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.

The thesis can be accessed as an online version hosted on gitlab pages with a few interactive graphs. Go to https://h.anandakumar.pages.gwdg.de/masterthesis/

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: 21854665

Address: Christophorusweg 12, Z609, Goettingen.

Phone:+49 176 42028164

E-Mail: harithaa.anand.125@gmail.com

First evaluator (Supervisor): Dr. Tim Meyer

Second evaluator (Supervisor): Dr. Harald Kusch

Date of Delivery: 01.06.2020

Statutory Declaration

I declare that this thesis entitled "Using RNA-Sequencing to Improve Characterisation and Production of iPSC Induced Cardiomyocytes for Heart Failure" was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.
I also state that I have produced my work according to the principles of good scientific practice in compliance with the valid “Richtlinen der Georg-August- Universitaet Goettingen zur Sicherung guter wissenschaftlicher Praxis”.

Date : 01.06.2020

Name: Harithaa Anandakumar

This thesis is dedicated to Snoopy!

ACKNOWLEDGEMENTS

I am extremely grateful to a number of people who have contributed to me being in this position and for the completion of the thesis. A few who deserve a special mention are:

My supervisor — Dr. Tim Meyer. It is to him that I owe the transformation of morphing from someone who did not even know the difference between "relative path" and "absolute path" in the realm of computing and bioinformatics (about a year back), to someone who can sort of make their way through a computer. My biggest take away and my deepest gratitude is to his warmhearted, unwavering support and guidance in making me gain trust in my own abilities and skills.

Second Reviewer — Dr. Harald Kusch. For accepting to review at a short notice and for his time.

Initiator into Computing — Dr. Manuel Nietert. For taking the chance with someone who had no experience in computing nor the capacity to speak German, to clean and work with clinical data in German and for that opportunity and for how it changed my outlook and forged potential paths, I am extremely grateful.

Group Leaders — Prof. Dr. Zimmermann, for his keen and thoughtful comments of the thesis and for the opportunity to be a part of the department. Dr. Malte Tiburcy, for his support throughout the project.

Faceless guides — To all those who posed questions and to all those who answered on stackoverflow, biostars, etc without whom this thesis would have truly not been materialized.

Gytis — for letting me sleep and for really trying to type softly at ungodly hours of the night, and for being the best lockdown companion.

Sudharsana — for the constant sisterly support and for sort of reducing the distance of home from about 6000 km to about 500 km

Friends — Kavya, Viswa, Nathan, Nishanth, Bharath for all their friendly duties

Parents and family — Parents for having had me and for everything else that followed, Family for the constant source of entertainment.

The Pandemic — for drilling down the fragility and resilience of humans and all our

constructs

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

cGMP Current good manufacturing practices

CM Cardiomyocytes

CVD Cardiovascular Diseases

DE Differential Expression

EHM Engineered Human Myocardium

HERV Human Endogenous Retro Viruses

hESC Human Embryonic Stem Cells

HF Heart Failure

hIPSC Human Induced Pluripotent Stem Cells

iPSC Induced Pluripotent Stem Cells

MHC Major Histocompatibility Complex

NCD Non-Communicable Diseases

NGS Next-Generation Sequencing

PC Principal Component

PCA Principal Component Analysis

RMM Rapid microbiological methods

RMSE Root mean squared error

RNA-Seq RNA-Sequencing

scRNA-Seq Single-cell RNA Sequencing

TEP Tissue engineered product

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 globally) and the same number is 72.6 years for an infant born in 2019[1]. In 1900, the top three causes of death were infectious diseases — flu (and pneumonia), tuberculosis and gastrointestinal infections[2]. 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 Figure 1.2). Taken together, we see an aging population strained by NCDs of which cardiovascular diseases (CVD) are the most pronounced (see Figure 1.1). Almost half of the deaths attributed to CVDs are caused by 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[3].
It is estimated that 1-2% of the healthcare budget is spent on HF[4], while the global economic burden is estimated at $108 billion per annum[5] and in Germany the annual prevalence-based costs for heart failure patients are around €25,532[6]. The increasing proportions of the elderly in western societies as well as the developing nations following the trend, 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. However, as per one study[7], 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[8].

Although there are myriad causes of HF, such as isichemic heart disease, aortic or mitral regurtitation (volume stress), aortic or mitral stenosis (pressure stress), congen-

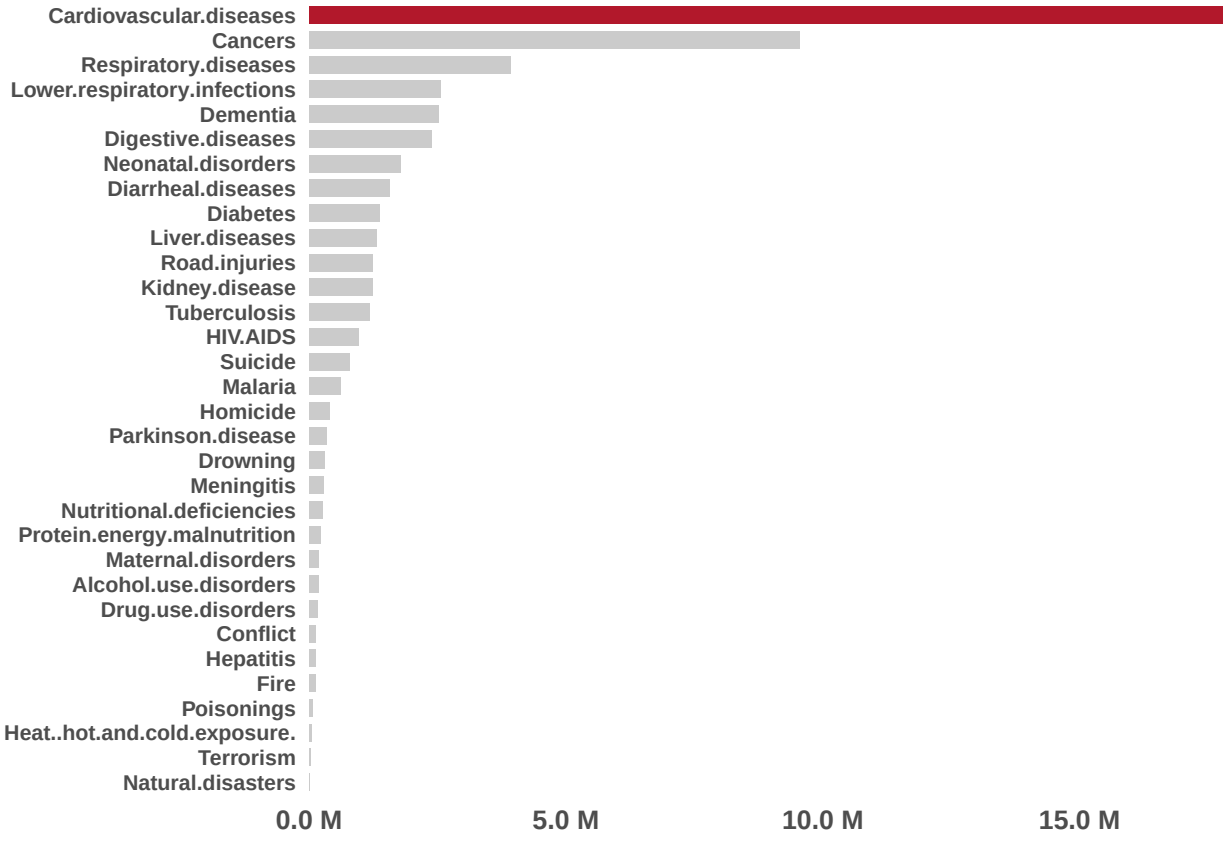

Figure 1.1: Number of Deaths by Cause in the world in 2017. The graph shows the various causes of death in 2017 in the y-axis and the number of deaths per cause in the x-axis in millions. Cardiovascular diseases were responsible for most deaths (15M). Data from: Max Roser and Esteban Ortiz-Ospina (2019)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

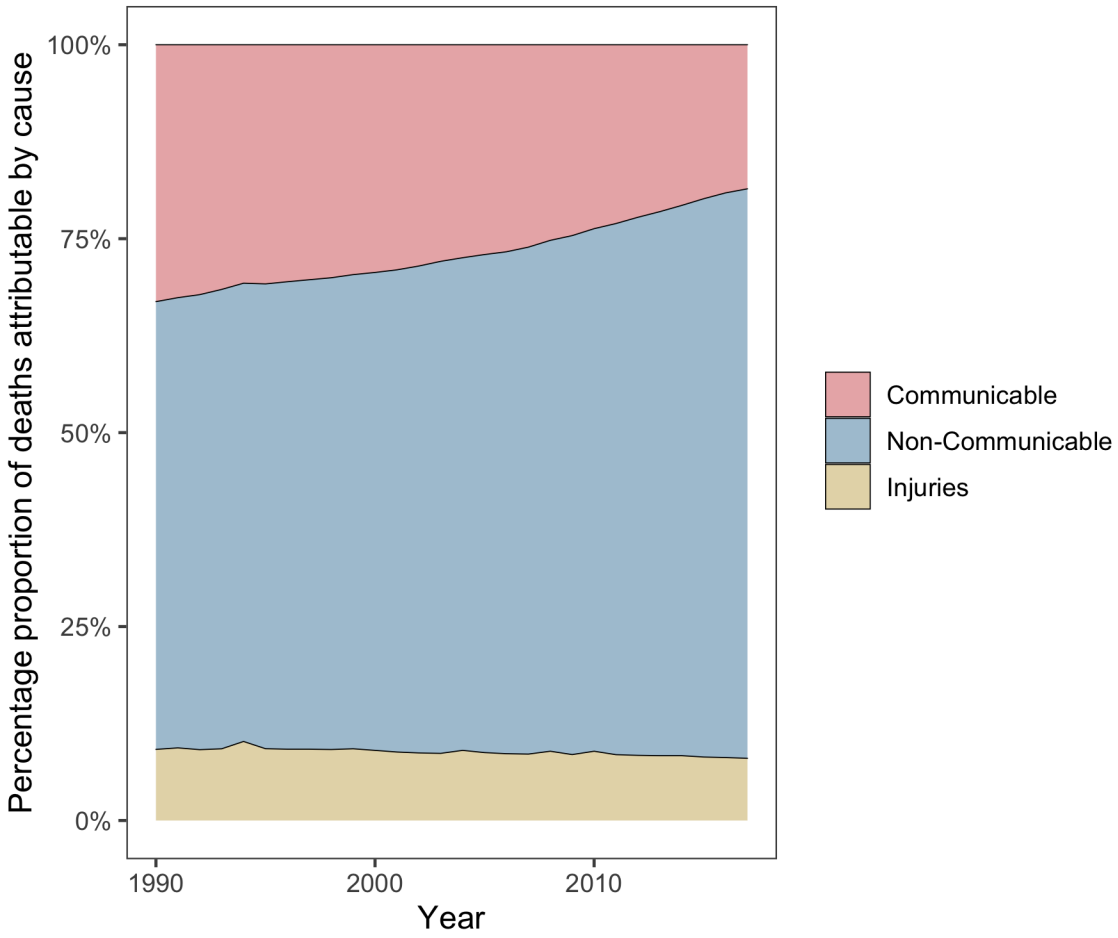

Figure 1.2: Three Major Causes of Death. Graph shows an increase in percentage proportion of deaths over the last three decades due to non-communicable diseases and a parallel reduction in deaths due to communicable diseases.

ital cardiomyopathy, constrictive pericarditis, alcohol excess, anemia, thyrotoxicosis, septicemia, acromegaly, they all commonly operate through the central mechanism of reduced ventricular function[9]. 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 maintanence of 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 paths 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 crippling disease that could immensely benefit from newer therapies. Adult human hearts are terminally differentiated and post-mitotic. A straight-forward approach in the treatment of HF would be to counteract the progressive loss of cardiomyocytes by supplementing the heart with fresh CMs[10]. This has been made possible largely due to the introduction of human embryonic[11] and induced pluripotent stem cells [12]. 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[13]. The straight-forward approach of direct supplementation of CMs by an 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)[14]. Several other strategies to strenghten/remuscularize the

heart such as, converting scar into muscle tissue by transdifferentiation[15], inducing endogenous cardiomyocyte regeneration and proliferation[16], and methods to save the remaining cardiomyocytes from cell death by modulating paracrine factors[17] have been investigated (see Figure 1.3). Despite the limitation in long term engraftment, cardiomyocyte implantation remains the most plausible option from a translational and mechanisitic perspective. 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[18]. 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. Additionally, these engineered patches have not shown to be associated with an increased propensity for arrhythmia[19–21]. 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)[22]. 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[23,24]. Allogenic transplantation[i] of thoroughly characterized iPSCs seems to be a more plausible approach to cell therapy[25]. 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

[i] Two main types of stem cell transplants. *Autologous* — uses a person's own stem cells. *Allogenic* — uses stem cells from an unrelated recipient.

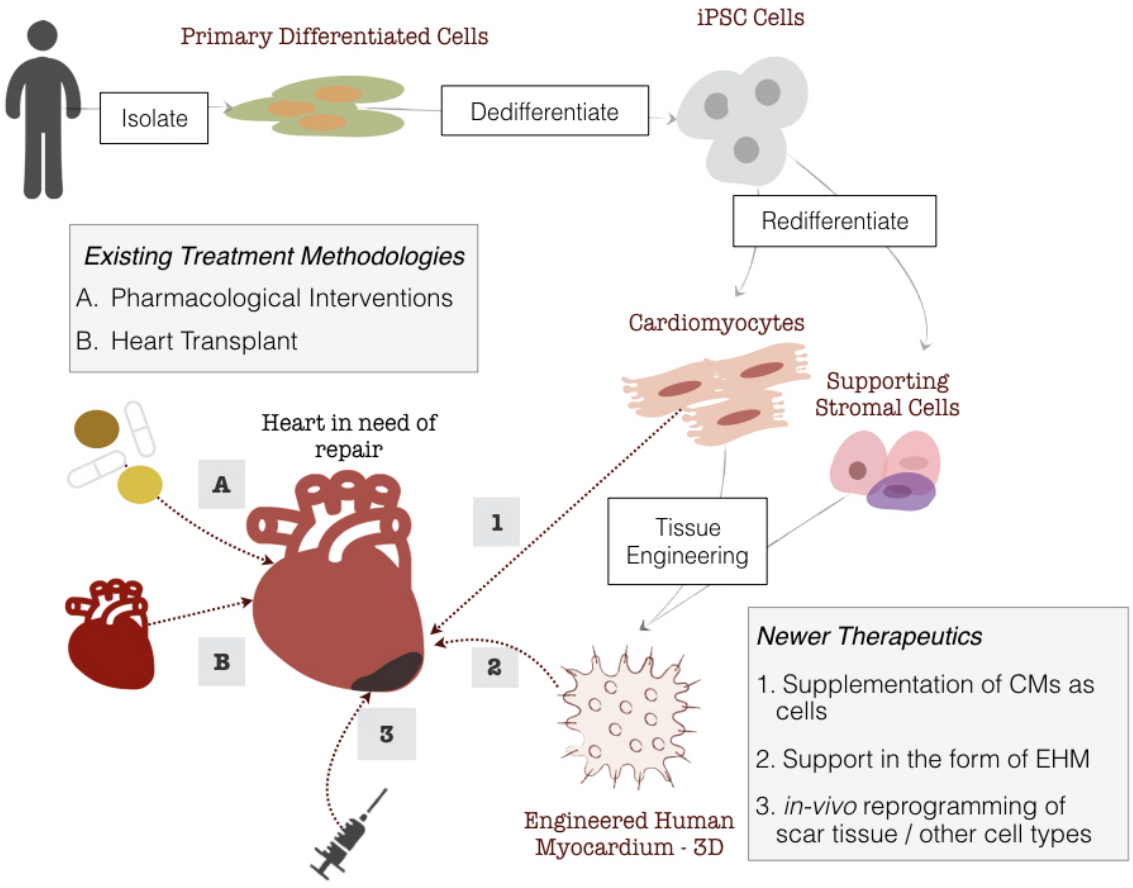

Figure 1.3: Delivery Strategies of iPSC-CMs and Treatment Options for Heart Failure. Production of EHMs start with isolating primary differentiated cells (e.g, fibroblasts) which are then dedifferentiated to iPSCs followed by redifferentiation to CMs and stromal cells, which are then combined in a collagen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

to establish efficiently chosen samples with sufficient HLA diversity to provide a reasonable HLA match for a large percentage of the target population[26]. For instance, a cell bank of 30 iPSC cell lines would enable to find a three-locus match in 82.2% of the Japanese population[27]. 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[28]. 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[29]. Once such cell-lines are authorized for clinical applications, 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 (see Figure 1.4). Several EHM patches may be stacked to make a muscle layer of optimum thickness that is sutured onto a 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[30].

1.2.1 cGMP and Quality Control of Tissue Engineered Products

Tissue engineered products (TEPs) are defined as "products developed for structural and functional repair of tissue/organ defects and their mode of action is to repair, restore or replace tissue structure/function"[31]. Any cGMP facility accredited for the manufacturing of TEPs is designed and organized according to the *Good Manufacturing Practice for Pharmaceutical Manufactures* complying with their *quality assurance* and *quality control* norms, all of which fall under an established *Quality System* approach that regulates all aspects of the collection, processing, storage and release of cell therapy products. The Quality System approach is that of a *risk-based* approach and hence it is of paramount importance to reduce the risk of potential contamination, both microbiological and cross contamination with other products from the same production plant. It is for this reason the Chapters 2.6.1, 2.6.27, 2.6.12 or 2.6.13 of

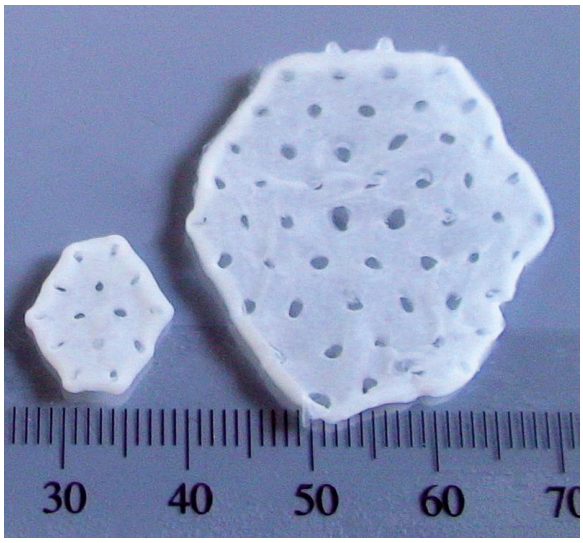

Figure 1.4: Various sizes of EHMs for clinical and experimental applications. Source: Tiburcy et al, 2017.

the *European Pharmacopoeia (Ph. Eur)* are dedicated to the various procedures of testing for microbial contamination that needs to be abided under a cGMP protocol[32]. This is because TEPs are engineered from live donor cells which can not be subjected to a sterilization process and may carry infective agents either on the cell surface or intra-cytoplasmic. They may also harbour latent infective agents integrated in the genome. The media and physical conditions used to manipulate and store the cells are by definition suited for survival and growth of certain tissues, but this could also allow for the latent and insidious growth of microbes. The methods currently in place for microbial testing involve inoculating the products in different media suitable for the growth of common aerobic and anaerobic growth of bacteria (Chapter 2.6.1 of Ph. Eur) where the growth of microbes is detected using turbidity as a measure or detected using the measure of carbon dioxide production (Chapter 2.6.27 of Ph. Eur). These are commonly known as *sterility tests* and although extensive, do not capture all possible bacterial contaminants that could thrive in the cellular growth media. Newer methods such as Rapid Microbiological Methods (RMM) are detection systems which yield equivalent or better results than conventional (microbiological) methods, in lesser time. In this context, an interesting possibility is to explore the potential microbial contamination using high-throughput sequencing data.

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. 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 transcribed 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 mainstay 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[33]. 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, all of which allow for understanding the dynamic and complex nature of the transcriptome.

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 mainstay in the toolkit of all life science research communities. A typical RNA-Seq experiment can be split into three parts[34]:

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 Analysis
3. Advanced Analysis
   - Interpretation (functional profiling)
   - 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 contains 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 1.5.

Single-cell RNA-Seq (scRNA-seq) addressess these issues as it measures the *distribution of expression levels* for each gene across a population of cells[35]. Even in diseases such as cystic fibrosis, which was considered to be well-studied and all potential cell types involved were known, scRNA-seq has revealed a new and unkown cell type, the ionocyte[36]. 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[37]. 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[38,39] are becoming outdated.

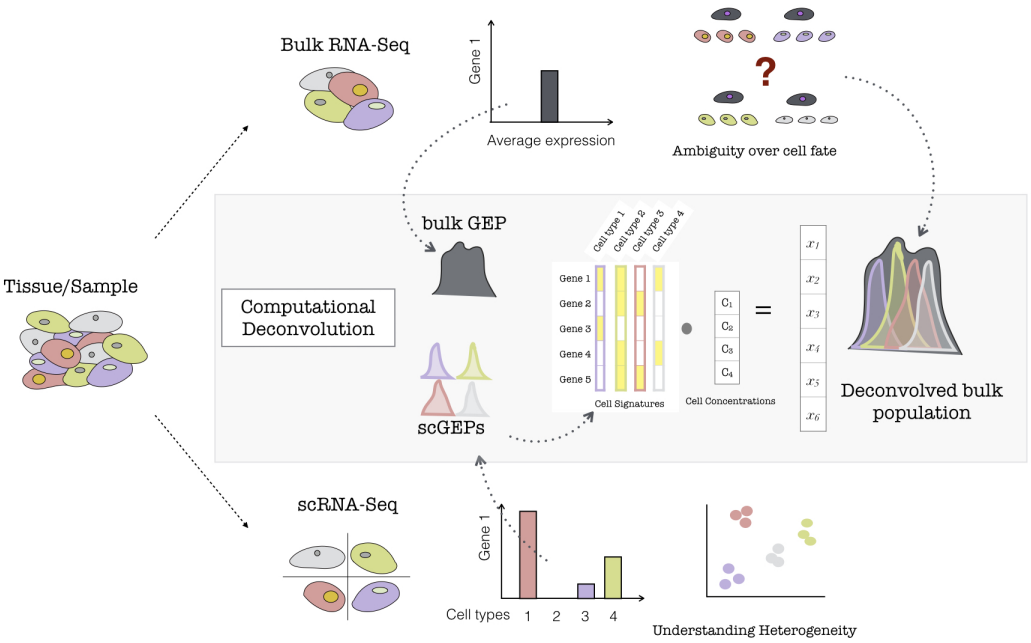

Figure 1.5: Example of Bulk and Single-Cell RNA-Seq. If a population of cells or tissue is considered, it can be sequenced either at the population/bulk level or the single-cell level (For simplicity, different coloured cells represent different sub-populations/cell-types within the sample). If the bulk 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

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[40–46]. Deconvolution, in the realm of 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 deconvolution, CIBERSORTx[47] 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 without 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 H are input requirements. Further explanation of deconvolution and the implementaion of the algorithms can be found at[47,48].
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 1.5.

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.,[49] 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 requir holistic interpretation of 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.[ii] Exploratory data analysis not only helps to find new ways of answering questions but it ultimately permits to detect unexpected patterns and formulate novel working hypotheses.

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 (PC) 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 relevant information in the dataset while uncorrelated noise is pushed to the last components. An example of the application of this method can be found in Witteveen et al.'s article[50]. The authors performed an observational study aiming to investigate the value of early systemic inflammation in predicting ICU-acquired weakness. Systemic inflammation

[ii] A confounding variable is a variable, other than the independent variable that you're interested in, that may affect the dependent variable. The existence of confounding variables in studies make it difficult to establish a clear causal link between treatment and outcome unless appropriate methods are used to adjust for the effect of the confounders.

can be represented by a variety of inflammatory cytokines such as interleukin (IL)-6, IL-8, IL-10, IL-13, tumor necrosis factor and interferon gamma. These cytokines are correlated with each other, and incorporation of them into a regression model will result in significant collinearity. One type of cytokine is regarded as one dimension, and there are dozens of dimensions in the original dataset. In the study, the authors employed PCA to reduce the dimension. They found that the variance of these ten cytokines can be accounted for by three PCs. As a result, the model was remarkably simplified. The goal is to reduce the features' dimensionality with minimal loss of information, for a simplistic example see Figure 1.6.

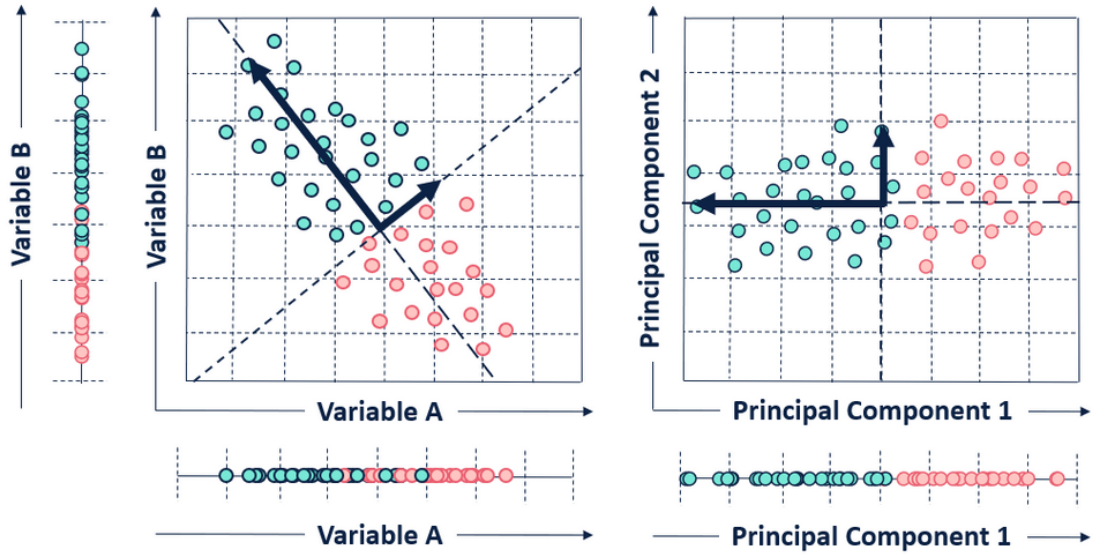

Figure 1.6: Illustration of PCA. Given two variables A and B, the plot on the left shows a scatter plot in its original place while the one on the right shows a PCA bi-plot of the variables. In this simplistic example, a 2D object (with 2 variables) which was not efficiently separated in 1D representation 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 potential to overcome the current donor organ shortage. Translation into clinics requires rigorous control and constant refinement of all processes 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. The two other areas relevant to this project, namely, fundamental knowledge about differentiation and refinement of differentiation protocols, have been vastly improved by the usage of sequencing technologies[13,51–55]. Wu et al[56,57] 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 have been performed by the group. The availability of bulk RNA-Seq data and public reference scRNA-Seq data sets like that of Friedman et al[58] along with accessible deconvolution techniques, like CIBERSORTx, allows within 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 Figure 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()`. 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 can be found here, 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 numbers of samples[59–63].

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[58] who investigated 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*. In this approach 5 distinct time points were sequenced, namely, on days 0 (hiPSC), 2 (germ layer

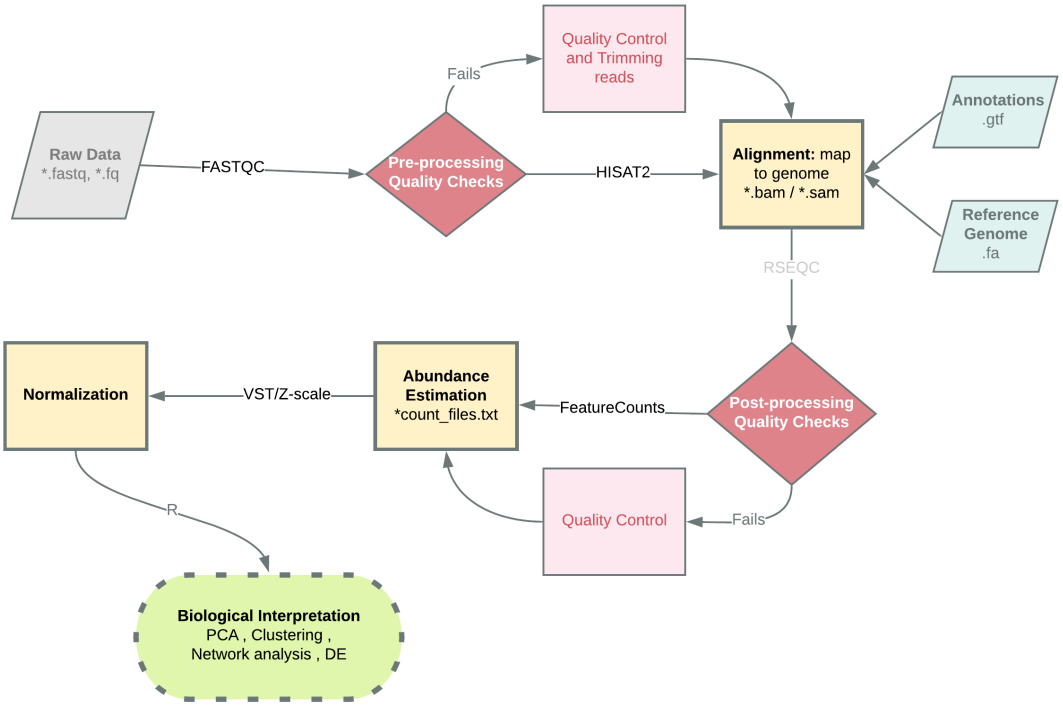

Figure 3.1: Basic analysis pipeline for Bulk RNA-Seq data used in this project. Briefly, raw sequenced data input as fastq files are run through FASTQC for basic quality checks, after which depending on the quality it either goes through additional steps of quality control or directly to an alignment tool like that of HISAT2. An optional post-alignment, quality control check exists, after which abundance of the transcripts is estimated using a tool such as FeatureCounts. This gived the raw read counts file which needs to then be normalized and then used for further analysis.
Shapes and their meanings: Parallelograms (inputs), rhombus (decision points), rectangles (processes), oval (termination).
Abbreviation: VST (variance stabilized transformation), PCA (principal component analysis), .fastq/.fa./fq (raw reads file format), .bam/.sam (binary alignment map, sequence alignment map – file formats for storing aligned sequence data), .gtf (gene transfer format – stores information on genes) .

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

specification), 5 (progenitor 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[47] 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 according to Friedman et al's paper[58]. 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 reliable 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 by Friedman et al.

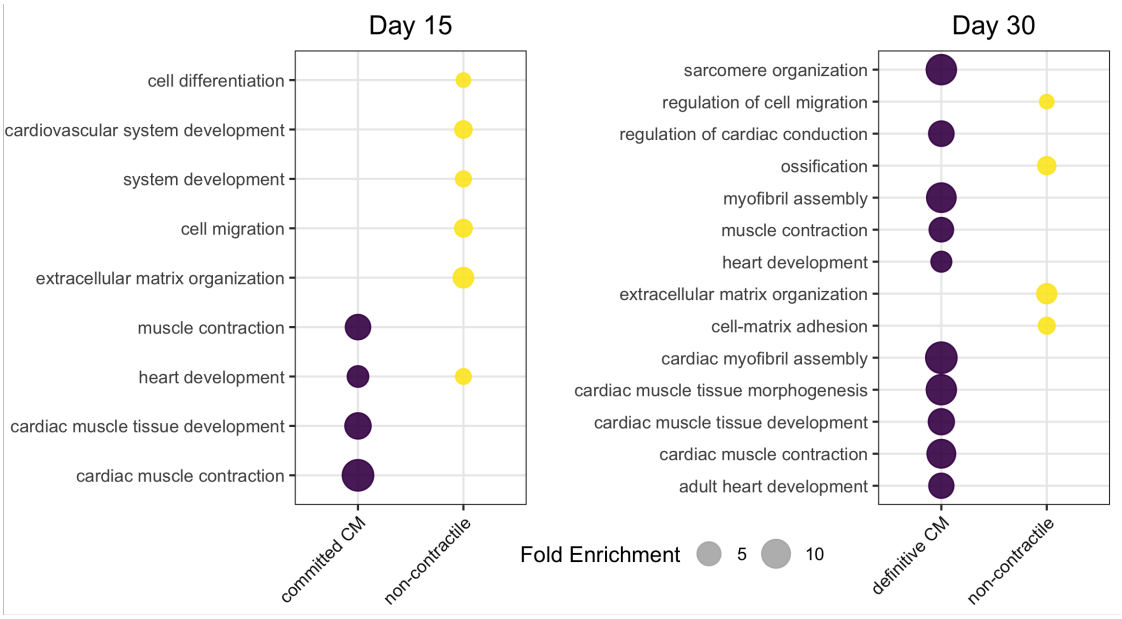

Figure 3.2: scRNA-Seq Reference Dataset. Post-processing and before feeding it into CIBERSORTx, the reference data set was analyzed to ensure its reliable reproduction of the sub-groups as defined by the paper. Here, at both time points there is a sub-group which is enriched for non-contractile features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

CIBERSORTX is an online tool with user-friendly GUI with detailed tutorials on the developer's 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 the `FeatureCounts` tool which requires a valid .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 outputs 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 were 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[64] 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 threshold (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 Figure 4.1.

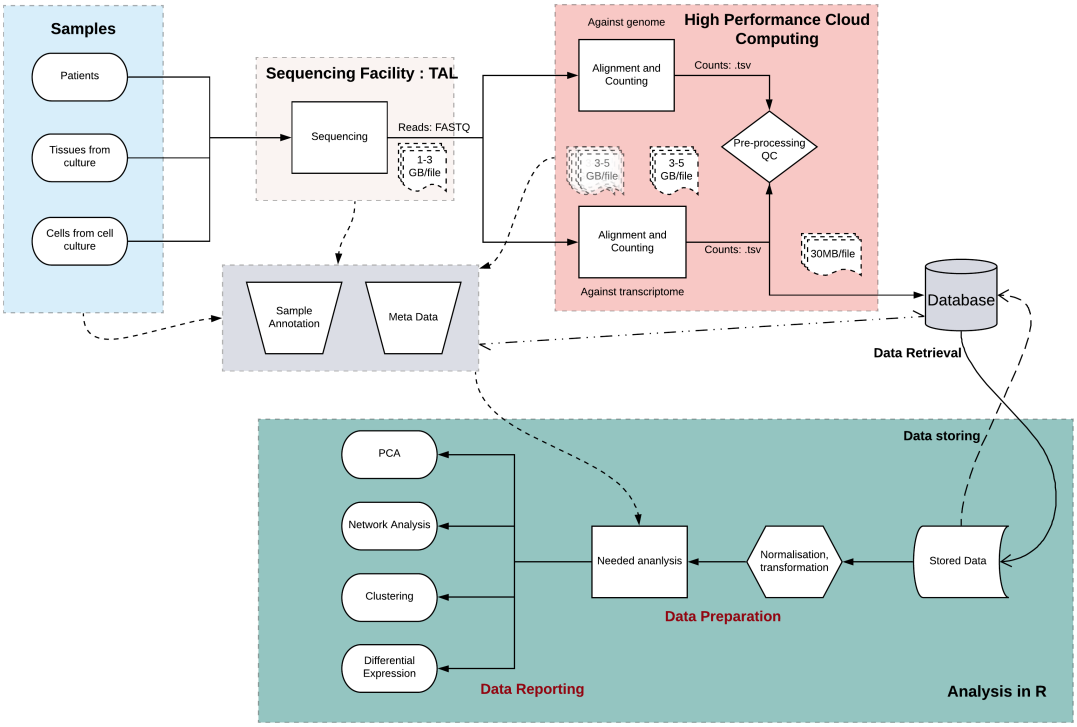

Figure 4.1: A General RNA-Seq Workflow. In general the source material of RNA-Seq is a sample either from patients, cell culture or tissues. After RNA extraction and optional enrichment, they get sequenced at a general sequencing (core) facility. This results in the generation of large files which are transferred to a HPC, where the first steps of alignment and abundance estimation is performed. Smaller count text files containing information on the transcripts and the quanitative amount of expression can then be easily stored in a local database for retrieval and further processing. The final steps of performing the required analysis specific for the question asked can be performed on any statistical environment (such as R) and visualized and deployed for publishing or internal use. Throughout this whole process it is vital to maintain a consistent and common, easy to use metadata collection system. Shapes and their meanings: cylinders (database), rhombus (decision points), rectangles (processes), oval (starting/ending points).

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 observed.

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 Table 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 Section 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 4.2A 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 Figure 4.2B.

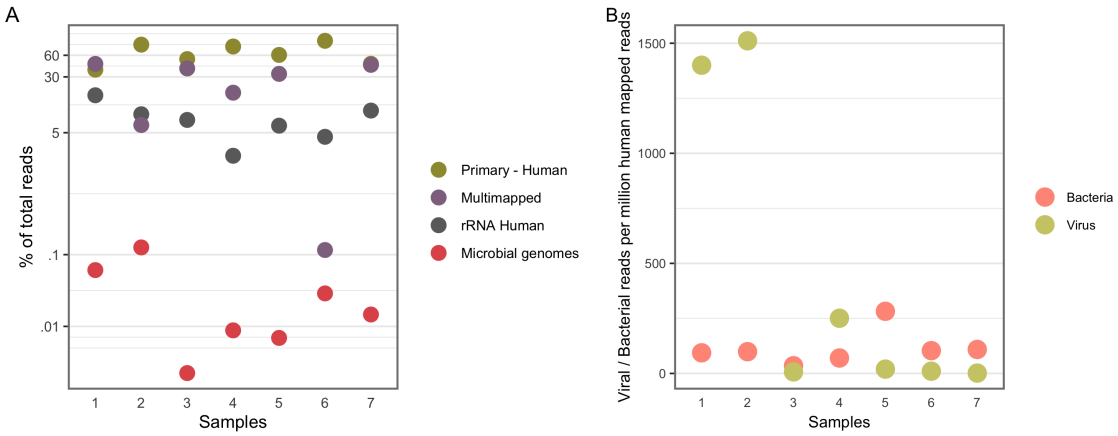

Figure 4.2: General Mapping Statistics. A shows samples and percentage of reads mapped to different groups. Y-aixs is in log scale to resolve low-expression reads. B shows shows separate bacterial and viral reads mapped per million human mapped reads per sample.

Absolute numbers of reads confidently assigned to different viral species across different samples is shown in Figure 4.3A. The same data is shown as relative abundance in Figure 4.3B. Samples 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). Proteus phage is the second entity with high number of absolute reads assigned to it (~7000).

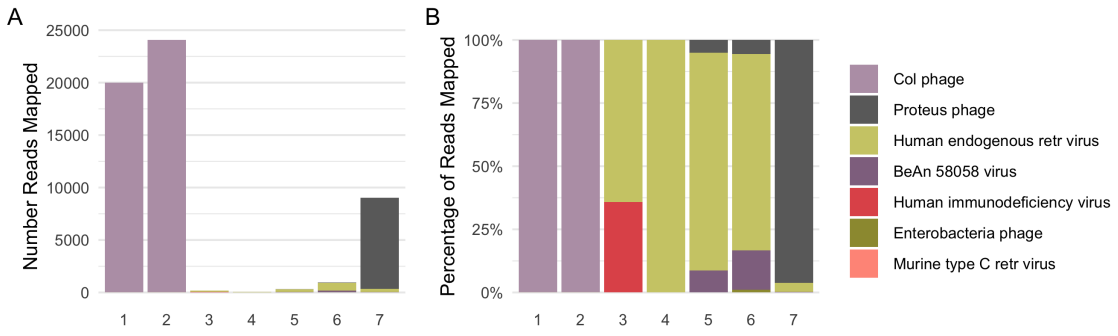

Figure 4.3: Possible viral contaminants. A shows the absolute number of reads confidently assigned to different viral species across different samples, while B represents them as relative abundances.

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[63] and HISEQ-4000[61] platforms while all in-house samples were sequenced on HISEQ-2000[30] which has a separate, dedicated lane for the PhiX spike-in quality control to avoid PhiX reads to appear in the FASTQ files.[i] Unusually high numbers 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 bacterium belonging to the *Enterobacteriaceae family*, which is the most common species responsible for catheter-associated urinary tract infections[65]. 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[66], 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 not well defined[67–69].

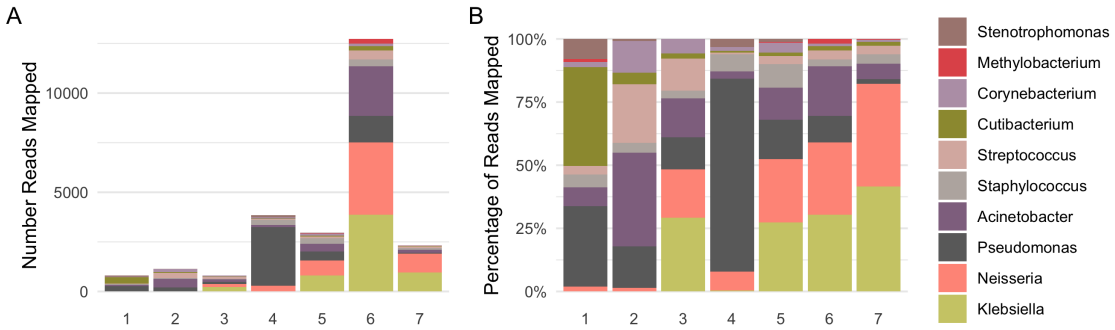

Figure 4.4: Possible bacterial contaminants at the Genus level. A shows the absolute number of reads confidently assigned to different bacterial genera across different samples, accounting for around 80 percent of all the reads assigned to bacteria, while B represents them as percentage proportions.

Unlike viral, the bacterial reads were analyzed at two levels, genus (Figure 4.4) and species (Figure 4.5). 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[70]. Sample 6 has ~13500 reads in

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

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 accounted 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[71–73]. 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[74]. 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[75]. 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.

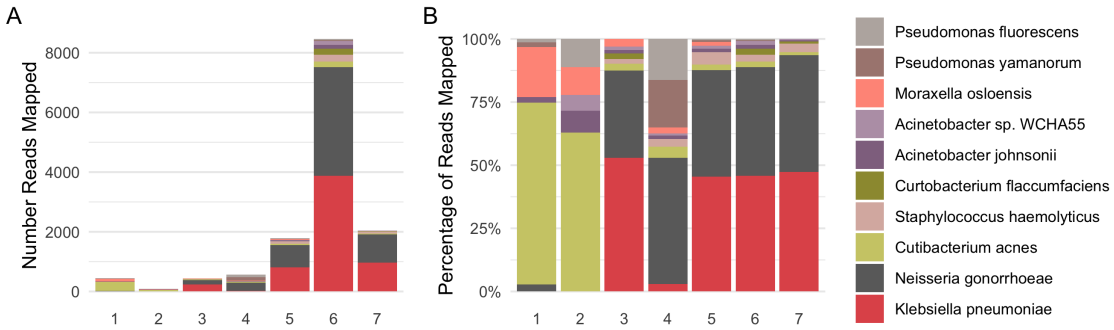

Figure 4.5: Possible bacterial contaminants at the species level. A shows the absolute number of reads confidently assigned to different bacterial species across different samples, accounting for around 50 percent of all the reads assigned to bacteria, while B represents them as percentage proportions.

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 Strong et al[70], showed that *Acinetobacter* contributed to the highest number of reads. While the study by Park et al[76], also picked up a large number of samples having PhiX phage reads in them. All the samples assessed

here have microbial reads less than 0.1% of total reads (including the first two samples if we account for the spike-in PhiX reads). This is considered very low and acceptible by other papers which have published on this topic[70], however, no rule/regulation stipulated by cGMP exists currently which deal with such non-standard exploration of microbial contamination in TEPs.
The samples also differed in their sequencing depths which would have influenced the extent to which metagenomic reads were picked up. This work is neither exhaustive nor confirmatory but since it has the possibility of working on already collected data, it could be a novel complementary and regulatory step that might one day be incorporated into a cGMP practice alongside other standard microbial detection techniques.

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 Table 3.1. To examine global trends in gene expression levels, the normalized data across all samples were visualized using PCA (Figure 4.7). 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 as a scree plot (Figure 4.6A). 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, 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 4.6B).

The second largest source of variation is captured by PC2 accounting for 14.3% of the total variance and separates the fetal heart samples from the rest of the sample types as shown by the ordering of sample types according to PC2, (see Figure 4.6C).

Cardiomyocyte samples which are >90% actinin+ in FACS are in-essence representative of a single cell type, while EHMs have the additional stromal cells. Apart from these two, fetal heart tissues also contain a variety of other cell types and sub-types including differentiated and differentiating cells along different lineages. The adult heart on the other hand 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 might be because the fetal samples were collected across different

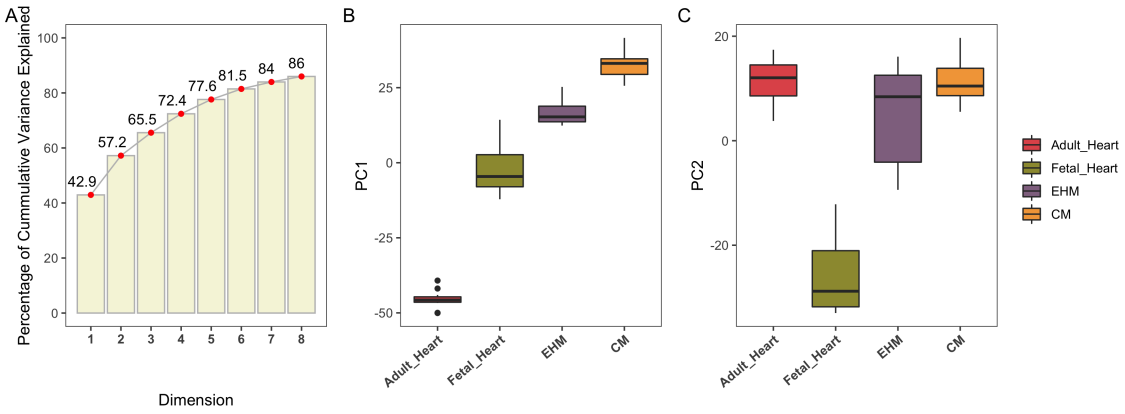

Figure 4.6: A shows the cummulative variance explained by the PCs. B and C show the different groups plotted against PC1 and PC2 respectively.

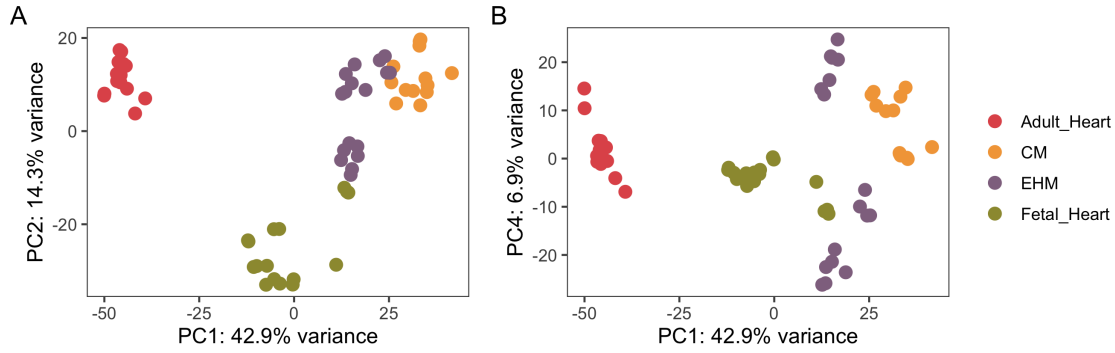

Figure 4.7: A and B show the PCA plots of all the samples againt PC1 and PC2/PC4. The samples are coloured based on their groups.

gestational periods, in the range of 9 weeks - 16 weeks (Figure 4.7A).

The EHM samples also show two semi-distinct clusters, corresponding to the two EHM sources — in-house and from the PRJNA362579 project. The source of this variation within the EHM samples appears to be strongly associated with PC4, see Figures 4.8 and 4.7B.

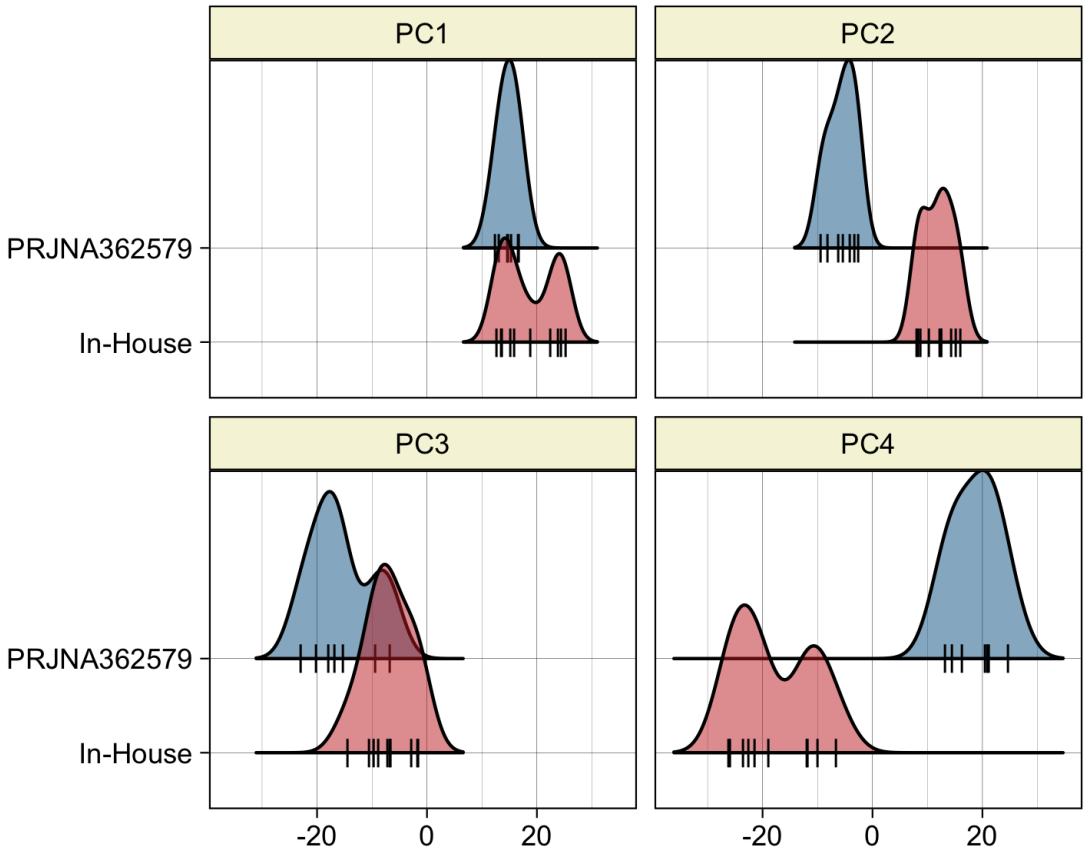

Figure 4.8: Separation of EHMs by the first 4 PCs. The two groups of EHMs and the varying degress of separation by different PCs, visualized by a density plot.

As explained in Section 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 Table 4.3.

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 in adult hearts biopsies like AQP7, possibly from the adipose tissue surrounding

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
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

the heart[77] and APOD from aortic valves of the adult heart[78]. These cell types would be atypical for bioengineered tissues and cultured cell and less common in the fetal heart samples. The current analysis pipeline, does not allow to quantify the relative closeness of the EHM groups to either fetal or adult phenotypes. Other biological and functional characterizations of the EHMs are necessary to complement the deconvolution approach.

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 Figure 4.9 A.2. 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). By choosing ~380 genes which are specifically detectable and enriched in the human heart, based on Human Protein Atlas data, the pearson correlation of the EHM and fetal samples becomes very comparable as shown in Figure 4.9 B.2. The median correlation of the EHM samples is 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[79]. Gene expression across different sample types as per the curated gene list appears to be more consistent and comparable, as visualized in Figure 4.9 B.1 in comparison to Figure 4.9 A.1. The genes in the heatmaps are hierarchically clustered.

These results show that the separation in PC1 of all the different groups are possibly not due to differences in the expression of genes pertaining to CMs, as they all have a higher correlation when a subset of heart-specific genes was used. And this high correlation was lost when the top 2000 genes with the highest loading in PC1 was used, indicating that the highest variance amongst the groups is possibly not due to biological differences in their CMs expression profile.

4.3.2 Gene-level analysis

Despite crucial insights derived 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

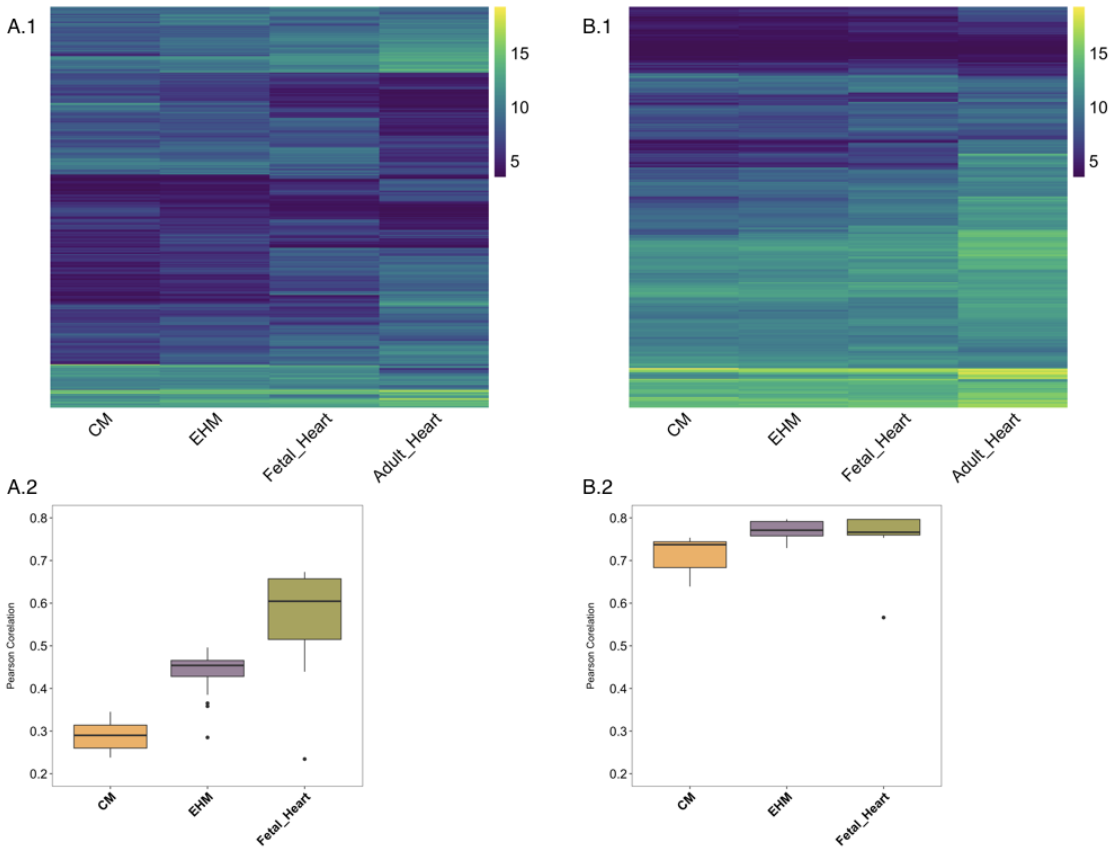

Figure 4.9: Correlation of samples. Heatmaps show VST normalized and group-wise clubbed data. A.1 shows the top 2000 genes with highest loading in PC1 and PC2. B.1 shows a heatmap drawn from a selection of 380 genes based on Human Protein Atlas. In both the heatmaps each row represents a gene and each column represents the average of the sample group. 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.

groups of genes, covering an example of metabolic, structural and general gene lists as shown in Figure 4.10. Panel A is a selected representation of the 10 highest and most specifically expressed genes based on information from 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[80], which is almost equally expressed in all samples. Panel B consists of 22 genes associated with 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[81,82]. 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[83], *Myl4* and *Myl7* are known to be expressed heavily by the developing heart[84–86], and *Tnni3* which is known to be expressed under specific developmental conditions[87]. 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 was of little help in the elucidation of the possible sub-populations within each sample. This was subsequently 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

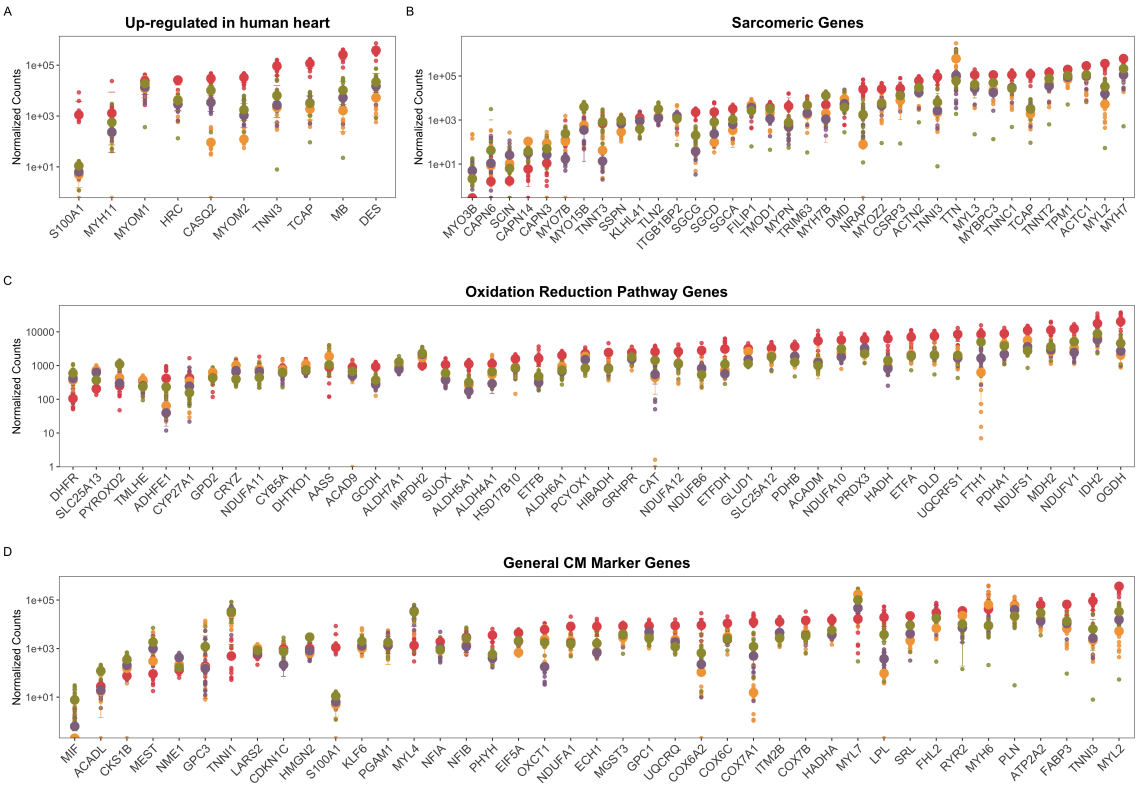

Figure 4.10: At individual gene level, the EHMs resemble the fetal heart expression levels. The graph shows normalized counts of different panels of gene sets. The bigger circles show the median of each group while the slightly transparent and smaller circles represent each sample for every gene. A shows 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

reference sample[58]. Based on 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 a *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)), as tabulated in B of 4.11. Methodologically we followed a state of the art protocol, as described in Section 3.2, where d15 and d30 reference data were processed and used in CIBERSORTx to generate a signature-matrix, which was then used to deconvolve the bulk samples, whose results are shown as a percentage proportion bar graph in Figure 4.11A.

CMs samples are considered highly (>90% cardiomyocytes) pure 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 EHMs, supporting the notion that CMs continue maturation when being exposed to the 3D EHM environment[88]. 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 or cardiomyocytes-like cells. This clearly validates the current differentiation protocol. The difference in composition of the EHM samples from that of CMs is also captured efficiently as seen by the increase in proportion of S2 sub-groups (the *non-contractile cells*) in the EHM samples.

As explained in Section 1.6, our 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 production of EHMs, it is vital that the batch-to-batch consistency is maintained. Currently four batches of CMs have been analyzed in Figure 4.12 post bulk-sequencing and subsequent deconvolution. The variation across samples within proportions sub-groups is not signficant (Figure 4.12A). Each sample and its corresponding deconvoluted putative sub-group percentages are shown in Figure 4.12B. It is evident that the 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

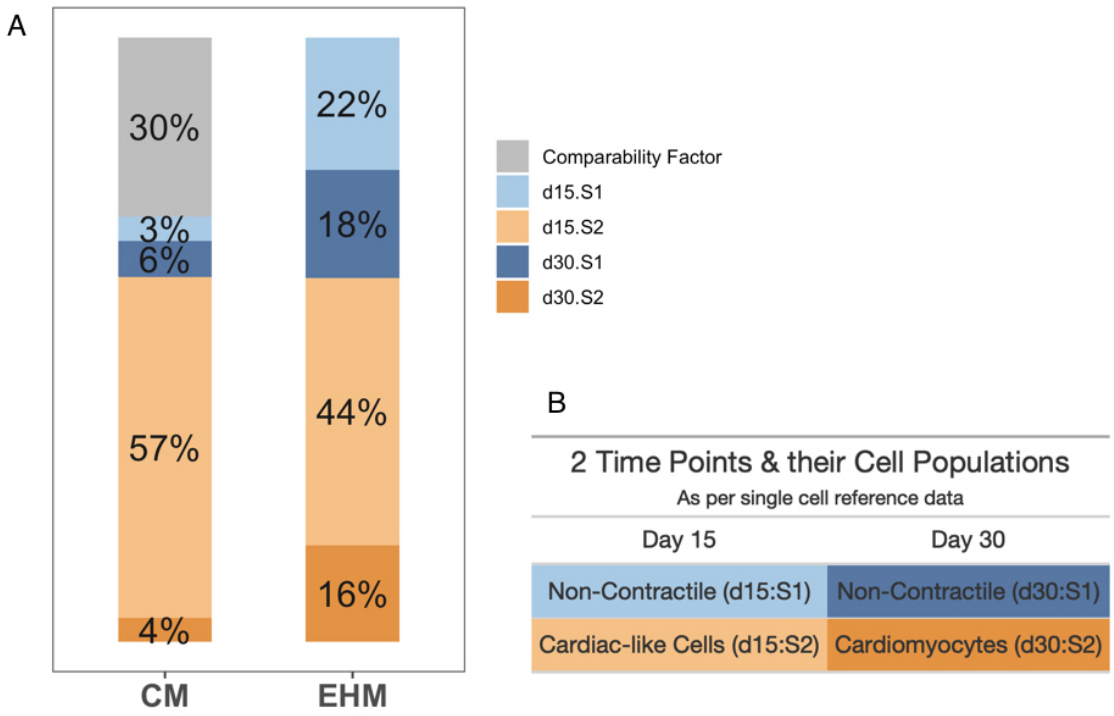

Figure 4.11: EHMs have a higher proportion of mature cardiomyocytes in comparison to CMs. A shows percentages of the various groups as per deconvolution, that are explained in B, along with a comparability factor added to the CM samples to allow for direct comparison of both CMs and EHMs.

and structural experiments of the group. Across every differentiation run, the efficiency of differentiation is assessed using flow cytometric markers, and percentage of cells positive for Actinin 2 marker indicate the approximate amount of cells differentiated into cardiomyocytes. In particular, one among the four samples was observed to have low percentages of Actinin positive cells (~40%), which was suspected to be an error in handling of the flow cytometer than actual lack of purity. The results from deconvolution show that indeed all the samples have a large proportion of cardiomyocytes (mean ~87%), and the single, potential outlier from flow cytometry data was possibly erroneous.

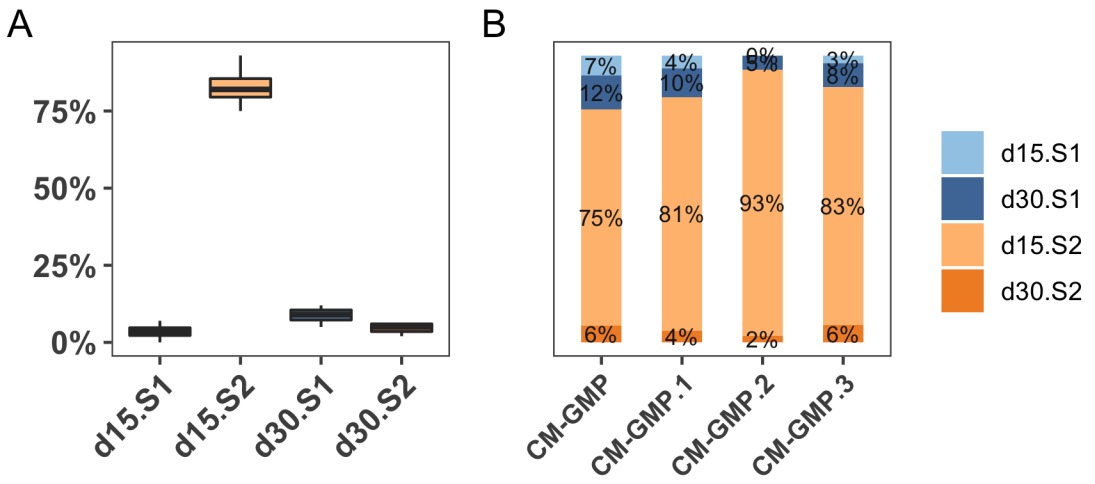

Figure 4.12: The GMP-CM samples are mostly consistent wrt subpopulations derived from deconvolution. A shows the box plots of the contribution of four sub groups across the four samples. B is a sample-wise representation of the sub-groups.

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 Figure 4.13. It is seen that the EHM samples from project PRJNA362579 have a higher average percentage of *mature-cardiomyocytes* (d30:S2 ~ 23%) and more specifically, 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). Those "mature" samples also have a higher combined percentage of d15:S1 and d30:S2 (i.e., all the cells committed to the cardiac lineage). 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. Yet, across the same cell-line or production batch, for instance the EHMs made from HES2 cell line, the consistency in putative sub populations are maintained.

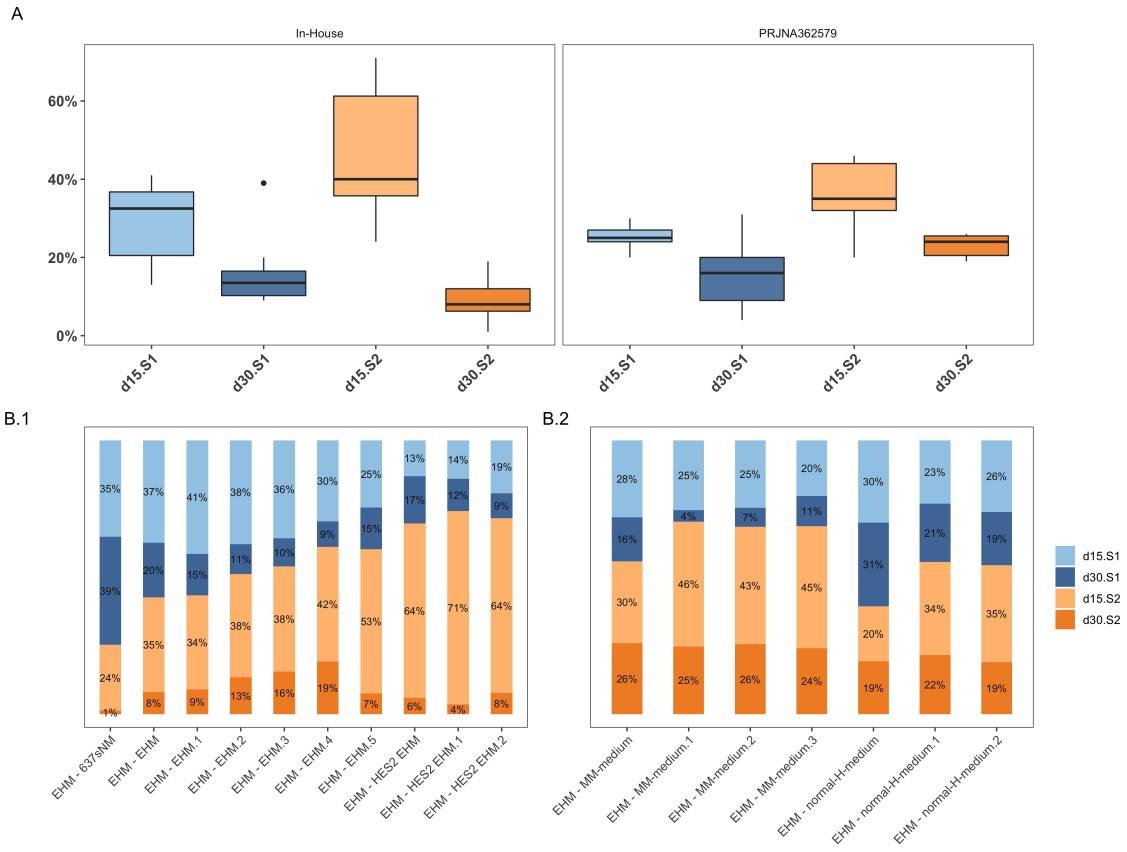

Figure 4.13: Differences in subpopulations amongst EHM samples. A shows the box-plots of percentage contributions of the various groups as per the source of the sample — in-house or from the PRJNA362579 project. B shows the corresponding sample-wise breakdown of the subpopulations in bulk data.

4.4.1 Limits of deconvolution

To address the reliability of the deconvolution technique, all sample groups and all samples within all groups were included (iPSCs, fibroblasts, CMs, EHMs, adult and fetal heart) and were deconvolved using the same reference dataset, as shown in Figure 4.15. Initially the validity of the deconvolution was evaluated by creating three different *pseudo bulk* samples, with known proportions of the sub-groups from the single cell data and these *bulk* samples were then deconvoluted, see Figure 4.14. Here

we see that the deconvolution recaptiulates the true proportions, with an average RMSE of 0.1.

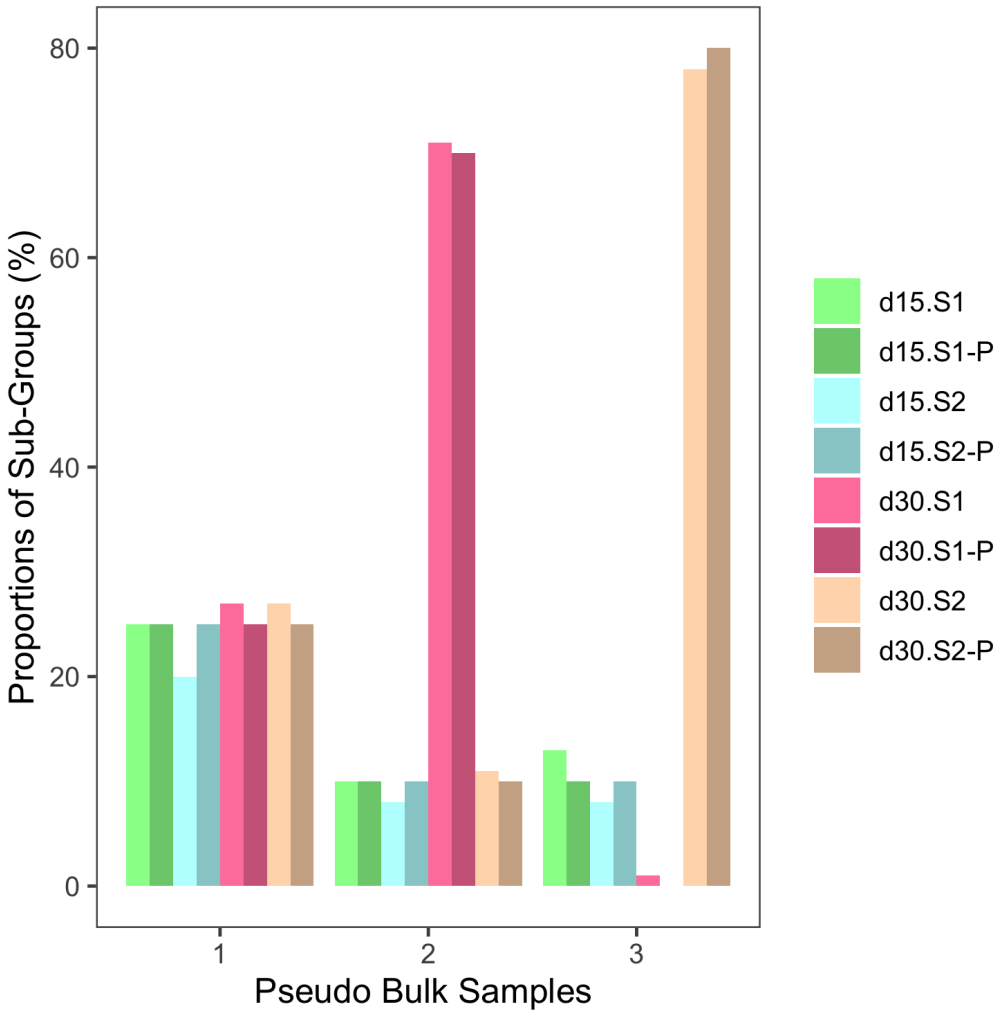

Figure 4.14: Estimating the validity of deconvolution. Three pseudobulk samples were created from the single cell reference data, the darker shade of the colours represent the known, true proportions of the four subgroups (denoted with a P). These samples were then deconvolved using CIBERSORTx and the resultant deconvolved estimated proportions are represented by the lighter shade bars adjacent to the true proportions.

On the other hand, when we use samples irrelevant or different from the reference data set, we can see that all the groups were *deconvolved*, yet, the reliability of these results varied widely. For instance, the root mean squared error (RMSE)[ii] of iPSCs and Fibroblasts are close to 1.0 while that of fetal heart group is around 0.3 (Figure 4.15B) . The correlation of the deconvolved results to the actual reference sample

[ii]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.

also had a large range, 0.19 for iPSC and 0.84 for the fetal heart. This shows the biggest limitation of such deconvolution techniques, and probably stands testimonial to the "garbage in, garbage out" situation, is the fact that deconvolution would work in any setting, however the usability of it depends on the relevance and similarity in composition of the cell types of the reference and the bulk data that needs to be deconvolved.

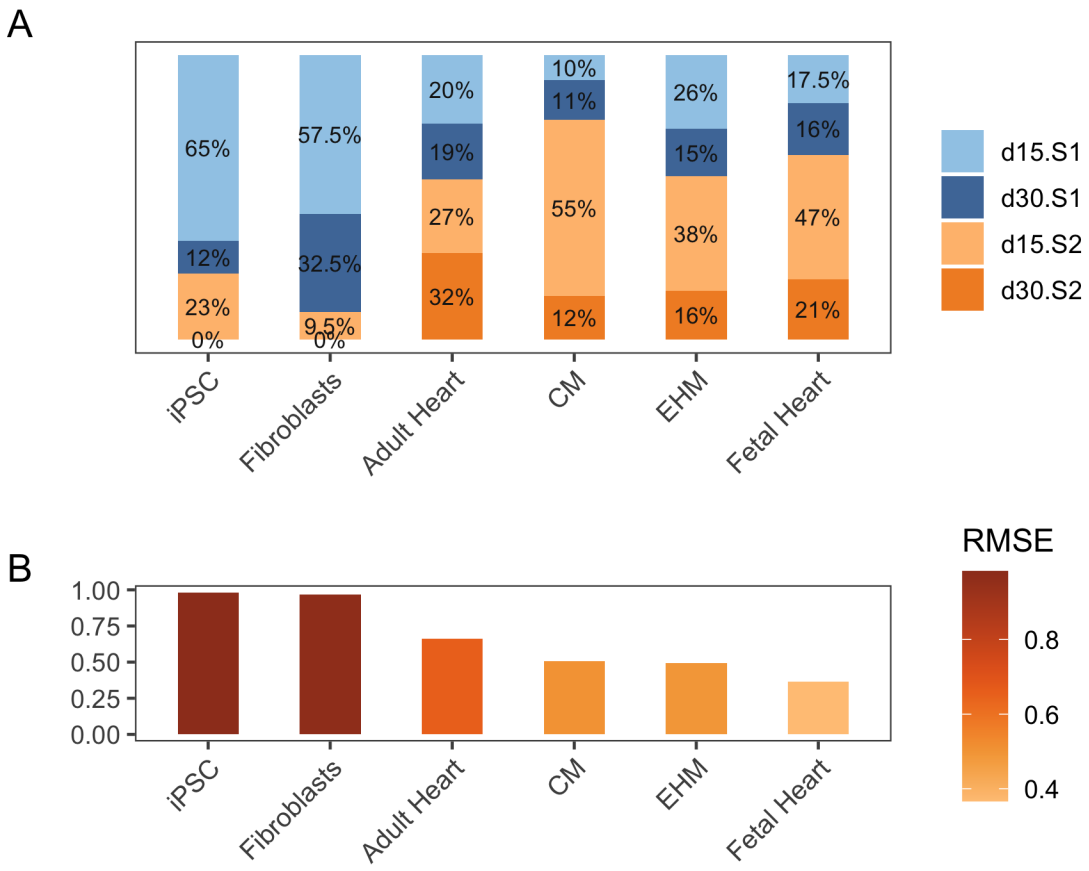

Figure 4.15: A shows the average deconvolution percentages of all groups to the four different subtypes. B shows the median RMSE of each group. Abbreviation: RMSE (root mean square error).

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 (Figure 4.16). 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.

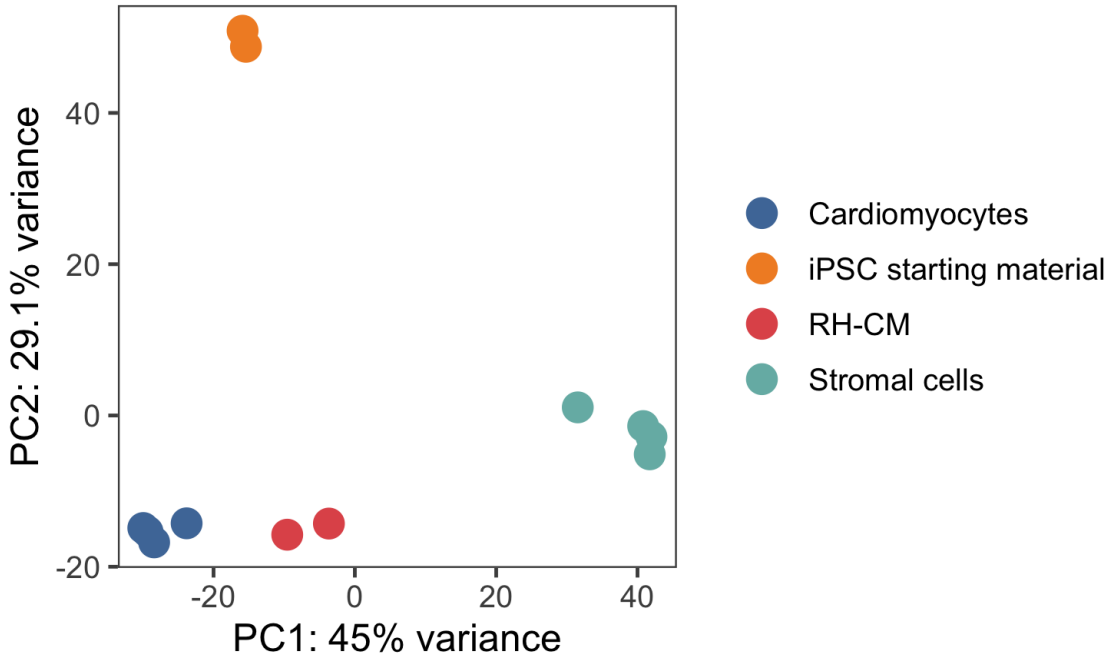

Figure 4.16: PCA of different samples, iPSCs, fibroblasts, human CMs and rh-CMs show that the rhesus CMs cluster with the human CMs.

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, validated the current protocol used in the production of CMs and EHMs by the group by showing that CM samples are mostly (~91%) composed of either committed cardiac cells or cardiomycytes, while EHMs possess about ~35% of non-myocyte population. Also, within this, the CMs in the pure CM samples are less mature than their counterparts found in EHMs. This is in-line with the view that CMs tend to mature in a 3D environment like that of an EHM. These results were based on an input with known and defined cell types and shows a supervised learning methodology. Using PCA, an unsupervised technique, we showed that the majority of the variation across different groups containing CMs — namely, adult heart, fetal heart, EHM and CM samples, as captured by PC1 does not explain the biological differences in the their CM populations, and the differences are possibly due to tissue complexity. We also showed that using this deconvolution technique and an easy workflow, could possibly be used to follow the efficiency and consistency of differentiation across different runs and batches of production. A key reason for the possibility of using this consistently is due to the cheaper costs of bulk sequencing compared with single cell sequencing. Here, using computational deconvolution it is possible to obtain the benefits of single cell sequencing (i.e., knowing the subtypes of populations within a sample in our use case) using just bulk sequencing, saving on costs and time. This could also complement the standard FACS based methods of tracking differentiation.

However, 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. Thus, new and unknown cell-types can not be characterised or assessed using this technique, limiting its use in exploration or discovery. 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 exactly the 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.

Evaluation of microbial contamination of engineered tissue in the context of clinical use is extensive and thorough, usually by microbiological methods. A part of the thesis also explored the presence of potential microbial contaminants using sequenced data. Here the results were in-line with most others' findings in this area and demonstrated the non-standard use of RNA-Seq data in microbial detection. Incidently, spike-ins were detected and confirmed as a testament to the method employed. However, with this generality of usage comes the limitation of specificity. It possibly can not be a standalone way of microbial estimation/detection, and could possibly hint towards extreme cases which warrant deeper, more focused exploration.

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. Although measures were applied to avoid potential batch effects (samples that were sequenced across different time points, different instruments, different depths), it still remains a 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.

Summary

Task At Hand

Targeted differentiation of hypoimmunogenic iPSCs into functional cell types and subsequent assembly into artificial tissues for organ repair and replacement holds great potential to overcome the current donor organ shortage. The main aim of the project was to find putative sub-populations within samples of iPSC-induced cardiomyocytes and EHMs. In transcriptomics, this is addressed using single cell RNA sequencing, yet, at the time of the start of thesis there was data only from bulk RNA sequencing. Here a *Computational Deconvolution* approach seemed capable of providing sub-population level information from bulk sequencing using a relevant single cell dataset.

Parts of thesis also focused explored the possibilty of using RNA sequencing data to identify potential microbial contaminants, as Prof. Zimmermann's group and their work is transitioning into the clinics and as such, would need to meet the highest of standards of *current Good Manufacturing Practices*, a part of which deals with minimising, and checking for microbial contamination.

Work Done

- An analysis pipeline was set up to analyse the bulk RNA sequencing data from raw sequencing files to count files, from which further analysis can be easily performed.
- Possible microbial contamination was investigated using the alignment of unmapped (to human genome) reads to bacterial and viral genomes.
- An exploratory data analysis was done and global trends were established which reiterated the group's previous findings. For example, using PCA across multiple datasets we confirmed that the cardiomyocytes within the adult heart, fetal heart, EHMs and iPSC-cardiomyocytes were globally similar and the majority of the differences between these groups are possibily based on their tissue complexity.
- Putative proportions of sub-populations within the iPSC-induced cardiomyocytes and EHMs was established using computatinal deconvolution. The findings corroborated with the current differentiation and production protocols.

References

1. Max Roser, E. O.-O. & Ritchie, H. Life expectancy. *Our World in Data* (2020).

2. Ritchie, H. & Roser, M. Causes of death. *Our World in Data* (2020).

3. 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).

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

5. 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).

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

7. 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).

8. Eurotransplant - Statistics.

9. Metra, M. & Teerlink, J. R. Heart failure. *The Lancet* **390**, 1981–1995 (2017).

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

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

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

13. 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).

14. 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).

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

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

17. 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).

18. 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).

19. 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).

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

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

22. 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).

23. 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).

24. 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).

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

26. 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).

27. Nakatsuji, N., Nakajima, F. & Tokunaga, K. HLA-haplotype banking and iPS

cells. *Nature Biotechnology* **26**, 739–740 (2008).

28. 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).

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

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

31. 2006, E. EMEA/chmp 2006. Guideline on human cell-based medicinal products. EMEA/chmp/410869/2006. (2006).

32. CD-P-TO. Guideline to the quality and safety of tissues and cells for human applications. *European Committee on Organ Transplantation,EDQM* (2017).

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

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

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

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

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

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

39. 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).

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

41. Becht, E. *et al.* Estimating the population abundance of tissue-infiltrating immune

and stromal cell populations using gene expression. *Genome Biology* **17**, 218 (2016).

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

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

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

45. 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.

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

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

48. 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).

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

50. Witteveen, E. *et al.* Increased Early Systemic Inflammation in ICU-Acquired Weakness; A Prospective Observational Cohort Study*. *Critical Care Medicine* **45**, 972–979 (2017).

51. 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).

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

53. 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.

54. 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).

55. 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).

56. 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).

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

58. 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).

59. 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).

60. 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).

61. 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).

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

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

64. 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).

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

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

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

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

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

70. Strong, M. J. *et al.* Microbial Contamination in Next Generation Sequencing: Implications for Sequence-Based Analysis of Clinical Samples. *PLoS Pathogens* **10**, (2014).

71. 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).

72. 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).

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

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

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

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

77. 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).

78. 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).

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

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

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

82. 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).

83. 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).

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

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

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

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

88. Machiraju, P. & Greenway, S. C. Current methods for the maturation of induced pluripotent stem cell-derived cardiomyocytes. *World Journal of Stem Cells* **11**, 33–43 (2019).