Exploring Gene Expression with Bioinformatics Using GLDS-288

Curricular Unit
Student Materials

EXPLORE

Space is an extremely harsh environment. NASA has been studying the effects of spaceflight on humans since 1961 when the human space program was started. After all this research, there is a lot that we still don't know about how space impacts biological processes or how these impacts are happening. As space biologists, we work on determining the mechanisms behind **phenotypic changes** we see during spaceflight. This is often due to changes in **gene expression**.

Spaceflight Factors:

There are five main spaceflight factors that are the stressors that cause the biological changes we see during spaceflight and are the biggest hurdle for us to overcome before considering long term space missions.

	Radiation	Isolation	Distance from Earth	Altered Gravity Fields	Hostile Environment
Brainstorm ways that these factors could impact biological processes					

We use model organisms, like mice, often during spaceflight experiments. Think about the positives and negatives of using mice as model organisms and what other examples of model organisms could be.

Pros for Using Mice	Cons for Using Mice	Other Examples of Model Organisms

The Goal: Use the GeneLab repository and Galaxy platform to analyze data to determine possible biological pathways contributing to **differential gene expression** in space in the spleens of mice.

Background:

The data we are going to analyze in GeneLab originated from a spaceflight experiment that flew on the Mouse Habitat 1 mission which launched with the JAXA to the ISS. To better understand the transcriptomic data, we need to get more information on how the experiment was setup. Visit this link for information on this experiment (GLDS-288).

1. Locate the Treatment Protocol information. Using this information, identify what independent variable(s) were investigated.

2. Given the independent variable(s), identify the treatment groups that were used.

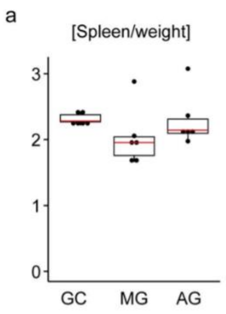
3. Browse through the remainder of the Protocol section. What is the dependent variable in this experiment?

4. Identify at least 3 controls that were used in the experiment and explain the importance of each.

Now that we understand more about the experimental design, we need to familiarize ourselves with the organ of interest: the spleen. After watching this video, answer the following questions and begin analyzing data the GLDS-288 researchers collected.

1. What are the two main functions of the spleen?

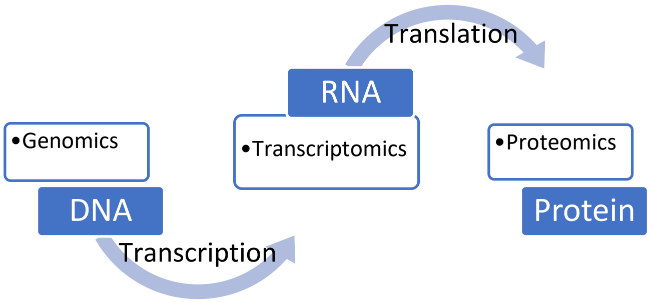
2. To the right is a graph of the weight of the mice from each of the experimental groups in our study. GC stands for Ground Control, AG is Artificial Gravity and MG is Microgravity. **Identify trends in the data and create a working hypothesis about why you think this might be happening.**



EXPLORE

RNA Sequencing

Often, when we refer to sequencing, we are talking about genetic sequencing—or determining the order of nucleotides in DNA. This is a helpful technique when you are interested in mutations or differences between individuals or alleles involved in certain conditions. However, in different environmental conditions, like spaceflight, we see changes in the phenotypes that are often not a result of mutations in DNA, but instead the result of genes being turned on or off due to the difference in the environment. This is **differential gene expression**; the DNA itself is not changed but environmental factors are turning the gene on so it is transcribed more often or turning the gene off so it is transcribed less often.



In order to study gene expression, information from mRNA is going to be most helpful. In order to collect this information, we use RNA sequencing. Using RNA sequencing to survey genes that are actively transcribed is called **transcriptomics**.

RNA sequencing is a complex process, but in effect, the mRNA transcripts from a cell, tissue or organism is isolated from the cells, transformed into complementary DNA (cDNA) which can then be sequenced. Each piece of cDNA is compared to a known genome to determine which genes the original mRNA transcripts were from in a process called **alignment**. In this way, we can use DNA sequencing methods to determine which genes were turned off or on which then gives us important information about the ways biological processes are changes in response to the environmental factors.

1. Explain the differences between genomics and transcriptomics. *Hint: think of the types of data that is collected in each.*

2. Think of two situations where genomics would be the most advantageous.

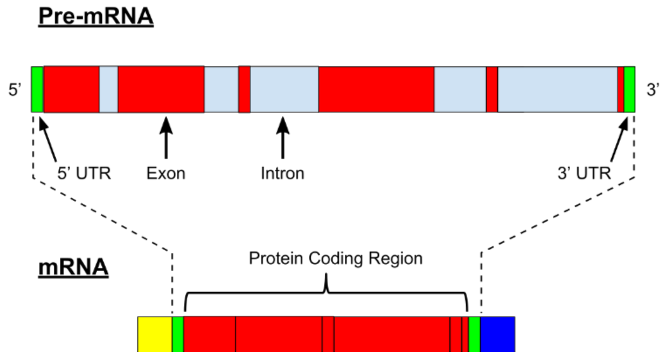
3. Think of two situations where transcriptomics would be the most advantageous.

RNA Processing and RNAStar

Much of the process of RNA sequencing happens after the sequencing has been completed. Much of this is done using bioinformatic algorithms that help show trends in the data, run alignments, and create visualizations that scientists need to make sense of all this information. RNAStar is one of the alignment algorithms that GeneLab uses on the Galaxy platform. What is particularly helpful about RNAStar is that it is **splice-aware**. To understand what this means and why it is so helpful, let's think about mRNA processing.

In eukaryotic genomes, one gene does not equal one protein. In other words, we can arrange the pieces of a gene in different ways through **RNA splicing** to create different transcripts which will then make different proteins.

Pre-mRNA is what the gene directly reads. It contains **exons** and **introns**. Exons are the portions that are expressed in the final protein. Introns are interrupting pieces that are removed. To create different protein products, different introns are removed or left in to create new transcripts.



USED WITH CREATIVE COMMONS LICENSE FROM HTTPS://COMMONS.WIKIMEDIA.ORG/WIKI/FILE:PRE-MRNA.SVG

1. How does RNA splicing complicate the process of alignment?

2. RNAStar is a splice-aware algorithm. What does this mean?

3. If we searched a genome for an entire mRNA transcript would we find a match? **Explain why or why not.**

4. How does RNA STAR accommodate splicing when it runs alignments?

Bioinformatics Analysis:

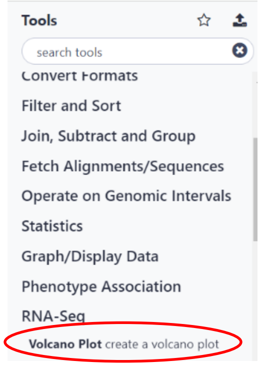
We will use a dataset from the GLDS-288 experiment that has already been aligned and gene expression in the flight group have been compared to the ground control group. We will not consider the artificial gravity group for our analysis. This data has been **normalized** to help us ignore genes that did not show significant changes in expression between the two groups. The first visualization we are going to create to begin analyzing data is a **volcano plot**.

To create a volcano plot for our dataset:

1. Open the Galaxy history here. Click on the plus sign in the top right corner to add it to your history. **Name the history GLDS-288 Analysis.**
2. In the tool bar on the left, Click on RNASeq > Volcano plot
3. Use the following settings:
   - o ***Specify an input file***: select Complete Gene Expression Results
   - o ***FDR (adjusted P value)****:* Column 7
   - o ***P value (raw):*** Column 6
   - o ***Log Fold Change***: Column 3
   - o ***Labels****:* Column 13
   - o ***Significance threshold***: Enter 0.05
   - o ***LogFC threshold to color***: Enter 1
   - o ***Points to label***: Significant
   - o ***Only label top-most significant***: Enter 20

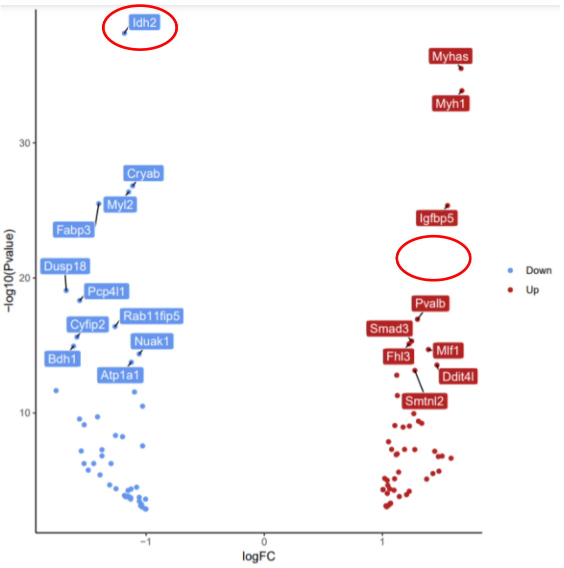
   <u>All other parameters to be left on default settings.</u>

   o Click Execute



The job should appear in your history and should be grey initially. As it is processed, it will turn orange and then green. While your job runs, read the information below about how to analyze a volcano plot and answer the questions.

Analyzing Volcano Plots:



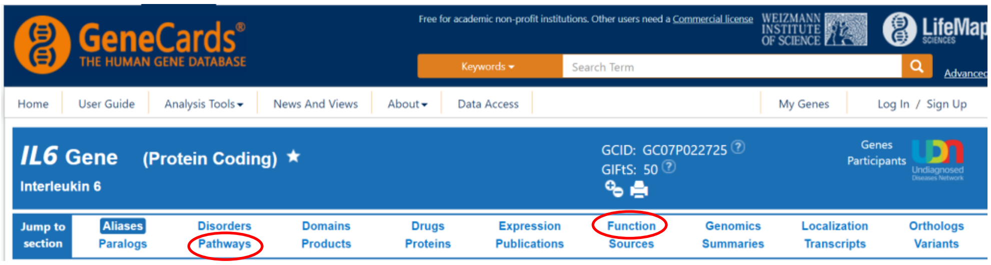
This is an example of what a volcano plot looks like. Your volcano plot will be different than this one.

- Each point represents one gene.
- The x-axis is showing the **fold change** which is how many times more or less the gene was expressed during spaceflight.
- The y-axis is showing statistical significance, or how confident we are that these genes changed because of the environment (space).
- Red points are genes that are **upregulated**.
- Blue points are genes that are **downregulated.**
- The labels are the 20 most significantly changed gene IDs.

1. Explain what a red point indicates regarding the gene is it associated with.

2. There are two genes circled in the example plot above. Which gene has a higher statistical significance? Explain what that tells you.

Once your volcano plot file turns green, view the file by clicking the eye icon. Pick 3 genes from the plot to search in genecards.org. Fill in the table below with as much information as you can find. To search in GeneCards, paste in the GeneID from the volcano plot into the Explore a Gene search bar. On the results page, use the Function and Pathways tabs to find more information.


Free for academic non-profit institutions. Other users need a Commercial license

GeneID	Up/Downregulated	Gene Function	Major Pathways

Share your table with another group that analyzed *different* genes. Describe any trends you see among the genes. **Create a hypothesis about what pathways or processes that are being impacted and why you think those changes may be happening given the conditions of spaceflight.**

Gene Ontology Analysis:

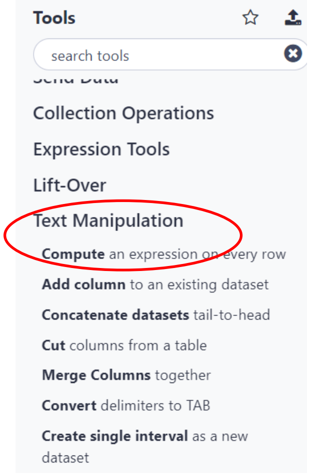
Gene Ontology is a database of gene functions and when we run data through goseq it decreases the complexity of the data so we can analyze it easier. Instead of looking up every individual gene that is differentially expressed, goseq runs the genes through Gene Ontology which determines if several genes function in similar biological pathways to better show us trends in expression. When we run our analysis, we will get a plot showing several significant pathways, how much each pathways' expression was changed in flight versus ground control and how statistically significant the changes were.

We have to first alter our files to run them in goseq:

1. Find Text Manipulation in the Toolbar > Compute
2. ***Add Expression:*** bool(c7<0.05)
3. ***As a new expression to:*** DESeq2 Results file
4. Execute

This will create a file with True or False added as a final column to our DESEq2 results. True will indicate that the expression change was statistically significant.

We need the Cut tool to get rid of all the columns from the file **except** for the gene (c1) and the true/false statements we added (c8)

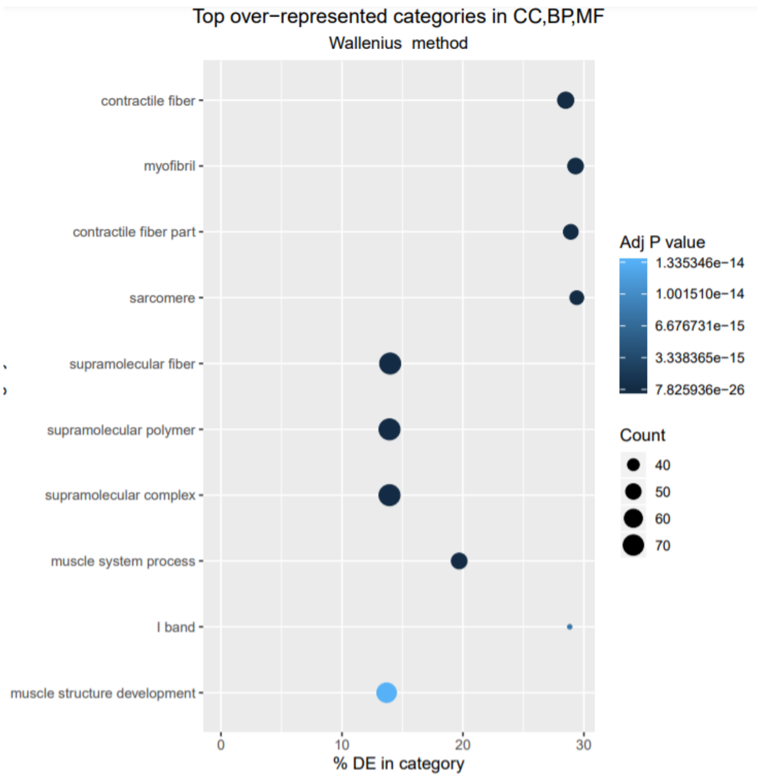

Tools
search tools
Collection Operations
Expression Tools
Lift-Over
Text Manipulation
Compute an expression on every row
Add column to an existing dataset
Concatenate datasets tail-to-head
Cut columns from a table
Merge Columns together
Convert delimiters to TAB
Create single interval as a new dataset

1. Find Cut under the same menu as Compute
2. ***Cut columns:*** c1,c8 from the Compute file we just generated
3. Execute
4. After the file is green in your history, Click the pencil icon to change the name of the file to **Genes and T/F**

Now we are ready to run goseq:

1. Search goseq in the toolbar menu
2. Choose the **Genes and T/F file** for the first file
3. ***Gene Lengths:*** Select the **GLDS-288 Gene lengths file**
4. ***Select a Genome:*** Mouse
5. Find ***Output Options***
   o ***Output Top GO Terms Plot?****:* Yes
   o ***Extract the DE genes?*** Yes
6. Execute

Goseq will generate 3 files in your history. We will be analyzing the Top over-represented terms plot. An example of this plot can be found on the next page. Read over the information on how to interpret this plot and then answer questions 1 and 2 for the GLDS-288 data when your files are done processing.



- Each dot represents ALL of the genes in one GO pathway
- Each pathway is labeled on the left
- The color of the dot is the P value (statistical significance)
- The darker the color, the MORE significant the changes are
- The x-axis how much change in gene expression there is
- The size of the points represents how many genes were altered in that pathway

For example, the sarcomere pathway has 28% differential gene expression and a high significance, but only 40 genes were changed.

1. Are there pathways that seem related to each other?

2. Discuss some of the tradeoffs in choosing which pathways to research further—should you choose pathways where lots of genes are interacting, higher DE changes or choose based on the P values?

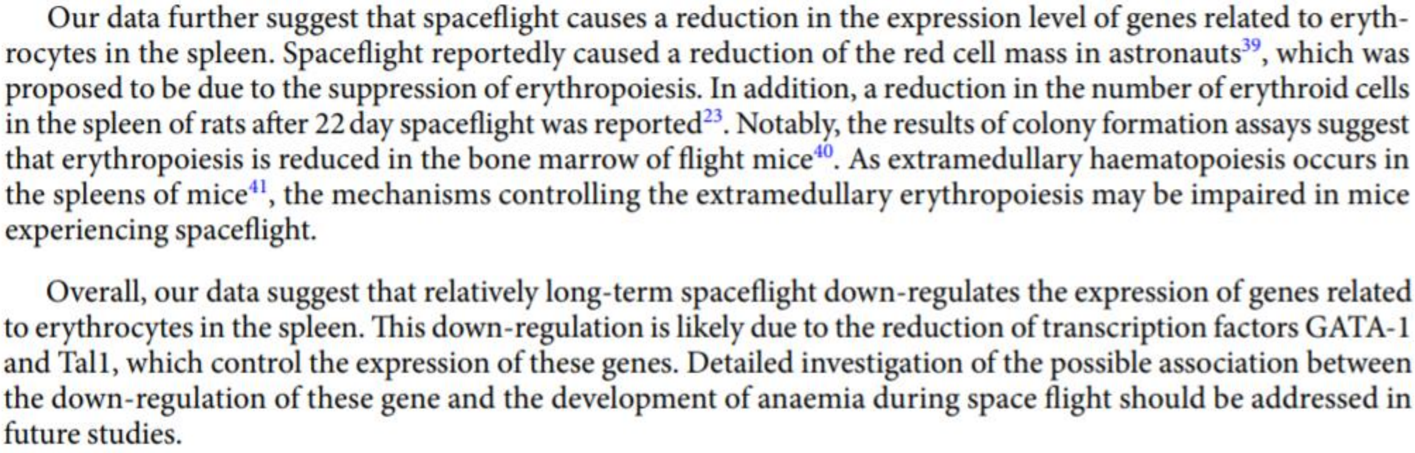
Choose 2 pathways to research using geneontology.org and complete the table to determine whether these pathways give you clues as to how space is impacting the spleen. Follow directions from the PowerPoint to help you navigate GeneOntology. You may also choose to do research on these pathways using other online resources.

GO Process	Description	Notes/Other Related Processes

Share your finding with another group to determine if there are overlaps in processes. Pick one GO term. **Revise your original hypothesis you created after analyzing the volcano plot.**

EXPLAIN

Now that you have created hypothesis for how the spleen is being affected by spaceflight, read the excerpts of this paper to see how Horie et al (2019) analyzed this data and the conclusions they came to.



Our data further suggest that spaceflight causes a reduction in the expression level of genes related to erythrocytes in the spleen. Spaceflight reportedly caused a reduction of the red cell mass in astronauts[39], which was proposed to be due to the suppression of erythropoiesis. In addition, a reduction in the number of erythroid cells in the spleen of rats after 22 day spaceflight was reported[23]. Notably, the results of colony formation assays suggest that erythropoiesis is reduced in the bone marrow of flight mice[40]. As extramedullary haematopoiesis occurs in the spleens of mice[41], the mechanisms controlling the extramedullary erythropoiesis may be impaired in mice experiencing spaceflight.

Overall, our data suggest that relatively long-term spaceflight down-regulates the expression of genes related to erythrocytes in the spleen. This down-regulation is likely due to the reduction of transcription factors GATA-1 and Tal1, which control the expression of these genes. Detailed investigation of the possible association between the down-regulation of these gene and the development of anaemia during space flight should be addressed in future studies.

1. What are erythrocytes and how are they related to the spleen's function?

2. What is erythropoiesis?

3. Explain what is happening to the spleen in terms of **differential gene expression**. (What genes are being turned on/off and how is that impacting the spleen?)

4. Based on this excerpt, what is a transcription factor? How do they contribute to gene expression?

Transcription factors is a vague term used to describe a lot of different proteins. Define the three main types of transcription factors below.

Activators-

Enhancers-

Repressors-

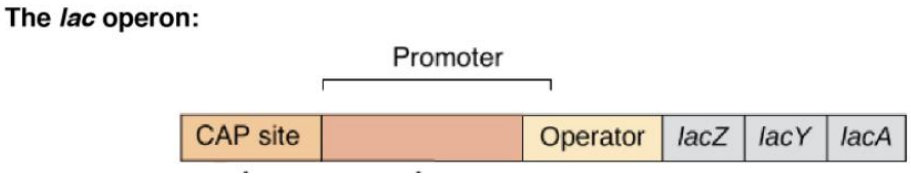
Explain how transcription factors could be contributing to the changes in gene expression we see in the GLDS-288 data.

EXTEND

Let’s apply what we learned about gene expression in the spleen to gene expression in bacteria by looking at the *lac* operon. This is a tightly controlled operon in bacteria because it determines what kind of nutrient they can break down—lactose or glucose. It is important to only express genes when the environment warrants it, otherwise the bacteria are wasting energy creating enzymes they don’t need.

Based on the description of the parts of the *lac* operon and our understanding of gene expression, determine how the transcription factors interact for the scenarios below.

- The operon has 3 genes, all are necessary to metabolize lactose
- CAP is an **activator**
- *lac* **repressor** is bound to the operator unless lactose is present
- CAP binds when glucose concentration is low



USED WITH CREATIVE COMMONS LICENSE FROM https://www.khanacademy.org/science/ap-biology/gene-expression-and-regulation/regulation-of-gene-expression-and-cell-specialization/a/the-lac-operon

Scenario 1
Bacteria have access to a high glucose/no lactose media.

Transcription Factors Bound	RNA Polymerase Active?	*lac* operon expressed?

Scenario 2:
Bacteria have no glucose available and high lactose.

Transcription Factors Bound	RNA Polymerase Active?	*lac* operon expressed?

Scenario 3:
Bacteria are in a media of equal parts glucose and lactose. They preferentially break down glucose and are now running low.

Transcription Factors Bound	RNA Polymerase Active?	*lac* operon expressed?

Scenario 4:
Bacteria have run out of both glucose and lactose in the media.

Transcription Factors Bound	RNA Polymerase Active?	*lac* operon expressed?

AUTHOR
Jennifer Callison-Bliss, Wheeler High School (Marietta, Georgia)
Edited by GL4HS Staff