

NASA GL4U: On-Demand Bioinformatics Training Using Space Biology Omics Data

National Aeronautics and
Space Administration



2024 ASGSR Educator Session 3

Amanda M. Saravia-Butler, Ph.D.
Science and GL4U Lead, NASA GeneLab
Contractor: Amentum

Biological & Physical Sciences



GL4U Background / Objectives



Background:

- NASA's GeneLab project empowers researchers with open access to space-relevant multi-omics data through the [Open Science Data Repository \(OSDR\)](#)
- GeneLab for Colleges and Universities (GL4U) offers **bioinformatics training for space biology**
- GL4U was conceptualized in 2020 through a collaboration with SJSU and USRA, and the first GL4U bootcamp was held virtually in June 2021



Objectives:

- **Educate and train** the next generation of scientists to process, analyze, and interpret space-relevant 'omics data using publicly available data and bioinformatics tools
- **Maximize** the number of scientists who understand Space Biology data and utilize NASA's open-source 'omics data and tools
- **Provide educators with the knowledge and resources** required to train and inspire their students using GeneLab bioinformatic analyses as an entree into Space Biology

GL4U materials are organized into introduction and omics-specific modules

Introduction Module

- Overview Lecture (NASA, SMD, BPS, OSDR, GeneLab)
- Jupyter Lab Tutorial
- Unix Jupyter Notebook (hands-on training)
- R Jupyter Notebook (hands-on training)

The Intro Module serves as a pre-requisite for any omics-specific module

Omics-specific Modules (RNAseq, Amplicon Seq, etc.)

- Omics Data Lectures
 - Experimental design
 - Sample preparation and quality control
 - Data processing tools and visualizations
 - Results analysis and interpretation
- Hands-on data processing and analysis of an OSDR dataset via Jupyter Notebooks (JNs)

Unix Intro JN

5. Running commands

Using the foundational rules described above, we will begin running some commands.

date

`date` is a command that prints out the date and time. This particular command doesn't require any arguments:

```
In [2]: date
Sat Jul 8 22:40:55 PDT 2023
```

When we run `date` with no arguments, it uses some default settings, like assuming we want to know the time in our computer's currently set time zone. But we can provide optional arguments to `date`.

Optional arguments most often require putting a dash in front of them in order for the program to interpret them properly.

Here, we are adding the `-u` argument to tell the `date` program to report UTC time instead of the local time – which will be the same if the computer we're using happens to be set to UTC time:

```
In [3]: date -u
Sun Jul 9 05:41:28 UTC 2023
```

Note that if we try to run the command above without the dash, we get an error (ignore the message that prints out highlighted in red, we wouldn't normally see that outside of a notebook):

```
In [4]: date u
date: invalid date 'u'
```

Note

Notice that the error above comes from the program `date`. So the program we wanted to use is actually responding to us, but it doesn't seem to know what to do with the letter `u` we gave it. And this is because it wasn't prefixed with a dash, like `-u`.

Let's see what happens if we try to enter this without the "space" separating `date` and the optional argument `-u`; the computer won't know how to break apart the command and we get a different error (again, ignoring the red output):

```
In [5]: date-u
date-u: command not found
```

R Intro JN

1d. Data frame manipulations

Much of the data we work with in bioinformatics is in the data frame or matrix format. For example, gene expression data is usually held in matrix format, with samples as columns and genes as rows, where each entry (or cell) in the matrix contains the expression of a particular gene in a particular sample.

When analyzing numerical data in table format, it can be useful to be able to perform mathematical functions on all cells in a data frame, such as adding a value to all cells or taking the log of all cells. Fortunately, R makes that easy for us to do.

Below are some examples of common mathematical manipulations we often perform on data frames in bioinformatics.

Add a value to all cells

In R, you can add, subtract, multiply, or divide the number in every cell of a data frame by a specific value very easily. Run the command in the next cell to add `1` to every value in your `myDF` data frame.

```
In [24]: myDF + 1
```

A data.frame: 10 x 3

	column1	column2	column3
	<dbl>	<dbl>	<dbl>
row1	2	3	4
row2	5	6	7
row3	8	9	10
row4	11	12	13
row5	14	15	16
row6	17	18	19
row7	20	21	22
row8	23	24	25
row9	26	27	28
row10	1	1	1

Use the next cell to subtract 2 from all values in your `myDF` data frame.

```
In [25]: myDF - 2
```

AmpSeq JN

Note

It's worth noting again that these are not interpretable as "real" numbers of anything (due to the nature of sequencing data), but they can still be useful as relative metrics of comparison within a study.

As a reminder:

- the left, "Chao1", is an estimate of total richness (total number of unique "things")
- the right, Shannon, is a metric of diversity – which incorporates "richness" and "evenness" (the relative proportions of all our unique things to each other)

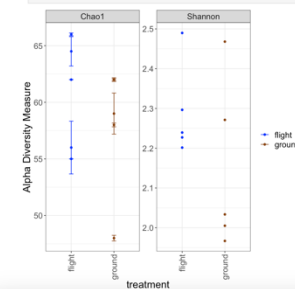
Looking at the plots above:

1. Do you notice any immediate differences between the flight and ground-control groups?

Some thoughts

We can also modify the parameters of the `plot_richness()` phyloseq function to group samples based on our treatment groups, which can be helpful sometimes:

```
In [32]: plot_richness(ASV_physeq, x = "treatment", color = "treatment", measures = c("Chao1", "Shannon")
scale_color_manual(values = unique(sample_info_treatment)) +
theme_bw() + theme(legend.title = element_blank(), text = element_text(size = 18),
axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```



RNAseq JN

4c. Volcano Plot

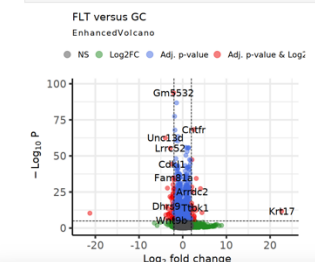
Finally, let's make a volcano plot to identify a few interesting genes. A volcano plot is a scatterplot which shows the relationship of the adjusted p-value to the log2 fold change. Genes with large fold changes that are also statistically significant by adjusted p-value are labeled.

First, we'll use the default settings from the `EnhancedVolcano()` function: log2 Fold Change cutoff > |2|, and the adjusted p-value cutoff is < 10e-6.

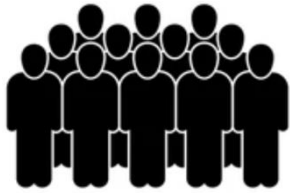
Note: You can read more about the `EnhancedVolcano()` function and see some examples by clicking [here](#)

```
In [68]: # Volcano plot showing genes differentially expressed in FLT vs GC
EnhancedVolcano(DGE_output_table,
lab = DGE_output_table$SYMBOL,
x = 'Log2fc_FLTvsGC',
y = 'Adj.p.value_FLTvsGC',
title = 'FLT versus GC',
legendLabel=c('NS','Log2FC','Adj. p-value'),
pCutoff = 10e-6,
FCcutoff = 2,
pointSize = 3.0,
labSize = 6.0,
colAlpha=0.5)

## Save your volcano plot
ggsave(file.path(DGE_plots,'GL08-104_volcano_DGE.png'), width = 6.5, height = 8.5, dpi = 300)
```



GL4U Direct Approach Students



2021 – Space biology & RNAseq

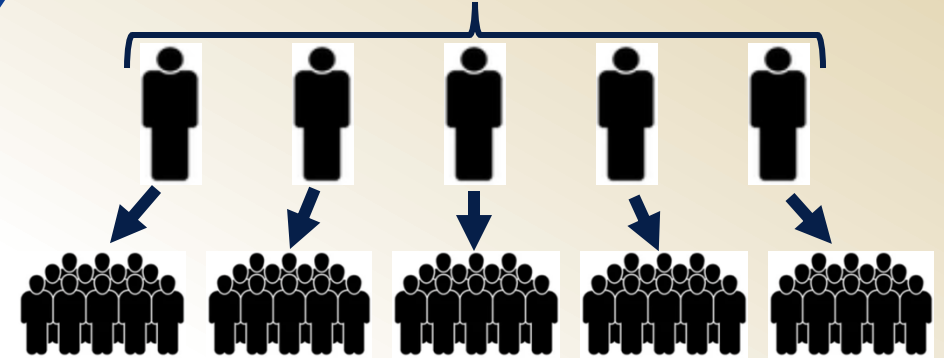
- Virtual 1-week long bootcamp with 14 **SJSU** students
- Collaboration with USRA/SJSU; SJSU compute system
- Space biology- and RNAseq-specific lectures and hands-on instruction using Jupyter Notebooks (JNs)

2023 – Space biology & Amplicon Seq

- In-person 4-day bootcamp with 24 **CSULA** students
- Collaboration with JPL/CSULA; NSF ACCESS compute system
- Space biology- and Amplicon Seq-specific lectures and JNs



GL4U Indirect Approach Educators



2022 – Space biology & RNAseq

- Virtual ~1.5-week long bootcamp
- Collaboration with JPL; SMCE compute system
- 6 professors and 4 graduate students from **4 Institutions**
- Space biology- and RNAseq-specific lectures and hands-on instruction using JNs
- Resources to teach at home institution

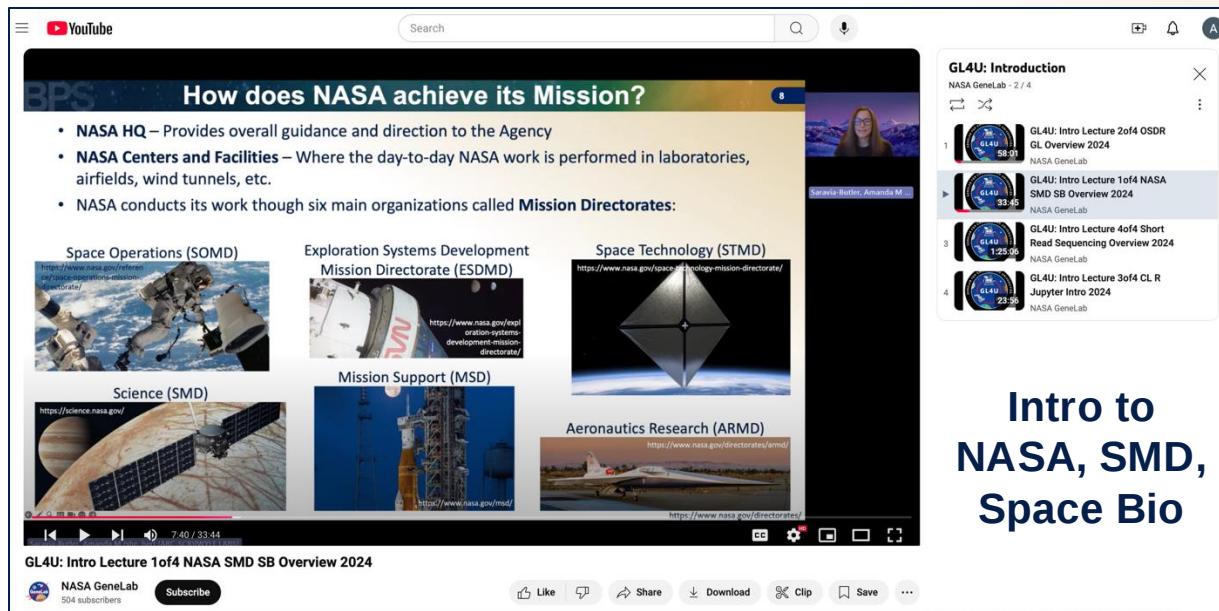
On-demand courses now available!

A hand holds a white microarray chip with blue and red lines. Below it, a petri dish contains green cells. A glowing atom and a diagram of a cell are overlaid on the scene. The background is a dark, starry space. The text "GL4U On-Demand" is centered in white.

GL4U On-Demand

Recorded lectures on YouTube

Hands-on activities via GitPod



How does NASA achieve its Mission?

- NASA HQ** – Provides overall guidance and direction to the Agency
- NASA Centers and Facilities** – Where the day-to-day NASA work is performed in laboratories, airfields, wind tunnels, etc.
- NASA conducts its work through six main organizations called **Mission Directorates**:

Space Operations (SOMD)

Exploration Systems Development Mission Directorate (ESMD)

Space Technology (STMD)

Science (SMD)

Mission Support (MSD)

Aeronautics Research (ARMD)

GL4U: Intro Lecture 1of4 NASA SMD SB Overview 2024

NASA GeneLab 504 subscribers

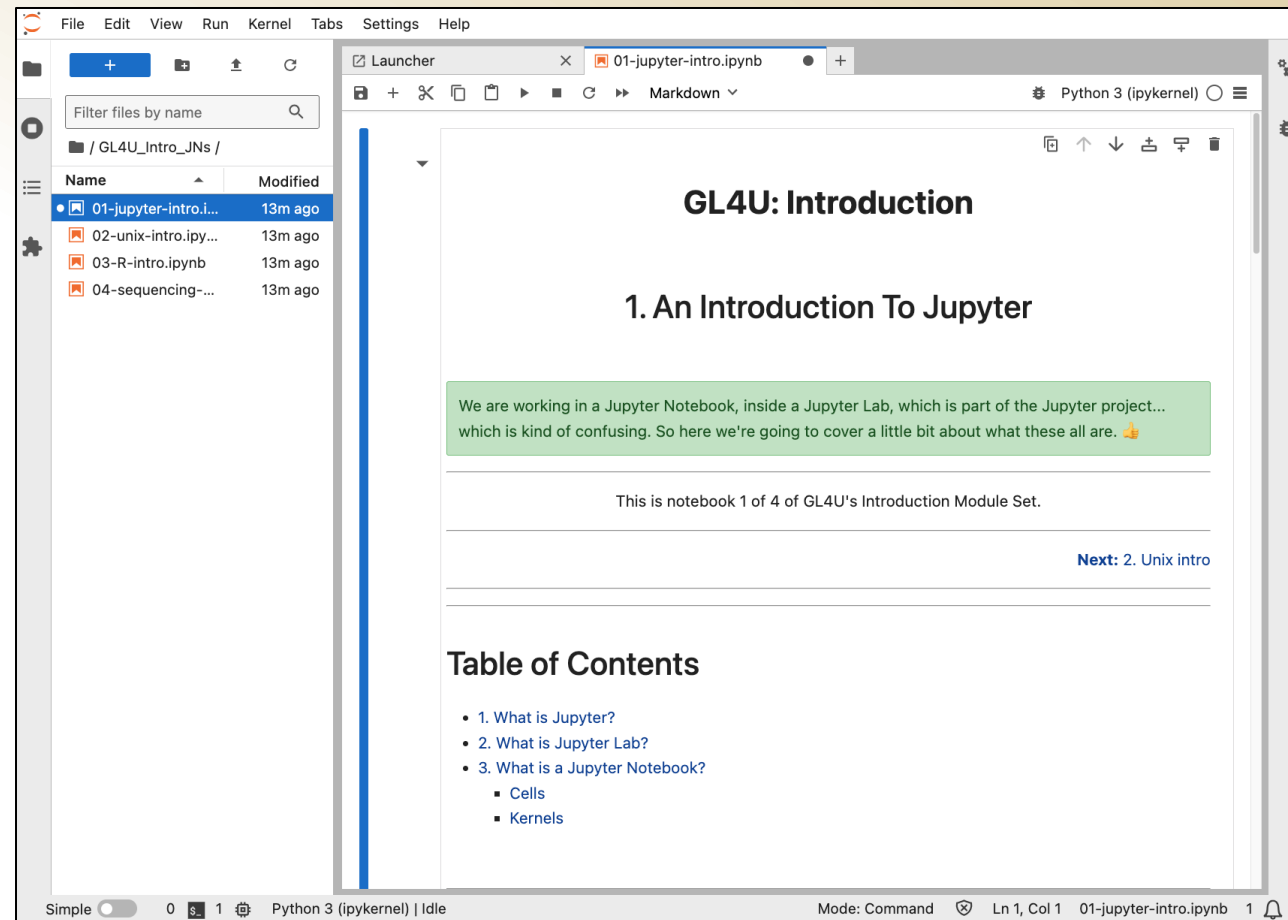
Like Share Download Clip Save

GL4U: Introduction

NASA GeneLab - 2 / 4

- GL4U: Intro Lecture 2of4 OSDR GL Overview 2024
- GL4U: Intro Lecture 1of4 NASA SMD SB Overview 2024
- GL4U: Intro Lecture 4of4 Short Read Sequencing Overview 2024
- GL4U: Intro Lecture 3of4 CL R Jupyter Intro 2024

Intro to NASA, SMD, Space Bio



GL4U: Introduction

1. An Introduction To Jupyter

We are working in a Jupyter Notebook, inside a Jupyter Lab, which is part of the Jupyter project... which is kind of confusing. So here we're going to cover a little bit about what these all are. 🙌

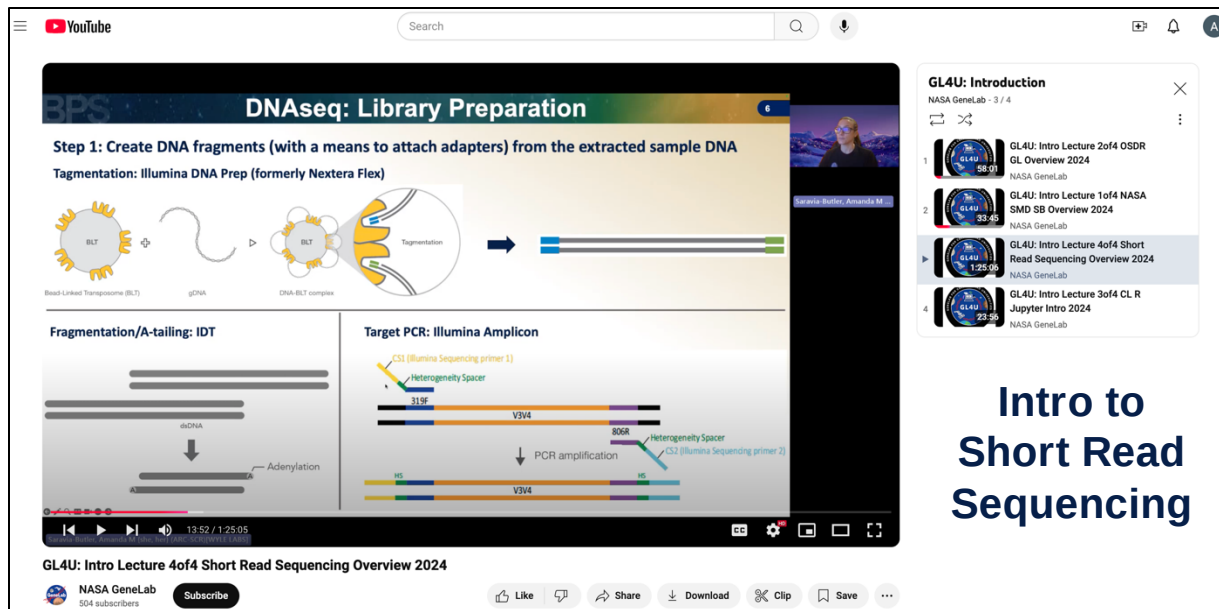
This is notebook 1 of 4 of GL4U's Introduction Module Set.

[Next: 2. Unix intro](#)

Table of Contents

- 1. What is Jupyter?
- 2. What is Jupyter Lab?
- 3. What is a Jupyter Notebook?
 - Cells
 - Kernels

Simple 0 1 Python 3 (ipykernel) | Idle Mode: Command Ln 1, Col 1 01-jupyter-intro.ipynb 1



DNAseq: Library Preparation

Step 1: Create DNA fragments (with a means to attach adapters) from the extracted sample DNA

Tagmentation: Illumina DNA Prep (formerly Nextera Flex)

Fragmentation/A-tailing: IDT

Target PCR: Illumina Amplicon

GL4U: Intro Lecture 4of4 Short Read Sequencing Overview 2024

NASA GeneLab 504 subscribers

Like Share Download Clip Save

GL4U: Introduction

NASA GeneLab - 3 / 4

- GL4U: Intro Lecture 2of4 OSDR GL Overview 2024
- GL4U: Intro Lecture 1of4 NASA SMD SB Overview 2024
- GL4U: Intro Lecture 4of4 Short Read Sequencing Overview 2024
- GL4U: Intro Lecture 3of4 CL R Jupyter Intro 2024

Intro to Short Read Sequencing

Recorded lectures on YouTube

YouTube

Search

RNA Sequencing Overview

Space
Ground

Samples of Interest

Isolate RNAs

Generate cDNA, fragment, size select, add linkers

Ribo-depletion or Poly(A) tail selection

Library QC

Sequence ends

100s of millions of paired reads
10s of billions bases of sequence

Remove low quality sequence data

Map to Genome, transcriptome and predicted exon junctions

Transcript
Short reads

Short reads split by intent

Downstream analysis

GL4U: RNAseq Lecture 1of2 RNAseq Overview part1of3 2024

NASA GeneLab
504 subscribers

Like Share Download Clip Save

GL4U: RNA Sequencing
NASA GeneLab - 4 / 4

- GL4U: RNAseq Lecture 1of2 RNAseq Overview part3of3 2024 NASA GeneLab 23:02
- GL4U: RNAseq Lecture 1of2 RNAseq Overview part2of3 2024 NASA GeneLab 1:28:38
- GL4U: RNAseq Lecture 2of2 Statistics Overview 2024 NASA GeneLab 1:05:05
- GL4U: RNAseq Lecture 1of2 RNAseq Overview part1of3 2024 NASA GeneLab 1:14:10

RNAseq Overview

YouTube

Search

Normalization with Median of Ratios Method

2. For every gene in a sample, the ratios (sample/pseudo-reference) are calculated. This is performed for each sample in the dataset. Since the majority of genes are not differentially expressed, the majority of genes in each sample should have similar ratios within the sample.

gene	Sample1	Sample2	pseudo-reference sample	ratio of Sample1/ref	ratio of Sample2/ref
EF2A	1489	906	1161.5	$1489/1161.5 = 1.28$	$906/1161.5 = 0.78$
ABCD1	22	13	16.9	$22/16.9 = 1.30$	$13/16.9 = 0.77$
MEPV	793	410	570.2	$793/570.2 = 1.39$	$410/570.2 = 0.72$
BAG1	76	42	56.5	$76/56.5 = 1.35$	$42/56.5 = 0.74$
MOV10	521	1196	883.7	$521/883.7 = 0.590$	$1196/883.7 = 1.35$

GL4U: RNAseq Lecture 2of2 Statistics Overview 2024

NASA GeneLab
504 subscribers

Like Share Download Clip Save

GL4U: RNA Sequencing
NASA GeneLab - 3 / 4

- GL4U: RNAseq Lecture 1of2 RNAseq Overview part3of3 2024 NASA GeneLab 23:02
- GL4U: RNAseq Lecture 1of2 RNAseq Overview part2of3 2024 NASA GeneLab 1:28:38
- GL4U: RNAseq Lecture 2of2 Statistics Overview 2024 NASA GeneLab 1:05:05
- GL4U: RNAseq Lecture 1of2 RNAseq Overview part1of3 2024 NASA GeneLab 1:14:10

Statistics Overview

Hands-on activities via GitPod

File Edit View Run Kernel Tabs Settings Help

Filter files by name

/ GL4U_RNAseq_JNs /

01-RNAseq_pro... 20m ago
02-RNAseq_ana... 20m ago
03-RNAseq_ana... 20m ago

GL4U: RNAseq

Pipeline for Processing GeneLab RNA-sequencing Data

The 2 GL4U RNAseq Jupyter Notebooks (JNs) are designed to teach students how to process RNA sequencing data using the GeneLab standard pipeline. Below are step-by-step instructions for processing samples derived from the soleus (aka "anti-gravity") muscle of mice that were either flown in space aboard the International Space Station (ISS) (spaceflight, FLT, animals) or kept in an environmental simulator on Earth to serve as ground controls (GC, animals) during NASA's Rodent Research - 1 mission. More information about the samples analyzed here can be found in the [Open Science Data Repository](#) under OSD-104.

RNAseq Pipeline Overview

This JN will cover the pipeline steps outlined in red. For more information about how GeneLab processes bulk RNAseq data, visit the [GeneLab Data Processing GitHub repository](#).

```

graph LR
    RawReads[Raw Reads] --> TrimFilter[Trim/Filter Raw Data  
(TrimGalore!)]
    TrimFilter --> AlignReads[Align Reads  
(STAR)]
    AlignReads --> EvaluateStrandedness[Evaluate Strandedness  
(RSeQC - Infer Exp)]
    EvaluateStrandedness --> CountAlignedReads[Count Aligned Reads  
(RSEM)]
    CountAlignedReads --> NormalizeCounts[Normalize Counts  
(DESeq2)]
    NormalizeCounts --> DGE[DGE  
(DESeq2)]
    DGE --> AddGeneAnnotations[Add Gene Annotations]
    AddGeneAnnotations --> DataVisualization[Data Visualization]
    DataVisualization --> PCA[PCA]
    DataVisualization --> VolcanoPlot[Volcano Plot]
    DataVisualization --> Heatmap[Heatmap]
    DataVisualization --> GSEA[GSEA]
  
```

1. RNAseq processing: fastq to counts

Here we are going to setup a directory structure to store the output data we'll generate, then we will follow the steps in the GeneLab standard RNAseq processing pipeline to generate raw count data. All of these steps are most easily done at a Unix-like command line, so this notebook uses a "Bash" kernel, a common language used in a Unix-like environment.

This is notebook 1 of 2 of GL4U's RNAseq Module Set. It is expected that GL4U's Introduction Module Set has been completed already.

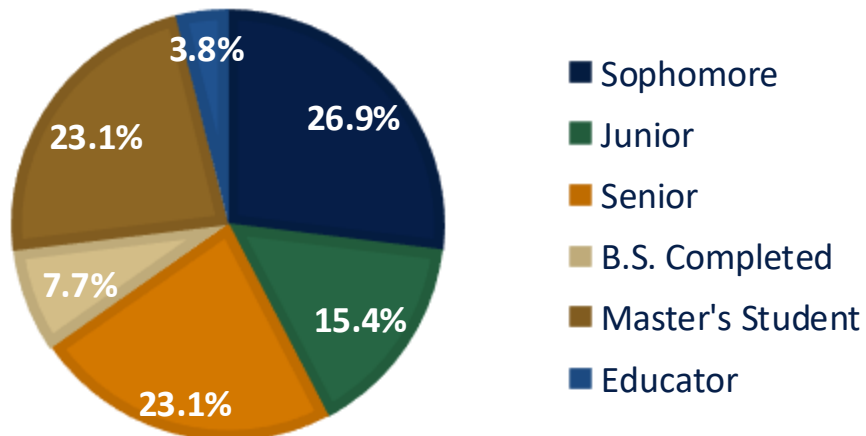
Table of Contents

- 0. Setup
- 1. Raw Data Quality Control (QC)
 - 1a. Raw Data QC with FastQC

Next: 2. RNAseq analysis

GL4U: RNAseq Virtual Bootcamp, 9/2024

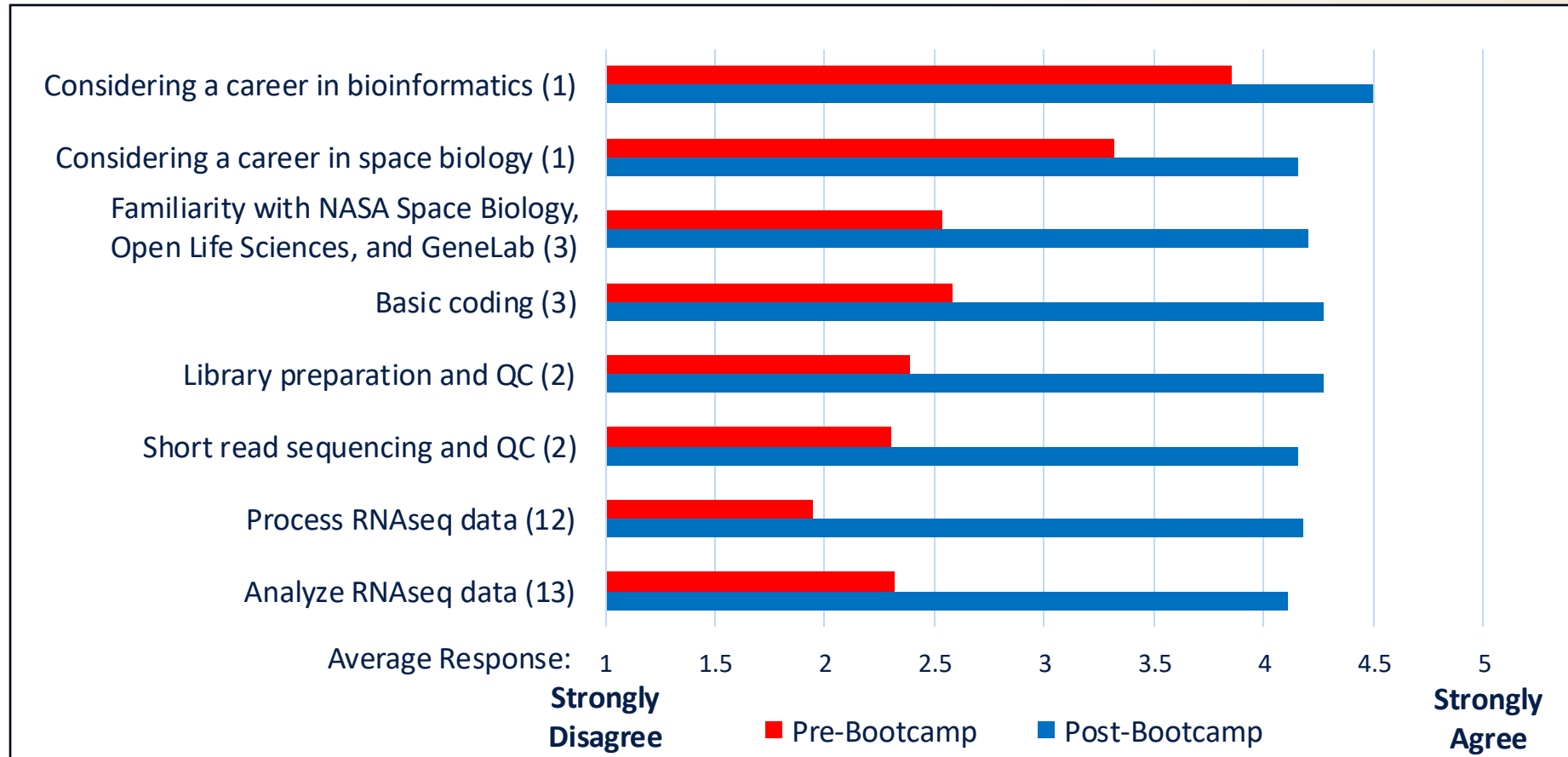
- Virtual 5-day bootcamp, 9/27 – 9/29; 10/4 – 10/5
- 34 students and 2 educators from 4 institutions
 - 14 students, 2 professors from AAMU
 - 10 students from CSULA
 - 1 student from UCSC
 - 9 students from UNM
- *Compute resources*: GitPod free compute resources
- Bootcamp covered GL4U Introduction RNA Sequencing modules



“I learned more about bioinformatics than I have in any other higher education class. This will greatly improve my comprehension when reading genetics papers. I cannot emphasize enough how valuable this workshop was, **it needs to continue for other students**, and I will absolutely be looking to attend more workshops such as this in the future.”

“This was an amazing experience. Amanda was terrific. It really opened up my eyes to space biology and really showed me this is something I would love to be able to pursue.”

“When I first took the survey, I was intimidated by some of the topics because I was unfamiliar with them, but after this bootcamp I have a better understanding of RNA sequencing. Dr. Saravia has inspired me to look into space biology careers, a career field I was not considering prior to this bootcamp.”



GL4U: Introduction

National Aeronautics and Space Administration



Certificate of Completion

This certifies that on the
5th day of October, 2024

FirstName LastName

has successfully completed the OSDR
GL4U: Introduction Module Set

Samrawit Gebre

Samrawit Gebre
Open Science Data Repository, Project Manager
NASA Ames Research Center

GL4U: RNA Sequencing

National Aeronautics and Space Administration



Certificate of Completion

This certifies that on the
5th day of October, 2024

FirstName LastName

has successfully completed the OSDR
GL4U: RNA Sequencing Module Set

Samrawit Gebre

Samrawit Gebre
Open Science Data Repository, Project Manager
NASA Ames Research Center

Open Science Analysis Working Groups (AWGs)



[Join an AWG!](#)

ANIMAL

226 members



MULTI-OMICS

351 members



MICROBIAL

216 members



PLANTS

176 members



AI/ML

323 members



ALSDA

(Physiological/BioMedical)



224 members

FEMALE REPRO

88 members



HUMAN

39 members



RADLAB

67 members



Collaborate on Data Mining/Publications



Cell Press Package 2020:
The Biology of Spaceflight
<https://www.cell.com/c/the-biology-of-spaceflight>



Nature Portfolio Collection 2024:
Space Omics and Medical Atlas across orbits (SOMA)
<https://www.nature.com/collections/ebdbcahdgc>

AWG Forum User Visits



Enroll on Canvas: <https://canvas.instructure.com/enroll/63FHH6>



GL4U: Intro > Syllabus

Home

Syllabus

Modules

Assignments

Files

Discussions

Grades

People

Lucid (Whiteboard)

Course Syllabus

Part I: Pre-Course Survey

Part II: Introduction to NASA Open Life Sciences

1. Intro to NASA, Science Mission Directorate, and Space Biology
2. Intro to NASA's Open Science Data Repository and GeneLab

Part III: Coding, The Basics

1. Setting up GitPod
2. Intro to Jupyter Lab environment
3. Intro to coding in a Unix-based command-line interface
4. Intro to coding in R

Part IV: Short Read Sequencing Overview

1. Library preparation overview
2. Sequencing by synthesis
3. Sequencing data quality assessment

Part V: Post-Course Survey

GL4U: Intro > Modules

Home

Syllabus

Modules

Assignments

Files

Discussions

Grades

People

Lucid (Whiteboard)

Collapse All

▼ MODULE 1: Pre-Course Survey

Complete All Items

GL4U Introduction On-Demand Pre-Course Survey

0 pts Submit

▼ MODULE 2: Introduction to NASA Sciences

Prerequisites: MODULE 1: Pre-Course Survey

Complete All Items

LECTURE 1: Introduction to NASA, Science Mission Directorate, and Space Biology

View

LECTURE 2: Introduction to NASA Open Life Sciences and GeneLab

View

▼ MODULE 3: Coding, The Basics

Prerequisites: MODULE 1: Pre-Course Survey, MODULE 2: Introduction to NASA Sciences

Complete All Items

LECTURE 3: Introduction to the Command Line, R, and Jupyter

View

HANDS-ON ACTIVITY 1: Getting Started With GitPod

0 pts Submit

HANDS-ON ACTIVITY 2: Jupyter Introduction

0 pts Submit

HANDS-ON ACTIVITY 3: Unix Introduction

GL4U: Intro Course Available On Canvas



Enroll on Canvas: <https://canvas.instructure.com/enroll/63FHH6>

Account

Dashboard

Courses

Calendar

Inbox

History

Help

GL4U: Intro > Discussions

- Home
- Syllabus
- Modules
- Assignments
- Files
- Discussions**
- Grades
- People
- Lucid (Whiteboard)

All

Search by title or author...

Add Discussion

Settings

Pinned Discussions

Questions about MODULE 1?

Questions about MODULE 2?

Questions about MODULE 3?

Questions about MODULE 4?

Questions about MODULE 5?

Discussions

Ordered by Recent Activity

GL4U: RNAseq Course Available On Canvas

Enroll on Canvas: <https://canvas.instructure.com/enroll/KYXWN6>



GL4U: RNAseq > Syllabus

Home

Syllabus

Modules

Assignments

Files

Discussions

Grades

People

Lucid (Whiteboard)

Course Syllabus

Part I: Pre-Course Activities

1. Verify Successful Completion of GL4U: Introduction Course
2. Pre-Course Survey

Part II: RNA Sequencing Overview and Data Pre-processing

1. What came before RNAseq
2. RNAseq experimental design
3. Raw data QC
4. Trim/filter data and QC

Part III: RNAseq Data Processing

1. Alignment and QC
2. Assess strandedness
3. Quantitation and QC

Part IV: RNAseq Data Analysis

1. Statistics overview
2. Count data normalization
3. Differential gene expression analysis
4. Data visualization

Part V (Optional): Apply Your Knowledge

1. Data analysis on any OSD dataset

Part VI: Post-Course Survey

GL4U: RNAseq > Modules

Home

Syllabus

Modules

Assignments

Files

Discussions

Grades

People

Lucid (Whiteboard)

▼ MODULE 1: Pre-Course Activities

Complete All Items

Verify Completion of GL4U: Introduction Course

0 pts Submit

GL4U RNAseq On-Demand Pre-Course Survey

0 pts Submit

▼ MODULE 2: RNA Sequencing Overview and Preprocessing

Prerequisites: MODULE 1: Pre-Course Activities

Complete All Items

LECTURE 1: RNA Sequencing Overview Part 1 of 3

View

HANDS-ON ACTIVITY 1: RNAseq Pre-processing

0 pts Submit

▼ MODULE 3: RNAseq Data Processing

Prerequisites: MODULE 1: Pre-Course Activities, MODULE 2: RNA Sequencing Overview and Preprocessing

Complete All Items

LECTURE 2: RNA Sequencing Overview Part 2 of 3

View

HANDS-ON ACTIVITY 2: RNAseq Processing

0 pts Submit

Enroll on Canvas: <https://canvas.instructure.com/enroll/KYXWN6>

<https://nasa.edgebioinformatics.org/home>

File Edit View Run Kernel Tabs Settings Help

03-RNAseq_analysis_anyOx

Launcher

Filter files by name

/ GL4U_RNAseq_JNs /

Name	Modified
01-RNAseq_pro...	17m ago
02-RNAseq_ana...	17m ago
03-RNAseq_ana...	17m ago

RNAseq Pipeline Overview

This JN will cover the pipeline steps outlined in red. For more information about how GeneLab processes bulk RNAseq data, visit the [GeneLab Data Processing GitHub repository](#).

```

graph LR
    RawReads[Raw Reads] --> Trim[Trim/Filter Raw Data  
(TrimGalore!)]
    RawReads --> RawDataQC[Raw Data QC  
(FastQC -> multiQC)]
    Trim --> TrimDataQC[Trimmed Data QC  
(FastQC -> multiQC)]
    Trim --> Align[Align Reads  
(STAR)]
    Align --> EvalStrand[Evaluate Strandedness  
(RSeQC - Infer Exp)]
    Align --> CountAlign[Count Aligned Reads  
(RSEM)]
    EvalStrand --> InferExpQC[Infer Exp Data QC  
(multiQC)]
    CountAlign --> RSEMDataQC[RSEM Data QC  
(multiQC)]
    CountAlign --> NormCounts[Normalize Counts  
(DESeq2)]
    NormCounts --> DGE[DGE  
(DESeq2)]
    DGE --> AddGene[Add Gene Annotations]
    AddGene --> PCA[PCA]
    AddGene --> DataVis[Data Visualization]
    AddGene --> Volcano[Volcano Plot]
    AddGene --> Heatmap[Heatmap]
    AddGene --> GSEA[GSEA]
  
```

3. RNAseq analysis of any OSD/GLDS dataset: DGE, annotations, and data visualization

Here we are going to download and format the input metadata and data files from any RNAseq dataset on OSD that contains the *genes.results output files from RSEM. We will next setup a directory structure to store the output data we'll generate. Then we will follow the steps in the GeneLab standard RNAseq processing pipeline to perform differential gene expression (DGE), add annotations, and generate data visualizations.

This is a modified version of the 2nd notebook of GL4U's RNAseq Module Set. It is expected that both the GL4U Introduction Module Set and the RNAseq Module Set have been completed already.

[Previous: 2. RNAseq analysis](#)

Table of Contents

- 0. Create Directory Structure for Processed Data
- 1. Download the Metadata and Data
 - 1a. Set up Directory Path Variables

GeneLab EDGE

My Projects My uploads Job Queue

AMS

Home

Public Projects

Upload Files

Run Workflow

EDGE bioinformatics is an open-source bioinformatics platform with a user-friendly interface that allows scientists to perform a number of bioinformatics analyses using state-of-the-art tools and algorithms. NASA EDGE takes an updated EDGE Bioinformatics framework and has only the NASA GeneLab Illumina amplicon sequencing data processing pipeline (AmplIllumina) integrated.

```

graph LR
    RawData[Raw data fastQC] --> Primer[Primer removal]
    Primer --> Quality[Quality filter/trim]
    Quality --> Filtered[Filtered data fastQC]
    Filtered --> Resolve[Resolve ASVs]
    Filtered --> Generate[Generate OTUs]
    Resolve --> Assign[Assign taxonomy]
    Generate --> Assign
  
```

Primary outputs:

- Recovered ASV/OTU sequences (fasta)
- Counts of sequences per sample
- Taxonomic classifications of sequences
- Alpha and beta diversity plots
- Taxonomic bar charts
- Differential abundance analysis and visualizations

Los Alamos NATIONAL LABORATORY

[Terms of Use, Privacy](#)

- Awarded ROSES F.7: Support for Open-Source Tools, Frameworks, and Libraries (3 years)

➤ **Enhancing Analysis Capabilities of Biological Data With the NASA EDGE Bioinformatics Platform**

Pending funding...

- Video-recorded walk-throughs of hands-on activities
- Automatic assignment grading
- Automatic GL4U certification upon course completion
- Set-up stable Jupyter Lab environment
 - GitPod has a 30min time-out, requiring users to re-run all previous commands upon GitPod restart
- Host live (weekly / biweekly) office hours
- Add additional omics datatypes (e.g. Amplicon seq, Metagenomics, scRNAseq, etc.)
- Expand internationally
- Incorporate NASA EDGE Bioinformatics training
- Continue to host live annual bootcamps

Open Science for Life in Space Teams



Open Science Analysis Working Group Members

Support



NASA Space Biology Program
NASA Science Mission Directorate
NASA Human Research Program
NASA Biological and Physical Sciences

Questions?

For more info about GL4U:



Sign up for the GL4U mailing list

- Stay up-to-date on future GL4U events / bootcamps:
Send an e-mail to GL4U-join@lists.nasa.gov
with the Subject: **subscribe**

Enroll in GL4U: Intro
On Canvas



Enroll in GL4U: RNAseq
On Canvas

