Spring 5-26-2021

Spaceflight and the Differential Gene Expression of Human Stem Cell-Derived Cardiomyocytes

Eugenie Zhu

Spaceflight and the Differential Gene Expression of Human Stem Cell-Derived Cardiomyocytes

A Project

Presented to

Department of Computer Science

San José State University

In Partial Fulfillment

Of the Requirements for the Degree

By

Eugenie Zhu

May 2021

ABSTRACT

The National Aeronautics and Space Administration (NASA) has performed many experiments on the International Space Station (ISS) to further understand how conditions in space can affect life on Earth. This project analyzed GLDS-258, a gene set from NASA's GeneLab repository which examines the impact of microgravity on human induced pluripotent stem-cell-derived cardiomyocytes (hiPSC-CMs). While many datasets have been run through NASA's RNA-Seq Consensus Pipeline (RCP) to study differential gene expression in space, a *Homo sapiens* dataset has yet to be analyzed using the RCP. The aim of this project was to run the first *Homo sapiens* dataset, GLDS-258, through the RCP on the San Jose State University College of Engineering High Performance Computing Cluster and investigate any biological significance from the results. In this study, a total of 18 hiPSC-CMs samples from ground control, flight, and post-flight groups are run through the RPC. The resulting differential gene expression data was further analyzed for biological significance using the Database for Annotation, Visualization, and Integrated Discovery (DAVID) and Gene Set Enrichment Analysis (GSEA). Results showed that most genes were differentially expressed in ground control versus flight groups, while post-flight groups and ground control groups did not have as many differentially expressed genes. Gene set analysis showed significant expression of genes in mitochondrial pathways as well as genes related to neurodegenerative diseases such as Alzheimer's, Huntington's, and Parkinson's disease. These results indicate that exposure to microgravity may play a role in altering expression of genes which are related to neurodegenerative pathways in cardiac cells. Our results demonstrate that it is possible to process *Homo sapiens* data through the RPC, and suggest that cardiomyocytes exposed to microgravity may exacerbate neurodegenerative disease progression.

Keywords: microgravity, space flight, cardiomyocytes, bioinformatics

ACKNOWLEDGEMENTS

I would like to thank the following group of people for their guidance in completing my project.

- Dr. Philip Heller for showing me how fascinating and complex bioinformatics can be, and for guiding me through each step of my project.
- Dr. Leonard Wesley and Dr. Thomas Austin, for their feedback and insight that helped me complete my project.
- Dr. Amanda Saravia-Butler from NASA GeneLab for going above and beyond by spending an entire day to show me how to properly process *Homo Sapiens* data through the RNA-Seq Pipeline.
- My peers for their support, patience, and insight when troubleshooting my project.
- My friends and family for their endless support throughout the years, and especially during the program.

Table of Contents

Table of Figures

Index of Tables

I. INTRODUCTION

A. The Human Body and Space

Space travel has long been one of humanity's greatest curiosities and persisting interests. As modern-day technologies advance, space exploration continues to become more accessible and space missions can increase in number. NASA's Perseverance Rover's successful landing on Mars on February 18, 2021 is a prime example of what is possible in the field of space exploration in years to come [1]. This is a monumental achievement for mankind and will undoubtedly pave the way for future missions, including bringing astronauts back to the Moon and to Mars [3], [4]. NASA plans to bring astronauts, including the first woman, to the Moon by the year 2024 on the Artemis Lunar Exploration Program [4]. The upcoming mission to the Moon will demonstrate new space travel technologies and possibilities, while also capturing the imagination of the younger generation and people around the world [4].

However, space travel can be accompanied by adverse impacts on the human body since environments in outer space are very different from the conditions in which life on Earth evolved. The main concerns that studies have been focused on include space radiation, isolation and confinement, distance from Earth, gravity fields, and hostile or closed environments [2]. Among these factors, exposure to radiation and microgravity have been studied extensively in a scientific setting. Exposure to space radiation can be a long-term issue for astronauts because it can harm cell DNA repair mechanisms and possibly lead to development of cancer or other detrimental illnesses [3]. Similarly, astronauts who were exposed to microgravity for extended periods of time exhibited changes to several systems in their body, including musculo-skeletal, cardiovascular, nervous, and immune systems [4]. While these have been many counter measures implemented to alleviate some of the effects of these hazards, more research is required to understand how to best protect our astronauts' health and safety.

B. Cardiomyocytes: Cardiac Cell Muscle

Cardiomyocytes are the muscular cells within the heart that allow the organ to contract and relax properly to provide blood throughout the body [5]. The contracting motions are regulated by ion channels that manage the entry and exit of Ca2+ into the cardiac cells [5].

Cardiomyocytes are difficult to study in culture for two major reasons. First, cardiomyocytes are not likely to divide after birth and therefore are also not likely to replicate in culture [5]. Second, to isolate the cardiomyocytes, the intercalated disks may become damaged that lead to increased intake of Ca2+ [7]. As a result, cell death will occur due to hypercontraction [7]. Figure 1 below depicts cardiomyocyte structure.

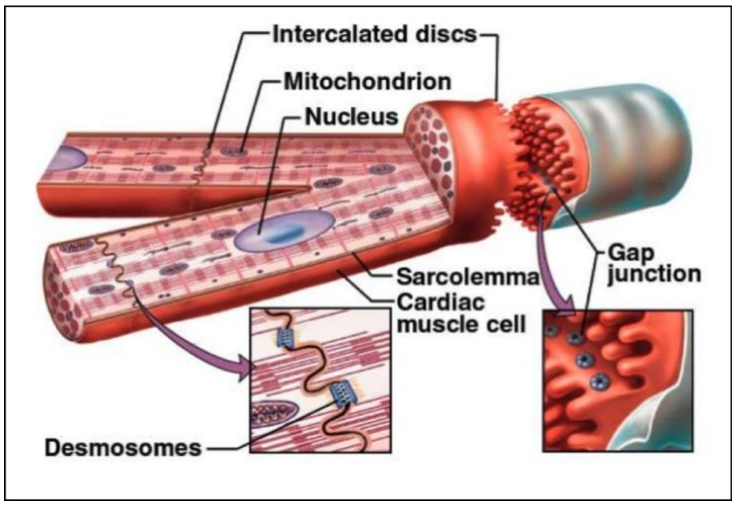

Figure 1: Graphical representation of cardiomyocyte structure [6]

C. Cardiomyocytes and Space Flight

NASA has performed an abundance of studies throughout the years regarding the cardiovascular system and human health. There has been substantial evidence to conclude that spaceflight will affect human cardiac function. One of the most famous studies, the NASA Twin Study, showed that arterial pressure will decrease while total cardiac output increases when an astronaut is exposed to a long-term microgravity environment [7]. In this study, identical twins Mark Kelly and Scott Kelly were observed after Scott lived in space for a year and Mark remained on Earth as a ground control. Other studies have confirmed similar findings, in which individuals have lowered arterial pressure and decreased heart rate [8]. While there have been several conclusive studies regarding the human cardiovascular system as a whole, studies of cardiac function at the cellular level have been limited due to the difficulty of studying cardiomyocytes in culture.

Animal models have historically been a popular choice when studying the relationship between microgravity and cardiomyocytes at the cellular level since human trials are complicated and costly. In particular, mice and rats have been used for several of these experiments. A study published in 1992 found that rat cardiac myosin, a protein that plays an important role in muscle contractions, had mRNA expression that was impacted when the rats were exposed to microgravity for 14 days [9]. A separate experiment concluded that short exposure to microgravity increases the expression of various rat cardiomyocyte mitochondrial enzymes [10]. These are two among many other animal model-based experiments performed in the last several decades. While these past experiments have provided valuable insight into the relationship between microgravity and mammalian cardiomyocytes, animal models cannot provide a completely exhaustive and conclusive results for the human body. NASA continues to seek more informative results by utilizing new technologies that allow us to perform experiments on human cardiomyocytes.

D. NASA GeneLab's RNA-Seq Consensus Pipeline

NASA's GeneLab project aims to create the GeneLab Data System (GLDS) repository, which hosts molecular data from spaceflight mission samples [11]. Additionally, GeneLab hopes to process these datasets to provide more knowledge about how life on Earth changes in space [12]. The current version of the repository, GLDS 3.0, contains over 200 datasets from around the world [12]. As more datasets are uploaded to GLDS, NASA has developed methods to visualize and analyze the data on their repository.

GeneLab's RNA-Sequence Consensus Pipeline (RCP) is one method NASA has developed to process datasets upload to GLDS. The pipeline is designed to analyze short-read RNA-sequence data, and includes several steps that will ultimately detect any differentially expressed genes [12]. A variety of datasets can be run through the pipeline, including *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, among others [13]. The three major steps of the pipeline are data pre-processing: quality control and trimming, data processing: read mapping and sample quantification, and finally differential gene expression calculations and gene annotations. GLDS contains more than 80 RNA sequencing datasets as of August 2020, and many of these sets have yet to be run through the pipeline. The process can take several hours and typically requires the user to manually run various scripts throughout the process [13]. A more streamlined, Nextflow implementation of the pipeline has been created by former San Jose State University student Jonathan Oribello; it requires little to no manual intervention throughout the process [13]. Since the Nextflow implementation of the pipeline was originally tailored to run *Mus musculus* datasets, this paper briefly examines how to run a *Homo sapiens* dataset through the Nextflow version of the pipeline. However, analysis will be focused around the output from manually running customized scripts from the pipeline since the Nextflow implementation for *Homo Sapiens* is still in development.

E. GeneLab Dataset 258

This paper focuses on the dataset GLDS-258, which contains RNA sequence data from human induced pluripotent stem cell-derived cardiomyocytes (hiPSC-CMs) provided by Joe Wu's group from Stanford University School of Medicine [14]. There are three distinct cell lines in this study: four and a half weeks of microgravity exposure, five and a half weeks of microgravity exposure with ten days of Earth gravity, and a control with five and a half weeks and 10 days of Earth gravity [15]. Each cell line had six samples for a total of 18 samples and were all generated from three individuals and differentiated into hiPSC-CMs using the 2D monolayer differentiation protocol [15]. Cell lines that were exposed to microgravity were sent to the ISS on a SpaceX Falcon 9 rocket during a resupply mission [15]. Figure 2 below depicts an overview of the process Wu's research group used to obtain the data, along with types of analysis used after obtaining the data.

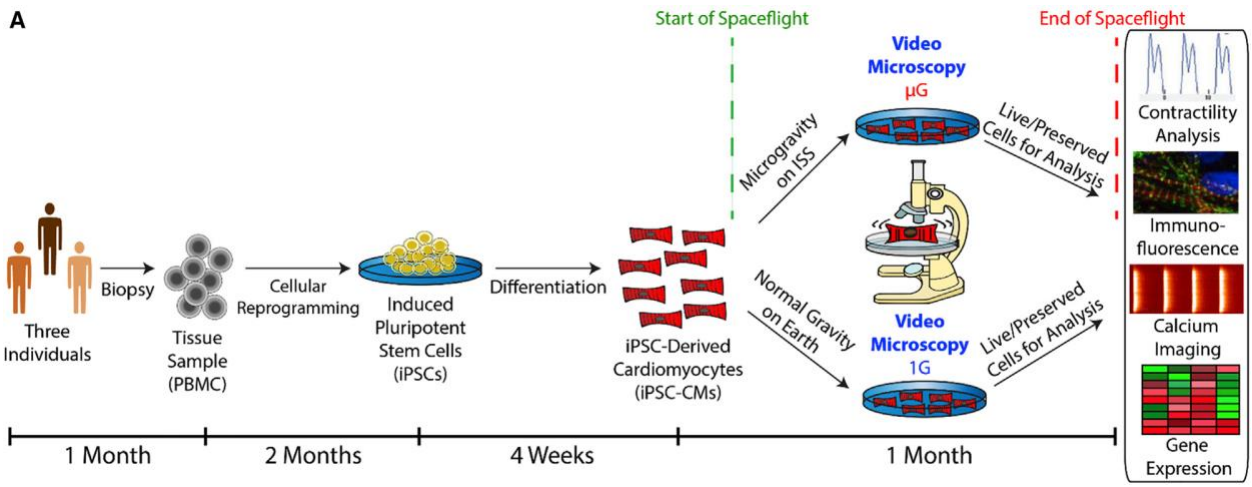

Figure 2: Graphical representation of the timeline for the entire experiment, Reprinted with permission obtained from Joe Wu from the Stanford University School of Medicine [15].

The raw reads gathered from these cell lines were analyzed using a different protocol from the RNA-Seq Consensus Pipeline [15]. Upon completion of the experiment, Wu's research group completed nucleic acid sequencing using Illumina HiSeq to create paired end reads with a 100 base pair length [15]. Their group aligned the raw reads to the hg38 human genome using HISAT2 [15]. The researchers then utilized featureCounts along with the Ensembl 85 annotation file to quantify the raw reads [15]. Finally, the genes were normalized and determined to be differentially expressed using the DESeq2 tool [15]. The results from this method of data analysis found that 2,635 genes were differently expressed based on a p-value of 0.05 [15].

While these methods work well, they are different from the protocol in NASA's RCP. The analysis in this paper of GLDS-258 will follow NASA's RCP and use more recent human reference genomes and annotations, since the hg38 human genome and Ensembl 85 references are from 2013 and 2016, respectively [15], [16]. This dataset will be the first *Homo Sapiens* dataset to be run through the RCP, and hopefully provide the groundwork for future *Homo Sapiens* datasets as well.

II. METHODS

A. Computational System

Analysis of the GLDS-258 dataset requires a large number of computational resources, and therefore cannot be performed on most personal machines. In order to successfully process the dataset in a timely manner, the San José State University (SJSU) College of Engineering (CoE) High Performance Computing (HPC) system was used to implement the RCP. The HPC cluster is a Linux based machine with 36 nodes, each with either 128GB or 256GB of RAM, for a total

memory of 6.7TB for the entire cluster [17]. The Slurm Batch scheduler was used to deploy jobs onto the HPC.

B. RCP Overview and Steps

Raw reads for GLDS-258 were obtained from NASA GeneLab's online repository, in the form of compressed fastq files. The raw data was prepared on the Illumina HiSeq platform and submitted by Dr. Joe Wu's research group. After downloading the data from NASA GeneLab's website, analysis using the RCP can begin.

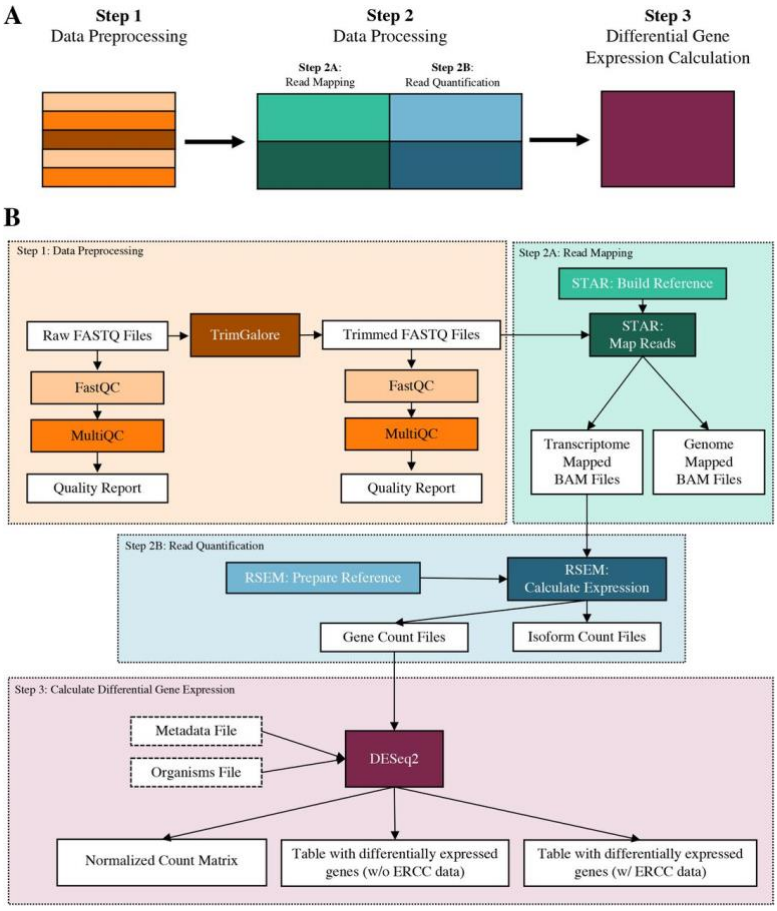

Figure 3: Overview of the RCP [11]

A high-level overview of the workflow for the RCP is shown in Figure 3. In general, the pipeline will take input in the form of compressed fastq files, and output csv files and diagrams with analyzed data [11]. Each of the steps below were completed by running an adapted Slurm script originally provided by Dr. Amanda Saravia-Butler from NASA GeneLab.

1) Raw Data Quality Control

The first step of the pipeline is to perform quality control with FastQC and MultiQC; these are two bioinformatics programs that were developed to streamline the process of assessing the quality of RNA sequence reads [18], [19]. FastQC and MultiQC versions 0.11.9 and 1.9 were used, respectively.

2) Pre-processing Data

The next step of the pipeline is to pre-process the raw sequence data. TrimGalore! Is a script that wraps around Cutadapt and FastQC to filter unnecessary or bad quality data [20]. Pre-processing is a necessary step to remove low quality reads, remove adapters, and remove any reads that have become too short in the trimming process [20]. TrimGalore! and Cutadapt version 0.6.6 and 3.2 were used, respectively.

3) Align and Count Reads

After pre-processing the data, the reads must be aligned to a reference genome by building a STAR index. The STAR tool requires a reference genome in the form of a fasta file, along with a GTF annotation file [21]. Both human genome fasta and GTF annotation files are from Ensembl release 102, and STAR version 2.7.7a was used.

RSEM is a weighted quantitation tool that uses a maximum likelihood estimation to match reads to genes [22]. RSEM version 1.3.1 was used for quantification.

4) Differential Gene Expression and Normalization Using R

A R script provided by Dr. Amanda Saravia-Butler was modified then used in order to perform differential gene expression using DESeq2. Analysis was completed using R version 4.0.5. Data visualization was also performed using R with the differential gene expression data output from DESeq2.

The R script followed the following steps. First, DESeqDataSet object is created. We filter out all genes with counts of 10 or less across all conditions. Then, `estimateSizeFactors()` is called on the DESeqDataSet object. This function estimates the size factors with a median ratio method. A size factor is created for every sample by dividing the median ratio of all the counts by the mean of the genes for all samples. Each sample's raw counts are then divided by the size factor created in the last step.

The next function to be called on the object is `estimateDispersions()`, which gathers dispersion estimates for a negative binomial distribution of data. The calculations will assume that the gene set counts will follow a negative binomial distribution, which is why this method is used. Each dispersion per gene is found by maximizing the Cox Reid-adjusted profile likelihood [23]. Essentially, the dispersion for each gene will describe how much of the expressed variance across the samples deviate from the mean of the gene's expression.

Finally, the function `nbinomWaldTest()` is called on the DESeqDataSet object. This function tests for any significance of coefficients in a Negative Binomial distribution, and will utilize the dispersion estimates from the `estimateSizeFactors()` and `estimateDispersions()` functions which were previously called. The Wald test will calculate the p-value that a gene's expression is significant when compared to a null hypothesis. The raw count and normalized data are then exported.

Adjusted p-values are calculated using the Benjamini-Hochberg technique. The false discovery rate used for this calculation is the default at 0.1. Finally, all data analysis is performed using the output table from DEDeq2. In the R script, multiple Principal Component Analysis (PCA) graphs were created, along with a heat map comparing all three groups.

C. Nextflow Implementation of RCP for Human Genome

The Nextflow implementation was initially created and tailored the analyze the GLDS-104 dataset, which is an experiment using mouse samples [12]. In order to run GLDS-258 using Nextflow, certain configuration files and changes need to be made since GLDS-258 uses human samples. Since the Nextflow implementation is not the official standard for analyzing NASA GeneLab data, the output which was used for analysis is from manually running the RCP for GLDS-258, in case there are any unknown discrepancies between the two analysis workflows. This section will briefly discuss and show examples of configuration files used for running the pipeline through Nextflow in order to demonstrate an alternative, more automated version of the RCP.

Two configuration files for the dataset of interest must be created. The first configuration file should include information about the dataset, and the second configuration file will specify how to run the workflow on the HPC. The values in the configuration file are parameters that will be passed into the scripts run in the pipeline. Any of these can be removed and inputted manually into the scripts, if needed. Other parameters can also be added to the configuration file, to make the scripts more parameterized and generic. Figure 4 shows an example configuration file that can be used for GLDS-258.

```
params {
  // used for debugging to limit samples used
  // default value here means no limit
  limiter = -1

  storeDirPath = '../workdir/GLDS-258/store'
  publishDirPath = '../workdir/GLDS-258/publish'

  // URLs to download data from GeneLab
  GLDS_URL_PREFIX = 'https://genelab-
data.ndc.nasa.gov/genelab/static/media/dataset/GLDS-258_rna_seq_'
  GLDS_URL_SUFFIX = '?version=1' // Assumption that these raw files
are always version 1

  // ensembl parameters for genome and annotations
  genomeFasta = 'path/to/Homo_sapiens.GRCh38.dna.
     primary_assembly.fa'
  genomeGTF = 'path/to/Homo_sapiens.GRCh38.102.gtf'
  ensembl_version = 102

  // source: https://genelab-
data.ndc.nasa.gov/genelab/accession/GLDS-258/
  // extracted from Sample Table, Column Sample Name
  samples = ['GSM4066596, 'GSM4066597', 'GSM4066598',
             'GSM4066599', 'GSM4066600', 'GSM4066601',
             'GSM4066602', 'GSM4066603', 'GSM4066604',
             'GSM4066605', 'GSM4066606', 'GSM4066607',
             'GSM4066609', 'GSM4066608', 'GSM4066610',
             'GSM4066611', 'GSM4066612', 'GSM4066613',]
```

Figure 4. Example configuration file for GLDS-258

Figure 5 shows the standard executor configuration file used, which shows instructions for running on the HPC. However, it is not optimal. Executor configuration files can be formatted to utilize the Slurm job scheduler for more intensive jobs. This would allow Nextflow to use resources required by open tasks, while the standard configuration shown in Figure 5 will reserve all resources in the beginning of execution, despite the differing demands of the pipeline throughout the workflow. For instance, when initially downloading the data from NASA Genelab, there is essentially no computational resources needed, and all the allocated computing resources will sit idle. The optimal executor example configuration file can be found on the Nextflow implementation repository at https://github.com/J-81/masterProject/blob/dev/config/executor/cos-hpc-4-node.config.

```
params {
  executor = 'local'

  withLabel: big_mem {
    memory = 70.GB
  }

  withLabel: maxCPU {
    cpus = 15
  }
}
```

Figure 5: Example of HPC executor configuration file used.

D. Gene Set Analysis

To understand the biological impact shown in the results, further analysis of the dataset can be performed using the results from the RCP. In particular, the Database for Annotation, Visualization, and Integrated Discovery (DAVID) and Gene Set Enrichment Analysis (GSEA) tools were used for additional analysis.

1) DAVID

DAVID allows users to input a gene list of interest and view any biological meaning behind the set as a whole. For this project, DAVID version 6.8 was used. Since we are the most interested in the flight versus ground control group, we inputted all significantly expressed genes with an adjusted p-value of less than or equal to 0.05 using their Ensembl gene identifiers. The original study performed a DAVID analysis using regular p-values instead of adjusted p-values; for comparison purposes, this study also analyzes the gene set with a regular p-value of less than or equal to 0.05 using DAVID.

2) GSEA

GSEA allows user to further understand the significance of complete pathways and relationships within their dataset. Analysis was completed with GSEA version 4.1.0 for Mac. GSEA was performed on 25,100 genes from the output of the DESeq2 analysis. MsigDB hallmark gene sets version 7.4 was used for comparison, along with MsigDB gene ontology biological process (GO BP) gene set version 7.2

3) Leading Edge Analysis

Leading Edge Analysis (LEA) was conducted with GSEA version 4.1.0 for Mac. The GO BP gene set was used for LEA. A total of 830 gene sets from the GO BP gene set were selected for LEA because they had a nominal p-value of less than 0.05.

III. RESULTS

A. Data Quality Control

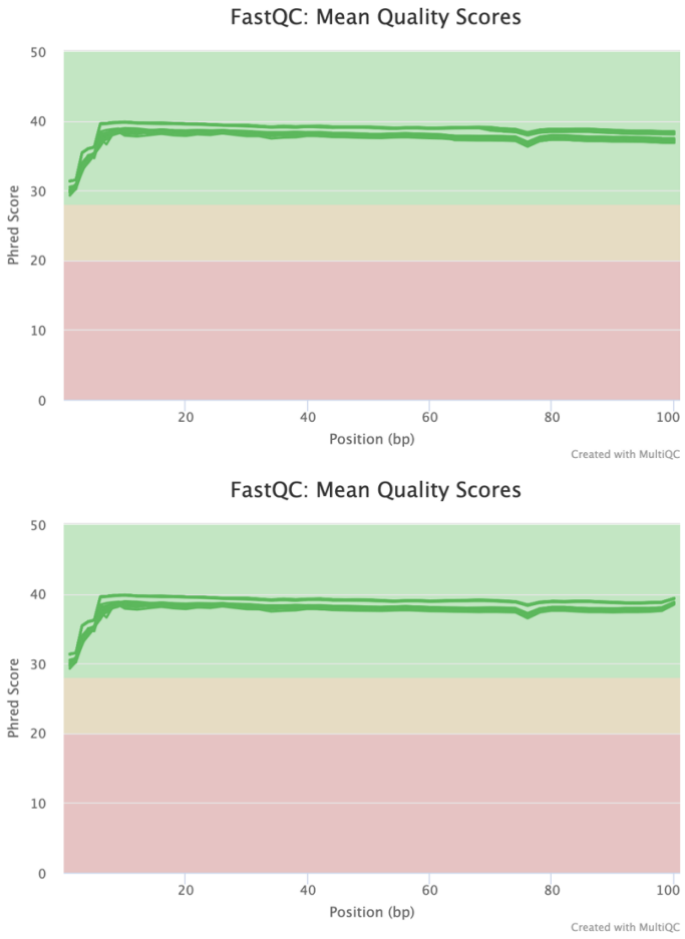

Figure 6: FastQC Mean Quality Score graphs for (A) raw reads and (B) trimmed reads.

A MultiQC report was generated for the raw reads and trimmed reads. Results from the MultiQC process showed that overall, all the reads are high quality since all samples had at least a read score of 28, indicated by the green region in the graphs in Figure 6A and Figure 6B. Each line on the graphs represents a unique sample. There are two distinct clusters of lines in both graphs in Figure 6. The clusters of lines with higher quality scores are the forward reads, while the clusters of lines with lower quality scores are the reverse reads. It is known that reverse reads tend to be lower quality than forward reads; however, the reason for this discrepancy is unknown

[24]. Even so, the quality for both the forward and reverse reads are high enough quality to be able to confidently continue with the rest of the analysis process.

In addition to the overall mean quality scores, the MultiQC report also reported other quality control scores. All samples had a sequence of 100 base pairs, as we expected. There were no samples with adapter contamination greater than 0.1%. Less than 1% of the reads for all samples were made up of overrepresented sequences. These results indicate that we can move forward with confidence for the rest of our analysis.

B. STAR Alignment Analysis

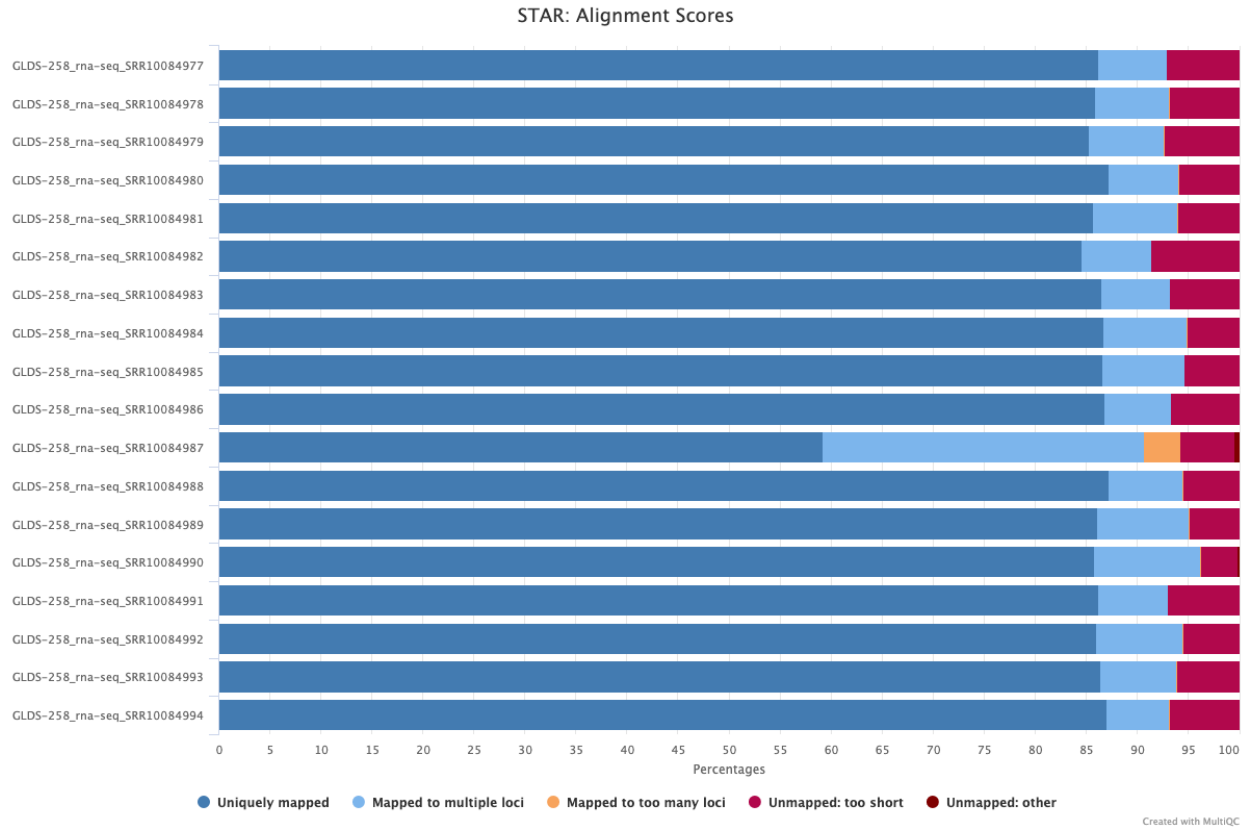

Figure 7: STAR Alignment Scores for all samples.

After STAR Alignment was completed, another round of MultiQC was run to understand the overall quality of the alignment process. The alignment results for each sample are shown above in Figure 7. All samples had 84.6 or greater percentage of uniquely mapped reads, except sample SRR10084987, which is a flight sample from line two. Additionally, unmapped parts of samples due to being too short ranged from 3.6 to 8.5 percent. All samples mapped less than 0.1 percent to too many loci, except for post the singular flight sample from line 2. There were no reads from any sample in the other unmapped category.

C. RSEM Counts Analysis

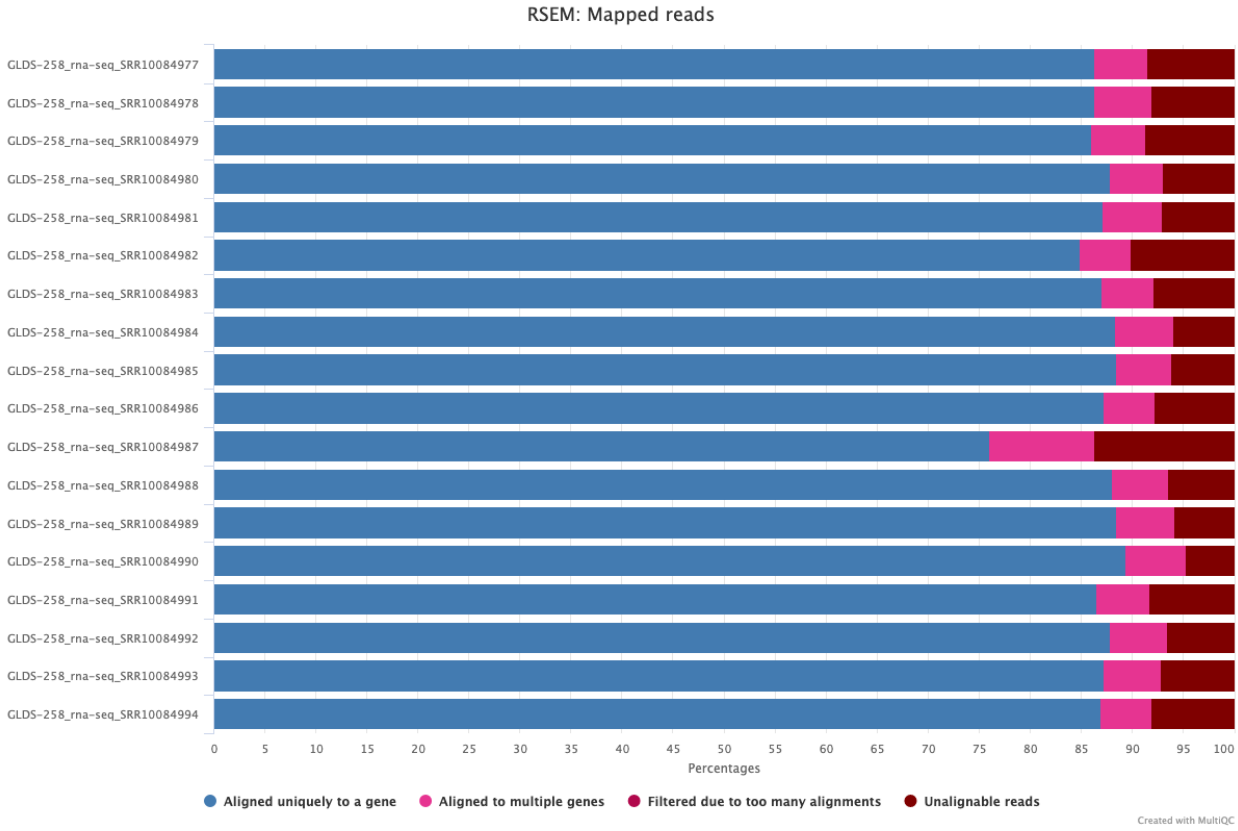

Figure 8: RSEM Mapped Reads for all samples.

After RSEM counting results were completed, another round of MultiQC was performed to analyze the quality of the counts. All the samples had at least 84.9 percent of reads align uniquely to a gene, except for the flight sample from line two, which only had 76.1 percent of samples map uniquely to a gene. For all samples, 4.9 percent to 5.9 percent of reads were aligned to multiple genes, besides the flight sample from line two, which has 10.2 percent of reads aligned to multiple genes.

D. DESeq2 Differential Gene Expression

1) General Comparison to the Previous Study

The original study conducted by Dr. Wu's research group from Stanford University School of Medicine found 3,008 genes differentially expressed between ground and flight groups, 2,026 genes differentially expressed between post-flight and flight groups, and 1,049 genes differentially expressed between post-flight and ground groups with $p <= 0.05$ using a two-tailed Student's t test [13]. This project found similar results, with 3,010 genes differentially expressed between ground and flight groups, 1,295 genes differentially expressed between post-

flight and flight groups, and 1,182 genes differentially expressed on a p-value of <= 0.05. The results of the two pipelines are shown in the bar graph in Figure 9.

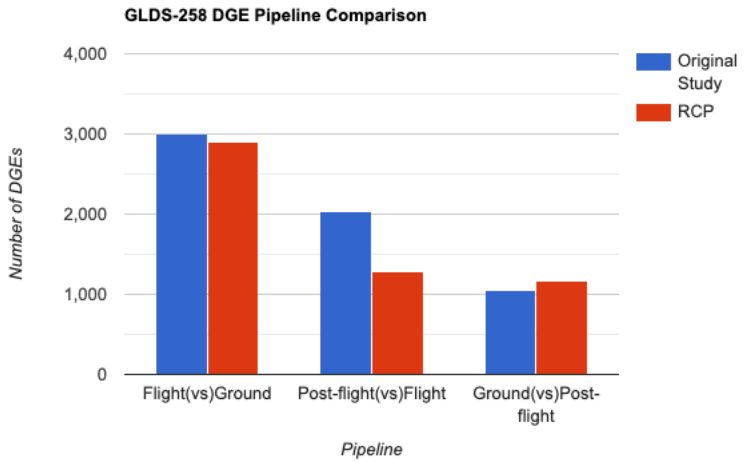

Figure 9: Comparison of differentially expressed genes between the RCP and Stanford's data analysis pipeline.

Additionally, this project focuses on the adjusted p-values instead of the p-values, to account for the false discovery rate due to multi-testing. The original study does not account for the false discovery rate. In terms of adjusted p-values for this project, there were 212 differentially expressed genes between flight and ground control groups, 30 differentially expressed genes between flight and post-flight groups, and zero differentially expressed genes between ground control and post-flight groups.

2) Principal Component Analysis

The raw counts for this analysis contain raw, unnormalized counts for each sample. This indicates that the samples may have different read depths, which are the number of reads that come off the sequencer for each sample. After normalization of data using the DESeq2 estimateSizeFactors() function, the new, normalized data can be plotted against the unnormalized counts data in two Principal Component Analysis (PCA) plots shown in Figure 10A and Figure 10B. There is no considerable difference between Fig. X (A) and Fig. X (B). In both graphs, there are three distinct clusters, which correlate to the three cell lines used to create the hiPSC-CMs. The original study also found that the samples clustered based on the three original cell lines [13].

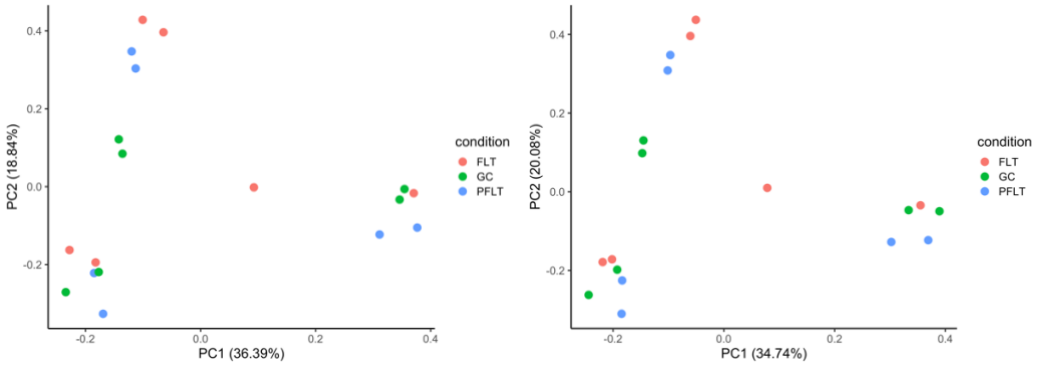

Figure 10: Principal Component Analysis for
(A) raw data and (B) normalized data.

In addition to plotting the unnormalized and normalized counts in PCA plots, the differentially expressed genes can be further visualized in a PCA plot. The normalized counts for the flight versus ground control group were filtered by an adjusted p-value of less than 0.05, then filtered by the log fold change greater than 1 and less than -1, and finally plotted. Figure 11 shows the resulting PCA from this filtering process. Instead of clustering by cell line as shown in the previous PCA plots, the samples are now more closely clustered by their condition group. Ground control and post-flight samples are clustered slightly closer together in comparison to the flight group.

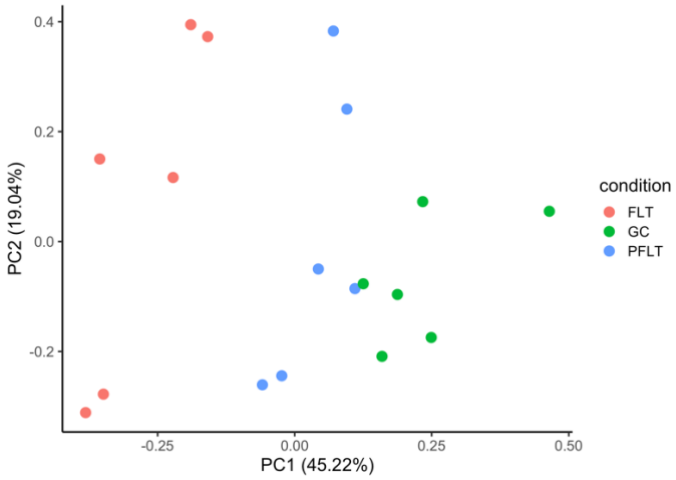

Figure 11: Principal Component Analysis (PCA) of differentially expressed genes between flight and ground control groups

3) Heatmap

A heatmap was created for the differentially expressed genes from the flight versus control group analysis. In Figure 12, the left most six samples represent the post-flight samples, the middle six samples represent the flight samples, and the last six samples represent the ground control samples. The differentially expressed genes number is 117, instead of the original 212,

because genes with not available names were removed. The post-flight and ground control samples columns on the heat map have similar levels of expression, while the flight samples have a visibly unique expression.

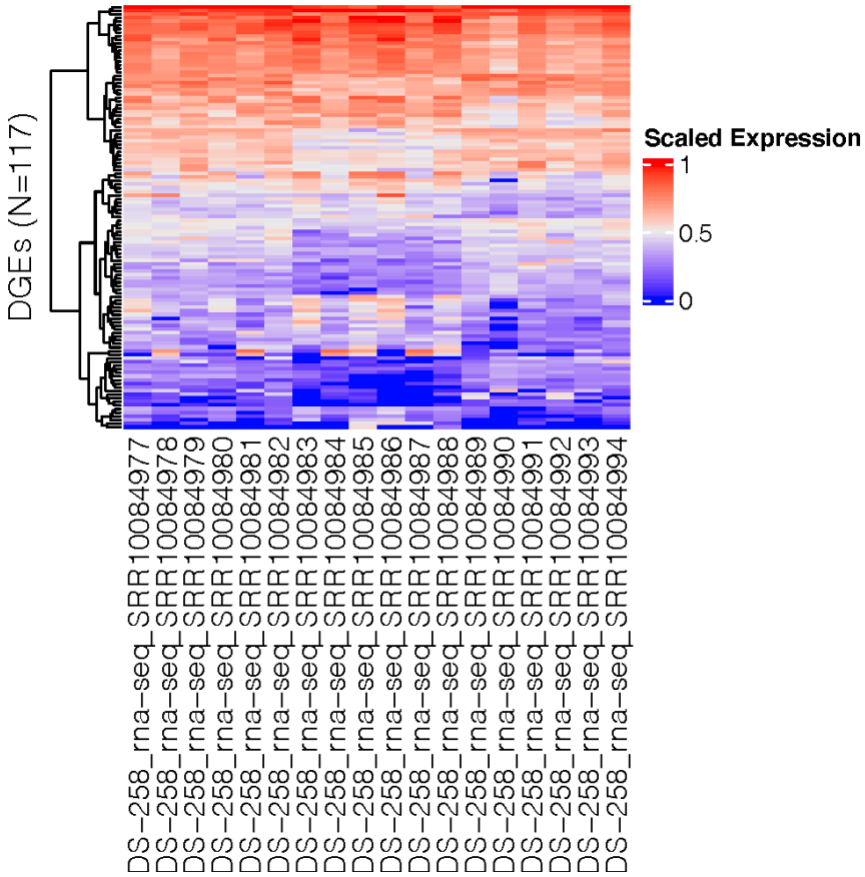

Figure 12: Heatmap of differentially expressed genes from flight versus control groups across all samples

E. Gene Set Analysis

1) DAVID Gene Set Clusters

Two gene sets were submitted to DAVID for analysis. First, the set of differentially expressed genes for fight versus control group p-values less than or equal to 0.05. This set was submitted to compare and contrast the results with the previous study conducted by Stanford University, since they also submitted a gene set with p-values less than 0.05. The second gene set submitted to DAVID was the list of differentially expressed genes between flight and post-flight groups, also with a p-value of less than or equal to 0.05.

The first set of genes submitted has 3,010 differentially expressed genes based on a p-value of less than or equal to 0.05. DAVID was able to recognize 2,506 genes as from the species *Homo Sapiens* and 503 genes were unknown. The top three enriched annotation clusters

displayed in Table I. The first cluster suggests that the Mitochondrion pathways were most heavily impacted from spaceflight and microgravity environments, with an enrichment score of 26.43. The second cluster shows many impacts on pathways within the KEGG Pathway, with an enrichment score of 15.16. The final cluster shows expected impacts on the cardiac muscle contraction, as well as additional mitochondrial group disruptions. Since we are most interested in the flight versus ground control groups, key terms from the significant DAVID clusters were submitted to the PubMed search function to analyze how well-researched these topics were. While mitochondrial mechanisms and DNA damage and repair searches yielded many results, the neurodegenerative diseases from KEGG's pathway do not seem as well researched. The query and their respective results are shown in Table III.

The second set of genes was 1,295 genes from the post-flight versus flight groups. DAVID was able to recognize 1,019 of these genes as *Homo Sapiens* genes, while 275 were unknown. The top three enriched annotation clusters for these results are shown in Table II. There were far fewer significantly enriched groups in the post-flight versus flight groups in comparison to the flight versus group control groups. Additionally, the enrichment scores for this gene set are all much smaller than the enrichment group scores from the ground versus flight control group. The most significant cluster, the first cluster, indicates that DNA repair and damage mechanisms were impacted. This cluster has an enrichment score of 2.82.

TABLE I.
TOP THREE ENRICHED ANNOTATION DAVID CLUSTERS
FOR FLIGHT VERSUS GROUND CONTROL GROUPS

Category	Term	Benjamini (FDR)
Cluster 1, Enrichment Score: 26.43		
UP_KEYWORDS	Mitochondrion	1.2E-41
GOTERM_CC_DIRECT	Mitochondrial inner membrane	2.4E-28
UP_SEQ_FEATURE	Transit peptide: Mitochondrion	1.9E-22
UP_KEYWORDS	Transit peptide	8.5E-24
GOTERM_CC_DIRECT	Mitochondrial matrix	8.1E-6
Cluster 2, Enrichment Score: 15.16		
UP_KEYWORDS	Mitochondrion inner membrane	1.2E-24
KEGG_PATHWAY	Parkinson's disease	2.6E-21
KEGG_PATHWAY	Huntington's disease	4.8E-19
KEGG_PATHWAY	Alzheimer's disease	2.5E-18
KEGG_PATHWAY	Oxidative phosphorylation	3.0E-18
UP_KEYWORDS	Electron transport	4.7E-14
UP_KEYWORDS	Respiratory chain	7.7E-14
KEGG_PATHWAY	Non-alcoholic fatty liver disease	7.7E-12
GOTERM_BP_DIRECT	Mitochondrial respiratory chain complex I assembly	2.4E-9
GOTERM_CC_DIRECT	Mitochondrial respiratory chain complex I	4.0E-8
GOTERM_BP_DIRECT	Mitochondrial electron transport, NADH to ubiquinone	6.0E-7
GOTERM_MF_DIRECT	NADH dehydrogenase (ubiquinone activity)	7.7E-7

TABLE I. (CONTINUED)

Cluster 3, Enrichment Score: 4.08		
KEGG_PATHWAY	Cardiac muscle contraction	7.8E-6
GOTERM_BP_DIRECT	Mitochondrial electron transport, cytochrome c to oxygen	2.7E-2
GOTERM_MF_DIRECT	Cytochrome-c oxidase activity	1.6E-1
GOTERM_BP_DIRECT	Hydrogen ion transmembrane transport	2.5E-1
GOTERM_CC_DIRECT	Mitochondrial respiratory chain complex IV	9.1E-2

TABLE II.
TOP THREE ENRICHED ANNOTATION DAVID CLUSTERS
FOR FLIGHT VERSUS POST-FLIGHT GROUPS

Category	Term	Benjamini (FDR)
Cluster 1, Enrichment Score: 2.82		
UP_KEYWORDS	DNA damage	8.3E-2
UP_KEYWORDS	DNA repair	9.3E-2
Cluster 2, Enrichment Score: 1.99		
UP_KEYWORDS	ATP-binding	1.5E-2
UP_KEYWORDS	Nucleotide-binding	1.5E-2
Cluster 3, Enrichment Score: 1.64		
UP_KEYWORDS	WD repeat	9.8E-2

TABLE III.
PUBMED RESULTS OF TOP DAVID CLUSTERS

Query	Number of Results
“mitochondrial” microgravity	109
“mitochondrial” spaceflight	92
“DNA damage” microgravity	72
“DNA damage” spaceflight	129
“DNA repair” microgravity	64
“DNA repair” spaceflight	86
"Parkinson" microgravity	8
"Parkinson" spaceflight	20
"Alzheimer" microgravity	6
"Alzheimer" spaceflight	16
"Huntington" microgravity	5
"Huntington" spaceflight	9

2) GSEA Results

There were 50 total gene sets in the MSigDB hallmark gene set used for comparison. These hallmark gene sets have well-known biological states and processes with consistent expression [25]. There were 24 gene sets that were significantly upregulated at a false discovery rate (FDR) of less than 25%, and five gene sets that were significantly downregulated, also at a FDR of less than 25%. This FDR is much higher than most gene set analyses, however, it was recommended by the GSEA software because the hallmark gene set has a small number of gene sets. Table IV show the significantly upregulated gene sets, and Table V shows the significantly downregulated gene sets.

Leading edge analysis was performed with the GO BP gene sets. There was a total of 830 gene sets with a p-value less than 0.05 that were chosen for leading edge analysis. The genes that had the highest number of counts across the gene sets are shown in Table VI. The leading edge analysis results can also be seen in Figure 12.

TABLE IV.
GSEA UPREGULATED HALLMARK GENE SETS

Gene Set	NES	NOM p-value	FDR	FWER p-val
HALLMARK_OXIDATIVE_PHOSPHORYLATION	3.53	0.000	0.000	0.000
HALLMARK_REACTIVE_OXYGEN_SPECIES_PATHWAY	2.45	0.000	0.000	0.000
HALLMARK_FATTY_ACID_METABOLISM	2.40	0.000	0.000	0.000
HALLMARK_ADIPOGENESIS	2.38	0.000	0.000	0.000
HALLMARK_MYOGENESIS	2.33	0.000	0.000	0.000
HALLMARK_XENOBIOTIC_METABOLISM	2.16	0.000	0.000	0.000
HALLMARK_MTORC1_SIGNALING	1.90	0.000	0.000	0.002
HALLMARK_TNFA_SIGNALING_VIA_NFKB	1.90	0.000	0.000	0.002
HALLMARK_HYPOXIA	1.87	0.000	0.001	0.005
HALLMARK_P53_PATHWAY	1.77	0.000	0.001	0.009
HALLMARK_UV_RESPONSE_UP	1.77	0.000	0.001	0.009
HALLMARK_HEME_METABOLISM	1.76	0.000	0.001	0.011
HALLMARK_CHOLESTEROL_HOMEOSTASIS	1.73	0.000	0.002	0.020
HALLMARK_PEROXISOME	1.63	0.000	0.026	0.066
HALLMARK_BILE_ACID_METABOLISM	1.49	0.011	0.025	0.285
HALLMARK_DNA_REPAIR	1.48	0.004	0.026	0.210
HALLMARK_APOPTOSIS	1.47	0.004	0.025	0.317
HALLMARK_COAGULATION	1.46	0.012	0.029	0.373
HALLMARK_ESTROGEN_RESPONSE_LATE	1.42	0.008	0.042	0.513
HALLMARK_MYC_TARGETS_V1	1.41	0.011	0.043	0.531
HALLMARK_GLYCOLYSIS	1.35	0.028	0.072	0.750
HALLMARK_IL2_STAT5_SIGNALING	1.28	0.042	0.127	0.127
HALLMARK_UNFOLDED_PROTEIN_RESPONSE	1.25	0.101	0.147	0.147
HALLMARK_COMPLEMENT	1.19	0.129	0.228	0.228

TABLE V.
GSEA DOWNREGULATED HALLMARK GENE SETS

Gene Set	NES	NOM p-value	FDR	FWER p-val
HALLMARK_MITOTIC_SPINDLE	-1.69	0.000	0.011	0.016
HALLMARK_G2M_CHECKPOINT	-1.56	0.000	0.027	0.071
HALLMARK_E2F_TARGETS	-1.52	0.110	0.029	0.113
HALLMARK_HEDGEHOG_SIGNALING	-1.30	0.058	0.150	0.569
HALLMARK_INTERFERON_ALPHA_RESPONSE	-1.30	0.223	0.121	0.571

TABLE VI.
TOP LEADING EDGE ANALYSIS GENES

Gene	Name	Expression	Number of Gene Sets
EDN1	Endothelin 1	Upregulated	97
AGT	Angiotensinogen	Upregulated	87
CAV3	Caveolin 3	Upregulated	84
APOE	Apolipoprotein E	Upregulated	83
PPARGC1A	PPARG Coactivator 1 Alpha	Upregulated	72
APOA1	Apolipoprotein A1	Upregulated	69
RYR2	Ryanodine Receptor 2	Upregulated	69

IV. DISCUSSION

A. Computational Resources

The combined total of all GLDS-258 raw RNA-Seq data is over 43 GB. This large amount of data creates a bottleneck effect in the beginning of the pipeline, causing the entire workflow to await the completion of this download. Thus, computing environments with sufficient download speed are required to complete the RCP in a timely manner.

Large amounts of RAM are needed to process data through the RCP. The process that required the most memory was STAR alignment, which required 80 GB of RAM. This value was reached through repeated trial and error by modifying the parameter `#SBATCH --mem=80000` in the Slurm script until it was able to run successfully. The `--mem` parameter in Slurm indicates the amount of real memory required per node [26]. In addition, the STAR alignment parameter `--limitBAMsortRAM` was increased to 70,000,000,000 for there to be enough memory for BAM sorting during STAR alignment. `--limitBAMsortRAM` allocates the maximum available RAM in bytes for sorting BAM [21]. This large amount of RAM needed is a trade-off for the fast mapping speed of the program [21], and suggests that this step in the RCP can also create a bottleneck effect.

Another process that required a large amount of RAM was the STAR reference building process. The parameter `#SBATCH --mem=60000` was used to build the *Homo Sapiens* reference. The reference must be recreated if the user wishes to use newer version of the Ensembl releases. Otherwise, a reference genome only needs to be built once for any organism, and the resulting output files can be reused for any dataset from that organism. The files produced from STAR reference building are extremely large, totaling up to over 30GB. If users wish to reuse the reference genome on another system, it may be not time efficient to repeatedly copy the files to multiple computers. In this case, it may just be faster to run the script and build the reference on the new system, which should take three to four hours.

An option that may be helpful in the future is to consider compressing the reference files and hosting it on cloud storage solution, such as Amazon Web Services Simple Storage Service. This option would allow users to build the reference for any organism once, persist the files onto cloud storage, and download them from any location or computer. Then, the STAR reference building process could be removed from the RCP and would decrease the total computational load significantly.

B. Raw and Trimmed Data Quality Analysis

The average quality of all the raw and trimmed reads were above 28; 28 is the lower boundary for a good quality read. There are two distinct clusters within both raw and trimmed mean quality score graphs in Figure 6A and Figure 6B. The cluster in both graphs with higher average quality scores are the forward reads, while the cluster with lower average quality scores

are the reverse reads. This is a known issue with Illumina sequencing, and the cause is not yet known, as previously mentioned [24]. This issue may be countered by using a different next generation sequencing technology. However, Illumina sequencing has almost two thirds of the market share in the next generation sequencing market because it's cost effective, even though it's less accurate than other technologies on the market [24]. Moreover, most of the existing GeneLab datasets were sequenced on Illumina technology.

FastQC and MultiQC are the technologies that are used for quality control in the RPC. These two technologies are largely popular in the bioinformatics field; however, there are other technologies that may provide more insight into the quality of sequencing data. Technologies for quality control, including FastQC, htSeqTools, and SAMState only focus on raw sequence metrics [27]. A possible alternative, RNA-SeQC is a tool that considers additional checks, such as saturation checking [27]. Thus, future versions of the RCP should consider expanding the use of quality control technologies to produce more insight on the quality of sequenced data.

C. STAR Alignment and RSEM Mapped Reads

1) Quality Analysis

STAR Alignment scores in Figure 7 all are good quality due to the high percentage of uniquely mapped reads. The only sample that does not have a high percentage of uniquely mapped reads is SRR10084987, which is the flight sample from line two. A lower number of uniquely mapped reads can indicate that there are more reads mapping to multiple loci. This can also be seen for sample SRR10084987 in Figure 7, which has many reads mapped to multiple loci in comparison to all the other reads.

MultiQC results of RSEM mapped reads in Figure 8 show that all samples have good quality mapping, except for sample SRR10084987. This is expected after viewing the poor alignment results for the same sample in Figure 7.

It is important for reads to be uniquely aligned to the reference for accurate quantification [28]. There are various biological mechanisms that may contribute to the appearance of sequence duplication, such as recombination, transposable elements, and alternative splicing. [29]. However, it seems unlikely that the duplication apparent in sample SRR10084987 would be due to biological mechanisms, since it is not present in any of the other samples. Instead, other external factors may have affected the mapped read results for sample SRR10084987.

2) Comparison to Previous Study

The study from Dr. Wu's research group at Stanford University used a variety of different technologies and references, which could account for some of the differences shown in Figure 9. Table VII shows the similarities and differences between this project and the Stanford University publication.

The first major difference between the two studies is the tool used for alignment. Based on a previous study comparing the performance of HISAT2 and STAR, the two programs tend to have a large number of overlapping reads that will map to the same area in the reference genome [30]. STAR, however, had a higher mapping rate and more mapped reads in comparison to HISAT2 in this study [30].

The next difference is the choice of quantification tool used. A previous study found that RSEM for isoform quantification is more aligned with the true count values, while featureCounts tends to undercounts the samples when using idealized data with no indels, no single nucleotide polymorphisms, and other modifications [31]. However, with more realistic data, featureCounts was found to have more accurate counts [31]. This is only a single study, and a definitive answer regarding which quantification tool is better is not possible. Both tools are different and have their respective advantages and disadvantages. These differences can explain some of the variation found in the results when comparing the two studies.

The final major difference between the two analysis workflows are the reference genome and annotations. Wu's study uses Ensembl 85, which was released in July 2016. This project uses Ensembl 102, which was released in December 2020. One major update in Ensembl 102 for the human genome was an update that translates any non-ATG start codons to Methionine. This difference is another factor that could have contributed to the difference in results for this project compared to the Stanford University publication.

TABLE VII.
COMPARISON OF RESOURCES USED BETWEEN STUDIES

Resource	**This Project**	**Wu [13]**
Alignment	STAR	HISAT2
Quantification	RSEM	featureCounts
Human genome	hg38	hg38
Annotation	Ensembl 103 (*Homo Sapiens*)	Ensembl 85 (*Homo Sapiens*)

D. Differential Gene Expression

There are several tools designed for differential gene expression analysis, including two popular choices of DESeq2 and edgeR [32]. Both tools are based on a negative binomial model distribution. When estimating dispersion factors, DESeq2 will use all genes with similar expression. In contrast, edgeR combines any common dispersions among genes that is estimated using a likelihood method and adds this two a gene-specific dispersion [32]. DESeq2 and edgeR are both solid choices for differential gene expression.

Further study into the DESeq2 package shows that there is a function DESeq() that calls the functions estimateSizeFactors(), estimateDispersions(), and nbinomWaldTest() automatically. There were no arguments included in these three function calls. Since the R script used in this project manually calls all of these functions, the script could be further optimized by directly calling the DESeq() function. Readability of the code would remain high since there are no additional arguments required for this project.

E. Gene Set Analysis

1) DAVID Enriched Clusters Results

The most significant enrichment found for flight versus ground control groups in Wu's study was the mitochondrion and transit peptide groups with an enrichment score of 48.87 [13]. This group was also found in this project's flight versus ground control groups to be the most enriched, with an enrichment score of 26.43 as shown in Table I. Other similar enrichment scores between the two include electron transport and mitochondrial respiratory chain scores at 14.7 and 15.16 for Wu's study and this project, respectively [13]. The results which indicate that mitochondrial metabolism is altered due to microgravity align Wu's study, as well as with previous studies [13]. Rat cardiac muscle cells have been shown to have increased expression in the mitochondrial metabolism pathway in microgravity environments [10]. Human cells have also shown different amounts of mitochondrial pathway gene expression in the NASA Twins Study [7]. However, Parkinson's disease, Huntington's disease, and Alzheimer's disease, three terms from the KEGG PATHWAY group in Table I cluster 1, had high enrichment scores of 15.16 but were not mentioned in Wu's study. The KEGG PATHWAY is a set of various known molecular pathways, including those about human diseases [33].

The top enriched clusters for post-flight versus flight groups, as shown in Table II, includes DNA damage and DNA repair groups, with an enrichment score of 2.82. This aligns well with the known knowledge that space radiation can cause DNA damage because of the interaction between charged particles and DNA [34]. Previous studies have also shown that DNA repair mechanisms expression is changed in microgravity environment, which can consequently increase the amount of DNA damage in human cells [35]. It is interesting, however, that the DNA damage and repair clusters are not present in Table I. It is possible that the hiPSC-CMs on the ISS return to similar levels of expression as ground controls after being returned to Earth [14]. This is supported by the similar levels of expression from the post-flight and ground control groups, shown in Figure 12. The DNA damage and repair expression may have been overshadowed by the large amounts of differential expression presented in the first three clusters in Table I, as supported by the large enrichment values. This suggests that much of the highly enriched mitochondrial mechanisms shown in Table I return to normal, less enriched levels upon return to Earth. DNA damage and repair pathways, however, may still be affected 10 days post-flight.

2) GSEA Upregulated Hallmark Gene Sets

GSEA is an important tool when studying differential expression because it analyzes an entire gene set, instead of only significantly expressed genes. This allows the program to see entire pathways or mechanisms that may have been impacted by microgravity and gives the user and opportunity to see the results as a whole picture, instead of gene by gene. The hallmark gene sets were chosen for this project because it is considered a good starting point for analyses that are looking for general insight into their dataset.

The most upregulated gene set is the oxidative phosphorylation gene set. Oxidative phosphorylation a process that takes place within the mitochondria and is the primary energy source for human cells [36]. The upregulation of this gene set aligns well with the results from the DAVID analysis, since the mitochondrial pathways had high enrichment scores. The second most upregulated gene set is the reactive oxygen species pathway. The reactive oxygen species pathway plays an important role in regulating various cellular processes [37]. Previous studies have shown that an overproduction of reactive oxygen species may promote neurogenerative disorders, such as Alzheimer's disease, Huntington's disease, and Parkinson's disease [38]. This upregulation in this hallmark gene set matches extremely well with the enriched neurodegenerative diseases shown in the DAVID cluster 1 in Table 1.

Upregulated genes shown in Table VI from the leading edge analysis shows upregulation in several genes including AGT and APOE. APOE, or Apolipoprotein E, may have isoforms that can affect the lipid metabolism, which can in turn affect degenerative processes [39]. This could also be a contributing factor for the third most upregulated hallmark gene set, the fatty acid metabolism, shown in Table IV. AGT, or Angiotensinogen, has been shown to increase oxidative stress, which is related to the oxygen reactive species, and cause neuroinflammation and neurodegeneration [40]. This is another example of how one of the upregulated hallmark gene sets, the oxygen species pathway, correlates with the leading edge analysis genes.

3) GSEA Downregulated Hallmark Gene Sets

The downregulated top three significant hallmark gene sets, including mitotic spindle, G2M checkpoint, and E2F targets are all related to the cell division cycle. This is particularly puzzling given that adult cardiomyocytes do not divide [5]. However, hiPSC-CMs have previously been shown to have had their cell cycles activated [41]. It is unclear whether the hiPSC-CMs used for this experiment are cell cycle activated cardiomyocytes. Further research is required particularly for the downregulated hallmark gene sets to reach a more definitive answer.

4) Spaceflight, Cardiomyocytes, and Neurodegenerative Diseases

The neurodegenerative disorders presented under Cluster 2 in Table I, are not mentioned in Wu's study. Neurodegenerative disorders typically include a steady loss of neurons, which can cause dementia, poor motor skills, among other symptoms [42]. The high enrichment scores of the neurodegenerative disorders indicate that exposure to microgravity may play a role in altering expression of genes which are related to neurodegenerative pathways in cardiac cells. Although many neurological disorders are not typically thought to have any association with the cardiovascular system, there is evidence that heart defects could aid in the progression of neurodegenerative diseases [43]. In particular, huntingtin, the protein responsible for Huntington's disease, is expressed in both cardiovascular and nervous systems [43]. Another study proposes that in order to prevent microgravity-induced neurodegeneration, possible hypergravity therapy could be performed on individuals who are in space [44]. There has also been evidence from multiple studies that cardiac sympathetic nerve degenerates in the beginning stages of Parkinson's disease [45], [46]. The exact correlation between cardiomyocytes and neurodegenerative diseases remains unclear. However, the possible connection between the two as shown in previous studies and this project set the groundwork for possible discoveries in the future.

Previous studies have also investigated the link between neurodegenerative diseases and mitochondrial pathways. This relationship is relevant to this project because of the significant amount of expression in mitochondrial pathway genes. One study suggests that damage to the mitochondrial oxidative process is a major change that typically happens in the later stages of neurodegenerative diseases [47], Apoptosis, or programmed cell death, are controlled by mitochondria and often modified in neurodegenerative diseases, thus leading to neurodegeneration [48]

Cardiomyocytes and neurodegenerative diseases seem to be linked through the close connections of the cardiovascular and nervous system. The interactions between the different pathways presented in the project are complex, but consistent. Cardiomyocytes in microgravity environments may have differential expression of certain genes and pathways, which could be related to accelerating the onset of some neurodegenerative diseases.

V. CONCLUSION

This project implemented the NASA RCP to analyze the GLDS-258 dataset, which studies hiPSC-CMs in a microgravity environment. There is an available Nextflow implementation of the RCP for *Mus musculus* datasets, and this project briefly explains what changes are required to run a *Homo Sapiens* dataset through the Nextflow implementation of the RPC. Analysis was completed on output data from running the pipeline manually since the Nextflow implementation is not official and still in development. The pipeline was run on the College of Engineering HPC, which provides sufficient resources for the computational tasks required. Most RNA-Seq

experiments have millions of base pairs to sequence, and therefore, will also require immense computing power.

Overall, the quality of the dataset was good, except for one sample when compared to the others 18. Even so, the quality of the worst sample was still acceptable and therefore the experiment was able to move forward. The next steps of the pipeline, STAR alignment and RSEM quantification, also produced good quality data. Thus, we were able to confidently run DESeq2 analysis was able to create a matrix of differentially expressed genes ready to be analyzed in gene set analysis.

Two types of gene set analysis were performed, including DAVID and GSEA. DAVID provided insight into the clusters of genes that were most enriched. GSEA provided information on upregulated and downregulated hallmark gene sets. There is a large amount of significant hallmark data sets in this project that have yet to be analyzed in depth and will require future study to fully understand.

When comparing to a previous study completed regarding GLDS-258, this study was able to have similar findings regarding significant expression of mitochondrial pathways. This project was also able to inspect some new information not mentioned in the previous study regarding neurodegenerative diseases. Previous works have touched on the relationship between the cardiovascular system and the nervous system, but no in-depth mechanisms have yet to be determined when relating pathways from the two systems. This project also explores the relationship between regulating mitochondrial pathways, and how they may affect neurodegenerative diseases. Overall, there seems to be an extremely complex web of interactions and gene expression cascades which can affect multiple parts of the human body.

This project was the very first *Homo Sapiens* dataset run through the RCP. Hopefully, this study can provide more insight into future analysis of *Homo Sapiens* datasets on the RCP, as well as possible links between the cardiovascular system and the nervous system. Analyzing the datasets from NASA GeneLab is an essential part to understanding how the human body works on Earth and in space. With more people traveling to space in the coming years, it is important to remain vigilant in trying to protect our astronauts' safety while they are in space. This effort will hopefully allow for increased space travel and more questions answered about the limitations of the human body for years to come.

REFERENCES

[1] "Touchdown! NASA's Mars Perseverance Rover Safely Lands on Red Planet – NASA's Mars Exploration Program." https://mars.nasa.gov/news/8865/touchdown-nasas-mars-perseverance-rover-safely-lands-on-red-planet/ (accessed Mar. 09, 2021).

[2] M. Whiting, "5 Hazards of Human Spaceflight," *Space Dly.*, Sep. 2018, Accessed: Mar. 10, 2021. [Online].

[3] S. Furukawa *et al.*, "Space Radiation Biology for 'Living in Space,'" *BioMed Res. Int.*, p. 25.

[4] G. Rea *et al.*, "Microgravity-driven remodeling of the proteome reveals insights into molecular mechanisms and signal networks involved in response to the space flight environment," *J. Proteomics*, vol. 137, pp. 3–18, Mar. 2016, doi: 10.1016/j.jprot.2015.11.005.

[5] E. A. Woodcock and S. J. Matkovich, "Cardiomyocytes structure, function and associated pathologies," *Int. J. Biochem. Cell Biol.*, vol. 37, no. 9, pp. 1746–1751, Sep. 2005, doi: 10.1016/j.biocel.2005.04.011.

[6] "Cardiomyocytes - The Cardio Research Web Project." http://www.cardio-research.com/cardiomyocytes (accessed Apr. 23, 2021).

[7] F. E. Garrett-Bakelman *et al.*, "The NASA Twins Study: A multidimensional analysis of a year-long human spaceflight," *Science*, vol. 364, no. 6436, Apr. 2019, doi: 10.1126/science.aau8650.

[8] N. M. Arzeno, M. B. Stenger, J. J. Bloomberg, and S. H. Platts, "Spaceflight-induced cardiovascular changes and recovery during NASA's Functional Task Test," *Acta Astronaut.*, vol. 92, no. 1, pp. 10–14, Nov. 2013, doi: 10.1016/j.actaastro.2012.05.023.

[9] D. B. Thomason, P. R. Morrison, V. Oganov, E. Ilyina-Kakueva, F. W. Booth, and K. M. Baldwin, "Altered actin and myosin expression in muscle during exposure to microgravity," *J. Appl. Physiol. Bethesda Md 1985*, vol. 73, no. 2 Suppl, pp. 90S-93S, Aug. 1992, doi: 10.1152/jappl.1992.73.2.S90.

[10] M. K. Connor and D. A. Hood, "Effect of microgravity on the expression of mitochondrial enzymes in rat cardiac and skeletal muscles," *J. Appl. Physiol. Bethesda Md 1985*, vol. 84, no. 2, pp. 593–598, Feb. 1998, doi: 10.1152/jappl.1998.84.2.593.

[11] "GeneLab Project Overview | NASA GeneLab." https://genelab.nasa.gov/overview (accessed Mar. 10, 2021).

[12] E. G. Overbey *et al.*, "NASA GeneLab RNA-Seq Consensus Pipeline: Standardized Processing of Short-Read RNA-Seq Data," Bioinformatics, preprint, Nov. 2020. doi: 10.1101/2020.11.06.371724.

[13] J. Oribello, "Differential Gene Expression Analysis of Rodents Exposed to Long-Term Space Flight and Insights into Physiological Effects," Master of Science in Bioinformatics, San Jose State University, San Jose, CA, USA, 2021.

[14] J. Wu, "Effects of Spaceflight on Human Induced Pluripotent Stem Cell-Derived Cardiomyocyte Structure and Function." NASA GeneLab, Nov. 07, 2019, doi: 10.26030/GFEW-5417.

[15] "GRCh38 - hg38 - Genome - Assembly - NCBI." https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.26/ (accessed Mar. 10, 2021).

[16] "Index of /pub/release-85/gtf/homo_sapiens/." http://ftp.ensembl.org/pub/release-85/gtf/homo_sapiens/ (accessed Mar. 10, 2021).

[17] "HPC Cluster." http://coe-hpc-web.sjsu.edu/ (accessed Apr. 08, 2021).

[18] P. Ewels, M. Magnusson, S. Lundin, and M. Käller, "MultiQC: summarize analysis results for multiple tools and samples in a single report," *Bioinformatics*, vol. 32, no. 19, pp. 3047–3048, Oct. 2016, doi: 10.1093/bioinformatics/btw354.

[19] "FastQC: A Quality Control tool for High Throughput Sequence Data." https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed Apr. 08, 2021).

[20] "Babraham Bioinformatics - Trim Galore!" https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/ (accessed Apr. 08, 2021).

[21] A. Dobin *et al.*, "STAR: ultrafast universal RNA-seq aligner," *Bioinformatics*, vol. 29, no. 1, pp. 15–21, Jan. 2013, doi: 10.1093/bioinformatics/bts635.

[22] B. Li and C. N. Dewey, "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome," *BMC Bioinformatics*, vol. 12, no. 1, p. 323, Dec. 2011, doi: 10.1186/1471-2105-12-323.

[23] "estimateDispersions function - RDocumentation." https://www.rdocumentation.org/packages/DESeq2/versions/1.12.3/topics/estimateDispersions (accessed Apr. 21, 2021).

[24] Sunyoung Kwon, Seunghyun Park, Byunghan Lee, and Sungroh Yoon, "In-depth analysis of interrelation between quality scores and real errors in illumina reads," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Osaka, Jul. 2013, pp. 635–638, doi: 10.1109/EMBC.2013.6609580.

[25] "GSEA | MSigDB | MSigDB Collections." http://www.gsea-msigdb.org/gsea/msigdb/collections.jsp (accessed Apr. 21, 2021).

[26] "Slurm Workload Manager - sbatch." https://slurm.schedmd.com/sbatch.html (accessed Apr. 22, 2021).

[27] L. Wang, S. Wang, and W. Li, "RSeQC: quality control of RNA-seq experiments," *Bioinformatics*, vol. 28, no. 16, pp. 2184–2185, Aug. 2012, doi: 10.1093/bioinformatics/bts356.

[28] T. J. Treangen and S. L. Salzberg, "Repetitive DNA and next-generation sequencing: computational challenges and solutions," *Nat. Rev. Genet.*, vol. 13, no. 1, pp. 36–46, Jan. 2012, doi: 10.1038/nrg3117.

[29] G. Deschamps-Francoeur, J. Simoneau, and M. S. Scott, "Handling multi-mapped reads in RNA-seq," *Comput. Struct. Biotechnol. J.*, vol. 18, pp. 1569–1576, 2020, doi: 10.1016/j.csbj.2020.06.014.

[30] S. Schaarschmidt, A. Fischer, E. Zuther, and D. K. Hincha, "Evaluation of Seven Different RNA-Seq Alignment Tools Based on Experimental Data from the Model Plant Arabidopsis thaliana," *Int. J. Mol. Sci.*, vol. 21, no. 5, p. 1720, Mar. 2020, doi: 10.3390/ijms21051720.

[31] D. Sarantopoulou, T. G. Brooks, S. Nayak, A. Mrcela, N. F. Lahens, and G. R. Grant, "Comparative evaluation of full-length isoform quantification from RNA-Seq," Bioinformatics, preprint, Jul. 2019. doi: 10.1101/698605.

[32] T. Wang, B. Li, C. E. Nelson, and S. Nabavi, "Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data," *BMC Bioinformatics*, vol. 20, no. 1, p. 40, Dec. 2019, doi: 10.1186/s12859-019-2599-6.

[33] "KEGG PATHWAY Database." https://www.genome.jp/kegg/pathway.html (accessed Apr. 22, 2021).

[34] M. Moreno-Villanueva, M. Wong, T. Lu, Y. Zhang, and H. Wu, "Interplay of space radiation and microgravity in DNA damage and DNA damage response," *Npj Microgravity*, vol. 3, no. 1, p. 14, Dec. 2017, doi: 10.1038/s41526-017-0019-7.

[35] R. Kumari, K. P. Singh, and J. W. DuMond, "Simulated microgravity decreases DNA repair capacity and induces DNA damage in human lymphocytes," *J. Cell. Biochem.*, vol. 107, no. 4, pp. 723–731, Jul. 2009, doi: 10.1002/jcb.22171.

[36] Y. Chaban, E. J. Boekema, and N. V. Dudkina, "Structures of mitochondrial oxidative phosphorylation supercomplexes and mechanisms for their stabilisation," *Biochim. Biophys. Acta BBA - Bioenerg.*, vol. 1837, no. 4, pp. 418–426, Apr. 2014, doi: 10.1016/j.bbabio.2013.10.004.

[37] V. Nogueira and N. Hay, “Molecular Pathways: Reactive Oxygen Species Homeostasis in Cancer Cells and Implications for Cancer Therapy,” *Clin. Cancer Res.*, vol. 19, no. 16, pp. 4309–4314, Aug. 2013, doi: 10.1158/1078-0432.CCR-12-1424.

[38] Z. Liu, T. Zhou, A. C. Ziegler, P. Dimitrion, and L. Zuo, “Oxidative Stress in Neurodegenerative Diseases: From Molecular Mechanisms to Clinical Applications,” *Oxid. Med. Cell. Longev.*, vol. 2017, pp. 1–11, 2017, doi: 10.1155/2017/2525967.

[39] V. Van Giau, E. Bagyinszky, S. S. An, and S. Kim, “Role of apolipoprotein E in neurodegenerative diseases,” *Neuropsychiatr. Dis. Treat.*, p. 1723, Jul. 2015, doi: 10.2147/NDT.S84266.

[40] O. A. Abiodun and M. S. Ola, “Role of brain renin angiotensin system in neurodegeneration: An update,” *Saudi J. Biol. Sci.*, vol. 27, no. 3, pp. 905–912, Mar. 2020, doi: 10.1016/j.sjbs.2020.01.026.

[41] J.-W. Rhee and J. C. Wu, “Cardiac Cell Cycle Activation as a Strategy to Improve iPSC-Derived Cardiomyocyte Therapy,” *Circ. Res.*, vol. 122, no. 1, pp. 14–16, Jan. 2018, doi: 10.1161/CIRCRESAHA.117.312287.

[42] B. N. Dugger and D. W. Dickson, “Pathology of Neurodegenerative Diseases,” *Cold Spring Harb. Perspect. Biol.*, vol. 9, no. 7, p. a028035, Jul. 2017, doi: 10.1101/cshperspect.a028035.

[43] B. J. Critchley, M. Isalan, and M. Mielcarek, “Neuro-Cardio Mechanisms in Huntington’s Disease and Other Neurodegenerative Disorders,” *Front. Physiol.*, vol. 9, p. 559, May 2018, doi: 10.3389/fphys.2018.00559.

[44] Y. Takamatsu *et al.*, “Protection against neurodegenerative disease on Earth and in space,” *Npj Microgravity*, vol. 2, no. 1, p. 16013, Dec. 2016, doi: 10.1038/npjmgrav.2016.13.

[45] S. Orimo *et al.*, “Axonal -synuclein aggregates herald centripetal degeneration of cardiac sympathetic nerve in Parkinson’s disease,” *Brain*, vol. 131, no. 3, pp. 642–650, Feb. 2008, doi: 10.1093/brain/awm302.

[46] S. Orimo *et al.*, “Degeneration of Cardiac Sympathetic Nerve Begins in the Early Disease Process of Parkinson?s Disease,” *Brain Pathol.*, vol. 17, no. 1, pp. 24–30, Jan. 2007, doi: 10.1111/j.1750-3639.2006.00032.x.

[47] P. H. Reddy, “Role of Mitochondria in Neurodegenerative Diseases: *Mitochondria as a Therapeutic Target in Alzheimer’s Disease*,” *CNS Spectr.*, vol. 14, no. S7, pp. 8–13, Aug. 2009, doi: 10.1017/S1092852900024901.

[48] Y. Wu, M. Chen, and J. Jiang, “Mitochondrial dysfunction in neurodegenerative diseases and drug targets via apoptotic signaling,” *Mitochondrion*, vol. 49, pp. 35–45, Nov. 2019, doi: 10.1016/j.mito.2019.07.003.