Spaceflight and Differential Gene Expression Analysis of Mice Quadriceps Exposed to Microgravity

Tommy Nguyen
San Jose State University

Follow this and additional works at: https://scholarworks.sjsu.edu/etd_projects

Part of the Bioinformatics Commons

Recommended Citation
DOI: https://doi.org/10.31979/etd.4tt8-8a9h
https://scholarworks.sjsu.edu/etd_projects/1053

This Master's Project is brought to you for free and open access by the Master's Theses and Graduate Research at SJSU ScholarWorks. It has been accepted for inclusion in Master's Projects by an authorized administrator of SJSU ScholarWorks. For more information, please contact scholarworks@sjsu.edu.
Spaceflight and Differential Gene Expression Analysis of Mice Quadriceps Exposed to Microgravity

A Project

Presented to

Department of Computer Science

San Jose State University

In Partial Fulfillment

Of the Requirements for the Degree

By

Tommy Nguyen

December, 2021
ABSTRACT

The trip to Mars and back is planned for the next 20 years. Improvement in technology and research has allowed data analysis, at a larger scale, on spaceflight specimen. However, research involving spaceflight is decentralized, as research is spread across laboratories with different methodologies. NASA’s GeneLab is a public repository for spaceflight-related omics data and promotes centralizing spaceflight RNA-Seq studies using NASA’s RNA-Seq pipeline. Jonathan Oribello, a SJSU’s Bioinformatics graduate and now an employee at GeneLab, has implemented Nextflow to NASA’s RNA-Seq pipeline. The Nextflow RNA-Seq pipeline was ran on San Jose State University’s College of Science High Performance Computing system, using study GLDS-103 from GeneLab. The results have shown the reproducibility of Nextflow, and new insights through customizable inputs. This project has shown possible insights to the skeletal muscle system, and the effects of microgravity on the circadian rhythm.

Keywords: Skeletal Muscle, Circadian Rhythm, Microgravity, Nextflow
I would like to thank the following people for their support:

- My advisor, Dr. Heller, for showing me the topic of spaceflight as my Master’s Project and guiding me through the process
- Dr. Andreopoulos and Dr. Lee for being great professors in the Bioinformatics program
- Jonathan Oribello for being available to answer questions regarding the pipeline
- The Bioinformatics program for being a welcoming community
- My family for being supportive of my education career
# TABLE OF CONTENTS

I. INTRODUCTION 8

A. Spaceflight Endeavors 8

B. Skeletal Muscle: a Brief Review 8

C. Studies Related to Spaceflight and Skeletal Muscle 9

D. NASA’s GeneLab and RNA-Seq Pipeline 10

E. GLDS-103 11

II. METHODS 12

A. Running the Nextflow RCP 12
   1) Nextflow 12
   2) Data Preprocessing 13
   3) Data Processing 14
   4) Differential Gene Expression Calculation 14

B. Gene Set Analysis 15
   3) Leading Edge Analysis 16
   4) Visualization of Gene Ontology Biological Processes by Cytoscape’s Enrichment Map 16

III. RESULTS 16

A. Nextflow RNA-Seq Pipeline 16

B. Quality Control Metrics 17

C. Alignment Metrics 19

D. DESeq2 analytical plots 20

E. Gene Set Analysis by DAVID and GSEA 22
   1) DAVID Gene Annotation Clusters 22
   2) Gene Set Enrichment Analysis 22

IV. DISCUSSION 29

A. Results for GLDS-103 29
   1) Reproducibility of Nextflow 29
   2) NRCP and Resource Requirements 29
3) Gene Set Analysis
   a) DAVID gene annotation clusters
   b) Investigation of upregulated hallmark gene sets
   c) Investigation of the downregulated hallmark gene sets
   d) Insights to GO biological processes terms through the C5 GO BP subset

V. CONCLUSION

REFERENCES

APPENDIX: SCRIPTS
TABLE OF FIGURES

Figure 1: A single mature muscle fiber, shown as a bundle of myofibrils, encased by the sarcolemma from [11].

Figure 2: NASA's consensus RNA-Seq pipeline from [28].

Figure 3: GitHub repository for NRCP under the user J-81 (Jonathan Oribello).

Figure 4: (A) Status panel from Nextflow Tower (B) Processes Panel from Nextflow Tower.

Figure 5: Venn diagram of genes with a p-value of less than 0.05. Non-overlapping regions are not to scale.

Figure 6: Total Read Counts for Each Sample.

Figure 7: MultiQC Mean Quality Score Plots (A) Untrimmed reads (B) Trimmed reads.

Figure 8: Star Alignment Metrics.

Figure 9: RSEM Counting Metrics.

Figure 10: Principal Component Analysis Plots of (A) NRCP and (B) NASA.

Figure 11: Dispersion plot. Black dots represent genes. Black dots encircled in a blue outline are outliers.

Figure 12: MA Plot, where every point represents a single gene. Points colored blue are significant under the Benjamini-Hochberg method (adjusted p-value < 0.05).

Figure 13: GO Biological Processes Gene Set Collection Enrichment Map with ‘Moderately conservative’ parameters. Red nodes indicate upregulation and blue nodes indicate downregulation. Connections (edges) mean overlap of significant genes shared between gene sets.

Figure 14: GO Biological Processes Gene Set Collection Enrichment Map with ‘Very conservative’ parameters only showing blue nodes.

Figure 15: GO Biological Processes Gene Set Collection Enrichment Map with ‘Very conservative’ parameters only showing red nodes.

Figure 16: Plot of amount of time elapsed for each process.

Figure 17: Genes that are prevalent in all annotated terms of cluster 2.

9
11
12
13
17
18
18
19
20
21
21
22
27
27
27
28
29
31
INDEX OF TABLES

Table I STATS OF NRCP.................................................................................................................................................. 17
Table II TOP THREE ENRICHED DAVID GENE ANNOTATION CLUSTERS ............................................................... 23
Table III GSEA UPREGULATED HALLMARK GENE SET .............................................................................................. 24
Table IV GSEA DOWNREGULATED HALLMARK GENE SET ....................................................................................... 25
Table V PREVALENT GENES FROM LEADING EDGE ANALYSIS ON 637 GO BP GENE SETS .............................. 26
I. INTRODUCTION

A. Spaceflight Endeavors

Exploring outer space is an endeavor for humankind. In 2017, President Trump issued a memorandum called “White House Space Policy Directive-1” [1]. This directive sets goals for humanity’s return to the Moon in 2024: to establish payloads on the surface, to assemble a Gateway outpost in orbit, and demonstrate the first lunar landing since 1972 [2]. In response to the directive, The National Aeronautics and Space Administration (NASA) created the Artemis Program that has two phases [3]. Phase 1 details the return to the Moon by landing the first woman and man at the lunar South Pole by 2024, and setting up two Gateway components, the Power and Propulsion Element (PPE), and the Habitation and Logistics Outpost (HALO), both launching in 2022 and 2023 respectively. Phase 2 advances technology for sustainability during space exploration—allowing for longer missions. Due to the COVID pandemic, NASA is expecting delays up to several years from the planned 2024 year [4]. Even with the projected delay, the excitement of new technology and research is never ending.

Taking humans to Mars and beyond has been in the scope for years and is planned for the next 20 years [5]. The trip to Mars and back would take several years, with hazards such as radiation, isolation, microgravity, confinement, and nutrition being concerns during long-duration space missions [6]. Two hazards, space radiation and microgravity, are heavily studied and are the two primary hazards preventing long-duration spaceflights [7]. Radiobiological research link exposure to radiation to growth of cancer, dysfunctional central nervous system, and break down of tissues [8]. Exposure to microgravity is known to affect cardiovascular, musculoskeletal, bone metabolic, and immune systems [9]. Understanding these hazards will provide solutions for astronauts during their trip to Mars and other long-duration spaceflights. With the rapid advancement of omics technologies, opens to higher volume of data analysis on spaceflight specimens. NASA’s Rodent Research (RR) aims to provide a better understanding of these hazards and offer solutions to counteract the negative effects [10]. In this project, data taken from NASA’s RR will be approached statistically differently, with the intent to provide more exploration and discoveries regarding microgravity and skeletal muscle.

B. Skeletal Muscle: a Brief Review

Skeletal muscle is a type of striated muscle which is composed of cells that require high energy [11]. This muscle type is the main driving force of contraction and movement. As depicted in Figure 1, skeletal muscle is an organized tissue, with bundles called muscle fibers (myofibers) that form bundles of fascicles, that of which form the muscle tissue. Myofibers are made of sarcomeres, units that make up the myofibers. Contraction of sarcomeres are important for force generation and rapid movement. The repeating units of sarcomeres are composed of actin and myosin and during contraction, Ca2+ is released from the sarcoplasmic reticulum—activating the actin-myosin cross bridging and contraction [12]. Skeletal muscle is
important for efficient energy production and cellular homeostasis, two processes that are part of metabolic and regulatory systems. Disruptions, by genetic or environmental factors, to the skeletal tissue results in lower muscle health and function (caused by muscle fiber loss) and probable death.

**Figure 1**: A single mature muscle fiber, shown as a bundle of myofibrils, encased by the sarcolemma from [11].

**C. Studies Related to Spaceflight and Skeletal Muscle**

The musculoskeletal system in relation to spaceflight have shown to produce many consequences. Microgravity decreases muscle mass, force, and shifts the muscle fiber type to a swifter contracting but less durable phenotype [13], [14]. In detail, inactive skeletal muscle during microgravity induce atrophy while active skeletal muscle during microgravity undergo hypertrophy [15], [16]. On average, 20% of inactive skeletal muscle mass is lost under normal conditions while 30%, on average, of skeletal muscle mass is lost under microgravity [17]. This loss of skeletal muscle mass is enough to hinder astronauts, preventing spacewalks, and doing repairs on the International Space Station (ISS). Therefore, studying the cellular and molecular mechanisms of skeletal muscle atrophy caused by microgravity will provide new ways to counteract the negative effects of microgravity. One new way to study the effects of microgravity on skeletal muscle is associating both bone and skeletal muscle systems [18].
These studies involving both bone and skeletal muscle tissue will allow full comprehension of the negative effects of microgravity. Therefore, this project will seek out biological processes involved in either bone or skeletal muscle tissue.

In vitro studies have shown that microgravity directly affects skeletal muscle cell differentiation and metabolism, allowing skeletal muscle cells to be potential countermeasure targets [16], [19], [20]. Additionally, studies involving Ca2+ signaling suggest that improper signaling of Ca2+ ions in skeletal muscle tissue leads to lower activity of myogenesis and skeletal muscle metabolism [21], [22]. For in vivo studies, astronauts exposed by microgravity have shown loss skeletal muscle mass and strength, mainly in areas from the hip to the toes and the torso [12]. Furthermore, in 6-9 days under microgravity conditions, skeletal muscle loss occurred rapidly in the quadriceps, gastrocnemius, and posterior back muscle, with values of 6%, 6%, and 10% respectively [23]–[25]. Longer duration missions of 6 months resulted to even greater muscle loss, with values of 12%, 13%, and 20% for the quadriceps, soleus-gastrocnemius, and posterior back skeletal muscles respectively [25]–[27]. The data suggests that muscle loss stabilizes after 3-4 months of exposure to microgravity and this loss is because of lower protein synthesis than degradation of the skeletal muscle tissue [25], [28]. Both in vitro and in vivo studies have shown substantial evidence of skeletal muscle loss and possible ways to alleviate microgravity conditions for both short and long duration missions. Further investigation into proteins and biological processes related to skeletal muscle tissue will open for more possibilities.

D. NASA’s GeneLab and RNA-Seq Pipeline

NASA’s GeneLab is a public database of space related studies and its goal is to collect data from biospecimens in both space and simulated space stressor conditions, while providing access to omics data [29]. GeneLab currently has 340 studies and are accessible through the home page https://genelab.nasa.gov/ [30]. NASA plans to centralize RNA-Seq data by minimizing variations through its RNA-Seq consensus pipeline (RCP). It is reasonable for NASA to encourage RNA-Seq data to follow a consensus pipeline since different software tools produce different results [31]–[33]. As depicted in Figure 2, RCP is split into three parts: Data Preprocessing, Data Processing, and Differential Gene Expression Calculation. A recent San Jose State University (SJSU) Bioinformatics graduate, Jonathan Oribello, has taken NASA’s RCP and implemented Nextflow, a workflow management system that allows for reproducibility [34]. Jonathan now works with NASA’s GeneLab and has continued to work on RCP with Nextflow for reproducibility. The software is hosted on GitHub and is found at https://github.com/J-81/Nextflow_RCP [35]. The Nextflow RCP (NRCP) was used for data processing in this project.
The study of interest, GLDS-103, was taken from GeneLab and a part of NASA’s RR project [36]. 12 mice of similar age, sex, and strain were used in study GLDS-103, where 6 of the mice were sent to the ISS and exposed to microgravity for 37 days. The other 6 were used for ground control and had matching ISS environmental conditions. Mice flown to the ISS were the experimental control group of this study, and are labeled as FLT, while ground control mice are labeled as GC, the control group of the study. FLT mice had their tissues extracted, with the left quadriceps muscle as the main muscle of study and transported back to Earth. The construction of the RNA library was done at UC Davis Genome Center, using Illumina TruSeq standard RNA library prep kit. The Illumina HiSeq 4000 platform was used on the RNA library, with 30 million reads per sample and sequenced by paired end 150 bp reads. The resulting raw data files are hosted under GLDS-103’s repository and are used in this project.
II. METHODS

A. Running the Nextflow RCP

NRCP was run on SJSU’s College of Science (CoS) High Performance Computing (HPC) system, ideal for large data processing due to its large data storage of 110 TB. SJSU’s CoS HPC system uses a resource scheduling program called SLURM, allowing users to submit batch jobs [37]. In Appendix: Scripts, Script 1 was used in this project for HPC. NRCP was ran on July 15th, 2021, of which the version of NRCP was last committed on July 11, 2021. Since the date of which NRCP was ran on for this project, several commits were done, with the last commit done on October 4, 2021. Figure 3 shows the publicly available NRCP and its contents on GitHub. The bin folder contains scripts that runs the various tools used throughout the pipeline, of which data preprocessing and data processing are written in Python and differential gene expression calculation is written in R [38], [39].

![GitHub repository for NRCP under the user J-81 (Jonathan Oribello)](image)

Figure 3: GitHub repository for NRCP under the user J-81 (Jonathan Oribello)

1) Nextflow

In previous SJSU Bioinformatics Master projects, a configuration file was needed to run Nextflow on CoS-HPC in order to properly extract data [40]–[42]. However, as of July 15, 2021, a configuration file is no longer needed. NRCP automatically retrieves necessary data found in GeneLab and ENSEMBL [43]. Version GRCh39 for the genome file and annotation file were used for reference assembly. Running NRCP requires a Linux environment with Nextflow and Conda installed. Connecting to the CoS-HPC requires access through a SSH client and a VPN. PuTTY
version 0.74, a SSH client, and Cisco AnyConnect Secure Mobility Client version 4.10.03104, a VPN, were both used to connect to the CoS-HPC [44], [45]. Script 1 was submitted as a job on CoS-HPC, with lines 13 and 14 tokenizing the job for Nextflow Tower and line 16 running NRCP. Nextflow Tower allows the user to observe the current state of the job and current processes that are running as shown in Figure 4. Other stats, such as run time and total memory, are also available on Nextflow Tower.

![Figure 4: (A) Status panel from Nextflow Tower (B) Processes Panel from Nextflow Tower](image)

2) **Data Preprocessing**

Data preprocessing is part of the first step of the NRCP pipeline, where raw RNA sequence data is checked for quality. Additionally, the raw RNA sequence data will be trimmed—filtering bad reads along with adaptor removal. FastQC version 0.11.9 quality control checks the raw FASTQ files from GLDS-103 [46]. MultiQC version 1.7 compiles the FastQC logs to allow better visualization of all raw FASTQ files [47]. Trim Galore version 0.6.6 and Cutadapt version 2.6 filters out bad reads (Phred scores under 20) and trims out adaptors from Illumina TruSeq [48]. The NRCP can identify between single and paired end reads and because paired end reads were used, both Trim Galore and Cutadapt have trimmed the data accordingly.
3) Data Processing

Data processing is the second step of the NRCP. STAR version 2.7.9a is a tool that aligns the trimmed reads from step 1 to the reference genome but requires the reference FASTA file and GTF file [49]. The necessary files were obtained from ENSEMBL version 104 [43]. After the reference genome has been indexed, using the files from ENSEMBL, the trimmed reads can now be mapped. RSEM version 1.3.1 was used to quantify the number of reads mapped to each annotated transcript and gene [50]. After RSEM has finished quantifying the samples, an unnormalized counts table is produced, containing the 12 samples and genes that were found. MultiQC version 1.7 was used to compile the log files produced from STAR alignment results.

4) Differential Gene Expression Calculation

The final step in pipeline is where differential gene expression (DGE) is calculated. The bin folder contains the R script file, _normcounts_noERCC_DGE_vis_ISA.R, which runs the DeSEQ2 version 1.30 R package on R version 4.0.3 [39], [51]. The R script file creates a DESeqDataSet object, which will be manipulated later in the script. Before DESeq2 is run on the RSEM count data, all genes with a sum of less than 10 counts from all samples are removed. When running DESeq2 after filtering out genes, three functions are called in this order on the DESeqDataSet object, estimateSizeFactors(), estimateDispersions(), and nbionmWaldTest().

estimateSizeFactors() estimates the size factors using a median ratio method, and a size factor is created for every sample by dividing the median ratio of all counts by the mean of the genes from all samples.

estimateDispersions() takes dispersion estimates of the data by using a negative binomial distribution (DESeq2 also assumes that the data follows a negative binomial distribution, thus the usage of the function). The dispersion per gene is found optimizing the Cox-Reid adjusted profile likelihood [51]. The dispersion parameter is the relationship between the variance of the observed count and its respective mean value.

Lastly, nbionmWaldTest(), which tests for significance of coefficients by calculated ‘sizeFactors’ and dispersion estimates. Wald Test calculates the statistical significance of a gene’s expression by producing a p-value. The raw count and normalized data are extracted into objects and then exported.

Adjusted p-values are also calculated using the Benjamini-Hochberg (BH) method, a method that accounts for the multiple testing problem by a set FDR value. The FDR value was 0.1, of which the p-values for genes that have passed the filter will be calculated as an adjusted p-value using the BH method.

5. Mapping gene ID

After DGE calculation is complete, an annotation database is built, of which a “keytype” is defined, indicating the primary type of annotation. By default, GeneLab uses ENSEMBL as its
keytype; therefore, ENSEMBL is then used to map annotations of the organism through Bioconductor AnnotationDbi version 1.52.0 package. Additional annotation columns from the package are added to the annotation database: SYMBOL, GENENAME, ENSEMBL, REFSEQ, and ENTREZID. Other annotation columns such as, STRING, and GOSLIM are also added into the annotation columns using STRINGDB version 2.2.0 and PANTHERdb version 1.0.10 R packages. After the annotation database is completed, the means and standard deviations of the normalized count data for each gene is calculated for all samples, and again but for samples within each respective group. The final DGE output table is ready to be assembled, combining all the data together in a .csv file.

B. Gene Set Analysis

To analyze every gene from the DGE table is time consuming, an alternative and powerful method is to analyze genes in gene sets, or groups of genes that share biological function, cellular location, or regulation [52]. GSEA and DAVID are tools used to perform gene set analysis, in which Leading Edge Analysis is another analytical tool in GSEA that can help offer insight on what genes are prevalent in gene sets.

1) DAVID

DAVID version 6.8 functional annotation clustering was used to analyze genes with adjusted p-value under 0.05, of which 773 genes were used [53], [54]. The tool generates gene set clusters based on enriched gene sets that have overlapping significant genes. This allows for insight into sets of annotations with a common theme.

2) GSEA

Gene set enrichment analysis (GSEA) version 4.1.0 was used to compare the total gene list produced from NRCP to two gene set collections in MsigDB version 7.4: hallmark and gene ontology biological processes (GOBP). Hall mark gene sets are excellent as a comparison since they are well-defined biological states or processes. A subset of the C5 (ontology gene sets) gene set collection, the GOBP gene sets, will be used as a comparison to explore significant biological processes. 22,211 genes were produced from NRCP and extracting only the ENSEMBL identifiers into a .txt file, of which is used as the expression dataset. Because each condition contains less than 7 samples, and that the experiment is comparing only two conditions, it was recommended to run the “Permutation type” parameter as gene_set, randomizing the gene set for each permutation instead of the phenotype label [52]. However, this causes a less stringent assessment of significance and users would want to use a stringent FDR cutoff such as 0.05 [52]. Due to GLDS-103 using mouse as the model organism, Mouse_ENSEMBL_Gene_ID_Human_Orthologs_MSigDB.v.7.4.chip was used as the Chip platform, converting the ENSEMBL gene identifiers to their respective human ortholog. The
expression data set was then ranked by Signal2Noise, the default ranking method by GSEA, and was recommended by Anthony Castanza, a member of the MsigDB team, since each group had more than 3 samples [55]. Figure 5 depicts the Signal2Noise calculation, where $\mu$ is the mean and $\sigma$ is the standard deviation. $\sigma$ has a minimum value of 2 times the absolute value of $\mu$ and $\mu$ is adjusted to 1 if $\mu$ is 0. As the signal-to-noise ratio increases, the difference of means increases, meaning, the more distinct the gene expression is in each phenotype [52]. The FDR threshold is less than 0.25 and was used for both runs.

$$\frac{\mu_A - \mu_B}{\sigma_A + \sigma_B}$$

Figure 5: Signal2Noise calculation, where subscript A is phenotype 1 and subscript B is phenotype 2 from [52]

3) **Leading Edge Analysis**

Leading Edge Analysis (LEA) was used for the C5 GOBP subset gene set collection, in comparison to the 22,211 genes. Gene sets with nominal p-values under 0.05 were taken, for a total of 637 significant gene sets.

4) **Visualization of Gene Ontology Biological Processes by Cytoscape’s Enrichment Map**

Cytoscape version 3.9.0 was used to create a visualization through the EnrichmentMap version 3.3.3 app. GSEA allows to automate the inputs for the EnrichmentMap app by using GSEA’s Enrichment Map Visualization tool. The only parameters needed to create the enrichment map are p-value, FDR q-value, and overlap coefficients, of which values of 0.05, 0.25, and 0.5 were used and were considered ‘Very permissive’ as stated by the EnrichmentMap app documentation. Conservative parameters were not used since it is recommended that 100-250 enriched terms at the very conservative thresholds. The Enrichment Map Visualization tool with ‘Moderately conservative’ parameters (p-value of 0.005 and FDR value of 0.075) was also ran to showcase the difference between results of the two parameters. Clusters were then annotated through AutoAnnotate version 1.3.4, using WordCloud version 3.1.4, and computing annotations based on the scoring of adjacent words across all the gene set GOBP terms [56], [57].

III. RESULTS

A. **Nextflow RNA-Seq Pipeline**

1) **NRCP run time**

Table 1 shows the run time of NRCP, of which two runs were done since run number 1 encountered an error, where the allotted memory to run STAR alignment was too low (STAR
alignment requires a minimum of 32 GB of memory); therefore, as shown in Script 1, the allotted memory for the job was increased to 120 GB and NRCP was able to run into completion.

<table>
<thead>
<tr>
<th>Run number</th>
<th>Run time</th>
<th>CPU time (hrs)</th>
<th>Total memory (GB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 h 42 m 8 s</td>
<td>15.1</td>
<td>37.20</td>
</tr>
<tr>
<td>2</td>
<td>1 d 17 h 40 m 36 s</td>
<td>240.0</td>
<td>465.45</td>
</tr>
</tbody>
</table>

2) **Comparison of results from NRCP and NASA**

Comparisons of NRCP and NASA was done by taking the expression data set and filtering genes that have a p-value of less than 0.05. This was done because publications for the Rodent Research project used p-value of less than 0.05. NRCP had 51 unique genes while NASA had 80 unique genes. This showcase that NRCP almost reproduces identical results.

![Figure 5: Venn diagram of genes with a p-value of less than 0.05. Non-overlapping regions are not to scale.](image)

**B. Quality Control Metrics**

Figure 6 displays the numbers of reads that were present in the raw FASTQ files. It is expected for RNA-Seq reads to show many duplicate reads. Figure 7 shows both untrimmed and trimmed reads and are labeled as A and B respectively. The compilation by MultiQC shows that there is a disparity between pair reads, of which the reverse reads are worse. However, this is a tendency by reverse reads. The Phred score of the trimmed reads remained above 30, indicating that the quality across the 150 base pair reads was excellent, with improvement toward the last few base pairs. There is a reported bias in the first few bases according to
FastQC. The cause of this bias is the random priming step in library production, however, the potential concerns are not problematic [58].

**Figure 6: Total Read Counts for Each Sample**

![FastQC: Sequence Counts](image)

**Figure 7: MultiQC Mean Quality Score Plots (A) Untrimmed reads (B) Trimmed reads**

![MultiQC Mean Quality Score Plots](image)
C. Alignment Metrics

For STAR alignment (Figure 8), MultiQC reported that at least 78.9% of reads for every sample (including reverse reads) were uniquely mapped to a location. About 12% of reads for every sample were mapped to multiple loci (2-10). About 9% of reads for every sample were unmapped due to alignment lengths being too short. About 0.1% of reads were mapped to too many loci (over 10 loci) and about 0.2% of reads for every sample were unmapped due to other reasons.

For RSEM counts (Figure 9), MultiQC reported that at least 77.1% of reads for every sample were uniquely mapped to a gene. About 10.9% of reads for every sample were mapped to multiple genes and about 12.1% of reads for every sample were unable to be aligned to genes. There were no instances of reads being aligned to too many genes for every sample.

Figure 8: Star Alignment Metrics
D. **DESeq2 analytical plots**

PCA plots is shown in Figure 10, where (A) is NRCP and (B) is NASA. The PCA plot for NRCP was made by first taking the log$_2$ of the normalized count data from DESeq2. The function, `prcomp()`, takes the log$_2$ data as an input and creates the PCA plot. Looking at the plots, both plots show identical variation between the two conditions.

The dispersion plot is a useful diagnostic to show outliers in the expression data set [51]. Outliers are found through several parameters. Gene-wise maximum likelihood estimates (MLE) are obtained using the respective gene data (black dot). A curve is then created (red) to fit the MLEs and to show the overall dispersion-mean dependence across all the genes. Genes are then fitted (understood as shrinkage) of which the gene-wise estimates head towards the consensus line. Black circles encircled by a blue outline are outliers and are not shrunk towards the consensus line. Figure 11 was created using the function `plotDispEsts()` and inputting the object that was saved when DESeq2 is ran on the DESeqDataSet object.

A MA plot was created to show the dispersion of gene expression between the two condition. It is useful to showcase genes that are upregulated in FLT mice and genes that are downregulated in FLT mice. Genes that show a positive log2 fold change value of 1 means that...
the gene is expressed 2 times more in FLT mice than in GC mice and genes that show a negative log2 fold change value of 1 is twice as under expressed in FLT mice than in GC mice. Points shaped as triangle are genes that are beyond the plot’s scale. Figure 12 was created using the plotMA() function and inputting the result object that was created using the results() function.

Figure 10: Principal Component Analysis Plots of (A) NRCP and (B) NASA

Figure 11: Dispersion plot. Black dots represent genes. Black dots encircled in a blue outline are outliers.
E. Gene Set Analysis by DAVID and GSEA

1) DAVID Gene Annotation Clusters

The top three enriched DAVID gene annotation clusters are shown in Table II. Annotation Cluster 1 detail disruptions in biological rhythms, particularly the circadian rhythm and its rhythmic process. Gene regulation regarding the circadian rhythm is also disrupted. Annotation Cluster 2 details disruption of transcription regulation. Annotation Cluster 3 show disruptions of the zinc finger, which the other terms, zinc, and metal-binding, are associated with. Zinc finger is an interesting term found in adjusted p-value gene list and will be investigated further. Annotation Cluster 1, Annotation Cluster 2, and Annotation Cluster 3, have an enrichment score of 8.84, 5.56, and 5.21 respectively.

2) Gene Set Enrichment Analysis

Table III and Table IV are the results from GSEA when comparing all genes to the Hallmark collection gene set with an FDR threshold of 0.25. Table V depicts the 10 most prevalent genes from the 637 significant gene sets when comparing all genes to the C5 GOBP subset. Figure 13 showcases the ‘Moderately conservative’ EnrichmentMap results. Figure 14 and 16 showcase the ‘Very permissive’ results. In total, Figure 14 and 16 displays 188 nodes, 56 of which are not shown due to the node not connected to another node. However, the 56 nodes that are not shown does not mean they are insignificant, and exclusion of these nodes are mainly for display purposes.
<table>
<thead>
<tr>
<th>Category</th>
<th>Term</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Annotation Cluster 1</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UP_KEYWORDS</td>
<td>Biological rhythms</td>
<td>1.26E-09</td>
</tr>
<tr>
<td>GOTERM_BP_DIRECT</td>
<td>GO:0007623~circadian rhythm</td>
<td>1.07E-07</td>
</tr>
<tr>
<td>KEGG_PATHWAY</td>
<td>mmu04710:Circadian rhythm</td>
<td>3.68E-08</td>
</tr>
<tr>
<td>GOTERM_BP_DIRECT</td>
<td>GO:0048511~rhythmic process</td>
<td>1.26E-06</td>
</tr>
<tr>
<td>GOTERM_BP_DIRECT</td>
<td>GO:0032922~circadian regulation of gene expression</td>
<td>1.52E-05</td>
</tr>
<tr>
<td>GOTERM_BP_DIRECT</td>
<td>GO:0042752~regulation of circadian rhythm</td>
<td>0.026716</td>
</tr>
<tr>
<td><strong>Annotation Cluster 2</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UP_KEYWORDS</td>
<td>Repressor</td>
<td>9.14E-08</td>
</tr>
<tr>
<td>UP_KEYWORDS</td>
<td>Transcription regulation</td>
<td>1.28E-07</td>
</tr>
<tr>
<td>UP_KEYWORDS</td>
<td>Transcription</td>
<td>1.29E-07</td>
</tr>
<tr>
<td>GOTERM_BP_DIRECT</td>
<td>GO:0006351~transcription, DNA-templated</td>
<td>4.92E-04</td>
</tr>
<tr>
<td>UP_KEYWORDS</td>
<td>Activator</td>
<td>5.48E-05</td>
</tr>
<tr>
<td>GOTERM_BP_DIRECT</td>
<td>GO:0006355~regulation of transcription, DNA-templated</td>
<td>0.002172</td>
</tr>
<tr>
<td>UP_KEYWORDS</td>
<td>DNA-binding</td>
<td>0.010013</td>
</tr>
<tr>
<td>GOTERM_MF_DIRECT</td>
<td>GO:0003677~DNA binding</td>
<td>0.071981</td>
</tr>
<tr>
<td>GOTERM_MF_DIRECT</td>
<td>GO:0003700~transcription factor activity, sequence-specific DNA binding</td>
<td>0.425386</td>
</tr>
<tr>
<td><strong>Annotation Cluster 3</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UP_KEYWORDS</td>
<td>Metal-binding</td>
<td>1.60E-05</td>
</tr>
<tr>
<td>UP_KEYWORDS</td>
<td>Zinc</td>
<td>2.17E-05</td>
</tr>
<tr>
<td>GOTERM_MF_DIRECT</td>
<td>GO:0008270~zinc ion binding</td>
<td>0.002475</td>
</tr>
<tr>
<td>GOTERM_MF_DIRECT</td>
<td>GO:0046872~metal ion binding</td>
<td>0.002475</td>
</tr>
<tr>
<td>UP_KEYWORDS</td>
<td>Zinc-finger</td>
<td>0.001091</td>
</tr>
</tbody>
</table>
# Table III

<table>
<thead>
<tr>
<th>NAME</th>
<th>ES</th>
<th>NES</th>
<th>NOM p-val</th>
<th>FDR q-val</th>
</tr>
</thead>
<tbody>
<tr>
<td>HALLMARK_INTERFERON_GAMMA_RESPONSE</td>
<td>0.482672</td>
<td>2.187474</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HALLMARK_CHOLESSTEROL_HOMEOSTASIS</td>
<td>0.481399</td>
<td>1.894577</td>
<td>0</td>
<td>0.002039</td>
</tr>
<tr>
<td>HALLMARK_HEDGEHOG_SIGNALING</td>
<td>0.547643</td>
<td>1.875297</td>
<td>0.001786</td>
<td>0.002016</td>
</tr>
<tr>
<td>HALLMARK_INTERFERON_ALPHA_RESPONSE</td>
<td>0.447914</td>
<td>1.835957</td>
<td>0</td>
<td>0.001512</td>
</tr>
<tr>
<td>HALLMARK_COMPLEMENT</td>
<td>0.365104</td>
<td>1.616055</td>
<td>0.001462</td>
<td>0.016717</td>
</tr>
<tr>
<td>HALLMARK_TNFA_SIGNALING_VIA_NFKB</td>
<td>0.35591</td>
<td>1.614682</td>
<td>0.001497</td>
<td>0.013931</td>
</tr>
<tr>
<td>HALLMARK_COAGULATION</td>
<td>0.361799</td>
<td>1.523447</td>
<td>0.00315</td>
<td>0.032807</td>
</tr>
<tr>
<td>HALLMARK_EPIHELIAL_MESENCHYMAL_TRANSITIONN</td>
<td>0.302906</td>
<td>1.376175</td>
<td>0.019802</td>
<td>0.128215</td>
</tr>
<tr>
<td>HALLMARK_MTORC1_SIGNALING</td>
<td>0.297977</td>
<td>1.367554</td>
<td>0.024024</td>
<td>0.123942</td>
</tr>
<tr>
<td>HALLMARK_SPERMATOGENESIS</td>
<td>0.30926</td>
<td>1.289498</td>
<td>0.001786</td>
<td>0.022789</td>
</tr>
<tr>
<td>HALLMARK/MITOTIC_SPINDLE</td>
<td>0.279182</td>
<td>1.281991</td>
<td>0.055394</td>
<td>0.215113</td>
</tr>
<tr>
<td>HALLMARK_ANIOGENESIS</td>
<td>0.37531</td>
<td>1.241098</td>
<td>0.172007</td>
<td>0.270664</td>
</tr>
<tr>
<td>HALLMARK_HYPOXIA</td>
<td>0.273634</td>
<td>1.239432</td>
<td>0.076135</td>
<td>0.25349</td>
</tr>
<tr>
<td>HALLMARK_APICAL_SURFACE</td>
<td>0.338806</td>
<td>1.21013</td>
<td>0.184801</td>
<td>0.297007</td>
</tr>
<tr>
<td>HALLMARK_KRAS_SIGNALING_DN</td>
<td>0.268661</td>
<td>1.183101</td>
<td>0.151917</td>
<td>0.339317</td>
</tr>
<tr>
<td>HALLMARK_KRAS_SIGNALING_UP</td>
<td>0.260746</td>
<td>1.179588</td>
<td>0.15188</td>
<td>0.326272</td>
</tr>
<tr>
<td>HALLMARK_IL6_JAK_STAT3_SIGNALING</td>
<td>0.286122</td>
<td>1.176885</td>
<td>0.177953</td>
<td>0.313153</td>
</tr>
<tr>
<td>HALLMARK_G2M_CHECKPOINT</td>
<td>0.24951</td>
<td>1.14806</td>
<td>0.181159</td>
<td>0.358497</td>
</tr>
<tr>
<td>HALLMARK_ESTROGEN_RESPONSE_LATE</td>
<td>0.251591</td>
<td>1.133096</td>
<td>0.192956</td>
<td>0.37597</td>
</tr>
<tr>
<td>HALLMARK_IL2_STAT5_SIGNALING</td>
<td>0.243075</td>
<td>1.10955</td>
<td>0.226387</td>
<td>0.412637</td>
</tr>
<tr>
<td>HALLMARK_GLYCOLYSIS</td>
<td>0.238116</td>
<td>1.087641</td>
<td>0.277778</td>
<td>0.453047</td>
</tr>
<tr>
<td>HALLMARK_ANDROGEN_RESPONSE</td>
<td>0.242173</td>
<td>0.987504</td>
<td>0.490291</td>
<td>0.720882</td>
</tr>
<tr>
<td>HALLMARK_ESTROGEN_RESPONSE_EARLY</td>
<td>0.217192</td>
<td>0.984076</td>
<td>0.48538</td>
<td>0.700853</td>
</tr>
<tr>
<td>HALLMARK_PI3K_AKT_MTOR_SIGNALING</td>
<td>0.232547</td>
<td>0.975665</td>
<td>0.499237</td>
<td>0.700114</td>
</tr>
<tr>
<td>HALLMARK_APICAL_JUNCTION</td>
<td>0.210236</td>
<td>0.955581</td>
<td>0.557833</td>
<td>0.73086</td>
</tr>
<tr>
<td>HALLMARK_NOTCH_SIGNALING</td>
<td>0.284838</td>
<td>0.95034</td>
<td>0.540448</td>
<td>0.718131</td>
</tr>
<tr>
<td>HALLMARK_UV_RESPONSE_DN</td>
<td>0.21378</td>
<td>0.933273</td>
<td>0.606928</td>
<td>0.738535</td>
</tr>
<tr>
<td>HALLMARK_E2F_TARGETS</td>
<td>0.204165</td>
<td>0.931846</td>
<td>0.609442</td>
<td>0.716154</td>
</tr>
<tr>
<td>HALLMARK_UV_RESPONSE_UP</td>
<td>0.20756</td>
<td>0.925569</td>
<td>0.641337</td>
<td>0.70831</td>
</tr>
<tr>
<td>HALLMARK_APOPTOSIS</td>
<td>0.199951</td>
<td>0.892449</td>
<td>0.721386</td>
<td>0.769917</td>
</tr>
<tr>
<td>HALLMARK_FATTY_ACID_METABOLISM</td>
<td>0.18256</td>
<td>0.798649</td>
<td>0.892128</td>
<td>0.944097</td>
</tr>
<tr>
<td>HALLMARK_PANCREAS_BETA CELLS</td>
<td>0.243311</td>
<td>0.779823</td>
<td>0.789946</td>
<td>0.941415</td>
</tr>
</tbody>
</table>
### TABLE IV
GSEA DOWNREGULATED HALLMARK GENE SETS

<table>
<thead>
<tr>
<th>NAME</th>
<th>ES</th>
<th>NES</th>
<th>NOM p-val</th>
<th>FDR q-val</th>
</tr>
</thead>
<tbody>
<tr>
<td>HALLMARK_MYC_TARGETS_V2</td>
<td>-0.4748</td>
<td>-1.91201</td>
<td>0.002525</td>
<td>0.003985</td>
</tr>
<tr>
<td>HALLMARK_MYC_TARGETS_V1</td>
<td>-0.36855</td>
<td>-1.8331</td>
<td>0</td>
<td>0.004309</td>
</tr>
<tr>
<td>HALLMARK_REACTIVE_OXYGEN_SPECIES_PATHWAY</td>
<td>-0.41817</td>
<td>-1.6178</td>
<td>0.007444</td>
<td>0.015261</td>
</tr>
<tr>
<td>HALLMARK_DNA_REPAIR</td>
<td>-0.30228</td>
<td>-1.43666</td>
<td>0.005848</td>
<td>0.074423</td>
</tr>
<tr>
<td>HALLMARK_UNFOLDED_PROTEIN_RESPONSE</td>
<td>-0.29583</td>
<td>-1.32713</td>
<td>0.054496</td>
<td>0.150923</td>
</tr>
<tr>
<td>HALLMARK_INFLAMMATORY_RESPONSE</td>
<td>-0.27016</td>
<td>-1.32149</td>
<td>0.019868</td>
<td>0.129248</td>
</tr>
<tr>
<td>HALLMARK_MYOGENESIS</td>
<td>-0.2662</td>
<td>-1.30867</td>
<td>0.020896</td>
<td>0.123763</td>
</tr>
<tr>
<td>HALLMARK_XENOBIOLOGIC_METABOLISM</td>
<td>-0.26224</td>
<td>-1.26928</td>
<td>0.050562</td>
<td>0.144889</td>
</tr>
<tr>
<td>HALLMARKHEME_METABOLISM</td>
<td>-0.2377</td>
<td>-1.17278</td>
<td>0.107744</td>
<td>0.268996</td>
</tr>
<tr>
<td>HALLMARK_P53_PATHWAY</td>
<td>-0.23869</td>
<td>-1.15876</td>
<td>0.132258</td>
<td>0.268841</td>
</tr>
<tr>
<td>HALLMARK_OXIDATIVE_PHOSPHORYLATION</td>
<td>-0.22071</td>
<td>-1.08125</td>
<td>0.240356</td>
<td>0.420551</td>
</tr>
<tr>
<td>HALLMARK_BILE_ACID_METABOLISM</td>
<td>-0.2274</td>
<td>-1.00891</td>
<td>0.408108</td>
<td>0.607098</td>
</tr>
<tr>
<td>HALLMARK_PEROXISOME</td>
<td>-0.21821</td>
<td>-0.98109</td>
<td>0.479381</td>
<td>0.653097</td>
</tr>
<tr>
<td>HALLMARK_PROTEIN_SECRETION</td>
<td>-0.21131</td>
<td>-0.92838</td>
<td>0.625337</td>
<td>0.779998</td>
</tr>
<tr>
<td>HALLMARK_ALLOGRAFT_REJECTION</td>
<td>-0.17894</td>
<td>-0.86684</td>
<td>0.846386</td>
<td>0.89971</td>
</tr>
<tr>
<td>HALLMARK_ADIPGENESIS</td>
<td>-0.15816</td>
<td>-0.78721</td>
<td>0.977419</td>
<td>0.979796</td>
</tr>
<tr>
<td>HALLMARK_WNT_BETA_CATENIN_SIGNALING</td>
<td>-0.21028</td>
<td>-0.75408</td>
<td>0.838863</td>
<td>0.951449</td>
</tr>
</tbody>
</table>
TABLE V
PREVALENT GENES FROM LEADING EDGE ANALYSIS ON 637 GO BP GENE SETS

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Description</th>
<th>Expression</th>
<th>Number of times gene is in a gene set</th>
</tr>
</thead>
<tbody>
<tr>
<td>WNT5A</td>
<td>Wnt family member 5A</td>
<td>Upregulated</td>
<td>108</td>
</tr>
<tr>
<td>EDN1</td>
<td>endothelin 1</td>
<td>Upregulated</td>
<td>107</td>
</tr>
<tr>
<td>PTK2B</td>
<td>protein tyrosine kinase 2 beta</td>
<td>Upregulated</td>
<td>82</td>
</tr>
<tr>
<td>GHRL</td>
<td>ghrelin and obestatin prepropeptide</td>
<td>Upregulated</td>
<td>77</td>
</tr>
<tr>
<td>F2RL1</td>
<td>F2R like trypsin receptor 1</td>
<td>Upregulated</td>
<td>74</td>
</tr>
<tr>
<td>SHH</td>
<td>sonic hedgehog signaling molecule</td>
<td>Upregulated</td>
<td>72</td>
</tr>
<tr>
<td>JAK2</td>
<td>Janus kinase 2</td>
<td>Upregulated</td>
<td>66</td>
</tr>
<tr>
<td>ANXA1</td>
<td>annexin A1</td>
<td>Upregulated</td>
<td>64</td>
</tr>
<tr>
<td>TLR4</td>
<td>toll like receptor 4</td>
<td>Upregulated</td>
<td>59</td>
</tr>
<tr>
<td>NOD2</td>
<td>nucleotide binding oligomerization domain 2</td>
<td>Upregulated</td>
<td>58</td>
</tr>
</tbody>
</table>
Figure 13: GO Biological Processes Gene Set Collection Enrichment Map with 'Moderately conservative' parameters. Red nodes indicate upregulation and blue nodes indicate downregulation. Connections (edges) mean overlap of significant genes shared between gene sets.

Figure 14: GO Biological Processes Gene Set Collection Enrichment Map with 'Very conservative' parameters only showing blue nodes.
Figure 15: GO Biological Processes Gene Set Collection Enrichment Map with ‘Very conservative’ parameters only showing red nodes.
IV. Discussion

A. Results for GLDS-103

1) Reproducibility of Nextflow

The results produced from NRCP in comparison to NASA is almost identical. Despite differences in number of genes when filtered by p-value, different versions of software and databases would produce different results and the objective of this project is to possibly offer new insights in GLDS-103. The PCA plots, depicted in Figure 10 is a case of reproducibility since the PCA plot produced from NRCP is near identical to the PCA plot from NRCP. However, identical PCA plots could also mean that unique genes found in NRCP did not cause a change in variance.

2) NRCP and Resource Requirements

CoS-HPC served as an incredible asset in this project due to its large memory capacity and its processing power. About 103 GB of raw data was downloaded from GeneLab to CoS-HPC and processed. As shown in Table I, about 2 days was needed to produce 502.65 GB of data with CPU hours clocking over 255 hours. Using a daily computer would take NRCP a lot longer than shown. Trim Galore, STAR alignment, and Count alignment, took the longest to finish, with Trim Galore averaging 136 minutes, STAR alignment averaging 241 minutes, and Count alignment averaging 306 minutes. Speed up time in these tools would be beneficial for future research/studies regarding spaceflight. For instance, the tool, fastp, does all preprocessing jobs and at a faster speed than Cutadapt [59].

![Figure 16: Plot of amount of time elapsed for each process](image-url)
3) Gene Set Analysis

The top three results for DAVID and GSEA Hallmark gene sets will be observed, with the exception to the downregulated Hallmark gene sets. This fulfils the FDR cut-off for GSEA since all top three gene sets produced from the Hallmark gene sets have an FDR q-value less than 0.05 (recommend FDR cut-off by GSEA due to permutation by gene sets). It is important to note that FDR q-values for myogenesis and inflammatory response are above the FDR q-value and it is quite possible that these gene sets were not relevant for the Quadriceps muscle under microgravity reasons. This is mainly for curiosity reasons and the chance for the gene sets for myogenesis and inflammatory response are, in fact, false positives.

a) DAVID gene annotation clusters

Circadian rhythm plays an important role in muscle atrophy. By definition, Circadian rhythm is a human’s regular physiological activities, regulated by diet, exercise, sleep, endocrine activity, and metabolism [60]. Major factors that affect circadian rhythm are, aging, diet, lack of exercise, external light sources, and disease, all of which induce muscle atrophy [61]–[64]. A recent study discovered a connection between circadian rhythm and skeletal muscle atrophy, of which Choi et al. discovered irregular working schedules increase skeletal muscle atrophy [65]. A mechanism that controls regulation of the musculoskeletal system is called the molecular clock, a series of interconnected transcriptional-translational feedback loops [66]. Further investigation of the molecular clock mechanism is shown to be present within all cells within the body, including both skeletal muscle and bone [67]–[70]. However, there seems to be little understanding of bone cell clock disruption and this is evident as Schroder et al. is the only study to describe the systematic effects of skeletal muscle rhythms to maintenance of bone [71].

Annotation Cluster 2 indicates that repressor, transcription regulation, and dna-binding are disrupted in microgravity. However, nothing clear sticks out on what exactly causes disruption in this annotation cluster. A deep dive on what genes is prevalent in all these keywords would open insights. As shown in Figure 17, Mafb, Nr1d1, Klf11, Nfil3, Zkscan17, and Nr1d2, are all genes prevalent in the annotated terms for Cluster 2. Mafb can either be a transcriptional activator or repressor and loss of this gene function results in a different type of muscle around the eye [72], [73]. Nr1d1, or commonly named in articles, Rev-erbA-alpha, is an interesting gene, as it is a transcriptional repressor that coordinates circadian rhythm and metabolic pathways [72], [74]. Klf11 is a suppressor of endothelial cell inflammatory activation [75]. Nfil3 is an important transcription factor for the development of natural killer cells and type 1 innate lymphoid cells [76, p. 3]. Not much is known about Zkscan17 (also called Nizp1) besides being a protein that interacts with NSD1 [77, p. 1]. Nr1d2 is known to be a repressor for circadian rhythm genes [78, p. 2]. These genes have shown to regulate transcription in some pathway, system, or mechanism, some of which to be repressors of the circadian rhythm.

Metal-binding, zinc, and zinc finger are terms clustered in Annotation Cluster 3. Looking at the genes that were connected to this cluster, Zinc and ring finger 2 and 3 (Znrf2, Znrf3) and ring finger protein 31, 114, 122, 128, 139, and 167 (RNF31, RNF114, RNF122, RNF128, RNF139, RNF167) were noticeable and prevalent across all cluster terms. It is possible that there is some connection between microgravity and zinc and ring finger genes. Zinc is important for every life on Earth, in fact, one out of ten human proteins are zinc related [79]. The ring domain
composition of the zinc finger can variate, providing different molecular function. CXXC5 and EGR3 are zinc-finger proteins that regulate skeletal myogenesis [80]. Lynch et al. found that zinc finger protein 593 (Zfp593), a protein-coding gene, is expressed in skeletal muscle and gets upregulated during skeletal muscle atrophy [81]. Unfortunately, the 3 aforementioned genes are not found in Cluster 3 but could be relevant in other skeletal muscle in the RR project.

![Figure 17: Genes that are prevalent in all annotated terms of cluster 2](image)

**b) Investigation of upregulated hallmark gene sets**

GSEA investigates the whole gene list, accounting for genes that have smaller expression values compared to significant genes of the whole gene list. This allows the user to see the gene list as a whole and not by a gene-by-gene basis. The hallmark gene set collection is used as it is an excellent gene set collection as a starting point of investigation.

In the Rodent Research project, NASA found that TGF-β1 is the most common regulator in microgravity and space radiation conditions [7]. TGF-β1 is an important cytokine and can cause fibrosis. Table III showcases the gene sets that are upregulated, or in this case, gene sets that are upregulated in FLT mice. Interferon gamma is known to be a TGF-β1 inhibitor, and because of this, interferon gamma is an antifibrosis agent [82], [83]. Cholesterol is stored in the T-tubule membranes in skeletal muscle tissue [11]. The cholesterol uptake to the skeletal muscle leads enrichment of muscle fiber membrane (sarcolemma) [84]. Cholesterol homeostasis upregulation is reasonable in microgravity conditions, as the skeletal muscle tries to maintain
its composition through cholesterol homeostasis by absorbing more cholesterol. Hedgehog signaling is upregulated in spaceflight, and a relevant gene, sonic hedgehog (SHH) is prevalent in significant gene sets as shown in Table V. Caradu et al. investigated the effects of SHH on ischaemic (restriction of blood flow to tissues) mice and found that ectopic SHH promoted ischaemia-induced angiogenesis [85].

c) Investigation of the downregulated hallmark gene sets

Table IV shows that the gene sets, MYC targets versions 2 and 1, are downregulated in FLT mice. MsigDB describes these two gene set collection as genes regulated by MYC. MYC is a proto-oncogene, thus, downregulation of MYC results in cell proliferation and cause inducement of cell apoptosis [86]. Referring to Table V, none of these genes are listed in the MYC target version 2 and 1 gene set. Reactive oxidative species (ROS) are produced as a by-product of cellular metabolic pathways [87]. Contraction of the skeletal muscle tissue produces ROS, where low levels is required for normal skeletal muscle force and high levels causes contractile dysfunction and fatigue [88]. This may suggest mice can produce less ROS under microgravity conditions. Uchida et al. studied the effects of microgravity on L6 myotubes rats and found increased ROS production signals the (Egr)1/2-Cbl-b signaling pathway, known to contribute to muscle volume lost [20]. This study is contradictory to the findings of GSEA since FLT mice experience downregulation of the ROS pathway, but an explanation could be the mice’ condition in the experiment. The downregulation of myogenesis and inflammatory response proves previous studies of spaceflight induces skeletal muscle atrophy [16]–[18]. There has been a study regarding astronauts being exposed to microgravity had their stress-response genes be suppressed, including genes that are important to DNA repair, oxidative stress, and protein folding [89].

d) Insights to GO biological processes terms through the C5 GO BP subset

Figure 13, 15, and 16 were figures created by the EnrichmentMap app, with Cytoscape as the platform. Figure 13 was manually annotated based on the GO BP term while Figure 14 and 16 were automatically annotated by AutoAnnotation and WordCloud. The different parameters suggest that there are not a lot of significant gene sets. Running EnrichmentMap with ‘Moderately conservative’ parameters excluded 237 gene sets, while running with ‘Very permissive’ parameters included the 237 gene sets. Interleukin 1 production cluster in Figure 13 have gene sets involving negative regulation of interleukin and production of interleukin-1 beta. Interleukin-1 beta is a proinflammatory cytokine upregulated in Parkinson disease, a disease linked to skeletal muscle atrophy [90], [91]. Interestingly, the same studies, have shown that tumor necrosis factor, a gene like interleukin-1 beta, to be also upregulated in Parkinson disease. In Figure 15, interleukin 1 production includes a gene set related to tumor necrosis factor production; however, this gene set was excluded from the ‘Moderately conservative’ enrichment map due to having an FDR q-value of 0.2312. This shows that, despite gene sets
being excluded, there is a chance for the gene set of not being a false positive. Future studies would require limiting the number of gene sets to compare with in C5 GO BP subset in order to increase the statistical power.

V. Conclusion

Results from analyzing GLDS-103 shows strong relations to the circadian rhythm and skeletal muscle health. In addition, a possible relation of musculoskeletal system to the bone system by the mechanism called the molecular clock. The results also shown cases of the zinc finger protein and its variants being disrupted but also shows that the genes disrupted are linked to circadian rhythm. NASA has found TGF-β1 to be present across the Research Rodent project, and this project has shown an inhibitor to TGF-β1, the interferon gamma gene. Additionally, this project has shown to validate other spaceflight research to skeletal muscle atrophy. Enrichment maps of the C5 GO BP subset show insight to possible future BP studies. As proven in this project, a gene set, that was excluded due to its high FDR q-value, was shown to be important in a study regarding the Parkinson disease.

This project has shown the reproducibility of Nextflow implementation, showing almost identical results to GeneLab. The greatest advantage of NRCP is that it can be rerun with different parameters, such as ENSEMBL version specification, allowing the pipeline to be ran with updated versions of the mouse genome and annotation. This was the case as GeneLab used an older version of ENSEMBL and this project showed differences of results when filtering genes to a p-value of less than 0.05. This project also highlights the importance of high memory and high processing power systems, as almost 2 days were required to process the entire raw data of GLDS-103. Quality checks of the raw data has shown no problems with the raw data and alignment results indicate enough reads being uniquely mapped to the genome.
REFERENCES


[58] “QC Fail Sequencing » Positional sequence bias in random primed libraries.”


Appendix: Scripts

```
#!/bin/bash
#SBATCH --nodes=1
#SBATCH --job-name=glide-103
#SBATCH --output=slurm_error.txt
#SBATCH --error=slurm_output.txt
#SBATCH --time=72:00:00

#SBATCH -p node

export TOWER_ACCESS_TOKEN=eyJ0eXAiOiJKV1QiLCJhbGciOiJIUzI1NiJ9.eyJzdWIiOiJUZXVuZXIiLCJpYXQiOjE2MDIyOTAyNjEsImV4cCI6MTY2MzI3OTI2N30.f40pZVzWhmz901j3YRs2r2cS8jg6TQl6UcEJ1ZNy0mGE

nextflow run J-S1/Nextflow_RCP -x vv_aar_update --glideAccession GLIS-103 --stageLocal --ensembleVersion 104 --ERCC=false --with-tower --resume

echo "Finished pipeline"

job_end="date +ts"

runtime=$(($job_end - $job_start))

hours=$(($runtime / 3600))

minutes=$(($runtime % 3600 / 60))

seconds=$(($runtime % 3600 % 60))

echo "Runtime: $hours:$minutes:$seconds (hh:mm:ss)"
```

Script 1: Bash Script for CoS-HPC