Using RNA-Sequencing to Improve Characterisation and Production of iPSC Induced Cardiomyocytes for Heart Failure

By

Harithaa Anandakumar

Master Thesis

University of Goettingen

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE (M.SC)

in

Cardiovascular Sciences

June 2020

Goettingen, Germany

Supervisor:

Tim Meyer, Ph.D.

DATA PAGE

Title of Thesis: Using RNA-Sequencing to Improve Characterisation and Production of iPSC Induced Cardiomyocytes for Heart Failure

Department: Department of Pharmacology and Toxicology

Name: Harithaa Anandakumar

Matriculation Number:

Address:

Phone:

E-Mail:

First evaluator (Supervisor):

Date of Delivery:

Second evaluator (Supervisor):

This thesis is dedicated to Snoopy!

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Life expectancy has drastically increased in the last century. For instance, an infant born in 1900 could expect to live upto 32.0 years (average life expectancy in 1900 globaly) and the same number is 72.6 years for an infant born in 2019. In 1900, the top three causes of death were infectious diseases — flu and pneumonia, tuberculosis and gastrointestinal infections. Enormous improvements in public health, sanitation, medical inventions and treatments such as vaccines and antibiotics led to a sharp reduction in infectious diseases which now account for less than 20% of deaths globally. In the same time frame, there has been a significant increase in the proportion of deaths caused by more chronic, non communicable diseases/conditions (NCD) (see 1.2). Taken together, we see an aging population strained by NCDs of which cardiovascular diseases (CVD) are the most pronounced (see 1.1). Almost half of the deaths attributed to CVDs are caused due to heart failure (HF). Despite impressive improvements in modern medicine, pharmacological interventions are capable of only alleviating the symptoms of HF, rendering it a progressive and terminal disease. Currently, the overall survival rate at one, five and ten years after a diagnosis with heart failure is estimated to be 75.9%, 45.5% and 24.5% respectively[1].
It is estimated that 1-2% of the healthcare budget is spent on HF[2], while the global economic budern is estimated at $108 billion per annum[3] and in Germany the annual prevalence-based costs for heart failure patients are around €25,532[4]. The increasing proportions of the elderly in western societies and with developing nations following suit, it is only expected that the incidence of HF would be on the rise. Yet, this debilitating and expensive disease's only viable treatment in terms of long-term life quality and mortality is a heart transplant. As per one study[5], 15% of patients died while waiting for a donor heart (at 180 days after listing), elucidating the severity of shortage of viable donor hearts. As of February 2020, there are a total of 1082 people on the heart transplant waitlist within the EuroZone as per Eurotransplant statistics[6].

Although there are myriad causes of HF, such as isichemic heart disease, aortic or mitral regurtitation (volume stress), aortic or mitral stenosis (pressure stress), congenital cardiomyopathy, constrictive pericarditis, alcohol excess, anemia, thyrotoxicosis,

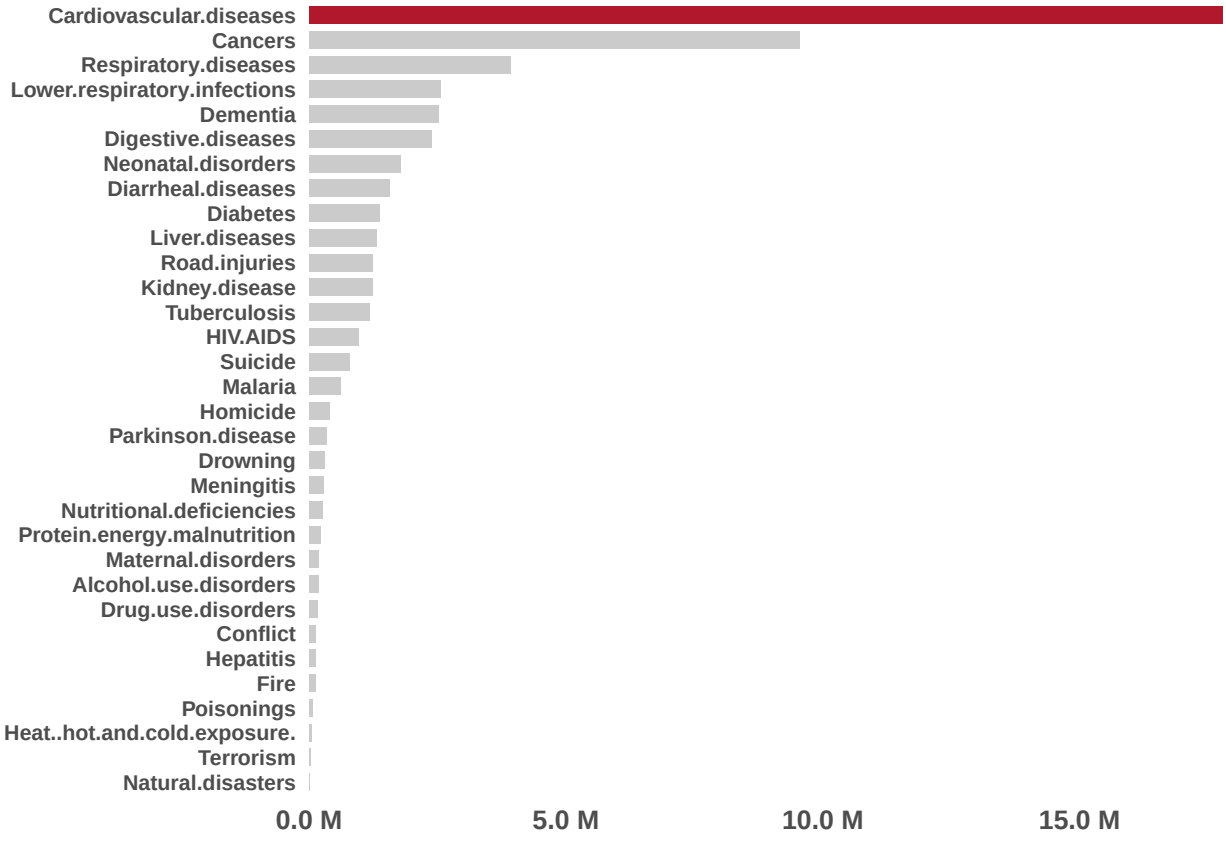


Figure 1.1: Number of Deaths by Cause in the world in 2017. The graph shows the various causes of death in 2017 in the y-axis and the number of deaths per cause in the x-axis in millions. Cardiovascular diseases were responsible for most deaths (15M). Data from: Max Roser and Esteban Ortiz-Ospina (2019) – "Causes of Death". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/causes-of-death' [Online Resource]

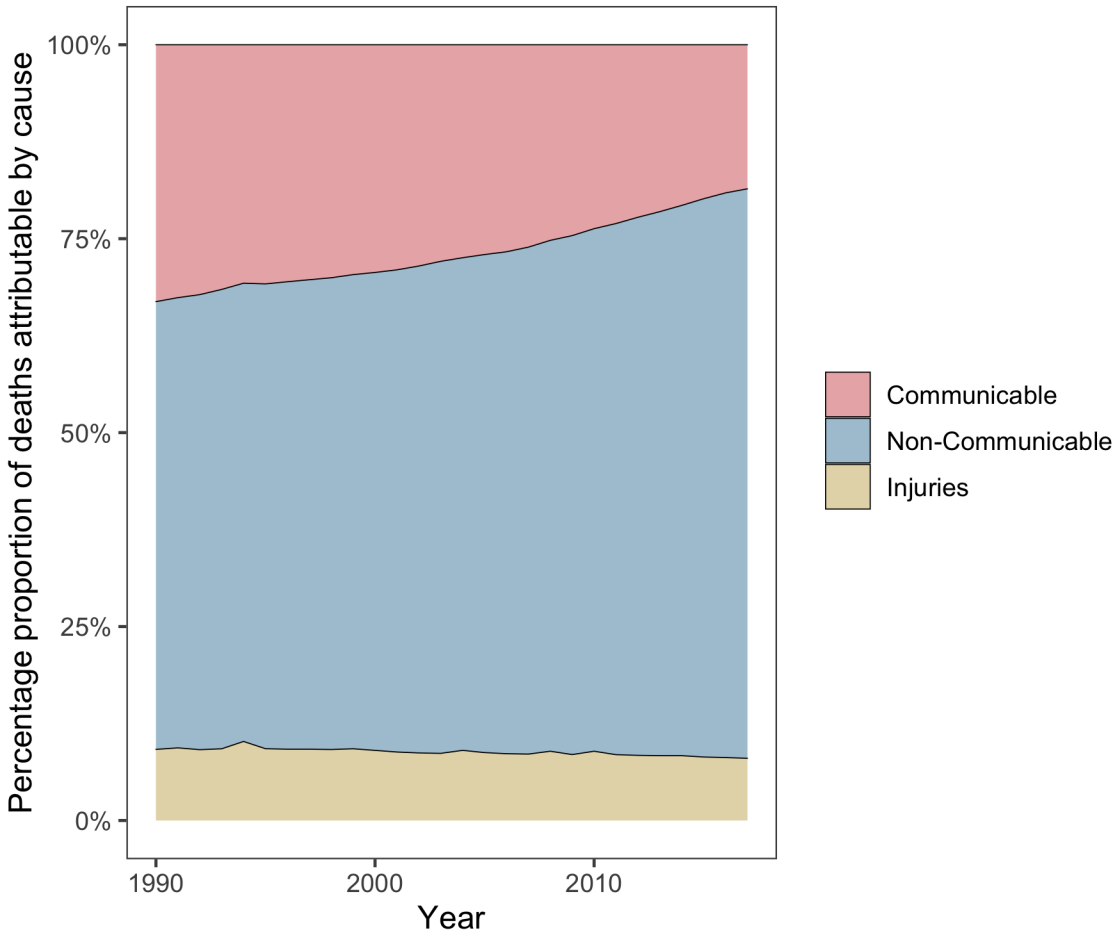


Figure 1.2: Three Major Causes of Death. Graph shows an increase in percentage proportion of deaths over the last three decades due to non-communicable diseases and a parallel reduction in deaths due to communicable diseases.

septicemia, acromegaly, they all commonly operate through the central mechanism of reduced ventricular function. Consequently, the heart is unable to adequately perfuse the tissues, resulting in a wide variety of clinical symptoms. Several compensatory measures are seen, for example, an initial phase of cardiac hypertrophy is seen to compensate for the loss of viable cardiomyocytes, resulting in a transient maintenance of the ejection fraction, sustainance of heart rate and blood pressure and thereby maintaining organ perfusion. Over time, these remodelling mechanisms become detrimental and end up worsening the left ventricular function. In effect, a negative feed-forward pathophysiological loop governed by a dissonant neurohormonal system and imapaired calcium signalling is established in late-stage HF. Most of the pharamacological treatments currently available for HF (diuretics, beta blockers, angiotensin receptor blockers, angiotensin converting enzyme inhibitors aldosterone antagonists, etc) do not halt or address the underlying pathophysiology. Device therapies are currently the alternatives to pharmacological drugs. These include cardioverter-defibrillator (ICD) which are implanted in severe cases as a means of primary or secondary prevention of sudden cardiac death. Ventricular assist device acts as a temporary bridge to a heart transplantation. Given this current scenario, it is vital to explore novel avenues for the treatment and management of HF.

1.1 Need for better therapeutics

Modern medicine has vastly improved the management of heart failure, yet it still remains a debilitating disease that would immensely benefit from newer therapies. Adult human hearts are terminally differentiated and post-mitotic. A straight-forward approach would be to counteract the progressive loss of cardiomyocytes by supplementing the heart with fresh CMs [7]. This has been made possible largely due to the introduction of human embryonic[8] and induced pluripotent stem cells [9]. iPSCs are defined by their unlimited proliferation capacity and ability to differentiate into any given cell type (derivatives of all three germ layers) upon adequate stimuli. Effective and defined protocols of directed differentiation of various iPSCs to a cardiac lineage/cell fate (apart from various other cell types) have been developed and covered in the review[10]. The straight-forward approach of direct supplementation of CMs by injection into the ventricular wall is fraught with its own key limitation: lack of long term engraftment of cardiomyocytes (varies based on the modality of delivery, covered below)[11]. Several other strategies to strenghten/remuscularize the heart such as, converting scar into muscle tissue by transdifferentiation[12], inducing endogenous

cardiomyocyte regeneration and proliferation[13], and methods to save the remaining cardiomyocytes from cell death by modulating paracrine factors[14] have been investigated (see 1.3). Despite the limitation in long term engraftment, cardiomyocyte implantation remains the most plausible option in a translational and mechanisitic stand point. It is currently known that cardiomyocytes supplemented as a cell injection have the lowest retention and epicardial delivery of cardiomyocytes as tissue engineered patches show an improved retention[15]. Animal studies indicate that transplantation of engineered heart muscle (EHM), made from human induced pluripotent stem cells (hIPSCs), to a failing heart as a means of remuscularization showed improved cardiomyocyte proliferation, vascularization, unimpaired electrical coupling and improved left ventricular function[16]. Additionally, these engineered patches have not shown to be associated with an increased propensity for arrhythmia[16–18]. More recently a macaque model of heart failure (with human-like cardiovascular physiology), showed near normal levels of contractile function after 3 months of transplantation of cardiomyocytes derived from human embryonic stem cells (hESCs)[19]. Collectively, these preclinical studies hold promise for the utilization of cardiomyocytes and EHMs thereby derived as a potential therapeutic source for failing human hearts.

1.1.1 Immunological Responses in Transplantations

Fully personalized cell therapy using autologous iPSCs for implantation circumvents problems associated with immune rejection. Yet, the cost and duration of obtaining clinical-grade iPSC cell lines along with their differentiation into required cell type for transplantation and verification of safety and efficacy have hampered autologous iPSC technology to move into clinical practice[20,21]. Allogenic transplantation[1] of thoroughly characterized iPSCs seems to be a more plausible approach to cell therapy[22]. Histocompatibility remains the main problem of using allogenic cells and tissues, including the ones that are derived as a result of iPSC differentiation. Roughly 20,000 HLA alleles are known www.ebi.ac.uk/imgt/hla/. This polymorphism is the reason why appropriate selection of donors for transplantation is crucial and difficult. A perfect donor match is unlikely, and there is always some degree of mismatch between the recipient's and donor's major histocompatility complex (MHC) genes necessitating the systemic administration of immunosuppressive drugs. To circumvent these problems, an HLA-haplotype bank of pluripotent stem cell lines was proposed to establish efficiently chosen samples with sufficient HLA diversity to provide a

[1] Two main types of stem cell transplants. *Autologous* — uses a person's own stem cells. *Allogenic* — uses stem cells from a donor whose HLA are acceptable matches to the patient's.

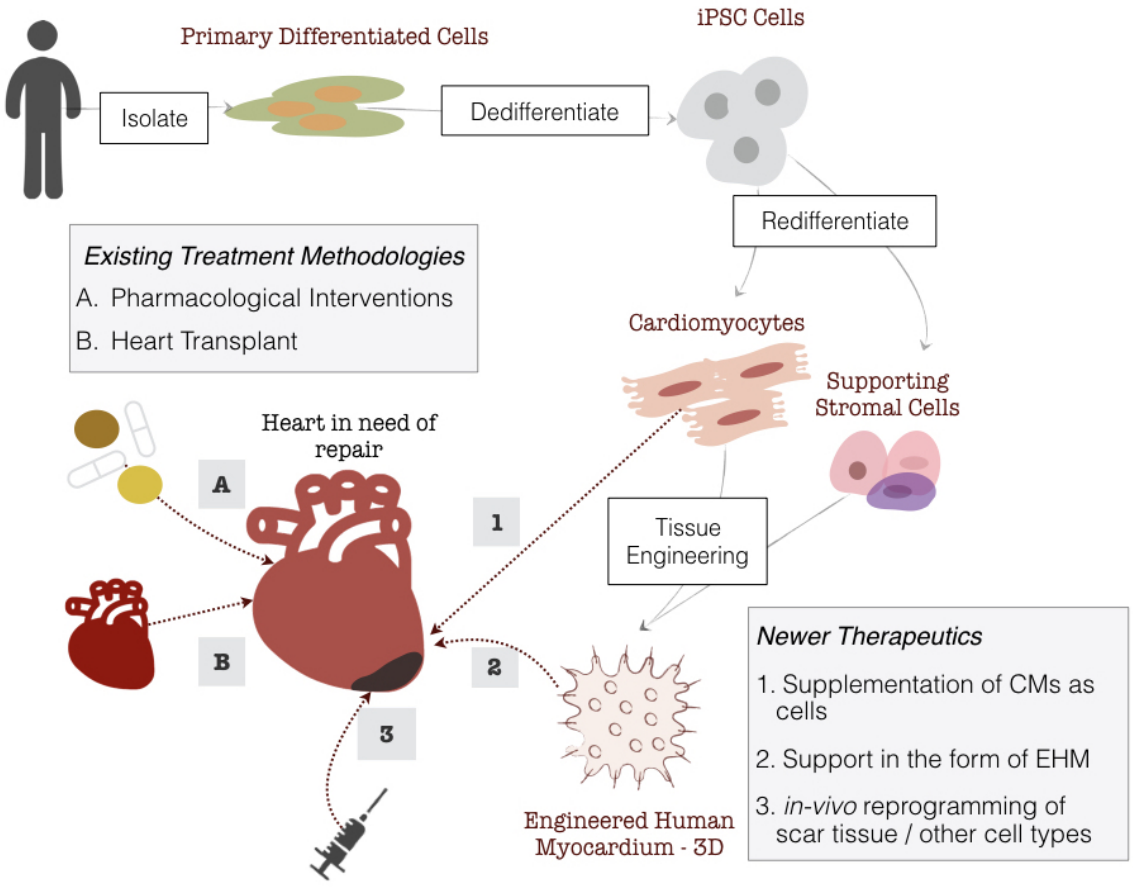


Figure 1.3: Delivery Strategies of iPSC-CMs and Treatment Options for Heart Failure. Production of EHMs start with isolating primary differentiated cells (e.g, fibroblasts) which are then dedifferentiated to iPSCs followed by redifferentiation to CMs and stromal cells, which are then combined in a collagen matrix forming the EHM.

reasonable HLA match for a large percentage of the target population[23]. For instance, a cell bank of 30 iPSC cell lines would be able to find a three-locus match in 82.2% of the Japanese population[24]. These numbers vary depending on the diversity of the population and inspite of these optimistic forecasts, such HLA haplotype banks may not completely prevent allogenic rejection as the minor mH antigens will still be inevitably different in unrelated donors and interactions of innate immunity is not accounted for[25]. An alternative strategy is the creation of an *universal stem cell line* by excision of highly polymorphic HLA class Ia and II molecules from iPSCs[26]. Once such cell-lines are authorized for clinical usage, using them to produce CMs and resultant EHMs would be feasible without drastic changes to the current protocols, allowing for faster and robust production of EHMs as a therapeutic option.

1.2 Engineered Human Myocardium

Engineered human myocardium is composed of human cardiomyocytes and supportive stromal cells both of which can by obtained by targeted differentiation from iPSCs using serum-free, GMP-compliant media and protocols. Differentiated cells are combined in an optimized ratio, embedded into a collagen matrix and casted 1.4. Several EHM patches may be stacked to make a muscle layer of optimum thickness that is sutured onto the failing myocardium to assist mechanically in pumping. For translation to clinics a production protocol that is compliant with current good manufacturing practices (cGMP) is required[27].

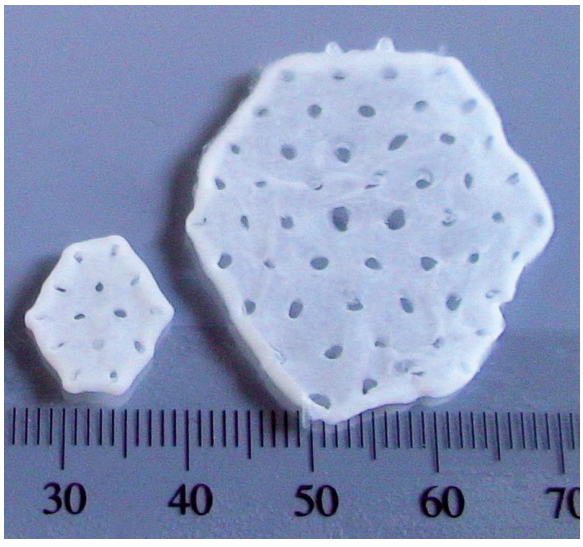


Figure 1.4: Various sizes of EHMs for clinical and experimental applications

1.3 RNA Sequencing

Information stored in genes as DNA is transcribed into RNA and ultimately translated into proteins, this is the central dogma of biology 1.5. The transcription of a subset of genes into RNA molecules gives a cell it's specificity and identity, along with regulating its activities. The term 'Transcriptome' refers to the total RNA at a given timepoint, whether from a population of cells or a single cell, and its analysis is known as transcriptomics. Microarrays, a hybridization based approach, were the main-stay of such transcriptomics until the recent advent of high-throughput next-generation sequencing (NGS) which revolutionized transcriptomics by enabling RNA analysis via the sequencing of complementary cDNA[28]. RNA sequencing (RNA-Seq) has several advantages over microarrays, namely its ability to detect transcripts that are not yet annotated, low background signal, a large dynamic range of expression level, higher sensitivity, higher reproducibility, all of which allow for understanding the dynamic and complex nature of the transcriptome.

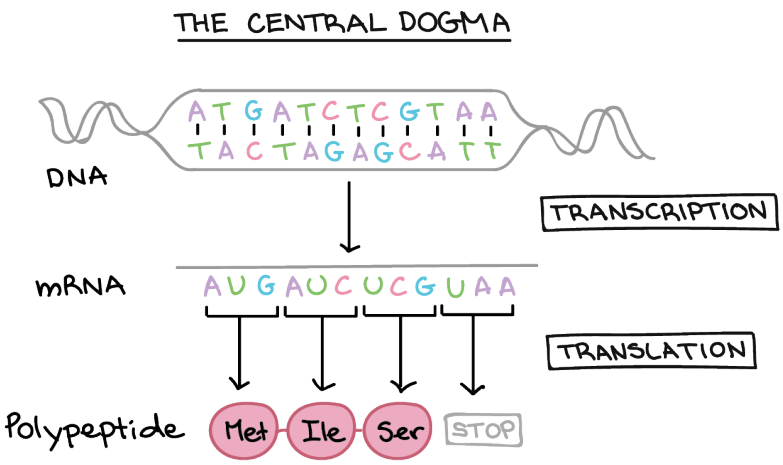


Figure 1.5: Illustration of the Central Dogma. Genetic material stored as DNA gets transcribed into RNA which then gets translated into proteins. RNA-Sequencing captures these RNA molecules.

The type of information that RNA-Seq provides can be broadly classified into two categories:

- Qualitative data which includes identifying transcripts, identifying intron/extron boundaries, poly-A sites and transcriptional start sites (TSS) which in RNA-Seq terminology is commonly referred to as "annotation".
- Quantitative data which includes measuring differences in expression, alternative TSS, alternative splicing, alternative polyadenylation between two or more treatments or groups.

This power of sequencing RNA has led to RNA-Seq not only being limited to the genomics community but also to it becoming a main-stay in the toolkit of all life science research communities. A typical RNA-Seq experiment can be split into three parts[29]:

1. Pre Analysis
    - Wet-Lab (Designing the project, RNA extraction, purification and enrichment of mRNA, cDNA synthesis, fragmentation, adaptor ligation and amplification, cDNA libraries to be sequenced)
    - Experimental Design (choosing the library type, sequencing legth, the number of replicates and sequencing depth). In the most common use-case of RNA-Seq analysis which is differential expression studies, two or more groups / conditions are defined. In this project, each differentiation run that produced CMs from iPSCs can be considered as a separate group.
    - Sequencing Design (spike-ins, randomization at library prep, randomization at sequencing run)
    - Quality Control (raw reads, read alignment, quantification, reproducibility)
2. Core Analysis
    - Transcriptomic Profiling (read alignment, transcript discovery, quantification level, quantification measure)
    - Normalization (Z-scale, variance stabilized transformation, etc)
    - Differential Expression
    - Interpretation (functional profiling)
3. Advanced Analysis
    - Visualization
    - Integration (eQTL, ATAC-seq, ChIP-Seq, proteomics/metabolomics)

The success of an RNA-Seq study depends on the choices and decisions made at each of these steps.

1.3.1 Single cell versus bulk RNA Seq

A single mammalian cell contain typically less than 1pg or 400K molecules of mRNA. The RNA to be sequenced may be collected from samples containing either multiple (bulk) or single cells. Data obtained from the more established bulk sequencing represents the *average expression level* for each gene across the large population of input. This bulk RNA-Seq which is the main work horse of gene expression studies is adequate for comparative transcriptomics, wherein samples of the same tissue are

compared across species, or for quantifying expression signatures from ensembles, such as in disease studies. However, it falls short in its ability to be an effective tool for studying heterogeneous systems, such as complex tissues (brain, heart, etc) or early developmental studies. It also fails to capture the stochastic nature of gene expression and spatial resolution can not be obtained, as ilustrated in 1.6.

Single-cell RNA-Seq (scRNA-seq) address this short coming as it measures the *distribution of expression levels* for each gene across a population of cells[30]. It has revealed new, unknown cell types in what were considered to be well-studied and established diseases, such as the discovery of ionocyte cells in cystic fibrosis[31]. Spatially resolved scRNA-Seq holds similar promises, revealing novel information on the extent of fetal marker gene expression in small populations of adult heart tissues[32]. Thus, novel biological questions addressing cell type identification, heterogeneity of cell responses, stochasticity of gene expression and inference of gene regulatory networks across cells can be studied. The applications of scRNA-Seq to novel biological questions and the computational and laboratory methods catering to it are advancing at such a rapid pace that even recent reviews[33,34] are becoming outdated.

1.4 Computational deconvolution

The usage of scRNA-Seq is still limited by its cost and impracticality with respect to analyses of large sample cohorts. Also, most clinical specimens are fixed, for example in formalin or embedded in paraffin, which renders its dissociation into intact single-cells impossible. To circumvent these limitations and utilize the specificity and accuracy of scRNA-Seq along with the ease of bulk of RNA-Seq, several groups have developed *deconvolution* computational techniques[35–41]. Deconvolution, in the realm of a sequencing, is a common umbrella term for a procedure that estimates the proportion of each cell type in a bulk sample. Flow cytometry and scRNA-Seq are experimental methods of deconvolution. Computational deconvolution leverages scRNA-Seq reference sets (or fluorescence-activated cell sorting (FACS)-sorted, purified bulk sets) for bulk gene expression deconvolution. Of various tools developed to perform deconvolutions, CIBERSORTx ([42]) became probably the most versatile because unlike other methods it can:

1. Leverage scRNA-Seq derived reference profiles for bulk tissue dissection
2. Overcome technical variation arising from different platforms (eg., bulk RNA-Seq, scRNA-Seq, microarrays) and tissue preservation techniques
3. Digitally "purify" cell-type specific expression profiles from bulk tissues with-

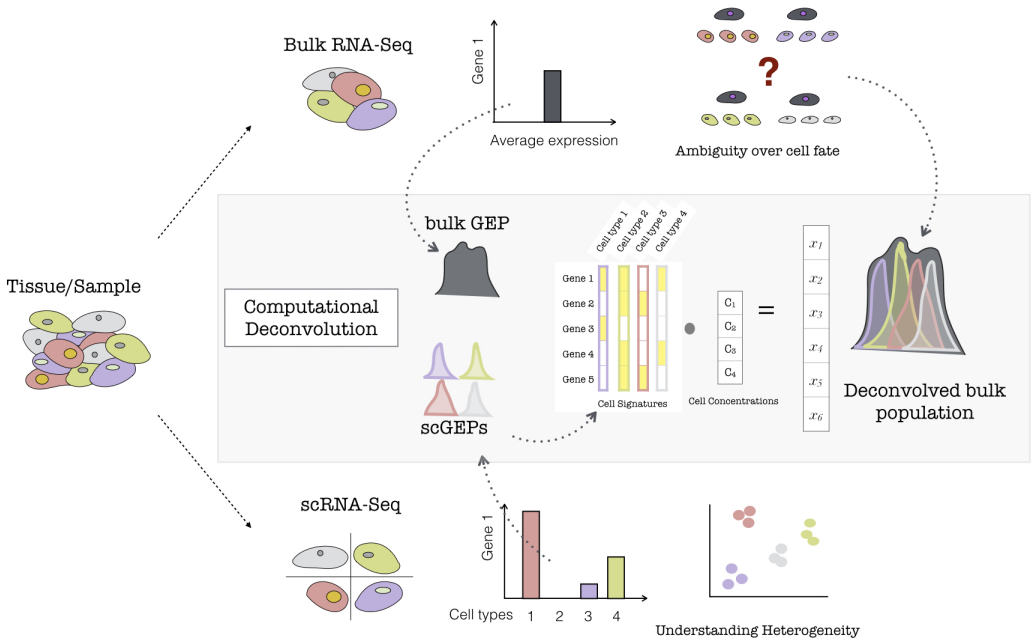


Figure 1.6: Example of Bulk and Single-Cell RNA-Seq. If a population of cells or tissue is considered, it can be sequenced either at the population/bulk level or the single-cell level (For simplicity, different coloured cells represent different sub-populations/cell-types within the sample). If the bulk analysis pathway is chosen, post-sequencing, for any arbitrary gene (here denoted as Gene 1), a single average expression level is measured, and subsequently is representative of a mixed cell population (a grey cell — combination of all the colored cells). On the otherhand, if the same population were to be sequenced using Single-Cell technologies, then each of these hypothetical cell types would record its own level of gene expression for the same arbitrary Gene 1, and this would allow to better understand the heterogenous tissue.
Illustration of Computational Deconvolution in the inner grey box. Based on the type of sequencing there is either a bulk gene expression profile for the entire population (a single grey profile) or distinct single cell profiles (shown by different coloured peaks accounting for the different cell types). The single cell profiles can be used to make a cell signature matrix wherein distinct set of genes are expressed by different cell types (a simplistic example of 5 genes and 4 cell types is shown). The idea of deconvolution is that with a signature matrix and a bulk GEP, the putative proportions of cell types within the bulk sample (denoted as cell concentrations) can be estimated, given that the scRNA-Seq is representative of the population of the Bulk RNA-Seq

out physical cell isolation. Briefly, most deconvolution algorithms, including CIBERSORTx, work to solve the following linear equations for $\mathbf{f}$:

$$m = Hf$$

m: mixture gene expression profile (GEP) (to be deconvolved)

f: a vector of fraction of each cell type in a signature matrix (the unknown)

H: a *signature matrix* containing signature genes for cell subsets of interest

Both m and B are input requirements. Further explanation of deconvolution and the implementaion of the algorithms can be found at[42,43].
With this framework, a relevant single-cell or bulk-sorted RNA sequencing data can be used to tease out molecular signatures of distinct cell types and these signatures can then be used to characterize cellular heterogeneity from bulk tissue transcriptomes without physical cell isolation, see 1.6.

1.5 Exploratory Data Analysis in RNA-Sequencing

High-throughput gene expression technologies have become a common choice for addressing systems-level and as well as molecular questions of biological phenomena. Yet, these approaches do not always meet the high expectations of the *sequencing revolution*, possibly due to the fact that the interpretation of the data is often lagging behind its generation. As discussed by Hudson et al.,[44] in their opinion article, the rampant usage of small/curated lists of differentially expressed (DE) genes are limiting and can possibly lead to misinterpretation or out-of-context conclusions. Unbiased exploratory data analysis techniques are required holistically interpret the data. Common techniques include unsupervised clustering (hierarchical, k-means, etc) and dimension reduction (discussed below), which are used to detect unbiased/unpredicted patterns, confounding variables. Exploratory data analysis not only help to find new ways of answering questions but it ultimately permits to detect unexpected patterns and formulate novel working hypothesis.

1.5.1 Principal Component Analysis (PCA)

High-dimensional data are common in today's biology as they arise when several features, like the expression of many genes, are measured for multiple samples. This kind of data holds several challenges such as high computational demand and an

increased error rate due to multiple test corrections when testing each feature for association with an outcome. PCA is an unsupervised dimension reduction technique, that on any given dataset performs linear transformation and fits the data to a new coordinate system in such a way that maximum variance is explained by the first coordinate, and each subsequent coordinate is orthogonal to the last and explains progressively lesser variance. Each principal component thus sums up a certain percentage of the total variation in the dataset. In this way, a set of x correlated variables over y samples is transformed to a set of p uncorrelated principal components over the same samples. Where many variables correlate with one another, they contribute strongly to the same principal component. PCA can find patters without prior knowledge about whether samples come from different treatment groups or have phenotypic differences. The first few principal components lend themselves to low-dimensional representation (eg, bi-plot) of the data, while retaining as much information as possible as they represent a large portion of the relevsnt information in the dataset while uncorrelated noise is pushed to the last components. The goal is to reduce the features' dimensionality with minimal loss of information, for a simplistic example see 1.7.

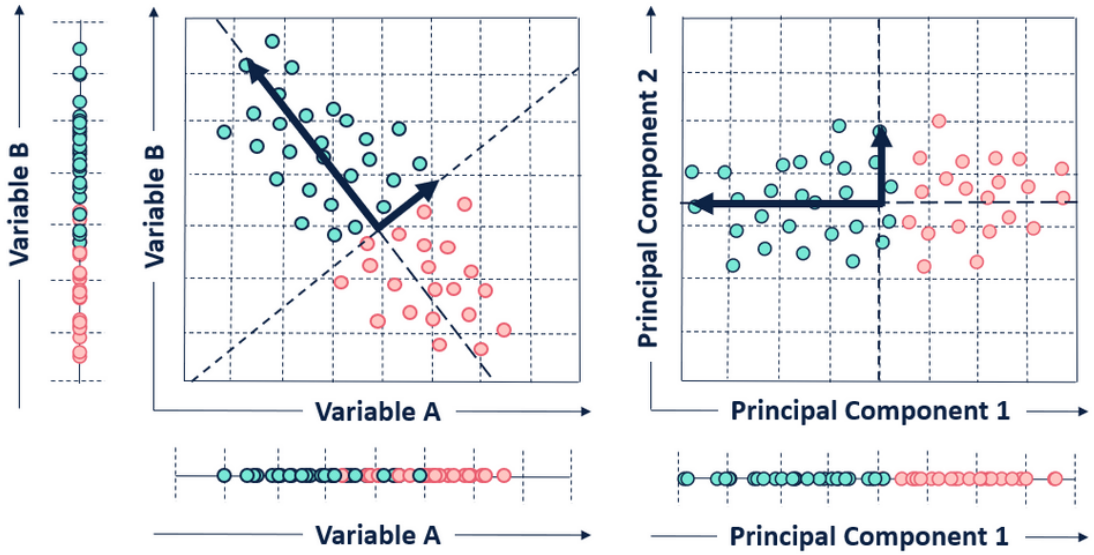


Figure 1.7: Illustration of PCA. Given two variables A and B, the plot on the left shows a scatter plot in its original place while the one on the right shows a PCA bi-plot of the variables. In this simplistic example, a 2D object (with 2 variables) which was not efficiently separated in 1D representation its original space is separated clearly in 1D across it's first PC. Here, 2D is efficiently reduced to 1D with minimal loss of information, this same technique can be applied to several dimensions to efficiently reduce it to smaller dimensions and be easily visualized.

1.6 Rationale for the current work

Targeted Differentiation of hypoimmunogenic iPSCs into functional cell types and subsequent assembly into artificial tissues for organ repair and replacement holds great promise to overcome the current donor organ shortage. Translation into clinics requires rigorous control and constant refinement of all process involved. RNA sequencing offers an in-depth view into the state of a cell (scRNA-seq) or a cell population (bulk RNA-Seq), and is ideally suited to describe the evolution of iPSCs along transient, morphologically not fully characterized states towards a terminally differentiated cell. Knowledge about differentiation and differentiation protocols are other areas relevant to this project, that have been vastly improved by the usage sequencing technologies[10,45–49]. Wu et al[50,51] for instance, evaluated current protocols to generate kidney organoids from hiPSCs (as source for tissue replacement) using scRNA-Seq. The study showed that the organoid-derived cell types were immature, and contained a significant percentage of non-renal cells. This proof-of-concept study showed the power of scRNA-Seq technologies to characterize and improve organoid differentiation. Prof.Zimmermann's research group has developed GMP compliant protocols for the differentiation of hiPSC to cardiomyocytes and stromal cells which are then used to make EHMs intended for tissue replacement therapy Currently scRNA-Seq is not available but multiple bulk RNA-Seq data across several differentiation runs has been performed by the group. The availability of bulk RNA-Seq data and public reference scRNA-Seq data sets like that of[52] along with accessible deconvolution techniques, like CIBERSORTx, allows the project to characterize hiPSC induced cardiomyocytes and EHM at a sub-population level based on transcriptomic data.

CHAPTER 2

AIMS AND OBJECTIVES

Stem cell and tissue engineering technologies allow for the potential treatment of heart failure with engineered tissue construtcs — the EHMs. We thus explored RNA-Seq data collected at two stages, as CMs and as EHMs, with the following aims and objectives:

1. To establish a reusable workflow to analyse the sequenced data — from raw FASTQ files to count files.
2. To check for the potential of RNA-Seq to identify microbial contamination.
3. To explore the data in context of maturity by comparing with adult and fetal samples from publically available datasets.
4. To identify the potential sub-populations via digital deconvolution techniques using a relevant scRNA-Seq dataset.

CHAPTER 3

METHODS

3.1 General Analysis Pipeline of Bulk RNA-Seq Data

The analysis pipeline used to process the bulk RNA-Seq data of both in-houses and downloaded datasets, is shown in 3.1. Briefly, the analysis of RNA-Seq started with assessing the quality of raw sequencing data as fastq files using `FASTQC` *(v0.11.4)*. Once the quality was deemed fit for further processing, the fastq files were mapped to `GRCh38/hg38` using `HISAT2` *(v2.1.0)*, resulting in BAM files. The coordinate sorted BAM files were then indexed using `SAMTOOLS` *(v1.9)*. The number of reads assigned to each feature of the genome was estimated using `FeatureCounts` of SUBREAD module *(v1.6.3)* with `Homo_sapiens.GRCh38.96.chr.gtf` as the reference genome `.gtf` file. The alignment, indexing and abundance estimation were performed on the *GWDG-high performance computing (HPC) cluster*. Count text files were imported into `R` *(v3.6.1)* running under macOS Mojave 10.14.5 for further processing. The data was normalized to either Z-scale or variable stabilized normalization in R using the DESeq2 package's *(v1.25.10)* `vst()` function. PCA plots were made using R's base function `prcomp()`. Most of the visualization was performed using the `ggplot2` package *(v3.2.1)*. Several other packages and few custom functions were used throughout this project. The bash and R scripts are attached in the supplementary, along with the output from `sessionInfo()` from R.

The bulk RNA-Seq data used in this project is collated from different sources, which are tabulated in table 3.1 along with their accession numbers and the number of samples[53–57].

3.2 Single Cell Reference Data and CIBERSORTX

Efficient deconvolution of bulk data requires a relevant single cell reference to estimate proportions of different cell types. For the current work we used reference data obtained by Friedman et al[52] who looked at cardiac differentiation of human pluripoten stem cells and performed single-cell transcriptomic analyses to map fate changes and analyze gene expression patterns during the differentation processes *in vitro*. They sequenced at 5 distinct time points on days 0 (hiPSC), 2 (germ layer specification), 5 (progenitor

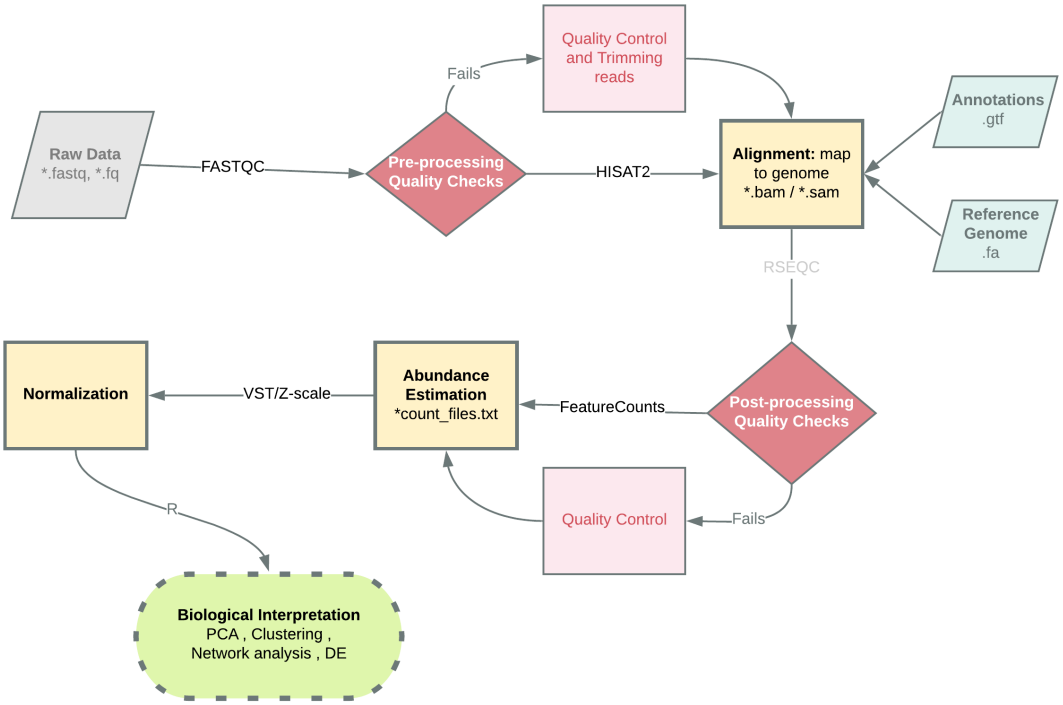


Figure 3.1: Basic analysis pipeline for Bulk RNA-Seq data used in this project

Table 3.1: Bulk RNA-Seq data and their sources

paper	Project_AccessionNumber	group	n
In-House	In-House	CM	20
		EHM	10
		Fetal_Heart	3
		Fib	4
		ipsc	2
		Rh	2
Kuppusamy KT 2015	PRJNA266045	Adult_Heart	2
		Fetal_Heart	2
Mills RJ 2017	PRJNA362579	Adult_Heart	1
		EHM	7
Pavlovic BJ 2018	PRJNA433831	Adult_Heart	12
Pervolaraki 2018	E_MTAB_7031	Fetal_Heart	9
Yan L 2016	PRJNA268504	Fetal_Heart	2

Note:
CM: cardiomyocytes, EHM: engineered heart muscle, Fib: iPSC-induced fibroblasts, Rh: rhesus iPSC-induced cardiomyocytes

cell), 15 (committed cardiac derivative) and 30 (definitive cardiac derivative) of their differentiation protocol. Relevant to this project are the last two timepoints — day 15 and day 30. Single-cell count data was downloaded from the ArrayExpress database maintained by EMBL-EBI, using the accession number E-MTAB-6268.
CIBERSORTX ([42]), reads a single cell reference input with each single-cell (every column) labelled according to the cell's phenotype or cluster identifier and bulk data with samples as columns and rownames as genes in both cases.

3.2.1 Processing of Single Cell Data

To create the reference file, clustering and *de novo* identification of cell types from scRNA data was performed as per Friedman et al's paper. Briefly, the outlier genes and cells (outside 3x median absolute deviation) of the number of cells with detected genes, mitochondrial reads, ribosomal genes were filtered out. Post filtering, `scran` *(1.12.1)* package was used for cell-to-cell normalization without quickClustering option. PCA and clustering was performed using `ascend` package *(v0.9.93)*, following the same parameters as the paper.

The differentially expressed genes between the clusters were then calculated by the `runDiffExpression()` from `ascend` package. Friedman et al identified two clusters at each of the last two time points. At Day 15, they define two sub-populations — non-contractile *(d15:S1)* and committed CM (cCM) *(d15:S2)* and likewise at Day 30 — non-contractile *(d30:S1)* and definitive CM (dCM) *(d30:S2)*. To verfiy the steps followed so far and validate the faithful reproduction of the paper, gene ontology analysis of differentially expressed genes within the sub clusters was performed. Figure 3.2 confirms that the clusters are consistent with the ones described in the paper.

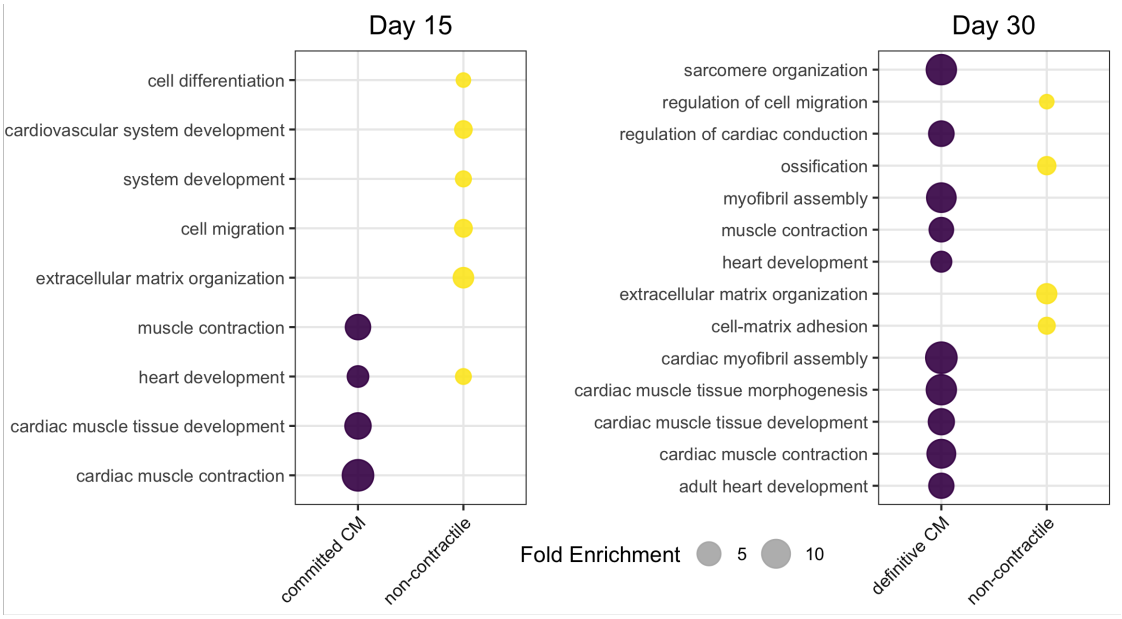


Figure 3.2: scRNA-Seq Reference Dataset. Post-processing and before feeding it into CIBERSORTx, the reference data set was analyzed to ensure its faithful reproduction of the sub-groups as defined by the paper. Here, at both time points there is a sub-group which is enriched for non-contractile features and another for cardiomyocyte features. The size of the circle corresponds to the fold enrichment seen. Reproduction of Figure 2 (J and M) from Friedman 2018.

CIBERSORTX is an online tool with user-friendly GUI with detailed tutorials on their webpage. Firstly, a `signature matrix` was created using this single cell reference file using the `Create Signature Matrix` function using `scRNA-Seq` as the input data type and all other settings were left at default. In the second step of deconvolution analysis, `mode` is set to `Impute Cell Fractions` and under `Custom` mode, the previously run signature matrix file is chosen from the drop down menu and a mixture file, previously uploaded bulk RNA-Seq data, is chosen. The option `enable batch correction` was used with `B-Mode`, which is advised for removing technical differences between the

platforms used for the signature and bulk matrices. Finally, for the `Permutations for significance analysis` option, the most stringent, `1000` option was chosen.

3.3 Analysis of Rhesus RNA-Seq

The bulk in-house samples from Rhesus were mapped using `HISAT2` with default parameters. There was no indexed reference genome readily available, so the entire genome was downloaded in from `USCS Genome Browser` — `rheMac10` assembly and converted from 2bit format to Fasta format using `twoBitToFa` available at USCS. Post alignment, abundance estimation was performed using `FeatureCounts` tool which requires a relevant .gtf file. The file was prepared using the following commands:

#Download

`wget -c -O mm9.refGene.txt.gz` filePathLinked

#Unzip the file and download the genePredToGtf tool from ucsc

```
cut -f 2- rheMac10.refGene.txt > refGene.input
```

```
./genePredToGtf file refGene.input rheMac10refGene.gtf
```

```
cat rheMac10refGene.gtf | sort -k1,1 -k4,4n > rheMac10refGene.gtf.sorted
```

This `rheMac10refGene.gtf.sorted` file was used as the input .gtf file for `FeatureCounts`. This gives the raw counts file of the *Rhesus macaque* sample mapped to it's own genome. To make comparisons with the human RNA-Seq samples relevant, orthologous genes between the two species was determined and only those with 1:1 orthology were used for further analysis. Orthologous genes were obtained from ensembl-biomart. The gene lengths of each gene was used for both species as a means of normalization within DESEQ2 by adding a matrix of gene lengths within the `assays(dds)[["avgTxLength"]] <- geneLengthMatrix` slot.

3.4 Estimating Bacterial and Viral Contaminants

DecontaMiner[58] was used to estimate the possible bacterial and viral contaminants in a representative subset of bulk samples of this project. Briefly, the *unmapped reads* i.e., those that failed to map to the reference genome were collected in a separate directory and mapped to bacterial and viral reads using the genome databases (NCBI nt) using MegaBLAST algorithm, specifying the number of allowed mismatches/gaps and the alignment length. The BLAST databases have been curated by downloading

the sequences of the complete genomes from the RefSeq repository. These .fasta files were assembled into blast databases by running the `makeblastdb` command. Files containing discarded reads along the pipeline are also generated — the low quality ones, ones mapped to mtRNA/rRNA and ambiguous and unaligned reads. The second part of the pipeline, involves setting a match count thershold (MCT) — minimum number of reads successfully mapped to a single organism to consider it a contaminant. This parameter was set at 100 (default is 5). The pipeline once run results in a table containing all the matches satisfying the alignment criteria.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 General Workflow and Mapping Statistics

Given that the analysis of RNA-Seq data is multi-faceted with distinct steps, a common work-flow modality was established as shown in 4.1.

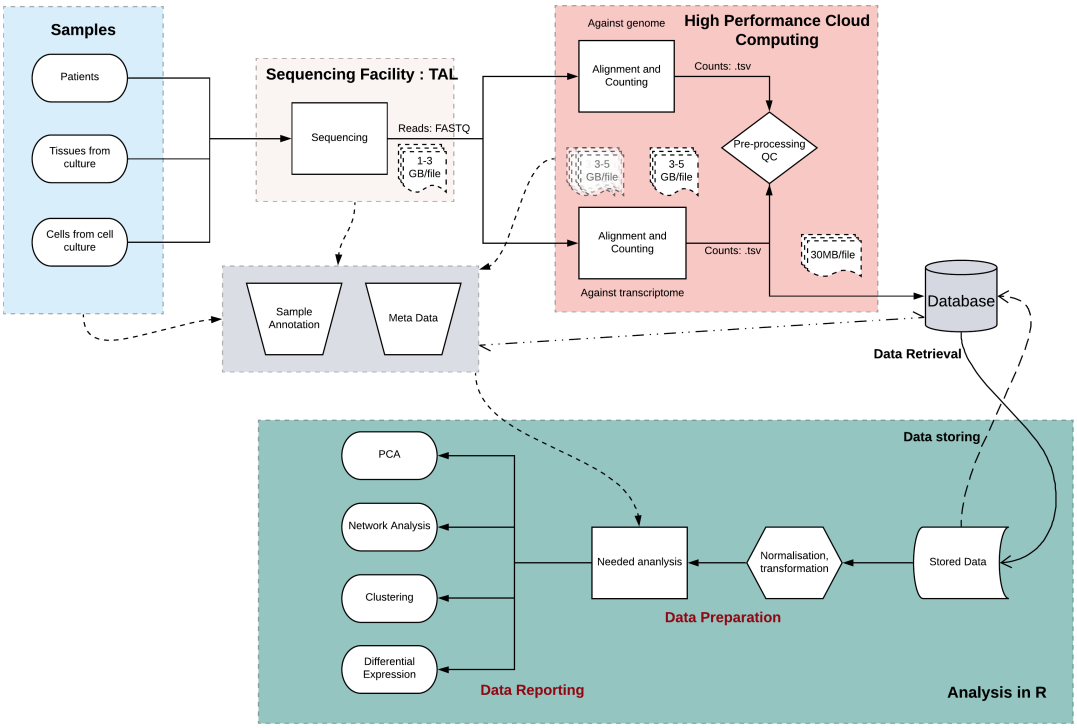


Figure 4.1: A General RNA-Seq Workflow.

The average alignment for *uniquely mapped reads* is ~70% across all samples used in this study, while the number of reads assigned to specific genomic coordinates by *featureCounts* abundance estimation tool is on average ~60%. These fall within normal acceptable ranges. There were no excessive adaptor content or duplicated reads seen.

4.2 Exploring Potential Microbial Contamination using RNA-Seq Data

To explore the potential microbial contaminants amongst the sample data sets, a representative subset of samples were chosen, see 4.1, such that every category i.e.,

adult heart, fetal heart, CM, EHM is represented by one sample in the subset, chosen across different sequencing runs, different years and different projects. Human and non-human reads were separated and the latter were used as possible microbial read candidates. After a series of filtering, as explained in 3.4, the non-host, high-quality and unique reads were aligned against the reference genomes of bacteria and virus.

Table 4.1: Samples chosen for in-depth analysis

Sample	Sample Number	Project Accession
SRR1663123_GSM1554465	1	PRJNA268504
SRR6706796_GSM2991857	2	PRJNA433831
Sample_r733sCDICM3	3	In-House
p556sCM10-3-4	4	In-House
p637sDiff6CM	5	In-House
p722s3C190604	6	In-House
p786sC190924A	7	In-House

Table 4.2: Sample Read Statistics

Sample Number	Total Reads	Primary	Multi-mapped	rRNA	Viral	Bacterial
1	35M	13M	16M	6M	20K	1K
2	20M	17M	1M	2M	24K	2K
3	51M	27M	20M	4M	190	949
4	53M	42M	10M	1M	70	5K
5	52M	31M	17M	3M	320	3K
6	51M	49M	60K	2M	966	14K
7	79M	36M	35M	8M	9K	3K

The general mapping statistics of the chosen samples can be seen in figure 4.2 A and tabulated in table 4.2. Samples vary in terms of their sequencing depth (akin to the total reads column) and the proportion mapped to the human genome, which varies between 40% to 91% across samples.
A large variance is found in the number of viral/bacterial reads mapped per million human mapped reads, see B of 4.2.

Absolute number of reads confidently assigned to different viral species across different samples is shown in figure 4.3 A. The same data is shown as relative percentages in 4.3 B. Sample 1 and 2 have a disproportionate number of reads, about 20,000 reads, mapped to a single viral genome — the col phage or phi-X174 (PhiX - NC_00 1422.1).

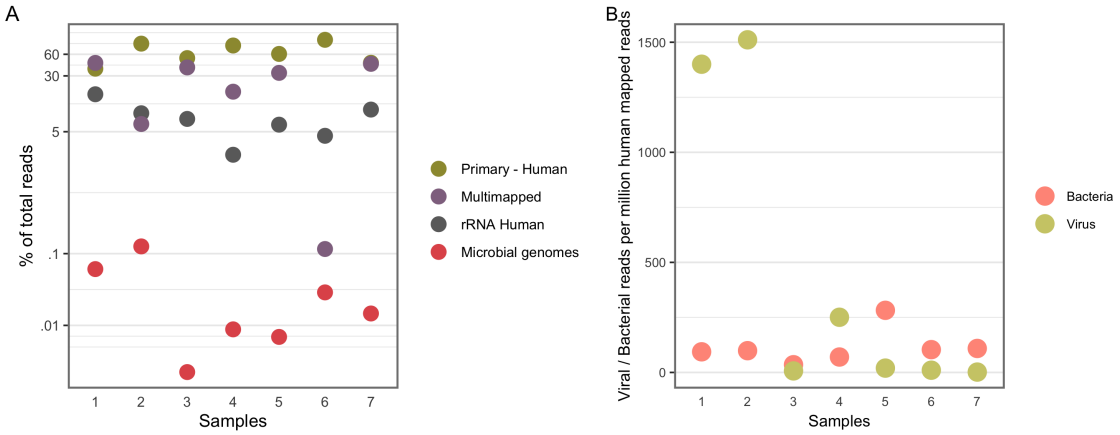


Figure 4.2: General Mapping Statistics. A) Samples and percentage of reads mapped to different groups. Y-aixs is in log scale to resolve low-expression reads. B) Shows separate bacterial and viral reads mapped per million human mapped reads per sample.

Proteus phage is the second organism with high number of absolute reads assigned to it (~7000).

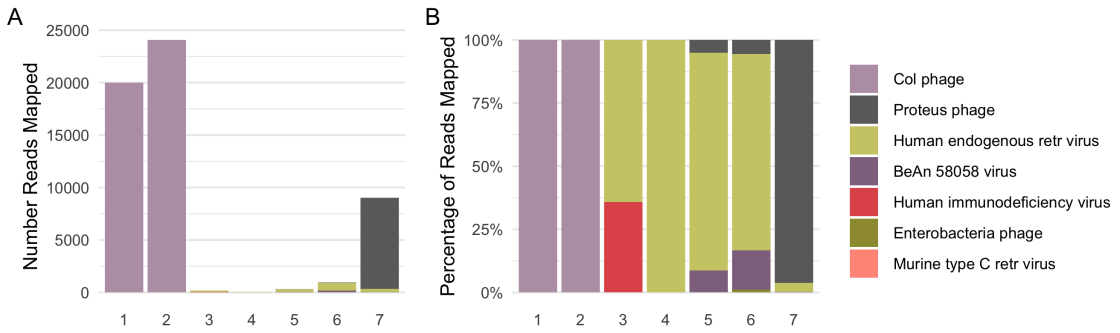
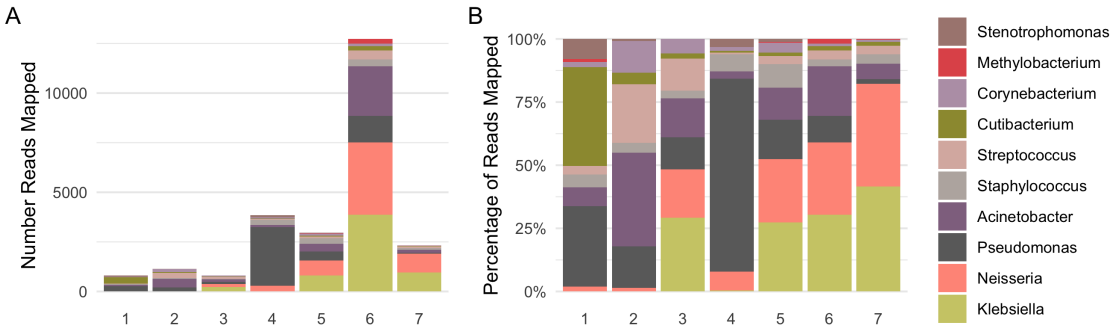


Figure 4.3: Possible viral contaminants. A shows the absolute number of reads confidently assigned to different viral species across different samples, while B represents them as relative abundances.

PhiX contaminants in samples 1 and 2 are most probably delibrately added as spike-in in illumina HISEQ platforms to increase nucleotide diversity. Samples 1 and 2 were sequenced on the HISEQ-2500[57] and HISEQ-4000[55] platforms while all in-house samples were sequenced on HISEQ-2000[27] which has a separate, dedicated lane for the PhiX spike-in quality control to avoid PhiX reads to appear in the FASTQ files.[1] Unusually high number of viral reads were seen in Sample 7, mostly that of Proteus phage VB_PmiS-Isfahan (NC_041925). Like other phage viruses, the Proteus phage also infects bacterial cells, specifically *Proteus mirabilis* a highly motile

[1]Information obtained from Oregon State University's Core facilities website Link.

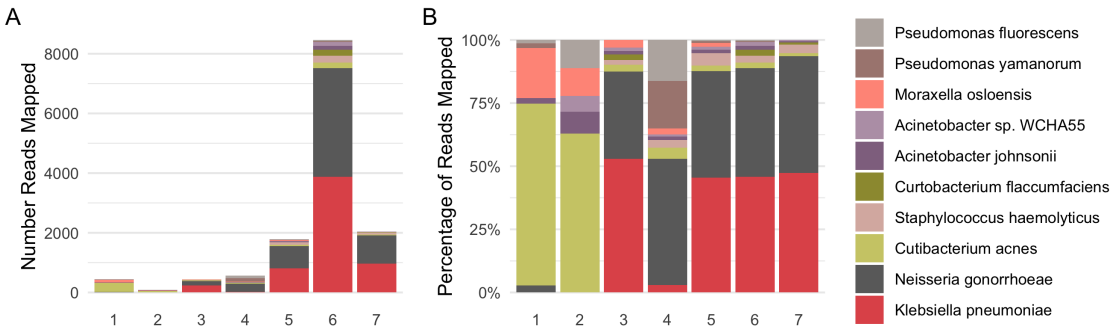
bacterium belonging to the *Enterobacteriaceae family*, which is the most common species responsible for catheter-associated urinary tract infections[59]. No reads were confidently assigned to the *Proteus* genera of bacteria, however, proteus phage is considered to be highly lytic and there are several other bacteria belonging to the *Proteus* genera which are ubiquitously present on and in human guts which could be infected by the phage[60], hence making this assignment plausible. All other viruses detected were less 200 reads per sample except those that mapped to the HERVs. These are viral sequences that represent ancient viral infections that affected the primates' germ line and became stably integrated into the host genome. This was an interesting find as ~ 8% of the human genome is said to be of viral origin, the HERVs. The reason behind its baseline expression in most of the adult tissues nor its role in different pathologies are well defined[61–63].



\begin{figure}
\caption[Analysis of the possible bacterial contaminants — Genus level]{Possible bacterial contaminants at the Genus level. A shows the absolute number of reads confidently assigned to different bacterial genera across different samples (accounting for ~80% of all the reads assigned to bacteria), while B represents them as percentage proportions.} \end{figure}

Unlike viral, the bacterial reads were analyzed at two levels, genus (figure 4.2) and species (figure 4.2). The 10 genera shown account for about ~80% of all reads mapped to bacteria while the 10 species shown account for about ~50% of all reads mapped. All samples except Sample 6, are within acceptable limits of bacterial reads/sample, less than 100 bacterial reads/million human mapped reads[64]. Sample 6 has ~13500 reads in absolute numbers and approximately 250 bacterial reads/million human mapped reads, most of which are accounted for by 4 genera — *Acinetobacter*, *Klebsiella*, *Neisseria*, *Pseudomonas*. This is also reflected at the species level, where *Klebsiella pneumoniae* and *Neisseria gonorrhoeae* account for 57% of the entire bacterial contamination found in the sample while the rest is account for by 189 other species. Low levels of

both these bacteria are also found in the other in-house samples. The presence of *Cutibacterium acnes*, *Pseudomonas* and *Acinetobacter* bacterial contamination has been well documented owing to their epidermal persence in the first case and to water associated presence, even ultra-purified, in the last two cases[65–67]. While *Klebsiella* has been associated with pathologies, it is also a known opportunistic pathogen which is a normal part of the microbial flora of mucosal surfaces such as the mouth and throat and found ubiquitously in nature/environment[68]. Likewise, although *Neisseria gonorrhoeae* is not a part of the normal flora, it could also present as benign/unnoticed infections of the mucosal surfaces urogenital tract, pharynx, and rectum, apart from causing a full-blown pathological disease[69]. The discovery of bacterial reads in cell line data and the finding of different bacterial taxa in data from different sequencing runs/groups/labs supports the idea that a good portion of bacterial reads are possibly not derived from the specimens themselves.



\begin{figure}
\caption[Analysis of the possible bacterial contaminants — Species level]{Possible bacterial contaminants at the species level. A shows the absolute number of reads confidently assigned to different bacterial species across different samples (accounting for ~50% of all the reads assigned to bacteria), while B represents them as percentage proportions.} \end{figure}

The results from these 7 representative samples are in-line with papers published which looked at the microbial contamination in RNA-Seq samples in terms of their diversity and the range of proportions of microbial contamination differing between different samples. For instance, work by[64], showed that *Acinetobacter* contributed to the highest number of reads. While the study by[70], also picked up a large number of samples having PhiX phage reads in them. This work is not exhaustive and the actual source of contamination was not actively sought after, however, it points towards a problem that could benefit from a more systematic address.

4.3 Global view of the transcriptomic data

Data from 68 samples were collected (20 iPSC-CMs, 17 EHM, 16 Fetal Heart samples, 15 Adult Heart Samples) across different studies as shown in 3.1. To examine global trends in gene expression levels, the normalized data across all samples were visualized using principal component analysis (figure 4.5). The major source of variation in the data is correlated with the sample type and is captured by the first PC accounting for 42.9% of the variation in the data set (see B of 4.4 and A of 4.4), where the first PC effectively separates the sub-groups with an almost uni-directional progression from cardiomyocytes to EHMs to fetal heart tissue and adult heart subtypes, showing that this vector captures the increase in complexity of the tissue types, figure 4.4 B.72.4% of total variation in the dataset can be explained by the first 4 PCs and the cumulative percentage of the variances explained by the first eight PCs is shown in the scree plot (A of 4.4).

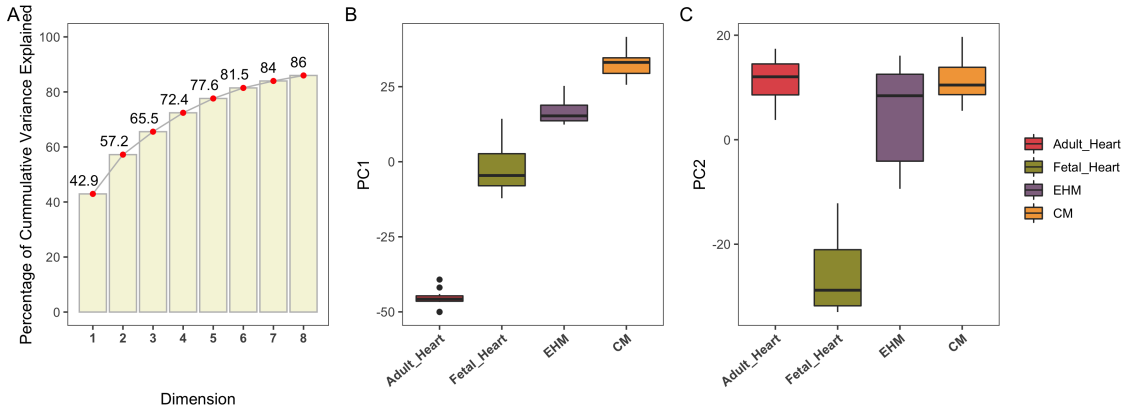


Figure 4.4: A shows the cummulative variance explained by the PCs. B and C show the different groups plotted against PC1 and PC2 respectively.

The second largest source of variation is captured by PC2 accounting for 14.3% of the toal variance and separates the fetal heart samples from rest of the sample types as shown by the ordering of sample types as per PC2, see C of 4.4.

Cardiomyocyte samples which are >90% actinin+ in FACS are in-essence representative of a single cell type, while EHMs have the additional stromal cells and fetal heart tissues constitute a variety of cell types and sub-types including differentiated and differentiating cells along different lineages while the adult heart is composed of not just cardiomyocytes and fibroblasts but also endothelial cells, immune cells, vascular smooth muscle cells and cells making up the conduction system. Fetal heart samples show relatively loose clustering compared to adult heart samples. This is not too

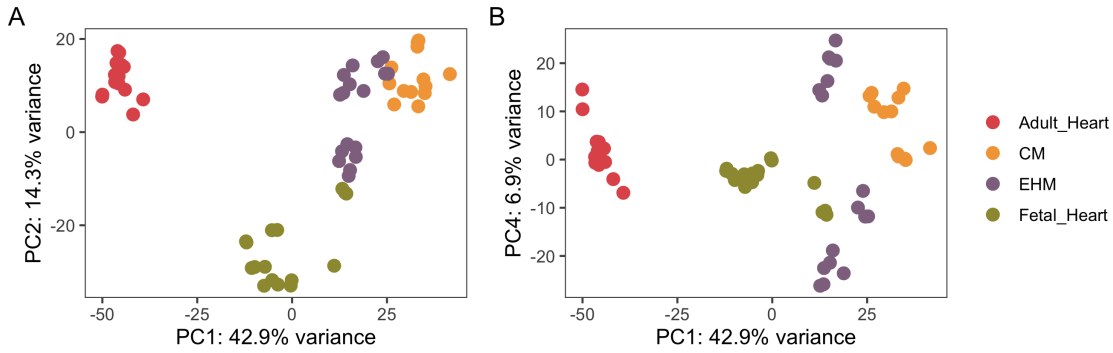


Figure 4.5: A and B show the PCA plots of all the samples againt PC1 and PC2/PC4. The samples are coloured based on their groups.

suprising because the fetal samples were collected across different gestational periods, in the range of 9 weeks - 16 weeks.

The EHM samples also show two semi-distinct clusters, corresponding to the two EHM sources — in-house and from PRJNA362579 project. The source of this variation within the EHM samples appears to be strongly associated with PC4, see figures 4.6 and 4.5 B.

As explained in 1.5.1, the loading scores of genes can be explored further to look for possible meaningful explanations of the basis of separations of the clusters. The top 50 genes with the highest absolute loading values is tabulated for the first 4 PCs in 4.3.

Table 4.3: Top 50 Genes with the highest absolute loading in the first 4 PCs

1		2		3		4	
Gene	PC	Gene	PC	Gene	PC	Gene	PC
GPR4	-0.033	ARHGAP33	-0.055	PLAT	-0.064	ANGPT2	-0.071
PDE2A	-0.033	PRRT2	-0.054	MMP14	-0.064	CLMP	-0.068
CD300LG	-0.033	MXD3	-0.054	ITIH3	-0.061	MEG8	-0.067
GIMAP1	-0.033	PIF1	-0.054	TNC	-0.060	F2RL2	-0.066
RAMP3	-0.033	KIF18B	-0.053	TGFBI	-0.059	DOCK10	-0.065
APOD	-0.033	CHTF18	-0.053	LIF	-0.059	SAMD9L	-0.062
SPAAR	-0.033	TROAP	-0.053	ENC1	-0.059	PI15	-0.062
RAI2	-0.033	HBG1	-0.052	LUM	-0.058	CTSK	-0.060
TMEM273	-0.033	SPTA1	-0.051	TIMP1	-0.057	PAMR1	-0.060
APOL3	-0.033	HBG2	-0.051	EFEMP1	-0.057	CCDC144B	-0.059
COL4A5	0.033	FAM95C	-0.051	FN1	-0.057	CLCA2	-0.059
SLC15A3	-0.033	EPB42	-0.051	SRPX2	-0.057	SAMD9	-0.058
SLC9A3R2	-0.033	HEMGN	-0.051	CFH	-0.056	CRB2	0.057

Table 4.3: Top 50 Genes with the highest absolute loading in the first 4 PCs *(continued)*

1		2		3		4	
Gene	PC	Gene	PC	Gene	PC	Gene	PC
PRELP	-0.033	TMEM155	-0.051	RHOU	-0.056	TMEM119	-0.057
FABP4	-0.032	PKMYT1	-0.050	CCL2	-0.056	FAP	-0.057
GPX3	-0.032	YJEFN3	-0.050	IL1R1	-0.056	CITED4	0.057
RORC	-0.032	LENG8	-0.050	VGLL3	-0.056	PCDH18	-0.056
C1QA	-0.032	FOXM1	-0.050	SERPINE2	-0.056	RPL10P6	-0.055
DIRAS1	-0.032	ESPL1	-0.050	BNC2	-0.055	PUS7L	-0.055
TYROBP	-0.032	GRIA2	-0.049	NTM	-0.055	LAMA5	0.055
HSPB6	-0.032	EMC10	-0.049	EMILIN1	-0.055	POSTN	-0.054
FGR	-0.032	SLC4A1	-0.049	SULF1	-0.055	PRRX1	-0.054
SOX18	-0.032	DDX39B	-0.049	FXYD5	-0.055	NT5E	-0.054
CD79B	-0.032	PLK1	-0.049	ZNF280D	0.054	KCNJ2	-0.054
CRYM	-0.032	ALAS2	-0.049	GPRC5A	-0.054	TWIST2	-0.054
RPL3L	-0.032	CDT1	-0.049	CYP1B1	-0.054	ADGRG7	-0.053
FCN3	-0.032	MYBL2	-0.049	COL1A2	-0.054	ZNF528	-0.052
TMEM143	-0.032	MIR503HG	-0.049	RCN3	-0.054	TMSB4XP8	-0.052
PTGDR2	-0.032	CDCA3	-0.049	DKK1	-0.054	LPAR1	-0.052
HLF	-0.032	AGAP6	-0.049	F2RL1	-0.053	LINC01405	0.052
NDRG4	-0.032	TMCC2	-0.049	IGFBP3	-0.053	GABRA4	-0.052
ADGRE5	-0.032	PLEKHG4B	-0.048	NLRP2	-0.053	THBS2	-0.052
ACKR1	-0.032	HBM	-0.048	LOX	-0.052	PRRX2	-0.051
AQP7	-0.032	SOD2	0.048	ITGA11	-0.052	DPP4	-0.051
IGF2BP3	0.032	CDC42	0.048	ZNF680	0.052	AC096664.2	0.051
PLIN4	-0.032	FKBP2	-0.048	SPHKAP	0.052	CXCL10	-0.051
ICAM2	-0.032	CHTF8	-0.048	PTX3	-0.052	ZNF248	-0.051
CD38	-0.032	TTYH3	-0.048	NNMT	-0.052	AC016739.1	0.051
C1QB	-0.032	IGSF9	-0.048	HGF	-0.052	MYH6	0.051
CCM2L	-0.032	FAUP1	0.048	KRT17	-0.052	ZNF736	-0.051
ADAM15	-0.032	BHLHB9	-0.048	CTHRC1	-0.052	DDR2	-0.051
P2RY8	-0.032	NPIPB5	-0.048	MMP9	-0.052	MEG3	-0.050
TNFSF12	-0.032	AP002884.1	0.048	PTGS2	-0.052	FOXP4	0.050
CBX7	-0.032	AHSP	-0.048	ITGA8	-0.052	CLEC2B	-0.050
LGALS9	-0.032	GTSE1	-0.048	COL3A1	-0.052	MMP3	-0.050
DMTN	-0.032	TYMS	-0.048	CDKN2B	-0.051	AL161787.1	0.050
CLEC3B	-0.032	PRRG3	-0.047	SERPINB2	-0.051	CMKLR1	-0.049
SPI1	-0.032	COL6A6	-0.047	FBN1	-0.051	TMEM176B	-0.049
GIMAP7	-0.032	KIFC1	-0.047	BASP1	-0.051	CCN4	-0.049
ECHDC3	-0.032	E2F1	-0.047	VWC2	0.051	CNTFR	0.049

Genes in the first PC account for the majority of separation of adult heart samples from the rest. The genes discriminatory for the adult heart samples not only account for the mature adult cardiomycoytes but also for the other cell types primarily found

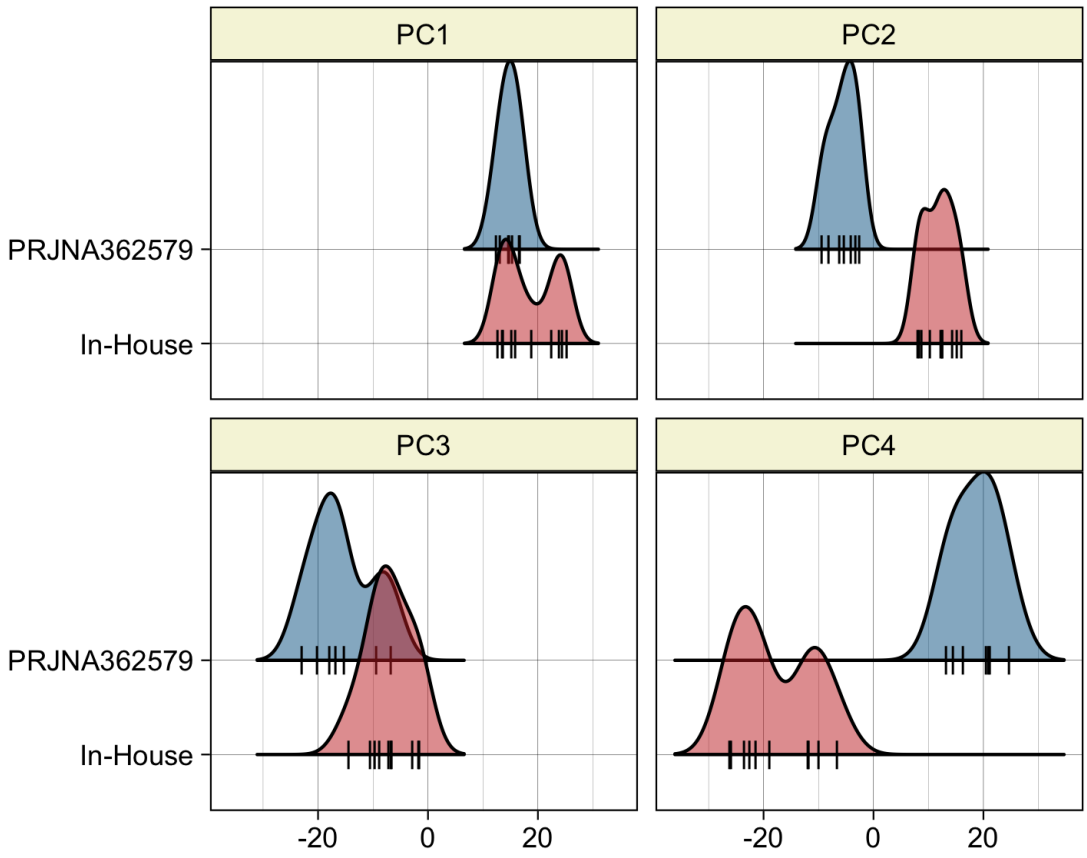


Figure 4.6: Separation of EHMs by the first 4 PCs. The two groups of EHMs and the varying degress of separation by different PCs, visualized by a density plot.

in adult hearts biopsies like AQP7, possibly from the adipose tissue surrounding the heart[71] and APOD from aortic valves of the adult heart[72]. These cell types would be atypical for bioengineered tissues and cultured cell and less common in the fetal heart samples.

While the explanation for the differences between the two EHM groups could be held by the genes in the 4th PC.

Table 4.4: Genes with high loadings and rankings in PC4 along with their known roles. Rows represent the genes and their corresponding ranks in each of the PC.

	PC (% variation explained by the PC)				
Gene	1 (42.9%)	2 (14.3%)	3 (9%)	4 (7%)	Role
Angpt2	1953	1870	1805	1	Involved in the initiation of cardiac angiogenesis
Dock10	1898	563	708	5	Remodeling and expression of Dock10 is vital for normal cardiac function
Pamr1	1560	1975	247	9	Upregulated in a mouse model of cardiac pathological remodeling
Cited4	1609	1277	1913	16	Associated with cardiogenic induction and proliferation capacity of ES cell-derived cardiomyocytes during in vitro cardiogenesis
Kcnj2	1514	1597	1954	24	One of the ion channels in heart
Twist2	1561	1475	530	25	Involved in adult cardiac maintenance
Gabra4	1361	1380	809	31	Detected in early human fetal heart development
Myh6	1801	1262	998	39	Fetal cardiomyocyte marker
Foxp4	1370	498	923	43	Observed before embryonic day 10.5 (E10.5) in cardiomyocytes and, later in the epicardium and endocardium of adult hearts
Cntfr	1944	280	1693	50	Present in adult heart

Gene	References
Angpt2	[73]
Dock10	[74]
Pamr1	[75]
Cited4	[76]
Kcnj2	[77]
Twist2	[78]
Gabra4	[79]
Myh6	[57]
Foxp4	[80]
Cntfr	[81]

Table 4.4 pin-points well studied genes that discriminate the two EHM populations, also note that their ranks are only high in PC4 while much lower in other PCs possibly showing its specificity in the separation of the EHM populations and not others.

The current information, does not allow to quantify the relative closness of the EHM groups to either fetal or adult phenotypes. Other biological and functional characterizations of the EHMs are useful and can be complemented with the deconvolution method to assess populations.

4.3.1 Correlation amongst groups

Pearson correlation of cardiomyocytes, EHMs and fetal heart samples' gene expression to that of adult heart based on the top 2000 genes with the highest absolute loadings in PC1 and PC2 was performed, to check for global similarity in expression patterns and is shown in A.2 of 4.7. These complement the observations from PCA and show that fetal heart samples have the highest similarity to adult heart samples (median ~0.6, with a notable range of 0.3 - 0.65), followed by EHMs (median ~0.45) and lastly by cardiomyocytes (median ~0.25). If we chose ~380 genes which are known to be specifically found and enriched in the human heart, as per Human Protein Atlas, the pearson correlation of the EHM and fetal samples become very comparable as shown in B.2 of 4.7. The median correlation of the EHM samples are even slightly higher than that of the fetal samples, which is in line with our estimation of the EHMs being comparable to (~13wk) fetal tissue[82]. Gene expression across different sample types as per the curated gene list appears to be more consistent and comparable, as visualized

in 4.7 B.1 in comparison to 4.7 A.1. The genes in the heatmaps are hierarchically clustered.

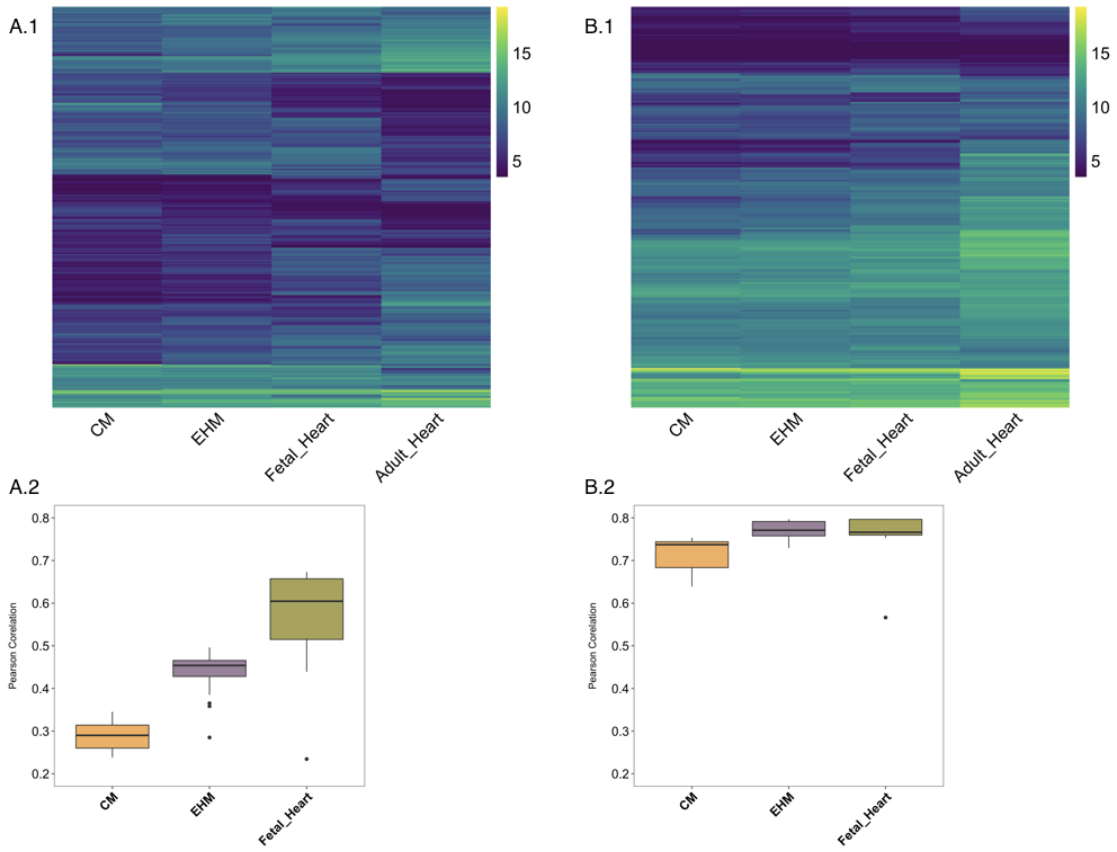


Figure 4.7: Correlation of samples. Heatmaps show VST normalized and group-wise clubbed data. A.1 shows the top 2000 genes with highest loading in PC1 and PC2. B.1 shows a heatmap drawn from a chosen 380 genes based on Human Protein Atlas. A.2 and B.2 show the pearson correlation of the different groups to the adult heart samples based on either the top 2000 genes or the curated 380 genes.

4.3.2 Gene-level analysis

Despite crucial insights garnered from the global view of samples, a view based on curated set of genes are also indispensible and remain one of the most common ways of comparing samples/groups using transcriptomic data. Here we selected four different groups of genes, covering an example of metabolic, structural and general gene lists as shown in Figure 4.8. Panel A is a select representation of the 10 highest and most specifically expressed genes as per human protein atlas. Unsurprisingly the adult heart samples show the highest expression for these genes, except the *Myom1* gene which has been recently found to be a putative marker for hPSC-induced cardiomyocytes' maturity[83], which is almost equally expressed by all samples. Panel

B is set of 22 genes defined to belong to the *sarcomere* gene ontology term. Genes such as *Myo3b*, *Capn3* have a lower expression in the adult heart samples as opposed to the others and are known to have distinct patterns of high expressions during the developmental timeline[84,85]. Panel C is representative of the genes belonging to the oxidation-reduction pathway, while D is a gene list derived from the group's prior publication which deduced that these 50 genes are differentially expressed and considered to be markers for cardiomyocytes. Notably, adult heart associated genes are generally highly expressed, followed by fetal and EHM samples. A few clear exceptions are genes such as *Gpc3* which has a clear and marked role in vertebrate development[86], *Myl4* and *Myl7* are known to be expressed heavily by the developing heart[87–89], and *Tnni3* which is known to be expressed under specific developmental conditions[90]. This gives a snapshot of the approximate level of general (panel A and D), structural (panel B) and metabolic (panel C) maturation/status of the EHMs and cardiomyocytes at a transcript level in comparison to the fetal and adult heart samples. These observations are in line with previous results showing that cardiomyocytes in 2D are less mature than EHMs and at a population/tissue level the EHMs are similar to fetal heart samples.

Exploring the data in the global context and at the gene-level did fortify pre-formed opinions but were of little help in the elucidation of the possibly sub-populations within each sample. This was then addressed by the deconvolution technique, as described below.

4.4 Deconvolution of Bulk CMs and EHMs RNA-Seq Data

Adult and fetal primary heart samples are natively heterogeneous with respect to cell composition, while bioengineered tissues consist of two main components and samples from targeted differentiations should contain one dominant cell type at beginning (iPSC) and end (CM) but go through stages with diverse transient cell populations in between. To rationalize and quantify the heterogeneity present within the in-house bulk samples, computational deconvolution was performed using a public single-cell reference sample[52]. As per the reference data, the last two time points sequenced during the differentiation of iPSCs to CMs are represented as Day 15 and Day 30 with two distinct sub-populations within each group. Day 15 is characterized by *non-contractile, fibroblast-like* sub-group (denoted as d15:S1) and a *contractile, cardiac-like/cCM* sub-group (denoted as d15:S2). Similarly day 30 is characterized by two sub-groups — *non-contractile cells* (d30:S1) and *cardiomyocytes/dCM* (d30:S2)

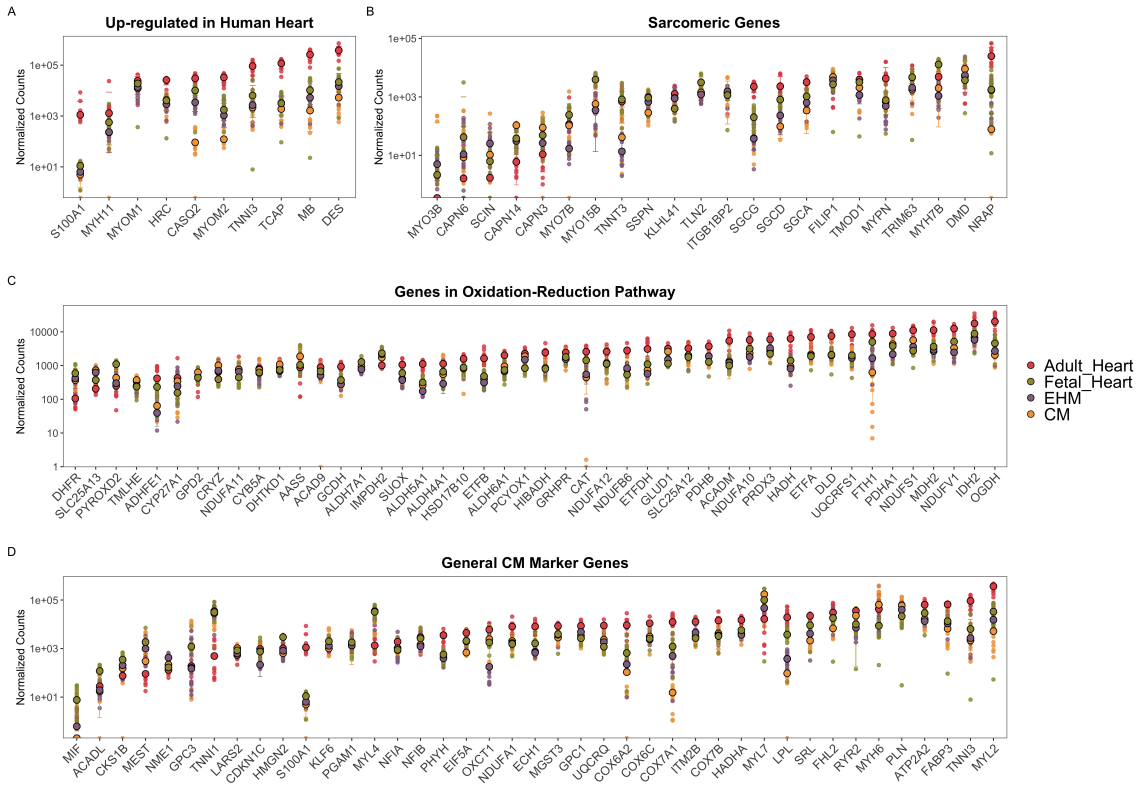


Figure 4.8: At individual gene level, the EHMs resemble the fetal heart expression levels. The graph shows normalized counts of different panels of gene sets. The filled in circles with a border show the median of each group while the slightly transparent and borderless circles represent each sample for every gene. A shows the 10 highest and most specifically expressed genes as per human protein atlas. B contains 22 genes belonging to the sarcomere gene ontology term. C is representative of the genes belonging to the oxidation-reduction pathway and D is a gene list derived from the group's prior publication which deduced that these 50 genes are differentially expressed and considered to be markers for cardiomyocytes

), as tabulated in B of 4.9. Methodologically we followed a state of the art protocol 3.2 where d15 and d30 reference data were processed and used in CIBERSORTx to generate a signature-matrix, which is then used to deconvolve the bulk samples, whose results are shown as a percentage proportion bar graph in A of 4.9. CMs samples are considered highly (>90%) pure cardiomyocyte state while EHM samples were prepared from 70% cardiomyocytes and 30% stromal cells (~30%). To scale the CM samples for the 30% FBs in the EHM sample a "virtual" cell type or comparability factor was added. The CM sample shows a larger proportion of d15:S1 phenotype-like cells (~57%) and a smaller proportion of d30:S1 (mature cardiomyocyte-like) cells as compared to while EHMs, supporting the assumption that CMs continue maturation when being exposed to the 3D EHM environment. The increase in duration of culture has been proposed to increase the maturity of the CMs, as also clearly evidenced by the deconvolution. It is also interesting to note that the deconvolution picked up the sparsity of non-cardiomyocyte-like cells in the CMs sample, only about ~9% of the sample is represented by the S2 sub-groups across both time-points, showing that the majority is composed by cardiomyocytes-like cells, a validation of the current differentiation protocol and at the same time, the increase in proportion of S2 sub-groups in the EHM samples accounting for its difference in composition.

As explained in 1.6, the current working group envisions to conduct clinical trials employing the usage of EHMs for treatment of heart failure. Given that CMs are a key component in the producing EHMs, it is vital that there is batch-to-batch consistency maintained. Currently four batches of CMs have been analyzed in 4.10 post bulk-sequencing and subsequent deconvolution. The variation across samples within proportions sub-groups is not signficant (4.10 A). Each sample and its corresponding deconvoluted putative sub-group percentages are shown in B of 4.10. It is evident that majority of any CM sample is accounted for by the d15:S2 (immature cardiomyocyte) phenotype (~ 83%) which is in-line with the evidences from functional and structural experiments of the group. Although the percentages of ACNT2+ cells, as established by flow cytometry, of these samples could not be correlated with any of the sub-groups' proportions, these results in their stand alone way show that across these four GMP production runs the CM samples are consistent.

To see whether the difference observed between the two EHM sub-groups (in-house and from the project PRJNA362579) in PCA could be addressed using the deconvolution technique, a sample-level deconvolution was performed, as shown in 4.11. It is seen that the EHM samples from project PRJNA362579 have a higher average percentage of

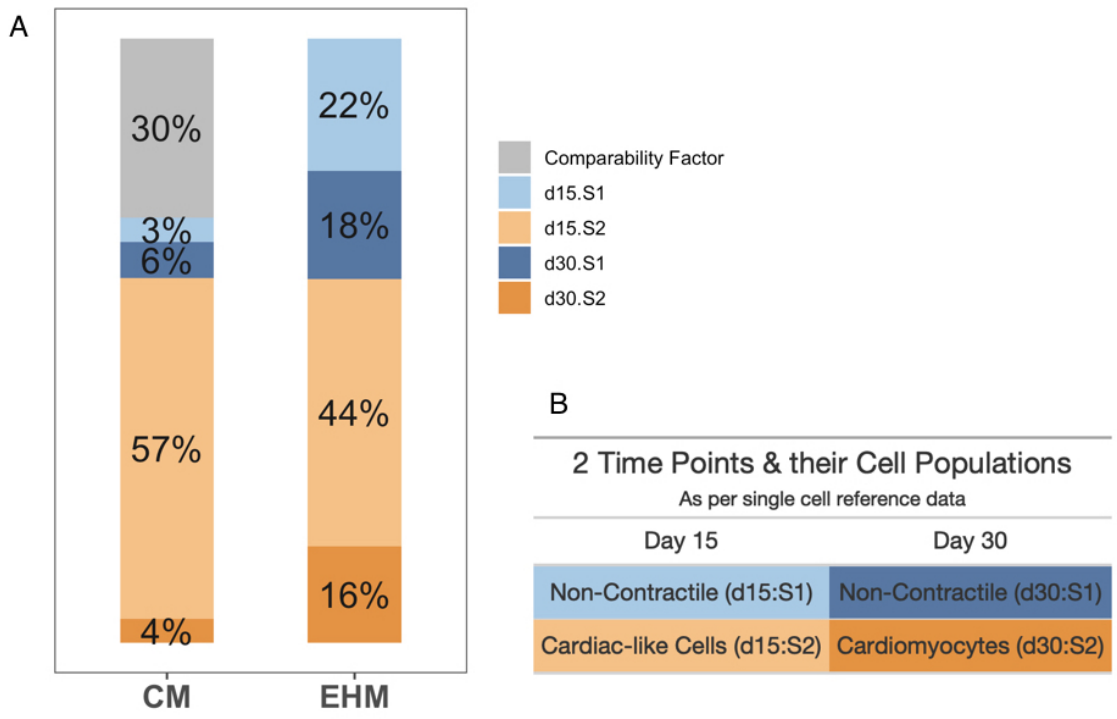


Figure 4.9: EHMs have a higher proportion of mature cardiomyocytes in comparison to CMs. A shows percentages of the various groups, that are explained in B, along with a comparability factor added to the CM samples to allow for direct comparison of both CMs and EHMs.

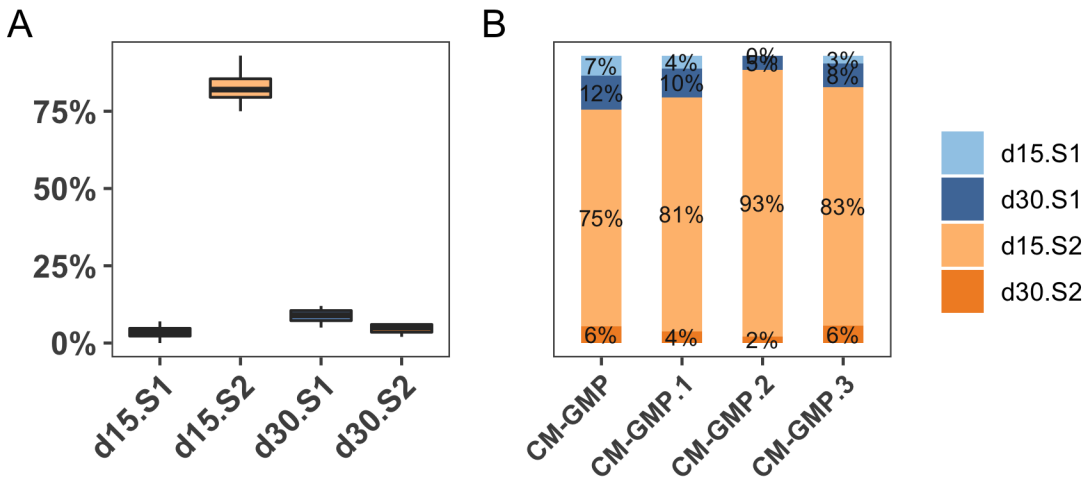


Figure 4.10: The GMP-CM samples are mostly consistent wrt subpopulations derived from deconvolution. A shows the box plots of the contribution of four sub groups across the four samples. B is a sample-wise representation of the sub-groups.

mature-cardiomyocytes (d30:S2 ~ 23%) and more specificially, a sub group of the EHMs within it which the authors consider were more mature (~25%) due to differences in media supplementation (those labelled with MM medium). The in-house samples show higher variation amongst the proportions of different sub-groups. This apparent consistency seen in the EHM samples from PRJNA362579, could be attributed to the fact that all these samples are from one study, primed for one hypothesis, while the in-house EHMs were sequenced across different years and different productions runs, along with changes in protocols.

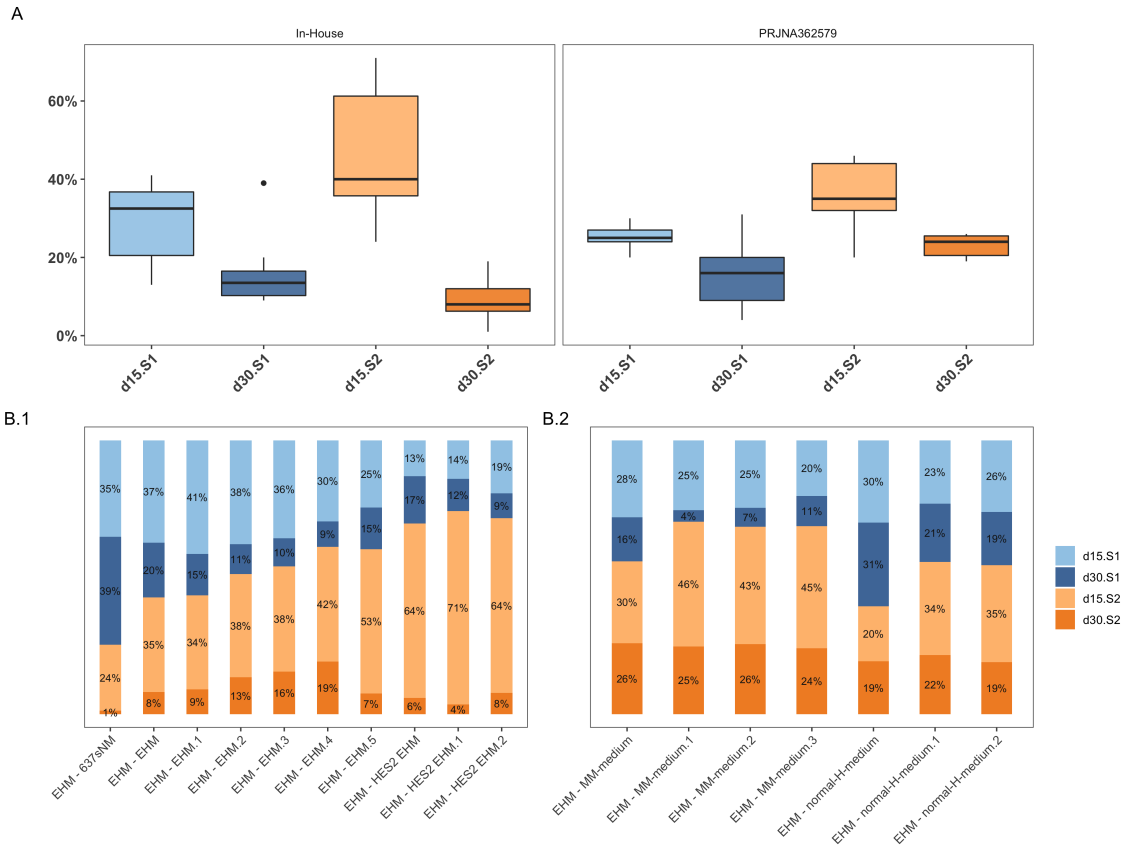


Figure 4.11: Differences in subpopulations amongst EHM samples. A shows the box-plots of percentage contributions of the various groups as per the source of the sample — in-house or from the PRJNA362579 project. B shows the corresponding sample-wise breakdown of the subpopulations in bulk data.

4.4.1 Limits of deconvolution

To address the reliability of the deconvolution technique, all sample groups (iPSCs, fibroblasts, CMs, EHMs, adult and fetal heart) were deconvolved using the same reference dataset, as shown in 4.12. We can see that all the groups were *deconvolved*,

yet, the reliability of the results varied widely. For instance, the root mean squared error (RMSE) of iPSCs and Fibroblasts are close to 1.0 while that of fetal heart group is around 0.3.[2] The correlation of the deconvolved results to the actual reference sample also had a large range, 0.19 for iPSC and 0.84 for the fetal heart. This shows that although deconvolution would work for any given sample and a putative subpopulation would be estimated for that sample, it could be totally irrelevant and unusable, as in the case of iPSCs or fibroblasts or fairly accurate as in the case of fetal heart samples. Cummulatively, adult, EHM and CM samples all have a mediocre correlation and RMSE.

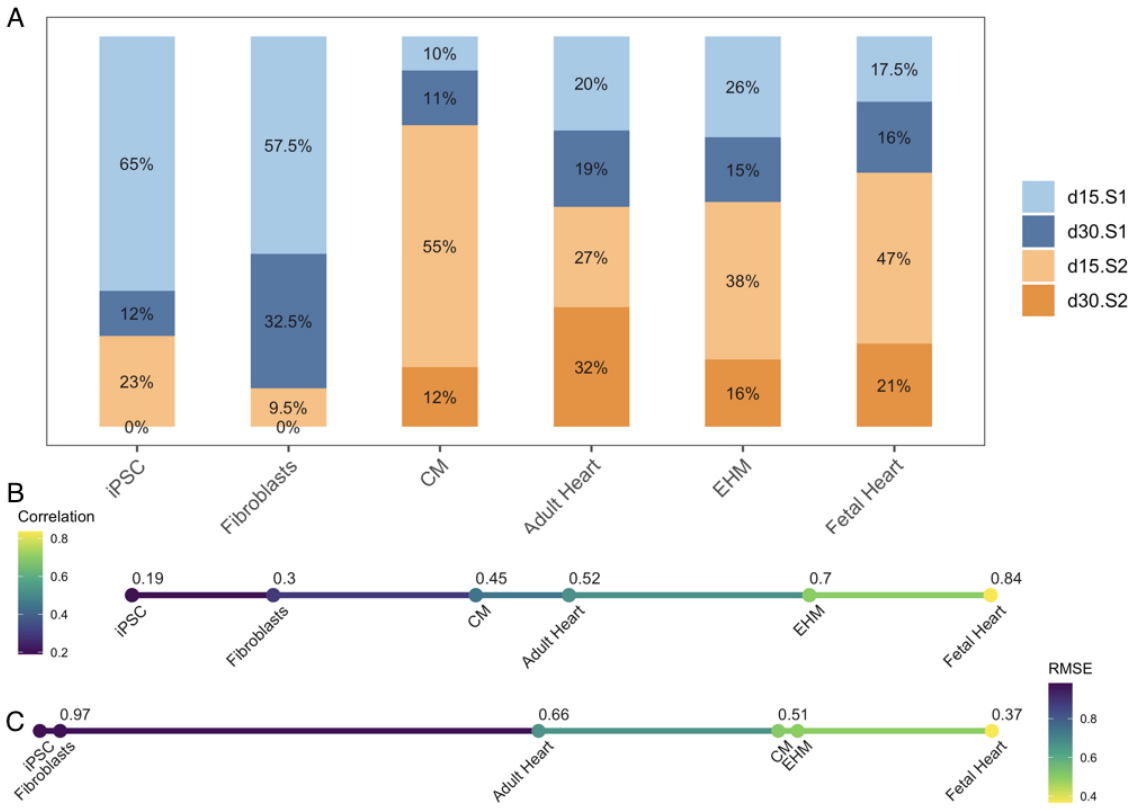


Figure 4.12: A shows the average deconvolution percentages of all groups to the four different subtypes. B and C show the median correlation and RMSE of each group.

[2]The RMSE is the square root of the variance of the residuals. It indicates the absolute fit of the model to the data — how close the observed data points are to the model's predicted values. Lower values of RMSE indicate better fit.

4.5 Basic characterisation of Rhesus Cardiomyocytes

To see where the cardiomyocytes produced from rhesus-iPSCs were placed in comparison to the human-iPSCs cardiomyocytes, the RH-CM samples were sequenced and processed. These rhesus cardiomyocytes were produced to produce EHM patches which were then transferred to a heart failure model of non-primate rhesus models. After retaining only the 1:1 orthologous genes and accounting for variations in gene lengths, the resultant samples were visualized in a basic PCA plot (4.13). All the samples are separated according to the cell type, starting with the first PC separating stromal cells and cardiomyocytes while the second PC separating iPSCs from the rest. Here, the human cardiomyocytes and rhesus cardiomyocytes (denoted as RH-CM) cluster together, and further analysis using DE showed that most of these differences were due to non-cardiac factors such as differences in signalling pathways and neuronal components possibly due to the fact that the rhesus cardiomyocytes were induced using the same protocol used for human cells which does not account for the differences in intricacies in the signalling pathways between the two.

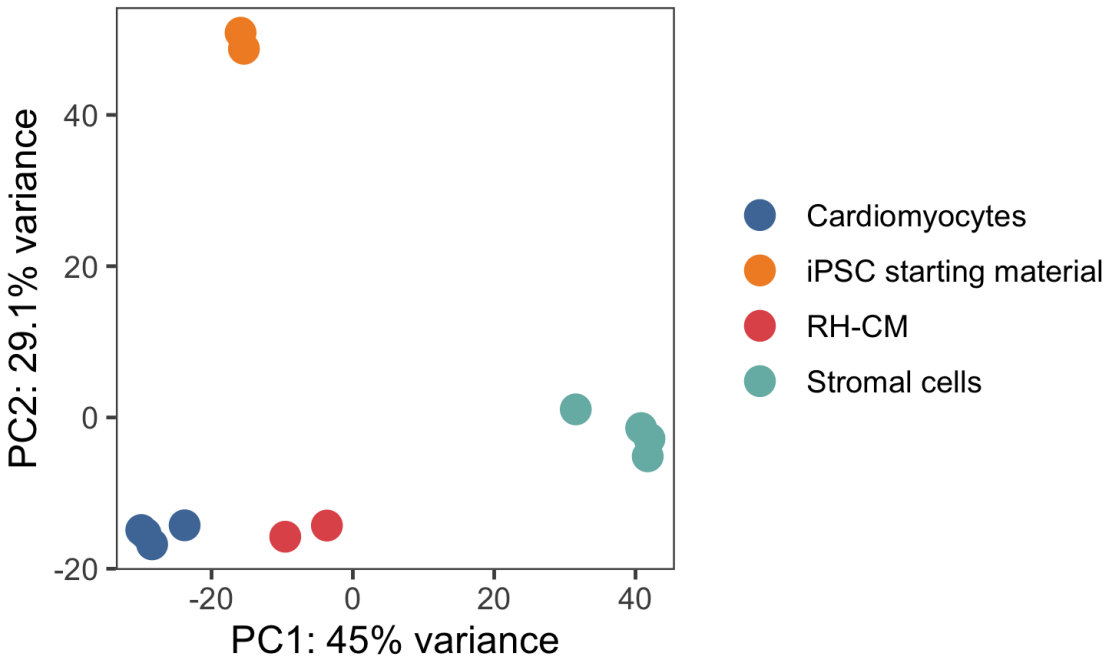


Figure 4.13: The rhesus CMs cluster with the human CMs.

CHAPTER 5

CONCLUSION AND FUTURE WORK

Through this project we sought to explore the potential of RNA-Seq to contribute to the characterization of iPSC induced CMs and EHMs for clinical applications. This was achieved by firstly establishing a workflow to analyse the sequenced data and secondly by using a computational deconvolution technique to gain sub-population level knowledge from bulk data.

The results obtained from computational deconvolution was a proof of principle wherein this *digital cytometry* technique, which has not yet been employed in this particular context, showed that the CMs in the pure CMs sample are less mature than their counterparts found in EHMs. Before this process becomes a regular part of quality control or characterization, limitations of the technique also need to be addressed, such as the fact that the results obtained by computational deconvolution is only *as good as* as the scRNA-Seq dataset. The potential granularity of information that can be obtained and its true revelance in deconvolution is mostly dependent on the scRNA-Seq reference dataset used. In this project, the scRNA-Seq reference used a protocol that is comparable and had the same end result, i.e., to produce CMs from iPSCs, yet, it is not the exact same protocol used in-house. To make this really robust, comparable and of true value in monitoring the CMs or EHMs across different production runs, it would be prudent to produce a standarized in-house scRNA-Seq reference dataset.

A part of the thesis also explored the presence of potential microbial contaminants as a part of quality checking and a representative subset of samples were analysed. Here the results were in-line with most others' findings in this area and although it does not significantly affect any downstream analysis, warrants a deeper investigation at a more systematic level.

This was an exploratory analysis work and used data from varied sources none of which were aimed particularly for the questions asked in this project, for instance, the experimental design and the choices therein, sample sizes, replicates etc., were not all the same nor for the chosen question. The best available was done to account for potential batch effects, yet, it remains a possible limitation to consider.

In summary, we demonstrated the potential of using computational deconvolution techniques to gain sub-population level information in bulk data and its possible role in aiding the refinement, quality control of the protocols to produce iPSC-induced CMs and EHMs.

References

1. Taylor, C. J. *et al.* Trends in survival after a diagnosis of heart failure in the United Kingdom 2000-2017: Population based cohort study. *BMJ* **364**, (2019).

2. Liao, L., Allen, L. A. & Whellan, D. J. Economic burden of heart failure in the elderly. *PharmacoEconomics* **26**, 447–462 (2008).

3. Cook, C., Cole, G., Asaria, P., Jabbour, R. & Francis, D. P. The annual global economic burden of heart failure. *International Journal of Cardiology* **171**, 368–376 (2014).

4. Lesyuk, W., Kriza, C. & Kolominsky-Rabas, P. Cost-of-illness studies in heart failure: A systematic review 20042016. *BMC Cardiovascular Disorders* **18**, 74 (2018).

5. Trivedi, J. R. *et al.* (574) - Risk Factors of Waiting List Mortality for Patients Awaiting Heart Transplant. *The Journal of Heart and Lung Transplantation* **35**, S214 (2016).

6. Eurotransplant - Statistics.

7. Bergmann, O. *et al.* Dynamics of Cell Generation and Turnover in the Human Heart. *Cell* **161**, 1566–1575 (2015).

8. Thomson, J. A. *et al.* Embryonic stem cell lines derived from human blastocysts. *Science (New York, N.Y.)* **282**, 1145–1147 (1998).

9. Takahashi, K. *et al.* Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131**, 861–872 (2007).

10. Burridge, P. W., Keller, G., Gold, J. D. & Wu, J. C. Production of de novo cardiomyocytes: Human pluripotent stem cell differentiation and direct reprogramming. *Cell Stem Cell* **10**, 16–28 (2012).

11. Nguyen, P. K., Neofytou, E., Rhee, J.-W. & Wu, J. C. Potential Strategies to Address the Major Clinical Barriers Facing Stem Cell Regenerative Therapy for Cardiovascular Disease: A Review. *JAMA cardiology* **1**, 953–962 (2016).

12. Inagawa, K. & Ieda, M. Direct reprogramming of mouse fibroblasts into cardiac myocytes. *Journal of Cardiovascular Translational Research* **6**, 37–45 (2013).

13. Kubin, T. *et al.* Oncostatin M is a major mediator of cardiomyocyte dedifferentiation and remodeling. *Cell Stem Cell* **9**, 420–432 (2011).

14. Gnecchi, M. *et al.* Paracrine action accounts for marked protection of ischemic heart by Akt-modified mesenchymal stem cells. *Nature Medicine* **11**, 367–368 (2005).

15. Sekine, H. *et al.* Cardiac cell sheet transplantation improves damaged heart function via superior cell survival in comparison with dissociated cell injection. *Tissue Engineering. Part A* **17**, 2973–2980 (2011).

16. Yang, T. *et al.* Cardiac engraftment of genetically-selected parthenogenetic stem cell-derived cardiomyocytes. *PloS One* **10**, e0131511 (2015).

17. Weinberger, F. *et al.* Cardiac repair in guinea pigs with human engineered heart tissue from induced pluripotent stem cells. *Science Translational Medicine* **8**, 363ra148 (2016).

18. Zimmermann, W.-H. *et al.* Engineered heart tissue grafts improve systolic and diastolic function in infarcted rat hearts. *Nature Medicine* **12**, 452–458 (2006).

19. Liu, Y.-W. *et al.* Human embryonic stem cell-derived cardiomyocytes restore function in infarcted hearts of non-human primates. *Nature Biotechnology* **36**, 597–605 (2018).

20. Neofytou, E., O'Brien, C. G., Couture, L. A. & Wu, J. C. Hurdles to clinical translation of human induced pluripotent stem cells. *The Journal of Clinical Investigation* **125**, 2551–2557 (2015).

21. Sayed, N., Liu, C. & Wu, J. C. Translation of Human-Induced Pluripotent Stem Cells: From Clinical Trial in a Dish to Precision Medicine. *Journal of the American College of Cardiology* **67**, 2161–2176 (2016).

22. Martin, U. Therapeutic Application of Pluripotent Stem Cells: Challenges and Risks. *Frontiers in Medicine* **4**, (2017).

23. Taylor, C. J. *et al.* Banking on human embryonic stem cells: Estimating the number of donor cell lines needed for HLA matching. *The Lancet* **366**, 2019–2025 (2005).

24. Nakatsuji, N., Nakajima, F. & Tokunaga, K. HLA-haplotype banking and iPS cells. *Nature Biotechnology* **26**, 739–740 (2008).

25. Bogomiakova, M. E., Eremeev, A. V. & Lagarkova, M. A. At Home among Strangers: Is It Possible to Create Hypoimmunogenic Pluripotent Stem Cell Lines? *Molecular Biology* **53**, 638–652 (2019).

26. Han, X. *et al.* Generation of hypoimmunogenic human pluripotent stem cells. *Proceedings of the National Academy of Sciences* **116**, 10441–10446 (2019).

27. Tiburcy, M. *et al.* Defined Engineered Human Myocardium with Advanced Maturation for Applications in Heart Failure Modelling and Repair. *Circulation* **135**, 1832–1847 (2017).

28. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: A revolutionary tool for transcriptomics. *Nature reviews. Genetics* **10**, 57–63 (2009).

29. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biology* **17**, 13 (2016).

30. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods* **6**, 377–382 (2009).

31. Montoro, D. T. *et al.* A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* **560**, 319–324 (2018).

32. Asp, M. *et al.* Spatial detection of fetal marker genes expressed at low level in adult human heart tissue. *Scientific Reports* **7**, (2017).

33. Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews. Genetics* **16**, 133–145 (2015).

34. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols* **13**, 599–604 (2018).

35. Aran, D., Hu, Z. & Butte, A. J. xCell: Digitally portraying the tissue cellular heterogeneity landscape. *Genome Biology* **18**, 220 (2017).

36. Becht, E. *et al.* Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biology* **17**, 218 (2016).

37. Kang, K. *et al.* CDSeq: A novel complete deconvolution method for dissecting heterogeneous samples using gene expression data. *PLOS Computational Biology* **15**, e1007510 (2019).

38. Newman, A. M. & Alizadeh, A. A. High-throughput genomic profiling of tumor-infiltrating leukocytes. *Current Opinion in Immunology* **41**, 77–84 (2016).

39. Quon, G. *et al.* Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction. *Genome Medicine* **5**, 29 (2013).

40. Racle, J., de Jonge, K., Baumgaertner, P., Speiser, D. E. & Gfeller, D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *eLife* (2017) doi:10.7554/eLife.26476.

41. Shen-Orr, S. S. & Gaujoux, R. Computational deconvolution: Extracting cell type-specific information from heterogeneous samples. *Current Opinion in Immunology* **25**, 571–578 (2013).

42. Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature Biotechnology* **37**, 773–782 (2019).

43. Chen, B., Khodadoust, M. S., Liu, C. L., Newman, A. M. & Alizadeh, A. A. Profiling tumor infiltrating immune cells with CIBERSORT. *Methods in molecular biology (Clifton, N.J.)* **1711**, 243–259 (2018).

44. Hudson, N. J., Dalrymple, B. P. & Reverter, A. Beyond differential expression: The quest for causal mutations and effector molecules. *BMC Genomics* **13**, 356 (2012).

45. Cuomo, A. S. E. *et al.* Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nature Communications* **11**, 1–14 (2020).

46. Han, X. *et al.* Mapping human pluripotent stem cell differentiation pathways using high throughput single-cell RNA-sequencing. *Genome Biology* **19**, 47 (2018).

47. McCracken, I. *et al.* Transcriptional dynamics of pluripotent stem cell-derived endothelial cell differentiation revealed by single-cell RNA sequencing. *Eur Heart J* (2019) doi:10.1093/eurheartj/ehz351.

48. Müller, G. A., Tarasov, K. V., Gundry, R. L. & Boheler, K. R. Human ESC/iPSC-based 'omics' and bioinformatics for translational research. *Drug Discovery Today: Disease Models* **9**, e161–e170 (2012).

49. Wesolowska-Andersen, A. *et al.* Analysis of Differentiation Protocols Defines a Common Pancreatic Progenitor Molecular Signature and Guides Refinement of Endocrine Differentiation. *Stem Cell Reports* **14**, 138–153 (2020).

50. Wu, H. *et al.* Comparative Analysis and Refinement of Human PSC-Derived Kidney Organoid Differentiation with Single-Cell Transcriptomics. *Cell Stem Cell* **23**, 869–881.e8 (2018).

51. Freedman, B. S. Better Being Single? Omics Improves Kidney Organoids. *Nephron* **141**, 128–132 (2019).

52. Friedman, C. E. *et al.* Single-Cell Transcriptomic Analysis of Cardiac Differentiation from Human PSCs Reveals HOPX-Dependent Cardiomyocyte Maturation. *Cell Stem Cell* **23**, 586–598.e8 (2018).

53. Kuppusamy, K. T. *et al.* Let-7 family of microRNA is required for maturation and adult-like metabolism in stem cell-derived cardiomyocytes. *Proceedings of the National Academy of Sciences of the United States of America* **112**, E2785–2794 (2015).

54. Mills, R. J. *et al.* Functional screening in human cardiac organoids reveals a metabolic mechanism for cardiomyocyte cell cycle arrest. *Proceedings of the National Academy of Sciences of the United States of America* **114**, E8372–E8381 (2017).

55. Pavlovic, B. J., Blake, L. E., Roux, J., Chavarria, C. & Gilad, Y. A Comparative Assessment of Human and Chimpanzee iPSC-derived Cardiomyocytes with Primary Heart Tissues. *Scientific Reports* **8**, 15312 (2018).

56. Pervolaraki, E., Dachtler, J., Anderson, R. A. & Holden, A. V. The developmental transcriptome of the human heart. *Scientific Reports* **8**, (2018).

57. Yan, L. *et al.* Epigenomic Landscape of Human Fetal Brain, Heart, and Liver. *The Journal of Biological Chemistry* **291**, 4386–4398 (2016).

58. Sangiovanni, M., Granata, I., Thind, A. S. & Guarracino, M. R. From trash to treasure: Detecting unexpected contamination in unmapped NGS data. *BMC Bioinformatics* **20**, 168 (2019).

59. Schaffer, J. N. & Pearson, M. M. Proteus mirabilis and Urinary Tract Infections. *Microbiology spectrum* **3**, (2015).

60. Drzewiecka, D. Significance and Roles of Proteus spp. Bacteria in Natural Environments. *Microbial Ecology* **72**, 741–758 (2016).

61. Grandi, N. & Tramontano, E. Human Endogenous Retroviruses Are Ancient Acquired Elements Still Shaping Innate Immune Responses. *Frontiers in Immunology* **9**, (2018).

62. Küry, P. *et al.* Human Endogenous Retroviruses in Neurological Diseases. *Trends in Molecular Medicine* **24**, 379–394 (2018).

63. Nelson, P. N. *et al.* Demystified . . . Human endogenous retroviruses. *Molecular Pathology* **56**, 11–18 (2003).

64. Strong, M. J. *et al.* Microbial Contamination in Next Generation Sequencing:

Implications for Sequence-Based Analysis of Clinical Samples. *PLoS Pathogens* **10**, (2014).

65. Kéki, Z., Grébner, K., Bohus, V., Márialigeti, K. & Tóth, E. M. Application of special oligotrophic media for cultivation of bacterial communities originated from ultrapure water. *Acta Microbiologica Et Immunologica Hungarica* **60**, 345–357 (2013).

66. Kulakov, L. A., McAlister, M. B., Ogden, K. L., Larkin, M. J. & O'Hanlon, J. F. Analysis of bacteria contaminating ultrapure water in industrial systems. *Applied and Environmental Microbiology* **68**, 1548–1555 (2002).

67. Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology* **12**, 87 (2014).

68. Bengoechea, J. A. & Sa Pessoa, J. Klebsiella pneumoniae infection biology: Living to counteract host defences. *FEMS Microbiology Reviews* **43**, 123–144 (2019).

69. Escobar, A., Rodas, P. I. & Acuña-Castillo, C. MacrophageNeisseria gonorrhoeae Interactions: A Better Understanding of Pathogen Mechanisms of Immunomodulation. *Frontiers in Immunology* **9**, (2018).

70. Park, S.-J. *et al.* A systematic sequencing-based approach for microbial contaminant detection and functional inference. *BMC Biology* **17**, 72 (2019).

71. Madeira, A., Camps, M., Zorzano, A., Moura, T. F. & Soveral, G. Biophysical Assessment of Human Aquaporin-7 as a Water and Glycerol Channel in 3T3-L1 Adipocytes. *PLOS ONE* **8**, e83442 (2013).

72. Nordquist, E., LaHaye, S., Nagel, C. & Lincoln, J. Postnatal and Adult Aortic Heart Valves Have Distinctive Transcriptional Profiles Associated With Valve Tissue Growth and Maintenance Respectively. *Frontiers in Cardiovascular Medicine* **5**, (2018).

73. Nader, M., Yaqinuddin, A. & Kvietys, P. Cardiac Angiogenesis: Role of Cardiomyocytes and Macrophages and Possible Therapeutic Approaches. *Current Angiogenesis (Discontinued)* (2014).

74. Segal, L. *et al.* DOCK10 is vital for normal cardiac function under neurohormonal activation. *Journal of Molecular and Cellular Cardiology* **120**, 18 (2018).

75. Wang, H.-B. *et al.* Identification of differentially expressed genes and preliminary validations in cardiac pathological remodeling induced by transverse aortic constriction. *International Journal of Molecular Medicine* **44**, 1447–1461 (2019).

76. Miake, J. *et al.* Cited4 is related to cardiogenic induction and maintenance of proliferation capacity of embryonic stem cell-derived cardiomyocytes during in vitro cardiogenesis. *PLOS ONE* **12**, e0183225 (2017).

77. Zhao, Z. *et al.* Ion Channel Expression and Characterization in Human Induced Pluripotent Stem Cell-Derived Cardiomyocytes. *Stem Cells International* (2018) doi:https://doi.org/10.1155/2018/6067096.

78. Min, Y.-L. *et al.* Identification of a multipotent Twist2-expressing cell population in the adult heart. *Proceedings of the National Academy of Sciences* **115**, E8430–E8439 (2018).

79. Iruretagoyena, J. I. *et al.* Metabolic gene profile in early human fetal heart development. *Molecular Human Reproduction* **20**, 690–700 (2014).

80. Zhang, Y. *et al.* Foxp1 coordinates cardiomyocyte proliferation through both cell-autonomous and nonautonomous mechanisms. *Genes & Development* **24**, 1746–1757 (2010).

81. Zheng, K., Zhang, Q., Sheng, Z., Li, Y. & Lu, H.-h. Ciliary Neurotrophic Factor (CNTF) Protects Myocardial Cells from Oxygen Glucose Deprivation (OGD)/Re-Oxygenation via Activation of Akt-Nrf2 Signaling. *Cellular Physiology and Biochemistry* **51**, 1852–1862 (2018).

82. Tiburcy Malte *et al.* Defined Engineered Human Myocardium With Advanced Maturation for Applications in Heart Failure Modeling and Repair. *Circulation* **135**, 1832–1847 (2017).

83. Cai Wenxuan *et al.* An Unbiased Proteomics Method to Assess the Maturation of Human Pluripotent Stem CellDerived Cardiomyocytes. *Circulation Research* **125**, 936–953 (2019).

84. Fougerousse, F. *et al.* Calpain3 expression during human cardiogenesis. *Neuromuscular disorders: NMD* **10**, 251–256 (2000).

85. Liu Qing *et al.* Genome-Wide Temporal Profiling of Transcriptome and Open Chromatin of Early Cardiomyocyte Differentiation Derived From hiPSCs and hESCs. *Circulation Research* **121**, 376–391 (2017).

86. Ng, A. *et al.* Loss of glypican-3 Function Causes Growth Factor-dependent Defects in Cardiac and Coronary Vascular Development. *Developmental biology* **335**, 208–215 (2009).

87. Pawlak, M. *et al.* Dynamics of cardiomyocyte transcriptome and chromatin landscape demarcates key events of heart development. *Genome Research* **29**, 506–519 (2019).

88. Schiaffino, S., Rossi, A. C., Smerdu, V., Leinwand, L. A. & Reggiani, C. Developmental myosins: Expression patterns and functional significance. *Skeletal Muscle* **5**, 22 (2015).

89. Wang, T. Y. *et al.* Human cardiac myosin light chain 4 (MYL4) mosaic expression patterns vary by sex. *Scientific Reports* **9**, 1–7 (2019).

90. Sheng, J.-J. & Jin, J.-P. TNNI1, TNNI2 and TNNI3: Evolution, Regulation, and Protein Structure-Function Relationships. *Gene* **576**, 385–394 (2016).