffer (EB, ONT SQK-PCS109). The purified cDNA samples were analysed with Genomic DNA ScreenTape

(next section) On average, the samples had a fragment length of 1500 bp and a concentration of 40 ng/µL. Elution Buffer (7 µL) was added to 5µL of sample to make up between 100-200 fmol of amplified cDNA.

3.4.3 Agilent Genomic DNA ScreenTape System

As part of the Nanopore library preparation, sizing analysis of the generated cDNA fragment was assessed with the Agilent Genomic DNA ScreenTape was used for this (used for fragments between 200bp to > 60,000bp).

All reagents were allowed to equilibrate to room temperature for 30 minutes. Genomic DNA ladder (1 µl) or sample (1 µl) was added to 10 µl Genomic DNA sample buffer, vortexed at 2000 rpm for 1 minute then spun to collect the liquid at the bottom of the tube. ScreenTape, tips and tubes were loaded onto the instrument and automated electrophoresis performed according to manufacturer's instructions. The instrument outputs a gel image (alongside the ladder) and an electropherogram from which the length of the fragment was assessed.

3.4.4 Adapter addition

To ligate sequencing adaptors to each transcript, 1 µL of Rapid Adapter (RAP, ONT SQK-PCS109) was added to each cDNA sample, the tube was mixed by tapping and briefly spun by microfuge, then incubated for 5 minutes at room temperature. The sample was loaded onto the gridION according to the ONT SQK-PCS109 kit instructions.

3.4.5 Bioinformatic analysis for Nanopore sequencing

Oxford Nanopore Technologies (ONT) reads were base called using the on-sequencer Guppy base calling software (v3.03 High accuracy) without the trimming option as this removes primers needed for the Pychopper software to work. Full length transcripts were identified using Pychopper (https://github.com/nanoporetech/pychopper) which is a software tool developed by Oxford Nanopore Technologies (ONT) to identify full length Nanopore cDNA reads. These reads were then mapped to a combined reference comprising the human genome

(hg38) and the sequins *in silico* chromosome (chrIS) using Minimap2 (v2.17-r941) [373] which is a general purpose alignment program to map long reads (parameters: –ax splice –secondary = no.)[373]. Reads were then processed using the Pinfish suite (https://github.com/nanoporetech/pinfish), a pipeline generated by ONT to generate annotations from long read cDNA (parameters: -p 1.0 -c 3 -d 10 -e 30). These full-length transcripts were then run through part of the bioinformatics pipeline to generate a list of putative novel transcripts identified by Nanopore sequencing. These transcripts were then compared to the putative novel transcripts identified from the Harvard data (which used Illumina short read sequencing).

3.5 Methods for Illumina short read sequencing of plasma RNA

RNA that had been extracted from plasma was sent to the Otago Genomics facility (Dunedin, New Zealand), so that sequencing libraries could be made and then sent to the Ramaciotti Centre for Genomics, Sydney, Australia for sequencing on the NovaSeq6000. The following sections (2.5.1-2.5.6) were carried out by Dr Aaron Jeffs at the Otago Genomics facility.
For Illumina total RNA library preparation, the SMARTer® Stranded Total RNA-Seq Kit v2 - Pico Input Mammalian (Takara Bio, USA) was used. A control total RNA sample was used (human brain, diluted to 0.25ng/µl), which was supplied with the kit. Briefly, the protocol involves first strand cDNA synthesis, addition of Illumina adapters and barcodes (for multiplexing), library purification and AMPure beads, a final PCR amplification and a second purification with AMPure beads. The following section has been abbreviated and the full protocol is in Appendix B-2

3.5.1 First strand synthesis

Option 2 (starting with degraded RNA) without fragmentation was used. RNA sample (8 µl) was mixed with 1 µl of SMART Pico Oligos Mix v2 on ice and incubated at 72 °C for 3 minutes, then chilled rack for 2 minutes. The First-Strand Master Mix was prepared with 4 µl 5x First-Strand Buffer, 4.5 µl SMART TSO Mix v2, 0.5 µl RNase Inhibitor and 2 µl

SMARTScribe Reverse Transcriptase, added to each reaction tube and vortexed and incubated at 42 °C for 90 minutes, 70 °C for 10 minutes and then held at 4 °C

3.5.2 Addition of Illumina adapters and indexes

A PCR master mix with 2 µL nuclease-free water, 25 µl SeqAmp CB PCR Buffer (2X), 1 µl SeqAmp DNA Polymerase was added to each sample. PCR Primer HT was added (1 µl), and PCR was performed: 94 °C for 1 minute, 5 cycles of 98 °C for 15 seconds, 55 °C for 15 seconds, 68 °C for 30 seconds, 68 °C for 2 minutes.

3.5.3 Purification of the RNA-Seq Library Using AMPure Beads

AMPure beads (40 µL/sample), were added to each sample tube and incubated for 8 minutes. Tubes were placed onto a magnetic separation device for 5 minutes, the supernatant was removed and discarded and 200 µl of freshly made 80% ethanol was added then removed and discarded, and the wash step was repeated. Pellets were air dried, 52 µl of nuclease-free water added, incubated for 5 minutes at room temperature, and placed back onto the magnetic separation device for 1 minute. Supernatant (50 µl) was transferred to a new tube, mixed with 40µL of AMPure beads and incubated for 8 minutes.

3.5.4 Depletion of Ribosomal cDNA with ZapR v2 and R-Probes v2

Tubes were placed on the magnetic separation device for 5 minutes, supernatant removed and discarded, and 200 µl of freshly made 80% ethanol added to each sample. Supernatant was removed and discarded, and the wash step was repeated. Tubes were aired dried on the magnetic separation device. A ZapR master mix was prepared: 16.8 µl Nuclease-Free Water, 2.2 µl 10X ZapR Buffer, 1.5 µl ZapR v2 and lastly 1.5 µl ‘activated’ R-Probes v2 (that bind rRNA and mitochondrial rRNA which had been preheated hot-lid thermal cycler at 72°C for 2 minutes and held at 4°C for at least 2 minutes). The dried beads were resuspended in 22 µl of the ZapR master mix, incubated at room temperature for 5 minutes and placed on the magnetic separation device for 1 minute. Supernatant was removed and transferred to a new

PCR tube. Tubes were incubated in a preheated hot-lid thermal cycler at 37 °C for 60 minutes followed by 72 °C for 10 minutes then held at 4°C.

3.5.5 Final RNA-Seq Library Amplification

Library fragments were further enriched in a second round of PCR. A PCR master mix was prepared: 26 µl Nuclease-Free Water, 50 µl SeqAmp CB PCR Buffer, 2 µl PCR2 Primers v2 and 2 µl SeqAmp DNA Polymerase. 80 µl of master mix was added to each sample tube. Tubes were incubated for 94 °C for 1 minute, 16 cycles of 98 °C for 15 seconds, 55 °C for 15 seconds, 68 °C for 30 seconds, and held at 4 °C.

3.5.6 Purification of Final RNA-Seq Library Using AMPure Beads

A second purification step with AMPure beads. Essentially the process was the same as Section 3.5.3 but instead of 40 µl/sample of AMPure beads, 100 µl AMPure beads were added to each sample. Also, the final elution was into 12 µl of Tris buffer. The samples were then sent to the Ramaciotti Centre for Genomics, Sydney, Australia for sequencing on the NovaSeq6000.

3.6 WGCNA: Weighted Correlation Network Analysis

For the Harvard data RNA-Seq analysis, to identify clusters of highly correlated genes that share a similar pattern of expression across patients and may be co-regulated, a weighted gene co-expression network analysis was performed using the WGCNA package [374] in R. WGCNA is a network approach to construct gene networks (modules) and identify modules that are correlated with a particular trait. Genes within a module that are highly interconnected are recognised as ‘hub’ genes – those genes which are thought to play a central role in the particular module. An unsigned weighted correlation network (negative as well as positive correlations were considered) was constructed for all genes across all samples (pre- and post-ischaemia), with the aim of identifying gene modules associated with ischaemia. WGCNA assigns genes to clusters based on correlation and shared network neighbours. Each module is

represented by a weighted average of the expression level of all genes within the module, referred to as the module eigengene (kME). Individual genes and modules were then tested for association with ischaemia by correlating module eigengenes with ischaemia. In this way, modules with the strongest association to ischaemia were identified, and the genes with the strongest association to each module indicated highly interconnected hub genes that may be driving ischaemia.

3.7 Ingenuity Pathway Analysis (IPA)

To identify the molecular pathways, networks and related biofunctions, differentially expressed genes within those modules identified by WGCNA as significantly associated with ischaemia were assessed with the Core Analysis Workflow with Ingenuity Pathway Analysis (IPA) software (http://www.ingenuity.com, Qiagen, Redwood City, CA, USA). Along with p-values, IPA gives a z-score which represents a statistical measure of the match between the expected vs observed direction of gene expression. The magnitude of the z-score is a measure of the proportion of genes with expression patterns consistent with activation of the pathway. Any z-score > 2 (indicating pathway activation) or < -2 (indicating pathway repression) was considered to be potentially biologically meaningful.

3.8 RNAscope

To identify the cellular localisation of mRNAs or lncRNAs associated with ischaemia, RNAscope (Advanced Cll Diagnostics, ACD, Newark, CA, USA) was used. RNAscope is a novel *in situ* hybridisation assay developed by for detection of target RNA within cells and tissue whilst preserving the structural integrity of both. Briefly, a specific double 'Z' shaped probe hybridises to the target RNA sequence (about 18-25 bases). The probe is then bound by amplifier probes, which bind the chromogenic label (fast-red + alkaline phosphatase).

3.8.1 RNAscope probes

Custom RNAscope probes for two annotated lncRNAs (PCAT 19 and VASH1-AS) and one novel lncRNA (MSTRG. 10265.1) that were associated with ischaemia by WGCNA analysis were ordered from ACD, along with NEAT1 – a lncRNA that has been demonstrated to be expressed predominantly in the nucleus of cells and which acted as a positive control and a negative control dapB probe. The RNAscope 2.5 HD fast red detection kit was used (CAT NO: 322360).

3.8.2 Tissue section preparation

Formalin-Fixed Paraffin-Embedded (FFPE) left ventricle tissue glass slides were kindly provided in collaboration with Professor Christine Moravec and Ms Wendy Sweet (Cleveland Clinic Foundation, Cleveland Ohio). Slides were baked for 1 hour at 65C°. They were then incubated for 5 mins at room temperature in xylene, followed by a second xylene incubation for 5 minutes, and then 1 minute in 100% ethanol (ETOH) twice. The slides were air dried at room temperature, before 5-8 drops of hydrogen peroxide (provided in the kit) were added to the slides and left for 10 minutes at room temperature. The slides were rinsed twice in distilled water. Slides were then placed in boiling(100-104°C) pre-treatment 2 solution (provided in the kit) for 30 minutes then rinsed in distilled water twice and once in 100% ETOH then left to air dry. A hydrophobic barrier was then drawn around the tissue section using an ImmEdge™ pen (Vector Laboratories Inc. Burlingame, CA, USA) to contain further treatments.

3.8.3 Hybridisation with RNA probes.

Hybridisation was conducted following the manufacturer's instructions in FFPE Detection kit (Red) Part 2. Protease plus solution was added to the slides (5 drops) and incubated at 40°C in the HybEZ™ Oven (ACD, Newark, CA, USA) for 30 minutes. The two lncRNA probes (PCAT 19 and VASH1-AS), a negative control (no antibody) and a positive control (NEAT1) (bacterial gene, dihydrodipicolinate reductase (DapB)) were added to separate slides (4

drops). All slides were then placed onto the HybEZ™ Humidity Control Tray, the lid placed on and placed in the HybEZ™ Oven and incubated at 40 ˚C for 2 hours. Slides were then rinsed with two changes of Wash Buffer for 2 minutes at room temperature.

Next, a series of incubation steps with pre-amplifier and several amplifier probes was carried out. AMP1-AMP4 (kit provided) were serially incubated at 40°C, AMP1 for 30 minutes, AMP2 for 15 minutes, AMP3 for 30 minutes, AMP4 for 15 minutes. AMP 5 was then added at room temperature for 60 minutes and lastly, AMP6, added at room temperature, for 15 minutes. Between each incubation step slides were washed twice in wash buffer, consisting of two-minute incubations at room temperature.

3.8.4 Signal detection

To detect the hybridisation probes, 120µl of a 1:60 ratio mix of Fast RED-B to Fast RED-A was pipetted onto each slide and incubated at room temperature in the HybEZ™ Humidity Control Tray for 10 minutes. Slides were rinsed twice in fresh distilled water.

3.8.5 Counterstaining

Slides were placed into a dish containing 50% Gills Haematoxylin II (Leica Biosystems, Richmond, IL, USA) for two minutes at room temperature after which the slides appeared purple, then washed in distilled water until the slides were clear whilst tissue sections were still purple. Slides were then rinsed in 0.02% Ammonia water, followed by tap water. Slides were dried in a 60°C oven for at least 15 minutes after which they were briefly dipped into fresh xylene. A cover mount was applied using 1-2 drops of SurgiPath DPX mounting medium. Slides were analysed and Tagged Image Format (.TIF) images captured using a Zeiss Apotome Microscope and associated software (AxioVersion 4.5. Apotome software, Carl Zeiss Microscopy, LLC, Thornwood, New York, USA).

3.9 Bioinformatic analysis of novel lncRNAs

3.9.1 Quality control to identify tissue origin of transcriptome

As a quality control step for the Harvard data RNA-Seq analysis the R package TissueEnrich was used [375]. TissueEnrich calculates the enrichment of tissue specific genes in a set of input genes and in this way identify the tissue most expected for that set of genes.

3.9.2 Conservation of novel lncRNAs

Evolutionary conservation of novel lncRNAs was assessed using precomputed nucleotide level calculations of evolutionary selection from phastCons score. PhastCons identifies evolutionary conserved elements in multiple-aligned sequences and assigns a score between 0 and 1; the closer the score is to 1, the more evolutionarily conserved the base [376]. A bigwig file containing phastCons base-by-base conservation scores across 20 mammalian species was downloaded from the UCSC site (http://hgdownload.soe.ucsc.edu/goldenPath/hg38/phastCons20way/hg38.phastCons20way.bw) and converted to bed format using the bigWigToBedGraph tool (http://hgdownload.soe.ucsc.edu/admin/exe/). Novel lncRNAs, annotated lncRNAs and coding mRNAs were aligned to the phastCons bed file and a mean phastCons score was extracted for each exon using bedtools map (bedtools map -a genes.bed -b phastCons.bedgraph -c 4 -o mean). To compare the conservation profile of exon sequences between novel lncRNAs, annotated lncRNAs and mRNAs, the frequency of scores were plotted for each group separately in R.

3.9.3 Novel lncRNAs overlapping Regulatory Features: SNPs, enhancers, promoters

Potential interactions between novel lncRNAs with regulatory elements were explored using Bedtools v2.27.1 [377]. Overlap of novel lncRNAs with disease-associated single nucleotide polymorphisms (SNPs) (https://www.ebi.ac.uk/gwas/downloads) curated for cardiovascular related traits and left ventricle specific enhancers, promoters, promoter flanks (defined as

transcription factor binding regions that flank promoters (Ensembl biomaRt https://www.ensembl.org/biomart/)) were interrogated along with nearest neighbouring genes, by comparing against annotated SNPs, regulatory elements and gene coordinates in GENCODE v29.

3.9.4 Novel lncRNAs overlapping expression quantitative trait locus (eQTLs)

Expression quantitative trait locus (eQTLs) analysis investigates associations between single nucleotide polymorphisms (SNPs) and gene expression levels. The novel lncRNAs were analysed to see if any overlapped eQTL SNPs, thereby potentially implicating the lncRNA in the mechanism underlying the eQTL association. Overlap of both novel and annotated lncRNA exon coordinates with cis-eQTL SNPs from GTEX v.8 left ventricle (https://gtexportal.org/home/datasets) was assessed using Bedtools Intersect (https://bedtools.readthedocs.io/en/latest/content/tools/intersect.html). LncRNAs found to be associated with an eQTL SNP were analysed to see whether the lncRNA and eQTL gene were members of the same WGCNA module (and therefore suggested to be co-regulated), to explore putative functional links between lncRNAs, eQTLs and ischaemia.

Chapter 4
The bioinformatic pipeline

4.1 The Bioinformatic pipeline explained

The raw data files from the sequencing instrument are essentially the starting point for the Illumina short-read sequencing bioinformatic pipeline developed in this thesis. The first part of this chapter describes the development of the pipeline, the software involved and justification for choosing the software. The second part of the chapter describes validation of the pipeline using publicly available data. For an overall schematic of the pipeline, which begins with raw RNA Sequencing reads in a fastq file see Figure 4-1.

4.1.1 The start of the pipeline QC - Trimming of adapters and low-quality reads

The primary limitation of short read sequencing is that, as the name suggests, only short fragments of RNA or DNA (up to 300 bp) can be sequenced. The length of the read is limited by the availability of reagents (one base is sequenced per cycle on the sequencing instrument) and so the number of cycles equals the number of bases that can be sequenced per fragment. For applications such as transcriptome sequencing, as in this thesis, 2 x 150bp is commonly used, representing a compromise between cost and maximising sequence length and accuracy. Mapping accuracy is increased with paired-end sequencing, as it has the advantage of producing twice the number of reads for the same fragment, which means that when the reads are aligned back to the reference genome we have additional information on how far apart the paired reads should be, helping with more accurate read alignment.

Preparation of RNA-Seq libraries involves RNA fragmentation followed by size selection and ligation of adapters. Adapters are necessary for forward and reverse priming during the sequencing reaction, barcoding and ligation to the flow cell of the sequencing machine.

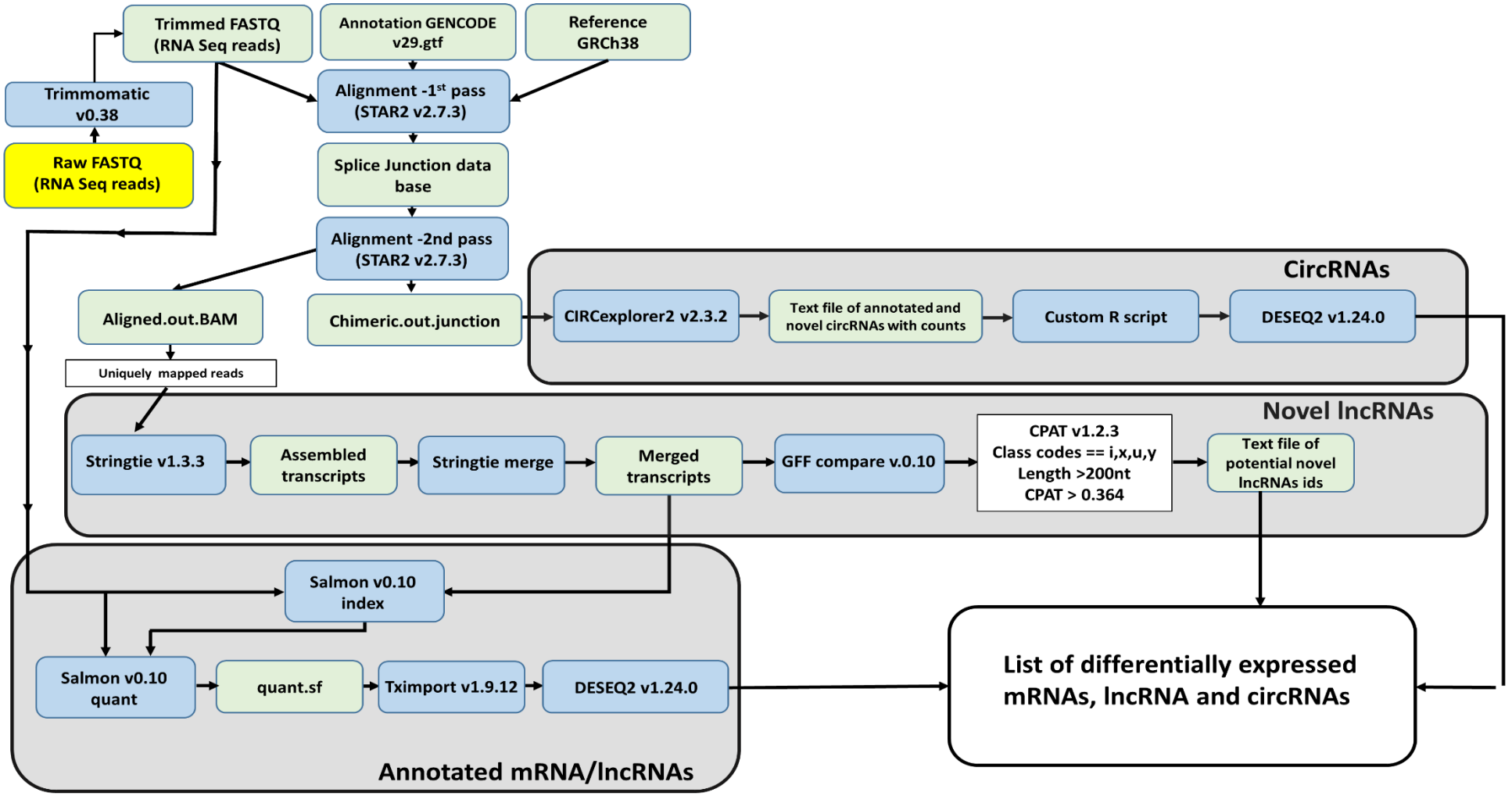


Figure 4-1 An overview of the bioinformatics pipeline developed in this thesis

Blue boxes indicate the software used, green boxes represent the input or output files. Each section of the pipeline and the justification as to the choice of software is explained in this chapter

Size selection is used to purify sequences that are the appropriate length for the sequencing chemistry, to maximise the information gained without redundancy (overlapped read pair sequences). For example, for 2 x 150 bp sequencing we would choose fragments at least 300 bases long so that 150 bases can be sequenced from each end of the fragment without overlapping bases in the middle (Figure 4-2). However, size selection (usually with solid phase reversible immobilisation beads) is an imperfect process and most RNA libraries include a range of fragment sizes. Inclusion of shorter fragments means that the number of cycles will exceed the fragment length and we will sequence into the adapter at the 3' end. This is especially true for low quality, fragmented RNA, and applied to RNA from plasma samples as in this thesis.

Read 'quality' is usually measured as a Phred quality score by the sequencing instrument. This score indicates the probability of the base being called correctly. It is calculated by the equation [378]as:

$Q = -10 \log_{10} E$, where E is the base-calling error probability.

The higher the score, the greater the probability that the base is correct. For example, a Phred score of 30 equates to a 1 in 1000 chance of the base being incorrect. Base errors can occur due to polymerase errors during library preparation or sequencing errors such as cluster density (where the imaging software cannot distinguish between two closely positioned clusters on the flowcell) and, for Illumina sequencing which uses the sequence by synthesis technique, phasing errors (whereby one fragment of the cluster can either be one cycle behind or in front when imaging takes place). Phasing tends to accumulate over time and so the 3 prime end of the read is more prone to quality loss.

Several software tools have been developed to minimise the number of 'unaligned reads' by ensuring that only high quality, non-adapter containing sequences are selected for alignment including Trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic), Trim Galore

(http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), bbduck from the BBMap suite (https://sourceforge.net/projects/bbmap/) and cutadapt (although this only trims adapter sequences https://opensource.scilifelab.se/projects/cutadapt/). For the pipeline described in this thesis, I used the Trimmomatic software [379] to trim the raw fastq files as it compares favourably to other trimming software, allows input of paired-end reads. A Q score of 20 was chosen as it appears to be the optimum score to maximise the percentage of mapped reads [380] and a default minimum length of 36 bases was chosen as this appears to be the length at which we can minimise mapping to multiple locations [381].

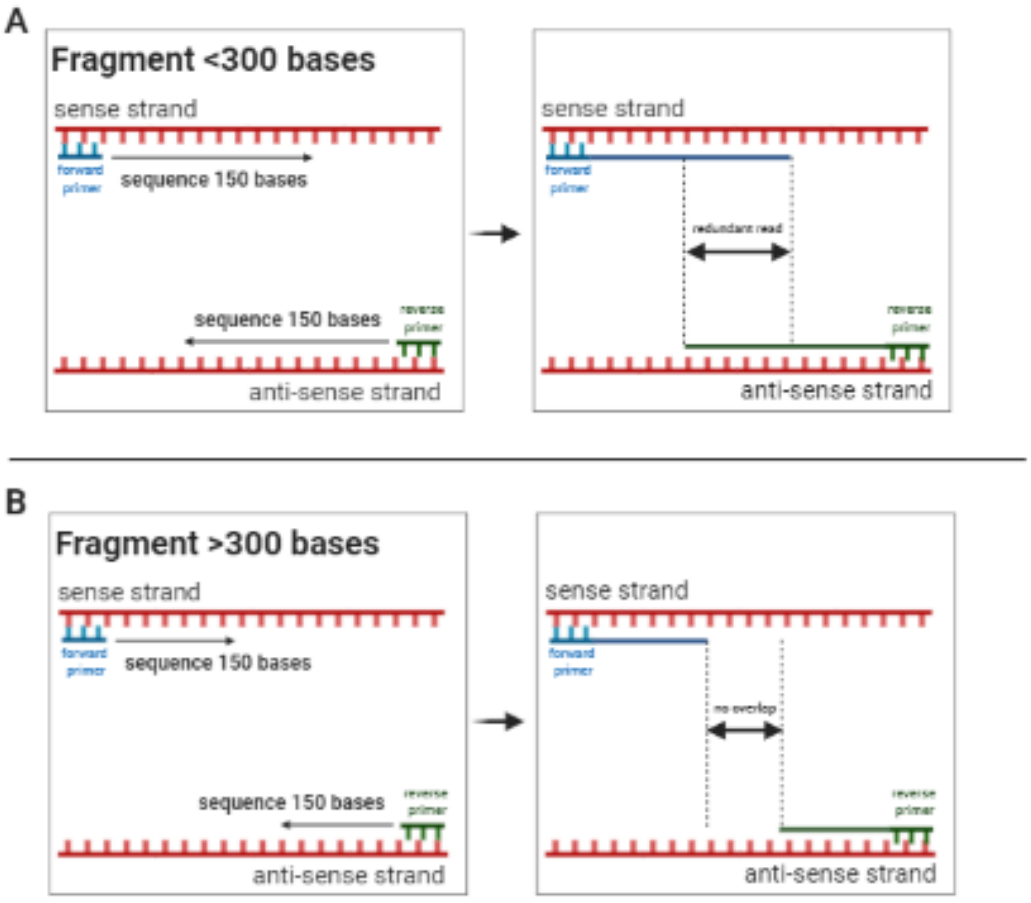


Figure 4-2 Why size selection during library preparation eliminates redundancy

The size selection step of RNA/cDNA during library preparation ensures enrichment of fragments of an appropriate length for sequencing. ***A)*** *If the fragment<300 bases the sequencing reads would*

overlap resulting in redundant (duplicated) part of the read ***B)*** *When the fragment is >300 bases then there is no overlap. Created in Biorender.com*

Trimmomatic has the additional feature of using a 'sliding window' where removal of the 3' end of a read occurs when the average quality of a group of bases drops below a user-defined quality threshold, thus preventing a single low quality base causing the removal of adjacent high quality bases [379] (see trimming figure 4-3 below).

Figure 4-3 QC plots showing reads before and after trimming

FastQC plots of a representative sample pre ***(A)*** *and post* ***(B)*** *trimming with Trimmomatic.* ***A & B Upper panel****): Boxplots indicating quality scores for bases at each position across all fragments: the x-axis shows the position in the read, the y-axis the Phred quality scores. For each boxplot, the central red line is the median value, the yellow box represents the inter-quartile range (25-75%), the upper and lower whiskers represent the 10% and 90% points, and the blue line represents the mean quality. The background of the graph divides the y axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red). Bases that fall below a Phred quality score of 30 are trimmed.* ***A & B Middle panel****): Line graphs showing the cumulative percentage of transcripts within the library which contain adapter sequences at each position. The software contains the sequences of the commonly used adapter sequences, in our case the Illumina universal adapter. Post trimming, all adapter sequences have been removed.* ***A & B Lower panel****) Line graphs showing the sequence length distribution :*

4.1.2 Alignment of spliced reads

After selecting short, 100-300 bp, high quality reads, the next task was to align them to the reference genome (GENCODE v29.gtf used in this thesis). The alignment step for RNA-Seq is less straightforward than that for DNA as the reads may not be mapped continuously. In other words, the reference contains the linear genome, including both introns and exons, whereas mature RNA transcripts have the introns spliced out and exons joined. Therefore, an individual RNA transcript may include sequences from more than one exon. For this reason, I chose the 'splice aware' alignment software, STAR (which stands for Spliced Transcripts Alignment to a Reference) [268].

When considering STAR against other aligners, STAR performed extremely well on several parameters including the percentage of reads aligned, accuracy of gene detection and the number of genes falsely quantified when compared against 8 and 13 other splice-aware aligners respectively [382, 383]. In addition to superior performance, STAR allows the identification of two alternative alignment strategies that were appropriate for the identification of both circRNAs and lncRNAs in this thesis. First, STAR allows the detection of 'chimeric' reads, which allows us to identify circRNAs. The reads are not 'chimeric' in the traditional sense of the word (i.e., fusion transcripts formed by trans splicing of mRNAs from

two different genes) but are 'back-spliced' (as discussed previously in section 1.3.3). These transcripts would normally be discarded due to a downstream 5' splice site joining to an upstream 3' slice site (rather than the usual linear splicing involving an upstream 5' splice site joining to a downstream 3' splice site). The choice of aligner dictates somewhat which software can be used to detect circRNAs.

Second, when aligning the linear reads, STAR allows two-pass mapping, which is important for detecting novel splice sites and enabling identification of putative novel transcripts. Briefly, transcripts were mapped using annotated splice junctions from the reference genome (first pass mapping). From this, highly confident novel splice junctions (e.g., the splice site had canonical splice sites, mapping scores were of high quality and it was seen in a minimum number of samples) were collected and added to the annotated junctions before re-running the alignment (second pass mapping), allowing STAR to map the reads to novel splice sites where appropriate.

The output from alignment of linear reads is a SAM file (Sequence Alignment Map) or its compressed binary version – a BAM file. The SAM/BAM file contains all of the information regarding alignment for each read in the sample such as read name, read sequence, read quality, where the read mapped in the genome (with genome co-ordinates), whether the read mapped uniquely or if it mapped multiple times in the genome or if it did not map at all, whether it mapped as a pair or if only the first or second read of the pair mapped (http://samtools.github.io/hts-specs/SAMv1.pdf). Using this information, the SAM/BAM file could be filtered, for example only taking the uniquely mapping reads (i.e. the reads only align to one place in the genome) and removing ambiguous multi-mapping reads, before any downstream analysis. At this point I also performed several very useful quality control checks to check that the library preparation had gone as planned using the RSeQC tool [384]. An overview of the quality control checks and plots are shown in Figure 4-4. First, I used the 'infer experiment' and 'Read distribution' modules to assess potential for DNA

contamination. The ‘Infer experiment’ module predicts the percentage of reads aligned to each strand by analysing a subset of reads from the BAM file. For uncontaminated stranded RNA libraries, the majority of reads will map to either the sense or the antisense strand (depending in which kit was used), whereas in an unstranded or contaminated library I would expect to see a 50%-50% distribution. The ‘Read distribution’ module determines the percentage of reads mapping to coding exons, 5′-untranslated region (UTR), 3′-UTR, introns, and intergenic regions. For RNA analyses we would expect to see the majority of reads mapping to the coding, 5’ and 3’-UTR regions; a more uniform distribution across all regions is indicative of DNA contamination. The RSeQC tool also plots the fragment length distribution of the libraries (discussed in Section 4.1.1). Along with the mapping statistics from the STAR aligner, RSeQC analysis can be nicely visualised with MultiQC software (Figure 4-4).

4.1.3 CircRNA read identification

CircExplorer2 was chosen to identify circRNA back spliced reads it is compatible with the STAR2 aligner and compares favourably on sensitivity and specificity when compared with other software [385]. Briefly, STAR outputs ‘chimeric’ or split reads which can then be filtered to be on the same chromosomes and within a certain distance. These are then input to CircExplorer2 which outputs a table of annotated circRNAs with raw counts which is then input with the linear reads into DESeq2 (section 4.1.6). Normalising for library size for circRNAs is challenging, as quantification of circRNAs is based only on back-spliced junction reads and therefore represents only a tiny fraction of the entire library size. Normalising using back-spliced junction reads alone would potentially ‘normalise away’ any difference in total circRNA production. Consequently, back spliced junction reads were normalised using the normalising factors generated from the *linear* reads (i.e., annotated mRNAs and lncRNAs) with DESeq2.

4.1.4 Transcript assembly – StringTie

After splice aware read alignment, the next bioinformatic challenge was to assemble these reads accurately into transcripts and genes for abundance estimation. This is a computational challenge: the software needs to correctly piece together the millions of short reads (~200-300 bp) produced in the sequencing run.

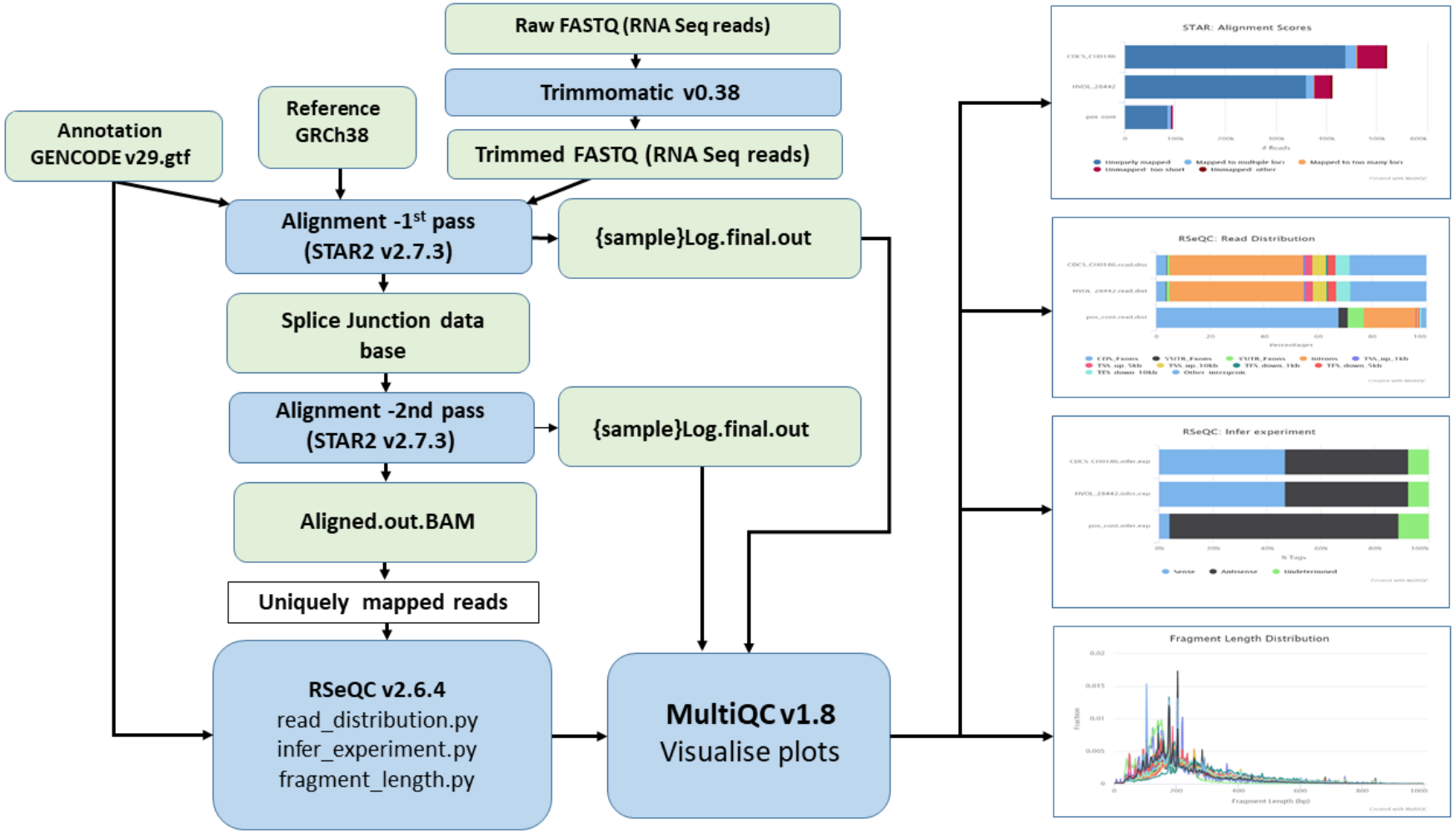
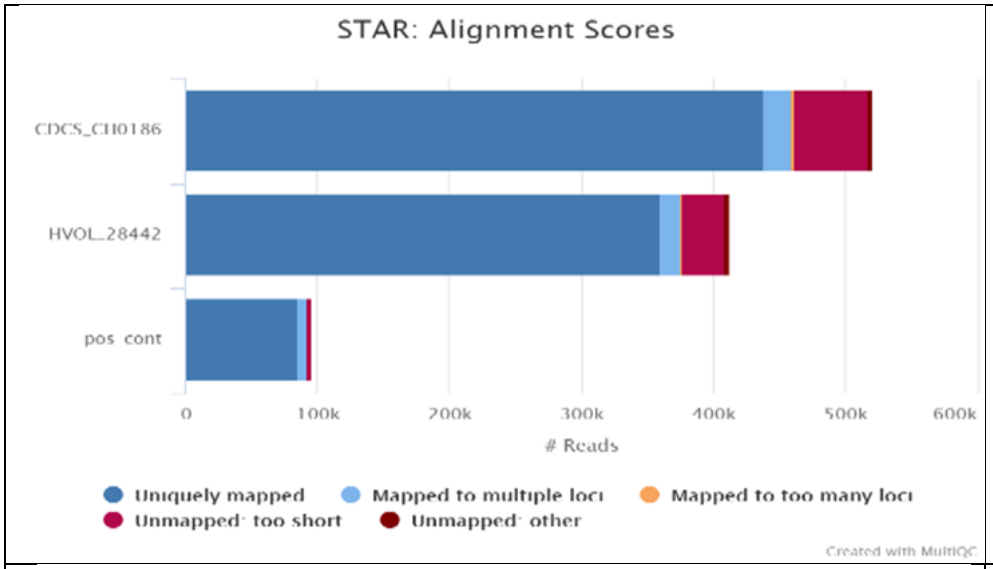
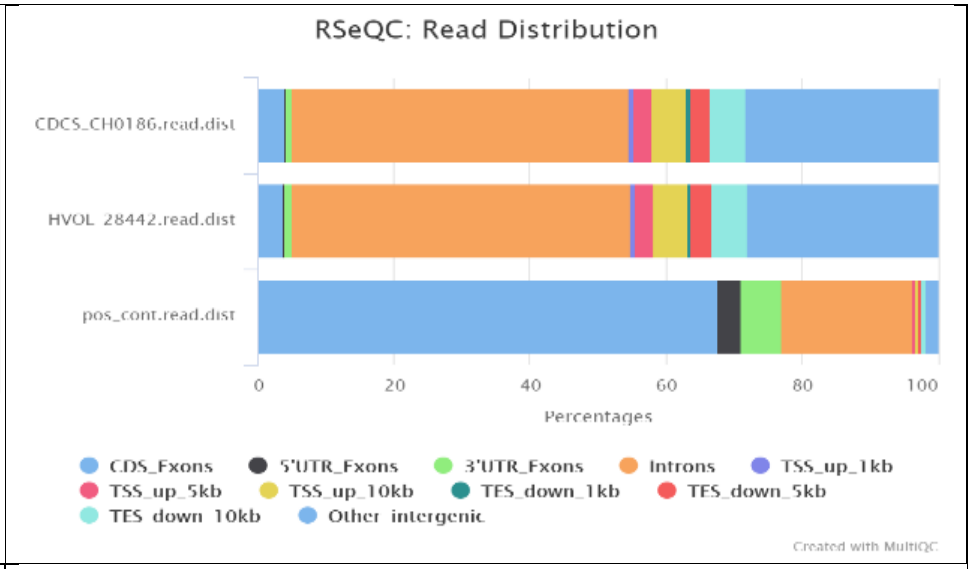
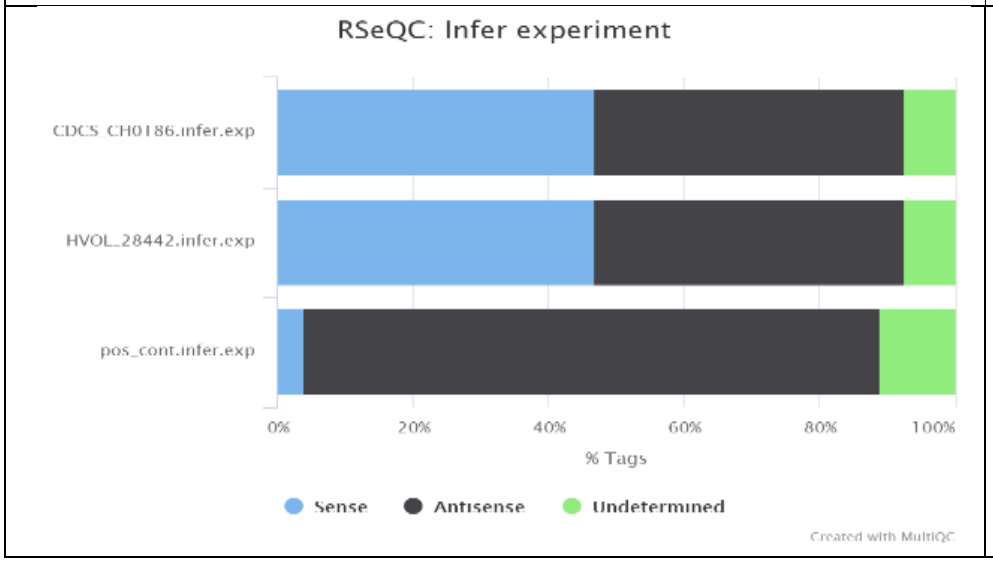


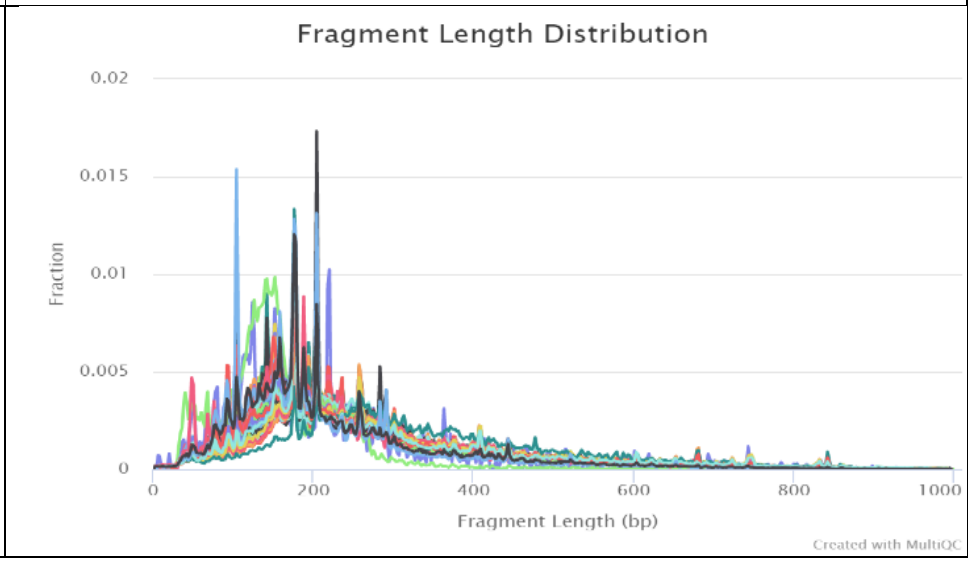
Figure 4-4 An overview of the quality control pipeline showing representative RNA samples

To plot the percentage of reads mapping uniquely, multiple times or not mapping, data is extracted from alignment log files and plotted with MultiQC software. For the remaining plots 'read distribution'- the percentage of reads mapping to each feature of the genome, 'infer experiment' – the percentage of reads mapping to each strand, and 'fragment length distribution' data from the BAM files are input to RSeQC and visualised with MultiQC (the plots are enlarged on next page)


STAR: Alignment Scores
CDCS_CH0186
HVOL_28442
pos_cont
0
100k
200k
300k
400k
500k
600k
Reads
Uniquely mapped
Mapped to multiple loci
Mapped to too many loci
Unmapped: too short
Unmapped: other
Created with MultiQC


RSeQC: Read Distribution
CDCS_CH0186.read.dist
HVOL_28442.read.dist
pos_cont.read.dist
0
20
40
60
80
100
Percentages
CDS_Exons
5'UTR_Exons
3'UTR_Exons
Introns
TSS_up_1kb
TSS_up_5kb
TSS_up_10kb
TES_down_1kb
TES_down_5kb
TES_down_10kb
Other_intergenic
Created with MultiQC


RSeQC: Infer experiment
CDCS_CH0186.infer.exp
HVOL_28442.infer.exp
pos_cont.infer.exp
0%
20%
40%
60%
80%
100%
% Tags
Sense
Antisense
Undetermined
Created with MultiQC


Fragment Length Distribution
0.02
0.015
0.01
0.005
0
Fraction
0
200
400
600
800
1000
Fragment Length (bp)
Created with MultiQC

Issues included correctly assigning reads to regions of the genome that are very similar to other regions, for example pseudogenes (which are sequences that have high homology to functional genes but are not coding into proteins) [386] or multicopy gene families (groups of similar genes that have originated from a common ancestral gene) where reads cannot be placed unambiguously. Additionally, different isoforms may share the same exons. Reference guided assembly clusters reads in the same location together, rather than using overlapping sequences. StringTie, which was chosen for transcript assembly in the pipeline, compares favourably when tested against other popular reference guided assemblers [387]. StringTie uses annotation as a guide but finds novel transcripts when the data dictates. StringTie clusters the reads and builds a splice graph model to represent all possible isoforms of a gene. Where StringTie differs is that it also estimates the 'heaviest path' of transcript abundance (i.e., identifies the most abundant isoform among all splice variants for each gene). StringTie then removes these reads from the splice graph and repeats the process until all reads have been assigned to a splice variant. In contrast, Cufflinks, one of the most popular assemblers, uses a different algorithm to calculate the minimum number of transcripts that explains all of the reads without taking abundance into account. After initial assembly, the assembled transcripts are merged together by a special StringTie module, which creates a uniform set of transcripts for all samples [388]

Of note, sequencing assembly will become less ambiguous as sequencing reads become longer with the evolution of third generation sequencers. As already mentioned, with longer read lengths, transcript assembly becomes more accurate as there are fewer ambiguous options for the assembler software to consider. Indeed, if the read is long enough it will cover the entire transcript and a more accurate representation of the different isoforms for each gene will be achieved.

4.1.5 Detection of novel transcripts

Once the transcripts were identified and mapped with genomic co-ordinates, I then used the GFF compare software (https://ccb.jhu.edu/software/StringTie/gffcompare.shtml) to compare these transcripts against all of the known transcripts in the reference genome (GENCODE annotation). GFF compare outputs a 'class code' for each transcript as to where it is in the genome with regards to annotated genes. For example, if it is a complete match to a reference transcript, that is all exons and introns match, then it is given a '=' class code (Figure 4-5A). Once each transcript was given a class code, potential novel lncRNAs were selected by filtering transcripts with the following class codes: 'u' which denotes an intergenic transcript (in between two transcripts/genes), 'i' which is an intronic transcript, 'x' which overlaps an exon of a known transcript/gene but the transcript is on the opposite strand of the reference and 'y' where the novel transcript straddles a known transcript/gene (Figure 4-5B).

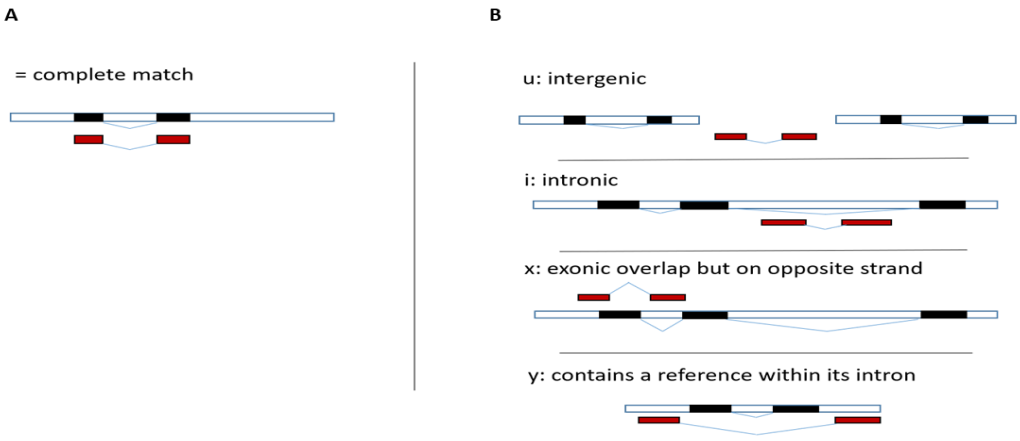


Figure 4-5 A schematic of the differing transcript class codes output by GFF compare.
A) *A complete match is given a '=' class code,* ***B)*** *Class codes for putative novel transcripts. The black boxes denote the reference exons on the genome (blue box), the red boxes denote the transcript in the test data set. The exons connecting intron(s) are denoted by the triangular lines. This is not an exhaustive list of class codes and a complete list can be found in the GFF compare manual* https://ccb.jhu.edu/software/StringTie/gffcompare.shtml

To identify putative novel lncRNAs, I also selected for transcripts longer than 200 bases (the length which differentiates long from short non-coding RNAs). These transcripts were then assessed for their lack of coding potential using the Coding-Potential Assessment Tool (CPAT)[389]. CPAT uses four features to predict coding potential: open reading frame (ORF) length, ORF coverage, Fickett Score and Hexamer usage bias. ORF length is used, as a long ORF indicates coding potential (although to complicate matters, it is now known that short ORFs can be translated from lncRNAs in what are termed micro peptides [102, 390]). Independent but complementary to this is the ORF coverage, which is defined as the ratio of ORF length to transcript length. Non-coding transcripts may contain ORFs by chance, but they usually have a shorter length relative to the overall length of the transcript. The Fickett score assesses the nucleotide composition of codons and considers the degree to which each base is favoured in a codon position and the percentage composition of each base. The score represents the probability that the nucleotide composition of that codon would be predicted to be coding or non-coding. Lastly, the Hexamer score considers certain dependencies of adjacent hexamers (6 bases or 2 codons) in protein-coding sequences. The software was developed using all four features on training data to optimise an algorithm for different species. A threshold of 0.364 gave the highest sensitivity and specificity (0.966 for both) for human data [389] This default threshold for human was used for the pipeline.

4.1.6 Gene and transcript quantification.

The final step of the pipeline for linear transcripts and genes was quantitation and differential expression analysis. Again, this was not straightforward, due to well-known biases in RNA-Seq data that need to be accounted for, mainly due to what is known as ‘positional bias’ (which stems from a preferential generation of fragments from either the start or end of the transcript during the fragmentation process) and under representation of transcripts with a higher GC content [391-394].

Positional bias stems from a preferential generation of fragments from either the start or end of the transcript during the fragmentation process of library preparation. There may also be positional bias due to degradation of the RNA itself. GC biases are thought to arise from the PCR amplification step (GC rich regions can form secondary loop structures that may impede the PCR reaction) during library preparation and results in fragments with high GC content being underrepresented. Either bias makes transcript estimation problematic as exons from different isoforms may not be represented accurately. Consequently, Salmon software was chosen over StringTie for gene and transcript abundance estimations in the pipeline [395] as it has a transcriptome-wide estimation method which can simultaneously estimate and correct for these biases and compares favourably to other transcript quantification software [396]. Similarly, the developer of StringTie also used Salmon for abundance estimation in a large-scale RNA-Seq experiment [397]. The tximport package [398] (developed by the authors of two of the best-known RNA-Seq statistical packages, EdgeR and DESeq2) was used to summarise the output of Salmon (as well as other transcript level quantifiers) to give estimated transcript levels counts and transcript lengths which were input to EdgeR or DESeq2 for gene level analysis [399].

Finally, once the number of each transcript in a sample had been counted (termed raw counts), DESeq2 [400] was used for differential expression analysis. DESeq2 is a popular statistical package for RNA-Seq and performs well when compared to other packages [401]. RNA-Seq experiments usually suffer from non-normality of data and the presence of outliers making inferential statistics difficult. DESeq2 performs a normalisation where it takes the geometric mean for each gene across all samples (the geometric is less affected by very small or vary large values in skewed data). Each gene's count in the sample is then divided by this mean to calculate a ratio. The median of these ratios is the size factor which corrects for

library size and composition bias (e.g., to account for a small number of genes being very highly expressed in one sample but not the other) [400]

When looking at differential gene expression between two groups the most common approach is to test the null hypothesis that the logarithmic fold change (LFC) between the two groups is zero [400]. This produces a list of genes passing multiple-test adjustment. This is difficult for noisy data such as RNA-Seq data which is likely with small biological replicates and lowly expressed genes (where small differences are exacerbated) as within group variance is likely to be high. DESeq2 tackles this problem by sharing information across genes - assuming that genes with similar expression levels have similar variance levels (or dispersion levels). DESeq2 builds a model of dispersion by plotting each genes dispersion and then fitting a smooth curve to capture the expected dispersion for genes at a given strength. It then shrinks noisy gene-wise dispersion towards this curve. The strength of the shrinkage depends on the number of samples: as the sample size increases, the shrinkage decreases in strength and also how close the dispersion is to the fit (curve).
DESeq2, the final step in the pipeline, generated a list of differentially expressed genes and transcripts, for subsequent analyses.

4.1.7 Summary

The bioinformatics pipeline was developed to analyse RNA-Seq data for differentially expressed annotated messenger (m)RNAs and long noncoding (lnc) RNAs, novel lncRNAs and circular (circ)RNAs combined trimming, quality control, alignment, mapping and quantification steps into a seamless analysis. Each step offered multiple different software options, were thoroughly assessed at the time of pipeline development and the best performing software selected. Different software produces different files, and each output

needed to be compatible with the next piece of software. Added to this was the challenge of amalgamating pipelines for mRNA, lncRNA and circRNA analysis.

The main advantages of this pipeline is that it is unique in that it is interrogating three different classes of RNAs simulataneously – annotated mRNAs and lncRNAs, novel lncRNAs and circRNAs.

The next section – the first results section, reviews the validation of this pipeline using publicly available data. Once each part of the pipeline was validated, the pipeline was combined into SnakeMake (https://snakemake.readthedocs.io/en/stable/) which is a Python language workflow management tool developed to create reproducible data analysis.

4.2 Pipeline Validation

4.2.1 Introduction

The pipeline was validated with three publicly available datasets – one dataset for annotated coding genes [95], one dataset for annotated and novel lncRNAs [402] and one dataset for circRNAs [403]. Differences between my pipeline and that of Yang *et al* [95], Mirsafian *et al* [402]and Memczak *et al* [403] are listed in Table 4-1 and discussed in the text.

For clarity I have combined the methods for all three branches of the pipeline into the following section. I then present the results of each data set sequentially and discuss the pipeline development and validation in the final section.

4.2.2 Methods:

The following methods are summarised in Figure 4-1 For all datasets the raw reads in FASTQ format were trimmed to remove adapter sequences and low-quality bases using Trimmomatic v0.38 [379]. A sliding window was used to trim bases with a Phred quality score < 20 with a minimum length filter of 36 bases. Quality trimmed raw reads were aligned

to the human reference genome sequence (GRCh38.p12) using STAR2 v2.6.0 [268] using ENCODE parameters (options for long RNA-Seq detection, see STAR v2.6.0 manual https://docplayer.net/91085649-Star-manual-2-6-0a-alexander-dobin-april-23-2018.html) providing GENCODE v29 [404] as a reference annotation.

Table 4-1 Differences between the bioinformatic analysis for the publicly available datasets versus the pipeline.

Annotated mRNA		
	Yang *et. al.* 2014	Pipeline
Trimming:	Adapter only	Adapter and quality > Q20 (Trimmomatic v0.38)
Human Reference:	Hg19 (University of California Santa Cruz UCSC February 2009)	GRCh38.p12 (Genome Reference Consortium December 2017)
Reference Annotation:	RefSeq (version unknown), Ensembl (version unknown) NONCODE 3.0 (January 2012)	GENCODE v29 (December 2017)
Aligner:	TopHat (version unknown)	STAR2 v2.6.0
Annotated and novel lncRNAs		
	Mirsafian *et. al.* 2016	
Trimming:	Adapter and quality > Q20 Trimmomatic (version unknown)	Adapter and quality > Q20 (Trimmomatic v0.38)
Human Reference:	GRCh38.79	GRCh38.p12 (Genome Reference Consortium December 2017)
Reference Annotation:	GENCODE v22	GENCODE v29 (December 2017)
Aligner:	HISAT v0.1.4	STAR2 v2.6.0
Transcript Assembler:	Stringtie v1.3.3	Stringtie v1.3.3
Coding Potential Software:	Coding Potential Assessment Tool (CPAT) v??? Cut-off 0.375	Coding Potential Assessment Tool (CPAT) v1.2.3 Cut-off 0.364

Quantification of genes/transcripts	CuffQuant, Cuffnorm	Salmon v0.10, Tximport v1.9.12, DESeq2 v1.21.23
CircularRNAs		
	Memczak *et. al.* 2015	Pipeline
Trimming:	-	Adapter and quality > Q20 (Trimmomatic v0.38)
Human Reference:	Hg19 (Feb 2009, GRCh37)	GRCh38.p12 (Genome Reference Consortium December 2017)
Aligner:	Bowtie2 v2.1.0	STAR2 v2.6.0
Annotation:	ENSEMBL release 75	Stringtie GTF (including GENCODE v29 and novel transcripts), hg38 RefSeq, hg38 KnownGenes
Circular detection	Reads not mapping continuously to the genome and head-to-tail splicing	CIRCexplorer v2.3.2

The STAR2 chimeric read alignment option was turned on for circRNA detection. The aligned reads were then assembled into transcripts using Stringtie v1.3.3 [388] to produce Gene transfer format (GTF) files for each sample. Transcripts with a minimum Fragments Per Kilobase of transcripts per Million mapped reads (FPKM) of 0.1 (as stipulated from the Mirsafian et al study I replicated) from individual sample GTF files were merged to form a single set of non-redundant transcripts using the "Stringtie merge" command. This merged Stringtie GTF was then compared against the reference genome using gffcompare v0.10 to classify annotated and novel transcripts. For novel lncRNA detection, transcripts that were non-homologous to any known coding or non-coding transcripts were filtered to only include transcripts (1) longer than 200 nucleotides; (2) with a low coding potential <0.364 (calculated using the Coding Potential Assessment Tool (CPAT) v1.2.3 [389]; (3) expressed at ≥0.1 FPKM in more than one sample (to exclude 'noise' and transcripts with low abundance) (4)

are multi- exonic (as stipulated by Mirsafian *et. al.*, this filter is for validation purposes only) and (5) are intergenic to any GENCODE transcript (as stipulated by Mirsafian *et. al.*, for validation purposes only).

For comparison with the published data FPKM values were assessed using the Bioconductor package Ballgown v2.13.1. Abundance estimation for differential expression analysis (not presented here) was generated using Salmon v0.10 [395]. Salmon requires an index built from the transcript sequences to quasi-map RNA-Seq reads in the quantification step. GFFREAD (http://ccb.jhu.edu/software/stringtie/gff.shtml) was used to build an index from the transcript sequences to quasi-map RNA-Seq reads in the quantification step. The raw sequencing reads were then used to generate abundance estimates for each sample. There was a more stringent filter for the annotated genes (mRNAs and lncRNAs) compared to the novel lncRNAs and circRNAs as the analysis for these was more exploratory.
For circRNA detection the chimeric reads that were output from the STAR2 alignment were input to CIRCexplorer2 v2.3.2 [242]. For annotation of the chimeric reads, the reads were compared against the Stringtie GTF. I chose CIRCexplorer2 as it was an established software is the analysis of circRNAs. When compared against all other circRNA detection software it was rated in the top 3 [385] and of the 3 software was the only one compatible with output from STAR.

4.2.3 Results

4.2.3.1 Annotated mRNAs

As no single dataset was available to validate all aspects of the pipeline the mRNAs, lncRNAs and circRNAs had to be validated separately with three different datasets. The following section describes the validation against the first dataset which tested the mRNAs, I could not test for lncRNAs also as the authors provided the annotations of these from NONCODE and not the Ensembl identifiers. Yang *et. al.* carried out deep RNA-Seq profiling

the transcriptome in failing human heart before and after mechanical support with a left ventricular assist device (LVAD). The transcriptome of 8 patients with ischemic heart failure before mechanical support with a left ventricular assist device (LVAD) was compared with that of 8 heart donors with no previously diagnosed heart disease.

There was a total of 233,838,517 input reads from Yang *et al* analysis compared to 216,298,854 from my analysis after trimming (Table 4-2) with 89% uniquely mapped to the human genome (hg19) (Yang *et al*) compared to 94% uniquely mapping to GRCh38 (my data).

Table 4-2 A comparison of total reads after trimming, read alignment and mRNA detection between Yang et al and the my bioinformatics pipeline developed here

	Yang et al	Thesis pipeline
Total reads after trimming	233,838,517	216,298,854
Total reads uniquely mapping	208,203,328 (89%)	202,997,711 (94%)
mRNAs detected (≥3 FPKM in ≥2 samples)	8831	6896

After converting the hg19 co-ordinates to hg38 via UCSC Batch Coordinate Conversion (lifOver) tool https://genome.ucsc.edu/util.html and using the same criterion as Yang *et al* that mRNAs had to be detected at ≥3 FPKM in at least 2 different samples, there were 8,831 mRNAs detected by Yang *et.al.* compared to 6,896 for my pipeline. A total of 6077 of these were detected by both analyses (69%). There were 819 mRNAs that were detected in my analysis that were not detected in Yang *et.al.* analysis with 2,754 mRNAs detected by Yang *et.al.* not detected in my pipeline (Figure 4-6).

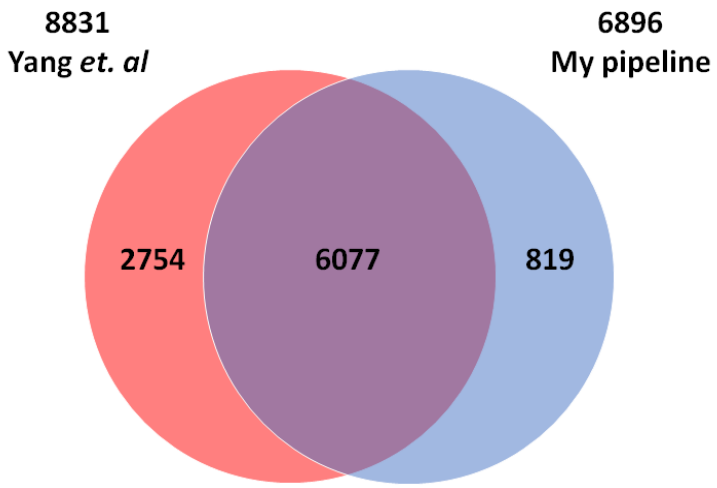


Figure 4-6 Comparison of the number of protein coding genes detected between the thesis pipeline and Yang et.al.

There was a strong correlation between the geometric mean FPKM values of each gene for the two analyses (Spearman correlation of 0.81 after removing genes with geometric means = 0, Figure 4-7).

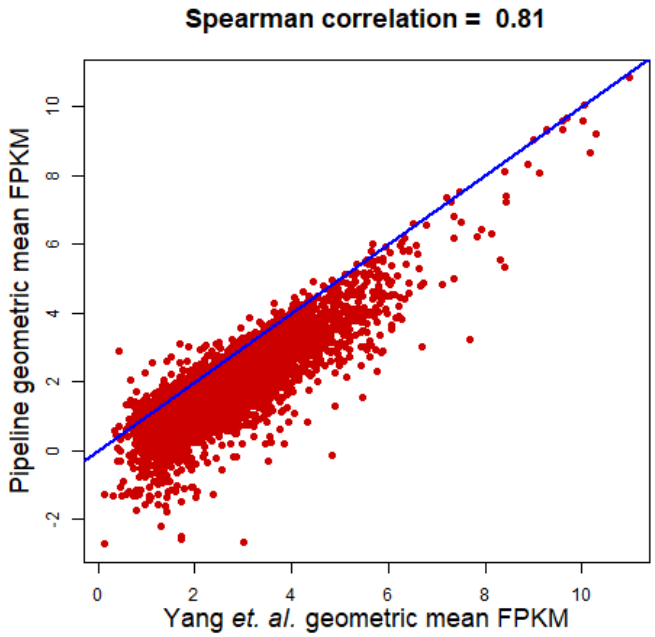


Figure 4-7 Scatter plot showing the geometric mean across all samples for protein coding genes that were identified in both analyses.

Low abundance genes were filtered out leaving only genes that had geometric means > 0 for both analyses, n = 5,570. N.B. Line = y=x

4.2.3.2 Annotated lncRNAs

Annotated lncRNA genes were validated using the dataset from Mirsafian *et. al*. Mirsafian *et. al* performed RNA-Seq in blood cells from four individuals and combined their data with eleven other publicly available datasets to characterise lncRNAs in human primary monocytes (as publicly available dataset for heart tissue could not be found). As the thesis pipeline was designed for paired-end analysis the single-ended Hrdlickova *et. al.* sample was not included in the analysis.

After applying the filter threshold of 0.1 FPKM in at least 1 sample (to replicate analysis from the paper), there were 2,214 (49%) genes that overlapped between the analysis (Figure 4-8). There were 1,392 lncRNA genes that were identified in my pipeline and not Mirsafian *et. al.* with 2,329 lncRNAs being detected by Mirsafian *et. al.* but not by my pipeline. If I did not apply any filter and looked at read counts normalised for sequencing depth only (and not convert to FPKM, which normalises for both sequencing depth and length) 4,473 (98%) genes overlapped which suggests a greater proportion of low abundance genes were being filtered out with the 0.1 FPKM filter when using my pipeline.

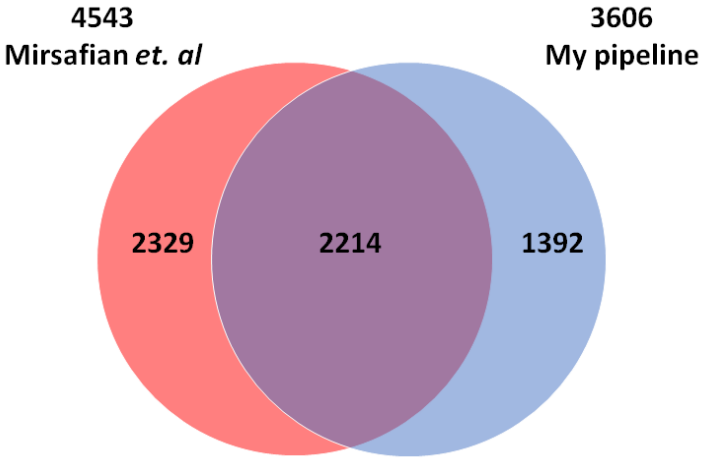


Figure 4-8 A comparison of the number of annotated lncRNA genes detected between the thesis pipeline and Misafian et.al.

2,214 (49% of lncRNAs identified by Mirsafian et. al.) lncRNA genes were detected in both analyses.

Again, genes that were robustly expressed (with geometric means > 0) and identified in both analyses were strongly correlated (Spearman correlation of 0.75, Figure 4-9).

4.2.3.3 Novel lncRNAs

As the Mirsafian *et.al.* dataset also identified novel lncRNA transcripts, I was able to use this dataset for the validation of the novel lncRNA part of the pipeline. After removing transcripts that mapped to annotated mRNAs and lncRNAs, passed an expression threshold of 0.1 FPKM, had a Coding-Potential Assessment Tool (CPAT) score of less than 0.375 and were intergenic Mirsafian *et.al.* were left with 1032 transcripts. For the same data set with the same filters (except the CPAT filter for which I used the human default threshold of 0.364) my pipeline identified 504 putative novel lncRNAs.

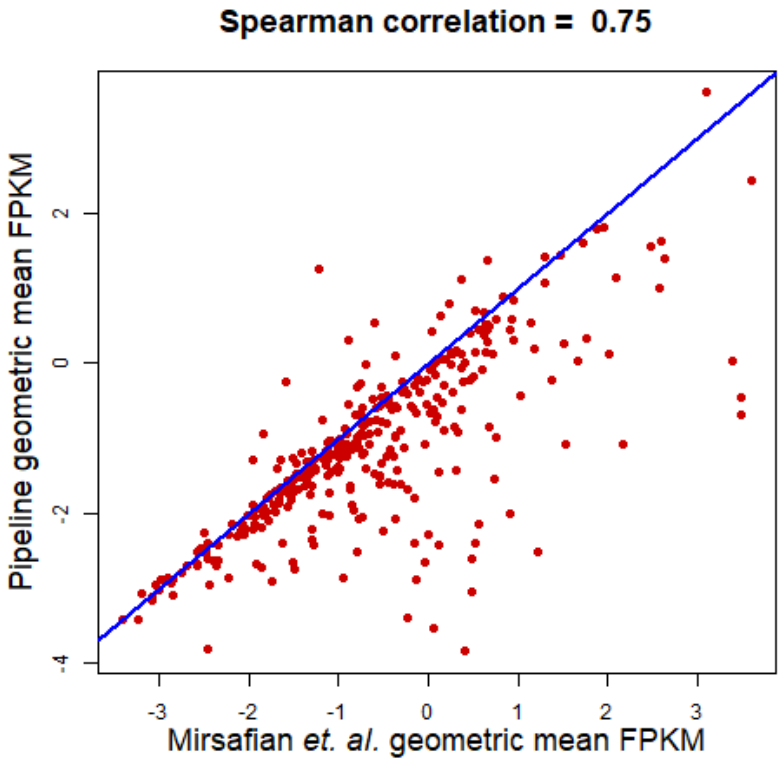


Figure 4-9 A plot showing the geometric mean for each lncRNA that was identified in both analyses.

Low abundance genes were filtered out leaving only genes that had geometric means > 0 for both analyses n = 342. Line = y=x

Comparing the sequences (using the BLAST software) of the putative novel lncRNAs from each analysis I extracted any transcripts that had an overlap of 70% of their length. For some transcripts this resulted in several shorter transcripts aligning to a longer one e.g. three transcripts identified by Mirsafian *et al* aligned to one of the transcripts identified by my pipeline. This meant that 404 (81%) of all putative novel lncRNAs identified by my pipeline overlapped 707 (69%) of putative novel lncRNAs identified by Mirsafian *et al* (Figure 4-10). For this reason the abundance of the overlapping novel transcripts could not be compared between the two analyses.

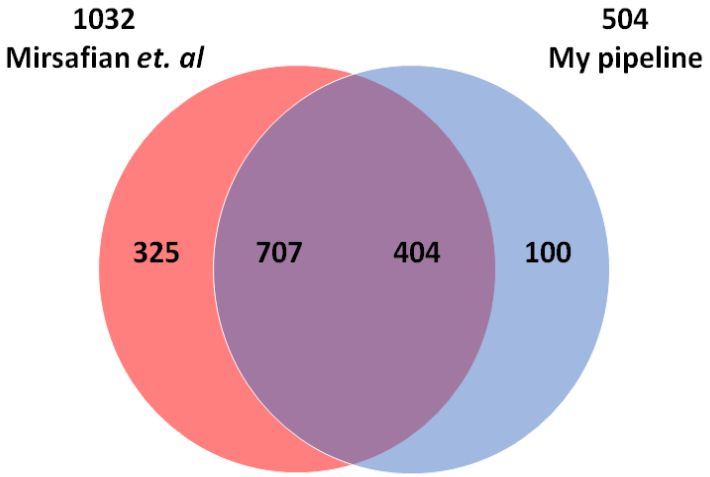


Figure 4-10 A comparison of the number of novel lncRNA transcripts detected between the thesis pipeline and Misafian et.al

707 (69%) of lncRNAs identified by Mirsafian et. al. overlapped 404 (81%) identified by my pipeline.

4.2.3.4 CircRNAs

The circRNA part of the thesis pipeline was validated with data from Memczak *et al* [403]. Memczak *et al* aimed to identify circulating circRNAs by sequencing RNA in human peripheral whole blood from five individuals, including one in duplicate.

Again, as the thesis pipeline trimmed for adapter and quality there were slightly fewer total reads compared with Memczak et. al. However, while the number of reads used for circRNA

detection was considerably lower in the thesis pipeline, the number of circRNAs identified was markedly higher in all samples. A summary of the mapping statistics can be seen in Table 4-3.

Table 4-3 A comparison of read numbers for circular RNA detection between Memczak et.al. and my pipeline.

	Memczak et. al.			My pipeline		
Sample	**Number of total reads (millions)**	**Number of reads used for circRNA detetction (millions)**	**Number of circRNA candidates**	**Number of total reads (millions)**	**Number of reads used for circRNA detetction**	**Number of circRNA candidates**
H1	57.85	9.45	4,550	56.63	242,287	13,066
H1_rep	169.86	28.27	9,996	164.93	701,386	23,746
H2	48.04	7.88	4,105	47.02	246,609	12,211
H3	164.93	24.44	11,113	160.63	888,659	25,942
H4	171.76	13.91	5,739	166.24	376,774	15,740
H5	170.20	29.02	10,002	165.67	618,166	23,977

There was an overlap of 77% circRNAs in both analyses. A total of 38,557 circRNAs were detected by the thesis pipeline that were not detected by Memczak *et. al.* (of which 575 had read counts for all samples), whereas only 4,249 circRNAs were detected by Memczak *et. al.* but not detected in my pipeline (of which 171 had read counts for all samples, Memczak *et. al.*

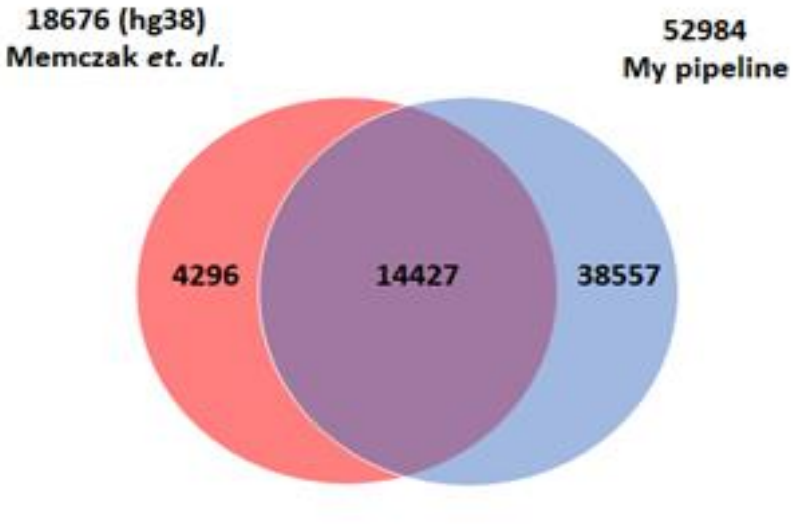


Figure 4-11 A comparison of the number of circRNAs detected between my analysis and Memczak et.al.

A Spearman correlation of 0.91was seen for the geometric mean of read counts for the 1709 circRNAs that were robustly expressed and detected by both analyses (Figure 4-12).

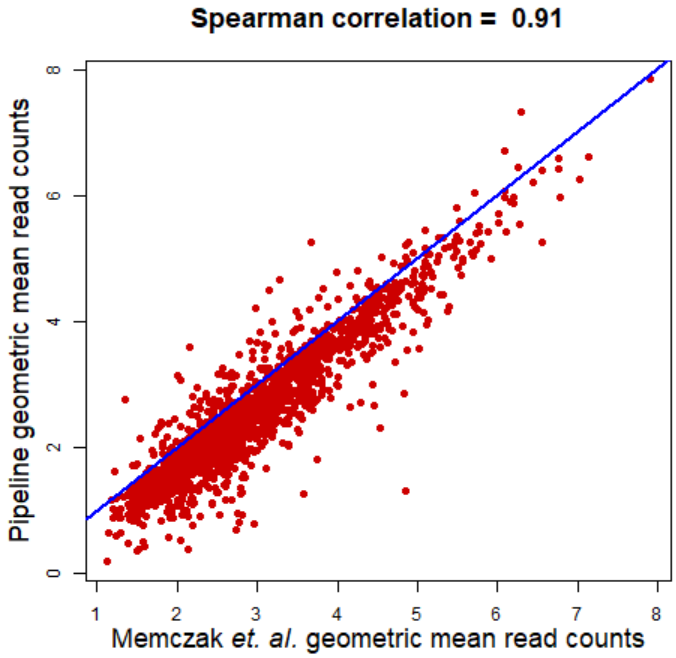


Figure 4-12 A plot showing the geometric mean for each circRNA that was identified in both analyses.

Low abundance genes were filtered out leaving only genes that had geomatric means > 0 for both analyses n = 1,709. Line = y=x.

4.2.4 Discussion

4.2.4.1 Overview of the validation principle

For each part of the pipeline (mRNA, lncRNA, novel lncRNA and circRNA), I searched for publically available data that was i) as close to my project dataset as possible (namely the use of paired-end RNA-Seq data in human heart tissue and, if possible, in human ischaemic and/or failing heart tissue) and ii) in a form conducive to reproducibility. Only the dataset from Yang *et al*. [95] originated from heart tissue with the other two datasets [402, 403] coming from whole blood highlighting the rarity of RNA-Seq data in human heart.

Each part of the pipeline was assesed using the raw fastq files from Yang *et al.*, Mirsafian *et.al* and Memczak *et. al* [95, 402] [403] These were run through the thesis pipeline which differed from the original analyses in several ways, including use of a more recent version of the reference genome (summarised in Table 4-1). In bioinformatics, software, reference sequences and annotations are constantly being updated. Because of these slight differences I was not expecting to achieve 100% reproducibility, but wanted to see significant overlap with the original analysis of the three datasets.

4.2.4.2 Performance of the pipeline for mRNAs

Yang *et. al.* carried out RNA-Seq on five groups of patients to look at the myocardial transcriptome in a non-failing heart group along with failing ischemic and non-ischemic human heart before and after mechanical support with a left ventricular device. To validate the thesis pipeline two groups that were most similar to my the CDCS HF + and HF -patients were selected, namely RNA from non-failing heart and ischemic heart pre-implantation of a left ventricular device.

After trimming for quality and adapters the number of input reads were slightly lower for my pipeline ~216 million compared to Yang *et. al* ~233 million. This was perhaps to be expected as the paper trimmed for adapter sequences only and not for quality. Adapters are ligated to

every fragment of DNA during library preparation and can be incorporated into the read sequence. It is important to remove these adapter sequences with trimming software to minimise the number of reads that do not align uniquely to the genome –the adapter sequence is synthetic and does not appear in any genomic sequence. The choice of trimming for both adapter and quality is to maximise the number of reads that unambigously and uniquely align to the genome. From the validation against the Yang *et al.* data, a higher percentage of uniquely mapping reads after trimming was achieved for the pipeline 94% compared to 89% for Yang *et. al.* As mentioned in section 1.4.2.1, the sequencer outputs a quality score for each base that it incorporates into the read. This score is represented as a PHRED (Q) score which is a probability that is logarithmically linked to the corresponding base being incorrect [378].

The choice of parameters for quality trimming is a trade off between maintaining the accuracy of the bases in the read (in order to maximise mappability) and preserving the number and length of reads for downstream analysis. Indeed, accurate gene expression estimates depend on getting this balance right with large estimation differences and biases being introduced if the balance is less than optimum. [380, 381].

For the alignment step STAR2 was chosen for the thesis pipeline compared with TopHat in pipeline used by Yang *et al*. STAR2 can discover both non-canonical splice sites and chimeric reads both of which the pipeline uses for the discovery of novel lncRNAs and circRNAs respectively. There is a 'two pass' mapping strategy in which the second alignment is supplied with junctions discovered from the first. STAR2 uses this splice junction database as a guide to more accurately align any non canonical (novel) junctions. However, if too many junctions are provided then the number of multi mappers – reads that can align in two or more places in the genome – increases at the expense of unique mappers which ultimately

affects the read counts. A filter was implemented that was strict enough to ensure that splice junctions did not occur by chance yet permissive enough to encourage detection of novel splice junctions and thus novel transcripts. Different filters were tested and showed that as the number of splice junctions passed to the second mapping increased (with a permissive filter) the number of splice sites increased but with fewer uniquely mapped reads which resulted in lower read counts per transcript. A strict filter, which resulted in fewer splice junctions being passed to the second mapping, increased the number of uniquely mapped reads at the expense of splice sites being seen which adversely effected novel transcript detection. A pragmatic filter was chosen: splice junctions had to be present in at least 3 samples with at least 3 uniqely mapped reads spanning the splice junction, in other words, a filter that was permissive enough to allow for novel transcript detection but would filter out junctions that could occur by chance.

After running the pipeline and applying the same filters for abundance as Yang *et. al.*, there was a detection overlap of 69%. However, the Yang et. al study included 827 annotated genes that were not present in GENCODE v29 (the pipeline's annotation) due to either name changes (e.g. *C14orf2* is a synonym for *ATP5MPL* in GENCODE v29), or an update as to the classification of the gene (e.g. *GPX1* listed as an mRNA in Yang *et. al.* is classified as a polymorphic pseudogene in GENCODE v29), this issue could be avoided somewhat by the use of Ensembl identifiers which were not provided by the authors . An overlap of 91% was seen if the analysis was run with no filter of abundance suggesting that the same genes were being detected by the thesis pipeline, albeit at a lower level. Additionally, if I did the analysis providing GENCODE gtf instead of the stringtie gtf there was a Spearman correlation of 0.94. For future analysis I will run the annoatated branch of the pipeline with the GENCODE gtf and the novel branch with the stringtie gtf.

Frustratingly, this dataset could not be used to validate the annotated lncRNA part of the pipeline. Unlike the annotated mRNA results, Yang *et. al.* created a lncRNA database, which instead of containing the lncRNA 'global' transcript name such as Non-Coding RNA Activated By DNA Damage (*NORAD),* used hg19 co-ordinates and transcript names from the NONCODE and Human Body Map lincRNA catalog database (e.g. n407887) along with the software ids of the 113 novels that they identified. After converting the hg19 coordinates to hg38 coordinates there was still no overlap between the lncRNAs identified by the thesis pipeline and Yang *et. al.* Worryingly, it was discovered that even for well known, highly expressed lncRNAs such as *NORAD,* the co-ordinates between GENCODE and NONCODE were different (e.g. in GENCODE there is only one isoform of *NORAD* – a single exon transcript ENST00000565493.1 with hg38 start position on chromosome 20 of 36045622 and stop position of 36050960; for NONCODE the same transcript has a hg38 start position of on chromosome 20 of 36045643 and a stop position of 36051016). Thus without a matching gene/transcript name or matching co-ordinates a reliable comparison could not be made. Also, Yang *et. al.* presented the annotated lncRNAs as *transcripts* and not as genes (there were 81 transcripts with duplicated start and stop co-ordinates – no exon information was given). These findings suggest that providing transcript co-ordinates is not always enough to ensure reproducible research.

As the dataset that was used for validation of novel lncRNAs used the current version of the genome (hg38), this dataset was also used to validate the annotated lncRNAs. The Yang et. al. dataset could not be used for novel lncRNAs pipeline validation as the published data contained truncated transcript co-ordinates, although, as demonstrated above, this is not always enough for reproducible research.

4.2.4.3 Performance of the pipeline for lncRNAs

A key difference between lncRNA analysis pipelines was the genome annotation. Mirsafian *et. al.* used GENCODE v22 which contained a total of 15,900 human lncRNAs genes, whereas the thesis pipeline used the latest (at the time of analysis) GENCODE v29 annotation which contained 16,066 lncRNAs. Mirsafian *et. al.* identified 6,382 lncRNAs, but on further inspection 460 genes were duplicated in the results table. This left 5,922 genes, of which, 4,543 were present in GENCODEv29. Once the same filter of FPKM >0.1 in at least 1 sample was applied, 4,543 lncRNAs remained for comparison, of which 2,214 (49%) overlapped with the lncRNAs identified by the thesis pipeline. However, normalising the thesis pipeline analysis for sequencing depth alone (rather than sequencing depth and length) gave 4,473 (98%) genes overlapping, suggesting the same lncRNAs were being detected in both analyses but at a lower abundance in the thesis pipeline.

One explanation for lower number of read counts from the thesis pipeline is that it uses relatively strict quality filters, such as filtering the BAM files (the files that contain the mapping information from the sequencing reads) for unique mapping reads only. Multi-mapping reads are ambigous and documentation for most software packages that quantify abundance provide insufficient detail on how these multi-mapping reads are dealt with; Are they assigned to only one position based on other quality metrics? Are they counted twice? Or are they not counted at all? By using this filter the thesis pipeline may be discarding reads but at least the remaining reads are of high quality and unambigous.

Another parameter affecting read count is the splice junction filter between the first and second pass alignment as discussed earlier. The pipeline wais designed to identify novel transcripts and this is reflected in the choice of splice junction filter. This, may be at the cost of introducing slightly more multi mapping reads, resulting in fewer reads passing the quality filters. Because lncRNAs are typically expressed at a lower abundance than mRNAs [405]

this lower read count appeared to affect the lncRNA detection rate more than the mRNA rate (although as already mentioned mRNA expression levels were also affected). Again, if I did the analysis providing GENCODE gtf instead of the stringtie gtf there was a much higher Spearman correlation. For future analysis I will run the annotated lncRNA branch of the pipeline with the GENCODE gtf and the novel branch with the stringtie gtf.

Normalising for length is only necessary when the expression of genes with different lengths are compared (because longer genes will have a greater number of reads by chance). Because the proposed analyses will only compare the expression of genes between groups (and not with other genes), the thesis pipeline does not need to normalise genes for length. Normalising for sequence depth alone is the approach recommended by DESeq2 [395], considered to be one of the best performers in RNA differential analysis [402].

For novel lncRNA identification, the thesis pipeline used CPAT as it compared favourably against other software [389, 406]. Mirsafian *et. al*. provided their novel lncRNA transcripts in FASTA format, which is convenient as it removes the introns and only gives the exonic, sequence therefore any differences between intron lengths are redundant. After applying the same filters as Mirsafian *et al* (retaining only multi-exonic, intergenic transcripts that were expressed at 0.1 FPKM in at least 1 sample), the thesis pipeline identified 504 putative novel lncRNAs compared to 1,032 for by Mirsafian *et. al.* A total of 707 (69% of the novel lncRNAs identified by Mirsafian et. al.) and 404 (81% of the novel lncRNAs identified by the pipeline) had over 70% alignment and were considered to match. The reason why the number is not the same for both is that some transcripts matched multiple transcripts from the other analysis. Overall, the putative novel transcripts identified by the thesis pipeline were longer than novel lncRNAs identified by Mirsafian *et. al*., which would suggest that more novel splice sites (and therefore a longer transcript) were being detected in the thesis pipeline.

Interestingly, Mirsafian *et. al.* neither used or justified not using the default coding probability cut-off used by CPAT (as used by the pipeline), which gives highest sensitivity and specificity for human data [389]. This default setting was tested against mock data (using known coding and non-coding sequences) and displayed the best true positive versus true negative count.

4.2.4.4 Performance of the pipeline against circRNA validation

For the circRNA detection CIRCexplorer2 was chosen as it performs well against other circRNA detection software [238, 385]. Memczak *et. al.* did not use specialist software to detect circRNAs but rather developed an in house pipeline that filtered reads that *did not* align continously to the genome. From these reads, the 20 nucleotide terminal sequences were re-aligned to the genome and any reads that were in reverse orientation were further filtered for mismatches, breakpoint detection, alignment scores, reads counts and distance. Additionally, they also filtered for only GT/AG flanking splice site signals (the canonical splice site motif), whereas the thesis pipeline does not filter out non-canoncial splice site motifs. Consquently, the two analsyses are quite different in the way they identify circRNAs - Memczak *et al*. built psuedo reference sequences containing back splice junctions and aligned the unmapped reads to these, whereas the pipeline used STAR and circExplorer2 uses split read alignment of chimeric reads. This perhaps explains why the mapping statistics are also very different. It is not clear whether the number of reads used for circRNA detection for the Memczak analysis includes *all* reads that do not uniquely align (*i.e.* multi mapping reads, reads that are too short or chimeric reads). This may be the case as although the Memczak analysis stated much higher read numbers used for circRNA detection, their analysis had lower numbers of circRNA candidates for each sample compared to the pipeline. Another difference was Memczak *et. al.* used the older hg19 version of the reference genome and so the circRNA co-ordinates had to be converted to the latest hg38 version. Even with these

differences there was an encouraging 77% overlap in circRNAs, which suggests that circRNAs are robustly detected by both methods. There were a large number of circRNAs detected in the pipeline that weren't detetced in the Memczak analysis although a large proportion of these were not detetced across all samples suggesting they may be of low abundance.

4.2.5 Conclusions

The development of the pipeline was an interesting exercise in demonstrating how reproducible bioinformatic research is. There are constant updates in hardware, software, reference files and methods of analysis (as highlighted with the circRNA analysis). The main challenge of genetic research is unambiguous annotation of coding and non-coding genes especially in the field of lncRNAs and circRNAs, where studies are discovering novel genes at what seems like an overwhelming rate. The challenge of pipeline development was to amalgamate three separate branches into one coherent and systematic method of analysis and to validate this using three separate datasets. It was encouraging that there was such a high degree of reproducibility for the datasets, giving confidence in the bioinformatic analysis in the subsequent chapters of the thesis.

Chapter 5
RNAs Associated with Ischaemia in Human Heart

5.1 Introduction

This chapter aimed to use the bioinformatics pipeline developed and validated (Chapter 4) to identify annotated mRNAs and lncRNAs and novel lncRNAs associated with myocardial ischaemia. Illumina RNA- Seq data from 85 paired samples collected from the left ventricle before and after a period of ischemia during cardiopulmonary bypass for aortic valve replacement was provided by Drs Muehlschlegel and Body at the Harvard Medical School [175]. My hypothesis was that lncRNAs help coordinate the post ischemic stress response and that some if these lncRNAs may not have been annotated previously. The idee was to both identify candidate genes and then to generate a set of genes altered in ischaemic cardiac tissue that could be used to look for overlap of genes altered by ischaemia in plasma.

5.2 Overview of research design

A summary of patient characteristics was provided in Section 3.2.1.1 and Table 3-1 gives a summary of patient demographics. Briefly, punch biopsies were taken from the left ventricle apex immediately after the start of cardiopulmonary bypass at the time of routine placement of a surgical vent (pre-ischaemia) and after a median of 74 minutes (interquartile range 61–93 minutes; post-ischaemia), during which time the heart was arrested with cold blood cardioplegia for myocardial protection. The purpose of the experimental design comparing paired pre- and post-ischaemic tissue from the same patient, was to reduce the confounding of underlying factors such as sex, or presence of CAD in some patients. Saddic *et al* checked post operation Creatine kinase-MB levels (an enzyme found in heart muscle test for MI) and these levels were very high confirming that ischemia took place. The raw data files from

Illumina RNA-Seq data files were run through my pipeline (described in Section 4.1) and lists of mRNAs, lncRNAs and novel lncRNAs, including those that were differentially expressed between pre- and post-ischaemia, were produced (circRNAs were not able to be studied owing to the library preparation method used). Quality control checks were carried out by plotting a Principal Component Analysis (PCA) for gene expression for the 85 paired samples and the R package TissueEnrich [375] was used to confirm tissue specificity. A conservative selection criterion was used to identify a 'high-confidence' set of differentially expressed genes/transcripts for downstream analysis: (i) p-value adjusted (padj) <0.001 after adjustment for multiple comparisons (Benjamini-Hochberg), (ii) absolute fold-change >1.2 (post-ischemic/pre-ischemic) and iii) for annotated genes, at least 90% of each 'pre' and 'post' group had to have an expression level of at least 0.5 transcripts per million (TPM). Putative novel lncRNAs detected from Illumina sequencing were validated with long read Nanopore sequencing. Nanopore sequencing sequences the full-length transcript so we can be more confident it is not a partial transcript. Additionally, by using Nanopore sequencing we could validate all putative novels rather than selecting a handful to validate with RT-PCR.

Nanopore sequencing was performed on RNA extracted from 8 donor heart tissues (left ventricle, provided by Cleveland Clinic, Section 3.2.2) with 1% RNA Sequins [371] spiked in as internal controls (Section 3.3.4). Bioinformatic analysis of Nanopore sequencing was performed (Section 3.4.5) to identify a list of putative novel lncRNAs. These were compared to the novel lncRNAs identified with Illumina sequencing (Figure 5-1) using gffcompare software. Differentially expressed mRNAs, lncRNAs and novel lncRNAs were analysed with a weighted gene correlation networks analysis (WGCNA [374], Section 3.6) and pathway analysis (Ingenuity Pathway Analysis, IPA Section 3.7) to identify gene modules, molecular pathways and biological functions associated with an early ischemic response.

In situ hybridisation with RNA Scope was performed on two annotated lncRNAs, one from each module that WGCNA identified as significantly associated with ischaemia to examine their subcellular location. Lastly, one of the differentially expressed novel transcripts that was identified by the pipeline, validated with Nanopore sequencing and identified in one of the modules associated with ischaemia was also analysed with RNAscope to further validate its expression and identify its subcellular location. RNAscope was performed on Formalin-fixed paraffin-embedded (FFPE) left ventricle tissue provided by the Cleveland Clinic.

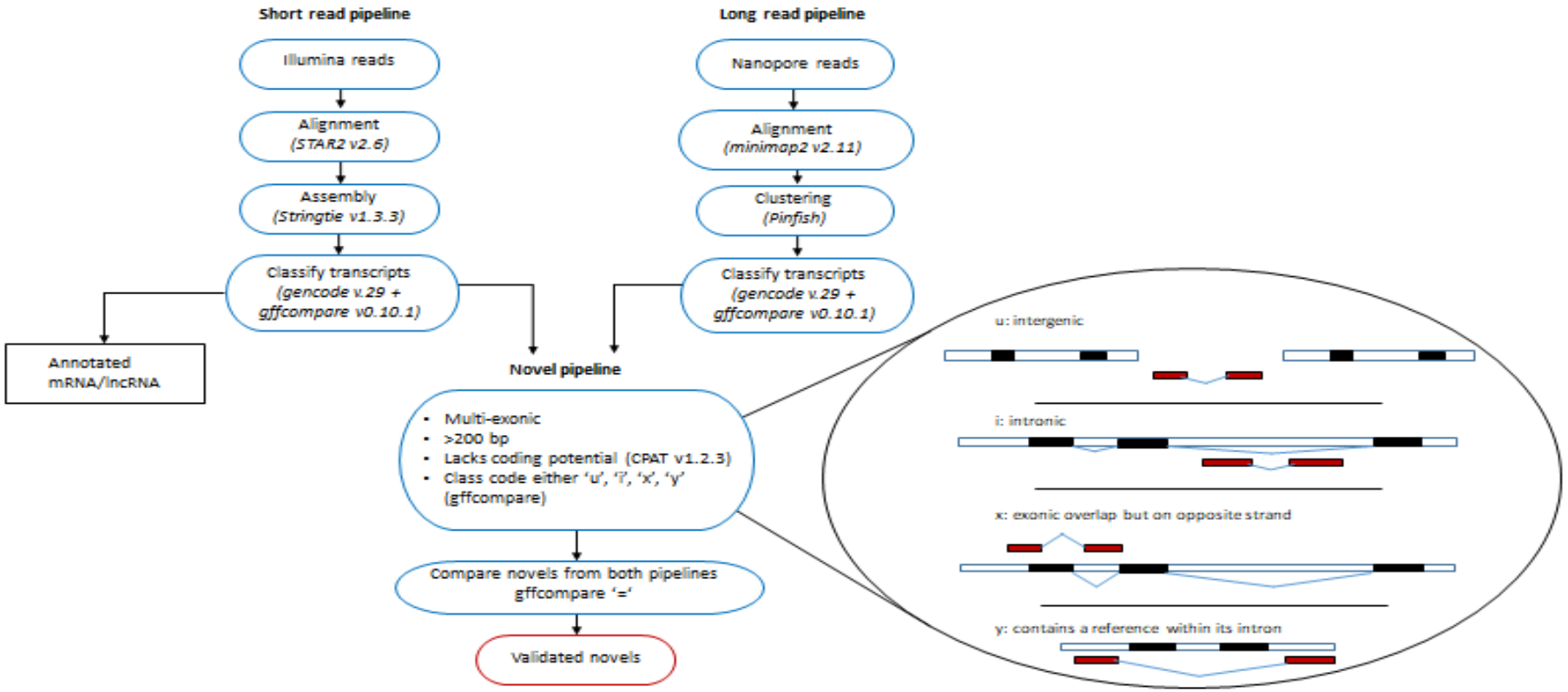


Figure 5-1 A schematic of the pipeline for novel lncRNA discovery and validation.

Illumina reads were aligned with STAR v2.6 and transcripts assembled with Stringtie v1.3.3; Nanopore long reads were aligned with minimap2 v2.11 and transcripts assembled with Pinfish. Transcripts were then assigned class codes with gffcompare v.0.10.1 using GENCODE v.29 as annotation. Transcripts were then filtered for class codes of either 'u' – intergenic; 'i' – intronic; 'x' – transcripts had an exon overlapping an annotated transcript but was on the opposite strand or 'y' – contained an annotated transcript within its intron. Transcripts were also filtered for length > 200 bases and passed the default filter for non-coding human transcript with CPAT v1.2.3 software. Transcripts that passed filters from both read technologies were then compared with each other using gffcompare and matching transcripts (class code '=') were deemed validated novel transcripts.

Positive controls using the lncRNA Nuclear enriched abundant transcript 1 (*NEAT1*) was used as examples of a nuclear enriched lncRNAs respectively [407] .

5.3 Results

5.3.1 Quality Control

Principal Component Analysis (PCA) for the 85 paired samples identified five outlying samples: 47V (-pre and -post), 56V (-post), 73V (-pre) and 107V (-post) (Figure 5-2). The fragment length distribution quality control plot showed that the cDNA libraries for patient 47V (-pre and -post ischaemia samples) had a much longer fragment lengths than the libraries for the rest of the samples (150-350 bases versus the majority <150 bases, Figure 5-3).

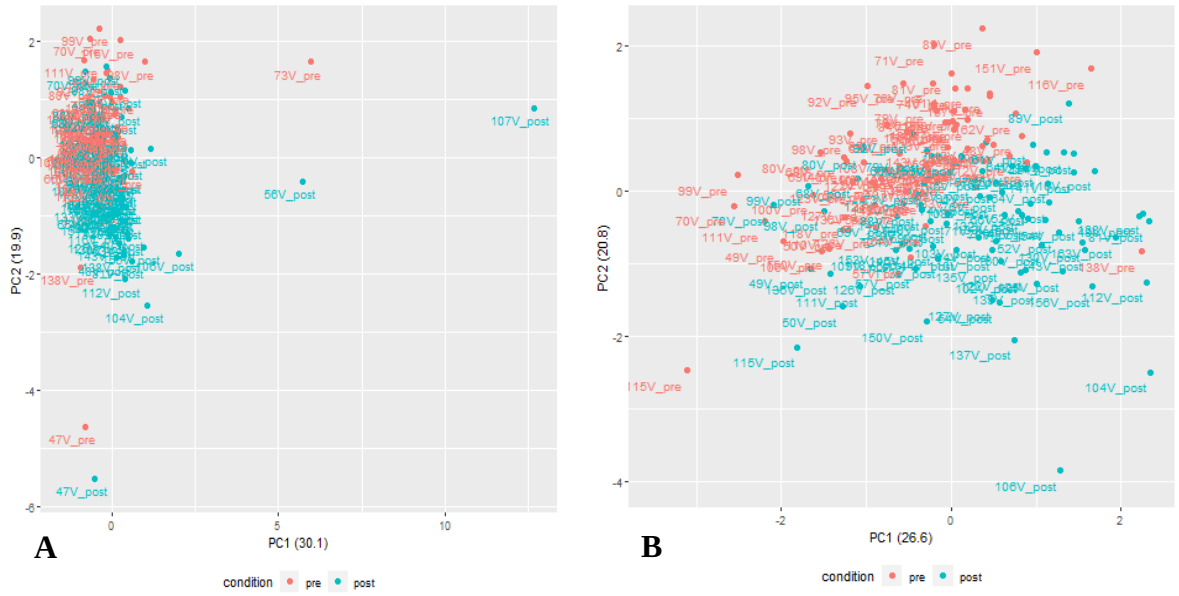


Figure 5-2 Principal Component Analysis (PCA) on the transcriptome of 85 paired pre- and post-ischaemia left ventricular samples.

A) The plot indicates four samples that are outliers B) Data re-plotted with the four outliers removed Red indicate pre ischemic samples, blue indicates post ischemic samples.

While fragment lengths of up to 350 bases are not too long for sequencing, it was concerning that this sample displayed a very different profile compared with the rest of the cohort when all fragments should be uniform, and the sample was discarded. For the remaining three outlying samples, the R package TisueEnrich showed that the probability that these samples

originated solely from heart or skeletal muscle was much lower (Figure 5-4). This suggests that there may have been other tissue taken when the biopsy was performed. These samples were also discarded, which left 81 pairs of samples for analysis.

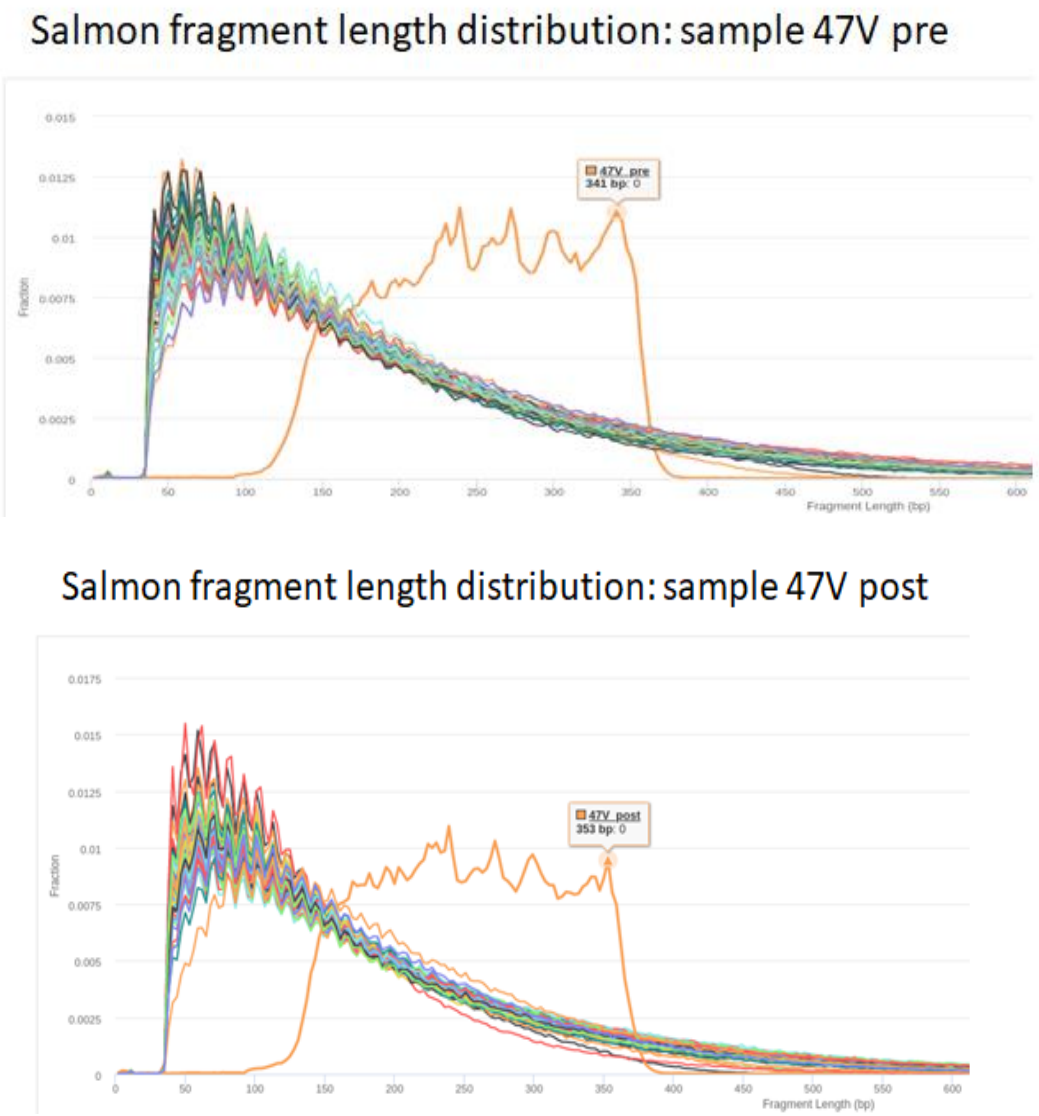
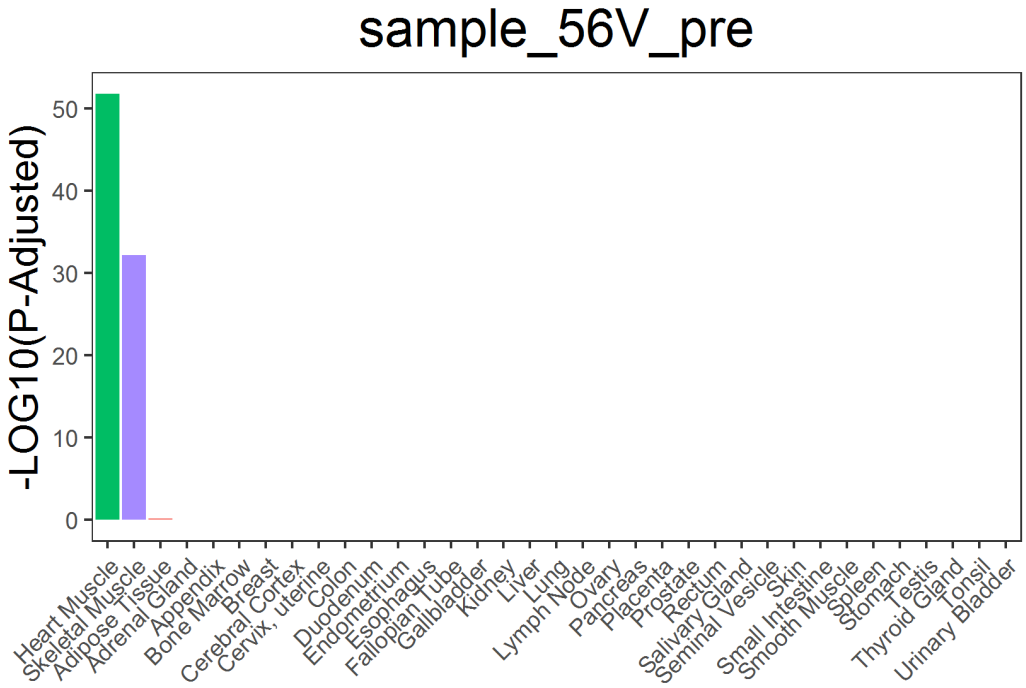
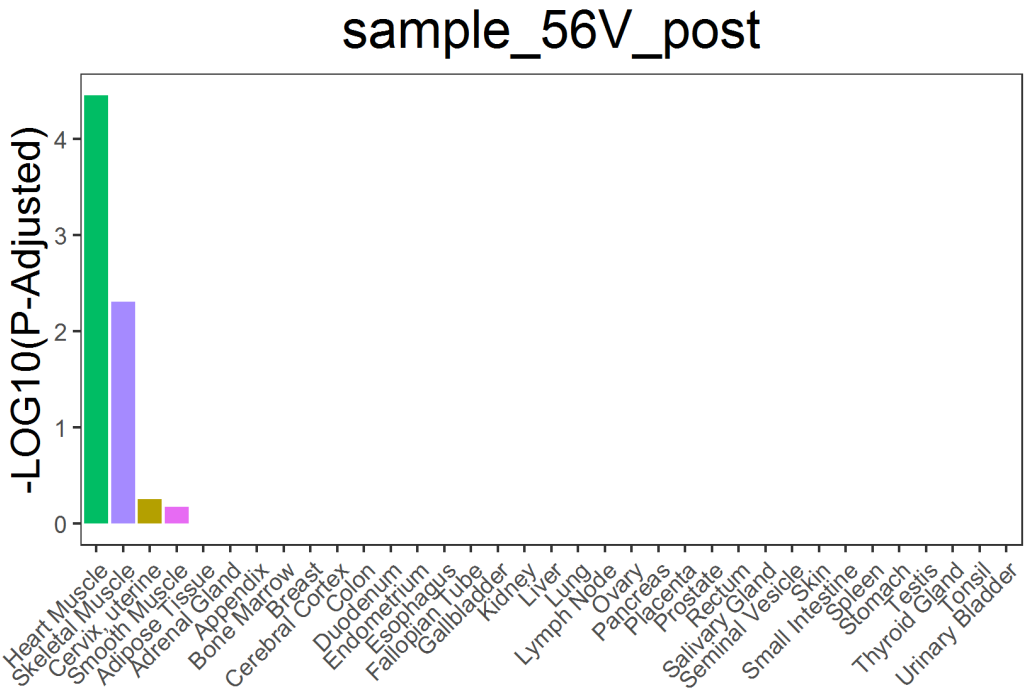


Figure 5-3 Fragment length distribution plots.

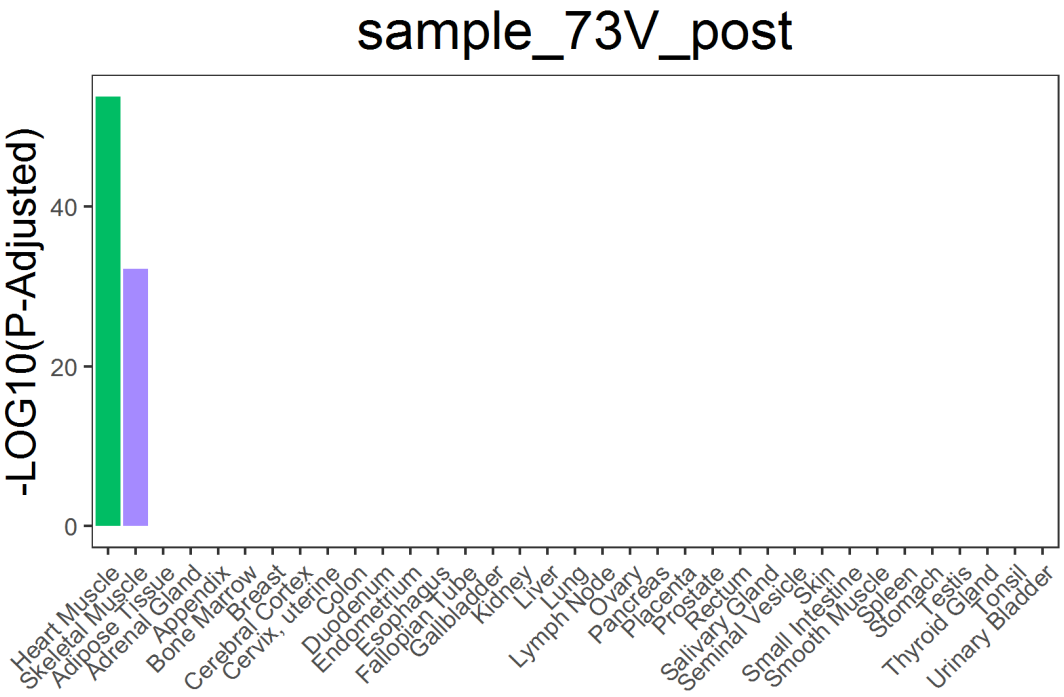
Sample 47V shows much larger fragment lengths compared to the rest of the sample, indicating a potential problem with library preparation.

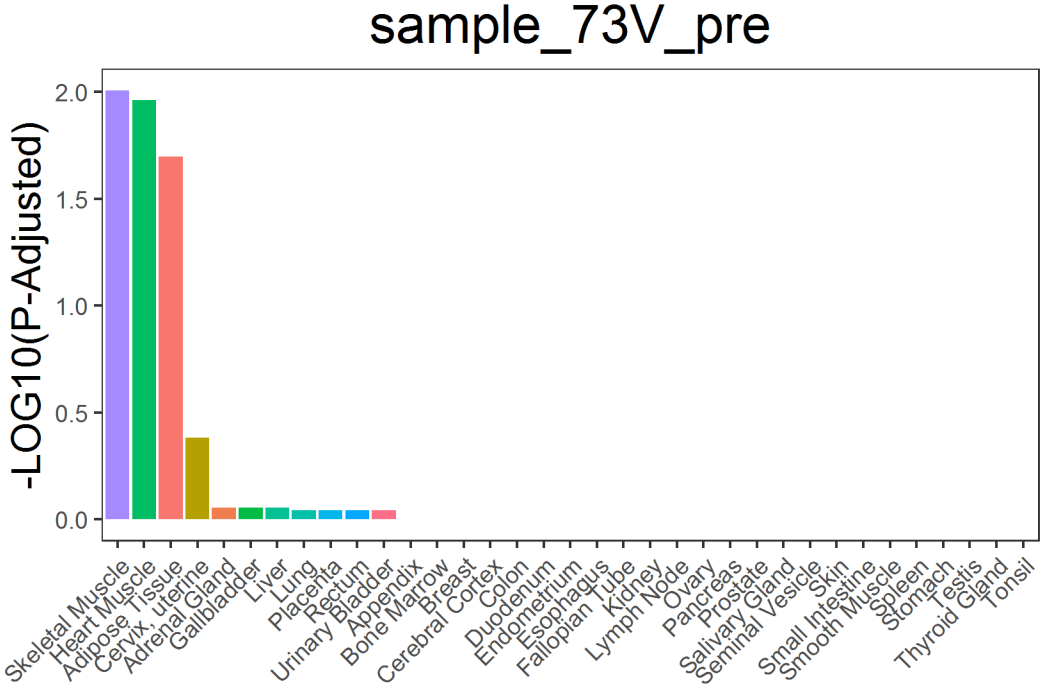
A


sample_56V_pre
-LOG10(P-Adjusted)
0
10
20
30
40
50
Heart Muscle
Skeletal Muscle
Adipose Tissue
Adrenal Gland
Appendix
Bone Marrow
Breast
Cerebral Cortex
Cervix, uterine
Colon
Duodenum
Endometrium
Esophagus
Fallopian Tube
Gallbladder
Kidney
Liver
Lung
Lymph Node
Ovary
Pancreas
Placenta
Prostate
Rectum
Salivary Gland
Seminal Vesicle
Skin
Small Intestine
Smooth Muscle
Spleen
Stomach
Testis
Thyroid Gland
Tonsil
Urinary Bladder


sample_56V_post
-LOG10(P-Adjusted)
0
1
2
3
4
Heart Muscle
Skeletal Muscle
Cervix, uterine
Smooth Muscle
Adipose Tissue
Adrenal Gland
Appendix
Bone Marrow
Breast
Cerebral Cortex
Colon
Duodenum
Endometrium
Esophagus
Fallopian Tube
Gallbladder
Kidney
Liver
Lung
Lymph Node
Ovary
Pancreas
Placenta
Prostate
Rectum
Salivary Gland
Seminal Vesicle
Skin
Small Intestine
Spleen
Stomach
Testis
Thyroid Gland
Tonsil
Urinary Bladder

B


sample_73V_post
-LOG10(P-Adjusted)
40
20
0
Heart Muscle
Skeletal Muscle
Adipose Tissue
Adrenal Gland
Appendix
Bone Marrow
Breast
Cerebral Cortex
Cervix, uterine
Colon
Duodenum
Endometrium
Esophagus
Fallopian Tube
Gallbladder
Kidney
Liver
Lung
Lymph Node
Ovary
Pancreas
Placenta
Prostate
Rectum
Salivary Gland
Seminal Vesicle
Skin
Small Intestine
Smooth Muscle
Spleen
Stomach
Testis
Thyroid Gland
Tonsil
Urinary Bladder


sample_73V_pre
-LOG10(P-Adjusted)
2.0
1.5
1.0
0.5
0.0
Skeletal Muscle
Heart Muscle
Adipose Tissue
Cervix, uterine
Adrenal Gland
Gallbladder
Liver
Lung
Placenta
Rectum
Urinary Bladder
Appendix
Bone Marrow
Breast
Cerebral Cortex
Colon
Duodenum
Endometrium
Esophagus
Fallopian Tube
Kidney
Lymph Node
Ovary
Pancreas
Prostate
Salivary Gland
Seminal Vesicle
Skin
Small Intestine
Smooth Muscle
Spleen
Stomach
Testis
Thyroid Gland
Tonsil

C

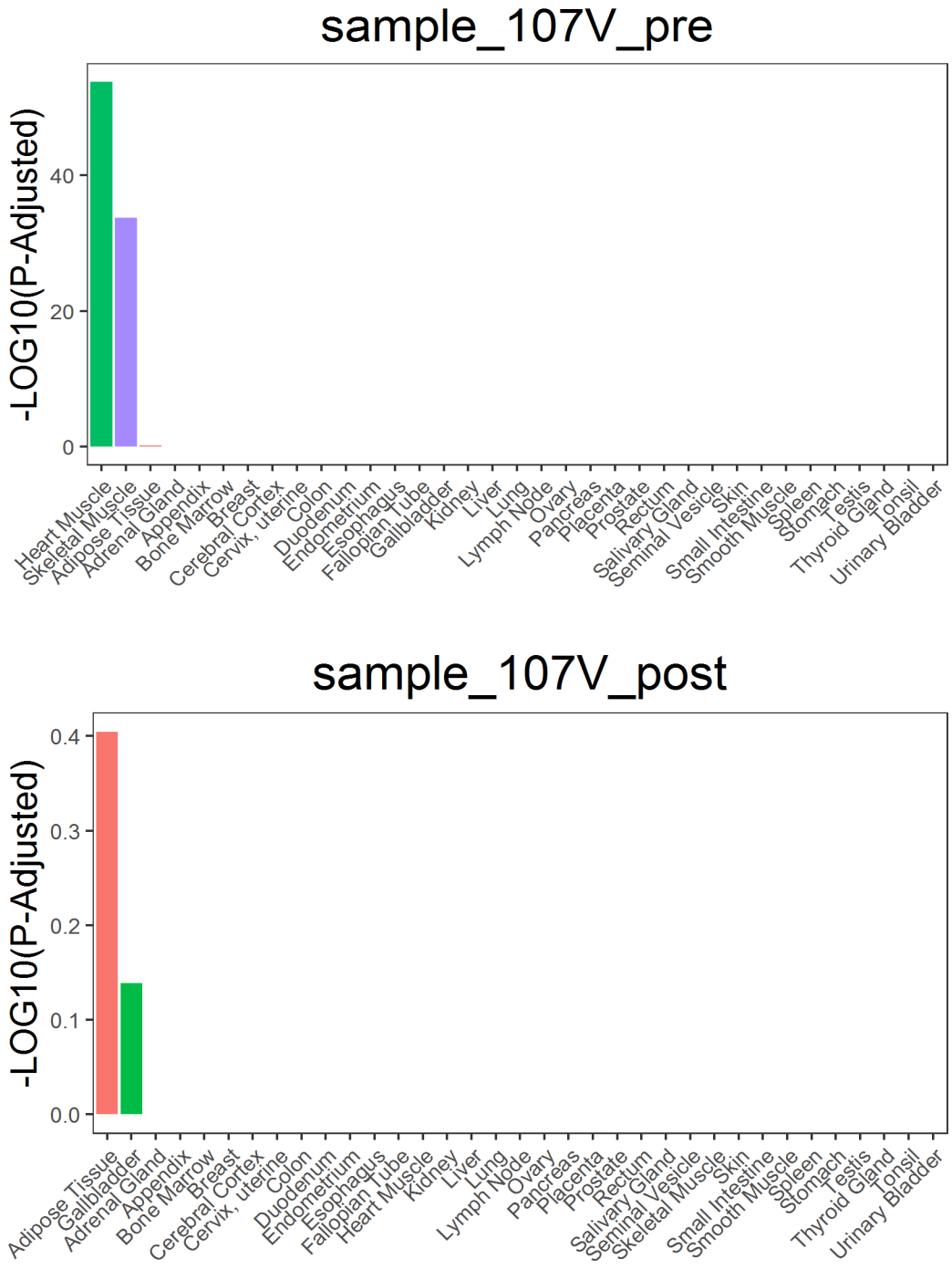


Figure 5-4 Plots from TissueEnrich for the three remaining outlying samples

A) *56V -post,* ***B)*** *73V -pre and* ***C)*** *107V -post (and their paired sample for comparison, shown in the upper panel). For the non-outlier sample for each patient the most probable tissue of origin for the expressed genes are heart or skeletal muscle (-log10 p-adjusted = ~50). In contrast, for the outlier sample, the heart and skeletal muscle p-adjusted value is much lower (-log10 p-adjusted = 0.4-4) and other tissue sources are shown to be similarly probable.*

5.3.2 Protein-Coding, Annotated and Novel lncRNAs Associated with Ischemia identified with Illumina Short Read Sequencing

Illumina RNA-Seq generated an average of 33.7 ± 12.6 million uniquely mapped reads per sample (approximately 85% of the total reads per sample). Unique reads mapped to 12,656 mRNAs and 1,488 annotated lncRNA genes and 10,567 putative novel lncRNAs in human left ventricle, including reported lncRNAs that are strongly expressed in cardiac tissue (but also expressed elsewhere), *MALAT1*, *NEAT1*, *H19*, *TUG1* [177, 408-410]. The overall expression levels of lncRNAs were lower than mRNAs (Table 5-1).

Table 5-1 Expression levels (median and IQR) of mRNAs and lncRNAs detected by my pipeline.

Gene type	Median TPM (IQR)
mRNAs	7.7 (3.4-17.8)
lncRNAs	1.9(1.3-3.3)

N.B. Putative novel lncRNAs could not be compared as no abundance filter was applied and novel lncRNAs were detected at the transcript level, rather than the gene level.

Expression of 2,446 mRNAs, 270 annotated lncRNAs and 1149 novel lncRNAs differed in response to ischemia (padj <0.001, absolute fold change >1.2, Figure 5-5). The 40 genes most differentially expressed genes from each class of RNA (ranked on fold-change) are shown in Appendix C. Of the differentially expressed mRNAs, 11 of the 20 most abundant genes were mitochondrial (*MT-CO1*, *MT-ND4*, *MT-ATP6*, *MT-ND1*, *MT-CYB*, *MT-ND4L*, *MT-CO2*, *MT-CO3*, *MT-ND2*, *MT-ND3*, *MT-ATP8*). Analysis of the RNA spike in controls (Sequins) showed a strong correlation (Spearman correlation = 0.82) between their measured abundance and their input concentration (Figure 5-6), confirming that library preparation and sequencing was successful.

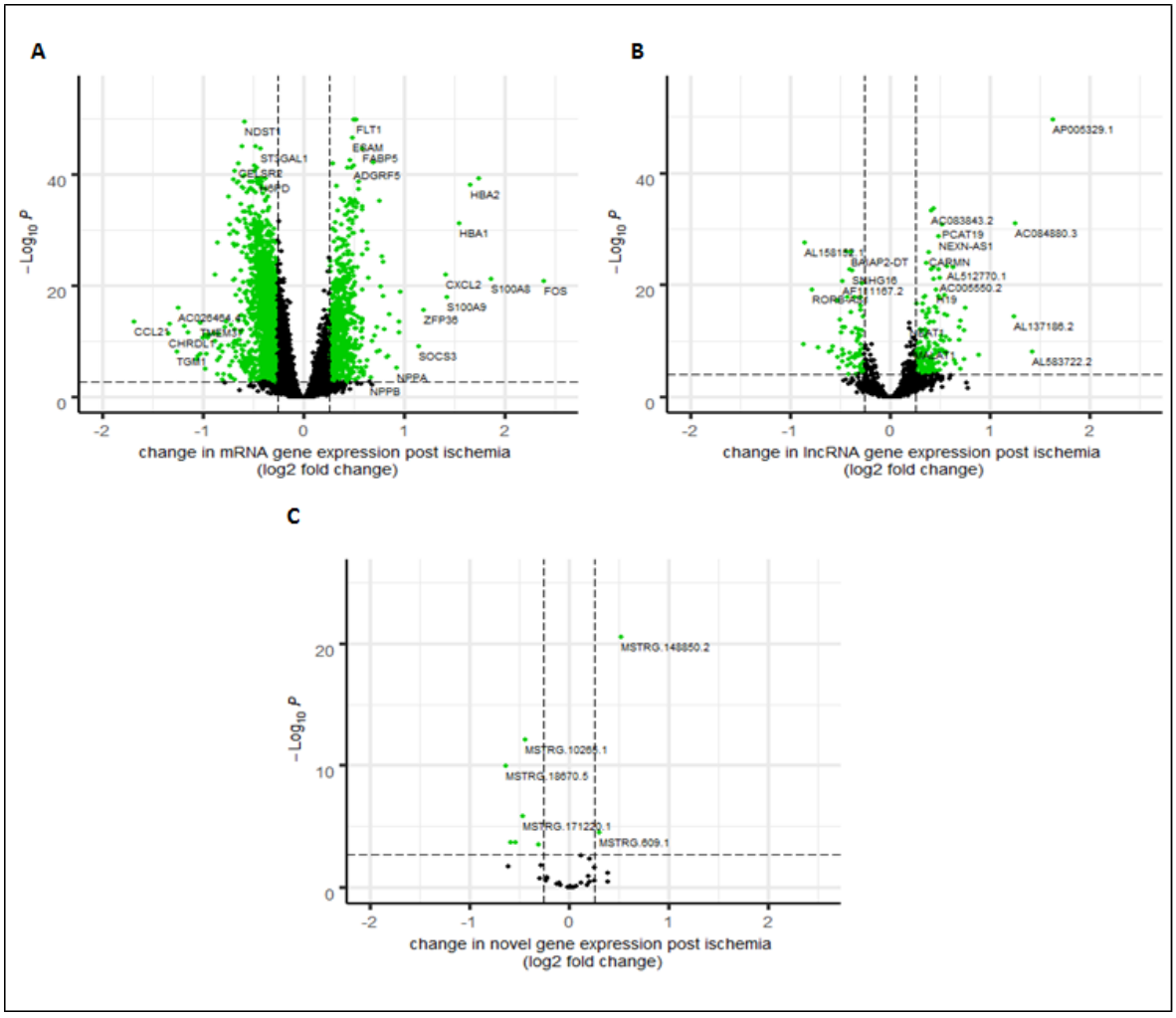


Figure 5-5 Volcano plots showing differential expression in 81 paired human left ventricle samples, comparing pre- versus post-ischemia

A) *mRNA genes,* ***B)*** *annotated lncRNA genes, and* ***C)*** *39 putative novel lncRNA transcripts identified by both Illumina and Nanopore technologies. They are annotated with arbitrary MSTRG identifiers by the Stringtie package in the Bioinformatics pipeline. Green indicates differentially expressed genes/transcripts with an adjusted p value (padj) <0.001 and an absolute fold change >1.2.*

5.3.3 Confirmation of Novel lncRNAs in Human Left Ventricle with Nanopore long read technology

Nanopore RNA-Seq yielded a total of 10,638,219 base-called reads of which 65.1% passed the QC filter of a minimum quality score of 7 (the nanopore base caller sets a minimum quality score of 7.0. These are not phred scores but the algorithms to calculate this score are not disclosed by Nanopore). Reads had a mean length of 547 nucleotides and a mean quality

score of 10 Of these, 2,710,486 full-length transcripts were identified, which had an alignment rate of 96.33% (75.6% to hg38 and 20.7% to chrIS). From these transcripts 153 potential multi-exonic, novel lncRNAs were identified.

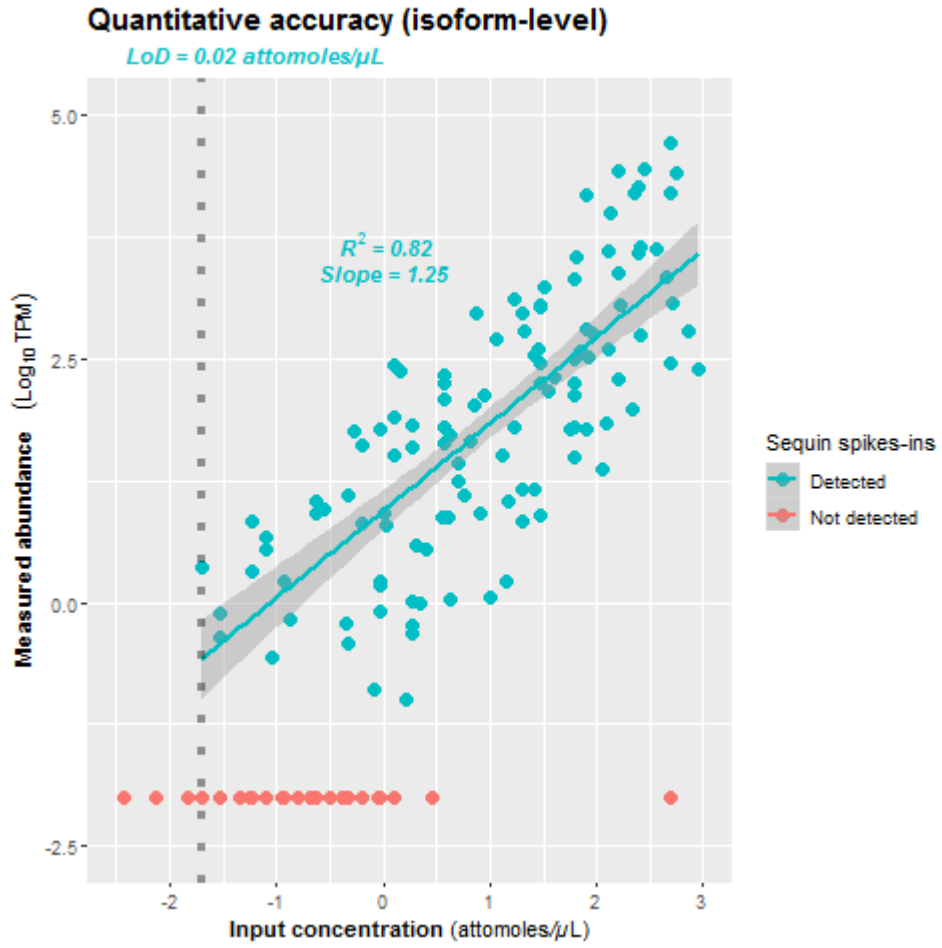


Figure 5-6 Detection of Sequin Controls.

Alignment and quantification of the internal sequin controls was carried out alongside heart tissue RNA. A strong correlation was seen between measured abundance (transcripts per million TPM) and input concentration (R^2 = 0.82).

Comparison of the putative novel lncRNA transcripts identified by Illumina RNA-Seq and Nanopore RNA-Seq identified 39 transcripts from 35 unique genes (four transcripts represented different isoforms) that shared complete intron chains (gffcompare class code "=", illustrated in figure 5-1). Of these, 20 were intergenic (gffcompare class code 'u'), 13

were intronic ('i'), five were novel antisense ('x') and one contained a reference gene within its intron ('y', class codes illustrated in Figure 5-1).

As the pipeline used the GENCODE v.29 annotation, confirmation as to whether the 39 putative novel lncRNAs were indeed novel was sought by searching the updated version of GENCODE (v.32), and the lncRNA databases, FANTOM CAT http://fantom.gsc.riken.jp/cat/?fd=source_data and NONCODE v.5. Of the 39 potentially novel lncRNAs, 12 transcripts (from 10 unique lncRNA genes) had subsequently been reported in GENCODE v.32, 21 transcripts (from 19 unique lncRNA genes) were reported in FANTOM CAT and 15 transcripts (from 12 unique lncRNA genes) were reported in NONCODE v.5 (28 unique lncRNAs in total).

The finding that 28 lncRNAs were now annotated in public databases confirmed the validity of our discovery pipeline, However, this left 10 remaining novel lncRNAs that were not previously described in any dataset (Figure 5-5C, Appendix C). Moreover, 1 of these was differentially expressed between pre- and post-ischaemia time points (padj. = 7.27 x 10^{-13}, fold change = 1.35, Figure 5-7).

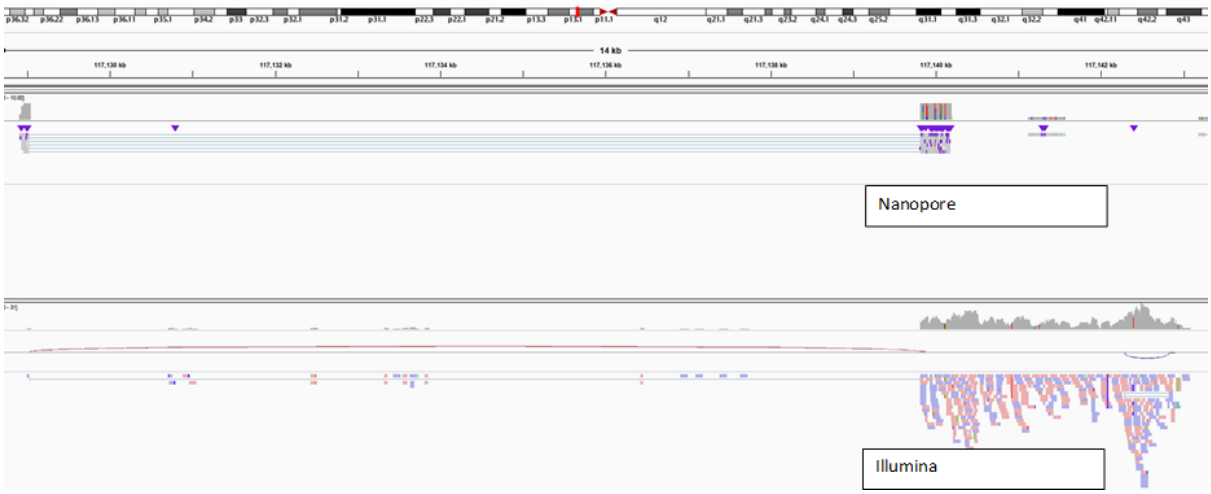


Figure 5-7 A screenshot from IGV showing RNA sequencing reads of the novel transcript MSTRG.10265.1 from a representative sample.

5.3.4 Evolutionary Conservation of lncRNAs

Previous reports have observed that annotated lncRNAs have lower primary sequence conservation compared to mRNAs [65, 106, 107]. To test whether the novel lncRNAs showed similar profiles of conservation to annotated lncRNAs rather than protein coding transcripts, averaged, pre-computed, per-base evolutionary conservation scores for each exon were compared. The novel lncRNAs showed lower primary conservation and had a similar profile to annotated lncRNAs compared with protein coding genes suggesting they are likely to be non-coding transcripts (Figure 5-8).

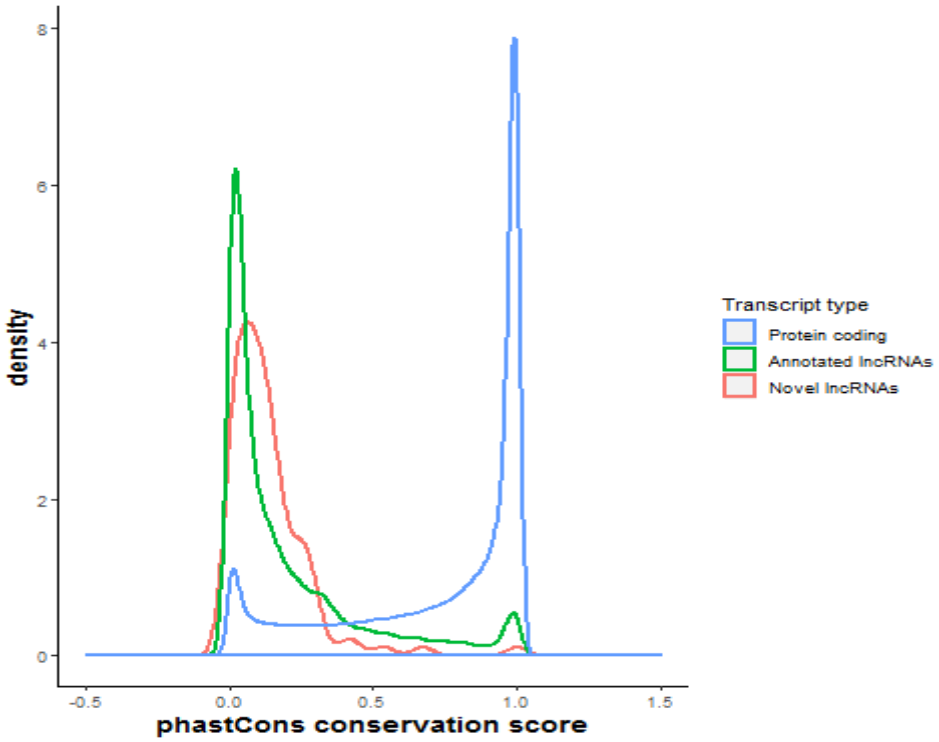


Figure 5-8 Geometric density plots showing the frequency of phastCons conservation scores for 20 mammalian species averaged base-wise for each exon.

Exons from protein-coding transcripts (expressed in this study) are shown in blue, the exons from annotated lncRNAs (expressed in this study) are shown in green and identified novel lncRNAs are shown in red. The phastCons score estimates the probability that a nucleotide is conserved; the closer the score is to 1, the more conserved the base.

5.3.5 Overlap of Putative Novel lncRNAs with Regulatory Elements in the Genome

To explore whether the novel lncRNAs had potential to regulate gene expression through known mechanisms, novel lncRNAs overlapping known regulatory elements and genome wide association study (GWAS) SNPs associated with cardiovascular traits were investigated. Promoters and enhancers were downloaded from Ensembl biomart with 'left ventricle' selected as the tissue filter. Of the 11 novel lncRNAs, 8 overlapped enhancers, 2 overlapped promoters with a further 7 overlapping promoter flanks. One novel (MSTRG.72507.1 which was later identified as AC100756.4 when comparing to a later version of GENCODE overlapped the SNP rs937741 which was associated with Blood pressure (smoking interaction) from the GWAS catalogue (Appendix C).

5.3.6 Overlap of novel lncRNAs with cis-eQTLs

eQTL SNPs are associated with changes in the expression level of a gene and are mostly located in non-coding regions [411]. Because an eQTL effect may be mediated by a lncRNA, the overlap between each of the novel lncRNAs and known eQTL SNPs in left ventricle was analysed. Of the 11 novel lncRNAs, the two exons of *MSTRG.8333.38* overlapped five cis-eQTL eSNPs (rs259352, rs259354, rs3820345, rs10783001 and rs35070110) associated with expression of the gene RWD Domain-Containing Sumoylation Enhancer (*RWDD3*) (GTEx SNP-gene associations $p < 1 \times 10^{-06}$. This lncRNA is situated immediately downstream of *RWDD3* and expression of *MSTRG.8333.38* - *RWDD3* was weakly correlated (Figure 5-9 Spearman -0.2 and p-value 0.002) providing a potential mechanistic link between these eSNPs and *RWDD3*.

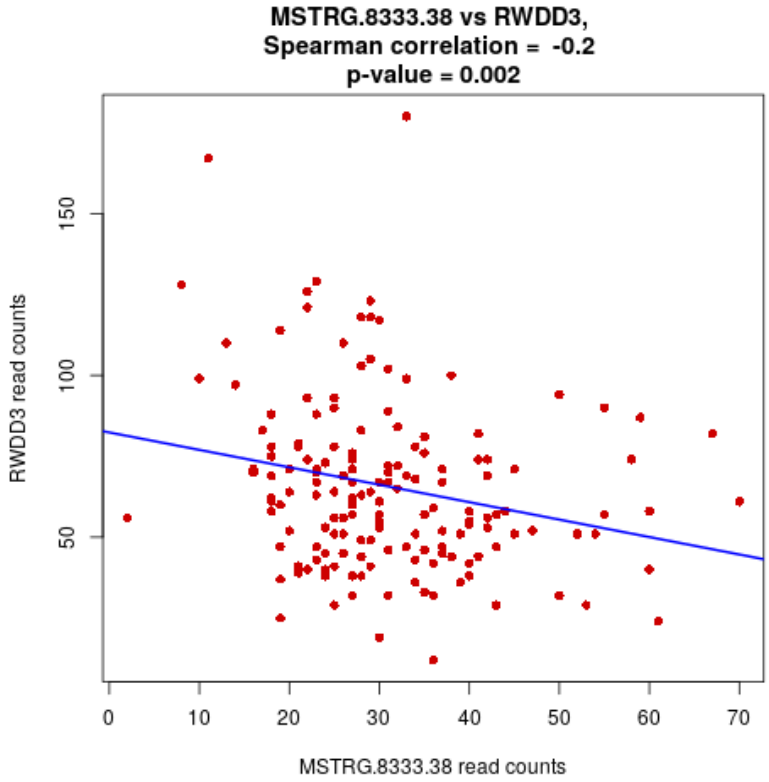


Figure 5-9 Spearman correlation of novel lncRNA-mRNA pair MSTRG.8333.38 - RWDD3. *MSTRG.8333.38 also overlapped cis-eQTLs associated with RWDD.*

The first exon of the two-exon novel MSTRG.10265.1 also overlapped an eQTL (chr1_117128873_A_G_b38) which is associated with V-Set Domain Containing T Cell Activation Inhibitor 1 (VTCN1) in left ventricle tissue. However, VTCN1 was not robustly expressed in this data set and was filtered out.

5.3.7 Identifying gene networks associated with ischemia

WGCNA interrogated all genes from the dataset and identified 18 modules of highly correlated genes with similar expression profiles across patients. Of these, two were strongly associated with ischemia (absolute correlation coefficient >0.6, p-value < $1x\ 10^{-17}$, Figure 5-10).

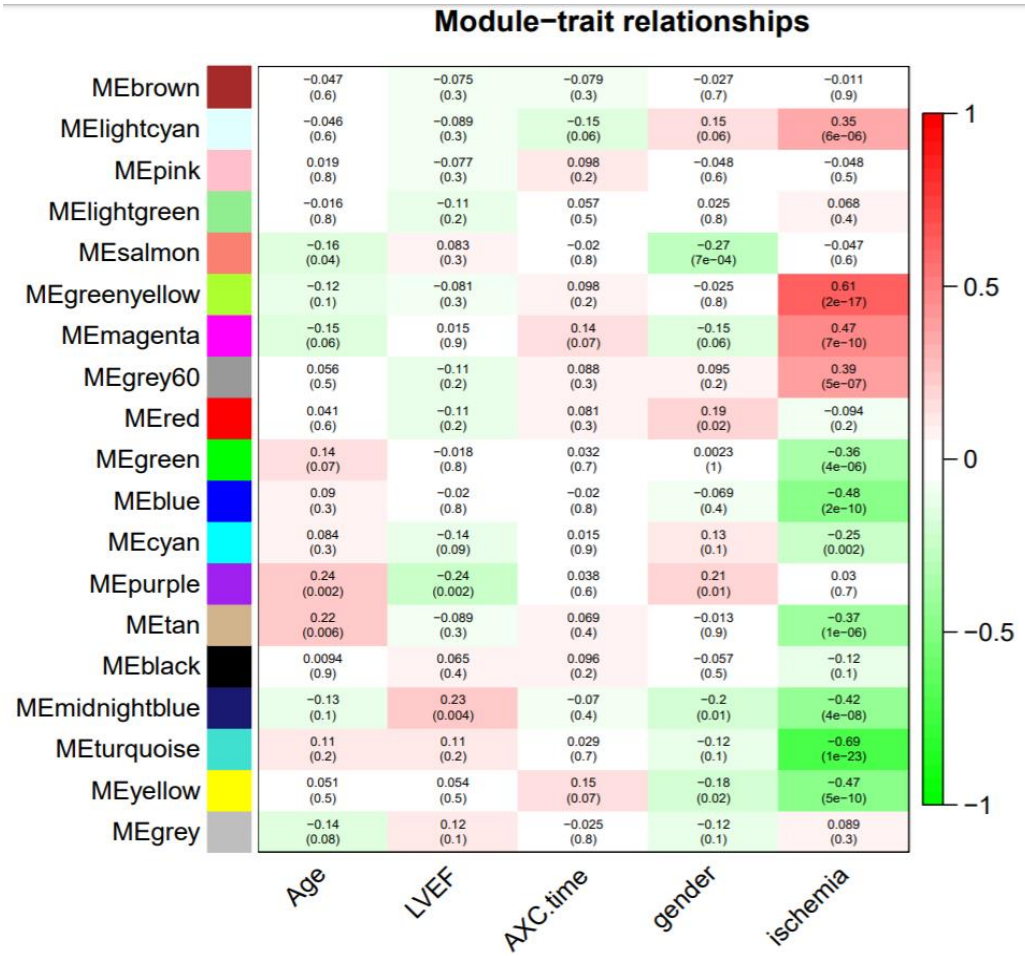


Figure 5-10 Module-trait relationships predicted by WGCNA.

Each cell shows the correlation and p-value between the module eigengene (row, with each module arbitrarily assigned a colour) and the trait (column). The table is color-coded by correlation according to the colour legend. ME: module eigengene.

Expression of genes in Module 1 (Turquoise Module) decreased from pre- to post-ischemia (Pearson correlation coefficient (PCC) =-0.69, p=1 x 10^{-23}). IPA analysis showed that the genes in this module (n = 2038) were enriched for pathways associated with cell death, apoptosis, and necrosis (Table 5-2). Module 1 also included one novel lncRNA which was moderately associated with ischemia (MSTRG.10265.1, PCC=-0.37, p= 2.46E-06)). Also, in Module 1, 39 annotated lncRNAs were moderately associated with ischemia (PCC > |0.4|, all padj<0.001). Three of these lncRNAs had module membership correlation values > 0.7, suggesting they may act as network hubs (genes with potential to co-regulate multiple genes in the Module and coordinate the early response to ischemia): AC005523.2, AF111167.2 and

CTBP1-DT. Module 1 mRNAs with the highest module membership and potential to serve as network hubs were Oxoglutarate Dehydrogenase (*OGDH*), Kinesin-like protein *(KIF1C)*, Mitofusion 2 (*MFN2*), and *MRPS27 (*module memberships of =0.95, 0.93, 0.93 and 0.92 respectively).

In contrast to Module 1, on average, the expression of genes in Module 2 (Green Yellow Module) increased from pre- to post-ischemia (Pearson Correlation coefficient (PCC) module eigengene with ischemia=0.67, $p=2 \times 10^{-17}$). Genes in Module 2 (n=250) were predicted to activate pathways involved in angiogenesis and vascular development in the cardiovascular system and proliferation of white blood cells and immune cells in response to ischemia (Table 5-2).

Potential network hubs (transcripts with the highest module membership with correlation values >0.7) included mRNA transcripts overwhelmingly associated with vasculogenesis and angiogenesis, consistent with pathway analysis. These include TEK Receptor Tyrosine Kinase (*TEK*, a member of the tyrosine kinase Tie2 family), ETS Transcription Factor (*ELK3)* Vascular Endothelium Cadherin 5 (*CDH5)*, and Eprhin B2.

Table 5-2 The top five disease or functions predicted by IPA (sorted by z-score) associated with the two modules most associated with ischemia (WGCNA).

Categories	Module	Diseases or Functions Annotation	p-value	Predicted Activation State	Activation z-score
Organismal Survival	**1**	Morbidity or mortality	9.2E-08	Increased	12.7
Organismal Survival	**1**	Organismal death	5.58E-08	Increased	12.7
Organismal Injury and Abnormalities	**1**	Organ Degeneration	4.3E-06	Increased	4.3
Cell Death and Survival	**1**	Apoptosis	9.03E-08	Increased	4.1
Cell Death and Survival	**1**	Necrosis	6.47E-09	Increased	3.4
Cardiovascular System Development and Function	**2**	Development of vasculature	1.02E-13	Increased	4.4
Cardiovascular System Development and Function, Organismal Development	**2**	Angiogenesis	6.31E-13	Increased	4.4
Cardiovascular System Development and Function, Organismal Development	**2**	Vasculogenesis	1.23E-12	Increased	4.3
Hematological System Development and Function, Lymphoid Tissue Structure and Development, Tissue Morphology	**2**	Quantity of lymphocytes	1.42E-06	Increased	3.7
Cellular Development, Cellular Growth and Proliferation,Hematological System Development and Function,Hematopoiesis,Lymphoid Tissue Structure and Development, Tissue Development	**2**	Hematopoiesis of mononuclear leukocytes	9.75E-07	Increased	3.4

Among the annotated lncRNAs in Module 2, five lncRNAs were moderately associated with ischemia (PCC > 0.4). Of these, *Prostate Cancer Associated Transcript 19* (*PCAT19)* and *AC093278.2* had module membership correlation values > 0.7, suggesting they may also serve as network hubs. There were no novel lncRNAs present in Module 2.

5.3.8 Overlap of annotated lncRNAs with cis-eQTLs

Investigation of lncRNAs overlapping cis-eQTLs was carried out to identify possible functional mechanisms. A total of 47 annotated lncRNAs overlapped 52 cis-eQTL snps (GTEx SNP-gene associations p-value < 10E-07).

For Module 1, 12 annotated lncRNAs overlapped significant cis-eQTL SNPs, of which 3 were differentially expressed in response to ischemia. These cis-eQTL eGene associations were: AC005523.2- Fem-1 homolog A (*FEM1A),* AC011476.3 - retinol dehydrogenase 13 (*RDH13*) and AC012313.1 - Zinc finger protein 84 (*ZNF584*). Expression of AC005523.2-*FEM1A* and AC011476.3 - *RDH13* was highly correlated (Spearman correlation 0.87, 0.81 and p-value 6.13 e-52, 2.63 e-38 respectively, Figure 5-11). However, the correlation of AC012313.1 - *ZNF584* was not significant. None of the annotated lncRNAs in Module 2 overlapped any eQTL eSNPs.

5.3.9 Subcellular localisation of ischaemia associated lncRNAs with RNA Scope

RNAscope – an *in-situ* hybridisation technique for RNA - was used to identify the subcellular localisation of two annotated lncRNAs as well as a novel lncRNA (VASH1-AS1, PCAT19 and MSTRG. 10265.1 Figure 5-12). These specific lncRNAs were chosen as they were significantly associated with the significant ischaemia-associated modules from WGCNA. They were either one of the highest differentially expressed (VASH1-AS1, turquoise module) or suggested to be a ‘hub’ gene and highly connected to other genes in that module (PCAT19,

green yellow module). All three transcripts appear to localise both within the nucleus and the cytoplasm.

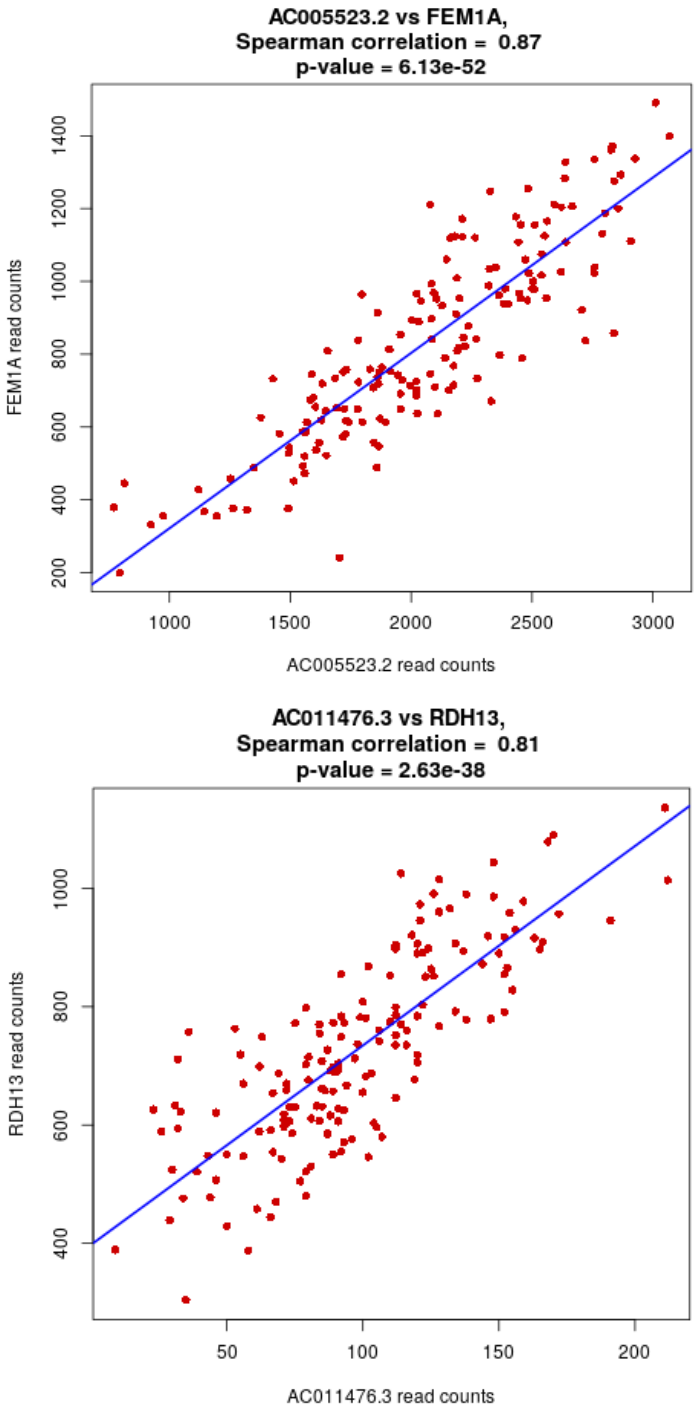
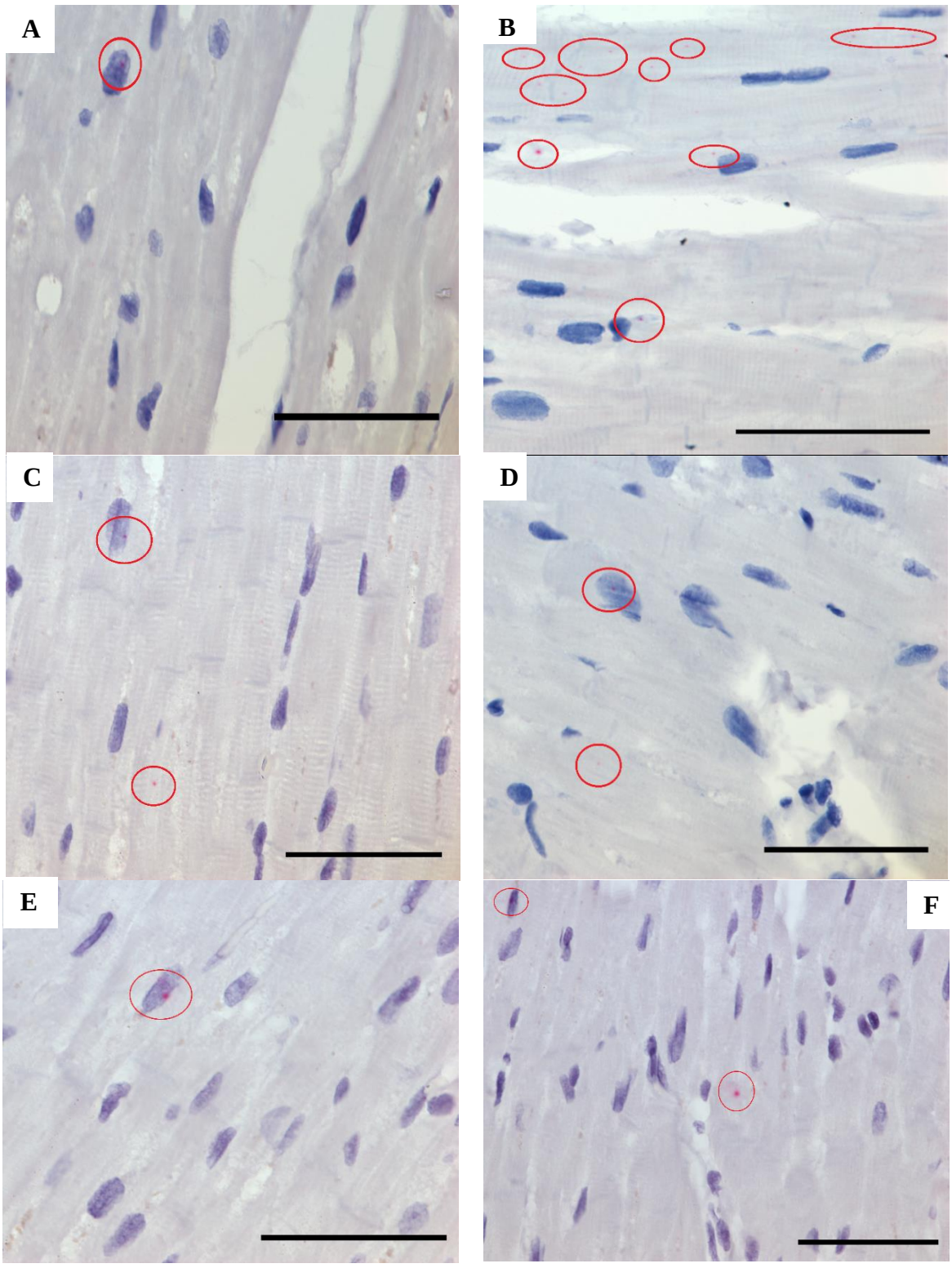


Figure 5-11 Spearman correlation of lncRNA-mRNA pairsAC005523.2- FEM1A (top panel) and AC011476.3 - RDH13 (bottom panel) where the lncRNA also overlapped cis-eQTLs associated with the mRNA.


A
B
C
D
E
F

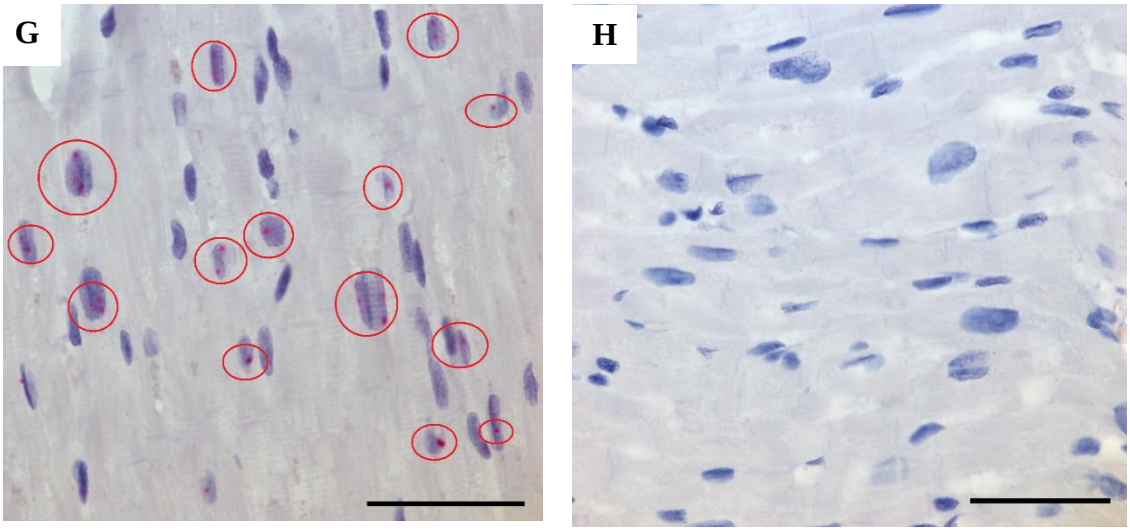


Figure 5-12 RNA Scope showing VASH1-AS1, PCAT19 and the novel MSTRG.10265.1 expression in cardiomyocytes.

The dark blue haematoxylin staining indicates the cell nucleus. Red circles indicate the location of the red RNA-Scope probes. ***A)*** *VASH1-AS1 probe nuclear localisation at x63 magnification,* ***B)*** *VASH1-AS1 probes cytoplasmic localisation at 63x magnification* ***C & D)*** *Shows PCAT19 probe nuclear and cytoplasmic localisation at x63 magnification* ***E & F)*** *Shows novel (MSTRG.10265.1) probe nuclear and cytoplasmic localisation at x63 magnification* ***G)*** *NEAT1 lncRNA positive control with an exclusively nuclear location* ***H)*** *Negative control Scale bar = 50µm*

5.4 Discussion

Before the *Saddic* paper [175], there were no studies that had examined the profile of lncRNAs of acute ischemia in the human heart. *Saddic et al* described the first lncRNA profile of acute ischemia in the human heart. This study builds on the *Saddic* analysis with addition of novel lncRNAs and whole genome correlation network analysis to examine the association of lncRNAs with the ischaemia heart and their potentially regulatory roles.

A total of 11 novel lncRNAs in human left ventricle were identified using a strategy of both short- and long-read sequencing. Of these, the expression of two novel lncRNAs, along with 270 annotated lncRNAs and 2,446 mRNAs, were altered in response to ischemia. Co-expression analysis of these lncRNAs and mRNAs identified two networks of genes that may promote angiogenesis, neovascularisation and cardiomyocyte cell death and form part of the early response to myocardial ischemia.

Accumulation of atherosclerotic plaque in the coronary arteries can lead to a prolonged reduction in blood flow and oxygen to the heart. At the cellular level, this can cause metabolism to switch from aerobic oxidative phosphorylation to anaerobic glycolysis, leading to increasing intracellular acidosis and radical oxygen species levels. High levels of intracellular calcium activate proteases that degrade the cytoskeleton. Cell death by necrosis, apoptosis, and autophagic mechanisms are the end result [412]. With cell death, cytokines are released to initiate an inflammatory response. Neutrophils, macrophages, and leukocytes infiltrate the area and cell adhesion molecules are secreted on the surface of endothelial cells and leukocytes to facilitate infarct healing, and subsequent ventricular remodelling. Data from the current study suggests that lncRNAs may help coordinate these immediate metabolic, inflammatory, and microscopic myocardial remodelling responses to ischemic injury, and may involve previously unannotated lncRNAs.

As Saddic *et al* had previously published analysis of mRNAs and annotated lncRNAs as markers for ischaemia, this analysis focussed more on novel lncRNAs as potential ischaemic markers. However, the WGCNA analysis also adds to the annotated genes with a novel approach to looking at gene network co-expression. The two novel lncRNAs identified here could serve as potential markers for myocardial ischemia if they are found to be exported into the circulation. Established biomarkers such as the troponins and creatine kinase are used for diagnosis and risk stratification of patients with chest pain and suspected acute coronary syndrome (ACS). However, they are unable to detect myocardial ischemia in the absence of necrosis and there remains a need for markers of subclinical atherosclerosis and ischemia [413]. My data suggest that lncRNAs along with mRNAs respond to ischemic changes in the left ventricle suggesting they may represent novel therapeutic targets or candidate biomarkers for early myocardial dysfunction.

WGCNA [374] collapses co-expressed groups or networks of genes into modules in an unsupervised manner, thereby minimising the huge multiple testing problem that plagues

whole transcriptome analysis. It can be used for finding clusters (modules) of highly correlated genes, relating modules to phenotypic traits and for calculating module membership and gene significance to identify important hub genes which may be strong drivers of that module. One of the novel lncRNAs (MSTRG.10265.1) belonged to a gene module (Module 1) of highly correlated, and potentially co-regulated, genes that included 39 annotated lncRNAs that were also correlated with ischemia. This module was associated with activation of cell death potentially through pathways involved in energy metabolism. Among the 39 annotated lncRNAs, three had high module membership which is indicative of highly interconnected 'hubs' (AC005523.2, AF111167.2, CTBP1-DT). While the mechanisms by which these lncRNAs might influence gene expression are unknown, it was observed that AC005523.2 lies antisense to FEM1A, which was in the same Module and also overlaps a *cis*-eQTL SNP associated with FEM1A suggesting AC005523.2 may be acting as a *cis*-regulator of FEM1A. FEM1A is localised within mitochondria of cardiac muscle and is increased in mouse hearts after myocardial infarction, [414] and may regulate apoptosis [415]. The other two lncRNAs may also regulate nearby genes: *AAF111167.2* is antisense to, and overlaps the active promoter of, *JDP2*, a transcription factor associated with maladaptive cardiac remodelling [416], and CTBP1-DT is antisense to CTBP1 which is involved in cell proliferation [417]. As well as AC005523.2-FEM1A, two other differentially expressed lncRNAs had significant cis-eQTL eGene associations, these were: AC011476.3 - retinol dehydrogenase 13 (*RDH13*) and AC012313.1 - Zinc finger protein 84 (*ZNF584).* *RDH13* is localized in the mitochondria and may function to protect the mitochondria against oxidative stress [418]. *ZNF584* has the highest level of expression in the left ventricle [419] and is one of the *KRAB-ZNF* transcriptional regulatory family of proteins which are transcriptional repressors. Despite their abundance in the genome little is known about their gene targets and biological functions. There is however, emerging evidence that they are involved in apoptosis, energy metabolism cell proliferation and differentiation [420].

In Module 1, mRNAs with the highest module membership were i) *OGDH*, which forms part of the 2-oxoglutarate dehydrogenase complex to catalyse conversion of 2-oxoglutarate (alpha-ketoglutarate) to succinyl-CoA and CO_2 during the Krebs cycle and is an important mitochondrial redox sensor [421], ii) *KIF1C*, which regulates actin-rich adhesion structures (podosomes) that remodel the extracellular matrix of cells including macrophages and synthetic vascular smooth muscle cells [422], iii) *MFN2*, which encodes a mitochondrial membrane protein that contributes to the maintenance and operation of the mitochondrial network, regulation of vascular smooth muscle cell proliferation and is pivotal during recovery from ischemia/reperfusion injury [423] and iv) MRPS27, which is required for the translation of mitochondrially encoded proteins and mitochondrial protein synthesis [424].

The second gene module strongly associated with ischemia was associated with angiogenesis and inflammation which have been demonstrated to be linked with atherosclerosis and ischaemia [425-427]. Potential hub genes (genes with the highest module membership) were overwhelmingly associated with vasculogenesis and angiogenesis. *TEK* mediates embryonic vascular development [428, 429] whereas *ELK3* regulates angiogenesis through the control of vascular endothelial growth factor (*VEGF*)[430], *CDH5*, is involved in vascular permeability and leukocyte transmigration and angiogenic sprouting [431] and Eprhin B2, present in arteries, and is involved in the angiogenesis process [432].

Among the annotated lncRNAs, 5 (PCAT19, AC093278.2, CARMN, AC005550.2, AC007743.1) were associated with ischaemia and were strongly correlated with other genes within the module suggesting they may play a significant role in the immediate response to ischaemia in the left ventricle. *PCAT19* has been shown to negatively regulate p53 in lung cancer [433] whist p53 has been shown to negatively regulate ischemia-induced angiogenesis [434]. While the function of *AC093278.2* has not been described, it lies 2.7Mb upstream from an enhancer that is active in the left ventricle and overlaps the 3'UTR of *ZNF366*. Among the remaining lncRNAs, *CARMN* is associated with a left ventricle 'super' enhancer

and has been shown to control cardiac specification and differentiation [435], *AC005550.2* is antisense and overlapping to Homeobox protein MOX-2 *MEOX2* (also in Module 1) which regulates cardiac energy metabolism via fatty acid uptake in heart capillary endothelium [436], and *AC007743.1* overlaps an active CCCTC-Binding factor (CTCF site thought to regulate the 3D structure of chromatin) and is antisense to Conserved Protein Domain Family 85A (*CCDC85A* also in Module 2).

In addition to the two lncRNAs associated with ischemia, 9 novel lncRNAs that were not associated with ischemia but were robustly detectable in left ventricle samples were also identified. These included MSTRG.8333.38, which overlapped cis-eQTLs associated with *RWDD3*. *RWDD3* is an enhancer of Small Ubiquitin-like Modifier (SUMO) conjugation which modifies post translation modification of proteins and is involved in heart specific development, metabolism, contractility, and protein quality control [437]. Both *RWDD3* and MSTRG.8333.38 were present in the same WGCNA module.

LncRNAs exert their functions via several mechanisms (reviewed in Section 2.2.4). This study identified several lncRNAs that overlapped *cis*-eQTLs in human left ventricle. One possibility is that this SNP(s) is affecting the secondary structure of the lncRNA and therefore its binding capacity either to chromatin, a transcription factor, a promoter, or the mRNA. However, currently this is speculative. It could be tested with analysis of the lncRNA with either allele (three genotypes).

By identifying the subcellular localisation of the lncRNA (nucleus versus cytoplasm), a biological mechanism can be hypothesised. There were initial technical issues when using RNAscope with heart tissue. Firstly, the kit provided positive control Peptidylprolyl Isomerase B (PPIB) did not appear to work with cardiomyocyte cells and, as the cardiomyocytes appeared to have a brown background stain inherent in the tissue, we had to use the fast red probe instead of the brown probe. From the RNA Scope analysis (Figure 5-

12) the two annotated lncRNAs (VASH1-AS1 and PCAT19) as well as the novel MSTRG.10265.1 appear to be located in both the nucleus and the cytoplasm. As the lncRNAs had high module membership which indicates those genes that have high connectivity with other genes in the module, it is possible that the lncRNAs assert their functions both inside the nucleus and in the cytoplasm by having more than one biological mechanism as has been seen with other lncRNAs [438]. Alternatively, it is possible is that the lncRNAs act in the cytoplasm and the probes identified in the nucleus are in the process of being transported to the cytoplasm and lastly it could be possible that the signal seen in the nucleus are in fact originating from cytoplasm on top of the nucleus. Further interrogation using z-stacks could confirm this. The handful of studies available for PCAT19 propose it may function through interactions with microRNAs, however this still does not rule out either location as miRNAs and Argonaute are also found in both [433, 439, 440]. There are currently no studies on VASH1-AS1. The successful demonstration of RNA Scope for the novel MSTRG.10265.1 is further validation that this lncRNA is a genuine transcript. MSTRG.10265.1 also overlaps an eQTL (chr1_117128873_A_G_b38) which is associated with *VTCN1* in the left ventricle, although this gene did not appear to be robustly expressed in the data and correlations of MSTRG.10265.1 and *VTCN1* expression could not be tested. As *VTCN1* is has been implicated in post-ischaemic cardiac remodelling [441], it could be that this gene is not turned on until a later, post ischaemic response. Additional work would be needed to gain insight into the functional role of these lncRNAs. Such experimental techniques may include RNA-binding protein immunoprecipitation (RIP) assays (which interrogate lncRNA and miRNA binding to Argonaute proteins within the cytoplasm) to test lncRNA/miRNA interaction, Chromatin Isolation by RNA Purification (ChIRP) where high throughput sequencing is used to find regions of the genome that are bound by a specific RNA, or RNA Antisense Purification (RAP) which uses biotinylated probes to capture interacting RNAs, followed by sequencing [442].

The study has several limitations. First, as the median ischaemic time was only 74 minutes, the gene expression changes seen here reflect the immediate response to a relatively short, mild ischaemia and lncRNA changes associated with prolonged ischaemia or a more severe response (e.g., after myocardial infarction) may have been missed. Second, although it is a good model for human myocardial ischaemia, it is possible some of the changes in gene expression may also be due to a cardioplegic cold response, rather than ischaemia *per se*. Finally, this data is largely descriptive and further work would be needed to confirm the role of these genes, particularly for potential hub genes, in the early response to myocardial ischaemia. Also, whether any of the transcripts can be detected in circulation and may have potential as biomarkers is unknown.

5.5 Conclusion.

Expression of 2,446 mRNAs, 270 annotated lncRNAs and 2 novel lncRNAs were associated with the early response to myocardial ischemia in human left ventricle (padj <0.001, absolute fold change >1.2). An additional 9 novel lncRNAs were identified in human left ventricle that were not altered by ischaemia. In addition to mRNAs, several lncRNAs appear to act as hub genes, potentially coordinating the expression of several genes within the same module. These findings suggest that both mRNAs and lncRNAs are altered in association with an early ischaemic stress response and may therefore have potential as therapeutic targets or circulating biomarkers for myocardial ischaemia.

Chapter 6
Establishing an RNA-Seq protocol to investigate cell-free RNA in plasma from heart patients and healthy volunteers

6.1 Introduction

Human biofluids contain a collection of cell-free nucleic acids. Historically, studies in biofluids have focussed either on circulating DNA of the foetus, to determine foetal sex, aneuploidies, micro-deletions and the detection of paternally inherited monogenic disorders [443], or on circulating tumour DNA, to characterise mutations and assess response to treatment [444]. More recently, there has been a growing body of evidence linking RNAs in biofluids with various diseases, including lncRNAs, circRNAs and miRNAs, making them excellent candidates for biomarkers [199, 445, 446]

Advances in RNA-Seq now allow us to study the transcriptome in 'liquid biopsies' such as plasma and other bodily fluids, to identify new diagnostic and prognostic markers for various diseases in a non-invasive manner [447-449] Circulating RNAs found in plasma include mRNAs, lncRNAs, circRNAs, miRNAs as well as other RNA species such as tRNAs, snoRNAs and piRNAs [450, 451]. They can be present in membrane-bound extracellular vesicles, such as exosomes, micro-vesicles and apoptotic bodies, they may be bound to ribonucleoprotein (RNP) complexes or high-density lipoproteins [HDLs], or circulating freely [452, 453]. Extracellular vesicles are highly abundant in biological fluids and are secreted by most cell types to transfer proteins, lipids and RNA between cells for the purpose of intercellular communication and signalling [454].

However, RNA-Seq of biofluids is technically demanding. The low input amounts of RNA, partial degradation of transcripts due to ribonucleases in blood that are not protected by protein complexes or microvesicles and the use of archival specimens of varying ages, result

in challenging technical hurdles. These reasons may explain why most RNA-Seq studies in biofluids have sequenced small-RNA or microarrays or used targeted strategies such as RT-qPCR. To date, only a handful of studies have carried out total RNA-Seq in bodily fluids [451-453, 455]. Despite the challenges, there are benefits to total RNA-Seq over microarrays and targeted strategies. These include an unbiased, genome-wide approach without *a priori* hypotheses, allowing capture of novel transcripts and isoforms. With the falling costs of RNA Sequencing, allowing samples to be sequenced to a much greater 'depth' (i.e. many more reads), and the development of specialist RNA library kits, enabling sequencing with smaller amounts of input RNA, we are in an exciting position to achieve a complete account of the circulating, cell-free transcriptome.

This chapter aimed to establish a method for detecting mRNAs, lncRNAs and circRNAs in archived plasma from heart patients and healthy controls. The bioinformatics analysis pipeline previously developed for identifying lncRNAs with altered expression in ischaemic cardiac tissue could be applied to RNA-Seq data acquired from plasma. Sequencing of plasma was carried out from 31 Healthy volunteers (HVOLs), 31 patients with unstable angina or myocardial infarction who remained free of heart failure for at least three years ('CDCS heart failure negative') and 30 patients with either unstable angina or myocardial infarction who were diagnosed with heart failure within three years ('CDCS heart failure positive').

Specific Objectives were to:

1. Determine the quality and quantity of RNA able to be extracted from human plasma (pilot 1)
2. Remove DNA contamination from RNA extracted from human plasma (pilot 2)
3. To ascertain if any mRNA/lncRNAs and heart-related mRNAs/lncRNAs could be detected when sequencing at a greater depth using the Illumina HiSeq (pilot 2a)
4. Determine the minimum volume of starting plasma required (pilot 3)

5. Assess the background level of RNA contamination present in the kit reagents by performing an RNA extraction on a no-template control (pilot 3)
6. Test plasma samples from a heart patient and a healthy control that have been stored for >10 years (pilot 4)
7. Test inclusion of artificial spike-in controls (Sequins, pilot 4).

6.2 Overview of research design

The protocols for recruitment of patient and healthy volunteer samples and the methods for RNA extraction, addition of artificial spike-in controls (Sequins) and library preparation are described in Chapter 2. Additions to, and deviations from, these protocols are summarised for each of the four pilot studies in Table 6-1. For the laboratory volunteer plasma consented blood was obtained in accordance with CHL, Endolab policy manual dated Sept 2020, section 12.3.13. RNA extraction from frozen plasma samples was carried out using the Norgen Plasma/Serum RNA purification Maxi kit (Thorold, Canada). Each pilot included a positive control RNA sample from human brain, which was provided with the Takara Bio SMARTer Stranded Total RNA-Seq Kit v2 (Pico Input Mammalian, Shiga, Japan). All sequencing was performed on an Illumina MiSeq sequencer using the Micro Kit v2, San Diego, California, U.S. except pilot 2a which used the Illumina HiSeq to sequence the same library from pilot2 but at a greater depth.

6.2.1 Bioinformatic Analysis

All bioinformatics analyses were carried out using the in-house pipeline as described in Section 4.1 At the time of running the plasma experiment, the GENCODE version had been updated to v33 so this newer version was used. Briefly, the raw FastQ files were trimmed using Trimmomatic software [379] and FastQC (Bioinformatics, Babraham Institute, Cambridge) was used to check that reads were of high quality. Reads were then mapped using

STAR2 splice aware software [268]. Transcripts were assembled with Stringtie [388] and quantified with Salmon [395].

Table 6-1 Overview of the samples used and the differences from the final protocol for the four pilot studies.

	Samples	RNA extraction	Artificial RNA spike-ins (Sequins)	Library preparation (Takara Bio SMARTer Stranded Total RNA-Seq Kit v2 (Pico Input Mammalian, Shiga, Japan).
Pilot 1	• CDCS patient plasma1 (5 mL) • HVOL plasma 1 (5 mL) • Positive control RNA	DNase step omitted to minimise loss of RNA	-	Positive control spiked into library pool at 10% v/v
Pilot 2	• HVOL plasma 2 (4.5 mL) • Positive control RNA	-	-	Positive control spiked into library pool at 15% v/v
Pilot 2a	• HVOL plasma 2 deep sequencing on HiSeq			Same library as pilot 2
Pilot 3	• Lab volunteer 1 (4 mL) • Lab volunteer 1 (3 mL) • Lab volunteer 1 (2 mL) • No template control (Nuclease free water) • Positive control RNA	-	-	Samples and positive control added to library pool at equimolar concentrations
Pilot 4	• CDCS patient plasma 2 (4 mL) • HVOL plasma 3 (5 mL) • Positive control RNA	-	1% Sequins added to each sample (calculation shown in Appendix D-1)	Samples and positive control added to library pool at equimolar concentrations

Sequencing metrics and mapping statistics from the BAM file were assessed with RSeQC [384], and visualised with MultiQC [456].

For pilot 4, which included artificial RNA spike-ins (Sequins), the Sequins reference sequence was added to the human reference prior to read alignment. The percentage of reads aligning to the Sequin reference was calculated by aligning the FastQ reads to the Human genome and Sequins reference sequences combined, as well as the Sequins reference alone. The BAM file was analysed using Samtools flagstat (http://www.htslib.org/doc/samtools-flagstat.html).

6.3 Results

6.3.1 Pilot 1 Determining the quality and quantity of RNA able to be extracted from human plasma

To test if enough RNA of sufficient quality could be extracted from archived patient and healthy volunteer plasma, pilot 1 included a sample from each of the CDCS and HVOL cohorts (Table 6-2).

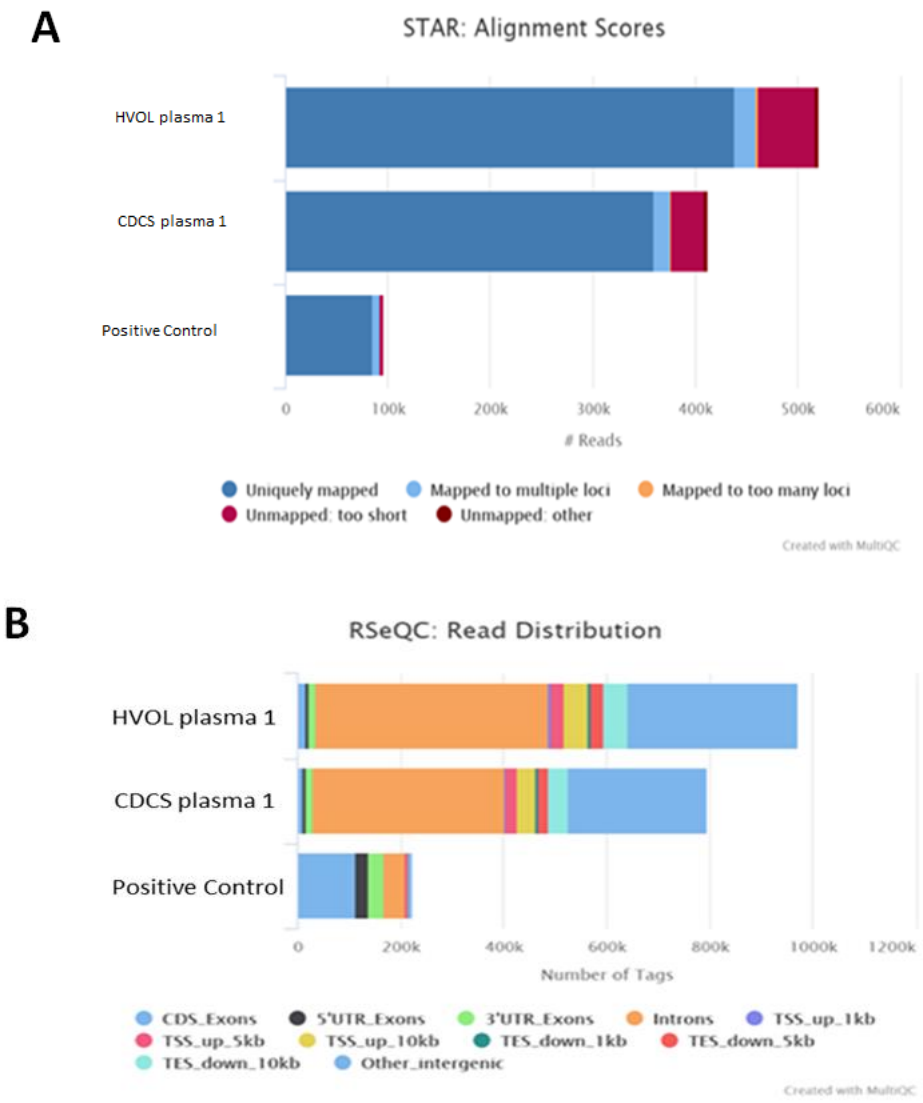
Table 6-2 Sample information for pilot 1

Sample ID	Sample information
CDCS plasma 1	CDCS cohort (heart failure negative) 5ml plasma starting volume
HVOL plasma 1	Healthy volunteer 5ml plasma starting volume
Positive control	Brain tissue RNA from the library preparation kit added at the sequencing facility (spiked into library pool at 10% v/v)

The STAR Alignment Scores showed the majority of reads uniquely mapped to the reference genome (~85% for all samples). The positive control was added at 10% v/v compared to the patient and healthy volunteers' plasma and this is reflected in the reduced number of reads aligning (Figure 6-1A). The percentage of reads mapping to coding regions, 5'UTR and 3'UTR for the CDCS and HVOL samples was much lower (3.7% and 3.6% respectively) than for the positive control (75%) (Figure 6-1B). In contrast, in the 'infer experiment' plots, which showed the 'strandedness' of the library, the positive control RNA sample had ~85% of reads aligning to the antisense strand, and the CDCS and HVOL samples had a near 50%-50% distribution of reads on sense/antisense strands (48% and 47%, respectively, Figure 6-1C).

These findings suggested the presence of DNA contamination in the CDCS and HVOL RNA samples. In the read distribution plot (Figure 6-1B), the reads mapped mostly to intronic and intergenic regions whereas an RNA sample would be expected to map predominantly to coding regions (CDS), the 5'UTR and the 3'UTR (as seen for the positive control). Furthermore, in contrast to the RNA positive control, half the reads mapped to the sense

strand and half the reads mapped to the antisense strand (Figure 6-1C). As this library was stranded, the majority of reads from an RNA sample would be expected to come from the antisense strand (as in the positive control). The presence of DNA was likely explained by the deliberate omission of the DNase step from the RNA extraction (in an attempt to maximise the RNA yield).



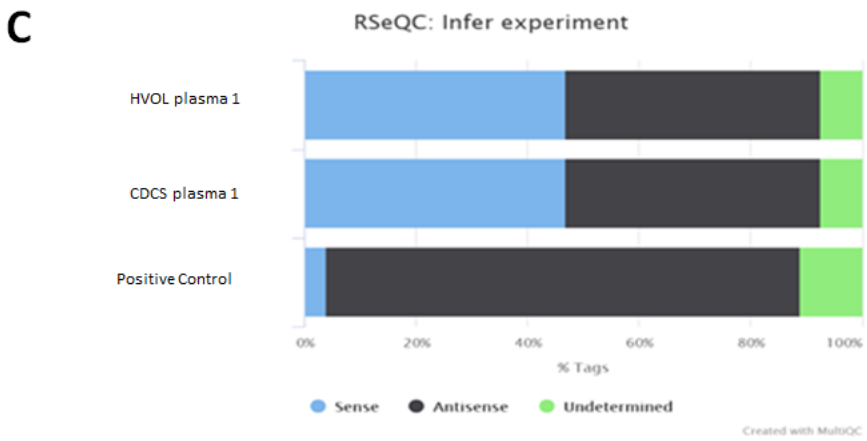


Figure 6-1 Pilot 1: Determining the quality and quantity of RNA able to be extracted from human plasma

A) *Alignment scores from STAR for pilot 1 showing the majority of reads were uniquely mapping to the genome. Each read is categorised as either 'uniquely mapping', 'mapping to multiple loci', 'mapping to too many loci' (by default STAR only outputs reads that map to <=10 loci, others are considered "mapped to too many loci") - unmapped as the proportion of the reads mapping is 'too short' or 'unmapped other'* ***B)*** *The distribution of reads over genomic features. In contrast to the positive control, the plasma samples show very few reads aligning to coding regions (Some reads are assigned to more than one category e.g. "TSS/TES_up_1kb" reads were also assigned to "TSS/TES_up/down_5kb" and "TSS/TES_up/down_10kb" Reads spliced once will be counted as 2 tags, reads spliced twice will be counted as 3 tags, etc. Therefore, "Total Tags" >= "Total Reads". CDS:coding sequence, UTR: untranslated region, TSS: transcription start site, TES: transcription end site)* ***C)*** *The percentage of reads aligning to the sense or antisense strand. For the plasma samples there is an approximate 50%/50% distribution to both strands whereas the majority of reads in the positive control map to the antisense strand. Note the fragment length distribution plots from RSeQC are not presented here or in the other RSeQC plots as they showed expected distributions.*

In summary, pilot 1 failed to show whether RNA could be extracted from archived patient and Healthy volunteer plasma. The next pilot was designed to test including the DNAse step in the RNA extraction protocol.

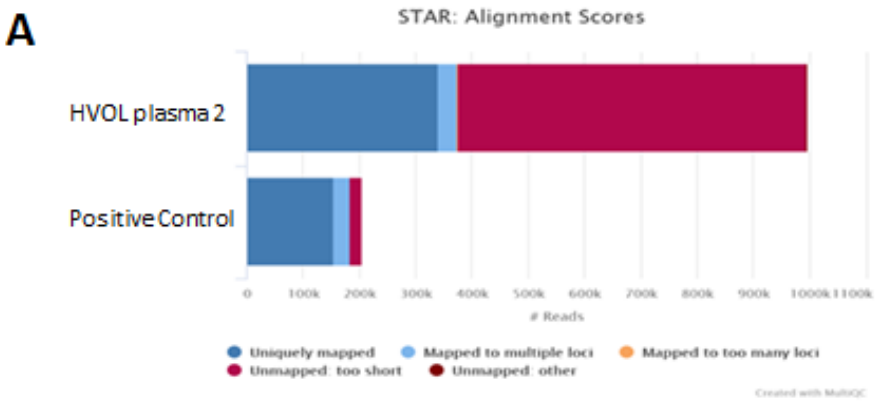
6.3.2 Pilot 2 Removing DNA contamination from RNA extracted from human plasma

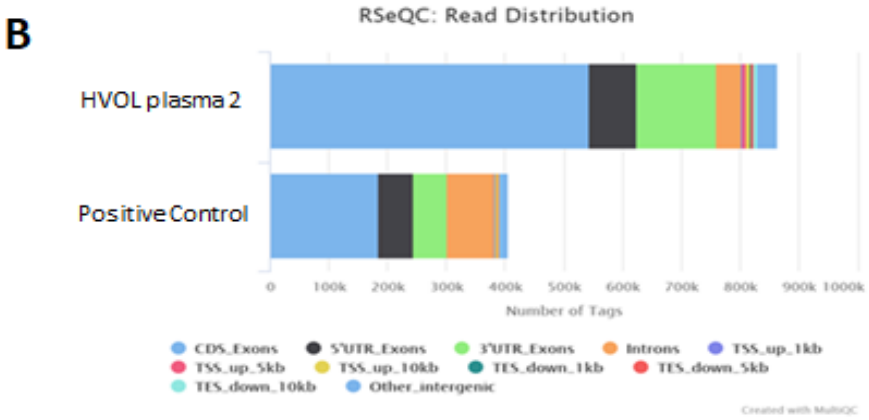
To test whether addition of a DNAse step could remove DNA contamination from the RNA samples while retaining sufficient RNA, pilot 2 tested a single HVOL sample alongside the positive control (Table 6-3).

Table 6-3 Sample information for pilot 2

Sample ID	Sample information
HVOL plasma sample 2	Healthy volunteer 4.5ml starting volume
Dunedin Positive Control	Brain tissue RNA from the library preparation kit added at the sequencing facility spiked into library pool at 15% v/v

MiSeq sequencing generated a total of 1 million reads for the HVOL sample and 200,000 reads for the positive control (reflecting the lower 15% v/v percentage that the positive control was spiked in).





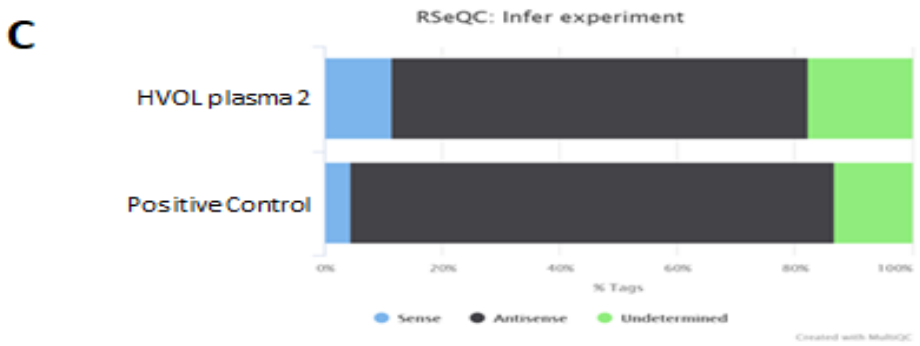


Figure 6-2 Pilot 2: Assessing DNA removal

A) *Alignment scores for pilot 2. A large percentage of reads were classified as 'Unmapped: too short' for the plasma samples* ***B)*** *The read distribution plot showing reads aligning mostly to coding regions which is indicative of RNA not DNA* ***C)*** *Consistent with this, the majority reads aligned to the antisense strand rather than a random alignment to both strands (as was seen in pilot 1).*

Only 33% of reads for the HVOL sample uniquely mapped to the human reference genome, with a large percentage (62%) being classified as 'Unmapped: too short' (Figure 6-2A). In contrast, the positive control had 75% reads mapping uniquely with only 11% being 'Unmapped: too short'. The DNAse step appeared to have been successful as the read distribution and 'infer experiment' plots gave the expected pattern for an RNA sample, namely the majority of reads aligning to coding regions and the majority of reads aligning to one strand (Figures 6-2B and 6-2C). In contrast to the HVOL sample in pilot 1, the pilot 2 HVOL sample closely resembled the positive control, with reads mostly aligning to coding regions rather than across the genome (Figure 6-3). To elucidate why a large percentage (~62%) of reads were classified as 'Unmapped: too short', the developer of the mapper STAR, Dr Alex Dobin (Cold Spring Harbor Laboratory), was contacted. He suggested three potential reasons:

- There was an issue with over trimming
- One of the pairs of the reads (either the forward or reverse read) may have mapping incorrectly.
- The RNA sample was contaminated

Over trimming was ruled out as the trimming software included a minimum length filter (confirmed by the QC plots). The second suggestion (incorrect mapping) was tested by mapping forward and reverse reads separately and again ruled out as the issue remained. Contamination with non-human transcripts was tested by mapping the unmapped reads from STAR (containing all but uniquely mapping reads) using Kraken2 software which assigns taxonomic labels to reads (https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1891-0) and visualises the data using Krona (https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-385). To reduce memory and storage requirements prebuilt bacterial, archaeal, and viral genomes, which contain 5% of sequences from the original databases, were used. For the HVOL sample, the majority of reads that were originally classified as 'Unmapped too short', mapped to bacterial genomes (63%), although there was still 13% mapping to the human reference (Figure 6-4). When several of the reads aligning to the human reference (selected at random) were searched in the human genome using BLAST, they appeared to be from regions that share high sequence similarity and are problematic to map, such as the human leukocyte antigen (HLA) complex (a group of highly related proteins). Among the remaining reads, 21% had 'no hits' and 3% were classified as 'other,' which may reflect that the database used contained only 5% of the original bacterial, archaeal, and viral genome database. For comparison, in the positive control, 96% of unmapped reads were classified as human with only 3% mapping to bacteria.

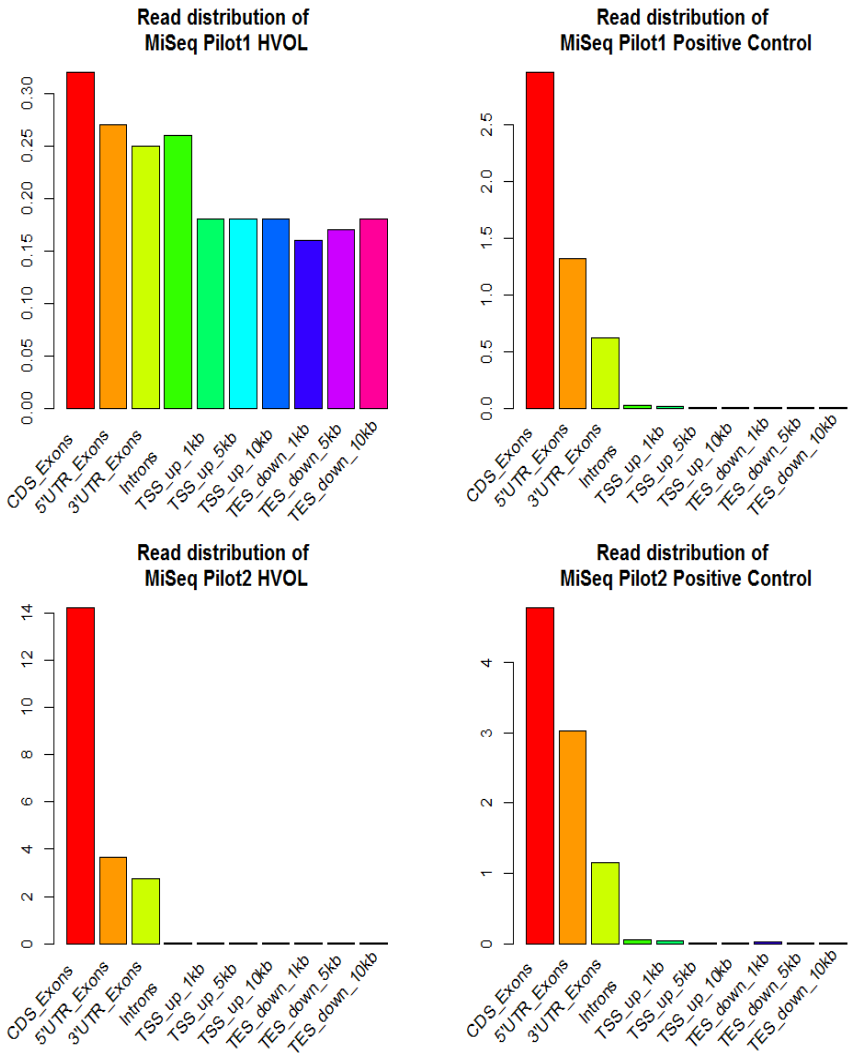
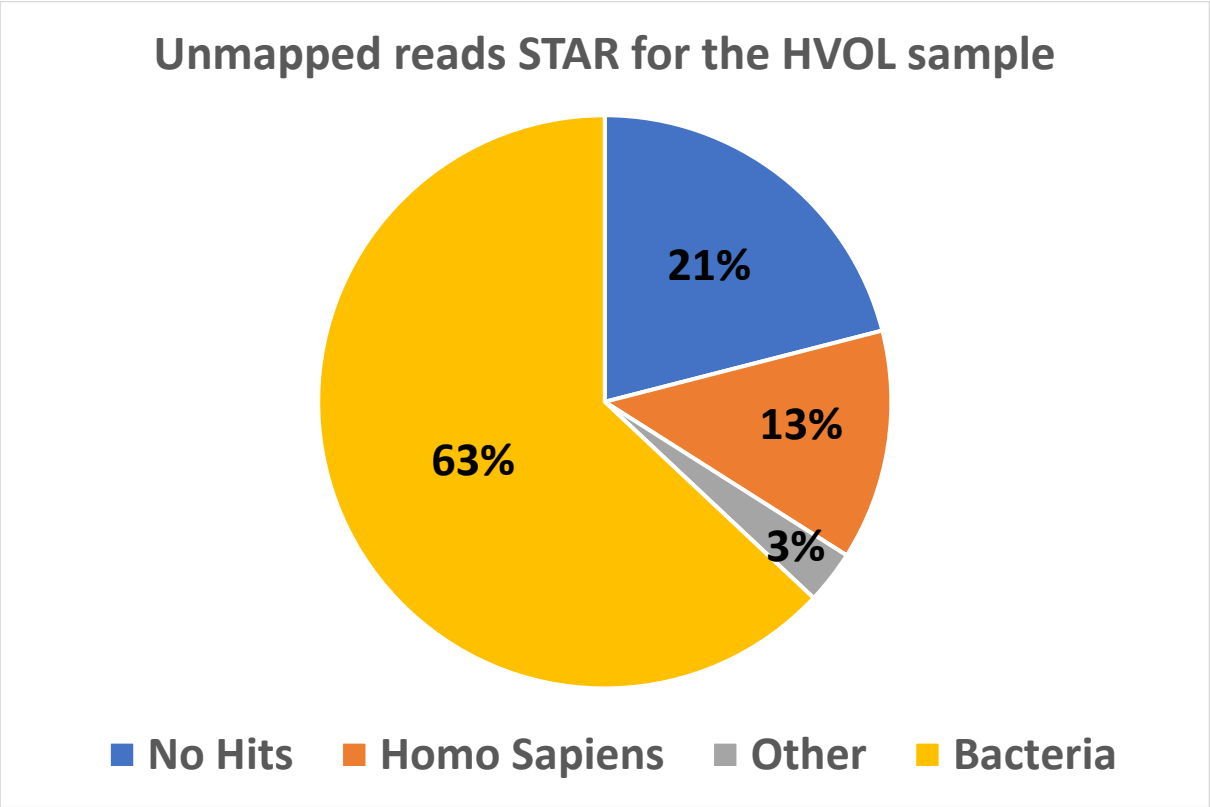


Figure 6-3 Bar plots comparing the read distributions across genomic features

Pilot 1 HVOL sample and brain tissue control (top panel) and pilot 2 HVOL sample and brain tissue control (bottom panel). The x-axis shows the different genomic features; the y-axis shows tags per kb.



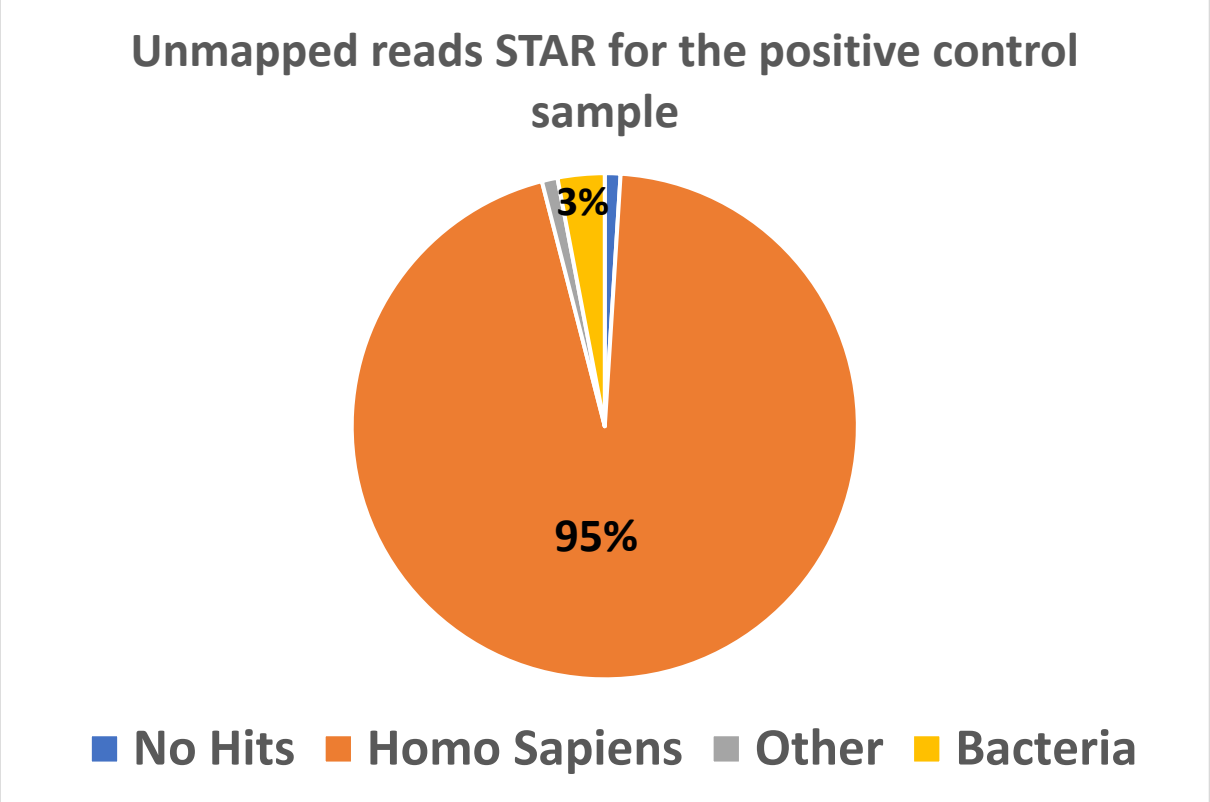


Figure 6-4 Visualisation of the taxonomic classification of the unmapped reads from STAR for pilot 2 using Kraken2 software.

A) *HVOL sample: 63% of the unmapped reads aligned to bacterial genomes (yellow), 13% were classified as Homo sapiens (orange), 21% did not align to any genomes ('no hits', blue), and 3% aligned to archaeal or fungal genomes ('other' grey) compared to* ***B)*** *positive control brain tissue sample, 3% bacteria (yellow) , 96% were classified as Homo sapiens (orange),*

0.9 'no hits' (blue), an 0.2% other (grey) . For the full output from the Krona visualisation software, which has a breakdown of the bacterial classifications, see Appendix D-2.

Thus, the issue of the high percent of 'Unmapped: too short' reads from the HVOL sample appears to be contamination or bacterial RNA inherent in the plasma sample. In summary, these data suggested addition of the DNAse step was successful, with the read distribution and 'strandedness' plots indicative of an RNA sample not DNA. However, there was also bacterial contamination or bacterial transcripts inherent in the plasma samples which resulted in a large percentage of reads not being mapped to the human reference.

6.3.3 Pilot 2a Ascertaining whether heart related mRNAs/lncRNAs could be detected in plasma

Table 6-4 Sample information for pilot 2a

Sample ID	Sample information
HVOL plasma sample 2	Healthy volunteer 4.5ml starting volume

As pilot 2 on the MiSeq pilot was successful in eliminating contamination with DNA, the same HVOL RNA library (Table 6-4) was run on the HiSeq to ascertain the level of mRNA and lncRNA detection and if any heart related mRNAs and lncRNAs could be detected when the sample was sequenced to a depth of 85 million reads. The QC metrics showed similar profiles to that of the MiSeq pilot 2 (Appendix D-3). The reads were imported into the R package DESeq2 [400] and transcript per million (TPM) values were determined for all annotated mRNA and lncRNA genes (comparing to GENCODE). A total of 5,161 annotated mRNAs and 553 annotated lncRNAs were detected at a minimum of 1 TPM.

To assess whether cardiac mRNAs may be present in plasma, expression levels of the 100 most abundant genes in heart left ventricle were downloaded from the GTEX portal (https://www.gtexportal.org/home/) and compared to the plasma mRNAs. Of the GTEX left

ventricle mRNAs, 25%were detected in plasma, and showed a high correlation when expression levels were compared. (Spearman correlation coefficient = 0.75, Figure 6-5).

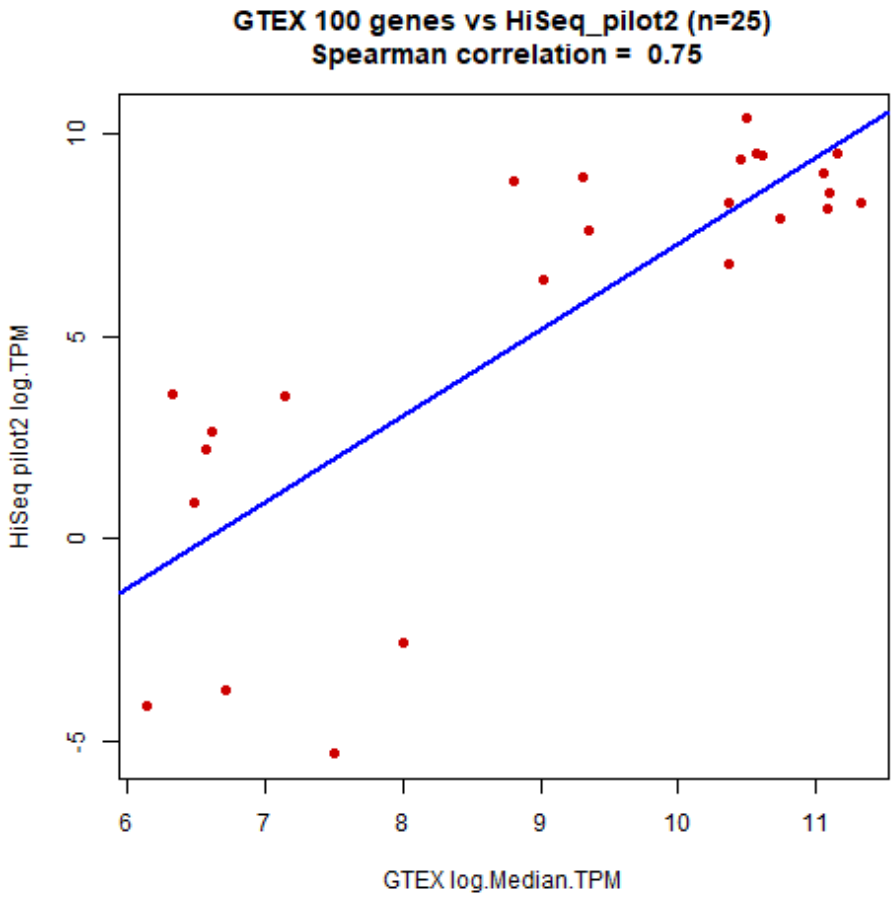


Figure 6-5 The most abundant left ventricle mRNAs are detectable in plasma

Similarly, to assess whether heart-related lncRNAs may be present in plasma, plasma lncRNAs were compared to ten human lncRNAs involved in heart disease [457] Of these, eight were detected in plasma in the current study, although four were below 1 TPM (Table 6-5).

In summary, these data suggested that despite only ~30% of reads uniquely aligning to the human reference, heart-related mRNAs and lncRNAs could be detected in plasma, when sequencing at a depth of ~85 million reads per sample.

Table 6-5 Eight out of ten human heart disease related lncRNAs were detected in plasma

gene	Ensembl ID	HVOL TPM
MALAT1	ENSG00000251562.8	319.060002
SENCR	ENSG00000254703.2	12.667896
FTX	ENSG00000230590.9	2.442054
H19	ENSG00000130600.18	1.866076
MEG3	ENSG00000214548.16	0.15
MIAT	ENSG00000225783.7	0.14
FENDRR	ENSG00000268388.5	0.07
HOTAIR	ENSG00000228630.5	0.05
ANRIL/CDKN2B-AS1	ENSG00000240498	
NRON	ENSG00000253079	

6.3.4 Pilot 3 Determining the minimum volume of starting plasma and assessing the level of RNA contamination present in the kit reagents

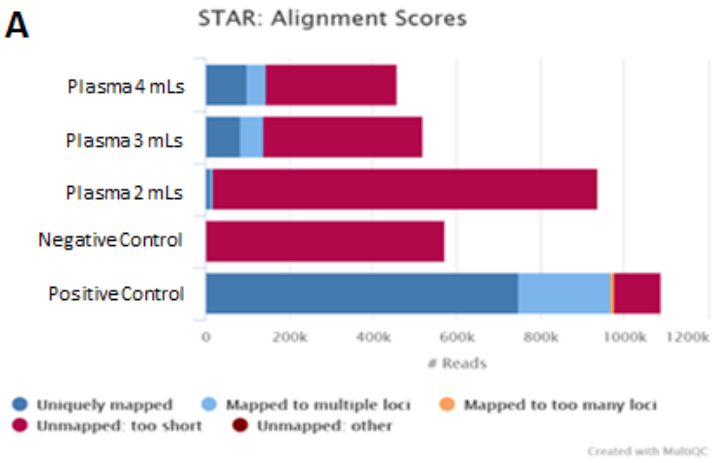
Table 6-6 Sample information for pilot 3

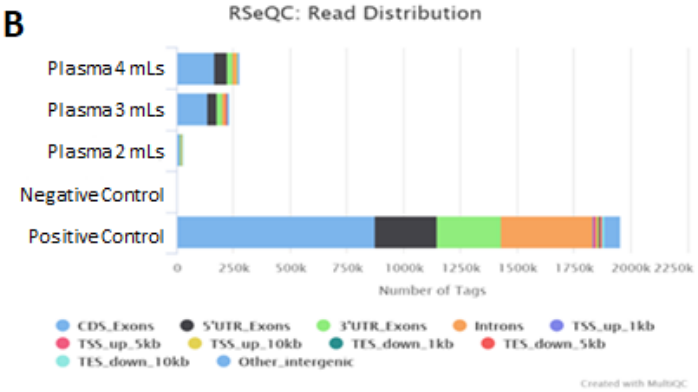
Sample ID	Sample information
Lab volunteer	4ml starting volume
Lab volunteer	3ml starting volume
Lab volunteer	2ml starting volume
Negative Control	Nuclease free water
Positive Control	Brain tissue RNA from the library preparation kit added at the sequencing facility at equimolar concentrations

As the volume of archived plasma was limited, with less than the advised starting volume of 5mL available, pilot 3 set out to test the minimum volume of plasma required using starting volumes of 4 mL, 3 mL, and 2 mL. A negative control with fresh nuclease free water was also tested to explore the origin of the bacterial RNA contamination (Table 6-6). Even though the samples were added to the library pool at equimolar concentrations, it was impossible to achieve this exactly, with the total number of aligned reads ranging from ~450 K to 1.1 million reads, as can be seen in the read alignment plots (Figure 6-6). Table 6-7 gives a summary of the read mapping and alignment statistics for human and bacterial genomes.

Table 6-7 A summary of the percentage of reads mapping to the human genome and bacterial genomes in pilot 3.

Sample ID	% Uniquely mapping reads to human genome	% Unmapped: too short	% reads aligning to coding regions	% reads aligning to antisense strand	% Unmapped reads aligning to bacterial genomes *	% Unmapped reads aligning to other genomes *
Plasma 4 mL	22	68	89.3	75	63	• 7 Human • 2 'other' • 23 'No hits'
Plasma 3 mL	16	73	88.5	78	68	• 9 Human • 1 'other' • 23 'No hits'
Plasma 2 mL	1.4	98	59.7	70	75	• 0.7 Human • 3 'other' • 21 'No hits'
Negative Control	0.2	99.7	9.9	51	79	• 0.3 Human • 0.8 'other' • 20 'No hits'
Positive Control	69	10	73.1	91	3	• 95 Human • 1 'other • 1 'No hits''





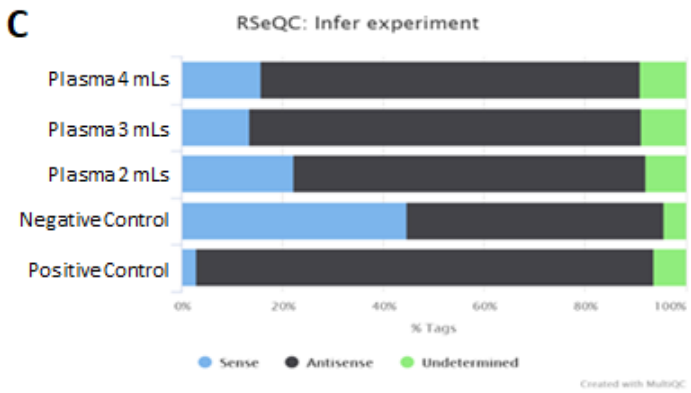


Figure 6-6 Pilot 3: Testing differing starting plasma volumes and a negative control
A) *Alignment scores. Starting volumes of 4 mL and 3 mL showed a similar percentage of uniquely mapping reads (22% and 16% respectively); The 2mL sample showed a very low percentage of uniquely mapping reads (1.4%). In contrast, a small percentage of reads were classified as 'Unmapped: too short' for the positive control* ***B)*** *Of the reads that did align, most aligned to coding regions which is indicative of RNA not DNA* ***C)*** *4 mL and 3 mL samples showed similar profiles with respect to the 'strandedness' of the reads; the 2mL sample showed a lower percentage of 'strandedness'.*

As in pilot 2a, Table 6-7 and Figure 6-6 show that RNA samples from plasma contain a high percentage of reads classified as 'Unmapped: too short' (subsequently shown to be of bacterial origin). The unmapped reads that were classified as human appeared to be reads from repetitive regions (after BLASTing these to the human genome) and may be from the 'mapped to multiple loci' category. The 4 mL and 3 mL samples showed read distribution and 'strandedness' plots indicative of RNA samples, whereas the 2mL sample showed a higher percentage of reads aligning to the sense strand, suggestive of DNA contamination.

To confirm that the plasma RNA profile was consistent between the 4 mL, 3 mL and 2 mL starting volumes, a Spearman's correlation was carried out for the 200 most abundant genes identified from each sample. In addition, to determine the degree of similarity in the RNA profile between individuals the RNA profile of the 4mL sample (lab volunteer) was compared with that of the 4.5ml HVOL sample from pilot 2. Strong correlations were found between the 4mL versus 3mL and 4mL versus 2mL starting volumes for pilot 3 (correlation coefficient 0.9

and 0.7 respectively, Figure 6-7A and 6-7B). I also saw a moderate correlation between the 4mL sample from the lab volunteer versus the 4.5ml HVOL sample from pilot 2 (correlation coefficient = 0.4, Figure 5-7C). This suggested that the RNA profile was reproducible for starting volumes ranging from 2ml-4ml. However, the 2 mL sample performed less well in quality metrics (as seen in Figure 6-6A), with only 1.4% of reads uniquely mapping. Consequently, it was determined that the minimum volume of plasma to yield an acceptable percentage of uniquely mapping reads was 3 mL. Moreover, the moderate correlation between the lab volunteer and healthy volunteer sample from pilot 2 suggested that the genes identified were relatively reproducible between individuals, despite different plasma preparation conditions and the healthy volunteer sample having been in storage for over ten years.

To explore if the bacterial reads were originating from the plasma (i.e. leakage from the gut microbiome) or there was exogenous contamination (i.e. from the RNA columns or other reagents), Spearman's correlation was calculated for the 100 most abundant bacterial reads in each sample, including the no template control (NTC) which had been treated as if it was a plasma sample (*i.e.* spun through columns and had reagents added to it). When comparing the bacterial reads in the NTC with the 4mL, 3mL and 2mL samples, the Spearman's correlation was 0.7, 0.7 and 0.4 respectively (Figure 6-8). The high degree of correlation suggested that the bacterial reads in the plasma samples originated from the kits or the nuclease free water itself (added to the sample to top up the starting volume to 5 mL), rather than being endogenous to the plasma. The bacterial reads from the positive control (brain tissue) were also moderately strongly correlated with the bacterial reads from the NTC (Spearman correlation = 0.5, Figure 6-8D).

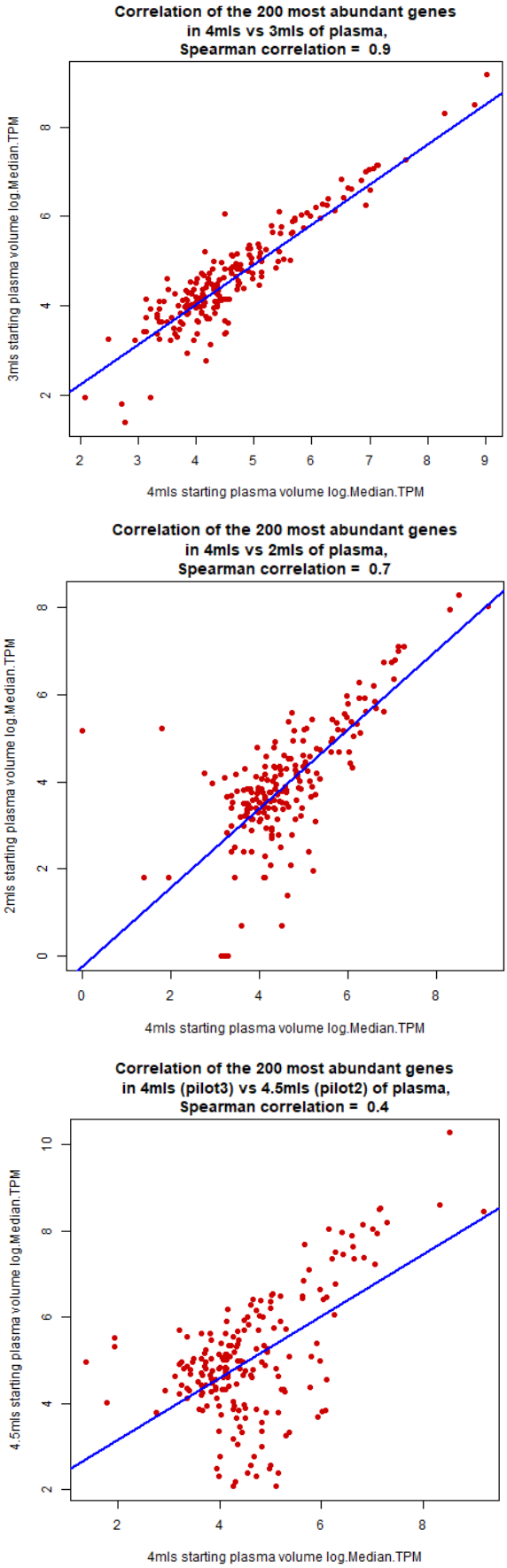


Figure 6-7 Spearman's correlation plots of the 200 most abundant genes

A) *4ml versus 3ml starting plasma volume,* ***B)*** *4ml versus 2ml starting plasma volume* ***C)*** *4ml versus 4.5ml (pilot2) starting plasma volume. X and Y axis show log normalised counts.*

However, the degree of contamination by bacterial reads in the brain sample was considerably less compared with the plasma samples: the most abundant bacterial transcript in the positive control that correlated with the NTC was only 0.4% of the total reads whereas this was between 12-20% of the total reads in the plasma samples (Figure 6-8A-D)

In summary, there have been several previous reports of bacterial contamination originating from laboratory reagents and consumables, and this contamination is especially exacerbated in samples with low endogenous content (such as plasma samples) where the contaminating nucleic acids can outnumber the nucleic acids within the sample [458-462]. This was not seen in the brain tissue positive control in which RNA is much more abundant than in the plasma samples, with only 3% of total unmapped reads mapping to bacteria.

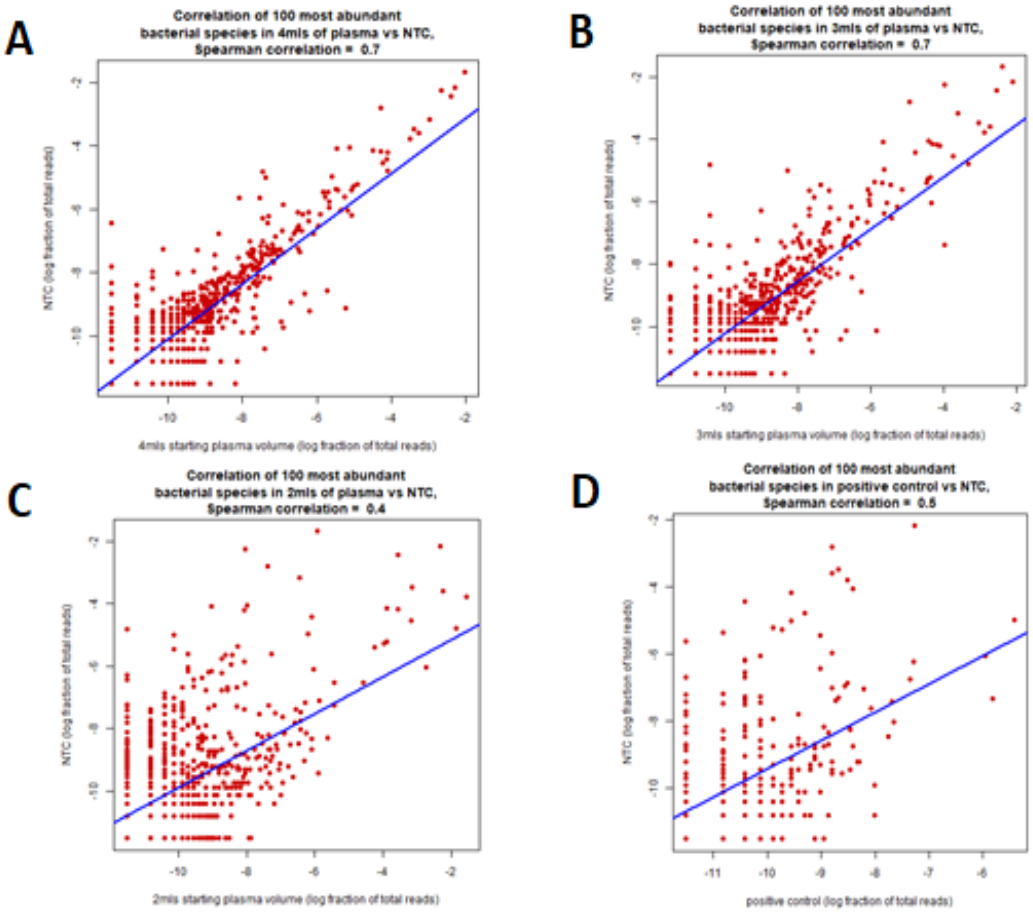


Figure 6-8 Spearman correlation plots comparing the 100 most abundant bacterial species in each sample against the NTC.

A) *4mL starting volume compared to the NTC* ***B)*** *3mL starting volume compared to the NTC* ***C)*** *2mL starting volume compared to the NTC* ***D)*** *Positive control compared to the NTC*

Because the Spearman correlation between the bacterial species in the positive control and the NTC was moderately strong (correlation coefficient = 0.5), it is likely that the relatively small number of bacterial species in the positive control originated from the kits or the nuclease free water itself. It seems that this bacterial contamination cannot be avoided for the plasma samples, in which the RNA is much less abundant. For future work these reads could be analysed to see if they align to bacterial ribosomal RNAs. If they do then oligonucleotide probes could be designed against the bacterial rRNAs to deplete them from the sample before sequencing [463]. The fact that 20-30% of the reads from plasma samples uniquely aligned to the human reference was encouraging and suggested that an accurate picture of the plasma transcriptome could still be achieved, even from starting volumes as little as 3 mL, provided the samples were sequenced deeply enough.

6.3.5 Pilot 4: Testing plasma samples that have been stored for >10 years and inclusion of artificial spike-in controls

Table 6-8 Sample information for pilot 4

Sample ID	Years in storage	Sample information
HVOL plasma 3	11	1% Sequins spiked in laboratory
HVOL plasma 3	11	No Sequins spiked
CDCS plasma 2	14	1% Sequins spiked in laboratory
CDCS plasma 2	14	No Sequins spiked
Positive Control	-	Brain tissue RNA from the library preparation kit added at the sequencing facility

The aims for the final pilot were to confirm that the RNA from patient and healthy control plasma samples that had been in storage over ten years could be sequenced and to test whether Sequins - artificial RNA spike-ins [371] - could be used as internal controls in plasma samples (Table 6-8). Sequins were spiked into each sample at the recommended 1% of the total RNA concentration prior to reverse transcription, library preparation and sequencing (Appendix D-1). As they share no homology with the human reference genome, they could be aligned back to their *in silico* genome and in this way, they could be used as internal controls to check that the library preparation and sequencing had worked successfully. For this

experiment, two plasma samples, with and without the addition of 1% v/v of Sequins, were compared.

From the QC plots (Figure 6-9) the samples that had been in storage over ten years displayed similar QC metrics to the samples from previous pilots. The read distribution and 'strandedness' plots showed there was no DNA contamination and there were similar percentages of uniquely mapping reads and reads that were 'Unmapped: too short'.

Detection of sequins suggested that the library preparation and sequencing had worked successfully. Table 6-9 shows the number of reads aligned to the human reference and Sequin reference combined as well as the number of reads aligned to the Sequin reference alone, from which the percentage of reads originating from the Sequins was calculated. The percentage of reads aligning to the Sequins reference in the HVOL sample was 33% and in the CDCS sample was 55%, substantially different from the expected 1%.

Table 6-9 Alignment statistics from the BAM file using Samtools flagstat.
Total reads aligning to the Human and Sequin reference combined and the Sequin reference alone, and the percentage of total reads aligning to Sequins are presented.

Sample ID	Total Reads aligned to both Human and Sequin reference	Total Reads aligned to Sequin reference	Percentage of reads aligning to Sequins (column 2/column 1)*100
HVOL plasma 3 (1% Sequins added)	280326	91654	33
CDCS plasma 2 (1% Sequins added)	500632	275424	55

In summary, samples archived for more than 10 years can be reliably sequenced using this method. However, the high percentage of sequin reads in the sample was unexpected. Sequins are usually added to RNA extracted from tissue and there is currently no information in the literature on adding Sequins to human plasma samples with such low RNA input amounts. Despite using the adapted protocol, [369] the high percentage of reads aligning to the Sequins suggested that the estimation of the quantity of (human) RNA in the plasma

samples was a gross overestimate, potentially owing to the large amount of bacterial contamination present, resulting in a higher percentage of sequins being added.

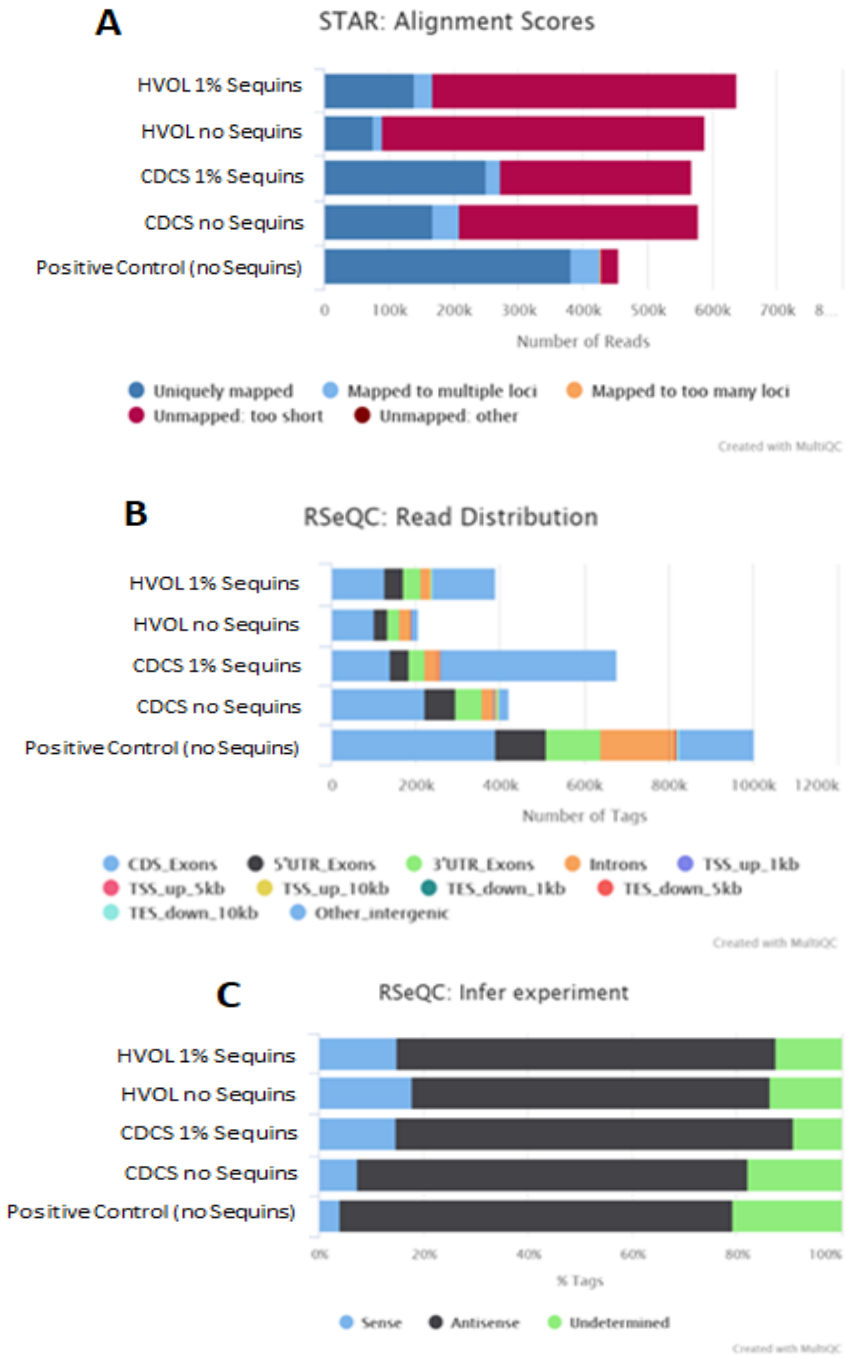


Figure 6-9 Testing whether synthetic spike ins can be added to the plasma sample.
A) *Alignment scores. This time reads were aligned to the human and in silico Sequin genomes combined, and hence the samples with no Sequins spiked in look to have fewer uniquely mapped reads aligning.* ***B)*** *Similarly, the samples with Sequins spiked in have a higher percentage of reads mapping to 'other intergenic' regions in the read distribution plot (as these are the reads aligning to the artificial Sequin genome)* ***C)*** *The Sequins do not affect the 'strandedness' of the libraries as they act*

the same as the endogenous RNA with respect to reverse transcription, library preparation and sequencing

Because the amount of human RNA could not be quantified, it was determined that the addition of Sequins to plasma samples would be omitted in the final project to avoid reducing the number of reads available for human RNA.

6.4 Conclusions.

RNA sequencing of plasma samples is technically challenging due to low RNA amounts, but the following conclusions can be drawn from the pilot studies.

- The addition of the DNAse step to the extraction protocol is critical and outweighs any potential loss of RNA.
- There was a large and unexpected proportion of reads that were bacterial, which appeared to originate from the kits.
- Heart disease-associated mRNAs and lncRNAs can be detected in human plasma when sequenced at a depth of 85 million reads.
- The minimum starting volume for RNA extraction was 3 mL plasma.
- Samples with storage time > 10 years were still able to be sequenced reliably.

While the addition of artificial RNA spike ins confirmed the sequencing run had performed as expected, they will be omitted in future experiments owing to the inability to reliably quantify the amount of human RNA in the sample and the potential for RNA sequins to take up a high proportion of reads.

Chapter 7
RNA-Sequencing of plasma from healthy volunteers and heart patients.

7.1 Introduction

Despite major advances in cardiovascular risk prediction over the last twenty years, we lack the ability to identify individuals at impending risk of MI or accurately identify all patients at risk of heart failure. To identify potential novel candidate RNA biomarkers, in this chapter I aimed to characterise the human plasma transcriptome and identify mRNAs, lncRNAs or circRNAs associated with the presence of Ischaemic Heart Disease (IHD) and progression to HF. Here, I describe a screening study to identify candidate diagnostic markers for follow up in larger cohorts.

7.1.1 Overview of research design

Building on the pilot studies in the previous chapter, this chapter presents the results from deep RNA-Seq of plasma from 31 Healthy volunteers (HVOLs), 31 patients with unstable angina or myocardial infarction who remained free of heart failure for at least three years ('CDCS heart failure negative') and 30 patients with either unstable angina or myocardial infarction who were diagnosed with heart failure within three years ('CDCS heart failure positive'). Patient selection and a summary of patient characteristics is provided in Section 3.2.3. Bloods were taken on average 4 months after the acute event when the patients were stable (i.e., gene expression changes associated with the acute injury had passed and expression changes would more likely to be associated with either cardiac re-modelling or compensatory mechanisms leading to heart failure). RNA extracted plasma from each of the three groups was sequenced to a depth of ~100M reads per sample and analysed with the bioinformatics pipeline to produce lists of mRNAs, lncRNAs, novel lncRNAs and circRNAs associated with ischaemic heart disease and progression to HF.

Prior to deep sequencing, all samples were barcoded, pooled at equimolar concentrations, and sequenced along with a positive (brain tissue) and negative control (RNase-free water), at a low read depth (~1 million reads) on an Illumina MiSeq to check that library preparation had worked satisfactorily. Quality control metrics showed similar distributions to the pilot experiments with only ~30% of reads uniquely mapping to the human genome (Appendix figure E-1A). The majority of reads aligned to coding regions and to the antisense strand indicating that the libraries consisted of RNA (Appendix Figure E-1B/C). The combined sample library was then sequenced to a greater depth (10 billion reads) on the Illumina NovaSeq 6000.

The raw reads were processed through the bioinformatics pipeline (presented in Section 4.1) to produce raw counts for annotated mRNA and lncRNA genes, novel lncRNA transcripts and back-spliced junction reads to identify circRNAs. The raw read counts were then analysed with DESeq2 to calculate differential expression (log2 fold change with Benjamin-Hochberg adjusted p-values) and transcript per million (TPM) values. For gene level analyses, a filter of at least 1 TPM for at least 90% of samples was applied.

7.2 Results

7.2.1 Quality assessment

RNA-Seq on the NovaSeq 6000 resulted in a median of 108M reads per sample, ranging from 5M-193M. However, upon closer inspection two libraries that had been prepared for a previous pilot and spiked into the final pool appeared to have much higher reads possibly due to being spiked in at a higher molarity. If we remove these two samples plus the outlying samples discussed in the next paragraph the range of reads was 84-138M. Analysis of variance (ANOVA) confirmed that read depth did not differ between the patient groups or controls, suggesting that read depth was not associated with disease phenotype or cohort and would be unlikely to confound differential expression analysis (p-value 0.58), see Tables 7-1 and 7-2 for summary statistics for all samples and for each group). The median number

(percentage) of uniquely aligned reads was 21M (19%) ranging from 1.5M-64M (2.2%-73.9%) (Figure 7-1, Table 7-1) Quality control metrics for the NovaSeq 6000 sequencing of plasma samples and brain tissue positive control showed similar distributions to the preliminary HiSeq sequencing, confirming that the NovaSeq run was successful (also Appendix Figures D-1 and D-2).

Table 7-1 Summary statistics for NovaSeq 6000 RNA-Seq for all samples

	Total reads (million)	% reads aligning	Reads uniquely aligning (millions)
Median	107.9	19.0	21.4
Minimum	5.4	2.2	1.5
Maximum	193.0	73.9	64.3

Table 7-2 Summary statistics for the NovaSeq 6000 RNA-Seq per group

	Median Total reads (millions)	Median % reads aligning	Median reads uniquely aligning (millions)
HVOL	111.2	20.0	22.5
CDCS HF -	100.1	19.1	21.9
CDCS HF +	110.3	13.5	14.5

Table 7-3 Summary statistics for the NovaSeq 6000 RNA-Seq outliers sequencing for all samples (outliers discarded)

	Total reads (millions)	% reads aligning	Reads uniquely aligning (millions)
Median	107.9	19.0	21.4
Minimum	84.0	3.3	3.4
Maximum	193.0	73.9	64.3

Table 7-4 Summary statistics for the NovaSeq 6000 RNA-Seq per group (outliers discarded)

	Median Total reads (millions)	Median % reads aligning	Median reads uniquely aligning (millions)
HVOL	111.2	18.9	23.9
CDCS HF -	102.0	19.0	22.4
CDCSHF +	110.2	13.6	15

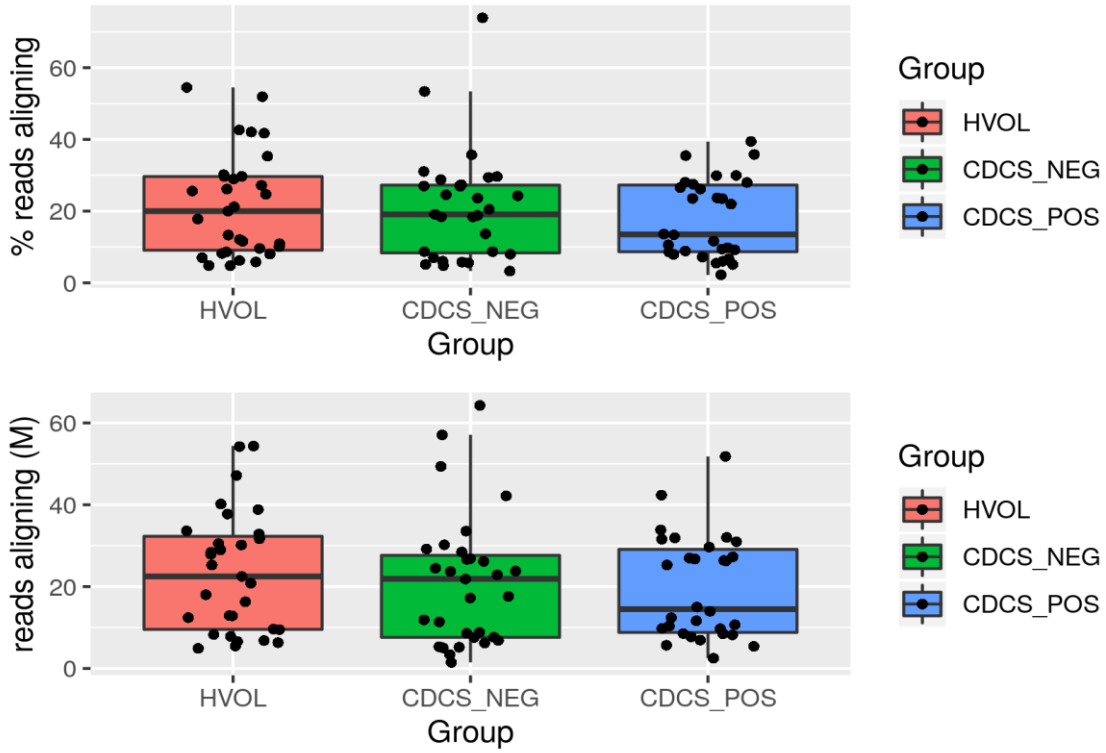


Figure 7-1 Plots of the three sample groups showing percentages of reads and millions of reads aligning for the three groups

Healthy Volunteer cohort, CDCS NEG: Coronary Heart Disease (Heart Failure negative), CDCS POS: Coronary Heart Disease (Heart Failure positive)

A Principal Component Analysis (PCA) identified three outlying samples, S25, S76 and S43 (Figure 7-2), coincidentally one outlier from each experimental group. Samples S25 and S76 had been flagged by the sequencing facility as failing the library preparation QC owing to a low yield of cDNA. The remaining sample, S43, had only 2.5 million reads (2.2%) aligning to the human genome and showed a 50/50 distribution on the infer experiment plot, indicative of a high degree of DNA contamination in the RNA library. All three samples were discarded, leaving 89 samples remaining for analysis (HVOL 30, CDCS HF negative 30, CDCS HF positive 29). The summary statistics were recalculated (Tables 7-3, 7-4) and this increased the minimum total reads per sample from 5.4M to 84M.

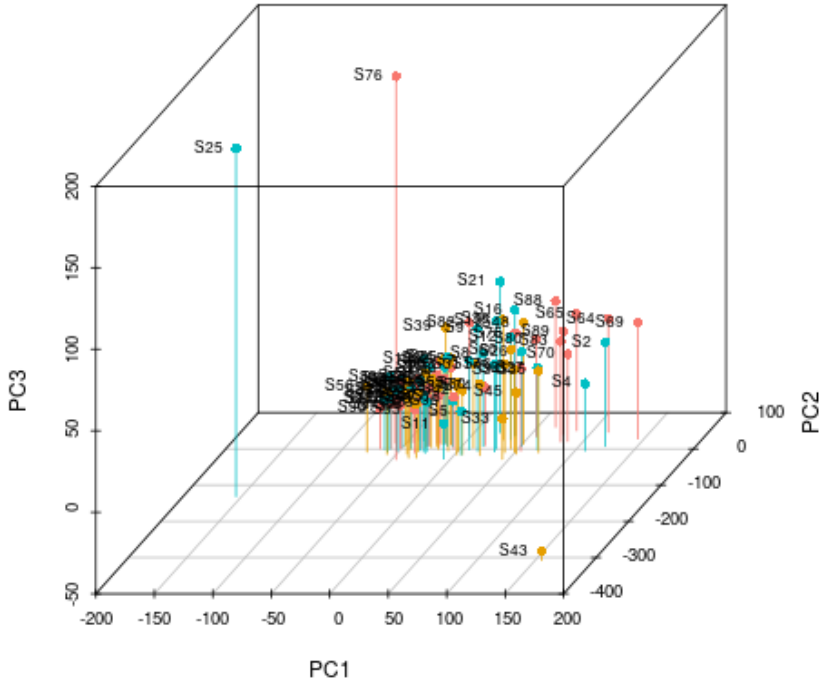


Figure 7-2 Principal Component Analysis plotting the normalised gene counts for each sample. *There are three outlying samples, S25, S76 (which were flagged by the sequencing facility as failing the minimum yield QC) as well as S43 which had the lowest number of reads aligning (second to S25). These three samples were discarded from the analysis.*

7.2.2 Differential expression analysis - annotated mRNA and lncRNAs

RNA-Seq of the plasma transcriptome identified 60,317 annotated genes (GENCODE.v33), of which only 4,153 genes remained after applying the filter of at least 90% of samples having ≥1 TPM. Of these, 3,986 were classified as mRNAs and 167 were classified as lncRNAs. Nearly a quarter of the genes (921, 22%) that were detected in the plasma were also detected in human left ventricle heart tissue (Harvard data, Section 5.3.2) with an abundance Spearman correlation of 0.33 Figure 7-3. To try to identify whether these were heart specific genes the genes detected in plasma were compared to a list of heart enriched genes generated through a median-based analysis of tissue-specific gene expression based on the GTEx data [464] (heart-specific genes were defined as genes with fold changes of the median expression levels

higher than 5.0 in the heart versus all other tissues). From this analysis one gene was present in both datasets (FGF23).

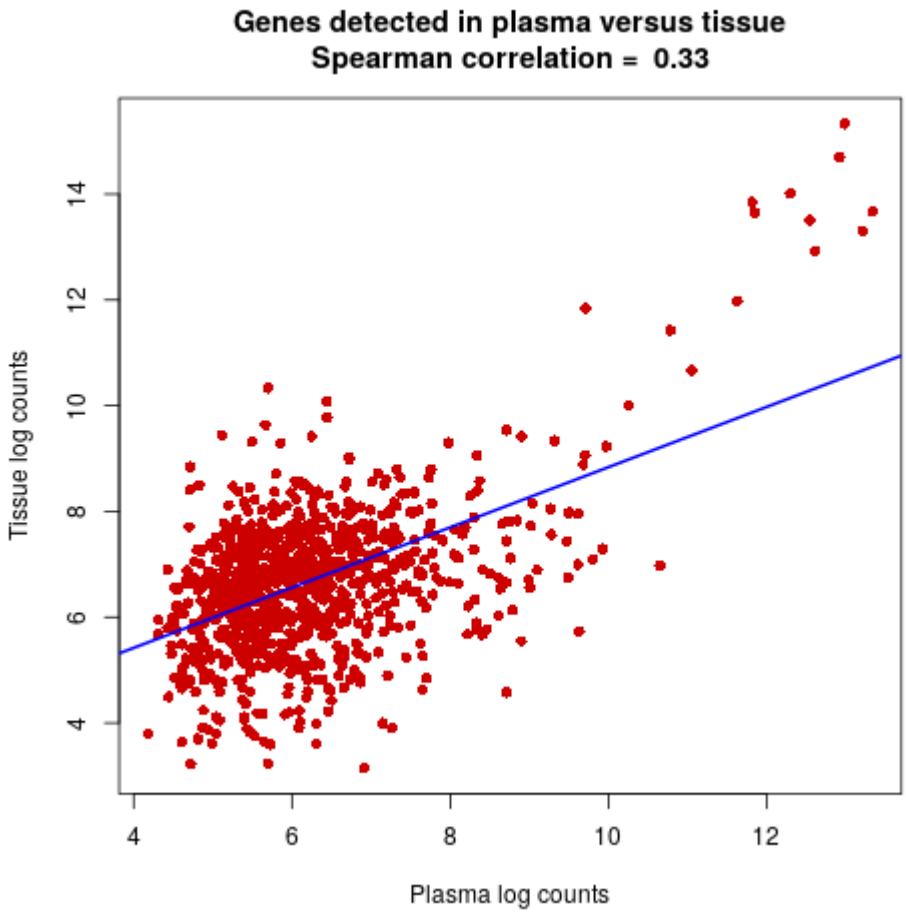


Figure 7-3 A scatter plot showing the correlation between normalised gene counts in plasma and left ventricle heart tissue

22% of genes detected in plasma were also detected in heart tissue.

In plasma, 13 out of the 20 most abundant protein-coding genes were mitochondrial (*MT-ND1, MT-ND2, MT-CO1, MT-ND4, MT-CO2, MT-ATP6, MT-ND5, MT-CYB, MT-CO3, MT-ND6, MT-ATP8, MT- ND3, MT-ND4L*) and 4 out of the 20 most abundant lncRNA genes had previously been shown to be expressed in the heart (*MALAT1, SNHG6, RMRP* and *ZFAS1*) [408, 465-468].

To identify candidate RNA biomarkers for the presence of IHD or progression from IHD to ischaemic HF, differential expression analysis was performed comparing gene expression i) between the patient groups and controls, and ii) between the two patient groups, separately.

Three patterns of differential expression were prioritised (Figure 7-4); namely that the CDCS heart failure positive group had significantly increased expression compared to the other two groups (indicating potential markers for progression from coronary heart disease to ischaemic heart failure, Scenarios 1 and 2, Figure 7-4) or that both CDCS groups had significantly increased expression compared to the HVOLs (indicating potential markers for the presence of IHD, Scenario 3, Figure 7-4).

Differential expression analysis (padj < 0.01 and absolute fold-change >1.2) between the CDCS HF positive versus CDCS HF negative showed no significantly differentially expressed genes (although there were four genes that were padj < 0.2 which suggest candidate markers for further investigation, Appendix Table E-1). This consequently ruled out scenarios 1 and 2, leaving Scenario 3.

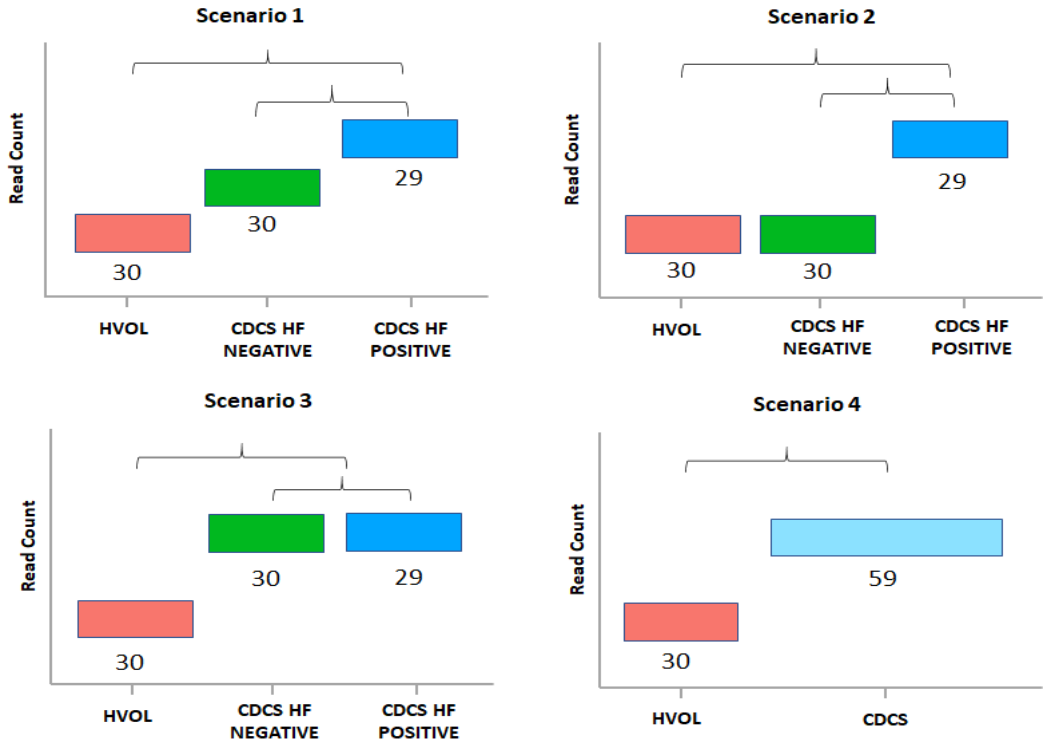


Figure 7-4 A schematic of the scenarios of gene expression between the three groups for biomarker analysis of acute coronary syndromes or progression from coronary heart disease to ischaemic heart failure.

Scenarios 1 -3 looked for upregulated genes for either or both of the CDCS groups compared to the HVOL group. Scenario 4 combined the two CDCS groups and looks for upregulated genes for the CDCS groups (combined) against the HVOL group.

In contrast, significantly differentially expressed genes were identified when both CDCS HF positive and CDCS HF negative were compared to the HVOLs (separately). To improve statistical power, it was decided to combine the two CDCS groups and look at differential expression between CDCS vs HVOL (equivalent to scenario 4). enabling identification of possible biomarkers of IHD (HVOL n=30 versus CDCS n=59).

This scenario distinguishes the CDCS cohort as a whole from the HVOL group, thereby DESeq2 analysis showed that 170 genes were differentially expressed between HVOLs and CDCS patients combined (padj < 0.01, absolute fold change > 1.2, with 96 genes up regulated in CDCS patients). Of these, 88 genes were mRNAs (Table 7-5), and 8 genes were lncRNAs (Table 7-6) which all appeared to directly overlap (n=7), or were very close to (7308 bases, n=1), CCCTC-binding transcription factor (CTCF)-binding sites (Appendix . Of the 20 mRNA genes with the highest fold change, 13 were mitochondrial genes (*MT-ND6, MT-ND1, MT-ND5, MT-CYB, MT-ND3, MT-CO2, MT-ATP8, MT-CO1, MT-ND4, MT-CO3, MT-ATP6, MT-ND2, MT-ND4L,* Table 7-5).

Table 7-5 A list of the top 20 protein coding genes that were higher in the CDCS versus HVOL cohort (padj < 0.01 CDCS versus HVOLs, sorted by fold change)

Gene name	Normalised Mean Read Counts	log2 Fold Change	padj
MT-ND6	112460	2.2	1.64E-27
MT-ND1	618841	2.2	1.65E-26
MTRNR2L12	654	2.1	8.15E-20
MT-ND5	218593	2.1	5.80E-28
MT-CYB	139778	2.1	3.29E-27
MT-ND3	48232	2.1	5.75E-29
MT-CO2	299436	2.1	1.00E-27
MT-ATP8	63002	2.1	3.25E-22
MT-CO1	434596	2.0	1.84E-25

MT-ND4	409106	2.0	9.42E-25
MT-CO3	135217	2.0	1.82E-24
CXCL14	145	2.0	1.40E-25
MT-ND2	547684	2.0	6.49E-25
MT-ND4L	16491	2.0	5.77E-26
NEUROD2	2257	1.9	1.48E-31
STAG2	2006	1.6	1.62E-36
FGF23	3154	1.6	1.14E-16
BTN3A2	1582	1.4	6.75E-24
ZNF302	4814	1.3	9.00E-09

Table 7-6 A list of lncRNAs that were higher in the CDCS versus HVOL cohort (padj < 0.01 CDCS versus HVOLs, sorted by fold change)

Gene name	Mean Counts	log2 Fold Change	padj
AL035078.1	673	1.8	6.48E-26
CCDC26	6006	1.7	4.51E-25
AC009404.1	230	1.6	1.50E-18
LINC02245	227	1.3	1.26E-10
RGPD4-AS1	193	1.3	1.01E-06
AL360012.1	1675	1.0	0.000156
MSC-AS1	148	0.8	0.000981
AD000090.1	14545	0.7	0.000381

7.2.3 Differential expression analysis - novel lncRNAs

RNA sequencing of the plasma transcriptome detected 28,725 coding and non-coding transcripts (after the abundance filter was relaxed for lncRNA detection to at least half of the samples having a read count above zero). Of these, 4,544 were classified as putative novel transcripts. To minimise false-positive detection of novel transcripts, only multi-exonic transcripts on the main chromosomes (not scaffold chromosomes) were selected for differential expression analysis (n=430). Not surprisingly, the novel transcripts were detected

at a much lower abundance than the annotated genes (median TPM of novel lncRNAs ranged from 0-50 compared to 2-33,374 for annotated lncRNAs) and only two putative novel lncRNAs were detectable in all 89 samples. Comparing the 430 putative novel transcripts to the putative novel transcripts identified in the Harvard data using short- (Hi-Seq) and long-read (Nanopore) sequencing with gffcompare (which can match transcripts with the same exons start and end co-ordinates) resulted in matches for two novel transcripts in the Harvard short-read data - MSTRG.74727.3 and MSTRG.202989.1, which overlapped a CTCF region (Table 7-7). While neither of these transcripts were differentially expressed, two other transcripts from the plasma dataset were significantly up regulated in CDCS patients compared with HVOLs (MSTRG.752033.2 log2 fold change 1.7, padj = 0.0017 and MSTRG.76602.1 log 2 fold change 2.4, padj = 0.003, Table 7-7).

No statistically significant differentially expressed putative novel transcripts were identified when comparing CDCS HF negative and CDCS HF positive patient groups.

7.2.4 CircRNA

For circRNA detection two different software packages were used: CircExplorer2 [242] (part of the bioinformatics pipeline) and CircTools [262]. CircExplorer2 detected a total of 66,184 circRNAs but after a filter of 90% of samples having reads above zero, 227 circRNAs remained. In contrast, CircTools detected 17,291 circRNAs (although this was with a filter of a minimum count of 5 reads in at least 6 samples - the default when running the software). After the same filter of 90% samples having reads above zero, 255 circRNAs remained. However, 50 genes appeared to be duplicated; in other words, there were 50 genes that had read counts for the same coordinates but on the opposite strand

strand. The developer of CircTools, Dr Jakobi, was contacted regarding this issue who suggested that it is not uncommon to see circRNAs also on the other strand with the same coordinates. Normally one of the strands is more dominant with respect to the read counts.

While this was true for most genes, for some genes the dominant strand was the opposite strand to CircExplorer2 and was classified as ‘not annotated’ for CircTools (example shown in Table 7-8).

Table 7-7 A list of interesting putative novel lncRNA transcripts.

The first two putative novel lncRNAs were expressed at higher levels in CDCS compared to HVOLs (padj < 0.01, sorted by fold change). The last two putative novel lncRNAs were also detected in the Harvard Heart Tissue analysis.

MSTRG ID	Chromosome	Start	Stop	Strand	Number of exons	Mean Counts	log2 Fold Change	padj
MSTRG.76602.1	chr10	3408557	3409012	+	2	6.8	2.4	0.003361
MSTRG.752033.2	chr7	113240035	113240441	-	2	7.6	1.7	0.001659
MSTRG.74727.3	chr1	247363146	247363578	-	2	27.3	0.5	0.897159
MSTRG.202989.1	chr13	20392432	20393540	-	2	9.8	0.5	0.753564

Table 7-8 An example of circTools detecting circRNA reads on both strands

Coordinates	CircExp2 Strand	CircExp2 Annotation	CircExp2 Counts	CircTools Strand	CircTools Annotation	CircTools Counts
13:99238426- 99244624	+	UBAC2	696	+	UBAC2	157
13:99238426- 99244624	+	UBAC2	696	-	not annotated	637

Consequently, the duplicates were discarded, leaving a total of 205 circRNAs detected by CircTools. A key difference between the two software packages is that CircTools detects circRNAs from the mitochondria (CircTools includes an option to filter out mitochondrial reads, whereas CircExplorer2 filters out mitochondrial reads without giving the option to disable the filter). In R, an overlap between the two software packages was calculated by extracting circRNAs that had the exact same co-ordinates. This identified 180 circRNAs that were consistently detected by both packages.

For CircExplorer2, no circRNAs were significantly differentially expressed between HVOLs and CDCS patients after normalisation (padj < 0.01, absolute fold change > 1.2), although three circRNAs had padj < 0.1 (*UBAC2*, padj 0.047; *CLNS1A*, padj 0.065 and *ASH2*L padj 0.09, Table 7-9).

Interestingly, CircTools, identified a potentially novel mitochondrial circRNA (detected in all 89 patients), that was significantly downregulated in CDCS patients compared to HVOLs (padj< 1.65E-20, log 2-fold change -2.03, table 6-10). However, this circRNA was annotated as being antisense to mitochondrially encoded tRNA valine (*MT-TV*) and because CircTools had an issue with the 'strandedness' of circRNAs the reads could be originating from the gene itself and the strand of origin would need to be validated. Analysis of the remaining circRNAs, identified four circRNAs that were potentially differentially expressed between CDCS and HVOL groups at padj<0.1 (unannotated antisense to *UBAC2*, *CLNS1A*, unannotated antisense to *FIP1L1*, unannotated antisense to *ZNF362*). These findings demonstrated that *circUBAC2* and *circCLNS1A* were differentially expressed in both software analysis.

There were no statistically significant differentially-expressed circRNAs when comparing CDCS HF negative and CDCS HF positive patient groups.

Table 7-9 A list of circRNAs identified from CircExplorer2

CircRNAs were expressed at higher levels in the CDCS cohort compared to HVOLs (N.B. no circRNAs were padj < 0.01 so the following are padj < 0.1, sorted by fold change).

chr	start	stop	strand	gene	Normalised Mean Read Counts	log2 Fold Change	padj	no Samples Null	no Samples Not null
chr11	77619605	77625818	-	CLNS1A	21	0.7	0.065036	6	83
chr13	99238426	99244624	+	UBAC2	696	0.4	0.047036	0	89
chr8	38114191	38119363	+	ASH2L	149	0.4	0.092905	0	89

Table 7-10 A list of circRNAs identified from CircTools

CircRNAs were expressed at higher levels in the CDCS cohort compared to HVOLs (Note no circRNAs were padj < 0.01 so the following are padj < 0.1, sorted by fold change).

chr	start	stop	strand	gene	base Mean	log2 Fold Change	padj	no Samples Null	no Samples Not null
chr11	77619605	77625818	-	CLNS1A	27	0.63	0.069782	3	86
chr13	99238426	99244624	-	not annotated	636	0.50	0.040348	0	89
chr4	53414614	53428183	-	not annotated	305	0.48	0.090983	0	89
chr1	33294936	33295305	-	not annotated	256	0.43	0.09515	1	88
chrM	1601	1671	-	not annotated	170	-2.03	1.65E-20	0	89

7.3 Discussion

Despite a large percentage of reads not mapping to the human genome when sequencing from human plasma, a median of ~20M uniquely mapping reads per sample can be obtained, provided the sample is sequenced to a sufficient depth (~100M reads per sample). From the results I am able to present a genome-wide screen from human plasma that identified promising candidates for genes upregulated in the coronary disease cohort compared to the healthy cohort. This list of upregulated genes was dominated by mitochondrial RNAs and mitochondrial associated genes which is an exciting discovery and worth follow up studies. Also, among this list of differentially expressed plasma genes was fibroblast growth factor 23 (*FGF23*) which is expressed in the heart, promotes hypertrophy and remodelling and has been identified as an independent marker for cardiovascular risk in various patient populations such as those with dilated cardiomyopathy, ischaemic heart disease, acute decompensated and chronic heart failure [469-473], along with STAG2 which is required both for proliferation and regulation of cardiac transcriptional programs [474].

A further gene which has one of the highest log fold changes between the coronary disease cohort compared to the healthy cohort, MTRNR2L12 - an isoform of Humanin, a 24 amino acid long molecule that is encoded by the mitochondrial 16S rRNA (MT-RNR2 gene) [475]. Humanin has been shown to protect against oxidative stress and apoptosis in the heart by reducing mitochondrial dysfunction [476-480].

Of downregulated genes, one lncRNA (which was the third most downregulated gene) was RNA Component Of Mitochondrial RNA Processing Endoribonuclease (RMRP) whose upregulation has been shown to aggravate myocardial I/R injury [465] and was demonstrated in mouse and human heart failure patients and displayed a nuclear intracellular localization [211]. Two out of the five most downregulated genes were also mitochondrial electron transport related genes – Cytochrome c oxidase assembly factor 6 (COA6) which is part of the

mitochondrial respiratory chain complex IV and absence of the protein leads to cardiomyopathy [481] and NADH:ubiquinone oxidoreductase complex assembly factor 2 (NDUFAF2) whose protein is involved in the assembly of NADH dehydrogenase (ubiquinone) also known as complex I. Deficiency of this is a common cause of mitochondrial oxidative phosphorylation disease and is also associated with cardiomyopathy [482].

Of the differentially expressed circRNAs identified (*UBAC2*, *CLNS1A*, and *ASH2*L, Table 7-9). These findings are consistent with a recent study that demonstrated that *circUBAC2* was upregulated in peripheral blood of MI patients and, when combined with four other circRNAs, had good sensitivity and specificity for MI diagnosis (preprint https://www.researchsquare.com/article/rs-33371/v1). Furthermore, circASH2L promotes tumour invasion, proliferation and angiogenesis by regulating miR-34a in pancreatic ductal adenocarcinoma [483] and miR-34a itself having been shown to play a key role in cardiac repair and regeneration following myocardial infarction [484].

The analysis of DNA and RNA liquid biopsies is an exciting and relatively new field that has received considerable attention in the past decade. The ability to detect circulating DNA and RNA disease biomarkers in a minimally invasive and relatively cheap manner highlights their clinical potential.

Initially, the focus of these studies was on circulating DNA but the potential of circulating cell-free RNA is slowly being realised. RNA-Seq tests have already been translated into clinical applications for cancer, for example, gene-fusion detection in blood [485, 486] and in prostate cancers [487].

RNA-Seq in plasma is technically challenging due to the very low abundance of RNA. In the current study, three samples failed quality control testing (reassuringly two of these had already been flagged as low quality by the sequencing facility). Whilst there has been success

for biomarker detection with small RNAs [488] microarrays and targeted analysis of lncRNAs with RT-qPCR [489, 490], very few studies have analysed total RNA in human plasma. However, the technical hurdles will diminish as RNA-Seq library preparations evolve. A recent study evaluated the performance in different biofluids of the strand-specific, total RNA library preparation kit used in this thesis (the SMARTer Stranded Total RNA-Seq Kit, Pico Input Mammalian). The kit was found to be highly accurate, allowing detection of several thousand transcripts from different classes, including mRNAs, lncRNAs and circRNAs [453]. Interestingly, that study found similarly low percentages of unique read alignment (~25%) in human plasma as the current study (median 21%). In contrast, the percentage of reads 'too short to map' and 'multi-mapping' was quite different (~50% and 25% in Everaert *et al* [453], versus median 81% and 2% in the current study). Everaert *et al* suggested that RNA that is 'too short to map' may represent degraded RNA, although they did not test these for alignment to bacterial genomes. The low percentage of uniquely mapping reads may be the reason why there are so few RNA-Seq studies in human plasma; if the sequencing was not deep enough, there would be too low a percentage of human reads aligning to be informative.

Among the 20 coding genes most differentially expressed between the HVOL and CDCS cohorts, (ranked on fold-change), 13 originated from mitochondria. In tissues with low energy demand such as the adrenal, thyroid and lung, mitochondria contribute ~5% of the total mRNA, whereas in the energy-demanding heart, mitochondrial transcripts make up 30% of total mRNA [491]. Mitochondria contain a circular genome of 16,569 bases which encodes 37 genes: 13 genes for four out of the five subunits of the respiratory complexes, 22 tRNAs and 2rRNAs [491]. Under anaerobic conditions (e.g., cardiac ischaemia), the heart cannot maintain essential cellular processes and oxidative stress ensues. Mitochondria are thought to be particularly susceptible to damage from reactive oxygen species and, as they undergo apoptosis, they release damage-associated molecular patterns (DAMPs) and RNA that can trigger an inflammatory response [492]. Findings from this study suggest that an

increase in mitochondrial degradation due to oxidative stress in cardiomyocytes could result in an increase in circulating, cell-free mitochondrial RNA in patients with IHD. There was a good correlation between the circulating levels of mRNA in the plasma and in the human heart tissue data from Harvard (Figure 7-3) suggesting that gene expression in heart tissue may be reflected in the plasma, at least for some transcripts.

The fact that the majority of mitochondrial mRNAs that were found to be upregulated in both cardiac tissue and in plasma are encoded by the mitochondrial genome (as opposed to the nuclear encoding mitochondrial genes), could be partly because of the abundance of mitochondrial genomes. But it could also be telling us something about CDCS pathology. These 13 upregulated mitochondrial genes are genes that encode for four out of the five subunits of the respiratory complexes which are responsible for cellular metabolism, producing energy in the form of ATP in the presence of oxygen. Mitochondrial are major consumers of oxygen are severely affected by ischaemic conditions and sensitive to oxidative stress. Consistent fold change across all of the genes suggest they are being released in a co-ordinated manner, being exported from the cytoplasm either passively as cell damage occurs or actively as danger signals in response to the ischemia.

In support of these findings, intact, cell-free mitochondria have been found in human blood (proposed to be released by cells for signalling purposes) [493] and elevated levels of mitochondrial *DNA* have been shown in plasma post myocardial infarction and is a predictor of mortality in patients with acute coronary syndromes [494, 495]. An intriguing recent study found mitochondria to be translocated from cell-to-cell in both physiological and pathophysiological conditions [496]. Mitochondria from distressed cardiomyocytes acted as "danger-signal organelles", which triggered anti-apoptotic and mitochondrial biogenesis in mesenchymal stem cells. In this way, mesenchymal stem cells were able to 'donate' their mitochondria to injured cells to prevent oxidative stress injury [496]. Together these studies

suggest that mitochondria may be detectable in plasma and, as they become degraded, increase the levels of circulating cell-free mitochondrial RNA, as seen in the CDCS cohort.

Circulating levels of eight lncRNAs were higher and two lncRNAs were lower (logfc >1.2, padj<0.01) in the CDCS cohort compared to HVOLs. Interestingly, lncRNAs with higher expression in patients directly overlapped (n=7), or were very close to (7308 bases, n=1), CCCTC-binding transcription factor (CTCF)-binding sites (www.ensembl.org). CTCF is a transcription factor that, along with the cohesion complex, creates large loop domains within nuclear DNA called topologically associated domains (TADs). Contacts of DNA within each TAD are strong (whereas they are weak between TADS) and in this way enhancers and promoters are brought into close contact for gene regulation [497]. A genome-wide analysis of CTCF proteins found that two sub-classes exists: those with an RNA-binding region (RBR) and those without that form RBR-independent and RBR-dependent loops [498]. A number of lncRNAs have been shown to regulate CTCF [499-502]. While the overlapping of CTCF-binding sites raises the possibility that the lncRNAs identified in the current study may regulate CTCF binding, further analysis of these lncRNAs with techniques such as CLIP sequencing (where the RNA-protein complex is crosslinked and then immunoprecipitated, then the RNA is sequenced) are needed. Of the two lncRNAs that were downregulated in the CDCS cohort, one lncRNA, RNA component of mitochondrial RNA processing endoribonuclease (RMRP), has previously been reported to be upregulated in myocardial ischaemia and heart failure and is thought to be involved in cardiac fibrosis and apoptosis [211, 467, 503].

Not surprisingly, the novel lncRNA transcripts were seen at a much lower abundance than the annotated lncRNAs. Only two novel lncRNAs were detected in all 89 samples, suggesting the majority of lncRNA transcripts are on the limit of detection for this dataset. Encouragingly, two transcripts identified in plasma had the exact same intron match as two transcripts from the Harvard human heart tissue (which strongly suggests they are the same novel transcript),

although these two were not seen in the Nanopore validation experiment. While identification of the same novel transcript in two tissue types (heart and plasma) is encouraging, these transcripts should be validated by a second method such as RT-qPCR because of their very low abundance. Two novel transcripts from the plasma dataset were higher in CDCS patients compared with controls, suggesting they could have potential as novel markers for the presence of IHD.

To increase the reliability of *in silico* circRNA prediction, two circRNA detection algorithms were combined. This approach balances the loss of true positives against the removal of false positives [504]. A similar number of circRNAs were detected with CircExplorer2 and CircTools software (227 versus 205); however, CircTools appeared to duplicate 50 loci. Either there is an issue with the software not annotating reads to the correct strand or there are genuinely sense and antisense circRNAs being expressed at the same loci.

While this might be resolved by manually analysing the read sequences and alignments, for the purposes of this project, it was decided to focus on the circRNAs detected by both algorithms

Despite this issue, there was extremely good overlap between the two algorithms, with 180 circRNAs in common, suggesting they would be the prime candidates to validate as circulating RNA biomarkers. In addition, CircTools software identified a circRNA derived from the mitochondrial genome that was not detected by CircExplorer2 - likely due to the mitochondrial genome being filtered out by default for CircExplorer2. This circRNA was detected in all 89 samples and was the only significantly differentially expressed circRNA between the groups, (log_2 fold change of -2.03 in CDCS patients compared with healthy controls, padj < 1.65E-20). Interestingly, a study which compared the performance of 11 different circRNA detection software packages did not consider any circRNA candidates derived from mitochondria [385]. While it is possible that historically, circRNA candidates

that arose from mitochondria were considered as false positives, a recent paper identified circRNAs encoded by the mitochondrial genome in both human and mouse and validated these with RT-qPCR, Northern blots and FISH. They proposed a model where these circRNAs facilitate the entry of nuclear encoded proteins into the mitochondria [505]. A second study has identified four upregulated mitochondrial encoded circRNAs in plasma of Chronic Lymphocytic Leukaemia patients using microarray [506]. The mitochondrial circRNA detected in this study lies antisense to the mitochondrial *tRNAVal* gene, *MT-TV* (although this would need to be validated). Mutations in *MT-TV* have been associated with cardiomyopathies and severe heart failure [507].

The circRNA that was most significantly higher in the CDCS cohort compared to the HVOL cohort from both software analyses (padj < 0.05, log2 fold change 0.4 (circExplorer2) log2 fold change 0.5 (circTools)) was circUBAC2 (although again, for CircTools this was annotated to be on the antisense strand). This is consistent with a previous report showing circUBAC2 was higher in blood of MI patients compared to healthy controls [508], suggesting *circUBAC2* has potential as a circulating marker for the presence of stable IHD. This study has identified several ischaemia-related circRNA candidates for further study, with 3 and 5 circRNAs having padj <0.1 (CircExplorer2 and CircTools respectively), and with two of these identified by both software packages.

7.4 Limitations of study

The primary limitation was small sample size, which limited the ability to detect statistically significant differences between the groups, particularly between CDCS patients who did or did not subsequently develop heart failure. Additionally, the plasma samples had been in -80°C storage for several years and potentially some RNA degradation had occurred. Despite this, the study provides important proof-of-principle that RNA-Seq is possible in human plasma. Due to the low starting amounts of RNA approaches to enrich for human RNA without compromising an *a priori* strategy for detecting novel genes in needed which is

already possible for circRNAs [239, 241] This approach to enrich for circRNAs was not used in the current study as a combined pipeline for linear and circRNA detection was desired.

7.5 Conclusion

This study has demonstrated that mRNAs, lncRNAs and circRNAs can be detected in human plasma and has identified several promising candidates for biomarkers for the presence of IHD. These candidates were higher in patients that were sampled 4 months after the index admission (angina or MI) and if they can be shown to also be higher prior to the index event then they could be promising candidates for underlying atherosclerosis. These genes could then be used as a panel of markers for predicting who may be at imminent risk of MI.

The sequencing of RNA from human plasma is technically challenging due to the low input amounts of RNA and unavoidable contamination from bacterial nucleic acids from the RNA purification kits. Results from this study confirm previous work showing that a substantial percentage of reads are lost to these contaminating reads and/or a high degree of RNA degradation [453]. This contamination has impact on the percentage of reads mapping to the human genome and therefore the ability to detection circulating cell free RNAs by RNA-Seq. This could be mitigated through use of enrichment or targeted procedures but may defeat genome wide *a priori* analyses to find novel RNA biomarkers (although this is already possible for circRNAs with protocols that biochemically remove of linear transcripts [239-241]. RNA extraction and sequencing kits are constantly evolving as is the sequencing technology itself. This issue may be alleviated with the third-generation technologies which offer kits that bypass the need for a PCR amplification step (which may exacerbate the contamination problem) and sequence RNA molecules directly [509]. Currently high amounts of input RNA are needed for these techniques (e.g. 500 ng of starting RNA, equivalent to 5mLs of plasma) which would not be possible for most plasma studies.

Chapter 8
Discussion

8.1 Introduction

The overall aim of this thesis was to identify non-coding RNAs associated with IHD. This was accomplished with a bioinformatics pipeline identifying differentially expressed mRNAs, lncRNAs, including putative novel lncRNAs, and, for the plasma experiment, circRNAs also, using short-read RNA Sequencing data in two different datasets. A fold change of 1.2 was selected (which is lower than the standard 1.5 fold change in most analyses) as I hypothesised if a gene is significantly differentially expressed then it is worth analysing (I put a stricter adjusted p-value so the change I saw was reproducible). A subtle difference in gene expression may already have a substantial impact. Some genes are more dosage-sensitive than others, if it is regulating other genes rather than having a biological impact on its own, there may not need to be a large fold change. It is not known for sure what a biological relevant fold change for lncRNAs/circRNAs is. The thesis was ultimately exploratory for biomarkers but also, I wanted to identify any lncRNAs circRNAs that were biologically interesting with ischemia – to give us a better understanding of the underlying biology.

Several promising candidate biomarkers for myocardial ischaemia including several novel lncRNAs (validated with Nanopore sequencing) were identified in ischaemic human heart tissue. The subcellular localisation of three promising lncRNA candidates (two annotated lncRNAs, one novel lncRNA) was performed.

This pipeline was then applied to plasma from patients with IHD and healthy controls to screen for candidate mRNA, lncRNA and circRNA biomarkers for progression from IHD to HF. Although candidate biomarkers for disease progression could not be detected in these patients, several additional lncRNA and circRNA candidates for the presence of ischaemic heart disease were identified.

The first results chapter (Chapter 4) assessed the outcomes of pipeline validation on publicly available data and will be discussed first. I will then discuss the application of the pipeline to human myocardial ischaemic tissue data provided by Harvard Medical School (Chapter 5). Next, I will discuss the RNA-Seq protocol developed to detect RNA in plasma and the results of the subsequent screening study, which used the pipeline to explore candidate circulating RNA markers for the presence of IHD and progression to HF (Chapters 6 and 7). Finally, I will describe the limitations of the study and future directions.

8.2 The performance of the bioinformatic pipeline

A fundamental principle of bioinformatics is reproducibility: to be able to take raw data files and replicate the results of a previous analysis. In reality, this is rarely achieved as reference genomes, annotations and software are constantly updated. Moreover, multiple software packages are developed for the same task and so different permutations of software along a bioinformatic workflow can compound the discrepancy in results. With these considerations in mind, the high degree of concordance with previously published analyses (both in terms of gene detection and expression levels) suggested my pipeline was successfully integrating software for detection and mapping of mRNAs, lncRNAs and circRNAs and was quantifying read counts for each class of RNA accurately.

Despite the high degree of concordance, some differences were observed. My pipeline appeared to be slightly less sensitive at detecting annotated genes of low abundance, owing to more stringent minimum abundance filters. This was apparent when I removed an abundance filter when applying my pipeline to the Yang *et al* mRNA dataset (Section 4.4.1) which increased the detection overlap from 69% (with abundance filter) to 91% (without abundance filter). As mentioned previously, when first aligning the reads I used a ‘first-pass’ alignment with STAR2, which in effect detects all possible splice sites (including novels ones). These are then used as a guide for the ‘second-pass’ alignment. These splice-sites need to be filtered sensibly as there is a trade-off between detection of annotated transcripts against novel

transcripts. The more possible splice junctions, the more possible multi-mapping reads and the less uniquely mapping reads, which affects downstream gene quantification. This was a tricky trade off as my pipeline was trying to detect both annotated and novel transcripts. These splice sites did not affect the circRNA expression as the alignment and detection of circRNA reads are separate to linear reads.

Another factor affecting detection rates was the use of a more up-to-date version of the reference genome (GENCODE v29) in my pipeline. The reference genome annotation is constantly being updated, with genes being added, reclassified, or even removed between releases. A notable example of this was seen in the dataset from Yang *et al* where 827 genes detected by the authors were no longer present in GENCODE v29. What was most striking was that a large part of reproducible research depends on the quality of the annotation. This was further highlighted with the fact that *NORAD,* a relatively well characterised lncRNA, had differing chromosome co-ordinates in GENCODE compared to NONCODE. Annotations and databases need to be regularly maintained and nomenclature needs to be exactly matching across them. This was discussed in section 1.2.2 for lncRNAs and was a criticism highlighted by Vromman *et al [272]* for circRNAs databases.

8.3 LncRNAs associated with myocardial ischaemia

With the pipeline validated, the next part of the project was to analyse data obtained from human hearts pre- and post-ischaemia (Chapter 4). This data had already been used for identifying annotated mRNAs and lncRNAs associated with ischaemia [175] although whereas Saddic *et al* looked at neighbouring lncRNA-mRNA relationships, I placed more emphasis on detecting novel lncRNAs associated with ischaemia and used a gene network approach to explore associations between lncRNAs and mRNAs. The reason for this (as discussed in Section 1.2) was that lncRNAs can act in *trans* as well as *cis* and a gene correlation network analysis may help identify potential regulatory lncRNAs that influence ischaemia by acting on one or many mRNAs through networks (modules). I propose that

lncRNAs identified as module 'hubs' may make good candidates for further analysis with functional experiments such as knock out studies in cardiomyocytes.

With no abundance filter, the pipeline identified 10,567 transcripts as putative novel lncRNAs. The pipeline abundance filter was turned off for novel lncRNAs, as my strategy was to validate these transcripts with long read Nanopore sequencing. LncRNAs were expressed at a lower level than mRNAs (Table 5-1) [65, 108] and if I had used the mRNA filter of 90% of samples having at least 0.5 TPM this would have excluded 62% of the putative novel lncRNAs, albeit the lowly expressed ones. However, without some sort of abundance filter or validation strategy false positives will vastly outnumber true positives. It was encouraging to see that out of 39 transcripts initially identified as novel, 28 of these had been documented as lncRNAs in a more recent version of the GENCODE annotation, confirming the validity of my pipeline. The number of genuinely novel transcripts detected by Illumina sequencing but not detected by Nanopore sequencing and vice versa remains unknown, and it is likely additional novel transcripts will continue to be discovered in human heart as sequencing technologies improve.

It can be difficult to decide how to prioritise candidate novel biomarkers for follow up. When assessing the 20 most up- and down-regulated putative novel lncRNAs detected by Illumina sequencing (based on log2 fold change, Appendix C-3) there were some interesting candidates. If one assumes 'guilt by association' based on the function of the nearest annotated gene, some interesting candidates appear. These include syndecan-4 (*SDC4*) which has been associated with cardiac remodelling [510], Ankyrin repeat domain 1 (*ANKRD1*) which codes for the cardiac stress-response protein, cardiac adriamycin-responsive protein [511], Aquaporin-1 (*AQP1*) involved with myocardial remodelling [512] Tristetraprolin (*ZFP36*) associated with regulating mitochondrial function in heart [513], *ACTC1* which encodes the cardiac muscle alpha actin associated with dilated or hypertrophic cardiomyopathy [514] and Human α-protein kinase 3 (*ALPK3*) associated with

cardiomyopathy [515]. This approach assumes the novel lncRNAs may be acting as *cis*-regulatory factors. However, before prioritising candidate novel biomarkers for follow up functional studies, all putative novel transcripts require further replication in independent samples, either by RNA sequencing or RT-qPCR in an independent set of human left ventricle samples.

Once validated, novel lncRNAs may be prioritised for further study based on their regulatory potential. Of the 11 novel transcripts that were validated with Nanopore sequencing (and not documented in the more recent annotations), the majority overlapped enhancers or promoter regions, indicating a potential mechanism for their biological function. Additionally, one of the novel transcripts overlapped five eQTL SNPs for the *RWDD3* gene. An eQTL SNP is a DNA sequence variant that has been associated with the expression level of an mRNA (in this case *RWDD3*). My data suggests that these variants are present in the exons of this lncRNA which could affect the secondary structure of the lncRNA or a target binding site for a miRNA, mRNA or protein. A moderate correlation was seen between the novel lncRNA and *RWDD3*. To test whether there is a direct regulatory effect of this novel lncRNA on *RWDD3* then repeat experiments would be needed or knockout studies looking to see if *RWDD3* expression levels are affected when the novel lncRNA is inhibited. A second novel lncRNA, which was present in one of the ischaemia-associated network modules identified by WGCNA also overlapped an eQTL SNP. This novel lncRNA was associated with *VTCN1* in left ventricle (although this gene was not expressed at detectable levels in the dataset). The novel lncRNA was not only validated with Nanopore sequencing but also with RNAscope (Figure 5-12), which further corroborated the utility of the pipeline.

LncRNAs may also be prioritised based on their degree of correlation with other genes. Among the mRNA, lncRNA and novel lncRNA genes identified in human heart tissues, WGCNA identified 18 networks of correlated genes, of which 6 were moderately to strongly associated with ischaemia. The module most strongly associated with ischaemia had 2038

genes including 39 annotated lncRNAs and one novel lncRNA (mentioned above) that were markedly enriched in cell death, apoptosis and necrosis pathways. Three of the annotated lncRNAs appeared to be highly correlated with other genes in the module (module membership correlation >0.7) and may therefore serve as network 'hubs' by co-regulating multiple genes in the module. One of these three potential hub lncRNAs (*AC005523.2*) also overlapped an eQTL for *FEM1A* and this time a strong (Spearman) correlation of 0.87 was seen suggesting *AC005523.*2 may have potential to regulate *FEM1A* directly. *FEM1A* is localised within mitochondria of cardiac muscle and is increased in mouse hearts after MI [414] and after ischaemia reperfusion injury [516].

Two other lncRNAs in Module 1 overlapped eQTL SNPs linked to mRNAs associated with mitochondria, apoptosis and energy metabolism (*AC012313.1 - ZNF584 and AC011476.3 - RDH13*, which had a Spearman correlation of 0.8). Additionally, three out of the four mRNAs with the highest module membership correlation values, *OGDH*, *MFN2* and *MRPS27*, were also associated with mitochondrial processes. It is well documented that mitochondria are vital to cellular metabolism but also play a central role in apoptotic cell death [517] and are intricately involved in MI and cardio protection [518]. It is tempting to propose that this module of highly correlated genes may promote activation of cell death, apoptosis, and necrosis in ischaemic cardiomyocytes by modulating mitochondrial function. Of note, mitochondrial genes also featured in the analysis of plasma RNA in patients with IHD in Chapter 7.

The second module most strongly associated with ischaemia was linked to angiogenesis and vasculogenesis, two processes involved in the cardiomyocyte response to ischaemia [519]. Cardiomyocytes are highly susceptible to cell death caused by ischaemia/hypoxia and several attempts using targeted therapies to induce angiogenesis and reduce apoptosis in the border zone of the perfused and non-perfused areas of post MI hearts have been made [519]. This second module included 242 mRNAs and 8 lncRNAs, with the transcripts with the highest

module membership overwhelmingly associated with pro-angiogenic functions. Although none of the lncRNAs in this module overlapped eQTLs, several (including *AC093278.2*, *CARMN* and *AC007743.1*) were in close proximity to enhancer elements and CTCF binding sites raising the possibility they have functional roles. This network of potentially co-regulated mRNAs and lncRNAs appears to activate angiogenesis as part of the compensatory response to myocardial ischaemia.

In summary, WGCNA analysis enabled several thousand genes to be condensed into modules that were most strongly associated with ischaemia and with each other and may therefore be co-regulated. This enabled identification of genes (including lncRNAs and novel lncRNAs) that appear to be highly connected within biologically relevant modules ('hub' genes). These genes may drive the early ischaemic response in human left ventricle may be good candidates for future functional and biomarker studies.

8.4 The plasma transcriptome

Only a handful of studies have analysed RNA-Seq from human plasma [451-453, 455, 485] and, to the best of my knowledge, this is the first study to analyse the plasma transcriptome of an ischaemic heart disease cohort. A common theme to emerge from these studies is that the results can vary wildly depending on the kit (and even the sequencing centre) [485], and from which plasma fraction the RNA is extracted [451, 453, 485].

For one experiment, Everaert *et al* [453] sequenced platelet-rich plasma and platelet-free plasma (similar to the plasma samples in Chapter 7) from two healthy donors Interestingly, for the platelet-free plasma, the authors saw a similar percentage (~25%) of uniquely mapped reads with ~50% of reads being unable to be mapped due to being classified as too short although the authors did not investigate these reads further (assuming the RNA was fragmented), my data suggests the unmapped reads originate from contaminating bacterial transcripts in the kits. This issue appears to arise when the input amounts of endogenous RNA are so low that it is overcome by bacterial genetic material. In future RNA-Seq studies, it

would be pertinent to enrich for human RNA so as not to lose such a large percentage of reads during sequencing. However, this can become tricky if novel transcripts are sought as enrichment usually requires knowledge of the target sequences you would like to enrich for. The issue of having such a low amount of starting RNA also meant that Sequins could not be used, as the amount of starting endogenous human RNA concentrations could not be reliably estimated.

Another interesting observation from Everaert *et al* [453]and Savelyeva *et al* [451] is that different plasma fractions had very different percentages of reads derived from nuclear RNA or mitochondrial RNA (Table 8-1). The percentage of mitochondrial- and nuclear-derived reads from my data are presented in Table 8-1 for comparison. The highest source of mitochondrial RNA appears to be the platelets (from Everaert *et al* platelet rich plasma) ~75% reads compared to ~ 20% from platelet free plasma compared to ~40% of reads in my data. Related to this, Rodosthenous *et al* [485] also saw differences in the percentage of reads derived from different RNA species (e.g., mRNAs, lncRNAs) depending on which RNA sequencing library preparation kit was used and whether the RNA was extracellular vesicle rich or poor. Incidentally, of the six kits tested, the kit that detected the greatest gene diversity and largest number of lncRNAs was the kit used in the current project - Switching Mechanism At 5' end of RNA Template (SMARTer) v2 Pico.

These studies, along with my own, highlight a lack of method standardisation and reproducibility in plasma RNA-Seq. Different plasma preparation protocols and library preparation kits may enrich for different plasma fractions which contain different RNA populations. This directly impacts on the detection of certain RNAs.

A filter of >= 1 TPM for at least 90% of samples is perhaps strict and we would be missing genes expressed in one experimental group and completely absent from the other for most clinical biomarkers do not fit this pattern. It is preferred that biomarkers are detected in

everyone (albeit at low levels) but which are at much higher levels in cases compared to controls. It is preferable to show the assay is working in everyone, but that the levels go above a threshold in the disease group.

A surprising finding was that 13 out of the top 20 most differentially expressed genes that were detected at higher levels in patients with ischaemic heart disease than controls were mitochondrial genes (Section 7.2.2).

On consideration, it is perhaps not surprising that mitochondrial genes are among those most differentially expressed in patients with IHD, given that the heart is such an energy demanding organ.

Table 8-1 Percentage of reads from mitochondrial or nuclear RNA from heart tissue and plasma RNA-Seq data

	My data Heart Tissue	My data plasma	Everaert *et al* platelet rich plasma †	Everaert *et al* platelet free plasma †	Everaert *et al* platelet free plasma extracellular vesicles †	Savelyeva *et al* platelet-poor blood plasma *	Savelyeva *et al* Intermediate fraction containing membrane vesicles*	Savelyeva *et al* Fraction enriched with exosome *	Savelyeva *et al* Vesicle-depleted plasma *
% Mitochondrial	46	40	~75	~20	~1	14	18	3	8
% Nuclear RNA	54	60	~25	~80	~99	86	82	97	92

† Adapted from [453]

*Adapted from [451]

Mitochondria are the site of cellular metabolism, producing energy in the form of adenosine triphosphate (ATP) in the presence of oxygen. They are major consumers of oxygen and can be severely affected by ischaemic conditions, which alter the composition of the electron transport chain complexes [520] (of which four out of five are encoded by 13 mitochondrial genes).

Comparing genes that were differentially expressed *both* in human heart (pre- and post-ischaemia) and plasma (IHD patients versus controls), showed that 11 out of 15 genes (all upregulated) were mitochondrial. These data suggest that circulating mitochondrial RNA can be detected in plasma and may be elevated in association with myocardial ischaemia and in established ischaemic heart disease. Importantly, all mitochondrial RNA transcripts were highly abundant making them ideal candidate biomarkers. Further work, in larger cohorts, will be needed to confirm these preliminary findings and test whether one (or a combination of) mitochondrial mRNAs have potential to identify those people who may be at risk of an imminent heart attack. It is currently difficult to unequivocally say whether the genes we see present in the plasma are solely originating from the heart but from a biomarker perspective this is not always needed. BNP was first discovered from brain extracts and CRP is a general inflammation marker but they both act as biomarkers for heart disease. A good biomarker, when taken in a clinical context, is specific to the phenotype regardless of its tissue of origin.

Although the differentially expressed lncRNAs in plasma were not altered in ischaemic heart tissue, they may also be good candidates to follow up. All overlapped or were in very close proximity to CTCF binding sites, advocating a potential regulatory function.

Prioritising those transcripts that are most abundant, and multi-exonic (less likely to be false positives due to the detection of the same splice junction in several samples) and strongly differentially expressed would be a way to prioritise lncRNAs further for possible RT-qPCR experiments. Furthermore, the fact that two novel lncRNAs identified in plasma were also

detected in human heart was very encouraging despite the fact they were not differentially expressed. Further knock out experiments could be carried out in cardiomyocytes to explore their functional roles.

Not only was this project one of the first to perform RNA sequencing of mRNAs and lncRNAs in human plasma, to the best of my knowledge only one other study has investigated the 'circRNAome' using RNA sequencing in plasma (studying exosomes from gastric cancer patients) [521]. CircRNAs are potentially excellent biomarkers in biofluids as, due to their closed, circular structure, they are protected from RNase R degradation and exist stably in exosomes [522].

In this project, I used two circRNA detection software packages and prioritised circRNAs detected in both. There was good overlap between the software (180 circRNAs detected in both), but this could be because they both used the same back spliced junction reads (detected by the STAR aligner software) as input files. However, because circRNAs can only be detected using reads that align across the back spliced junction (rather than sequencing reads aligning to the whole transcript, as for their linear counterparts) only high abundance circRNAs will be detected. Even if the circRNA is derived from one exon we cannot unambiguously assign the reads that align to the 'body' to the circRNA as they could equally be originating from the linear transcript. By only counting the back spliced junction reads, the proportion of reads assigned to circRNA transcripts was a very small percentage of the total reads. This makes normalising to library size/read depth problematic, with the best strategy being to use the normalisation factor from the linear reads.

To address this problem, several methods have been developed to degrade linear transcripts and enrich for circRNAs [239] [241]. With these methods a more quantitative and qualitative analysis of circRNAs can be achieved. Full length circRNA sequences can be achieved with reads that align to the 'body' of the circRNA and adjoin the back spliced reads. However, as

my pipeline was set up to analyse both linear RNAs and circRNAs simultaneously, these methods could not be used. I also could not analyse the heart tissue data for circRNAs owing to the use of polyA selection during library preparation (circRNAs do not have poly-A tails). Despite not enriching for circRNAs, the pipeline robustly detected hundreds of circRNAs. No significantly differentially expressed circRNAs were identified when my conservative filter of p-adjusted < 0.01 was applied, but two promising candidates were detected by both software for a p-adjusted < 0.1. These candidates require confirmation in a larger cohort with RT-qPCR to confirm their expression. One validated, RNA immunoprecipitation (RIP) or pulldown assays could be used to identify proteins interacting with the circRNA. Most intriguingly, one of the software packages detected a highly differentially expressed mitochondrial circRNA in all of the samples (the other package did not detect mitochondrial circRNAs). Historically mitochondrial derived circRNAs may have been discarded as potential artefacts but several studies this year have reported the discovery of mitochondrial circRNAs [505, 506, 523]. Potentially, if this circRNA is confirmed experimentally, then not only would it be discovery of a novel biomarker associated with ischaemia but also discovery of a novel circRNA.

8.5 Limitations of the study and future directions.

For robust statistical measurements large enough sample numbers and ideally paired samples are desired. When analysing human tissue, sample numbers are usually limited either by costs or accessibility to patient samples – especially human heart tissue. For this thesis I had the rare opportunity to access paired human heart tissue samples. As the cohort sizes for both the heart tissues and plasma experiments were small (n=81 Harvard data, n=92 for the plasma experiment) the statistical power may not be sufficient to identify differentially expressed genes.

As RNA is actively transcribed, RNA-Seq has the advantage of providing a ‘real time’ insight into the ischaemic repsonse. It is a ‘snapshot’ of gene expression and the timing of sampling

reflects gene expression that is transient (unless the genes are continously up- or down-regulated). This was not so much an issue for the plasma experiment as bloods were taken on average 4 months after the acute event when the patients were stable (i.e., gene expression changes associated with the acute injury had passed and expression changes would more likely to be associated with either cardiac re-modelling or compensatory mechanisms). The ischaemic heart tissue (Harvard) data was more suscpetible to the importance of timing. Gene expression changes seen here (after a median of 74 minutes of ischaemia) reflect the immediate response to a relatively short, mild ischaemia and changes associated with prolonged ischaemia or a more severe response (e.g., after myocardial infarction) may have been missed.Also, it is possible some of the changes in gene expression may also be due to a cardioplegic cold response, rather than ischemia per se. Of course another limitation is that there is no perfect model for ischemia and perhaps there is a confounder effect by other apsects of coronary heart disease.

With the apparent contamination by bacteria from the kits which sequester a large proprtion of the reads it is possbile that we are still not sequencing deep enough to get a true picture of the plasma transcritpome. This may be being exacerbated by the large number of read also being taken up by mitochondria. If we could deplete the sample of bacterial contamination and the mitochondrial reads we would be left with many more reads resulting in more read diversity. This would enable us to have a better idea of which genes in the plasma are coming directly from the heart. Of course, we may not want to deplete the sample of the mitochondrial reads as these could be a true reflection of the underlying biology. Another way of perhaps overcoming the read enrichment issue would be to enrich for extra-cellular vesicles which would allow deep sequencing if this sub-population within plasma and would be a good future experiment. This would also help to overcome to issue of having to have such a large starting volume of plasma.

Candidates identified in this thesis worth following up are the upregulated mitochondrial genes, the lncRNAs that are in the modules most significantly associated with ischaemia identified by WGCNA – especially the ones suggested to be acting as 'hub' genes and the putative novel circRNA that was strongly downregulated in the plasma of CDCS patients. The fold change seen for some were relatively small but it was encouraging that they were consistently being changed in the same direction (strict adjusted p-values). The hope would be to improve areas under the curve (AUC) in Receiver Operating Characteristic (ROC) plots when combined with other markers as a panel.

For the ischaemic associated genes identified in this study, due to the sample numbers, this analysis acts as a screen with the requirement of validation of these in larger cohorts. Further validation of the candidate biomarkers would need to be carried out via qRT-PCR in larger cohorts. This is technically demanding due to the low amounts of starting RNA but there is emerging evidence that this is possible [524-526]. In addition, the tissue analysis used tissue blocks containing a heterogeneous combination of cells. In future, single cell RNA-Seq may enable lncRNA profiling of specific cardiac cell types. Finally, the tissue origins of the lncRNAs and circRNAs that were identified in plasma could be explored through regional sampling studies [527] or by induced pluripotent stem cells from cardiomyocytes where we could look for changes in the cells and media.

8.6 Concluding remarks

This study has established a bioinformatics pipeline and methodology for identifying and validating putative novel lncRNAs and circRNAs in human heart tissue and plasma. The pipeline was designed to identify mRNAs, lncRNAs, circRNAs and novel lncRNAs and detection of all four RNA classes was validated with published data. Several promising candidates from each of these classes have been identified for follow up studies not only for biomarker potential but also for functional studies to verify their functional roles and establish their potential as therapeutic targets in IHD.

In human heart tissue, eleven novel lncRNAs were discovered and network analysis identified several lncRNAs that may be acting as key regulators in the ischaemic heart. One of the novel lncRNAs was further validated with RNA Scope showing its subcellular localisation in both the nucleus and cytoplasm. In human plasma, novel candidate mRNA, lncRNA and circRNA biomarkers for IHD were discovered. While larger sample sizes may be needed to discover biomarkers for progression to HF, comparing data from ischaemic heart tissue and plasma from an IHD cohort may be a powerful way to prioritise candidate biomarkers for follow-up studies in larger cohorts and functional analysis.

This thesis demonstrates that cutting-edge bioinformatics analysis of RNA sequencing can be used to interrogate the whole transcriptome in human heart tissue and plasma in the discovery of candidate biomarkers for myocardial ischaemia and IHD.

References

1. Malakar, A.K., et al., *A review on coronary artery disease, its risk factors, and therapeutics.* J Cell Physiol, 2019. **234**(10): p. 16812-16823.
2. Health., M.o. *Mortality 2015 data tables*. 2018; Available from: https://www.health.govt.nz/publication/mortality-2015-data-tables
3. Tobias, M., Turley, M., *Health Loss in New Zealand: A report from the New Zealand Burden of Diseases, Injuries and Risk Factors Study 2006-2016.* 2013.
4. Grey, C., et al., *First and recurrent ischaemic heart disease events continue to decline in New Zealand, 2005-2015.* Heart, 2018. **104**(1): p. 51-57.
5. Robson, B., Harris, R.,, *Hauora: Mäori Standards of Health IV. A study of the years 2000-2005*, T.R.R.H.a.E. Pōmare, Editor. 2007: Wellington.
6. Organization., W.H., *Global Atlas on cardiovascular disease prevention and control*, S. Mendis, Puska, P., Norrving, B. , Editor. 2011, World Health Organization: Geneva.
7. Health., M.o. *Cardiovascular disease*. 2018; Available from: https://www.health.govt.nz/our-work/populations/maori-health/tatau-kahukura-maori-health-statistics/nga-mana-hauora-tutohu-health-status-indicators/cardiovascular-disease.
8. Cahill, T.J. and R.K. Kharbanda, *Heart failure after myocardial infarction in the era of primary percutaneous coronary intervention: Mechanisms, incidence and identification of patients at risk.* World Journal of Cardiology, 2017. **9**(5): p. 407-415.
9. Ponikowski, P., et al., *2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC). Developed with the special contribution of the Heart Failure Association (HFA) of the ESC.* Eur J Heart Fail, 2016. **18**(8): p. 891-975.
10. Yancy, C.W., et al., *2013 ACCF/AHA Guideline for the Management of Heart Failure: A Report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines.* Journal of the American College of Cardiology, 2013. **62**(16): p. e147-e239.
11. Velagaleti, R.S. and R.S. Vasan, *Heart failure in the twenty-first century: is it a coronary artery disease or hypertension problem?* Cardiol Clin, 2007. **25**(4): p. 487-95; v.
12. Savarese, G. and L.H. Lund, *Global Public Health Burden of Heart Failure.* Cardiac Failure Review, 2017. **3**(1): p. 7-11.
13. Heart Foundation. *NZ-led heart failure findings debunk world medical view*. 2018; Available from: https://www.heartfoundation.org.nz/about-us/news/media-releases/new-heart-failure-findings.
14. Badimon, L. and G. Vilahur, *Thrombosis formation on atherosclerotic lesions and plaque rupture.* J Intern Med, 2014. **276**(6): p. 618-32.
15. Libby, P., *Inflammation in atherosclerosis.* Arterioscler Thromb Vasc Biol, 2012. **32**(9): p. 2045-51.
16. Bentzon, J.F., et al., *Mechanisms of plaque formation and rupture.* Circ Res, 2014. **114**(12): p. 1852-66.
17. Santos-Gallego, C.G., B. Picatoste, and J.J. Badimón, *Pathophysiology of Acute Coronary Syndrome.* Current Atherosclerosis Reports, 2014. **16**(4): p. 401.

18. Frangogiannis, N.G., *Pathophysiology of Myocardial Infarction.* Compr Physiol, 2015. **5**(4): p. 1841-75.
19. Smolin, L.A., Grosvenor, Mary B., *Nutrition: Science and Applications*. 2019: John Wiley & Sons, Inc.,.
20. Fuster, V., et al., *Hurst's the heart.* 2017.
21. Lopaschuk, G.D., *AMP-activated protein kinase control of energy metabolism in the ischemic heart.* International Journal of Obesity, 2008. **32**(4): p. S29-S35.
22. Bergmann, O., et al., *Evidence for cardiomyocyte renewal in humans.* Science (New York, N.Y.), 2009. **324**(5923): p. 98-102.
23. Gabriel-Costa, D., *The pathophysiology of myocardial infarction-induced heart failure.* Pathophysiology, 2018.
24. Frangogiannis, N.G., *The inflammatory response in myocardial injury, repair and remodeling.* Nature reviews. Cardiology, 2014. **11**(5): p. 255-265.
25. Talman, V. and H. Ruskoaho, *Cardiac fibrosis in myocardial infarction—from repair and remodeling to regeneration.* Cell and Tissue Research, 2016. **365**(3): p. 563-581.
26. Prabhu, S.D. and N.G. Frangogiannis, *The Biological Basis for Cardiac Repair After Myocardial Infarction: From Inflammation to Fibrosis.* Circulation research, 2016. **119**(1): p. 91-112.
27. Anzai, T., *Post-infarction inflammation and left ventricular remodeling: a double-edged sword.* Circ J, 2013. **77**(3): p. 580-7.
28. Yellon, D.M. and D.J. Hausenloy, *Myocardial Reperfusion Injury.* New England Journal of Medicine, 2007. **357**(11): p. 1121-1135.
29. Azevedo, P.S., et al., *Cardiac Remodeling: Concepts, Clinical Impact, Pathophysiological Mechanisms and Pharmacologic Treatment.* Arquivos Brasileiros de Cardiologia, 2016. **106**(1): p. 62-69.
30. Kemp, C.D. and J.V. Conte, *The pathophysiology of heart failure.* Cardiovasc Pathol, 2012. **21**(5): p. 365-71.
31. Kalogeropoulos, A.P., et al., *Characteristics and outcomes of adult outpatients with heart failure and improved or recovered ejection fraction.* JAMA Cardiology, 2016. **1**(5): p. 510-518.
32. Chen, Y.T., et al., *Heart Failure with Reduced Ejection Fraction (HFrEF) and Preserved Ejection Fraction (HFpEF): The Diagnostic Value of Circulating MicroRNAs.* Cells, 2019. **8**(12).
33. Ho, J.E., et al., *Predicting Heart Failure With Preserved and Reduced Ejection Fraction: The International Collaboration on Heart Failure Subtypes.* Circulation. Heart failure, 2016. **9**(6): p. 10.1161/CIRCHEARTFAILURE.115.003116 e003116.
34. Bloom, M.W., et al., *Heart failure with reduced ejection fraction.* Nature Reviews Disease Primers, 2017. **3**: p. 17058.
35. Lilly, L.S., *Pathophysiology of heart disease : a collaborative project of medical students and faculty*. 2010, Philadelphia, Pa; London: Lippincott Williams & Wilkins.
36. Chatterjee, K., *Neurohormonal activation in congestive heart failure and the role of vasopressin.* Am J Cardiol, 2005. **95**(9a): p. 8b-13b.
37. Diez, J., *Chronic heart failure as a state of reduced effectiveness of the natriuretic peptide system: implications for therapy.* Eur J Heart Fail, 2017. **19**(2): p. 167-176.
38. Pervez, M.O., et al., *Prognostic and diagnostic significance of mid-regional pro-atrial natriuretic peptide in acute exacerbation of chronic obstructive*

pulmonary disease and acute heart failure: data from the ACE 2 Study. Biomarkers, 2018. **23**(7): p. 654-663.
39. Yasue, H., et al., *Localization and mechanism of secretion of B-type natriuretic peptide in comparison with those of A-type natriuretic peptide in normal subjects and patients with heart failure.* Circulation, 1994. **90**(1): p. 195-203.
40. Nadar, S.K. and M.M. Shaikh, *Biomarkers in Routine Heart Failure Clinical Care.* Card Fail Rev, 2019. **5**(1): p. 50-56.
41. Weber, M. and C. Hamm, *Role of B-type natriuretic peptide (BNP) and NT-proBNP in clinical routine.* Heart, 2006. **92**(6): p. 843-9.
42. Hall, C., *NT-ProBNP: the mechanism behind the marker.* J Card Fail, 2005. **11**(5 Suppl): p. S81-3.
43. Maisel, A.S., et al., *Rapid Measurement of B-Type Natriuretic Peptide in the Emergency Diagnosis of Heart Failure.* New England Journal of Medicine, 2002. **347**(3): p. 161-167.
44. Davis, M., et al., *Plasma brain natriuretic peptide in assessment of acute dyspnoea.* Lancet, 1994. **343**(8895): p. 440-4.
45. Januzzi, J.L., et al., *NT-proBNP testing for diagnosis and short-term prognosis in acute destabilized heart failure: an international pooled analysis of 1256 patients: the International Collaborative of NT-proBNP Study.* Eur Heart J, 2006. **27**(3): p. 330-7.
46. Masson, S., et al., *Direct comparison of B-type natriuretic peptide (BNP) and amino-terminal proBNP in a large population of patients with chronic and symptomatic heart failure: the Valsartan Heart Failure (Val-HeFT) data.* Clin Chem, 2006. **52**(8): p. 1528-38.
47. Zile, M.R., et al., *Prognostic Implications of Changes in N-Terminal Pro-B-Type Natriuretic Peptide in Patients With Heart Failure.* J Am Coll Cardiol, 2016. **68**(22): p. 2425-2436.
48. Khanam, S.S., et al., *Prognostic value of short-term follow-up BNP in hospitalized patients with heart failure.* BMC Cardiovasc Disord, 2017. **17**(1): p. 215.
49. Richards, A.M., et al., *B-type natriuretic peptides and ejection fraction for prognosis after myocardial infarction.* Circulation, 2003. **107**(22): p. 2786-92.
50. Coppola, G., et al., *Short term prognostic role of NT-proBNP in patients after myocardial infarction.* Minerva Cardioangiol, 2009. **57**(1): p. 13-21.
51. Krim, S.R., et al., *Racial/Ethnic differences in B-type natriuretic peptide levels and their association with care and outcomes among patients hospitalized with heart failure: findings from Get With The Guidelines-Heart Failure.* JACC Heart Fail, 2013. **1**(4): p. 345-352.
52. Gaggin, H.K. and J.L. Januzzi, Jr., *Biomarkers and diagnostics in heart failure.* Biochim Biophys Acta, 2013. **1832**(12): p. 2442-50.
53. Patibandla, S., K. Gupta, and K. Alsayouri, *Cardiac Enzymes*, in *StatPearls*. 2020, StatPearls Publishing
Copyright © 2020, StatPearls Publishing LLC.: Treasure Island (FL).
54. Daubert, M.A. and A. Jeremias, *The utility of troponin measurement to detect myocardial infarction: review of the current findings.* Vasc Health Risk Manag, 2010. **6**: p. 691-9.
55. Xue, Y., et al., *Serial changes in high-sensitive troponin I predict outcome in patients with decompensated heart failure.* Eur J Heart Fail, 2011. **13**(1): p. 37-42.
56. Berk, B.C., W.S. Weintraub, and R.W. Alexander, *Elevation of C-reactive protein in "active" coronary artery disease.* Am J Cardiol, 1990. **65**(3): p. 168-72.

57. Ridker, P.M., et al., *Inflammation, aspirin, and the risk of cardiovascular disease in apparently healthy men.* N Engl J Med, 1997. **336**(14): p. 973-9.
58. Pokharel, Y., et al., *High-sensitivity C-reactive protein levels and health status outcomes after myocardial infarction.* Atherosclerosis, 2017. **266**: p. 16-23.
59. Ridker, P.M., *From C-Reactive Protein to Interleukin-6 to Interleukin-1: Moving Upstream To Identify Novel Targets for Atheroprotection.* Circ Res, 2016. **118**(1): p. 145-56.
60. Kang, S., et al., *Relationship of High-Sensitivity C-Reactive Protein Concentrations and Systolic Heart Failure.* Curr Vasc Pharmacol, 2017. **15**(4): p. 390-396.
61. Araújo, J.P., et al., *Prognostic value of high-sensitivity C-reactive protein in heart failure: a systematic review.* J Card Fail, 2009. **15**(3): p. 256-66.
62. Gehlken, C., et al., *Galectin-3 in Heart Failure: An Update of the Last 3 Years.* Heart Fail Clin, 2018. **14**(1): p. 75-92.
63. Januzzi, J.L., Jr., et al., *Measurement of the interleukin family member ST2 in patients with acute dyspnea: results from the PRIDE (Pro-Brain Natriuretic Peptide Investigation of Dyspnea in the Emergency Department) study.* J Am Coll Cardiol, 2007. **50**(7): p. 607-13.
64. Yancy, C.W., et al., *2017 ACC/AHA/HFSA Focused Update of the 2013 ACCF/AHA Guideline for the Management of Heart Failure.* Journal of the American College of Cardiology, 2017. **70**(6): p. 776.
65. Derrien, T., et al., *The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression.* Genome Res, 2012. **22**.
66. Eissa, S., et al., *Evaluation of urinary miRNA-96 as a potential biomarker for bladder cancer diagnosis.* Med Oncol, 2015. **32**(1): p. 413.
67. Yin, Q., et al., *Elevated serum lncRNA TUG1 levels are a potential diagnostic biomarker of multiple myeloma.* Exp Hematol, 2019. **79**: p. 47-55.e2.
68. Huang, M., et al., *Circular RNA hsa_circ_0000745 may serve as a diagnostic marker for gastric cancer.* World J Gastroenterol, 2017. **23**(34): p. 6330-6338.
69. Ohno, S., *So much "junk" DNA in our genome.* Brookhaven Symp Biol, 1972. **23**: p. 366-70.
70. Eddy, S.R., *The C-value paradox, junk DNA and ENCODE.* Current Biology, 2012. **22**(21): p. R898-R899.
71. Hahn, M.W. and G.A. Wray, *The g-value paradox.* Evol Dev, 2002. **4**(2): p. 73-5.
72. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
73. Taft, R.J., M. Pheasant, and J.S. Mattick, *The relationship between non-protein-coding DNA and eukaryotic complexity.* Bioessays, 2007. **29**(3): p. 288-99.
74. Guttman, M., et al., *Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals.* Nature, 2009. **458**(7235): p. 223-7.
75. Carninci, P., et al., *The transcriptional landscape of the mammalian genome.* Science, 2005. **309**(5740): p. 1559-63.
76. Birney, E., et al., *Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.* Nature, 2007. **447**(7146): p. 799-816.
77. Dunham, I., et al., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.

78. Derrien, T., et al., *The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression.* Genome Res, 2012. **22**(9): p. 1775-89.
79. Morris, K.V. and J.S. Mattick, *The rise of regulatory RNA.* Nat Rev Genet, 2014. **15**(6): p. 423-37.
80. Gao, S., et al., *Two novel lncRNAs discovered in human mitochondrial DNA using PacBio full-length transcriptome data.* Mitochondrion, 2017.
81. Hardwick, S.A., et al., *Targeted, High-Resolution RNA Sequencing of Non-coding Genomic Regions Associated With Neuropsychiatric Functions.* Front Genet, 2019. **10**: p. 309.
82. Lagarde, J., et al., *High-throughput annotation of full-length long noncoding RNAs with Capture Long-Read Sequencing (CLS).* bioRxiv, 2017.
83. Huarte, M., *The emerging role of lncRNAs in cancer.* Nat Med, 2015. **21**(11): p. 1253-61.
84. Fenoglio, C., et al., *An emerging role for long non-coding RNA dysregulation in neurological disorders.* Int J Mol Sci, 2013. **14**(10): p. 20427-42.
85. Pullen, T.J. and G.A. Rutter, *Roles of lncRNAs in pancreatic beta cell identity and diabetes susceptibility.* Front Genet, 2014. **5**: p. 193.
86. He, J.H., Z.P. Han, and Y.G. Li, *Association between long non-coding RNA and human rare diseases (Review).* Biomed Rep, 2014. **2**(1): p. 19-23.
87. Hrdlickova, B., et al., *Genetic variation in the non-coding genome: Involvement of micro-RNAs and long non-coding RNAs in disease.* Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease, 2014. **1842**(10): p. 1910-1922.
88. Consortium, G. *Gencode v26*. 2017; Available from: https://www.gencodegenes.org/.
89. Iyer, M.K., et al., *The landscape of long noncoding RNAs in the human transcriptome.* Nat Genet, 2015. **47**(3): p. 199-208.
90. Volders, P.J., et al., *An update on LNCipedia: a database for annotated human lncRNA sequences.* Nucleic Acids Res, 2015. **43**(Database issue): p. D174-80.
91. Zhao, Y., et al., *NONCODE 2016: an informative and valuable data source of long non-coding RNAs.* Nucleic Acids Research, 2016. **44**(Database issue): p. D203-D208.
92. Moran, I., et al., *Human beta cell transcriptome analysis uncovers lncRNAs that are tissue-specific, dynamically regulated, and abnormally expressed in type 2 diabetes.* Cell Metab, 2012. **16**(4): p. 435-48.
93. Trimarchi, T., et al., *Genome-wide mapping and characterization of Notch-regulated long noncoding RNAs in acute leukemia.* Cell, 2014. **158**(3): p. 593-606.
94. Paralkar, V.R., et al., *Lineage and species-specific long noncoding RNAs during erythro-megakaryocytic development.* Blood, 2014. **123**(12): p. 1927-1937.
95. Yang, K.C., et al., *Deep RNA sequencing reveals dynamic regulation of myocardial noncoding RNAs in failing human heart and remodeling with mechanical circulatory support.* Circulation, 2014. **129**(9): p. 1009-21.
96. Bell, R.D., et al., *Identification and initial functional characterization of a human vascular cell-enriched long noncoding RNA.* Arterioscler Thromb Vasc Biol, 2014. **34**(6): p. 1249-59.
97. Xu, J., et al., *A comprehensive overview of lncRNA annotation resources.* Brief Bioinform, 2017. **18**(2): p. 236-249.

98. Alberti, A., et al., *Comparison of library preparation methods reveals their impact on interpretation of metatranscriptomic data.* BMC Genomics, 2014. **15**: p. 912.
99. Chakraborty, S., et al., *LncRBase: an enriched resource for lncRNA information.* PLoS One, 2014. **9**(9): p. e108010.
100. Quek, X.C., et al., *lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs.* Nucleic Acids Res, 2015. **43**(Database issue): p. D168-73.
101. Zhou, B., et al., *EVLncRNAs: a manually curated database for long non-coding RNAs validated by low-throughput experiments.* Nucleic Acids Research, 2017: p. gkx677-gkx677.
102. Anderson, D.M., et al., *A micropeptide encoded by a putative long noncoding RNA regulates muscle performance.* Cell, 2015. **160**(4): p. 595-606.
103. Cohen, S.M., *Everything old is new again: (linc)RNAs make proteins!* Embo j, 2014. **33**(9): p. 937-8.
104. Nelson, B.R., et al., *A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle.* Science, 2016. **351**(6270): p. 271-5.
105. Pueyo, J.I., E.G. Magny, and J.P. Couso, *New Peptides Under the s(ORF)ace of the Genome.* Trends Biochem Sci, 2016. **41**(8): p. 665-678.
106. Quinn, J.J. and H.Y. Chang, *Unique features of long non-coding RNA biogenesis and function.* Nat Rev Genet, 2016. **17**(1): p. 47-62.
107. Cabili, M.N., et al., *Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses.* Genes Dev, 2011. **25**(18): p. 1915-27.
108. Batista, P.J. and H.Y. Chang, *Long noncoding RNAs: cellular address codes in development and disease.* Cell, 2013. **152**(6): p. 1298-307.
109. De Santa, F., et al., *A large fraction of extragenic RNA pol II transcription sites overlap enhancers.* PLoS Biol, 2010. **8**(5): p. e1000384.
110. Novikova, I.V., S.P. Hennelly, and K.Y. Sanbonmatsu, *Structural architecture of the human long non-coding RNA, steroid receptor RNA activator.* Nucleic Acids Res, 2012. **40**(11): p. 5034-51.
111. Louro, R., et al., *Conserved tissue expression signatures of intronic noncoding RNAs transcribed from human and mouse loci.* Genomics, 2008. **92**(1): p. 18-25.
112. Lee, C. and N. Kikyo, *Strategies to identify long noncoding RNAs involved in gene regulation.* Cell & Bioscience, 2012. **2**(1): p. 37.
113. Zhang, X.O., et al., *Complementary sequence-mediated exon circularization.* Cell, 2014. **159**(1): p. 134-47.
114. Rinn, J.L. and H.Y. Chang, *Genome regulation by long noncoding RNAs.* Annu Rev Biochem, 2012. **81**: p. 145-66.
115. Clark, B.S. and S. Blackshaw, *Long non-coding RNA-dependent transcriptional regulation in neuronal development and disease.* Front Genet, 2014. **5**: p. 164.
116. Mattick, J.S. and J.L. Rinn, *Discovery and annotation of long noncoding RNAs.* Nat Struct Mol Biol, 2015. **22**(1): p. 5-7.
117. Wu, H., L. Yang, and L.L. Chen, *The Diversity of Long Noncoding RNAs and Their Generation.* Trends Genet, 2017.
118. Andersson, R., et al., *An atlas of active enhancers across human cell types and tissues.* Nature, 2014. **507**(7493): p. 455-61.

119. Villegas, V.E. and P.G. Zaphiropoulos, *Neighboring Gene Regulation by Antisense Long Non-Coding RNAs.* International Journal of Molecular Sciences, 2015. **16**(2): p. 3251-3266.
120. Katayama, S., et al., *Antisense transcription in the mammalian transcriptome.* Science, 2005. **309**(5740): p. 1564-6.
121. St Laurent, G., et al., *Intronic RNAs constitute the major fraction of the non-coding RNA in mammalian cells.* BMC Genomics, 2012. **13**: p. 504.
122. Bertone, P., et al., *Global identification of human transcribed sequences with genome tiling arrays.* Science, 2004. **306**(5705): p. 2242-6.
123. Kampa, D., et al., *Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22.* Genome Res, 2004. **14**(3): p. 331-42.
124. Chen, L.L., *Linking Long Noncoding RNA Localization and Function.* Trends Biochem Sci, 2016. **41**(9): p. 761-72.
125. Dykes, I.M. and C. Emanueli, *Transcriptional and Post-transcriptional Gene Regulation by Long Non-coding RNA.* Genomics Proteomics Bioinformatics, 2017.
126. Geisler, S. and J. Coller, *RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts.* Nature reviews. Molecular cell biology, 2013. **14**(11): p. 699-712.
127. Guttman, M. and J.L. Rinn, *Modular regulatory principles of large non-coding RNAs.* Nature, 2012. **482**(7385): p. 339-46.
128. Mishra, A. and R.D. Hawkins, *Three-dimensional genome architecture and emerging technologies: looping in disease.* Genome Medicine, 2017. **9**(1): p. 87.
129. West, J.A., et al., *The long noncoding RNAs NEAT1 and MALAT1 bind active chromatin sites.* Molecular cell, 2014. **55**(5): p. 791-802.
130. Engreitz, J.M., et al., *RNA-RNA Interactions Enable Specific Targeting of Noncoding RNAs to Nascent Pre-mRNAs and Chromatin Sites.* Cell, 2014. **159**(1): p. 188-199.
131. Wang, K.C., et al., *Long noncoding RNA programs active chromatin domain to coordinate homeotic gene activation.* Nature, 2011. **472**(7341): p. 120-124.
132. Helbig, R. and F.O. Fackelmayer, *Scaffold attachment factor A (SAF-A) is concentrated in inactive X chromosome territories through its RGG domain.* Chromosoma, 2003. **112**(4): p. 173-82.
133. McHugh, C.A., et al., *The Xist lncRNA directly interacts with SHARP to silence transcription through HDAC3.* Nature, 2015. **521**(7551): p. 232-236.
134. Hasegawa, Y., et al., *The Matrix Protein hnRNP U Is Required for Chromosomal Localization of Xist RNA.* Developmental Cell, 2010. **19**(3): p. 469-476.
135. Engreitz, J.M., et al., *The Xist lncRNA Exploits Three-Dimensional Genome Architecture to Spread Across the X Chromosome.* Science, 2013. **341**(6147).
136. Simon, M.D., et al., *High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation.* Nature, 2013. **504**(7480): p. 465-469.
137. Wang, K.C. and H.Y. Chang, *Transcription coactivator and lncRNA duet evoke Hox genes.* PLoS Genet, 2017. **13**(6): p. e1006797.
138. Engreitz, J.M., N. Ollikainen, and M. Guttman, *Long non-coding RNAs: spatial amplifiers that control nuclear structure and gene expression.* Nat Rev Mol Cell Biol, 2016. **17**(12): p. 756-770.

139. Guttman, M., et al., *Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs.* Nat Biotechnol, 2010. **28**.
140. da Rocha, S.T., et al., *Jarid2 Is Implicated in the Initial Xist-Induced Targeting of PRC2 to the Inactive X Chromosome.* Mol Cell, 2014. **53**(2): p. 301-16.
141. Wutz, A., T.P. Rasmussen, and R. Jaenisch, *Chromosomal silencing and localization are mediated by different domains of Xist RNA.* Nat Genet, 2002. **30**(2): p. 167-74.
142. Tsai, M.C., et al., *Long noncoding RNA as modular scaffold of histone modification complexes.* Science, 2010. **329**(5992): p. 689-93.
143. Hajjari, M. and A. Salavaty, *HOTAIR: an oncogenic long non-coding RNA in different cancers.* Cancer Biology & Medicine, 2015. **12**(1): p. 1-9.
144. Wu, Y., et al., *Long noncoding RNA HOTAIR involvement in cancer.* Tumour Biol, 2014. **35**(10): p. 9531-8.
145. Hacisuleyman, E., et al., *Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre.* Nat Struct Mol Biol, 2014. **21**(2): p. 198-206.
146. Sun, L., et al., *Long noncoding RNAs regulate adipogenesis.* Proc Natl Acad Sci U S A, 2013. **110**(9): p. 3387-92.
147. Clemson, C.M., et al., *An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles.* Mol Cell, 2009. **33**(6): p. 717-26.
148. Imamura, K., et al., *Long noncoding RNA NEAT1-dependent SFPQ relocation from promoter region to paraspeckle mediates IL8 expression upon immune stimuli.* Mol Cell, 2014. **53**(3): p. 393-406.
149. Shevtsov, S.P. and M. Dundr, *Nucleation of nuclear bodies by RNA.* Nat Cell Biol, 2011. **13**(2): p. 167-173.
150. van Heesch, S., et al., *Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes.* Genome Biology, 2014. **15**(1): p. R6-R6.
151. Ulitsky, I. and David P. Bartel, *lincRNAs: Genomics, Evolution, and Mechanisms.* Cell, 2013. **154**(1): p. 26-46.
152. Gong, C. and L.E. Maquat, *lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements.* Nature, 2011. **470**(7333): p. 284-8.
153. Faghihi, M.A., et al., *Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase.* Nat Med, 2008. **14**(7): p. 723-30.
154. Chu, C.Y. and T.M. Rana, *Translation repression in human cells by microRNA-induced gene silencing requires RCK/p54.* PLoS Biol, 2006. **4**(7): p. e210.
155. Carrieri, C., et al., *Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat.* Nature, 2012. **491**(7424): p. 454-7.
156. Wang, J., J. Sun, and F. Yang, *The role of long non-coding RNA H19 in breast cancer.* Oncol Lett, 2020. **19**(1): p. 7-16.
157. Dey, B.K., K. Pfeifer, and A. Dutta, *The H19 long noncoding RNA gives rise to microRNAs miR-675-3p and miR-675-5p to promote skeletal muscle differentiation and regeneration.* Genes Dev, 2014. **28**(5): p. 491-501.
158. Espinosa, J.M., *Revisiting lncRNAs: How Do You Know Yours Is Not an eRNA?* Mol Cell, 2016. **62**(1): p. 1-2.

159. Li, W., D. Notani, and M.G. Rosenfeld, *Enhancers as non-coding RNA transcription units: recent insights and future perspectives.* Nat Rev Genet, 2016. **17**(4): p. 207-23.
160. Kim, T.K., et al., *Widespread transcription at neuronal activity-regulated enhancers.* Nature, 2010. **465**(7295): p. 182-7.
161. Consortium, T.E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**.
162. Arner, E., et al., *Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells.* Science, 2015. **347**(6225): p. 1010.
163. Schaukowitch, K., et al., *Enhancer RNA facilitates NELF release from immediate early genes.* Mol Cell, 2014. **56**(1): p. 29-42.
164. Soutourina, J., *Transcription regulation by the Mediator complex.* Nat Rev Mol Cell Biol, 2018. **19**(4): p. 262-274.
165. Li, W., et al., *Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation.* Nature, 2013. **498**(7455): p. 516-20.
166. The Wellcome Trust Case Control, C., *Genome-wide association study of copy number variation in 16,000 cases of eight common diseases and 3,000 shared controls.* Nature, 2010. **464**(7289): p. 713-720.
167. McPherson, R., et al., *A common allele on chromosome 9 associated with coronary heart disease.* Science, 2007. **316**(5830): p. 1488-91.
168. Helgadottir, A., et al., *A common variant on chromosome 9p21 affects the risk of myocardial infarction.* Science, 2007. **316**(5830): p. 1491-3.
169. Samani, N.J., et al., *Genomewide association analysis of coronary artery disease.* N Engl J Med, 2007. **357**(5): p. 443-53.
170. Holdt, L.M., et al., *ANRIL expression is associated with atherosclerosis risk at chromosome 9p21.* Arterioscler Thromb Vasc Biol, 2010. **30**(3): p. 620-7.
171. Congrains, A., et al., *Genetic variants at the 9p21 locus contribute to atherosclerosis through modulation of ANRIL and CDKN2A/B.* Atherosclerosis, 2012. **220**(2): p. 449-55.
172. Tsai, P.C., et al., *Additive effect of ANRIL and BRAP polymorphisms on ankle-brachial index in a Taiwanese population.* Circ J, 2012. **76**(2): p. 446-52.
173. Holdt, L.M. and D. Teupser, *Recent studies of the human chromosome 9p21 locus, which is associated with atherosclerosis in human populations.* Arterioscler Thromb Vasc Biol, 2012. **32**(2): p. 196-206.
174. Holdt, L.M., et al., *Alu elements in ANRIL non-coding RNA at chromosome 9p21 modulate atherogenic cell functions through trans-regulation of gene networks.* PLoS Genet, 2013. **9**(7): p. e1003588.
175. Saddic, L.A., et al., *The Long Noncoding RNA Landscape of the Ischemic Human Left Ventricle.* Circ Cardiovasc Genet, 2017. **10**(1).
176. Zhang, Z., et al., *Increased plasma levels of lncRNA H19 and LIPCAR are associated with increased risk of coronary artery disease in a Chinese population.* Scientific Reports, 2017. **7**(1): p. 7491.
177. Liu, L., et al., *The H19 long noncoding RNA is a novel negative regulator of cardiomyocyte hypertrophy.* Cardiovasc Res, 2016. **111**(1): p. 56-65.
178. Michalik, K.M., et al., *Long noncoding RNA MALAT1 regulates endothelial cell function and vessel growth.* Circ Res, 2014. **114**(9): p. 1389-97.
179. Tang, Y., et al., *The lncRNA MALAT1 protects the endothelium against ox-LDL-induced dysfunction via upregulating the expression of the miR-22-3p target genes CXCR2 and AKT.* FEBS Letters, 2015. **589**(20): p. 3189-3196.
180. Liu, J.Y., et al., *Pathogenic role of lncRNA-MALAT1 in endothelial cell dysfunction in diabetes mellitus.* Cell Death Dis, 2014. **5**: p. e1506.

181. Mao, Y.S., B. Zhang, and D.L. Spector, *Biogenesis and function of nuclear bodies.* Trends Genet, 2011. **27**(8): p. 295-306.
182. Min, L., et al., *Antidifferentiation Noncoding RNA Regulates the Proliferation of Osteosarcoma Cells.* Cancer Biother Radiopharm, 2016. **31**(2): p. 52-7.
183. Zangrando, J., et al., *Identification of candidate long non-coding RNAs in response to myocardial infarction.* BMC Genomics, 2014. **15**: p. 460.
184. Vausort, M., D.R. Wagner, and Y. Devaux, *Long noncoding RNAs in patients with acute myocardial infarction.* Circ Res, 2014. **115**(7): p. 668-77.
185. Ishii, N., et al., *Identification of a novel non-coding RNA, MIAT, that confers risk of myocardial infarction.* J Hum Genet, 2006. **51**(12): p. 1087-99.
186. Yan, B., et al., *lncRNA-MIAT regulates microvascular dysfunction by functioning as a competing endogenous RNA.* Circ Res, 2015. **116**(7): p. 1143-56.
187. Qu, X., et al., *MIAT Is a Pro-fibrotic Long Non-coding RNA Governing Cardiac Fibrosis in Post-infarct Myocardium.* 2017. **7**: p. 42657.
188. Zhu, X.H., et al., *LncRNA MIAT enhances cardiac hypertrophy partly through sponging miR-150.* Eur Rev Med Pharmacol Sci, 2016. **20**(17): p. 3653-60.
189. Li, X., et al., *Down-regulation of lncRNA KCNQ1OT1 protects against myocardial ischemia/reperfusion injury following acute myocardial infarction.* Biochemical and Biophysical Research Communications, 2017. **491**(4): p. 1026-1033.
190. Kurreck, J., *Antisense technologies. Improvement through novel chemical modifications.* Eur J Biochem, 2003. **270**(8): p. 1628-44.
191. Parasramka, M.A., et al., *Long non-coding RNAs as novel targets for therapy in hepatocellular carcinoma.* Pharmacology & Therapeutics, 2016. **161**: p. 67-78.
192. Maeder, M.L., et al., *CRISPR RNA-guided activation of endogenous human genes.* Nat Meth, 2013. **10**(10): p. 977-979.
193. Hewson, C., et al., *Extracellular vesicle associated long non-coding RNAs functionally enhance cell viability.* Non-coding RNA Research, 2016. **1**(1): p. 3-11.
194. Dong, L., et al., *Circulating Long RNAs in Serum Extracellular Vesicles: Their Characterization and Potential Application as Biomarkers for Diagnosis of Colorectal Cancer.* Cancer Epidemiol Biomarkers Prev, 2016. **25**(7): p. 1158-66.
195. Eissa, S., et al., *Rapid detection of urinary long non-coding RNA urothelial carcinoma associated one using a PCR-free nanoparticle-based assay.* Biomarkers, 2015. **20**(3): p. 212-7.
196. Holdt, L.M., et al., *Circular non-coding RNA ANRIL modulates ribosomal RNA maturation and atherosclerosis in humans.* Nat Commun, 2016. **7**: p. 12429.
197. Zhou, X., et al., *Long non-coding RNA ANRIL regulates inflammatory responses as a novel component of NF-κB pathway.* RNA Biology, 2016. **13**(1): p. 98-108.
198. Ballantyne, M.D., et al., *Smooth Muscle Enriched Long Noncoding RNA (SMILR) Regulates Cell Proliferation.* Circulation, 2016. **133**(21): p. 2050-2065.
199. Yang, Y., et al., *Plasma long non-coding RNA, CoroMarker, a novel biomarker for diagnosis of coronary artery disease.* Clin Sci (Lond), 2015. **129**(8): p. 675-85.
200. Uchida, S. and S. Dimmeler, *Long Noncoding RNAs in Cardiovascular Diseases.* Circulation Research, 2015. **116**(4): p. 737.

201. Xuan, L., et al., *Circulating long non-coding RNAs NRON and MHRT as novel predictive biomarkers of heart failure.* J Cell Mol Med, 2017.
202. Yan, Y., et al., *Circulating Long Noncoding RNA UCA1 as a Novel Biomarker of Acute Myocardial Infarction.* Biomed Res Int, 2016. **2016**: p. 8079372.
203. Wang, K., et al., *APF lncRNA regulates autophagy and myocardial infarction by targeting miR-188-3p.* 2015. **6**: p. 6779.
204. Wang, K., et al., *CARL lncRNA inhibits anoxia-induced mitochondrial fission and apoptosis in cardiomyocytes by impairing miR-539-dependent PHB2 downregulation.* Nat Commun, 2014. **5**: p. 3596.
205. Li, X., J. Zhou, and K. Huang, *Inhibition of the lncRNA Mirt1 Attenuates Acute Myocardial Infarction by Suppressing NF-κB Activation.* Cellular Physiology and Biochemistry, 2017. **42**(3): p. 1153-1164.
206. Wang, K., et al., *The long noncoding RNA NRF regulates programmed necrosis and myocardial injury during ischemia and reperfusion by targeting miR-873.* Cell Death and Differentiation, 2016. **23**(8): p. 1394-1405.
207. Wang, Z., et al., *The long noncoding RNA Chaer defines an epigenetic checkpoint in cardiac hypertrophy.* Nat Med, 2016. **22**(10): p. 1131-1139.
208. Viereck, J., et al., *Long noncoding RNA Chast promotes cardiac remodeling.* Science Translational Medicine, 2016. **8**(326): p. 326ra22.
209. Wang, K., et al., *The long noncoding RNA CHRF regulates cardiac hypertrophy by targeting miR-489.* Circ Res, 2014. **114**(9): p. 1377-88.
210. Tao, H., et al., *LncRNA GAS5 controls cardiac fibroblast activation and fibrosis by targeting miR-21 via PTEN/MMP-2 signaling pathway.* Toxicology, 2017. **386**: p. 11-18.
211. Greco, S., et al., *Long noncoding RNA dysregulation in ischemic heart failure.* Journal of Translational Medicine, 2016. **14**: p. 183.
212. Han, P., et al., *A long noncoding RNA protects the heart from pathological hypertrophy.* Nature, 2014. **514**(7520): p. 102-6.
213. Jiang, F., X. Zhou, and J. Huang, *Long Non-Coding RNA-ROR Mediates the Reprogramming in Cardiac Hypertrophy.* PLoS One, 2016. **11**(4): p. e0152767.
214. Micheletti, R., et al., *The long noncoding RNA Wisper controls cardiac fibrosis and remodeling.* Sci Transl Med, 2017. **9**(395).
215. Kumarswamy, R., et al., *Circulating long noncoding RNA, LIPCAR, predicts survival in patients with heart failure.* Circ Res, 2014. **114**(10): p. 1569-75.
216. Nigro, J.M., et al., *Scrambled exons.* Cell, 1991. **64**(3): p. 607-13.
217. Cocquerelle, C., et al., *Splicing with inverted order of exons occurs proximal to large introns.* The EMBO Journal, 1992. **11**(3): p. 1095-1098.
218. Jeck, W.R., et al., *Circular RNAs are abundant, conserved, and associated with ALU repeats.* Rna, 2013. **19**(2): p. 141-57.
219. Yuan, C., et al., *EMT related circular RNA expression profiles identify circSCYL2 as a novel molecule in breast tumor metastasis.* Int J Mol Med, 2020. **45**(6): p. 1697-1710.
220. Fu, X., et al., *Circular RNA MAN2B2 promotes cell proliferation of hepatocellular carcinoma cells via the miRNA-217/MAPK1 axis.* J Cancer, 2020. **11**(11): p. 3318-3326.
221. Akhter, R., *Circular RNA and Alzheimer's Disease.* Adv Exp Med Biol, 2018. **1087**: p. 239-243.
222. Rybak-Wolf, A., et al., *Circular RNAs in the Mammalian Brain Are Highly Abundant, Conserved, and Dynamically Expressed.* Mol Cell, 2015. **58**(5): p. 870-85.

223. Ambrose, J.A. and M. Singh, *Pathophysiology of coronary artery disease leading to acute coronary syndromes.* F1000Prime Rep, 2015. **7**: p. 08.
224. Zhang, Y., et al., *Circular Intronic Long Noncoding RNAs.* Molecular Cell, 2013. **51**(6): p. 792-806.
225. Bachmayr-Heyda, A., et al., *Correlation of circular RNA abundance with proliferation – exemplified with colorectal and ovarian cancer, idiopathic lung fibrosis, and normal human tissues.* 2015. **5**: p. 8057.
226. Vilades, D., et al., *Plasma circular RNA hsa_circ_0001445 and coronary artery disease: Performance as a biomarker.* Faseb j, 2020. **34**(3): p. 4403-4414.
227. Song, Z., et al., *Identification of urinary hsa_circ _0137439 as potential biomarker and tumor regulator of bladder cancer.* Neoplasma, 2020. **67**(1): p. 137-146.
228. Fanale, D., et al., *Circular RNA in Exosomes.* Adv Exp Med Biol, 2018. **1087**: p. 109-117.
229. Tan, W.L., et al., *A landscape of circular RNA expression in the human heart.* Cardiovasc Res, 2017. **113**(3): p. 298-309.
230. Zhang, Y., et al., *The Biogenesis of Nascent Circular RNAs.* Cell Rep, 2016. **15**(3): p. 611-624.
231. Starke, S., et al., *Exon circularization requires canonical splice signals.* Cell Rep, 2015. **10**(1): p. 103-11.
232. Ashwal-Fluss, R., et al., *circRNA biogenesis competes with pre-mRNA splicing.* Mol Cell, 2014. **56**(1): p. 55-66.
233. Jeck, W.R. and N.E. Sharpless, *Detecting and characterizing circular RNAs.* Nat Biotechnol, 2014. **32**(5): p. 453-61.
234. Ivanov, A., et al., *Analysis of intron sequences reveals hallmarks of circular RNA biogenesis in animals.* Cell Rep, 2015. **10**(2): p. 170-7.
235. Conn, S.J., et al., *The RNA binding protein quaking regulates formation of circRNAs.* Cell, 2015. **160**(6): p. 1125-34.
236. Errichelli, L., et al., *FUS affects circular RNA expression in murine embryonic stem cell-derived motor neurons.* Nat Commun, 2017. **8**: p. 14741.
237. You, X. and T.O.F. Conrad, *Acfs: accurate circRNA identification and quantification from RNA-Seq data.* Scientific Reports, 2016. **6**(1): p. 38820.
238. Szabo, L. and J. Salzman, *Detecting circular RNAs: bioinformatic and experimental challenges.* Nat Rev Genet, 2016. **17**(11): p. 679-692.
239. Pandey, P.R., et al., *RPAD (RNase R treatment, polyadenylation, and poly(A)+ RNA depletion) method to isolate highly pure circular RNA.* Methods, 2019. **155**: p. 41-48.
240. Xiao, M.S. and J.E. Wilusz, *An improved method for circular RNA purification using RNase R that efficiently removes linear RNAs containing G-quadruplexes or structured 3' ends.* Nucleic Acids Res, 2019. **47**(16): p. 8755-8769.
241. Panda, A.C., et al., *High-purity circular RNA isolation method (RPAD) reveals vast collection of intronic circRNAs.* Nucleic Acids Research, 2017. **45**(12): p. e116-e116.
242. Zhang, X.O., et al., *Diverse alternative back-splicing and alternative splicing landscape of circular RNAs.* Genome Res, 2016. **26**(9): p. 1277-87.
243. Ma, X.K., et al., *CIRCexplorer3: A CLEAR Pipeline for Direct Comparison of Circular and Linear RNA Expression.* Genomics Proteomics Bioinformatics, 2019. **17**(5): p. 511-521.
244. Gao, Y., J. Wang, and F. Zhao, *CIRI: an efficient and unbiased algorithm for de novo circular RNA identification.* Genome Biol, 2015. **16**: p. 4.

245. Gao, Y., J. Zhang, and F. Zhao, *Circular RNA identification based on multiple seed matching.* Brief Bioinform, 2018. **19**(5): p. 803-810.
246. Cheng, J., F. Metge, and C. Dieterich, *Specific identification and quantification of circular RNAs from sequencing data.* Bioinformatics, 2016. **32**(7): p. 1094-6.
247. Memczak, S., et al., *Circular RNAs are a large class of animal RNAs with regulatory potency.* Nature, 2013. **495**(7441): p. 333-8.
248. Wang, K., et al., *MapSplice: accurate mapping of RNA-seq reads for splice junction discovery.* Nucleic Acids Res, 2010. **38**(18): p. e178.
249. Hoffmann, S., et al., *A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection.* Genome Biol, 2014. **15**(2): p. R34.
250. Song, X., et al., *Circular RNA profile in gliomas revealed by identification tool UROBORUS.* Nucleic Acids Res, 2016. **44**(9): p. e87.
251. Jia, G.Y., et al., *CircRNAFisher: a systematic computational approach for de novo circular RNA identification.* Acta Pharmacol Sin, 2019. **40**(1): p. 55-63.
252. Szabo, L., et al., *Statistically based splicing detection reveals neural enrichment and tissue-specific induction of circular RNA during human fetal development.* Genome Biol, 2015. **16**: p. 126.
253. Menzel P, M.I., *BIQ: a method for searching circular RNAs in transcriptome databases by indexing backsplice junctions.* bioRxiv, 2019.
254. Li, X. and Y. Wu, *Detecting circular RNA from high-throughput sequence data with de Bruijn graph.* BMC Genomics, 2020. **21**(1): p. 749.
255. Li, X., et al., *CircMarker: a fast and accurate algorithm for circular RNA detection.* BMC Genomics, 2018. **19**(6): p. 572.
256. Sekar, S., et al., *ACValidator: A novel assembly-based approach for in silico verification of circular RNAs.* Biol Methods Protoc, 2020. **5**(1): p. bpaa010.
257. Sun, P. and G. Li, *CircCode: A Powerful Tool for Identifying circRNA Coding Ability.* Frontiers in Genetics, 2019. **10**(981).
258. Gaffo, E., et al., *CirComPara: A Multi-Method Comparative Bioinformatics Pipeline to Detect and Study circRNAs from RNA-seq Data.* Noncoding RNA, 2017. **3**(1).
259. Zhong, S., et al., *CircPrimer: a software for annotating circRNAs and determining the specificity of circRNA primers.* BMC Bioinformatics, 2018. **19**(1): p. 292.
260. Meng, X., et al., *CircPro: an integrated tool for the identification of circRNAs with protein-coding potential.* Bioinformatics, 2017. **33**(20): p. 3314-3316.
261. Li, L., D. Bu, and Y. Zhao, *CircRNAwrap - a flexible pipeline for circRNA identification, transcript prediction, and abundance estimation.* FEBS Lett, 2019. **593**(11): p. 1179-1189.
262. Jakobi, T., A. Uvarovskii, and C. Dieterich, *circtools-a one-stop software solution for circular RNA research.* Bioinformatics, 2019. **35**(13): p. 2326-2328.
263. Andrés-León, E., R. Núñez-Torres, and A.M. Rojas, *miARma-Seq: a comprehensive tool for miRNA, mRNA and circRNA analysis.* Scientific Reports, 2016. **6**(1): p. 25749.
264. Chen, C.Y. and T.J. Chuang, *NCLcomparator: systematically post-screening non-co-linear transcripts (circular, trans-spliced, or fusion RNAs) identified from various detectors.* BMC Bioinformatics, 2019. **20**(1): p. 3.
265. Zhao, J., et al., *ReCirc: prediction of circRNA expression and function through probe reannotation of non-circRNA microarrays.* Mol Omics, 2019. **15**(2): p. 150-163.
266. Humphreys, D.T., et al., *Ularcirc: visualization and enhanced analysis of circular RNAs via back and canonical forward splicing.* Nucleic Acids Res, 2019. **47**(20): p. e123.

267. Gao, Y. and F. Zhao, *Computational Strategies for Exploring Circular RNAs.* Trends Genet, 2018. **34**(5): p. 389-400.
268. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.
269. Kim, D., B. Langmead, and S.L. Salzberg, *HISAT: a fast spliced aligner with low memory requirements.* Nat Methods, 2015. **12**(4): p. 357-60.
270. Kim, D., et al., *TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.* Genome Biol, 2013. **14**(4): p. R36.
271. Horiuchi, T. and T. Aigaki, *Alternative trans-splicing: a novel mode of pre-mRNA processing.* Biol Cell, 2006. **98**(2): p. 135-40.
272. Vromman, M., J. Vandesompele, and P.J. Volders, *Closing the circle: current state and perspectives of circular RNA databases.* Brief Bioinform, 2020.
273. Hansen, T.B., et al., *Natural RNA circles function as efficient microRNA sponges.* Nature, 2013. **495**(7441): p. 384-388.
274. Piwecka, M., et al., *Loss of a mammalian circular RNA locus causes miRNA deregulation and affects brain function.* Science, 2017. **357**(6357): p. eaam8526.
275. Hansen, T.B., et al., *miRNA-dependent gene silencing involving Ago2-mediated cleavage of a circular antisense RNA.* Embo j, 2011. **30**(21): p. 4414-22.
276. Guo, J.U., et al., *Expanded identification and characterization of mammalian circular RNAs.* Genome Biol, 2014. **15**(7): p. 409.
277. Xia, P., et al., *A Circular RNA Protects Dormant Hematopoietic Stem Cells from DNA Sensor cGAS-Mediated Exhaustion.* Immunity, 2018. **48**(4): p. 688-701.e7.
278. Li, Z., et al., *Exon-intron circular RNAs regulate transcription in the nucleus.* Nat Struct Mol Biol, 2015. **22**(3): p. 256-64.
279. Chen, N., et al., *A novel FLI1 exonic circular RNA promotes metastasis in breast cancer by coordinately regulating TET1 and DNMT1.* Genome Biol, 2018. **19**(1): p. 218.
280. Zeng, Y., et al., *A Circular RNA Binds To and Activates AKT Phosphorylation and Nuclear Localization Reducing Apoptosis and Enhancing Cardiac Repair.* Theranostics, 2017. **7**(16): p. 3842-3855.
281. Du, W.W., et al., *Induction of tumor apoptosis through a circular RNA enhancing Foxo3 activity.* Cell Death Differ, 2017. **24**(2): p. 357-370.
282. Stagsted, L.V., et al., *Noncoding AUG circRNAs constitute an abundant and conserved subclass of circles.* Life Sci Alliance, 2019. **2**(3).
283. Mo, D., et al., *A universal approach to investigate circRNA protein coding function.* Scientific reports, 2019. **9**(1): p. 11684-11684.
284. Godet, A.C., et al., *IRES Trans-Acting Factors, Key Actors of the Stress Response.* Int J Mol Sci, 2019. **20**(4).
285. Chen, C.Y. and P. Sarnow, *Initiation of protein synthesis by the eukaryotic translational apparatus on circular RNAs.* Science, 1995. **268**(5209): p. 415.
286. Wang, Y. and Z. Wang, *Efficient backsplicing produces translatable circular mRNAs.* Rna, 2015. **21**(2): p. 172-9.
287. Fan, X., Y. Yang, and Z. Wang, *Pervasive translation of circular RNAs driven by short IRES-like elements.* bioRxiv, 2019: p. 473207.
288. Meyer, K.D., et al., *5' UTR m(6)A Promotes Cap-Independent Translation.* Cell, 2015. **163**(4): p. 999-1010.
289. Yang, Y., et al., *Extensive translation of circular RNAs driven by N(6)-methyladenosine.* Cell Res, 2017. **27**(5): p. 626-641.

290. Legnini, I., et al., *Circ-ZNF609 Is a Circular RNA that Can Be Translated and Functions in Myogenesis.* Mol Cell, 2017. **66**(1): p. 22-37.e9.
291. Yang, Y., et al., *Novel Role of FBXW7 Circular RNA in Repressing Glioma Tumorigenesis.* JNCI: Journal of the National Cancer Institute, 2017. **110**(3): p. 304-315.
292. Zhang, M., et al., *A peptide encoded by circular form of LINC-PINT suppresses oncogenic transcriptional elongation in glioblastoma.* Nature Communications, 2018. **9**(1): p. 4475.
293. Zhang, M., et al., *A novel protein encoded by the circular form of the SHPRH gene suppresses glioma tumorigenesis.* Oncogene, 2018. **37**(13): p. 1805-1814.
294. Zheng, X., et al., *A novel protein encoded by a circular RNA circPPP1R12A promotes tumor pathogenesis and metastasis of colon cancer via Hippo-YAP signaling.* Molecular Cancer, 2019. **18**(1): p. 47.
295. Liang, W.-C., et al., *Translation of the circular RNA circβ-catenin promotes liver cancer cell growth through activation of the Wnt pathway.* Genome Biology, 2019. **20**(1): p. 84.
296. Geng, H.H., et al., *The Circular RNA Cdr1as Promotes Myocardial Infarction by Mediating the Regulation of miR-7a on Its Target Genes Expression.* PLoS One, 2016. **11**(3): p. e0151753.
297. Han, D., et al., *Circular RNA circMTO1 acts as the sponge of microRNA-9 to suppress hepatocellular carcinoma progression.* Hepatology, 2017. **66**(4): p. 1151-1164.
298. Lim, T.B., et al., *Targeting the highly abundant circular RNA circSlc8a1 in cardiomyocytes attenuates pressure overload induced hypertrophy.* Cardiovasc Res, 2019. **115**(14): p. 1998-2007.
299. Santer, L., C. Bär, and T. Thum, *Circular RNAs: A Novel Class of Functional RNA Molecules with a Therapeutic Perspective.* Mol Ther, 2019. **27**(8): p. 1350-1363.
300. Zhang, Y., et al., *Elevated serum circ_0068481 levels as a potential diagnostic and prognostic indicator in idiopathic pulmonary arterial hypertension.* Pulm Circ, 2019. **9**(4): p. 2045894019888416.
301. Chen, C., et al., *The Circular RNA CDR1as Regulates the Proliferation and Apoptosis of Human Cardiomyocytes Through the miR-135a/HMOX1 and miR-135b/HMOX1 Axes.* Genet Test Mol Biomarkers, 2020.
302. Bahn, J.H., et al., *The landscape of microRNA, Piwi-interacting RNA, and circular RNA in human saliva.* Clin Chem, 2015. **61**(1): p. 221-30.
303. Hulstaert, E., et al., *Charting extracellular transcriptomes in The Human Biofluid RNA Atlas.* bioRxiv, 2020: p. 823369.
304. Wang, L., et al., *Identification of circular RNA Hsa_circ_0001879 and Hsa_circ_0004104 as novel biomarkers for coronary artery disease.* Atherosclerosis, 2019. **286**: p. 88-96.
305. Zhao, Z., et al., *Peripheral blood circular RNA hsa_circ_0124644 can be used as a diagnostic biomarker of coronary artery disease.* Scientific Reports, 2017. **7**(1): p. 39918.
306. Vausort, M., et al., *Myocardial Infarction-Associated Circular RNA Predicting Left Ventricular Dysfunction.* J Am Coll Cardiol, 2016. **68**(11): p. 1247-1248.
307. Salgado-Somoza, A., et al., *The circular RNA MICRA for risk stratification after myocardial infarction.* IJC Heart & Vasculature, 2017. **17**: p. 33-36.
308. Shen, L., et al., *CircRNA0044073 is upregulated in atherosclerosis and increases the proliferation and invasion of cells by targeting miR107.* Mol Med Rep, 2019. **19**(5): p. 3923-3932.

309. Yang, L., et al., *Circular RNA circCHFR Facilitates the Proliferation and Migration of Vascular Smooth Muscle via miR-370/FOXO1/Cyclin D1 Pathway.* Mol Ther Nucleic Acids, 2019. **16**: p. 434-441.
310. Zhuang, J.B., et al., *Circ_CHFR expedites cell growth, migration and inflammation in ox-LDL-treated human vascular smooth muscle cells via the miR-214-3p/Wnt3/beta-catenin pathway.* Eur Rev Med Pharmacol Sci, 2020. **24**(6): p. 3282-3292.
311. Mao, Y.Y., et al., *Circ-SATB2 upregulates STIM1 expression and regulates vascular smooth muscle cell proliferation and differentiation through miR-939.* Biochem Biophys Res Commun, 2018. **505**(1): p. 119-125.
312. Kong, P., et al., *circ-Sirt1 controls NF-κB activation via sequence-specific interaction and enhancement of SIRT1 expression by binding to miR-132/212 in vascular smooth muscle cells.* Nucleic acids research, 2019. **47**(7): p. 3580-3593.
313. Zhang, S., et al., *Circular RNA circ_0003204 inhibits proliferation, migration and tube formation of endothelial cell in atherosclerosis via miR-370-3p/TGFbetaR2/phosph-SMAD3 axis.* J Biomed Sci, 2020. **27**(1): p. 11.
314. Sun, J., Z. Zhang, and S. Yang, *Circ_RUSC2 upregulates the expression of miR-661 target gene SYK and regulates the function of vascular smooth muscle cells.* Biochem Cell Biol, 2019. **97**(6): p. 709-714.
315. Song, C.L., et al., *Effect of Circular ANRIL on the Inflammatory Response of Vascular Endothelial Cells in a Rat Model of Coronary Atherosclerosis.* Cellular Physiology and Biochemistry, 2017. **42**(3): p. 1202-1212.
316. Liu, H., et al., *Circular RNA has_circ_0003204 inhibits oxLDL-induced vascular endothelial cell proliferation and angiogenesis.* Cell Signal, 2020. **70**: p. 109595.
317. Burd, C.E., et al., *Expression of linear and novel circular forms of an INK4/ARF-associated non-coding RNA correlates with atherosclerosis risk.* PLoS genetics, 2010. **6**(12): p. e1001233-e1001233.
318. Michiels, C., *Physiological and pathological responses to hypoxia.* Am J Pathol, 2004. **164**(6): p. 1875-82.
319. Konstantinidis, K., R.S. Whelan, and R.N. Kitsis, *Mechanisms of cell death in heart disease.* Arterioscler Thromb Vasc Biol, 2012. **32**(7): p. 1552-62.
320. Turer, A.T. and J.A. Hill, *Pathogenesis of myocardial ischemia-reperfusion injury and rationale for therapy.* Am J Cardiol, 2010. **106**(3): p. 360-8.
321. Jin, Q. and Y. Chen, *Silencing circular RNA circ_0010729 protects human cardiomyocytes from oxygen-glucose deprivation-induced injury by up-regulating microRNA-145-5p.* Mol Cell Biochem, 2019. **462**(1-2): p. 185-194.
322. Chen, L., et al., *circDLPAG4/HECTD1 mediates ischaemia/reperfusion injury in endothelial cells via ER stress.* RNA Biol, 2020. **17**(2): p. 240-253.
323. Ji, X., et al., *MicroRNA-31-5p attenuates doxorubicin-induced cardiotoxicity via quaking and circular RNA Pan3.* J Mol Cell Cardiol, 2020. **140**: p. 56-67.
324. Wang, K., et al., *Circular RNA mediates cardiomyocyte death via miRNA-dependent upregulation of MTP18 expression.* Cell Death Differ, 2017. **24**(6): p. 1111-1120.
325. Li, M., et al., *A circular transcript of ncx1 gene mediates ischemic myocardial injury by targeting miR-133a-3p.* Theranostics, 2018. **8**(21): p. 5855-5869.
326. Gan, J., et al., *Circular RNA_101237 mediates anoxia/reoxygenation injury by targeting let-7a-5p/IGF2BP3 in cardiomyocytes.* Int J Mol Med, 2020. **45**(2): p. 451-460.

327. Sun, L.Y., et al., *Circ_LAS1L regulates cardiac fibroblast activation, growth, and migration through miR-125b/SFRP5 pathway.* Cell Biochem Funct, 2020. **38**(4): p. 443-450.
328. Du, W.W., et al., *Foxo3 circular RNA promotes cardiac senescence by modulating multiple factors associated with stress and senescence responses.* European Heart Journal, 2017. **38**(18): p. 1402-1412.
329. Garikipati, V.N.S., et al., *Circular RNA CircFndc3b modulates cardiac repair after myocardial infarction via FUS/VEGF-A axis.* Nat Commun, 2019. **10**(1): p. 4317.
330. Huang, S., et al., *Loss of Super-Enhancer-Regulated circRNA Nfix Induces Cardiac Regeneration After Myocardial Infarction in Adult Mice.* Circulation, 2019. **139**(25): p. 2857-2876.
331. Zhou, L.Y., et al., *The circular RNA ACR attenuates myocardial ischemia/reperfusion injury by suppressing autophagy via modulation of the Pink1/ FAM65B pathway.* Cell Death Differ, 2019. **26**(7): p. 1299-1315.
332. Bai, M., et al., *CircHIPK3 aggravates myocardial ischemia-reperfusion injury by binding to miRNA-124-3p.* Eur Rev Med Pharmacol Sci, 2019. **23**(22): p. 10107-10114.
333. Cai, L., et al., *Circular RNA Ttc3 regulates cardiac function after myocardial infarction by sponging miR-15b.* J Mol Cell Cardiol, 2019. **130**: p. 10-22.
334. Zhao, B., et al., *CircMACF1 Attenuates Acute Myocardial Infarction Through miR-500b-5p-EMP1 Axis.* J Cardiovasc Transl Res, 2020.
335. Zhang, Q., et al., *The circular RNA hsa_circ_0007623 acts as a sponge of microRNA-297 and promotes cardiac repair.* Biochem Biophys Res Commun, 2020. **523**(4): p. 993-1000.
336. Shao, Y., et al., *Circular RNA circDENND2A protects H9c2 cells from oxygen glucose deprivation-induced apoptosis through sponging microRNA-34a.* Cell Cycle, 2020. **19**(2): p. 246-255.
337. Werfel, S., et al., *Characterization of circular RNAs in human, mouse and rat hearts.* Journal of Molecular and Cellular Cardiology, 2016. **98**: p. 103-107.
338. Bai, M., et al., *CircRNA 010567 improves myocardial infarction rats through inhibiting TGF-β1.* Eur Rev Med Pharmacol Sci, 2020. **24**(1): p. 369-375.
339. Burchfield, J.S., M. Xie, and J.A. Hill, *Pathological ventricular remodeling: mechanisms: part 1 of 2.* Circulation, 2013. **128**(4): p. 388-400.
340. Wu, Q.Q., et al., *Mechanisms contributing to cardiac remodelling.* Clin Sci (Lond), 2017. **131**(18): p. 2319-2345.
341. Cohn, J.N., R. Ferrari, and N. Sharpe, *Cardiac remodeling--concepts and clinical implications: a consensus paper from an international forum on cardiac remodeling. Behalf of an International Forum on Cardiac Remodeling.* J Am Coll Cardiol, 2000. **35**(3): p. 569-82.
342. Hinderer, S. and K. Schenke-Layland, *Cardiac fibrosis - A short review of causes and therapeutic strategies.* Adv Drug Deliv Rev, 2019. **146**: p. 77-82.
343. Han, J., et al., *Circular RNA-Expression Profiling Reveals a Potential Role of Hsa_circ_0097435 in Heart Failure via Sponging Multiple MicroRNAs.* Front Genet, 2020. **11**: p. 212.
344. Wen, J., et al., *Circular RNA HIPK3: A Key Circular RNA in a Variety of Human Cancers.* Frontiers in Oncology, 2020. **10**: p. 773.
345. Ni, H., et al., *Inhibition of circHIPK3 prevents angiotensin II-induced cardiac fibrosis by sponging miR-29b-3p.* Int J Cardiol, 2019. **292**: p. 188-196.
346. Wang, K., et al., *A circular RNA protects the heart from pathological hypertrophy and heart failure by targeting miR-223.* European Heart Journal, 2016. **37**(33): p. 2602-2611.

347. Tang, C.M., et al., *CircRNA_000203 enhances the expression of fibrosis-associated genes by derepressing targets of miR-26b-5p, Col1a2 and CTGF, in cardiac fibroblasts.* Sci Rep, 2017. **7**: p. 40342.
348. Zhou, B. and J.W. Yu, *A novel identified circular RNA, circRNA_010567, promotes myocardial fibrosis via suppressing miR-141 by targeting TGF-β1.* Biochem Biophys Res Commun, 2017. **487**(4): p. 769-775.
349. Deng, Y., et al., *Circ-HIPK3 Strengthens the Effects of Adrenaline in Heart Failure by MiR-17-3p - ADCY6 Axis.* Int J Biol Sci, 2019. **15**(11): p. 2484-2496.
350. Zhu, Y., et al., *Upregulation of Circular RNA CircNFIB Attenuates Cardiac Fibrosis by Sponging miR-433.* Front Genet, 2019. **10**: p. 564.
351. Li, H., et al., *Circular RNA circRNA_000203 aggravates cardiac hypertrophy via suppressing miR-26b-5p and miR-140-3p binding to Gata4.* Cardiovasc Res, 2020. **116**(7): p. 1323-1334.
352. Benjak, A., C. Sala, and R.C. Hartkoorn, *Whole-genome sequencing for comparative genomics and de novo genome assembly.* Methods Mol Biol, 2015. **1285**: p. 1-16.
353. Konnick, E., C.M. Lockwood, and D. Wu, *Targeted Next-Generation Sequencing of Acute Leukemia.* Methods Mol Biol, 2017. **1633**: p. 163-184.
354. Podnar, J., et al., *Next-Generation Sequencing RNA-Seq Library Construction.* (1934-3647 (Electronic)).
355. Billmeier, M. and P. Xu, *Small RNA Profiling by Next-Generation Sequencing Using High-Definition Adapters.* Methods Mol Biol, 2017. **1580**: p. 45-57.
356. Song, Y., et al., *A comparative analysis of library prep approaches for sequencing low input translatome samples.* BMC Genomics, 2018. **19**(1): p. 696.
357. Cao, J., et al., *The single-cell transcriptional landscape of mammalian organogenesis.* Nature, 2019. **566**(7745): p. 496-502.
358. Wheeler, E.C., E.L. Van Nostrand, and G.W. Yeo, *Advances and challenges in the detection of transcriptome-wide protein-RNA interactions.* Wiley Interdiscip Rev RNA, 2018. **9**(1).
359. Lister, R., et al., *Highly integrated single-base resolution maps of the epigenome in Arabidopsis.* Cell, 2008. **133**(3): p. 523-36.
360. Nagalakshmi, U., et al., *The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing.* Science, 2008. **320**(5881): p. 1344.
361. Wilhelm, B.T., et al., *Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution.* Nature, 2008. **453**(7199): p. 1239-1243.
362. Mortazavi, A., et al., *Mapping and quantifying mammalian transcriptomes by RNA-Seq.* Nature Methods, 2008. **5**(7): p. 621-628.
363. Heather, J.M. and B. Chain, *The sequence of sequencers: The history of sequencing DNA.* Genomics, 2016. **107**(1): p. 1-8.
364. Mardis, E.R., *Next-Generation Sequencing Platforms.* Annual Review of Analytical Chemistry, 2013. **6**(1): p. 287-303.
365. Zhao, S., et al., *Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion.* Scientific Reports, 2018. **8**(1): p. 4781.
366. Leggett, R.M. and M.D. Clark, *A world of opportunities with nanopore sequencing.* Journal of Experimental Botany, 2017. **68**(20): p. 5419-5429.
367. Dohm, J.C., et al., *Benchmarking of long-read correction methods.* NAR Genomics and Bioinformatics, 2020. **2**(2).

368. Lu, H., F. Giordano, and Z. Ning, *Oxford Nanopore MinION Sequencing and Genome Assembly.* Genomics Proteomics Bioinformatics, 2016. **14**(5): p. 265-279.
369. Li, X., et al., *Lowering the quantification limit of the QubitTM RNA HS Assay using RNA spike-in.* BMC Molecular Biology, 2015. **16**(1): p. 9.
370. Schroeder, A., et al., *The RIN: an RNA integrity number for assigning integrity values to RNA measurements.* BMC Molecular Biology, 2006. **7**(1): p. 3.
371. Hardwick, S.A., et al., *Spliced synthetic genes as internal controls in RNA sequencing experiments.* Nature Methods, 2016. **13**(9): p. 792-798.
372. Hardwick, S.A., I.W. Deveson, and T.R. Mercer, *Reference standards for next-generation sequencing.* Nature Reviews Genetics, 2017. **18**: p. 473.
373. Li, H., *Minimap2: pairwise alignment for nucleotide sequences.* Bioinformatics, 2018. **34**(18): p. 3094-3100.
374. Langfelder, P. and S. Horvath, *WGCNA: an R package for weighted correlation network analysis.* BMC Bioinformatics, 2008. **9**: p. 559.
375. Jain, A. and G. Tuteja, *TissueEnrich: Tissue-specific gene enrichment analysis.* Bioinformatics, 2018. **35**(11): p. 1966-1967.
376. Siepel, A., et al., *Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.* Genome Res, 2005. **15**(8): p. 1034-50.
377. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-842.
378. Ewing, B. and P. Green, *Base-calling of automated sequencer traces using phred. II. Error probabilities.* Genome Res, 1998. **8**(3): p. 186-94.
379. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data.* Bioinformatics, 2014. **30**(15): p. 2114-20.
380. Del Fabbro, C., et al., *An extensive evaluation of read trimming effects on Illumina NGS data analysis.* PLoS One, 2013. **8**(12): p. e85024.
381. Williams, C.R., et al., *Trimming of sequence reads alters RNA-Seq gene expression estimates.* BMC Bioinformatics, 2016. **17**(1): p. 103.
382. Yang, C., et al., *The impact of RNA-seq aligners on gene expression estimation.* ACM-BCB ... ... : the ... ACM Conference on Bioinformatics, Computational Biology and Biomedicine. ACM Conference on Bioinformatics, Computational Biology and Biomedicine, 2015. **2015**: p. 462-471.
383. Baruzzo, G., et al., *Simulation-based comprehensive benchmarking of RNA-seq aligners.* Nature Methods, 2016. **14**: p. 135.
384. Wang, L., S. Wang, and W. Li, *RSeQC: quality control of RNA-seq experiments.* Bioinformatics, 2012. **28**(16): p. 2184-2185.
385. Zeng, X., et al., *A comprehensive overview and evaluation of circular RNA detection tools.* PLoS Computational Biology, 2017. **13**(6): p. e1005420.
386. Tutar, Y., *Pseudogenes.* Comp Funct Genomics, 2012. **2012**: p. 424526.
387. Voshall A., M.E., *Next-Generation Transcriptome Assembly: Strategies and Performance Analysis*, in *Bioinformatics in the Era of Post Genomics and Big Data*. 2018, IntechOpen.
388. Pertea, M., et al., *StringTie enables improved reconstruction of a transcriptome from RNA-seq reads.* Nature Biotechnology, 2015. **33**: p. 290.
389. Wang, L., et al., *CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model.* Nucleic acids research, 2013. **41**(6): p. e74-e74.
390. Choi, S.-W., H.-W. Kim, and J.-W. Nam, *The small peptide world in long noncoding RNAs.* Briefings in bioinformatics, 2019. **20**(5): p. 1853-1864.
391. Roberts, A., et al., *Improving RNA-Seq expression estimates by correcting for fragment bias.* Genome biology, 2011. **12**(3): p. R22-R22.

392. Tuerk, A., G. Wiktorin, and S. Güler, *Mixture models reveal multiple positional bias types in RNA-Seq data and lead to accurate transcript concentration estimates.* PLoS Comput Biol, 2017. **13**(5): p. e1005515.
393. Benjamini, Y. and T.P. Speed, *Summarizing and correcting the GC content bias in high-throughput sequencing.* Nucleic Acids Res, 2012. **40**(10): p. e72.
394. Risso, D., et al., *GC-content normalization for RNA-Seq data.* BMC Bioinformatics, 2011. **12**: p. 480.
395. Patro, R., et al., *Salmon provides fast and bias-aware quantification of transcript expression.* Nature methods, 2017. **14**(4): p. 417-419.
396. Zhang, C., et al., *Evaluation and comparison of computational tools for RNA-seq isoform quantification.* BMC Genomics, 2017. **18**(1): p. 583.
397. Pertea, M., et al., *Thousands of large-scale RNA sequencing experiments yield a comprehensive new human gene list and reveal extensive transcriptional noise.* bioRxiv, 2018: p. 332825.
398. Soneson, C., M. Love, and M. Robinson, *Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences [version 2; referees: 2 approved].* F1000Research, 2016. **4**(1521).
399. Love, M.I., C. Soneson, and R. Patro, *Swimming downstream: statistical analysis of differential transcript usage following Salmon quantification.* F1000Research, 2018. **7**: p. 952-952.
400. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome biology, 2014. **15**(12): p. 550-550.
401. Costa-Silva, J., D. Domingues, and F.M. Lopes, *RNA-Seq differential expression analysis: An extended review and a software tool.* PloS one, 2017. **12**(12): p. e0190152-e0190152.
402. Mirsafian, H., et al., *Long non-coding RNA expression in primary human monocytes.* Genomics, 2016. **108**(1): p. 37-45.
403. Memczak, S., et al., *Identification and Characterization of Circular RNAs As a New Class of Putative Biomarkers in Human Blood.* PloS one, 2015. **10**(10): p. e0141214-e0141214.
404. Harrow, J., et al., *GENCODE: the reference human genome annotation for The ENCODE Project.* (1549-5469 (Electronic)).
405. Ard, R., R.C. Allshire, and S. Marquardt, *Emerging Properties and Functional Consequences of Noncoding Transcription.* Genetics, 2017. **207**(2): p. 357-367.
406. Han, S., et al., *Long Noncoding RNA Identification: Comparing Machine Learning Based Tools for Long Noncoding Transcripts Discrimination.* BioMed Research International, 2016. **2016**: p. 14.
407. Hirose, T., T. Yamazaki, and S. Nakagawa, *Molecular anatomy of the architectural NEAT1 noncoding RNA: The domains, interactors, and biogenesis pathway required to build phase-separated nuclear paraspeckles.* Wiley Interdiscip Rev RNA, 2019. **10**(6): p. e1545.
408. Huang, S., et al., *Long noncoding RNA MALAT1 mediates cardiac fibrosis in experimental postinfarct myocardium mice model.* J Cell Physiol, 2019. **234**(3): p. 2997-3006.
409. Gast, M., et al., *Long noncoding RNA NEAT1 modulates immune cell functions and is suppressed in early onset myocardial infarction patients.* Cardiovasc Res, 2019.
410. Zhu, Y., et al., *Long noncoding RNA TUG1 promotes cardiac fibroblast transformation to myofibroblasts via miR29c in chronic hypoxia.* Mol Med Rep, 2018. **18**(3): p. 3451-3460.

411. Hindorff, L.A., et al., *Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.* Proc Natl Acad Sci U S A, 2009. **106**(23): p. 9362-7.
412. Gourdin, M.D.P., *Impact-of-ischemia-on-cellular-metabolism*, in *Artery Bypass*, W.S. Aronow, Editor. 2013: New York Medical College and Westchester Medical Center.
413. Chacko, S., et al., *The role of biomarkers in the diagnosis and risk stratification of acute coronary syndrome.* Future Sci OA, 2018. **4**(1): p. Fso251.
414. Cambier, L., et al., *Fem1a is a mitochondrial protein up-regulated upon ischemia-reperfusion injury.* (1873-3468 (Electronic)).
415. Gupta, I., et al., *Delineating Crosstalk Mechanisms of the Ubiquitin Proteasome System That Regulate Apoptosis.* Frontiers in Cell and Developmental Biology, 2018. **6**: p. 11.
416. Kalfon, R., et al., *JDP2 and ATF3 deficiencies dampen maladaptive cardiac remodeling and preserve cardiac function.* PLoS One, 2019. **14**(2): p. e0213081.
417. Blevins, M.A., M. Huang, and R. Zhao, *The Role of CtBP1 in Oncogenic Processes and Its Potential as a Therapeutic Target.* Mol Cancer Ther, 2017. **16**(6): p. 981-990.
418. Belyaeva, O.V., et al., *Human retinol dehydrogenase 13 (RDH13) is a mitochondrial short-chain dehydrogenase/reductase with a retinaldehyde reductase activity.* (1742-464X (Print)).
419. Bastian, F., et al., *Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species*. 2008. p. 124-131.
420. Lupo, A., et al., *KRAB-Zinc Finger Proteins: A Repressor Family Displaying Multiple Biological Functions.* (1389-2029 (Print)).
421. Mailloux, R.J., D. Craig Ayre, and S.L. Christian, *Induction of mitochondrial reactive oxygen species production by GSH mediated S-glutathionylation of 2-oxoglutarate dehydrogenase.* Redox Biol, 2016. **8**: p. 285-97.
422. Efimova, N., et al., *Podosome-regulating kinesin KIF1C translocates to the cell periphery in a CLASP-dependent manner.* Journal of Cell Science, 2014. **127**(24): p. 5179.
423. Chun, S.K., et al., *Loss of sirtuin 1 and mitofusin 2 contributes to enhanced ischemia/reperfusion injury in aged livers.* Aging Cell, 2018. **17**(4): p. e12761.
424. Davies, S.M., et al., *MRPS27 is a pentatricopeptide repeat domain protein required for the translation of mitochondrially encoded proteins.* FEBS Lett, 2012. **586**(20): p. 3555-61.
425. Shohet, R.V. and J.A. Garcia, *Keeping the engine primed: HIF factors as key regulators of cardiac metabolism and angiogenesis during ischemia.* J Mol Med (Berl), 2007. **85**(12): p. 1309-15.
426. Arslan, F., D.P. de Kleijn, and G. Pasterkamp, *Innate immune signaling in cardiac ischemia.* Nat Rev Cardiol, 2011. **8**(5): p. 292-300.
427. Gisterå, A. and G.K. Hansson, *The immunology of atherosclerosis.* Nat Rev Nephrol, 2017. **13**(6): p. 368-380.
428. Ward, N.L. and D.J. Dumont, *The angiopoietins and Tie2/Tek: adding to the complexity of cardiovascular development.* Semin Cell Dev Biol, 2002. **13**(1): p. 19-27.
429. Teichert, M., et al., *Pericyte-expressed Tie2 controls angiogenesis and vessel maturation.* Nature Communications, 2017. **8**(1): p. 16106.
430. Park, J.I., et al., *Novel function of E26 transformation-specific domain-containing protein ELK3 in lymphatic endothelial cells.* Oncol Lett, 2018. **15**(1): p. 55-60.

431. Sauteur, L., et al., *Cdh5/VE-cadherin promotes endothelial cell interface elongation via cortical actin polymerization during angiogenic sprouting.* Cell Rep, 2014. **9**(2): p. 504-13.
432. Rudno-Rudzinska, J., et al., *A review on Eph/ephrin, angiogenesis and lymphangiogenesis in gastric, colorectal and pancreatic cancers.* Chin J Cancer Res, 2017. **29**(4): p. 303-312.
433. Zhang, X., et al., *lncRNA PCAT19 negatively regulates p53 in non-small cell lung cancer.* Oncol Lett, 2019. **18**(6): p. 6795-6800.
434. Pfaff, M.J., et al., *Tumor suppressor protein p53 negatively regulates ischemia-induced angiogenesis and arteriogenesis.* J Vasc Surg, 2018. **68**(6s): p. 222S-233S.e1.
435. Ounzain, S., et al., *CARMEN, a human super enhancer-associated long noncoding RNA controlling cardiac specification, differentiation and homeostasis.* (1095-8584 (Electronic)).
436. Coppiello, G., et al., *Meox2/Tcf15 heterodimers program the heart capillary endothelium for cardiac fatty acid uptake.* (1524-4539 (Electronic)).
437. Mendler, L., T. Braun, and S. Muller, *The Ubiquitin-Like SUMO System and Heart Function: From Development to Disease.* (1524-4571 (Electronic)).
438. Miao, H., et al., *A long noncoding RNA distributed in both nucleus and cytoplasm operates in the PYCARD-regulated apoptosis by coordinating the epigenetic and translational regulation.* PLoS Genet, 2019. **15**(5): p. e1008144.
439. Xie, Y.H. and J. Hu, *Suppression of long non-coding RNA PCAT19 inhibits glioma cell proliferation and invasion, and increases cell apoptosis through regulation of MELK targeted by miR-142-5p.* Genes Genomics, 2020. **42**(11): p. 1299-1310.
440. Xu, S., J. Guo, and W. Zhang, *lncRNA PCAT19 promotes the proliferation of laryngocarcinoma cells via modulation of the miR-182/PDK4 axis.* J Cell Biochem, 2019. **120**(8): p. 12810-12821.
441. Santos-Zas, I., et al., *Adaptive Immune Responses Contribute to Post-ischemic Cardiac Remodeling.* Front Cardiovasc Med, 2018. **5**: p. 198.
442. Turner, A.W., et al., *Multi-Omics Approaches to Study Long Non-coding RNA Function in Atherosclerosis.* Front Cardiovasc Med, 2019. **6**: p. 9.
443. Breveglieri, G., et al., *Non-invasive Prenatal Testing Using Fetal DNA.* Mol Diagn Ther, 2019. **23**(2): p. 291-299.
444. Butler, T.M., P.T. Spellman, and J. Gray, *Circulating-tumor DNA as an early detection and diagnostic tool.* Curr Opin Genet Dev, 2017. **42**: p. 14-21.
445. Ding, H., et al., *Combined detection of miR-21-5p, miR-30a-3p, miR-30a-5p, miR-155-5p, miR-216a and miR-217 for screening of early heart failure diseases.* Biosci Rep, 2020.
446. Liu, X.X., et al., *A two-circular RNA signature as a noninvasive diagnostic biomarker for lung adenocarcinoma.* J Transl Med, 2019. **17**(1): p. 50.
447. Bazzell, B.G., et al., *Human Urinary mRNA as a Biomarker of Cardiovascular Disease.* Circ Genom Precis Med, 2018. **11**(9): p. e002213.
448. Yuan, T., et al., *Plasma extracellular RNA profiles in healthy and cancer patients.*
449. Reddy, L.L., et al., *Circulating miRNA-33: a potential biomarker in patients with coronary artery disease.* Biomarkers, 2019. **24**(1): p. 36-42.
450. Danielson, K.M., et al., *High Throughput Sequencing of Extracellular RNA from Human Plasma.* PloS one, 2017. **12**(1): p. e0164644-e0164644.
451. Savelyeva, A.V., et al., *Variety of RNAs in Peripheral Blood Cells, Plasma, and Plasma Fractions.* Biomed Res Int, 2017. **2017**: p. 7404912.

452. Qin, Y., et al., *High-throughput sequencing of human plasma RNA by using thermostable group II intron reverse transcriptases.* Rna, 2016. **22**(1): p. 111-28.
453. Everaert, C., et al., *Performance assessment of total RNA sequencing of human biofluids and extracellular vesicles.* Scientific Reports, 2019. **9**(1): p. 17574.
454. Turchinovich, A., O. Drapkina, and A. Tonevitsky, *Transcriptome of Extracellular Vesicles: State-of-the-Art.* Frontiers in Immunology, 2019. **10**(202).
455. Galvanin, A., et al., *Diversity and heterogeneity of extracellular RNA in human plasma.* Biochimie, 2019. **164**: p. 22-36.
456. Ewels, P., et al., *MultiQC: summarize analysis results for multiple tools and samples in a single report.* Bioinformatics, 2016. **32**(19): p. 3047-3048.
457. Hobuß, L., C. Bär, and T. Thum, *Long Non-coding RNAs: At the Heart of Cardiac Dysfunction?* Frontiers in Physiology, 2019. **10**: p. 30.
458. Salter, S.J., et al., *Reagent and laboratory contamination can critically impact sequence-based microbiome analyses.* BMC Biol, 2014. **12**: p. 87.
459. Glassing, A., et al., *Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples.* Gut Pathogens, 2016. **8**(1): p. 24.
460. Eisenhofer, R., et al., *Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations.* Trends Microbiol, 2019. **27**(2): p. 105-117.
461. Shen, H., S. Rogelj, and T.L. Kieft, *Sensitive, real-time PCR detects low-levels of contamination by Legionella pneumophila in commercial reagents.* Mol Cell Probes, 2006. **20**(3-4): p. 147-53.
462. Weyrich, L.S., et al., *Laboratory contamination over time during low-biomass sample analysis.* Mol Ecol Resour, 2019. **19**(4): p. 982-996.
463. Culviner, P.H., C.K. Guegler, and M.T. Laub, *A Simple, Cost-Effective, and Robust Method for rRNA Depletion in RNA-Sequencing Studies.* mBio, 2020. **11**(2): p. e00010-20.
464. Ahn, J., H. Wu, and K. Lee, *Integrative Analysis Revealing Human Heart-Specific Genes and Consolidating Heart-Related Phenotypes.* Front Genet, 2020. **11**: p. 777.
465. Kong, F., et al., *Long noncoding RNA RMRP upregulation aggravates myocardial ischemia-reperfusion injury by sponging miR-206 to target ATG3 expression.* Biomedicine & Pharmacotherapy, 2019. **109**: p. 716-725.
466. Jiang, Y., et al., *Long noncoding RNA SNHG6 contributes to ventricular septal defect formation via negative regulation of miR-101 and activation of Wnt/β-catenin pathway.* Pharmazie, 2019. **74**(1): p. 23-28.
467. Zhang, S.Y., et al., *Upregulation of lncRNA RMRP promotes the activation of cardiac fibroblasts by regulating miR-613.* Mol Med Rep, 2019. **20**(4): p. 3849-3857.
468. Sun, R. and L. Zhang, *Long non-coding RNA MALAT1 regulates cardiomyocytes apoptosis after hypoxia/reperfusion injury via modulating miR-200a-3p/PDCD4 axis.* Biomed Pharmacother, 2019. **111**: p. 1036-1045.
469. Leifheit-Nestler, M. and D. Haffner, *Paracrine Effects of FGF23 on the Heart.* Front Endocrinol (Lausanne), 2018. **9**: p. 278.
470. Shibata, K., et al., *Association between circulating fibroblast growth factor 23, α-Klotho, and the left ventricular ejection fraction and left ventricular mass in cardiology inpatients.* PLoS One, 2013. **8**(9): p. e73184.

471. Isakova, T., et al., *Associations between fibroblast growth factor 23 and cardiac characteristics in pediatric heart failure.* Pediatr Nephrol, 2013. **28**(10): p. 2035-42.
472. Parker, B.D., et al., *The associations of fibroblast growth factor 23 and uncarboxylated matrix Gla protein with mortality in coronary artery disease: the Heart and Soul Study.* Ann Intern Med, 2010. **152**(10): p. 640-8.
473. Mirza, M.A., et al., *Serum intact FGF23 associate with left ventricular mass, hypertrophy and geometry in an elderly population.* Atherosclerosis, 2009. **207**(2): p. 546-51.
474. De Koninck, M., et al., *Essential Roles of Cohesin STAG2 in Mouse Embryonic Development and Adult Tissue Homeostasis.* Cell Rep, 2020. **32**(6): p. 108014.
475. Bik-Multanowski, M., J.J. Pietrzyk, and A. Midro, *MTRNR2L12: A Candidate Blood Marker of Early Alzheimer's Disease-Like Dementia in Adults with Down Syndrome.* Journal of Alzheimer's Disease, 2015. **46**: p. 145-150.
476. Thummasorn, S., et al., *Humanin exerts cardioprotection against cardiac ischemia/reperfusion injury through attenuation of mitochondrial dysfunction.* Cardiovasc Ther, 2016. **34**(6): p. 404-414.
477. Thummasorn, S., et al., *High-dose Humanin analogue applied during ischemia exerts cardioprotection against ischemia/reperfusion injury by reducing mitochondrial dysfunction.* Cardiovasc Ther, 2017. **35**(5).
478. Thummasorn, S., et al., *Humanin directly protects cardiac mitochondria against dysfunction initiated by oxidative stress by decreasing complex I activity.* Mitochondrion, 2018. **38**: p. 31-40.
479. Kumfu, S., et al., *Humanin Exerts Neuroprotection During Cardiac Ischemia-Reperfusion Injury.* J Alzheimers Dis, 2018. **61**(4): p. 1343-1353.
480. Charununtakorn, S.T., et al., *Potential Roles of Humanin on Apoptosis in the Heart.* Cardiovasc Ther, 2016. **34**(2): p. 107-14.
481. Pacheu-Grau, D., et al., *Cooperation between COA6 and SCO2 in COX2 maturation during cytochrome c oxidase assembly links two mitochondrial cardiomyopathies.* Cell Metab, 2015. **21**(6): p. 823-33.
482. Ogilvie, I., N.G. Kennaway, and E.A. Shoubridge, *A molecular chaperone for mitochondrial complex I assembly is mutated in a progressive encephalopathy.* J Clin Invest, 2005. **115**(10): p. 2784-92.
483. Chen, Y., et al., *Circ-ASH2L promotes tumor progression by sponging miR-34a to regulate Notch1 in pancreatic ductal adenocarcinoma.* J Exp Clin Cancer Res, 2019. **38**(1): p. 466.
484. Yang, Y., et al., *MicroRNA-34a Plays a Key Role in Cardiac Repair and Regeneration Following Myocardial Infarction.* Circ Res, 2015. **117**(5): p. 450-9.
485. Rodosthenous, R.S., et al., *Profiling Extracellular Long RNA Transcriptome in Human Plasma and Extracellular Vesicles for Biomarker Discovery.* iScience, 2020. **23**(6): p. 101182.
486. Intlekofer, A.M., et al., *Integrated DNA/RNA targeted genomic profiling of diffuse large B-cell lymphoma using a clinical assay.* Blood Cancer Journal, 2018. **8**(6): p. 60.
487. McKiernan, J., et al., *A Novel Urine Exosome Gene Expression Assay to Predict High-grade Prostate Cancer at Initial Biopsy.* JAMA Oncology, 2016. **2**(7): p. 882-889.
488. Perdas, E., et al., *Potential of Liquid Biopsy in Papillary Thyroid Carcinoma in Context of miRNA, BRAF and p53 Mutation.* Curr Drug Targets, 2018. **19**(14): p. 1721-1729.

489. Shi, J., et al., *The Plasma LncRNA Acting as Fingerprint in Hilar Cholangiocarcinoma.* Cell Physiol Biochem, 2018. **49**(5): p. 1694-1702.
490. Zheng, R., et al., *Genome-wide long non-coding RNAs identified a panel of novel plasma biomarkers for gastric cancer diagnosis.* Gastric Cancer, 2019. **22**(4): p. 731-741.
491. Mercer, T.R., et al., *The human mitochondrial transcriptome.* Cell, 2011. **146**(4): p. 645-58.
492. Ilic, Z., et al., *Control (Native) and oxidized (DeMP) mitochondrial RNA are proinflammatory regulators in human.* Free Radic Biol Med, 2019. **143**: p. 62-69.
493. Al Amir Dache, Z., et al., *Blood contains circulating cell-free respiratory competent mitochondria.* Faseb j, 2020. **34**(3): p. 3616-3630.
494. Wang, L., et al., *Plasma nuclear and mitochondrial DNA levels in acute myocardial infarction patients.* Coron Artery Dis, 2015. **26**(4): p. 296-300.
495. Sudakov, N.P., et al., *The level of free circulating mitochondrial DNA in blood as predictor of death in case of acute coronary syndrome.* European Journal of Medical Research, 2017. **22**(1): p. 1.
496. Mahrouf-Yorgov, M., et al., *Mesenchymal stem cells sense mitochondria released from damaged cells as danger signals to activate their rescue properties.* Cell Death Differ, 2017. **24**(7): p. 1224-1238.
497. Ghirlando, R. and G. Felsenfeld, *CTCF: making the right connections.* Genes Dev, 2016. **30**(8): p. 881-91.
498. Hansen, A.S., et al., *Distinct Classes of Chromatin Loops Revealed by Deletion of an RNA-Binding Region in CTCF.* Mol Cell, 2019. **76**(3): p. 395-411.e13.
499. Yao, H., et al., *Mediation of CTCF transcriptional insulation by DEAD-box RNA-binding protein p68 and steroid receptor RNA activator SRA.* Genes Dev, 2010. **24**(22): p. 2543-55.
500. Saldaña-Meyer, R., et al., *CTCF regulates the human p53 gene through direct interaction with its natural antisense transcript, Wrap53.* Genes Dev, 2014. **28**(7): p. 723-34.
501. Yang, F., et al., *The lncRNA Firre anchors the inactive X chromosome to the nucleolus by binding CTCF and maintains H3K27me3 methylation.* Genome Biol, 2015. **16**(1): p. 52.
502. Amaral, P.P., et al., *Genomic positional conservation identifies topological anchor point RNAs linked to developmental loci.* Genome Biol, 2018. **19**(1): p. 32.
503. Kong, F., et al., *Long noncoding RNA RMRP upregulation aggravates myocardial ischemia-reperfusion injury by sponging miR-206 to target ATG3 expression.* Biomed Pharmacother, 2019. **109**: p. 716-725.
504. Hansen, T.B., *Improved circRNA Identification by Combining Prediction Algorithms.* Frontiers in Cell and Developmental Biology, 2018. **6**(20).
505. Liu, X., et al., *Identification of mecciRNAs and their roles in the mitochondrial entry of proteins.* Sci China Life Sci, 2020.
506. Wu, Z., et al., *Mitochondrial Genome-Derived circRNA mc-COX2 Functions as an Oncogene in Chronic Lymphocytic Leukemia.* Mol Ther Nucleic Acids, 2020. **20**: p. 801-811.
507. Toyoshima, Y., Y. Tanaka, and K. Satomi, *MELAS syndrome associated with a new mitochondrial tRNA-Val gene mutation (m.1616A>G).* BMJ Case Rep, 2017. **2017**.

508. Lin, F., et al., *[Network correlation of circRNA-miRNA and the possible regulatory mechanism in acute myocardial infarction].* Zhonghua Yi Xue Za Zhi, 2018. **98**(11): p. 851-854.
509. Garalde, D.R., et al., *Highly parallel direct RNA sequencing on an array of nanopores.* Nat Methods, 2018. **15**(3): p. 201-206.
510. Echtermeyer, F., et al., *Syndecan-4 signalling inhibits apoptosis and controls NFAT activity during myocardial damage and remodelling.* Cardiovasc Res, 2011. **92**(1): p. 123-31.
511. Ling, S.S.M., et al., *Ankyrin Repeat Domain 1 Protein: A Functionally Pleiotropic Protein with Cardiac Biomarker Potential.* Int J Mol Sci, 2017. **18**(7).
512. Montiel, V., et al., *Inhibition of aquaporin-1 prevents myocardial remodeling by blocking the transmembrane transport of hydrogen peroxide.* Sci Transl Med, 2020. **12**(564).
513. Sato, T., et al., *mRNA-binding protein tristetraprolin is essential for cardiac response to iron deficiency by regulating mitochondrial function.* Proceedings of the National Academy of Sciences, 2018. **115**(27): p. E6291.
514. Teng, G.Z. and J.F. Dawson, *The Dark Side of Actin: Cardiac actin variants highlight the role of allostery in disease development.* Arch Biochem Biophys, 2020. **695**: p. 108624.
515. Çağlayan, A.O., et al., *ALPK3 gene mutation in a patient with congenital cardiomyopathy and dysmorphic features.* Cold Spring Harb Mol Case Stud, 2017. **3**(5).
516. Cambier, L., et al., *Fem1a is a mitochondrial protein up-regulated upon ischemia-reperfusion injury.* FEBS Lett, 2009. **583**(10): p. 1625-30.
517. Bock, F.J. and S.W.G. Tait, *Mitochondria as multifaceted regulators of cell death.* Nature Reviews Molecular Cell Biology, 2020. **21**(2): p. 85-100.
518. Ramachandra, C.J.A., et al., *Mitochondria in acute myocardial infarction and cardioprotection.* EBioMedicine, 2020. **57**: p. 102884.
519. Kobayashi, K., et al., *Dynamics of angiogenesis in ischemic areas of the infarcted heart.* Scientific Reports, 2017. **7**(1): p. 7156.
520. Fuhrmann, D.C. and B. Brüne, *Mitochondrial composition and function under the control of hypoxia.* Redox Biol, 2017. **12**: p. 208-215.
521. Rao, M., et al., *Circular RNA profiling in plasma exosomes from patients with gastric cancer.* Oncol Lett, 2020. **20**(3): p. 2199-2208.
522. Li, Y., et al., *Circular RNA is enriched and stable in exosomes: a promising biomarker for cancer diagnosis.* Cell Res, 2015. **25**(8): p. 981-4.
523. Zhao, Q., et al., *Targeting Mitochondria-Located circRNA SCAR Alleviates NASH via Reducing mROS Output.* Cell, 2020. **183**(1): p. 76-93.e22.
524. Ouyang, Q., et al., *Using plasma circRNA_002453 as a novel biomarker in the diagnosis of lupus nephritis.* Mol Immunol, 2018. **101**: p. 531-538.
525. Chen, F., et al., *Circular RNAs expression profiles in plasma exosomes from early-stage lung adenocarcinoma and the potential biomarkers.* J Cell Biochem, 2020. **121**(3): p. 2525-2533.
526. Zhang, R.F., et al., *LncRNA UCA1 affects osteoblast proliferation and differentiation by regulating BMP-2 expression.* Eur Rev Med Pharmacol Sci, 2019. **23**(16): p. 6774-6782.
527. Palmer, S.C., et al., *Regional clearance of amino-terminal pro-brain natriuretic peptide from human plasma.* Eur J Heart Fail, 2009. **11**(9): p. 832-9.

Appendix A

Summary statistics of HVOL, CDCS HF – and CDCS HF + patient groups

Summary statistics to review the matching of patient and control groups.

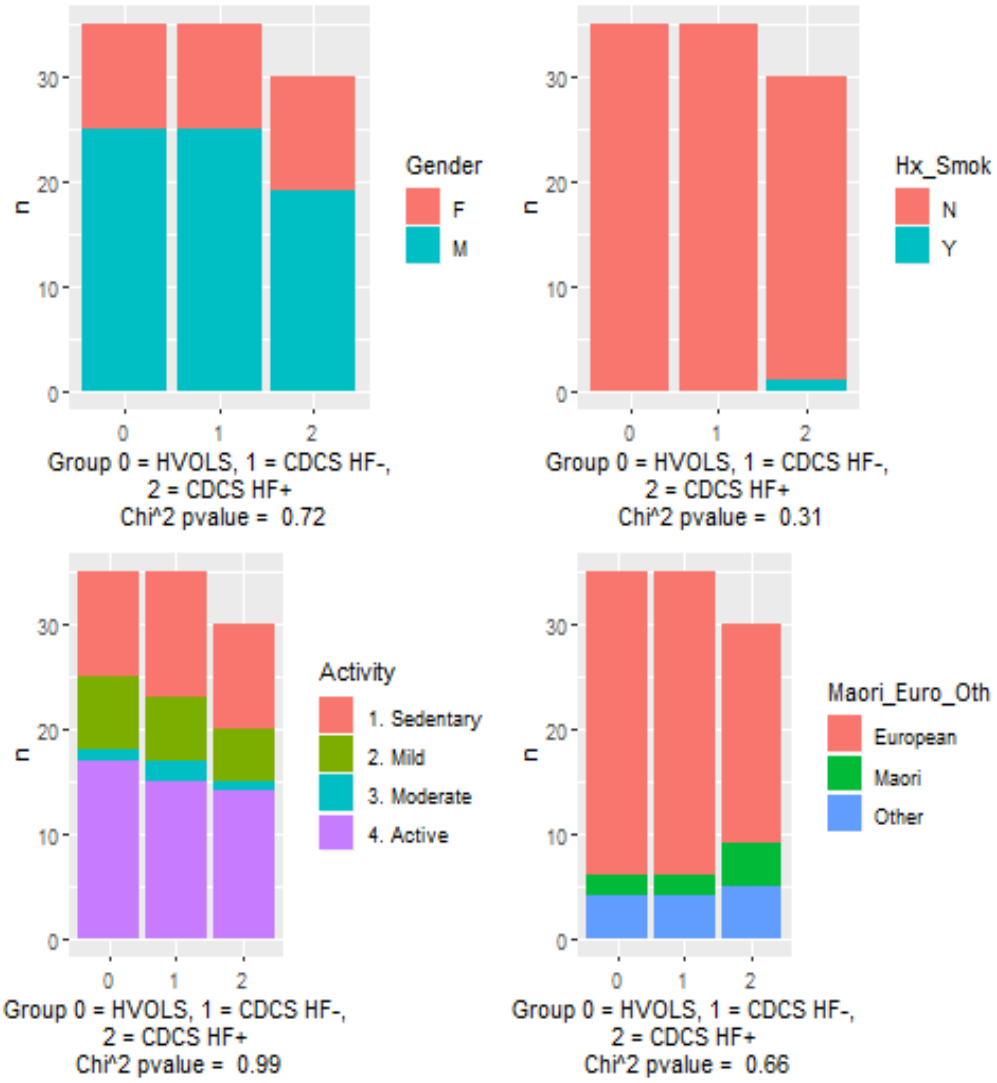


Figure A-1 Chi-square tests carried out for gender, physical activity, history of smoking and ethnicity

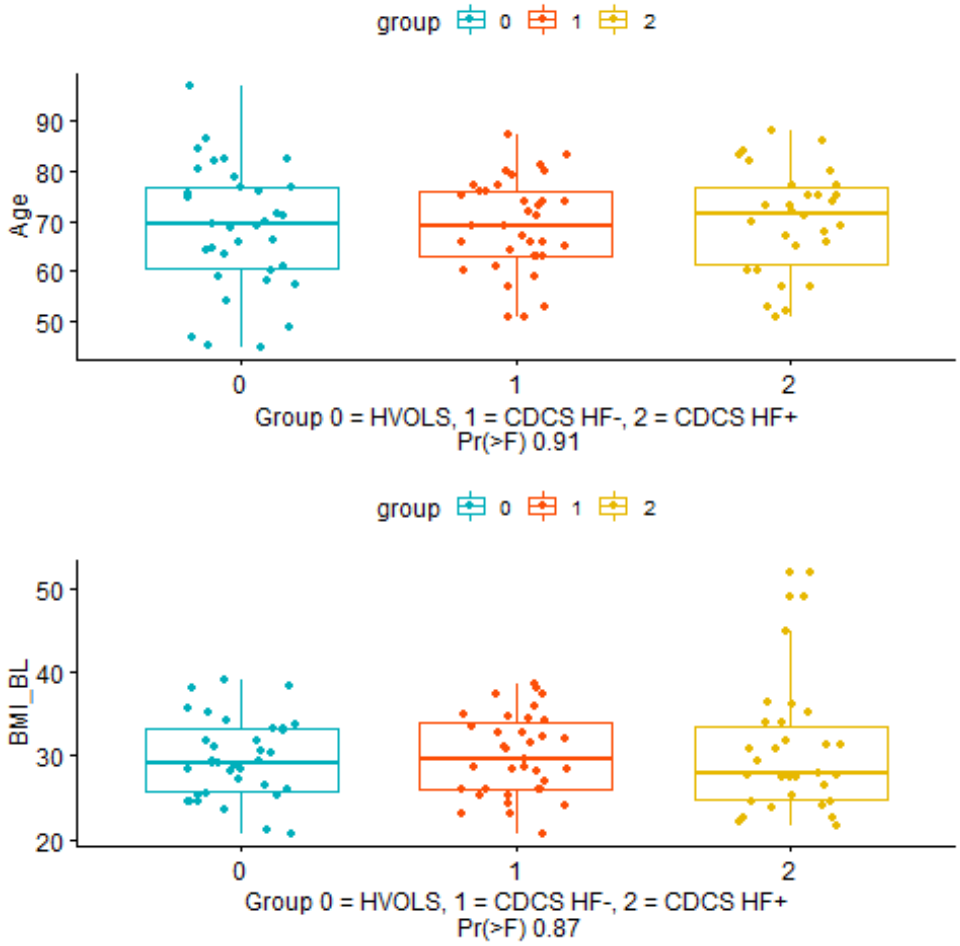
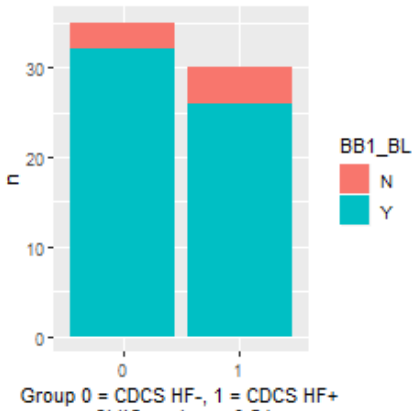
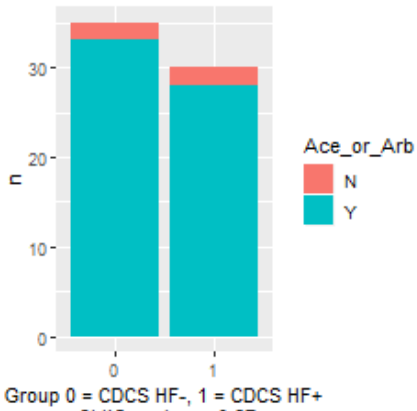


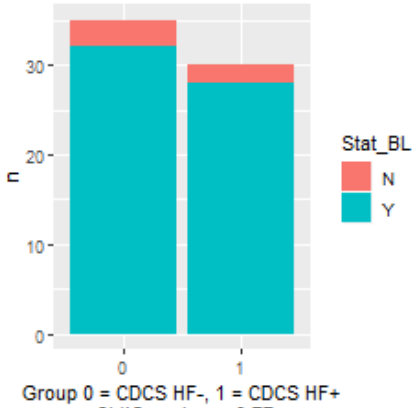
Figure A-2 ANOVA tests were carried out for age and body mass index (BMI) at baseline.



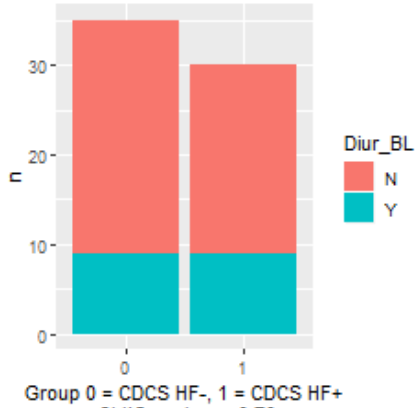
Group 0 = CDCS HF-, 1 = CDCS HF+
Chi^2 pvalue = 0.54



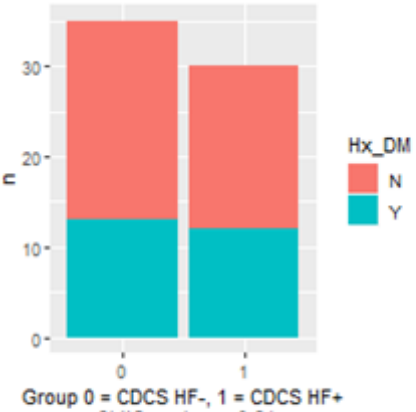
Group 0 = CDCS HF-, 1 = CDCS HF+
Chi^2 pvalue = 0.87



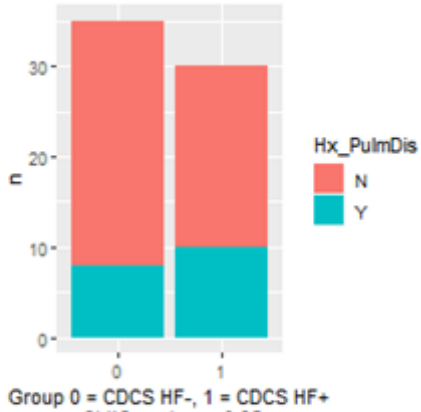
Group 0 = CDCS HF-, 1 = CDCS HF+
Chi^2 pvalue = 0.77



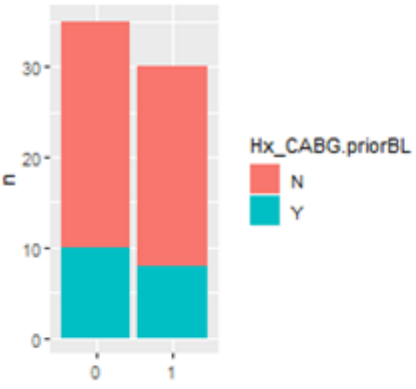
Group 0 = CDCS HF-, 1 = CDCS HF+
Chi^2 pvalue = 0.70



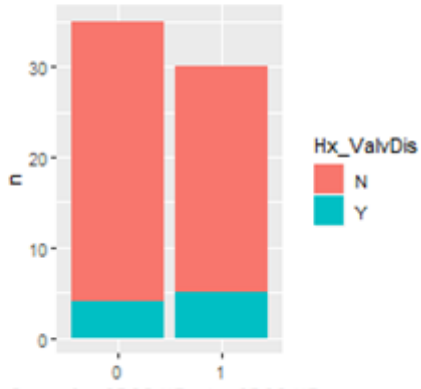
Group 0 = CDCS HF-, 1 = CDCS HF+
Chi^2 pvalue = 0.81



Group 0 = CDCS HF-, 1 = CDCS HF+
Chi^2 pvalue = 0.35



Group 0 = CDCS HF-, 1 = CDCS HF+
Chi^2 pvalue = 0.86



Group 0 = CDCS HF-, 1 = CDCS HF+
Chi^2 pvalue = 0.54

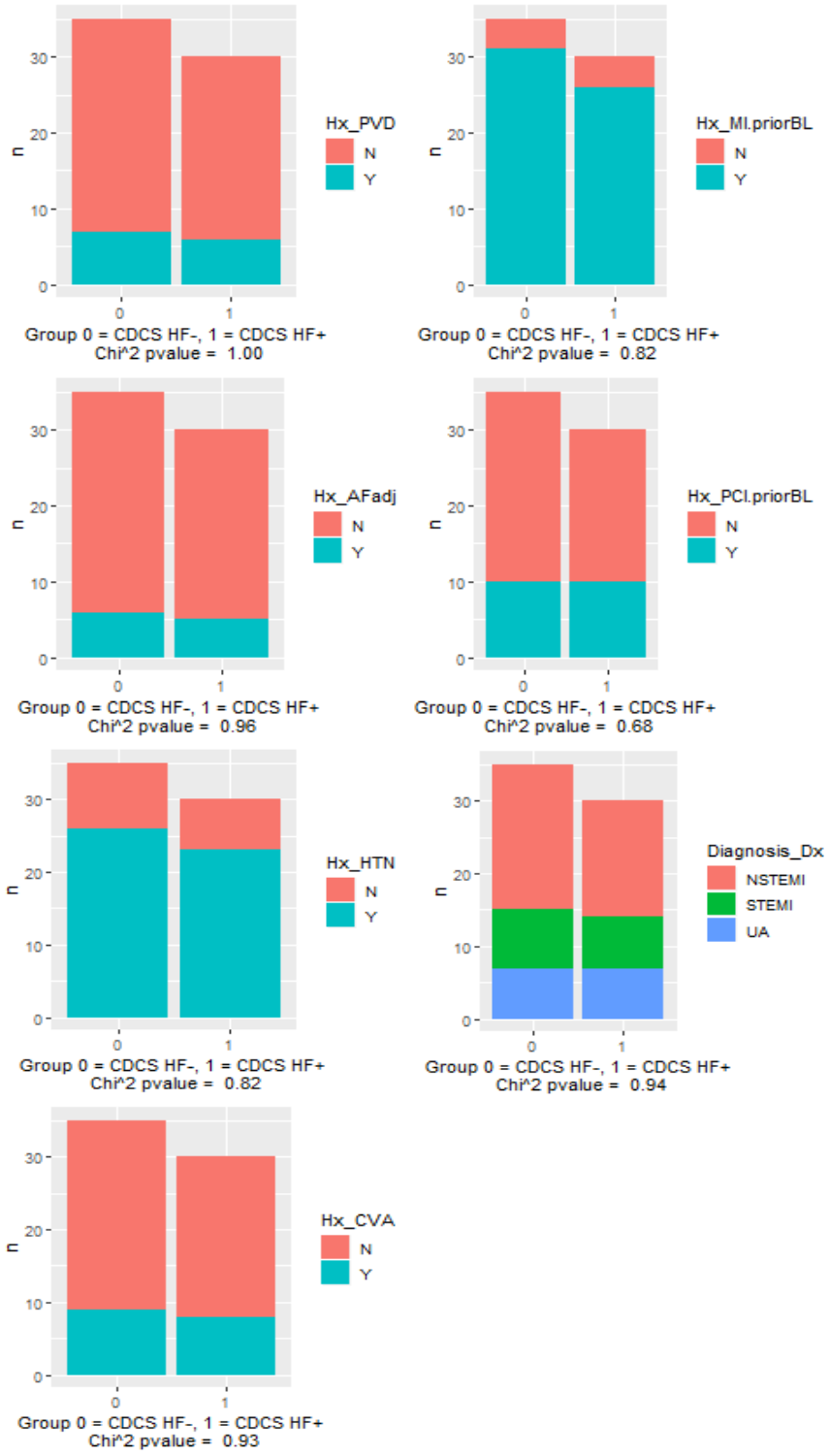


Figure A-3 Chi-square tests for the selected variables present in the CDCS groups.

BB1, Beta Blocker 1, ACE, angiotensin converting enzyme, ARB, angiotensin-receptor blockers, Stat – Statins, Diur – Diuretics, Hx_DM – History of Diabetes mellitus, Hx_PulmDis- History of pulmonary disorder, Hx_CABG – History of Coronary artery bypass grafting, Hx_ValvDis – History of (heart) valve disorder, Hx_PVD – History of Peripheral vascular disease, Hx_MI.priorBL – History of myocardial infarction prior to baseline, , Hx_AFadj – History of Atrial Fibrillation, Hx_PCI.priorBL - History of Percutaneous Coronary Intervention prior to baseline, HxHTN – History of hypertension, Hx_CVA – History of cerebrovascular accident.

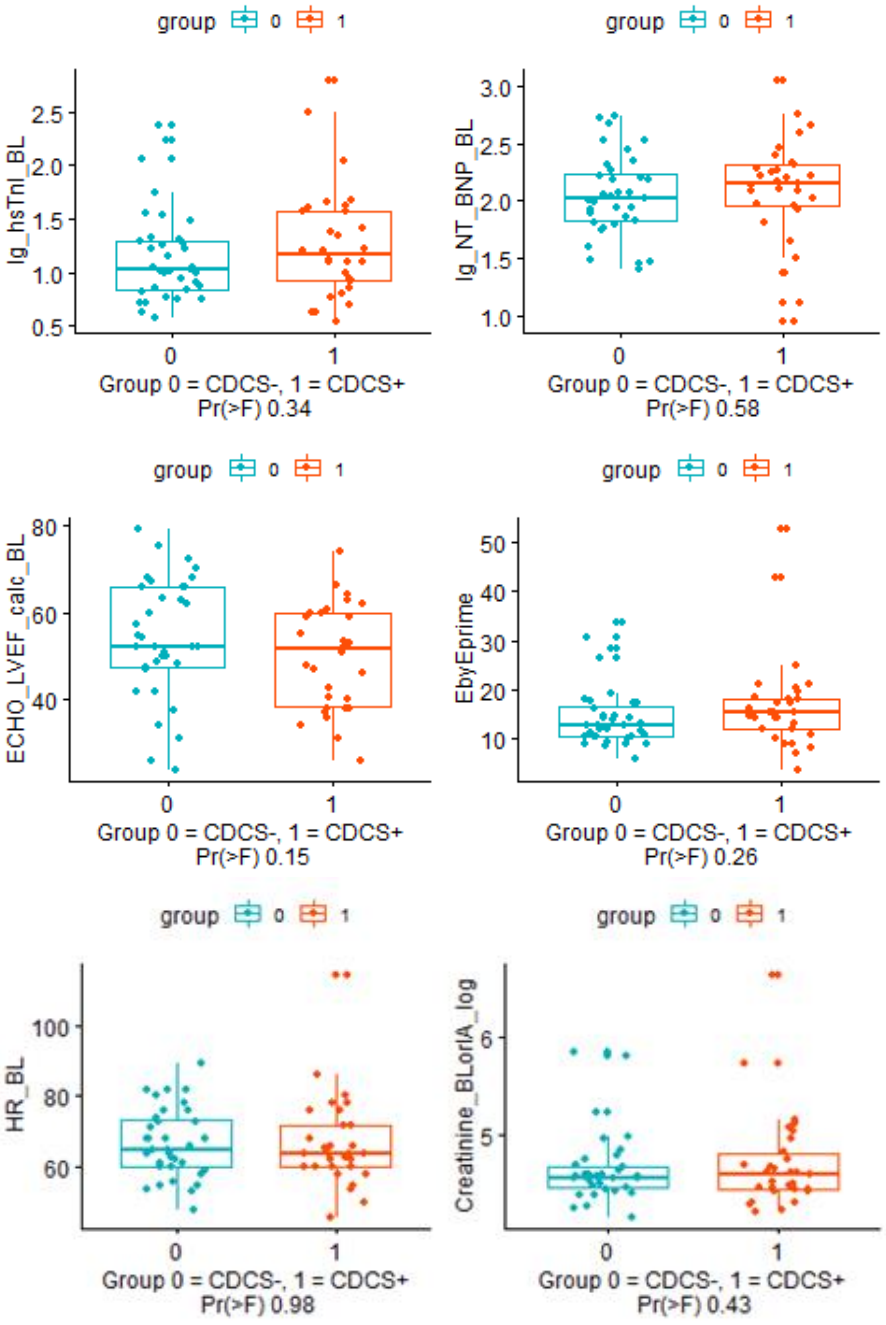


Figure A-4 ANOVA tests for the selected variables present in the CDCS groups.

hsTni, High sensitivity Troponin at baseline, NT-BNP, N-Terminal Brain Natriuretic Peptide at baseline, LVEF, Left Ventricular Ejection Volume at baseline, EbyEprime - ECHO index of diastolic dysfunction, HR_BL – Heart rate at baseline, Creatinine_BLorIA – Creatinine at baseline or index admission

Appendix B

B-1 Laboratory Methods

B1-1 RNA extraction from tissue

For the validation experiment using Nanopore sequencing (Section 3.4), RNA from the Cleveland donor heart samples was used (previously extracted by Dr Anna Pilbrow). Left ventricular tissue from donor hearts was broken into small blocks (100-150 mg) in liquid nitrogen using a pestle and mortar and placed into pre-chilled tubes on dry ice. Automated grinding was performed using a Retsch Mixer Mill MM301, (Haan, Germany) at a frequency of 30Hz for 10 minutes in 800µL pre-chilled TRIzol® (Invitrogen, Carlsbad, CA). Samples were mixed with 160µL chloroform by vigorous shaking for 15 seconds, incubated at room temperature for 2-3 minutes and then centrifuged at 12,000g for 15 minutes. The supernatant, containing the RNA, was transferred to a fresh 1.5mL Eppendorf tube.

RNA clean-up was performed with the Norgen Biotek CleanAll Kit according to manufacturer's instructions (cat #23800, Norgen Biotek Corporation, Thorold, Canada). This kit provides a rapid method for purification, clean-up, and concentration of RNA. RNase-free 70% ethanol (500µl) was added to the supernatant and vortexed for 10 seconds. This was applied to a 600µl spin column with collection tube and centrifuged at 14,000g for 1 minute, to bind the RNA. The flow-through was discarded and the collection tube and column reassembled. This was repeated until the entire ethanol-supernatant mix had been loaded onto the column. 500µl of wash solution was added and the column was centrifuged at 14,000g for 1 minute. This step was repeated to wash the column a second time. A third wash was carried out by adding 500µl of wash solution and the column centrifuged for 14,000g for 2 minutes. The column was transferred to a fresh elution tube. 50µl of elution buffer was added and centrifuged for at 200g for 2 minutes then at 1,400g for 1 minute. RNA (1.5µl) was quantified using the Nanodrop 8000 (ThermoFisher Scientific Inc, Waltham, USA).

B-2 RNA library preparation, the SMARTer® Stranded Total RNA-Seq Kit v2 – Pico Input Mammalian (Takara Bio, USA)

B2-1 First strand synthesis

There are two options for this part of the protocol – option 1 is for high quality RNA (and includes various time options for fragmentation of the RNA); option 2 is for RNA that are already degraded. As it was suspected that the RNA from the plasma samples would be partially degraded, option 2 (starting with degraded RNA) without fragmentation was used. Each RNA sample (8µl) was mixed with 1µl of SMART Pico Oligos Mix v2 on ice and incubated at 72°C in a preheated, hot-lid thermal cycler (Eppendorf® Mastercycler®, Hamburg, Germany) for 3 minutes, then immediately placed on an ice-cold PCR chiller rack for 2 minutes. The First-Strand Master Mix was prepared by mixing (per sample) 4µl 5x First-Strand Buffer, 4.5µl SMART TSO Mix v2, 0.5µl RNase Inhibitor and 2µl SMARTScribe Reverse Transcriptase. First-Strand Master Mix (11µl) was added to each reaction tube and vortexed for 2 seconds, and tubes were spun briefly to collect the contents. Tubes were incubated in a thermal cycler at 42°C for 90 minutes, 70°C for 10 minutes and then held at 4°C

B2-2 Addition of Illumina adapters and indexes

Next, the indexes (barcodes) used to distinguish samples from each other when sequencing a pooled library were added. A PCR master mix was prepared by mixing (per sample) 2µL nuclease-free water, 25 µl SeqAmp CB PCR Buffer (2X), 1µl SeqAmp DNA Polymerase, and 28µl of this mix was added to each sample. Each 5’ and 3’ PCR Primer HT was added (1µl), and each sample was vortexed gently and spun briefly. The tubes were placed in a preheated hot-lid thermal cycler and PCR was performed using the following conditions: 94°C for 1 minute (initial denaturation), followed by 5 cycles of 98°C for 15 seconds (denaturation),

55°C for 15 seconds (annealing) and 68°C for 30 seconds (extension), and a final extension of 68°C for 2 minutes. After thermal cycling samples were held at 4°C.

B2-3 Purification of the RNA-Seq Library Using AMPure Beads

AMPure beads (40µL/sample), were equilibrated at room temperature, added to each sample tube, mixed by pipetting ~10 times and incubated for 8 minutes to allow the DNA to bind to the beads. Tubes were spun briefly then placed onto a magnetic separation device for 5 minutes until the solution was totally clear. While the tubes were on the magnetic device, the supernatant was removed and discarded and 200µl of freshly made 80% ethanol was added without disturbing the beads, to wash away contaminants. After 30 seconds the supernatant was carefully removed and discarded, and the wash step was repeated. Tubes were briefly spun at 2,000g to collect the remaining ethanol, placed onto the magnetic separation device for a further 30 seconds, and the remaining ethanol was carefully removed. The tubes were opened at room temperature for 3–5 minutes until the pellets were dry. The cDNA was eluted from the beads by resuspended them in 52µl of nuclease-free water and mixed thoroughly by pipetting until all beads were washed off the side of the tube. Tubes were incubated for 5 minutes at room temperature to rehydrate then briefly spun and placed back onto the magnetic separation device for 1 minute or until the solution was clear. Supernatant (50µl) was transferred to a new tube and mixed well by vortexing, with a fresh aliquot (40µL) of AMPure beads. Tubes were incubated for 8 minutes to allow the DNA to bind to the beads.

B2-4 Depletion of Ribosomal cDNA with ZapR v2 and R-Probes v2

Library fragments originating from rRNA (18S and 28S) and mitochondrial rRNA were depleted with ZapR v2 in the presence of R-Probes v2 (mammalian-specific). The R-probes hybridise to rRNA and mitochondrial rRNA. After AMPure bead purification, tubes were briefly spun and placed on the magnetic separation device for 5 minutes or longer, until the solution was completely clear. Following 5 minutes incubation on the magnetic separation device, whilst the tubes were still on the magnet, the supernatant was removed and discarded,

and 200 µl of freshly made 80% ethanol was added to each sample. After 30 seconds the supernatant was removed and discarded, and the wash step was repeated. The tubes were spun briefly (~2,000g) and placed on the magnetic separation device for 30 seconds to remove any remaining ethanol as described above. The open sample tubes were rested at room temperature until the pellets were dry. A ZapR master mix was prepared by combining the following reagents at room temperature in order: 16.8 µl Nuclease-Free Water, 2.2 µl 10X ZapR Buffer, 1.5 µl ZapR v2 and lastly 1.5 µl 'activated' R-Probes v2 that had been incubated in a preheated hot-lid thermal cycler at 72°C for 2 minutes and held at 4°C for at least 2 minutes. The dried, DNA-bound AMPure beads were resuspended in 22 µl of the ZapR master mix with thorough mixing. Tubes were incubated at room temperature for 5 minutes to rehydrate the beads, briefly spun to collect the liquid and placed on the magnetic separation device for 1 minute or longer, until the solution was completely clear. Supernatant (20 µl) was removed and transferred to a new PCR tube. Tubes were incubated in a preheated hot-lid thermal cycler at 37°C for 60 minutes followed by 72°C for 10 minutes then held at 4°C.

B2-5 Final RNA-Seq Library Amplification

Library fragments not cleaved by the ZapR reaction were further enriched in a second round of PCR. A PCR master mix was prepared by adding (per sample) 26 µl Nuclease-Free Water, 50 µl SeqAmp CB PCR Buffer, 2 µl PCR2 Primers v2 and 2 µl SeqAmp DNA Polymerase. 80µl of master mix was added to each sample tube, mixed by gently tapping then spun down. Tubes were incubated in a preheated hot-lid thermal cycler under the following conditions: 94°C for 1 minute (initial denaturation), followed by 16 cycles of 98°C for 15 seconds (denaturation), 55°C for 15 seconds (annealing), 68°C for 30 seconds (extension), and then held at 4°C.

B2-6 Purification of Final RNA-Seq Library Using AMPure Beads

The amplified RNA-seq library then went through a purification step with AMPure beads. Essentially the process was the same as section 2.5.3 but instead of 40µ/sample of AMPure beads, 100µl AMPure beads were added to each sample. Also, the final elution was into12µl of Tris buffer. The samples were then sent to the Ramaciotti for sequencing on the NovaSeq6000.

Appendix C

mRNAs and lncRNA genes and novel lncRNA transcripts associated with ischaemia in heart tissue

Table C-1 The 20 most up- and down-regulated mRNAs in response to ischemia in human left ventricle.

Gene	Ensembl id	log2 Fold Change	p-adjusted	Median TPM
FOS	ENSG00000170345.9	2.39	1.21E-21	13.75
S100A8	ENSG00000143546.9	1.86	5.96E-22	2.06
HBB	ENSG00000244734.4	1.74	5.22E-40	130.01
HBA2	ENSG00000188536.13	1.66	6.73E-39	156.32
HBA1	ENSG00000206172.8	1.54	5.74E-32	81.77
S100A9	ENSG00000163220.10	1.43	9.61E-19	4.72
CXCL2	ENSG00000081041.8	1.41	1.14E-22	5.05
ZFP36	ENSG00000128016.5	1.20	2.49E-16	19.67
SOCS3	ENSG00000184557.4	1.15	6.25E-10	2.75
GADD45B	ENSG00000099860.8	0.97	1.45E-19	9.05
C2CD4B	ENSG00000205502.3	0.96	2.34E-14	1.08
SLC11A1	ENSG00000018280.16	0.96	2.88E-12	1.13
NPPA	ENSG00000175206.10	0.92	6.53E-06	20.15
GJA5	ENSG00000265107.2	0.86	1.58E-15	2.94
AC118553.2	ENSG00000283761.1	0.84	3.64E-08	1.58
JUNB	ENSG00000171223.5	0.83	5.41E-08	14.03
DUSP1	ENSG00000120129.5	0.80	6.84E-13	42.84
NRGN	ENSG00000154146.12	0.80	4.35E-25	3.05
KLF4	ENSG00000136826.14	0.79	6.23E-19	8.40

BEST1	ENSG00000167995.15	0.78	5.89E-26	1.80
LYVE1	ENSG00000133800.8	-0.89	3.72E-12	4.55
CD68	ENSG00000129226.13	-0.94	4.89E-12	5.29
MLXIPL	ENSG00000009950.15	-0.95	2.54E-11	1.62
AC005943.1	ENSG00000267059.2	-0.96	1.93E-08	7.94
ZNF385B	ENSG00000144331.19	-0.98	4.89E-12	1.86
PROB1	ENSG00000228672.3	-0.98	1.02E-05	1.70
PTPRF	ENSG00000142949.16	-1.00	2.08E-11	1.88
CFB	ENSG00000243649.8	-1.02	1.69E-08	2.80
TMEM37	ENSG00000171227.6	-1.02	5.02E-14	0.88
SCD	ENSG00000099194.5	-1.05	3.50E-08	1.30
HP	ENSG00000257017.8	-1.06	1.79E-07	3.78
KRT18	ENSG00000111057.10	-1.14	1.93E-12	2.22
C3	ENSG00000125730.16	-1.18	1.37E-13	36.28
AC026464.4	ENSG00000260914.3	-1.24	8.96E-17	3.91
TGM1	ENSG00000092295.11	-1.25	7.08E-09	0.79
NUDT4B	ENSG00000177144.7	-1.33	8.40E-14	15.01
CHRDL1	ENSG00000101938.14	-1.34	3.79E-12	1.08
CCL21	ENSG00000137077.7	-1.68	3.01E-14	5.60
PRG4	ENSG00000116690.12	-2.09	1.31E-16	1.92
PLA2G2A	ENSG00000188257.11	-2.12	3.70E-17	6.23

Table C-2 The 20 most up- and down- regulated annotated lncRNAs in response to ischemia in human left ventricle.

Gene	Ensembl id	log2 Fold Change	p-adjusted	Median TPM
AP005329.1	ENSG00000264235.5	1.63	3.44E-50	5.53
AL583722.2	ENSG00000258430.1	1.43	7.19E-09	6.77
AC084880.3	ENSG00000256609.1	1.26	1.21E-31	1.44
AL137186.2	ENSG00000232807.2	1.25	4.58E-15	4.23
AC241644.3	ENSG00000274415.1	0.89	3.35E-08	1.82
AC087588.2	ENSG00000274976.1	0.75	1.36E-16	4.06
ACTA2-AS1	ENSG00000180139.11	0.75	1.09E-12	1.70
UTAT33	ENSG00000231851.5	0.74	1.53E-07	0.99
AC012615.2	ENSG00000267007.1	0.71	7.97E-06	2.46
VASH1-AS1	ENSG00000258301.3	0.70	2.06E-14	3.82
CRNDE	ENSG00000245694.9	0.69	3.32E-13	7.29
AC135178.1	ENSG00000226871.1	0.69	6.22E-11	1.29
AC008040.5	ENSG00000268220.1	0.65	1.24E-06	1.94
AL136164.4	ENSG00000279312.1	0.65	2.35E-07	1.62
AC091588.3	ENSG00000266283.1	0.63	6.10E-24	1.48
AC009087.1	ENSG00000260252.1	0.62	0.0001	1.08
AC025569.1	ENSG00000258168.5	0.60	6.40E-11	0.96
SNHG25	ENSG00000266402.3	0.60	0.0004	4.32
AL512770.1	ENSG00000228302.2	0.58	3.83E-24	0.95
LINC01004	ENSG00000228393.3	0.58	2.50E-15	3.39
CTBP1-DT	ENSG00000196810.4	-0.46	3.49E-13	4.40
SH3RF3-AS1	ENSG00000259863.1	-0.46	2.29E-08	0.79
AP001972.5	ENSG00000279117.1	-0.47	0.0009	1.84

AC011476.3	ENSG00000267265.5	-0.47	4.03E-09	3.93
LINC00957	ENSG00000235314.1	-0.47	3.23E-13	3.96
AC010980.2	ENSG00000267034.1	-0.47	4.61E-07	1.07
AC090515.4	ENSG00000259353.1	-0.48	2.51E-08	1.36
AF111167.2	ENSG00000259319.1	-0.48	2.08E-21	5.75
AL365259.1	ENSG00000237742.6	-0.48	3.09E-12	1.78
AL450326.1	ENSG00000230555.2	-0.50	1.82E-14	2.41
DNAAF4-CCPG1	ENSG00000261771.5	-0.51	6.17E-06	1.95
AC090644.1	ENSG00000285731.1	-0.52	0.088066	1.04
LINC01018	ENSG00000250056.5	-0.53	0.057317	1.60
AC113189.2	ENSG00000262880.1	-0.57	0.084906	1.16
AC022440.1	ENSG00000285914.1	-0.59	0.091336	3.19
AL021068.1	ENSG00000213062.4	-0.61	0.09738	0.89
LINC02432	ENSG00000248810.1	-0.72	0.108691	1.47
RORB-AS1	ENSG00000224825.2	-0.78	0.081297	1.69
AL158152.1	ENSG00000269929.1	-0.85	0.072942	1.92
AC008760.2	ENSG00000276980.1	-0.87	0.126728	1.80

Table C-3 The 20 most up-and down-regulated novel lncRNAs in response to ischemia in human left ventricle.

MSTRG_ID	Chromosome	Start	Stop	Strand	Class code	Nearest Gene	# exons	log2 Fold Change	p-adjusted	Median TPM
MSTRG.131093.1	chr20	45325338	45325752	+	x	SDC4	2	4.173926	2.12E-15	0.54
MSTRG.86175.11	chr17	447902	489122	+	u	-	3	2.917532	3.05E-06	8.72
MSTRG.104032.1	chr19	39406895	39409136	-	x	ZFP36	4	2.457344	1.63E-45	1.50
MSTRG.83300.1	chr16	58046102	58046839	-	x	MMP15	2	2.152208	1.11E-14	3.34
MSTRG.83567.1	chr16	58707134	58708232	+	x	GOT2	2	2.073946	1.43E-14	13.30
MSTRG.27879.1	chr10	90141461	90921003	+	x	ANKRD1	3	1.89843	3.87E-19	1.28
MSTRG.205582.1	chr7	30895368	30925524	-	x	AQP1	2	1.844036	5.94E-06	1.82
MSTRG.100871.1	chr19	3122969	3123965	-	x	GNA11	2	1.792638	2.22E-07	2.61
MSTRG.43021.2	chr12	4303785	4305242	-	i	AC008012.1	2	1.782057	6.61E-11	10.70
MSTRG.104750.46	chr19	45306413	45307690	+	x	CKM	2	1.772354	3.03E-05	0.62
MSTRG.73900.1	chr15	40035995	40036576	+	x	SRP14	2	1.76072	1.01E-06	0.78
MSTRG.18214.1	chr1	228097691	228098964	-	x	ARF1	2	1.508408	6.64E-14	15.30
MSTRG.9631.7	chr1	109669340	109675230	-	x	GSTM2	2	1.407911	1.14E-05	5.54
MSTRG.35134.1	chr11	46677392	46690160	+	x	ARHGAP1	2	1.349519	7.75E-08	6.06

MSTRG.117955.1	chr2	133549002	133615658	+	x	NCKAP5	2	1.220788	0.000273	2.54
MSTRG.1633.1	chr1	12006733	12013252	-	x	MFN2	3	1.220345	2.29E-17	2.50
MSTRG.43173.1	chr12	6236930	6241095	-	x	CD9	2	1.210054	1.57E-13	0.62
MSTRG.54028.2	chr12	119178965	119180265	-	x	HSPB8	2	1.116854	4.43E-17	0.61
MSTRG.192545.1	chr6	44250403	44253799	-	x	HSP90AB1	3	1.107886	9.04E-25	0.52
MSTRG.1910.25	chr1	16013977	16018044	+	x	HSPB7	3	1.090551	1.03E-11	6.55
MSTRG.70918.1	chr14	92940915	92941414	+	i	ITPK1	2	-1.04909	9.31E-07	6.53
MSTRG.124395.1	chr2	206075916	206182843	+	x	INO80D	3	-1.06122	6.46E-06	1.64
MSTRG.27420.1	chr10	80155701	80169347	+	x	ANXA11	2	-1.09287	1.66E-11	33.00
MSTRG.104154.5	chr19	40315848	40319525	+	u	-	2	-1.09853	1.39E-09	1.39
MSTRG.11996.12	chr1	160213705	160214998	-	x	PEA15	3	-1.13159	2.52E-10	14.70
MSTRG.74119.69	chr15	42413380	42417167	+	x	ZNF106	4	-1.1572	3.13E-12	18.50
MSTRG.18420.37	chr1	229431992	229432867	+	x	ACTA1	3	-1.23328	4.09E-05	1.06
MSTRG.130515.14	chr20	38128922	38130073	+	x	TGM2	3	-1.24771	0.000105	12.70
MSTRG.184244.1	chr5	137807996	137809722	-	x	NPY6R	2	-1.24808	6.31E-09	8.08
MSTRG.191314.4	chr6	30704436	30705720	+	x	MDC1	3	-1.26659	2.55E-07	1.46

MSTRG.1365.2	chr1	8360724	8361221	+	i	RERE	2	-1.31139	1.76E-11	6.79
MSTRG.192637.2	chr6	45902065	45902966	+	i	CLIC5	2	-1.37	3.62E-05	6.71
MSTRG.36725.1	chr11	67424160	67424917	-	i	CARNS1	2	-1.43	2.52E-06	2.14
MSTRG.93175.6	chr17	75845805	75847288	+	x	WBP2	3	-1.45	1.70E-08	10.2
MSTRG.142094.1	chr3	30691524	30693583	-	x	TGFBR2	3	-1.53	0.014268	0.61
MSTRG.86360.1	chr17	1714053	1715842	+	i	MIR22HG	2	-1.69	3.53E-47	0.89
MSTRG.214045.1	chr7	135928089	135928528	+	x	MTPN	2	-1.79	3.57E-05	1.45
MSTRG.155590.4	chr3	196234960	196236269	+	x	PCYT1A	2	-1.84	0.000114	1.25
MSTRG.68787.1	chr14	69352103	69353652	-	i	GALNT16	4	-2.27	7.24E-11	1.85
MSTRG.77609.2	chr15	84857074	84858196	-	x	ALPK3	2	-2.86	1.30E-12	5

Table C-4. Novel lncRNA transcripts originally identified from Illumina short read RNA-Seq and validated with long read Nanopore technology. After comparisons with updated annotations 11 lncRNAs were considered truly novel (highlighted in red).

MSTRG_ID	C'me	Start	Stop	log2FoldChange	padj	txLength	txClass	# Exons	# Samples Null	# Samples Not null	mean TPM
MSTRG.8333.38	**chr1**	**95247358**	**95256066**	**0.20**	**0.005**	**4158**	**i**	**2**	**0**	**162**	**0.36**
MSTRG.10127.1	chr1	115356664	115364913	0.07	0.74	402	i	3	29	133	0.51
MSTRG.10265.1	**chr1**	**117128696**	**117143589**	**-0.43**	**7.27E-13**	**4121**	**x**	**2**	**0**	**162**	**1.68**
MSTRG.10779.1	chr1	146387133	146388149	0.02	0.79	794	i	2	3	159	1.66
MSTRG.18670.2	chr1	231795748	231854307	0.12	0.00	25443	i	3	0	162	0.50
MSTRG.18670.5	chr1	231813360	231848025	-0.63	1.21E-10	1594	i	3	0	162	2.61
MSTRG.18670.6	chr1	231813373	231848080	-0.54	0.00	1481	i	2	3	159	0.97
MSTRG.109423.2	**chr2**	**37489457**	**37605898**	**0.21**	**0.30**	**945**	**y**	**3**	**21**	**141**	**0.22**
MSTRG.116760.1	chr2	119723168	119759827	0.06	0.99	438	x	3	56	106	0.31
MSTRG.139608.1	chr3	112054	131466	0.19	0.12	759	u	3	3	159	2.51
MSTRG.140910.1	**chr3**	**15894181**	**16137554**	**0.39**	**0.34**	**399**	**u**	**4**	**60**	**102**	**0.32**
MSTRG.148850.2	chr3	119666102	119674319	0.52	2.96E-21	4645	i	2	0	162	0.90

MSTRG.149055.1	**chr3**	**123335278**	**123338361**	**0.26**	**0.02**	**1183**	**i**	**3**	**1**	**161**	**0.47**
MSTRG.150894.1	chr3	147688931	147732462	0.03	0.99	310	u	2	43	119	0.49
MSTRG.161102.1	chr4	62930250	62953640	-0.30	0.00	9526	u	2	0	162	0.44
MSTRG.165657.4	chr4	114334360	114365102	-0.60	0.02	1078	u	3	27	135	0.40
MSTRG.165657.8	chr4	114334812	114364973	-0.09	0.65	563	u	4	5	157	2.96
MSTRG.171220.1	chr4	173692903	173699621	-0.47	1.42E-06	5519	u	2	1	161	1.80
MSTRG.181743.2	**chr5**	**107778856**	**107781422**	**0.11**	**0.66**	**841**	**u**	**2**	**35**	**127**	**0.34**
MSTRG.201880.1	**chr6**	**157328269**	**157363141**	**0.19**	**0.60**	**476**	**u**	**4**	**49**	**113**	**0.37**
MSTRG.224927.1	**chr8**	**94223489**	**94228144**	**-0.10**	**0.43**	**2298**	**u**	**3**	**0**	**162**	**0.30**
MSTRG.233013.2	chr9	38949702	38968353	-0.28	0.01	868	u	4	2	160	2.22
MSTRG.233013.1	chr9	38949702	38998504	-0.22	0.17	1048	u	4	10	152	1.39
MSTRG.233934.2	chr9	68947242	68975934	0.01	1	8161	i	3	0	162	6.02
MSTRG.29120.1	chr10	105460553	105473362	0.39	0.07	1260	u	2	40	122	0.16
MSTRG.29422.1	chr10	106648504	106677289	-0.12	0.45	1787	x	2	1	161	0.40
MSTRG.31895.1	chr11	6384901	6390323	0.02	0.87	731	u	2	45	117	0.24
MSTRG.40006.1	**chr11**	**104071819**	**104093201**	**-0.29**	**0.19**	**1516**	**i**	**4**	**19**	**143**	**0.33**

MSTRG.44658.2	chr12	20104401	20109918	0.25	0.28	697	u	3	10	152	1.02
MSTRG.50494.1	chr12	83456471	83472641	-0.59	0.00	2864	u	2	10	152	0.37
MSTRG.52812.1	chr12	104858841	104871639	0.07	0.67	1545	x	2	11	151	0.20
MSTRG.61980.1	chr13	90273741	90276422	0.26	0.28	685	u	2	32	130	0.49
MSTRG.65104.1	**chr14**	**24271210**	**24299055**	**-0.09**	**0.55**	**390**	**x**	**3**	**0**	**162**	**2.03**
MSTRG.65423.1	chr14	33958672	33965015	-0.01	0.91	1543	i	2	4	158	0.42
MSTRG.72507.1	chr15	23401880	23407745	-0.21	0.16	3015	u	2	2	160	0.17
MSTRG.79975.1	chr16	10621150	10622567	-0.23	0.29	573	u	2	39	123	0.35
MSTRG.80356.1	chr16	14486764	14492242	0.02	0.91	4575	i	2	0	162	1.10
MSTRG.240803.1	chrX	10039516	10045783	0.13	0.37	650	i	2	8	154	0.54
MSTRG.609.1	**KI270742.1**	**6**	**110029**	**0.31**	**3.29E-05**	**4805**	**u**	**3**	**0**	**162**	**0.55**

Table continued

MSTRG_ID	GWAS SNP	Associated SNP / MAF	Ensembl enhancer	Ensembl promoter	Ensembl promoter flank	Gencode v32	Noncode v5	FANTOM CAT	FANTOM data	Associated with (tissue type FANTOM)	FANTOM regulatory overlap
MSTRG.8333.38	**x**	x	✓	**x**	✓						
MSTRG.10127.1	x	x	✓	x	✓	AL512638.3		CATG00000072671.1	http://fantom.gsc.riken.jp/cat/v1/#/genes/CATG0	cardial valve	Enhancer

									0000072671.1		
MSTRG.10265.1	**x**	x	✓	**x**	✓						
MSTRG.10779.1	x	x	x	x	x	AC245407.2					
MSTRG.18670.2	x	x	✓	✓	✓	AL136171.2	NONHSAG004626.3	CATG00000098132.1	http://fantom.gsc.riken.jp/cat/v1/#/genes/CATG0000098132.1	cardial valve	Promoter
MSTRG.18670.5	x	x	✓	✓	✓	AL136171.2		CATG00000098132.1	http://fantom.gsc.riken.jp/cat/v1/#/genes/CATG0000098132.1	cardial valve	Promoter
MSTRG.18670.6	x	x	✓	✓	✓	AL136171.2	NONHSAG004626.3	CATG00000098132.1	http://fantom.gsc.riken.jp/cat/v1/#/genes/CATG0000098132.1	cardial valve	Promoter
MSTRG.109423.2	**x**	x	✓	✓	✓						
MSTRG.116760.1	x	x	✓	✓	✓			CATG00000049811.1	http://fantom.gsc.riken.jp/cat/v1/#/genes/CATG0000049811.1	neuron projection bundle	Promoter
MSTRG.139608.1	x	x	x	x	✓	CHL1-AS2	NONHSAG034274.2	CHL1-AS2	http://fantom.gsc.riken.jp/cat/v1/#/genes/ENSG00000224318.1	nervous system	

MSTRG.140910.1	x	x	✓	x	✓						
MSTRG.148850.2	x	x	x	x	✓	AC023494.1					
MSTRG.149055.1	x	x	x	x	✓						
MSTRG.150894.1	x	x	✓	x	✓		NONHSAG036322.2				
MSTRG.161102.1	x	x	✓	x	✓			CATG00000068763.1	http://fantom.gsc.riken.jp/cat/v1/#/genes/CATG00000068763.1	Cardiac chamber	
MSTRG.165657.4	x	x	✓	x	✓		NONHSAG038694.3				
MSTRG.165657.8	x	x	✓	x	✓		NONHSAG038694.3	CATG00000073509.1	http://fantom.gsc.riken.jp/cat/v1/#/genes/CATG00000073509.1	smooth muscle tissue/aortic smooth muscle cell	Enhancer
MSTRG.171220.1	x	x	✓	x	x		NONHSAG039364.2	CATG00000074547.1	http://fantom.gsc.riken.jp/cat/v1/#/genes/CATG00000074547.1		
MSTRG.181743.2	x	x	✓	x	x						
MSTRG.201880.1	x	x	✓	x	✓						
MSTRG.224927.1	x	x	✓	x	✓						

MSTRG.233013.2	x	x	✓	x	✓		NONHSAG052209.2				
MSTRG.233013.1	x	x	✓	x	✓	AL953883.1	NONHSAG052209.2				
MSTRG.233934.2	x	x	✓	x	✓		NONHSAG052505.2				
MSTRG.29120.1	x	x	x	x	x			CATG00000000711.1	http://fantom.gsc.riken.jp/cat/v1/#/genes/CATG0000000711.1		
MSTRG.29422.1	x	x	✓	x	x	AL133395.1		CATG00000116352.1	http://fantom.gsc.riken.jp/cat/v1/#/genes/CATG0000116352.1		Promoter
MSTRG.31895.1	x	x	✓	✓	✓		NONHSAG106428.1	CATG00000004684.1	http://fantom.gsc.riken.jp/cat/v1/#/genes/CATG0000004684.1	eosinophil	Promoter
MSTRG.40006.1	x	x	✓	x	✓						
MSTRG.44658.2	x	x	✓	x	x	AC126468.1	NONHSAG010640.3	CATG00000011418.1	http://fantom.gsc.riken.jp/cat/v1/#/genes/CATG0000011418.1	cardial valve	Enhancer
MSTRG.50494.1	x	x	✓	x	x		NONHSAG011825.2	CATG00000012608.1	http://fantom.gsc.riken.jp/cat/v1/#/genes/CATG0		Promoter

									0000012608.1		
MSTRG.52812.1	x	x	x	x	✓	AC089985.1		CATG00000010265.1	http://fantom.gsc.riken.jp/cat/v1/#/genes/CATG0000010265.1	cardiac muscle myoblast	Enhancer
MSTRG.61980.1	x	x	x	x	x			CATG00000017744.1	Not in database?		
MSTRG.65104.1	x	x	x	✓	✓						
MSTRG.65423.1	x	x	x	x	✓			CATG00000018504.1	Not in database?		
MSTRG.72507.1	✓	rs937741 / 0.3191	✓	x	x	AC100756.4	NONHSAG016274.2	CATG00000022295.1	http://fantom.gsc.riken.jp/cat/v1/#/genes/CATG0000022295.1	cardiac chamber	Promoter
MSTRG.79975.1	x	x	✓	x	✓		NONHSAG018574.2	CATG00000028691.1	http://fantom.gsc.riken.jp/cat/v1/#/genes/CATG0000028691.1		Enhancer
MSTRG.80356.1	x	x	x	x	x			CATG00000026724.1	http://fantom.gsc.riken.jp/cat/v1/#/genes/CATG0000026724.1		Enhancer
MSTRG.240803.1	x	x	x	x	✓			WWC3-AS1	http://fantom.gsc.riken.jp/cat/v1/#/g		Enhancer

									enes/ENSG00000225076.1		
MSTRG.609.1	**x**	x	**x**	**x**	**x**						

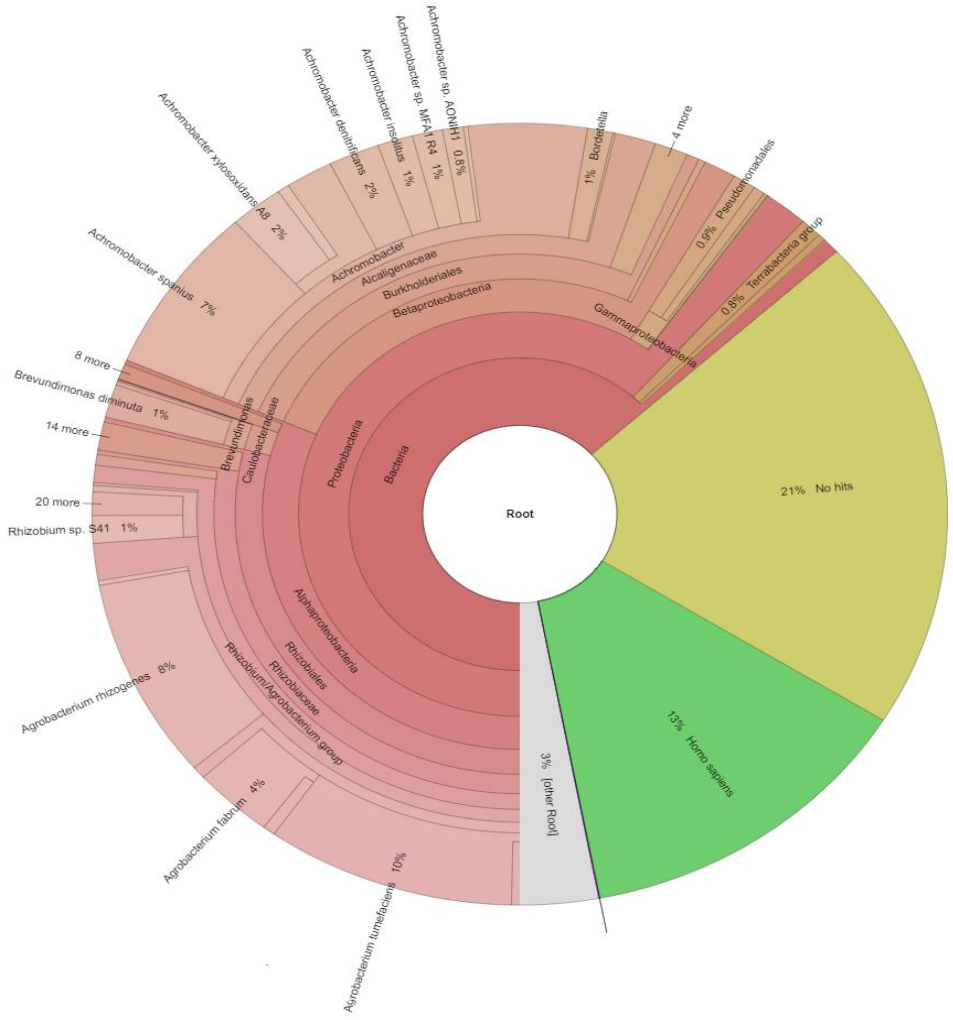
Appendix D

Plasma Pilot Studies

Table D-1. Alignment statistics from the BAM file using Samtools flagstat for pilot 4. Total reads aligning to the Human and Sequin reference combined and the Sequin reference alone, and the percentage of total reads aligning to Sequins are presented.

Sample ID	Total Reads aligned to both Human and Sequin reference	Total Reads aligned to Sequin reference	Percentage of reads aligning to Sequins (column 2/column 1)*100
HVOL plasma 3 (1% Sequins added)	280326	91654	33
CDCS plasma 2 (1% Sequins added)	500632	275424	55

A


Root
Bacteria
Proteobacteria
Alphaproteobacteria
Rhizobiales
Rhizobiaceae
Rhizobium/Agrobacterium group
Agrobacterium rhizogenes 8%
Agrobacterium fabrum 4%
Agrobacterium tumefaciens 10%
Rhizobium sp. S41 1%
20 more
14 more
Brevundimonas diminuta 1%
8 more
Brevundimonas
Caulobacteraceae
Betaproteobacteria
Burkholderiales
Alcaligenaceae
Achromobacter
Achromobacter spanius 7%
Achromobacter xylosoxidans A8 2%
Achromobacter denitrificans 2%
Achromobacter insolitus 1%
Achromobacter sp. MFA1 R4 1%
Achromobacter sp. AONIH1 0.8%
Gammaproteobacteria
Bordetella 1%
4 more
Pseudomonadales 0.9%
Terrabacteria group 0.8%
21% No hits
13% Homo sapiens
3% [other Root]

B

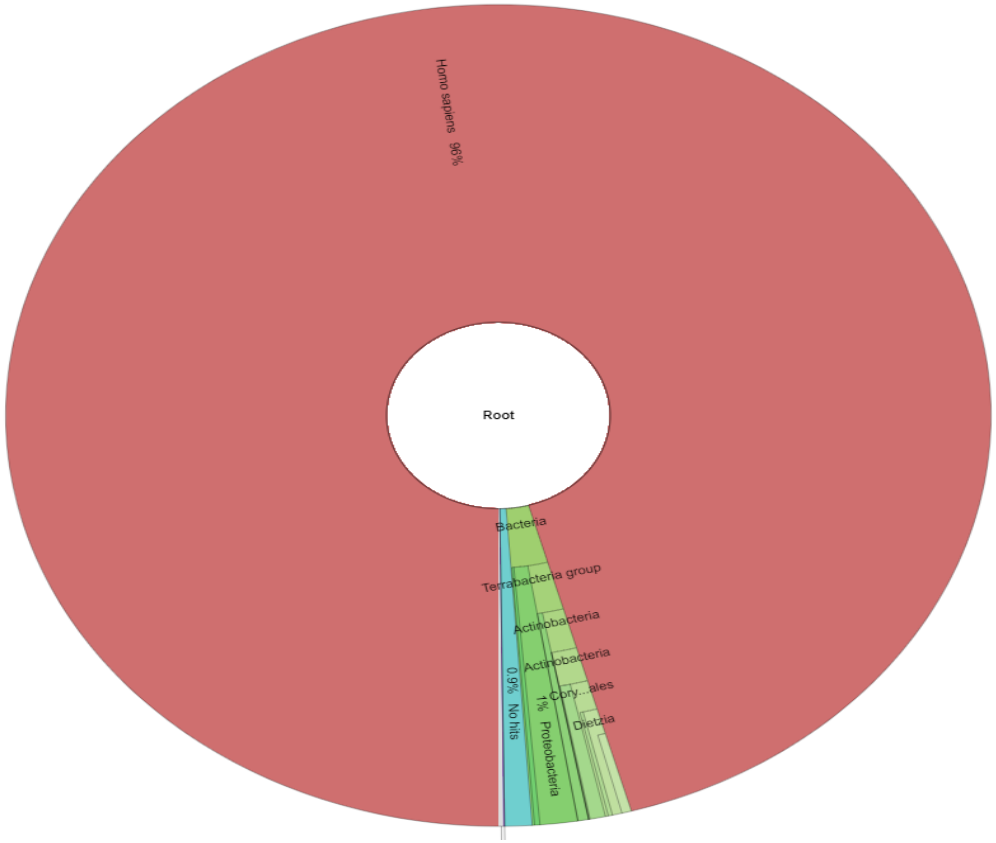
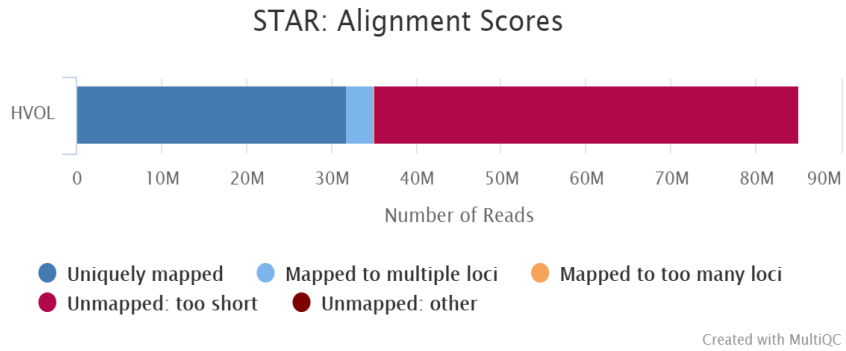
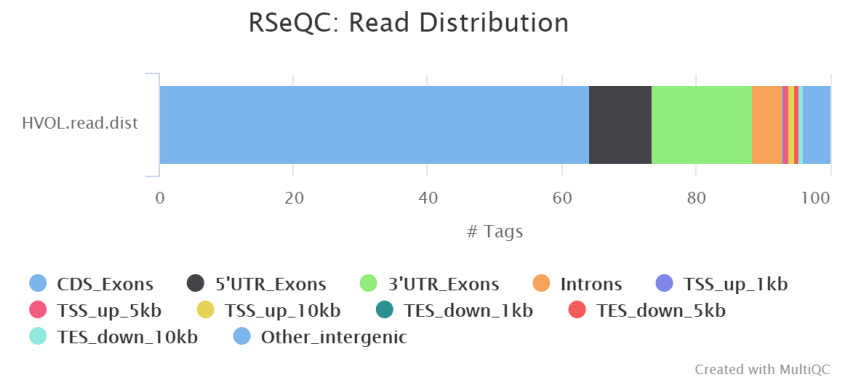


Figure D-2 Visualisation of the taxonomic classification of the unmapped reads from STAR for pilot 2 using Kraken2 and Krona software. A) HVOL sample: 63% of the reads were classified as bacteria (red),13% as Homo sapiens (green), 21% had 'no hits' (light green) and 3% were classified as 'other' (grey) B) positive control sample: 96% were classified as Homo sapiens (red), 3% bacteria (green), 0.9 'no hits' (turquoise) an 0.2% other (grey)





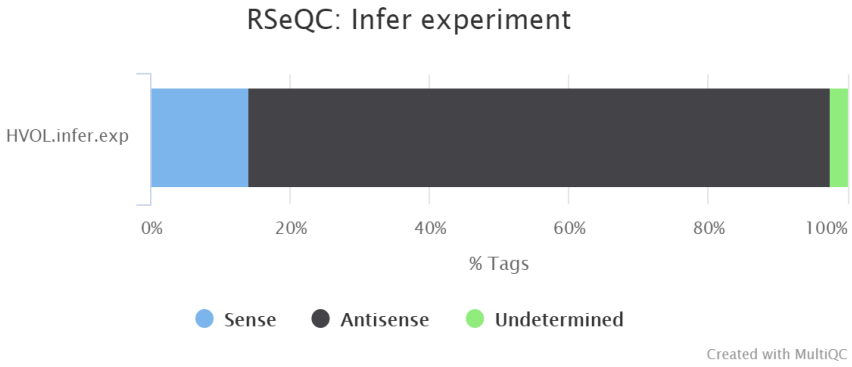


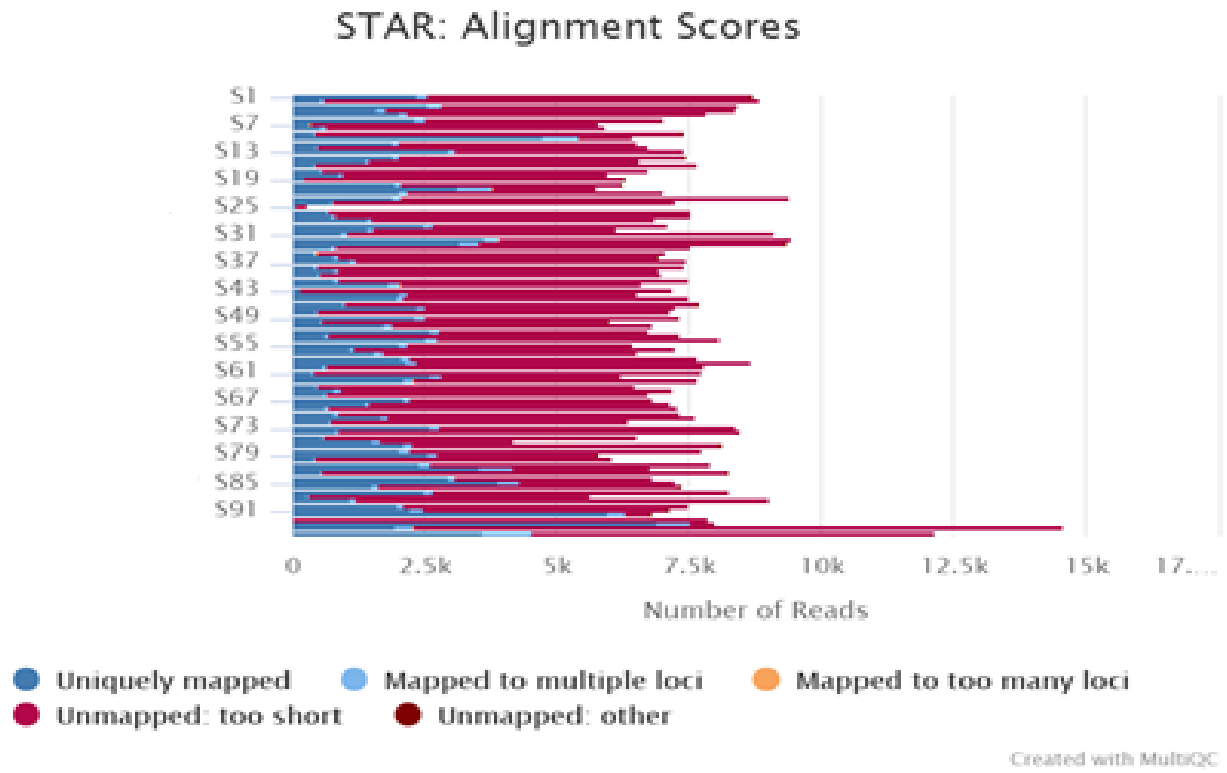
Figure D-3 Ascertaining if any lncRNAs and heart related lncRNAs can be detected in plasma (pilot 2a).

A-C) *The QC metrics show similar profiles to the same sample run at a lower depth on the MiSeq*

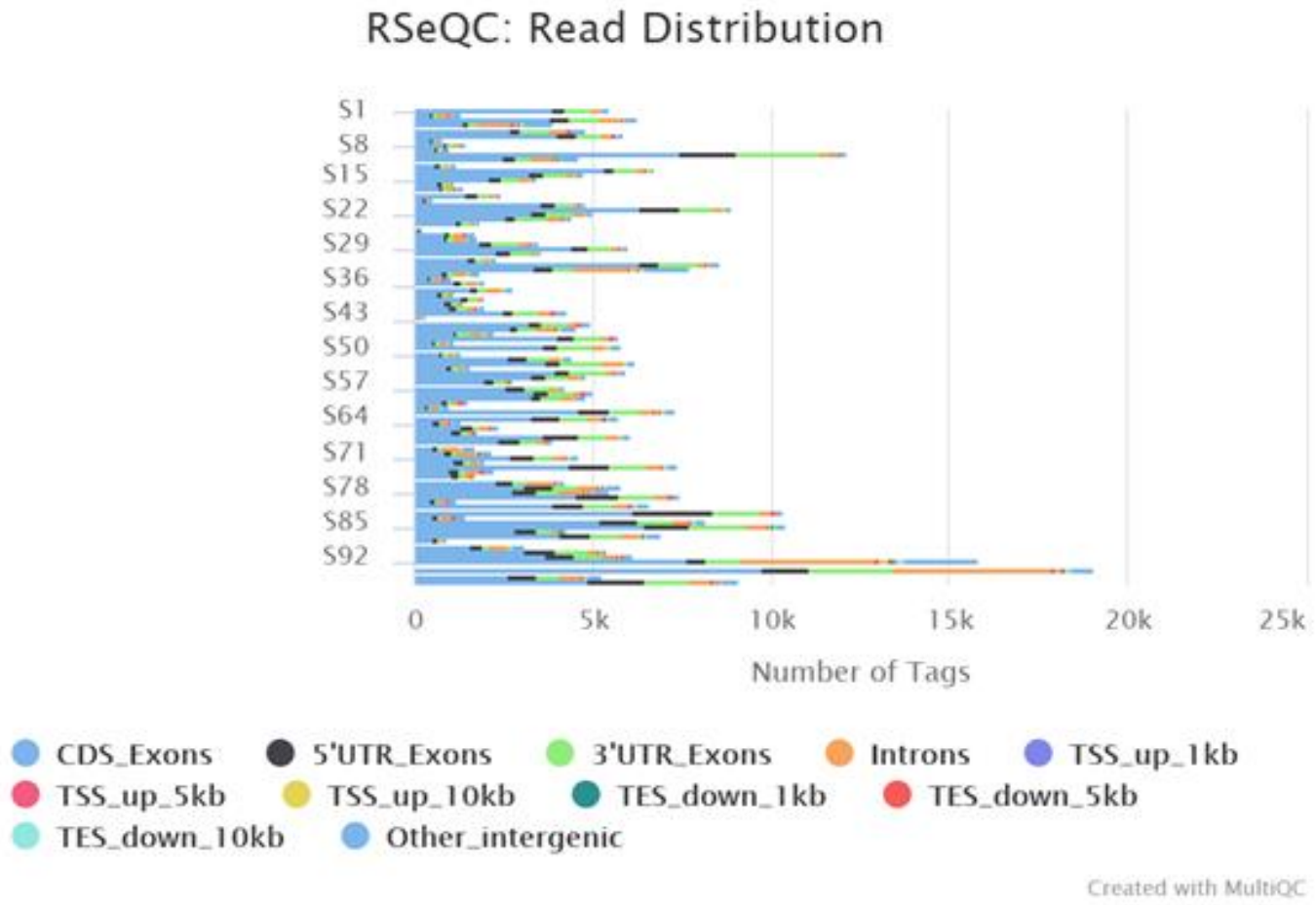
Appendix E

QC Metrics for plasma RNA Sequencing

A



B


RSeQC: Read Distribution
S1
S8
S15
S22
S29
S36
S43
S50
S57
S64
S71
S78
S85
S92
0
5k
10k
15k
20k
25k
Number of Tags
CDS_Exons
5'UTR_Exons
3'UTR_Exons
Introns
TSS_up_1kb
TSS_up_5kb
TSS_up_10kb
TES_down_1kb
TES_down_5kb
TES_down_10kb
Other_intergenic
Created with MultiQC

C

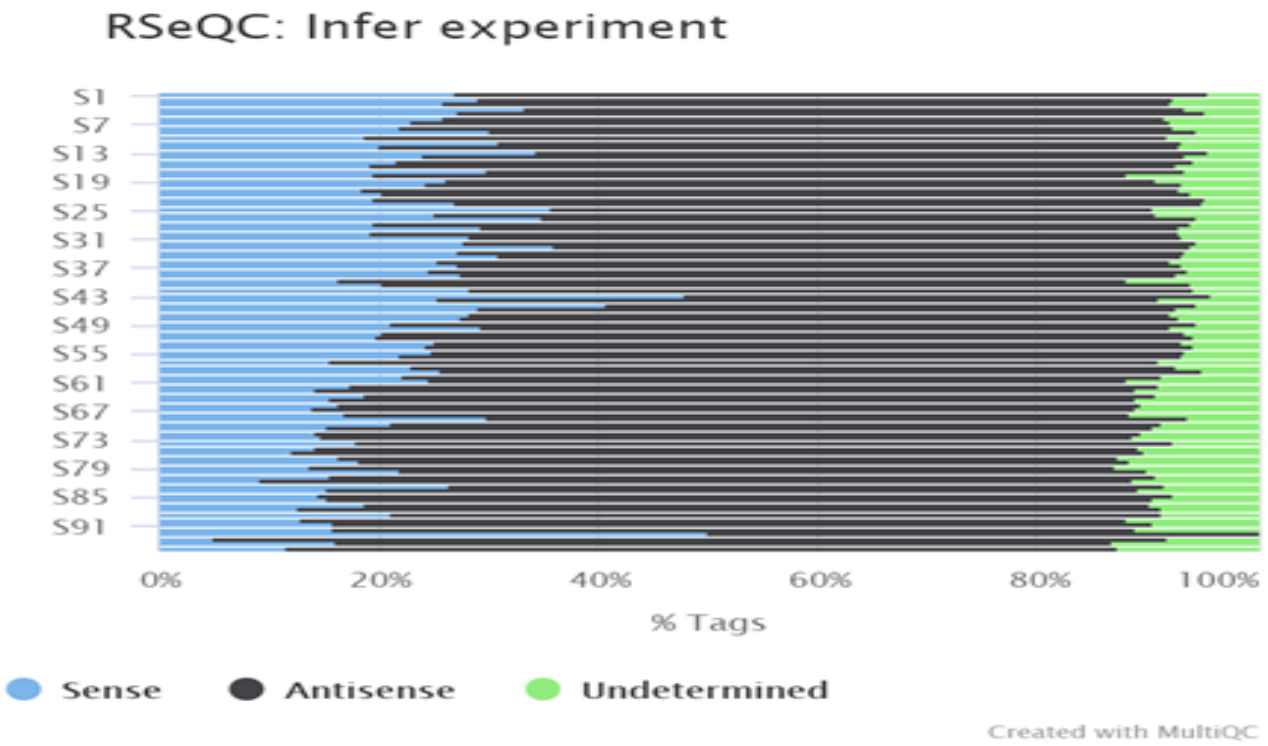
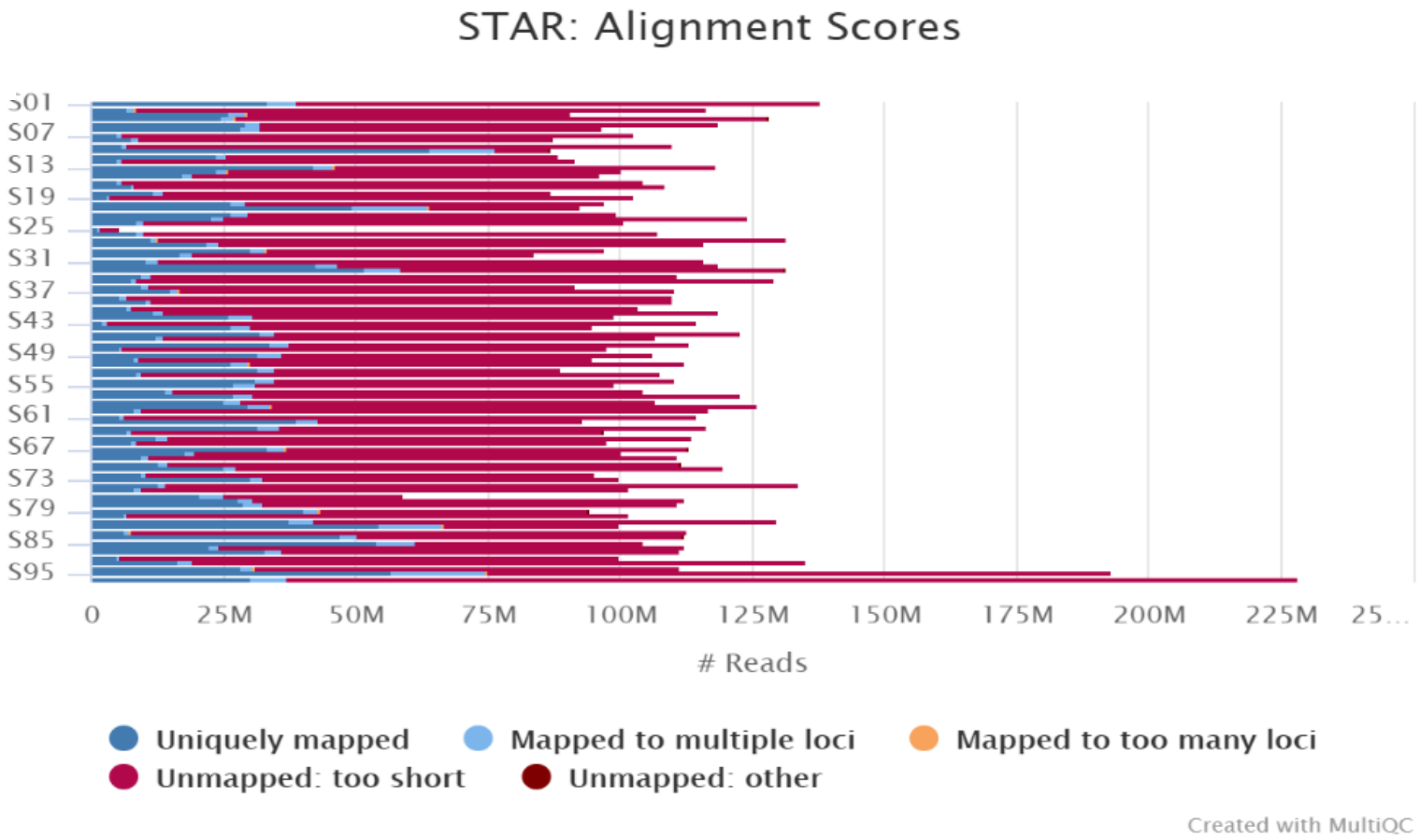


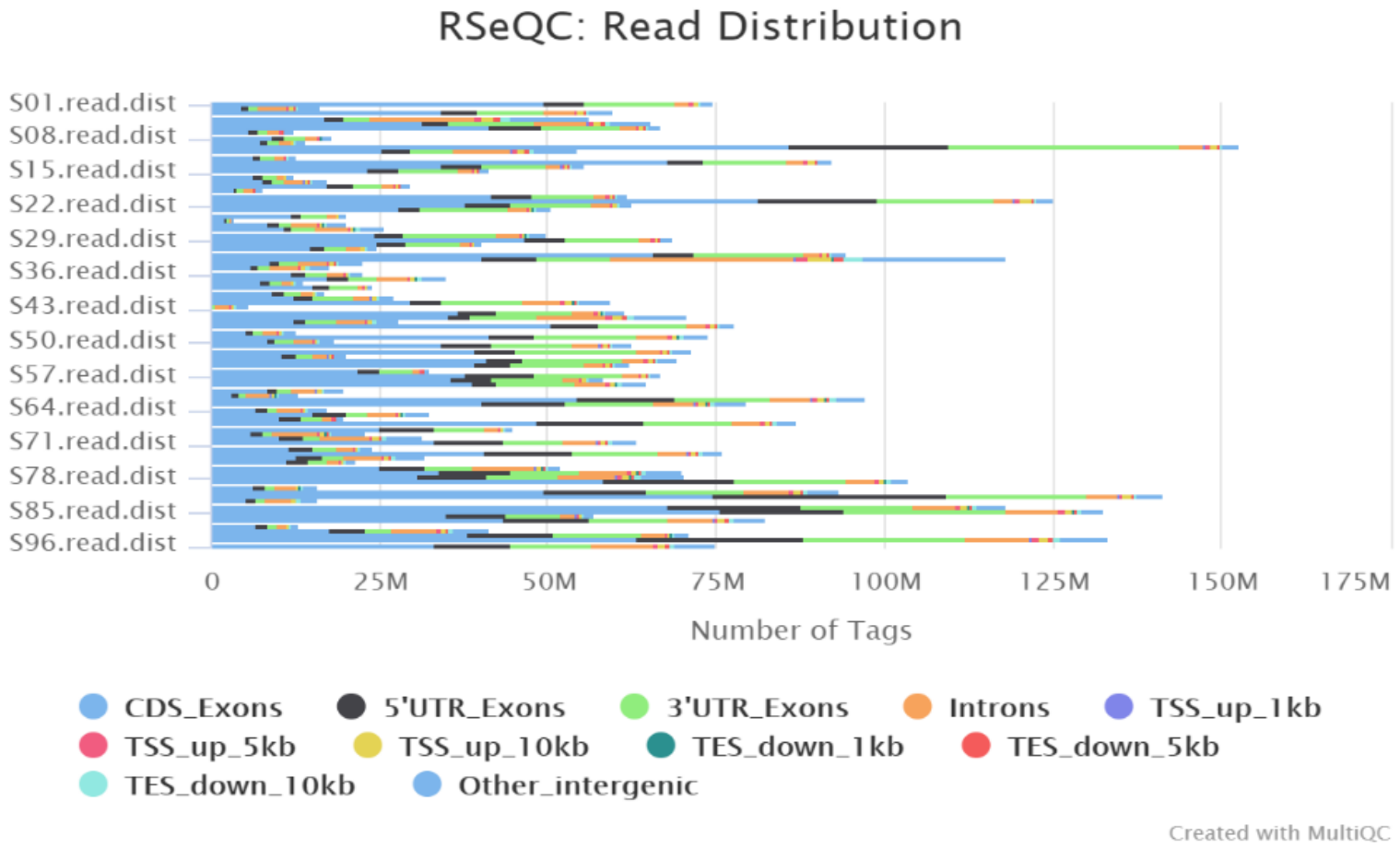
Figure E-1 Alignment scores for the HiSeq plasma experiment.

A) *The scores show similar distributions to the pilot plots with ~30% and 70% of reads uniquely mapping and unmapped: too short, respectively.* ***B)*** *Read distribution plots showing most reads aligning to coding regions* ***C)*** *Infer experiment plots showing the majority of reads aligning to the antisense strand.*

A


STAR: Alignment Scores
S01
S07
S13
S19
S25
S31
S37
S43
S49
S55
S61
S67
S73
S79
S85
S95
0
25M
50M
75M
100M
125M
150M
175M
200M
225M
25...
Reads
Uniquely mapped
Mapped to multiple loci
Mapped to too many loci
Unmapped: too short
Unmapped: other
Created with MultiQC

B


RSeQC: Read Distribution
S01.read.dist
S08.read.dist
S15.read.dist
S22.read.dist
S29.read.dist
S36.read.dist
S43.read.dist
S50.read.dist
S57.read.dist
S64.read.dist
S71.read.dist
S78.read.dist
S85.read.dist
S96.read.dist
0
25M
50M
75M
100M
125M
150M
175M
Number of Tags
CDS_Exons
5'UTR_Exons
3'UTR_Exons
Introns
TSS_up_1kb
TSS_up_5kb
TSS_up_10kb
TES_down_1kb
TES_down_5kb
TES_down_10kb
Other_intergenic
Created with MultiQC

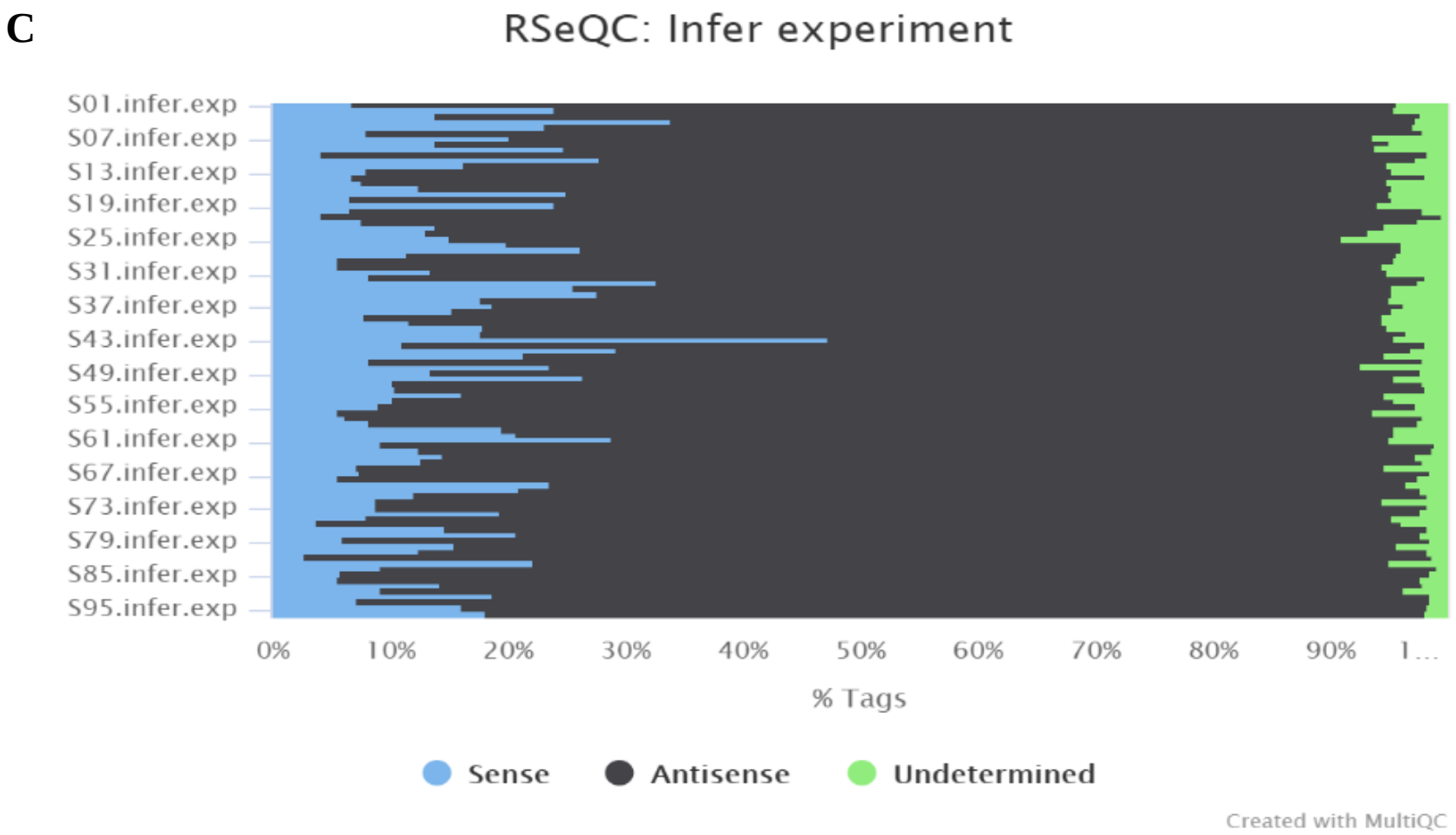


Figure E-2A) Alignment scores for the NovaSeq 6000 plasma experiment.

The scores show similar distributions to the pilot plots with ~30% and 70% of reads uniquely mapping or unmapped: too short, respectively. ***B)*** *Read distribution plot showing most reads aligning to coding regions* ***C)*** *Infer experiment plot showing the majority of reads aligning to the antisense strand (N.B. the obvious exception is sample 43 which was one of the outliers and consequently discarded from analysis).*

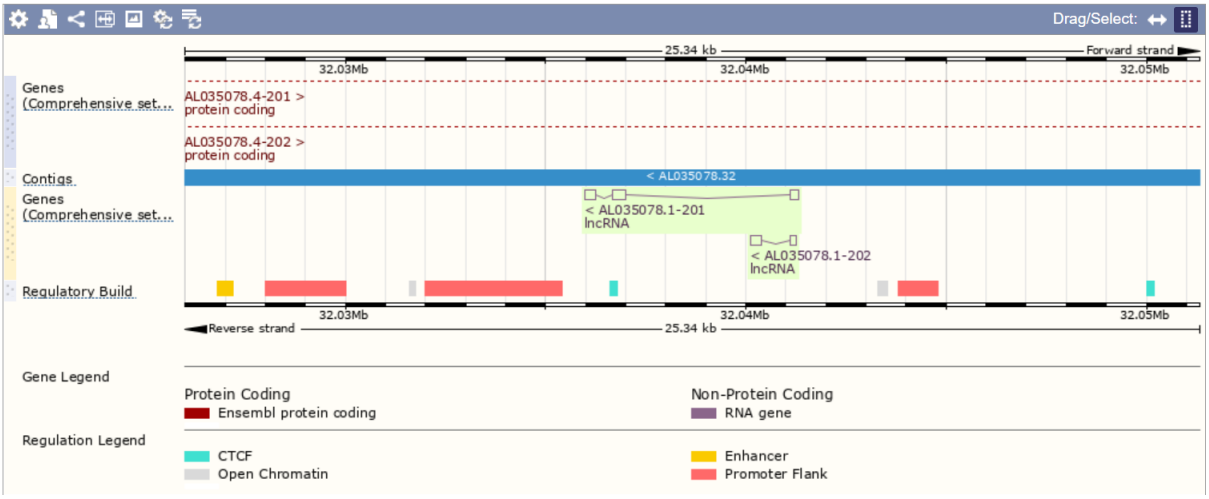
Table E-1 Genes differentially expressed between CDCS HF negative and CDCS HF positive (padj<0.2) suggesting potential genes for further investigation.

Ensemble id	Gene name	Gene type	Base Mean	log2 Fold Change	padj
ENSG00000135905.19	DOCK10	protein coding	381.3008	-1.16	0.041267
ENSG00000249307.7	LINC01088	lncRNA	370.2495	1.25	0.121078
ENSG00000068885.15	IFT80	protein coding	113.9025	0.94	0.127449
ENSG00000125779.22	PANK2	protein coding	258.7731	-0.64	0.127449

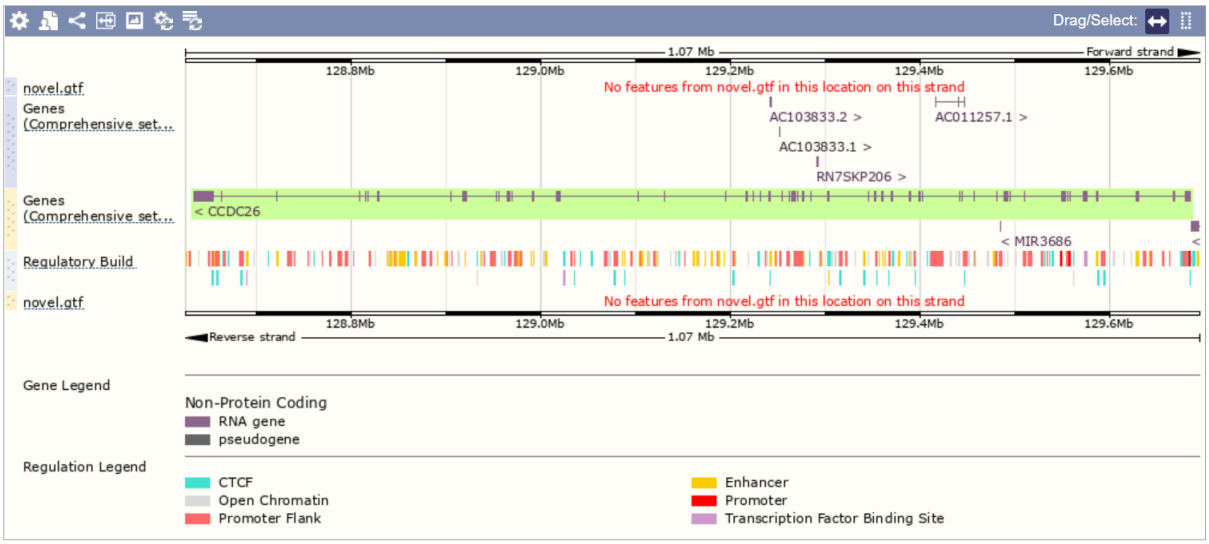
Appendix F

Plasma RNA Sequencing

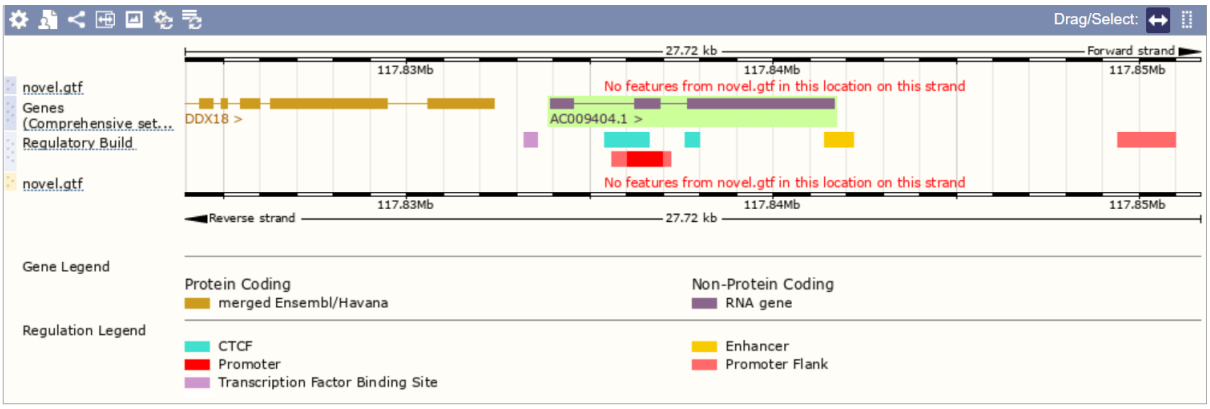
AL035078.1



CCDC26

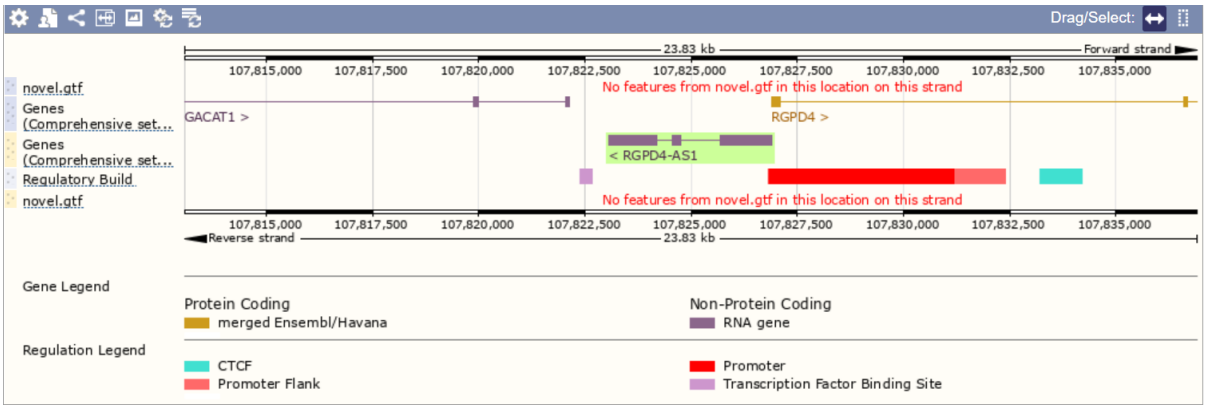


AC009404.1

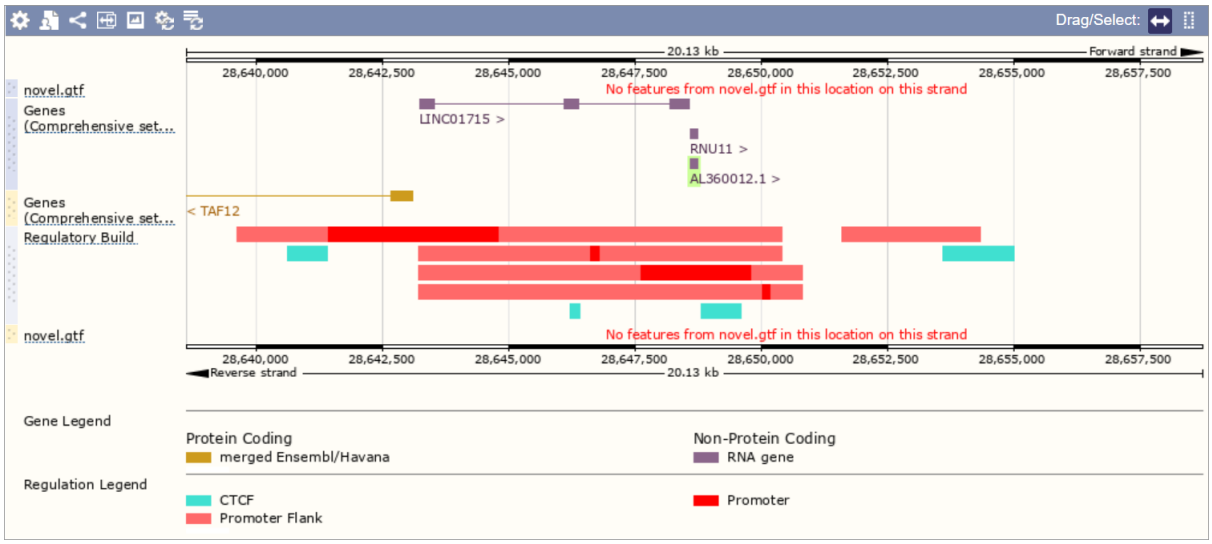


LINC02245

RGPD4-AS1



AL360012.1



MSC-AS1

AD000090.1

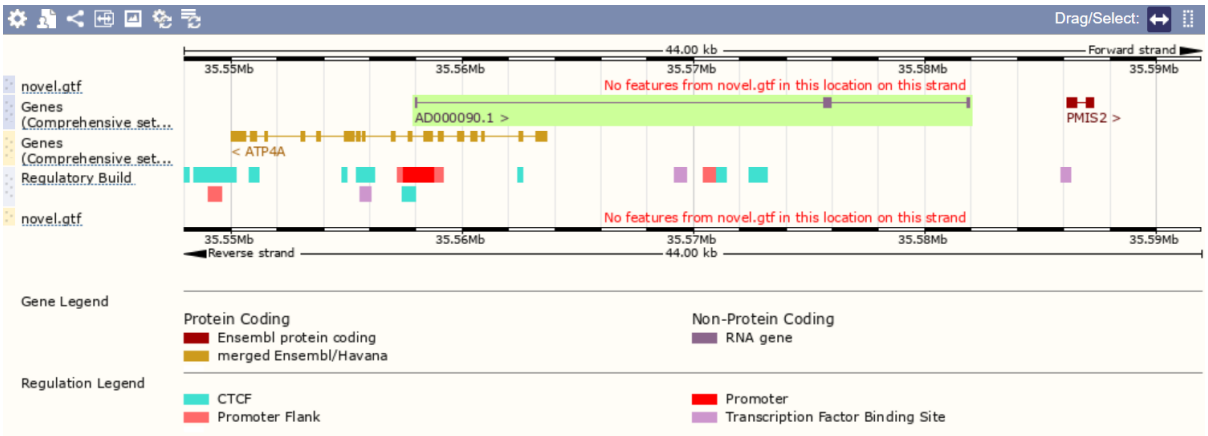


Figure F-1) Overlap of the upregulated lncRNAs with CTCF binding sites from the CDCS vs HVOL plasma RNA Sequencing