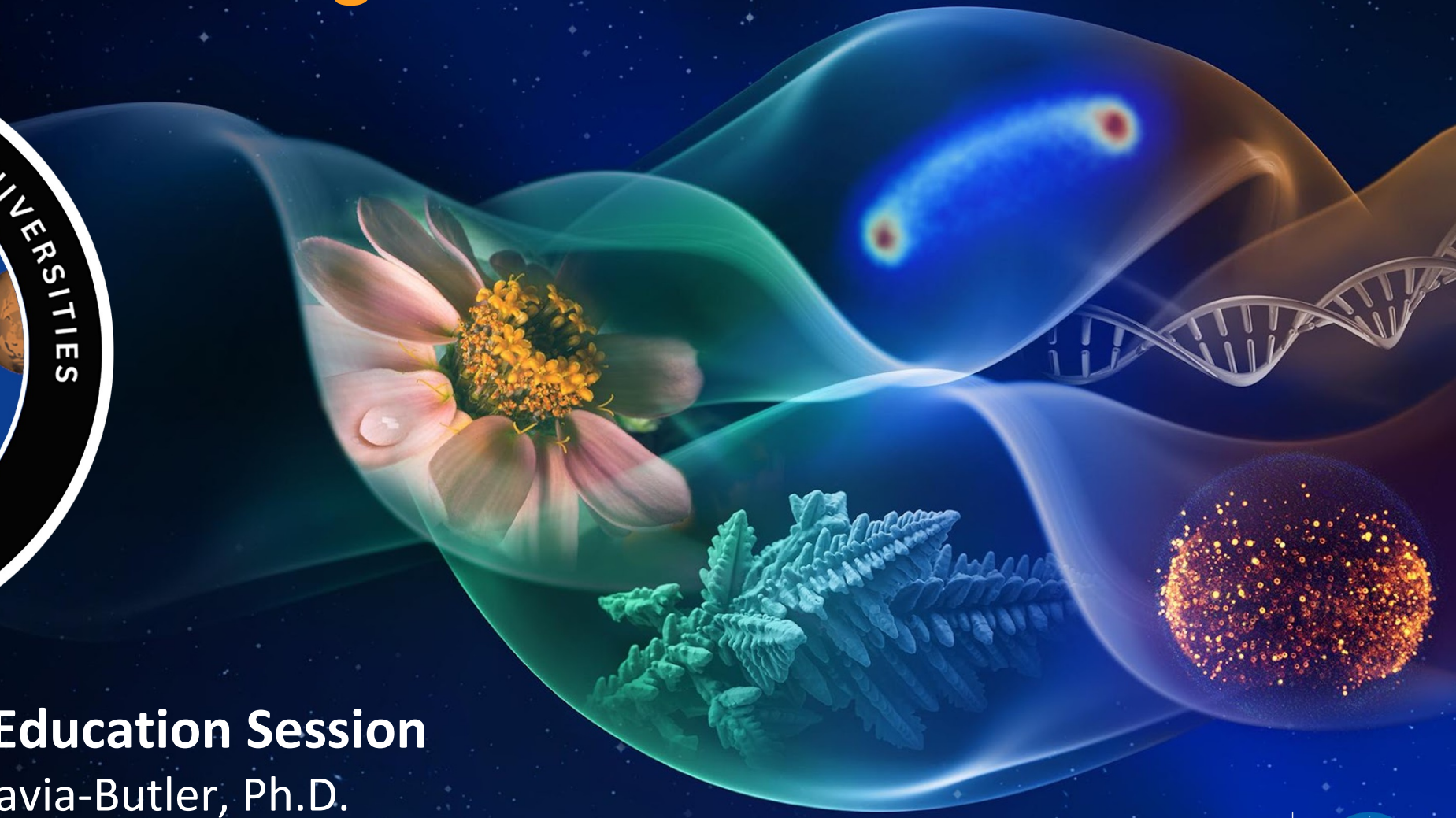
GL4U: Using Space Biology Omics Data to Provide Bioinformatics Training for Students and Educators



2023 ASGSR Education Session

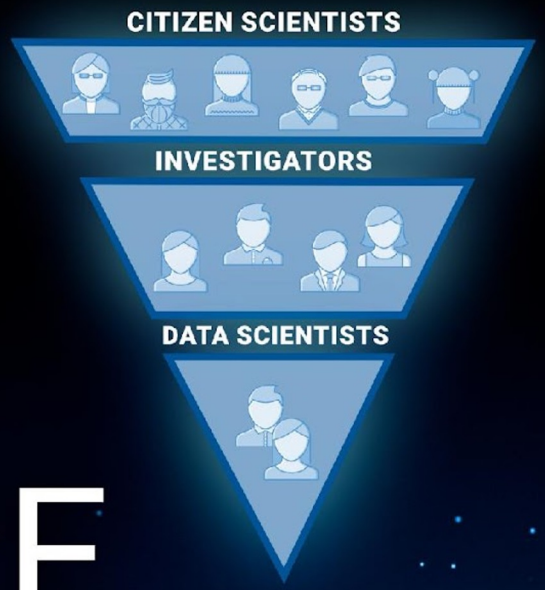
Amanda M. Saravia-Butler, Ph.D.
Science, Data Processing, and GL4U Lead, NASA GeneLab
Contractor: KBR

National Aeronautics and Space Administration

OSDR/GeneLab Overview

TISSUE REPOSITORY

SCIENTIFIC COMMUNITY



OPEN SCIENCE

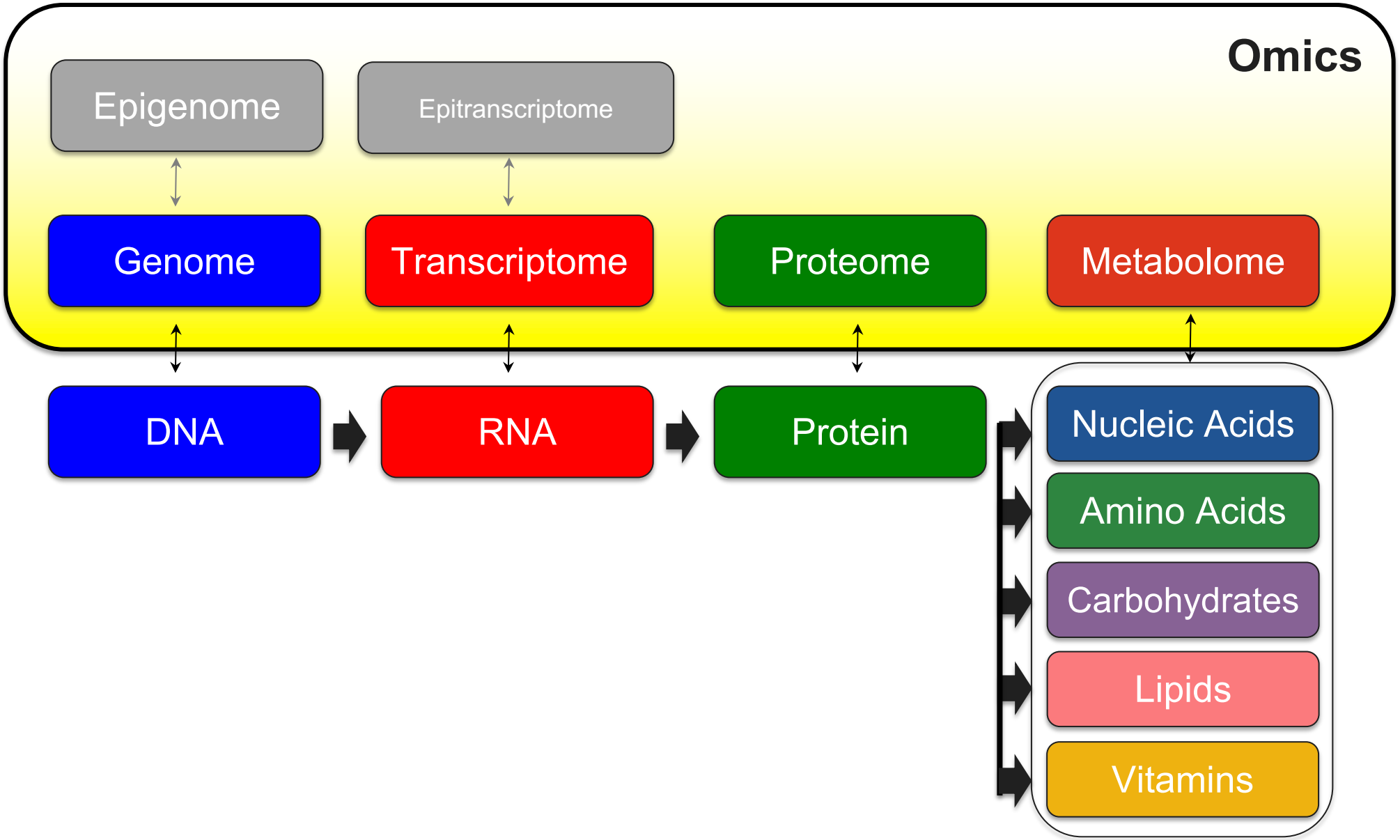
DATA REPOSITORY

GENELAB
OMICS

+

ALSDA
PHENOTYPIC

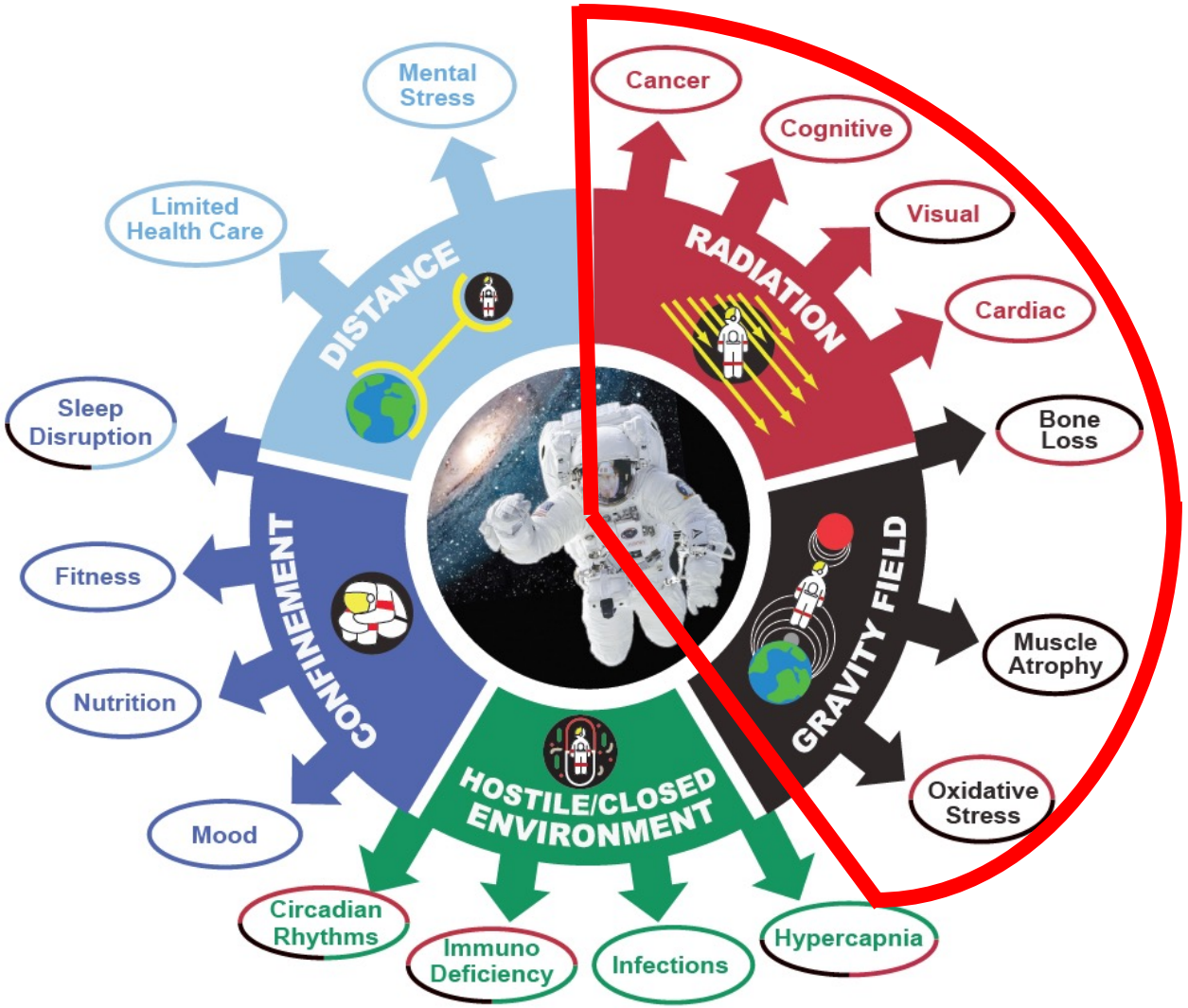
What Are Omics?



Why Is Studying Omics Important For Spaceflight?

- What, when, and where genes are expressed allow for cell type diversity and enable living organisms to respond and adapt to surroundings
- Gene expression is primarily regulated by environmental factors both micro (cell's micro-environment) and macro (organism's external stimuli or stressors)
- Spaceflight alters the transcriptional patterns and molecular signaling networks within our cells, which in turn causes physiological changes
- Understanding such changes will enable development of mitigation strategies to better withstand the rigors of long-duration spaceflight

Primary Stressors of Spaceflight



Interpreting Omics Data

Raw Sequence Data

```
@A00654:12:HFY5YDSXX:1:1101:2121:1000_CTCGCGTTT 1:N:0:NGGAGAGT+TCCTACCT
AATATCAGTGATATTTAGAAACCACATAGTAAGCTAACAACTAATAATGGAATGGTTTTAATATCCTGTGACAAGTTAATGTGGATACTATGCGGTCTTCCTTAAAATGCTGTATGGTACTGTCCTCACCTCTTCTTTGTGCTGCTGTA
+
FFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFF:FFFFF:FF:FFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFF
@A00654:12:HFY5YDSXX:1:1101:5466:1000_CCTCCCTCG 1:N:0:NGGAGAGT+TCCTACCT
GGCGGGTCTGAGCAAGAACACTCTGCCATCGCCGGCTCGGGACGGGAGGCCCACCCGTGGCCGAGCACGCAGACCTGGATGAACCGAGGCCTCGATGGGCCTGGACGAAGGGGAGAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF,F:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00654:12:HFY5YDSXX:1:1101:14109:1000_TGATCTACG 1:N:0:NGGAGAGT+TCCTACCT
TGTCACTACAACCAGCTGTGCCTGTGCTATTGCAGTTACACAGTGTCACTACAACCAACTGTGCCTGTGCTATTGCAGTTACACAGTGTCACTACAACCAGCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGGAGAGTTGATC
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF,:FFFFFFF:FF:F:FF:FFFFFFF:FFFFFFFF:FFFFFFFFFF:FFFFFFFFFFFF:
```

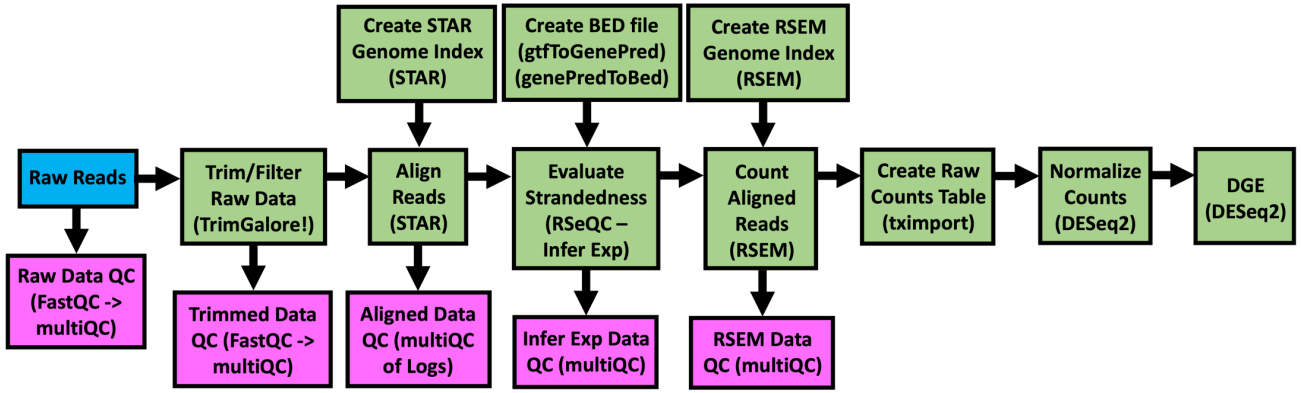
Processed RNA Sequence Data: Differential Gene Expression

SYMBOL	GENENAME	Group.Mean_(FLT)	Group.Mean_(GC)	Log2fc_(FLT)v(GC)	P.value_(FLT)v(GC)	Adj.p.value_(FLT)v(GC)
Timm17a	translocase of inner mitochondrial membrane 17a	3.91	15.11	-2.67	0.00	0.00
NA	NA	166.47	108.32	0.63	0.00	0.00
NA	NA	325.78	171.64	0.93	0.00	0.00
Dnajc7	DnaJ heat shock protein family (Hsp40) member C7	22.31	42.16	-0.95	0.00	0.00
Slc15a4	solute carrier family 15, member 4	2.75	9.38	-2.32	0.00	0.00
Ckap5	cytoskeleton associated protein 5	18.24	46.64	-1.39	0.00	0.00
NA	NA	104.49	67.33	0.64	0.00	0.00

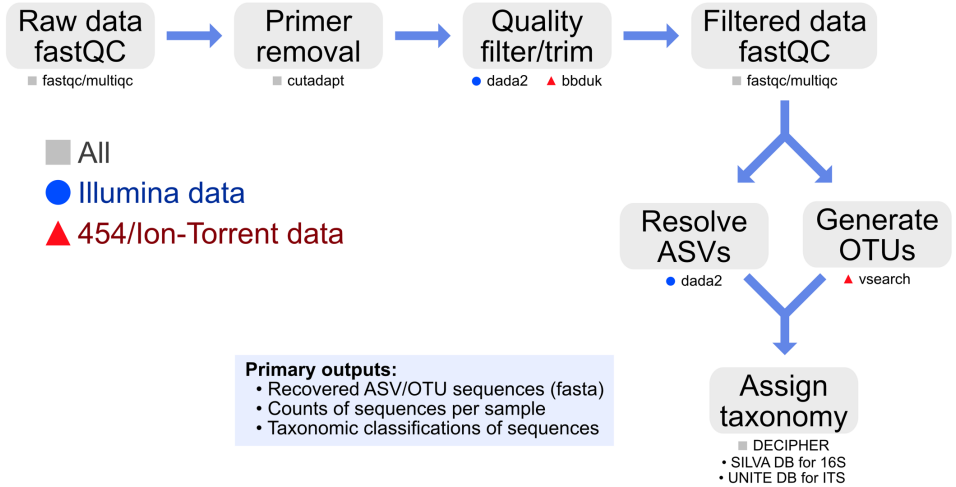
GeneLab Data Processing Pipelines

Build consensus data processing pipelines with the scientific community

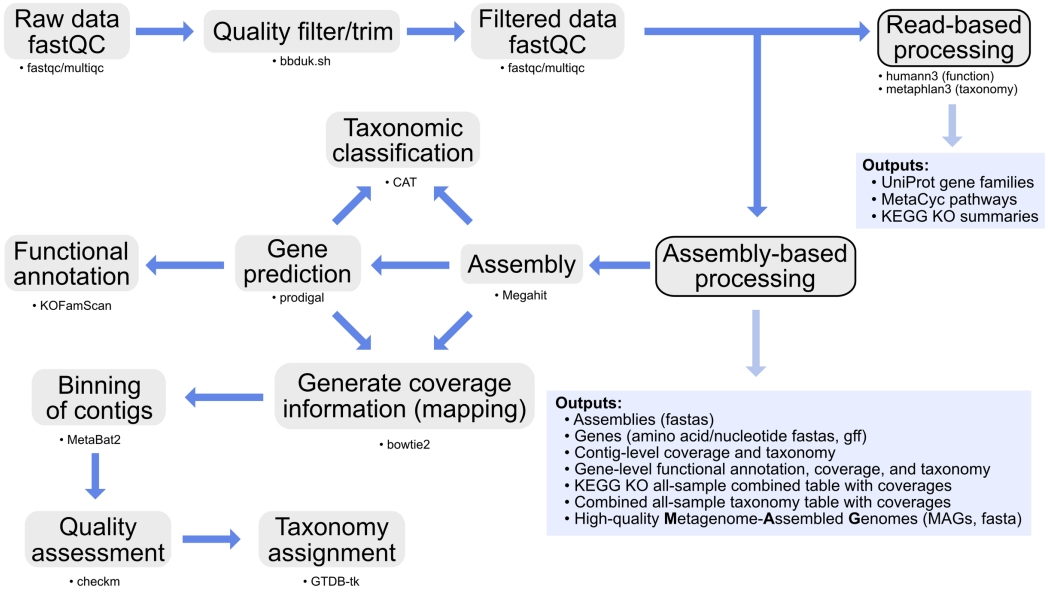
RNA Sequencing Data



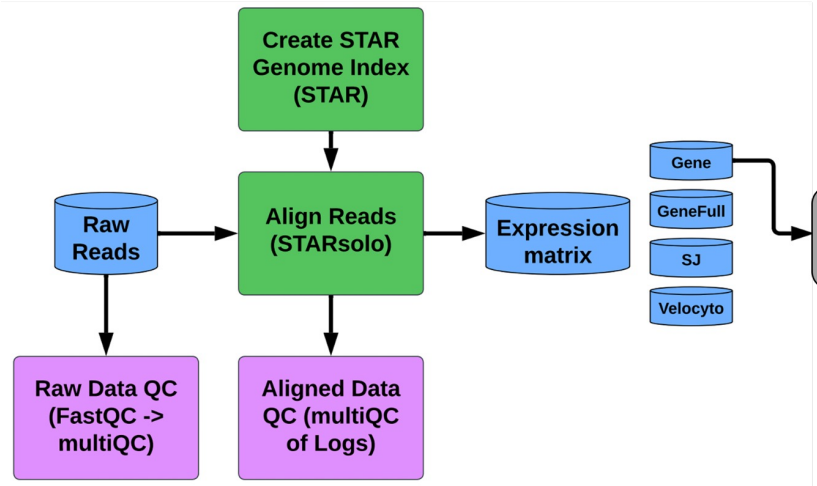
Amplicon Sequencing Data



Metagenomics Data



Single Cell RNA Sequencing Data



Open Science Data Repository (https://osdr.nasa.gov/bio/repo/)

Open Science for Life in Space

Home | About | Data & Tools | Working Groups | Help

General Search Filters

Data Source

- [x] GeneLab
- [x] ALSDA
- [ ] NIH GEO
- [ ] EBI PRIDE
- [ ] ANL MG-RAST

Data Type

- [x] Study
- [ ] Experiment
- [ ] Subject
- [ ] Biospecimen
- [ ] Payload

Show more

Study Search Filters

Project Type

- [ ] Ground
- [ ] Spaceflight
- [ ] High Altitude

Assay Type

- [ ] Amplicon Sequencing Assay
- [ ] Bisulfite Sequencing
- [ ] ChIP-Seq
- [ ] Behavior (Gait)
- [ ] Gel Electrophoresis

Show more

Organism

Open Science Data Repository Search

Search Datasets

Sort By: Release Date

Items per page: 25 | 1 – 25 of 435

Study

OSD-385

Persistence of Escherichia coli in the microbiomes of red Romaine lettuce (Lactuca sativa cv. 'Outredgeous')- does seed sanitization matter?

Organisms	Factors	Assay Types	Release Date	Description
Microbiota	Treatment Seed Sanitization Tissue	Amplicon Sequencing	19-Apr-2024	Seed sanitization via chemical processes removes/reduces microbes from the external surfaces of the seed and thereby could have an impact on the plants,health or productivity. To determine the impact ...

Highlights: *cgene*

Study

OSD-580

Transcriptional profiling of heart tissue from mice flown on the RRRM-2 mission

Organisms	Factors	Assay Types	Release Date	Description
Mus musculus	Spaceflight Age Euthanasia Location	transcription profiling	03-Jan-2024	In the Rodent Research Reference Mission (RRRM-2), forty female C57BL/6NTac mice were flown on the International Space Station. To assess differences in outcomes due to age, twenty 12 week-old and twe...

Highlights: *cgene*

Study

OSD-576

Transcriptional profiling of tibialis anterior muscle from mice flown on the RR-23 mission

Organisms	Factors	Assay Types	Release Date	Description
Mus musculus	Spaceflight	transcription profiling	12-Dec-2023	The objective of the Rodent Research-23 mission (RR-23) was to better understand the effects of spaceflight on the eyes, specifically on the structure and function of the arteries, veins, and lymphati...

Highlights: *cgene*

Study

OSD-524

Ionizing radiation induces transgenerational effects of DNA methylation in zebrafish

Organisms	Factors	Assay Types	Release Date	Description
Danio rerio	Ionizing Radiation Generation	DNA methylation profiling	31-Aug-2023	Ionizing radiation is known to cause DNA damage, yet the mechanisms underlying potential transgenerational effects of exposure have been scarcely studied. Previously, we observed effects in offspring ...

GeneLab Data Processing GitHub Repo

Search or jump to... / Pull requests Issues Marketplace Explore

nasa / GeneLab_Data_Processing Public

Edit Pins

README.md

Open Science for Life in Space

GeneLab_Data_Processing

About

The NASA GeneLab Data Processing team and Analysis Working Group members have created standard pipelines for processing omics data from spaceflight and space-relevant experiments. This repository contains the processing pipelines that have been standardized to date for the assay types indicated below. Each subdirectory in this repository holds current and previous pipeline versions for the respective assay type, including detailed descriptions and processing instructions as well as the exact processing commands used to generate processed data for datasets hosted in the GeneLab Data Repository.

Assay Types

Click on an assay type below for data processing information.

- Create GeneLab Reference Annotations
- Amplicon Sequencing
  - Illumina
  - 454 and Ion-Torrent
- Metagenomics
  - Removing human reads
  - Illumina
- (bulk) RNAseq
- single cell RNAseq

https://github.com/nasa/GeneLab_Data_Processing

GeneLab for Colleges and Universities (GL4U)
Overview

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** to:
  - ➢ Increase accessibility and interpretability of multi-omics Space Biology data
  - ➢ Train a diverse next generation of Space Biology researchers
  - ➢ Enhance awareness and understanding of Space Biology data

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 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 Content Design

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 × 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)

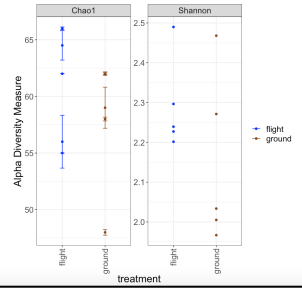
Looking at the plots above:

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_tab$color)) +
    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

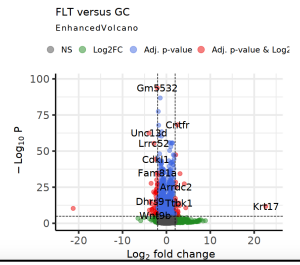
Finally, let's make a volcano plot to identify a few interesting genes. A volcano plot is a scatterplot which shows the relationship of the adjusted p-value to the log2 fold change. Genes with large fold changes that are also statistically significant by adjusted p-value are labeled.

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

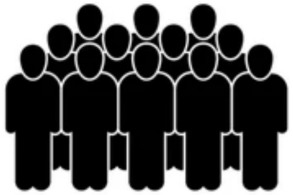
```
In [68]: # Volcano plot showing genes differentially expressed in FLT vs GC
EnhancedVolcano(DGE_output_table,
    lab = DGE_output_table$SYMBOL,
    x = 'Log2fc_(FLT)v(GC)',
    y = 'Adj.p.value_(FLT)v(GC)',
    title = 'FLT versus GC',
    legendLabels=c('NS','Log2FC','Adj. p-value',
      'Adj. p-value & Log2FC'),
    pCutoff = 10e-6,
    FCcutoff = 2,
    pointSize = 3.0,
    labSize = 6.0,
    colAlpha=0.5)

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



GL4U Approach

GL4U Direct Approach

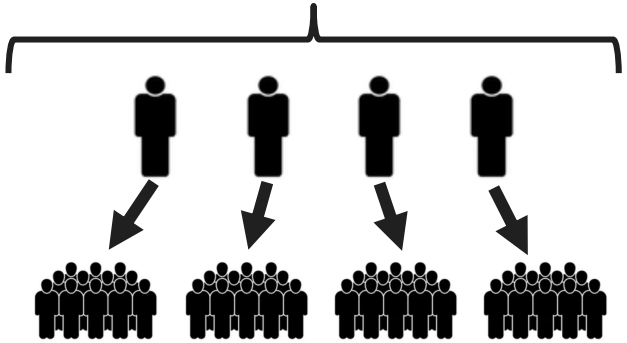


MSI/HBCU Students

- Space biology-relevant training in bioinformatics
- Uses **direct** (training students) and **indirect** (training educators) approaches
- The GL4U Introduction module and one omics-specific module is taught to students or educators during a 1-2 week-long bootcamp
- Training focused on MSIs and HBCUs
- Compute resources are provided to all participants
  - ➢ For indirect bootcamps: educators receive materials and training to enable them to run the bootcamp at their home institutions
- All GL4U modules are made publicly available on GitHub (https://github.com/nasa/GeneLab-Training/tree/main/GL4U) and virtual bootcamps are recorded to enable independent learning

GL4U Indirect Approach

MSI/HBCU Educators



GL4U RNAseq Certification

GL4U Introduction Module
- ☑ Pre-Intro Module Survey
- ☑ Intro Lectures
- ☑ Intro JNs
- ☑ Post-Intro Module Survey

+

GL4U RNAseq Module
- ☑ Pre-RNAseq Module Survey
- ☑ RNAseq Lectures
- ☑ RNAseq JNs
- ☑ Post-RNAseq Module Survey

- Pre- and post- bootcamp surveys used to assess participant knowledge before and after training and to collect feedback
- GL4U certification offered for completing the GL4U Introduction module and one omics-specific module
  - ➢ *GL4U RNAseq certification shown as an example*

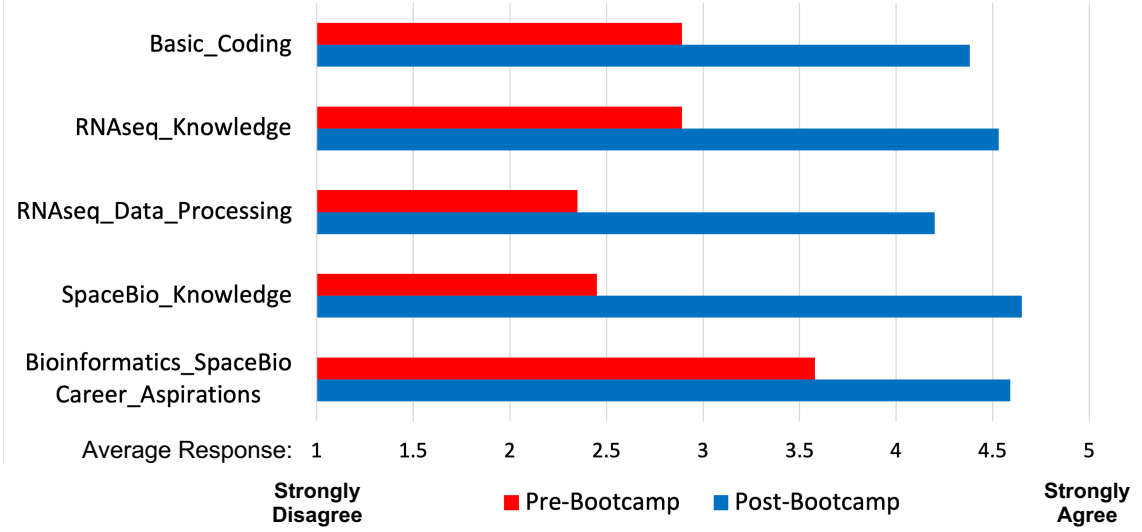
GL4U RNAseq Bootcamps

GL4U RNAseq Student / Educator Bootcamps

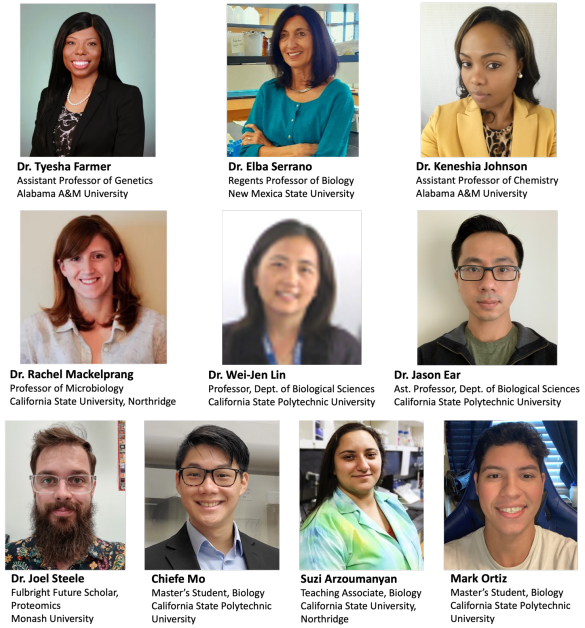
GL4U: RNAseq Bootcamp with SJSU, 06/2021

- Virtual 1-week long bootcamp with 17 **SJSU** students
- *Compute resources*: SJSU HPC
- Bootcamp covered GL4U Introduction and RNA Sequencing (RNAseq) modules

Student survey results

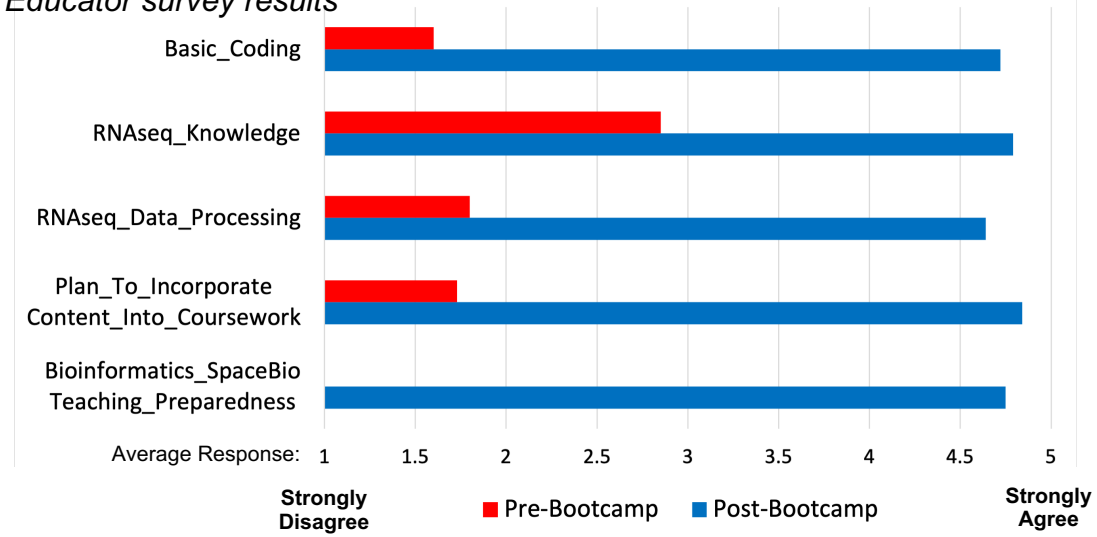


GL4U: RNAseq Bootcamp with MSIs/HBCUs, 06/2022



- Virtual 9-day long bootcamp with 6 professors and 4 graduate students from **4 HBCUs/MSIs**:
  - CalState Northridge
  - Alabama A&M
  - New Mexico State University
  - CalPoly Pomona
- *Compute resources*: **NASA NCCS SMCE**
- Bootcamp covered GL4U Introduction and RNAseq modules

Educator survey results



Student / Educator Feedback

GL4U: RNAseq Student Bootcamp with SJSU, 06/2021

"**The bootcamp was very informative and lectures were very well structured**. … Overall, the bootcamp experience was amazing, loved the fact that we had guest speakers as well. Thank you so much, especially to Amanda and Lauren for being so helpful and patient in explaining difficult concepts."

"This was such a great experience, **I learned lots of new information and I can't wait to explore more into Unix and R** and their functionalities!"

"The bootcamp was great. …the JNs helped reinforce the learning from the lectures. ...I will definitely save **the RNAseq lecture material** to study in the future because it **is bursting with great information**."

"I found the Bootcamp extremely well organized and its content very thorough."

"Considering the amount of content they had to teach us in a week, I thought the speakers did an amazing job. … **This is an experience I would recommend to anyone** who is willing to put the time and effort into learning about RNA sequencing."

GL4U: RNAseq Educator Bootcamp with MSIs/HBCUs, 06/2022

"**The camp was exceptionally well structured has gone over content a lot better than other courses I have participated in. The ability to run the code and understand it is much more valuable than plug and play GUI environments**…"

"Thank you! The content and delivery was easy to understand. The instructors are knowledgeable and were considerate of the differences in people's expertise… **I felt very comfortable with the material, I think it was the right amount of information for college-level students.** …"

"I thought this was really great. One issue I struggle with is developing material to teach bioinformatics to my students. **The material provided here is fantastic and will greatly improve how I teach bioinformatics and the scope of what I'll be able to teach**..."

"It was super informative and the instructors were extremely patient and passionate."

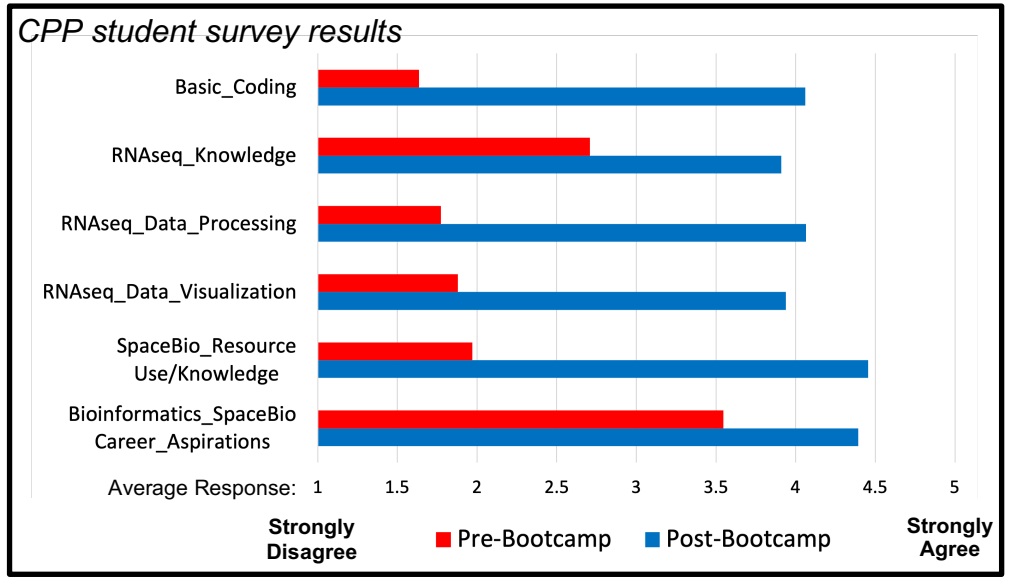
"**This was an amazing workshop and the absolute best bioinformatics bootcamp that I've ever participated in**. The material was easy to follow and the instructors did a fantastic job facilitating the discussions and providing thorough answers. … **I look forward to developing ways to implement the things learned for the benefit of students that I teach.**"

Educators Teach GL4U At Home Institution

CPP-GL4U RNAseq Independent Study Course, Fall 2023

Dr. Wei-Jen Lin
Professor, Dept. of Biological Sciences
California State Polytechnic University, Pomona

- Dr. Lin taught GL4U RNAseq content for 11 junior & senior CalPoly Pomona (CPP) undergraduate students
- *Compute resources*: **NASA NCCS SMCE**



CPP-GL4U RNAseq Independent Study Course Feedback

From the Educator

"The course has been extremely well supported by Amanda and the GeneLab/SMCE team. Amanda responded to my questions and requests promptly and passionately. Students really enjoyed learning the materials and the aspiration from NASA Biology and the guest talks. **Many students told me that this course is the best course they have had**. … **one student in this course just got accepted to a spring internship at JPL**... Overall, **the CPP-GL4U workshop has turned out well beyond my expectation**."

From the Students

"Thanks a lot. I can only imagine how complex it must have been to compile this knowledge into a bootcamp and try to make it as clear as possible for us students to learn from and **it was a clear success**. I understand it nicely so thank you again."

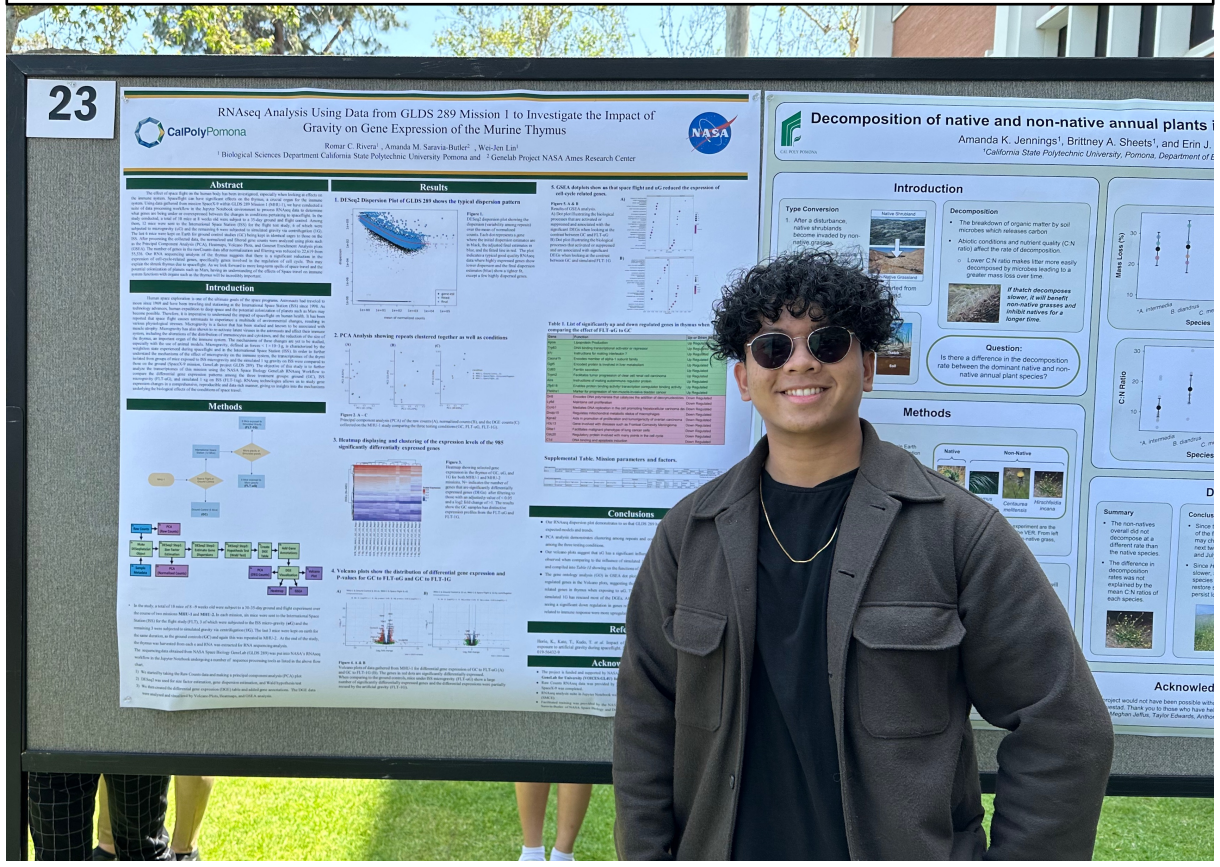
"While there was quite a bit of information, I believe the structure and pace of the course was reasonable. …I appreciate this opportunity and am quite happy I took it because **I feel significantly more confident about pursuing a field related to bioinformatics as well as more comfortable with the idea of continuing to learn some basic coding to complement my biology background**."

Students Generate New Knowledge Using GL4U On SMCE

To date, 3 students have used the knowledge gained from the GL4U: RNAseq Bootcamp and the modified RNAseq DGE JN on SMCE for their student projects

- Each student was able to generate new knowledge about the effects of spaceflight on terrestrial biology while gaining a valuable skillset

CPP Student, Romar Rivera Presents His Findings
2023 College of Science Symposium

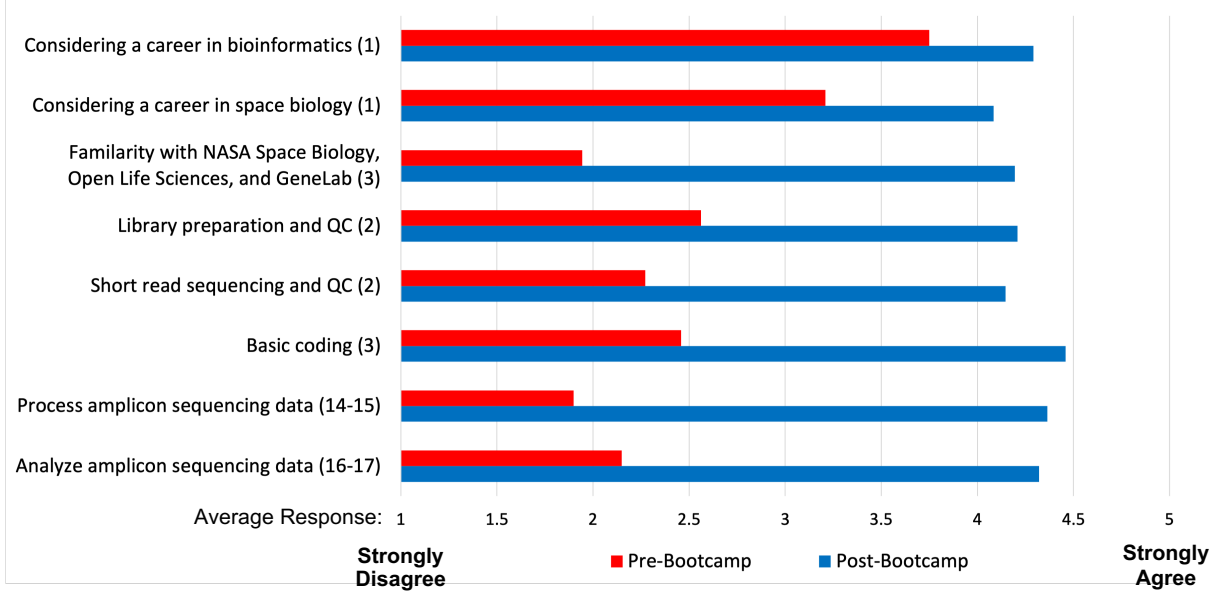


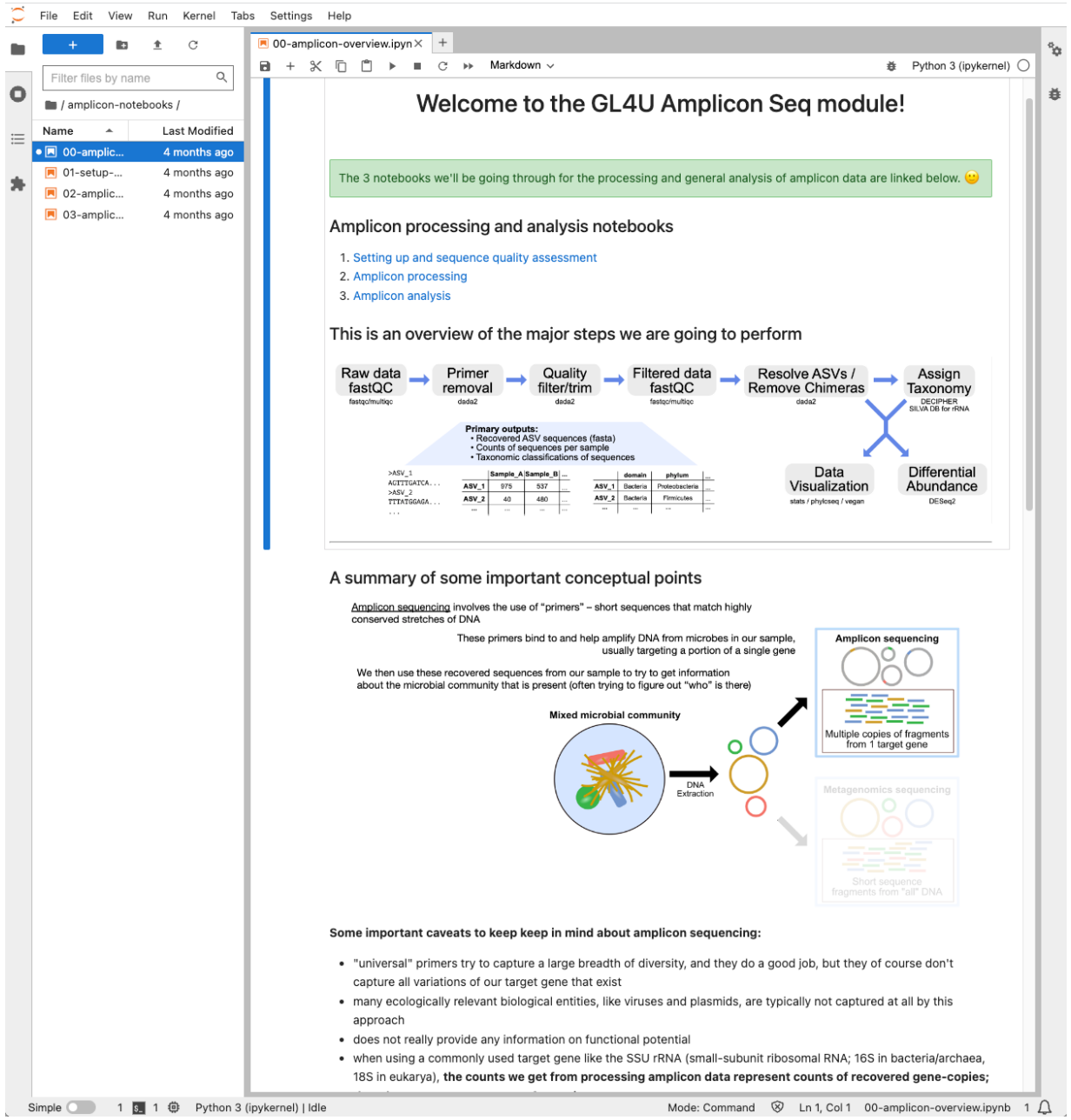
GL4U Amplicon Seq Bootcamp

GL4U AmpSeq Student Bootcamp

GL4U: AmpSeq Bootcamp at CSULA, 07/2023

- In-person 4-day bootcamp with 24 **CSULA** students
- *Compute resources*: NSF ACCESS
- Bootcamp covered GL4U Introduction and Amplicon Sequencing (AmpSeq) modules





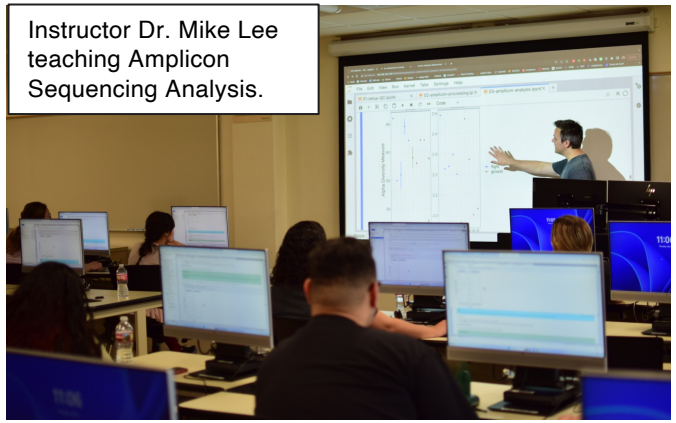
Student Feedback

GL4U: AmpSeq Bootcamp at CSULA, 07/2023

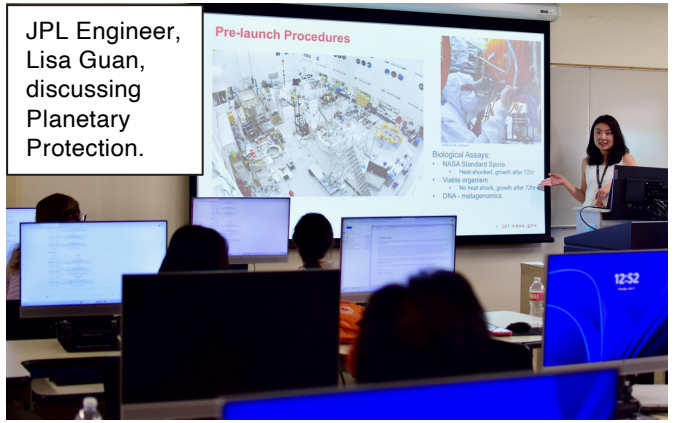
"**Very informative and well structured**. Going to use this as a reference when practicing amplicon sequencing to become more familiar with it."

"**This bootcamp** was very informative, opened up my eyes more to bioinformatics, and **is inspiring me to pursue a career in this field**."

"**Before the bootcamp, I had no idea that space biology was a thing** and I'm really excited to learn and explore more!"



Instructor Dr. Mike Lee teaching Amplicon Sequencing Analysis.



JPL Engineer, Lisa Guan, discussing Planetary Protection.

"…I personally learned a lot. **I really now know where to begin in order to get myself even deeper into this area**."

"I'm so glad I took this bootcamp, it was very supportive and very encouraging. …**this bootcamp was really able to demystified concepts and gave me the resources to soon become more confident in this field of science**."

"Mike and Amanda were great speakers and teachers this week! **They made the information very easy to understand and made it interesting as well**. …**Overall it is an awesome bootcamp**, and I can definitely see myself using this skill I learned throughout my Masters and future PhD!"

Accessing GL4U: AmpSeq Content

nasa / GeneLab-Training

Type / to search

Code | Issues | Pull requests | Actions | Projects | Wiki | Security | Insights | Settings

GL4U: Amplicon Sequencing Bootcamp 2023 Pilot at CSULA

Mike Lee edited this page on Jul 13 · 20 revisions

Welcome to the main page for the GeneLab for Colleges and Universities (GL4U) Amplicon Sequencing Bootcamp 2023 Pilot taking place at California State University, Los Angeles (CSULA)!

This course takes place Monday, 10-Jul-2023 to Thursday, 13-Jul-2023 in-person at CSULA 😊

The course utilizes lectures and hands-on coding in Jupyter Notebooks. It begins with the GL4U: Introduction Modules, which includes an overview of NASA Space Biology, NASA Open Life Sciences and the Open Science Data Repository (OSDR), GeneLab, and key foundational coding skills and concepts related to bioinformatics and working with sequencing data. Then it moves onto the GL4U: Amplicon Sequencing Modules, which provides an overview of amplicon sequencing and hands-on processing of samples derived from the Rodent Research - 6 mission, OSD-249, using the GeneLab Amplicon Sequencing Data Processing Pipeline followed by analysis and visualization of the processed data.

Lecture materials are linked in the schedule below, while Jupyter notebooks will be linked after the bootcamp.

Schedule Links

- 10 July - Monday
- 11 July - Tuesday
- 12 July - Wednesday
- 13 July - Thursday

Accessing the NSF ACCESS cloud computers

Links to accessing the National Science Foundation (NSF) Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) cloud computers that will be used during the bootcamp can be found on this page.

Requesting ACCESS resources and how to set up for teaching

Information on requesting ACCESS resources for teaching these bootcamp materials can be found on this page.

Individual computer setup

Information on how to set up an environment on a local computer capable of running the Jupyter notebooks can be found on this page.

nasa / GeneLab-Training

Type / to search

Code | Issues | Pull requests | Actions | Projects | Wiki | Security | Insights

GL4U: Amplicon Seq 2023 Pilot Individual Setup

Mike Lee edited this page on Jul 12 · 2 revisions

This page details one path to set up the required environment for running the GL4U Amplicon Seq 2023 Pilot Jupyter notebooks. The processing is light enough to be run on a typical laptop. It requires a Unix-like environment and utilizes conda for installing all required tools as detailed below.

We will be helping folks who want to do this during the bootcamp, but if you try to pursue this later and run into any issues, feel free to reach out to Mike (Mike.Lee@nasa.gov) and/or Amanda (Amanda.M.Saravia-Butler@nasa.gov) for help 🙂

Page contents

- Accessing a Unix-like environment
- Installing conda
- Creating the needed environment
- Downloading and launching the notebooks

Accessing a Unix-like environment

A Unix-like environment is required.

- On a Mac or Linux, this can be accessed by searching for and opening the "Terminal" app.
- On a Windows computer, installing the Windows Subsystem for Linux (WSL) is required. You can try opening "PowerShell" and running `wsl --install` and following along with the process. After the installation is complete, you would want to open "Ubuntu" on the Windows computer to access your Unix-like environment.

Installing conda

Conda is a package and environment manager, and it is the method used here to setup the required environment to run the Jupyter notebooks. One place you can learn more about conda is this page at Happy Belly Bioinformatics, or you can just skip to this section and follow the installation instructions. Be sure to start with the `curl` command there that is specific to if you are on a Mac, Windows, or Linux machine (these commands should be run in your Unix-like environment).

After finishing that installation (and there is a "(base)" at the start of your prompt at the command line), the first thing we are going to install with conda is mamba, to enable faster installations, by running the following:

```
conda install -y -n base -c conda-forge mamba
```

Creating the needed environment

The following command will create a conda environment called "GL4U-amplicon-2023", and may take a few minutes to complete:

```
mamba create -n GL4U-amplicon-2023 -y -c conda-forge -c bioconda -c defaults \
            jupyterlab=3.6.0 bash_kernel=0.9.0 r-irkernel=1.3.2 coreutils=9.1 \
            r-base=4.1.3 r-tidyverse=1.3.2 r-vegan=2.6_4 r-dendextend=1.16.0 \
            bioconductor-dada2=1.22.0 bioconductor-decipher=2.22.0 \
            bioconductor-phyloseq=1.38.0 bioconductor-deseq2=1.34.0 \
            fastqc=0.11.9 multiqc=1.12 jupyter_contrib_nbextensions=0.7.0
```

Downloading and launching the notebooks

Still in our Unix-like environment, running this next codeblock will download and unpack the Jupyter notebooks into locations in our home directory:

```
# downloading notebooks
curl -L -o ~/GL4U-2023-amplicon-bootcamp-notebooks.zip https://figshare.com/ndownl

# unpacking
unzip ~/GL4U-2023-amplicon-bootcamp-notebooks.zip -d ~/
# that includes:
    # 00-overview.ipynb
    # intro-notebooks/
    # amplicon-notebooks/

# removing zip
rm ~/GL4U-2023-amplicon-bootcamp-notebooks.zip
```

Then we can activate the conda environment we created above, and launch the Jupyter notebooks like so:

```
# changing into home directory
cd ~/

# activating conda enviroment
conda activate GL4U-amplicon-2023

# launching jupyter lab
jupyter lab 00-overview.ipynb
```

That conda environment will always need to be active (so our prompt should start with "(GL4U-amplicon-2023)") if we want to run these Jupyter notebooks.

Accessing GL4U: AmpSeq Content

nasa / GeneLab-Training

Type / to search

Code | Issues | Pull requests | Actions | Projects | Wiki | Security | Insights | Settings

GL4U: Amplicon Sequencing Bootcamp 2023 Pilot at CSULA

Mike Lee edited this page on Jul 13 · 20 revisions

Welcome to the main page for the GeneLab for Colleges and Universities (GL4U) Amplicon Sequencing Bootcamp 2023 Pilot taking place at California State University, Los Angeles (CSULA)!

This course takes place Monday, 10-Jul-2023 to Thursday, 13-Jul-2023 in-person at CSULA 😊

The course utilizes lectures and hands-on coding in Jupyter Notebooks. It begins with the GL4U: Introduction Modules, which includes an overview of NASA Space Biology, NASA Open Life Sciences and the Open Science Data Repository (OSDR), GeneLab, and key foundational coding skills and concepts related to bioinformatics and working with sequencing data. Then it moves onto the GL4U: Amplicon Sequencing Modules, which provides an overview of amplicon sequencing and hands-on processing of samples derived from the Rodent Research - 6 mission, OSD-249, using the GeneLab Amplicon Sequencing Data Processing Pipeline followed by analysis and visualization of the processed data.

Lecture materials are linked in the schedule below, while Jupyter notebooks will be linked after the bootcamp.

Schedule Links

- 10 July - Monday
- 11 July - Tuesday
- 12 July - Wednesday
- 13 July - Thursday

Accessing the NSF ACCESS cloud computers

Links to accessing the National Science Foundation (NSF) Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) cloud computers that will be used during the bootcamp can be found on this page.

Requesting ACCESS resources and how to set up for teaching

Information on requesting ACCESS resources for teaching these bootcamp materials can be found on this page.

Individual computer setup

Information on how to set up an environment on a local computer capable of running the Jupyter notebooks can be found on this page.

nasa / GeneLab-Training

Type / to search

Code | Issues | Pull requests | Actions | Projects | Wiki | Security | Insights | Settings

GL4U: Amplicon Seq 2023 Pilot ACCESS Request and Content Setup

Mike Lee edited this page on Jul 13 · 8 revisions

Page contents

- Summary
- Requesting resources
- Transferring credits
- Requesting an increase in allotted quotas
- Setting up instances
- Creating links for participants
- Deleting instances

Summary

ACCESS is an NSF-funded resource that can provide cloud computing for research and educational purposes. If you are an educator that wants to run the GL4U: Amplicon Seq 2023 bootcamp for your class, this page will help walk you through the process of requesting resources and then how to utilize them. If you have any questions or hit any problems, please don't hesitate to reach out to Mike.Lee@nasa.gov for help 🙂

The process is detailed below, but all that is required for submitting a request is filling out a form that includes:

- A couple of paragraphs providing an overview of the purpose (here for educational purposes) and how ACCESS would be used
- A CV of the submitting PI

Primary resources utilized

- ACCESS - an NSF-funded program that we apply to in order to get cloud-computing resources
- Jetstream2 - one of the primary computing infrastructures ACCESS works with (and the one most suitable for requests like this)
- Exosphere - a web-interface for managing Jetstream2 computing resources

Accessing GL4U: AmpSeq Content

Requesting resources

1. Login to ACCESS

Click to Login to ACCESS at the top right of this page https://access-ci.org/. There is no account creation step here, but rather you can choose from a few methods to authenticate your identity, such as ORCID or Google. You also may need to set up dual-factor authentication.

2. Submit a request

Once logged in, the next step is to submit a request. There are different types offered, listed here. At the time of putting this page together, "Explore ACCESS" is the most appropriate for educational purposes.

To begin the process, while logged in, go to the opportunities page, and click to "SUBMIT AN EXPLORE ACCESS REQUEST".

There you will need to enter a few things. An example of each is presented below, but these should be adjusted for your scenario. Anything not listed here can be skipped on the form:

Title

> Running an Amplicon Sequencing Bioinformatics Course

Public Overview

> General overview
> I am a professor and would like to utilize the NASA GeneLab (https://genelab.nasa.gov/) GeneLab for Colleges and Universities (GL4U; https://github.com/nasa/GeneLab-Training/tree/main/GL4U) amplicon sequencing training materials with my students. This will likely take place over just a week or two, and an EXPLORE ACCESS allocation would provide sufficient resources to be able to provide the same computing environment to all participants.
>
> How I plan to use ACCESS
> I intend to use Indiana Jetstream2, managed through the Exosphere website, and to create individual m3.medium instances for each participant based off the publicly available "GL4U-amplicon-2023" image.
>
> Thank you for your consideration and any help!

Transferring credits

Once the request has been approved, you will need to transfer the credits from ACCESS to JetStream2. Once logged in at https://access-ci.org/, go to https://allocations.access-ci.org/requests in order to transfer credits.

For the appropriate allocation, select "Choose New Action", then "Exchange". On the next screen, choose Indiana Jetstream2 as Resource, click "Add Resource", enter all credits, add anything to the comment box (as it is required), then click Submit.

Requesting an increase in allotted quotas

The starting allotted quotas will typically only allow up to maybe 10 concurrent instances to be created. If you are going to have more than that actively participating, you need to submit a request to increase the allotted quotas.

To do this, log into JetStream2 using the same identity authentication as used above to log in to ACCESS, then click "Add allocation", then "Add ACCESS Account", then once verified, select the allocation to be added to JetStream2.

After it is present on JetStream2 when you are logged in, then go to this support page to build an email as described next: https://jetstream2.exosphere.app/exosphere/getsupport

Select the button for "An Allocation", then modify the following text to specific your specific allocation (e.g., "BIO######", which you can get from this ACCESS page) and how many concurrent instances you will need (an instance is a computer, so 1 for each planned participant):

> Hi there,
>
> We plan to use this allocation (<YOUR ALLOCATION ID>) with <YOUR TOTAL STUDENTS/PARTICIPANTS> concurrent m3.medium instances for a bioinformatics course we are running.
>
> Could you please help with increasing the allotted quotas so that we will be able to run up to <YOUR TOTAL STUDENTS/PARTICIPANTS> m3.medium instances concurrently on this allocation, including cores, ram, volume, ports, available IP addresses, and whatever else would be required?
>
> Thank you for any help!

Then click to "Build Support Request", copy the contents of the text window, and paste it in an email to "help@jetstream-cloud.org" with the subject header "[Jetstream2] Support Request From Exosphere for Jetstream2".

Accessing GL4U: AmpSeq Content

Setting up instances

Once the above is all taken care of, you can begin setting up instances.

There is extensive documentation on Jetstream2 here: https://docs.jetstream-cloud.org/ui/exo/exo/

It is a lot, and the Jetstream2 folks are super-responsive to requests for help, but, as mentioned above, feel free to reach out to Mike.Lee@nasa.gov too.

Log into JetStream2, select the appropriate allocation, then:

- choose "Create" at the top-right, then "Instance"
- select "By Image", select the text window to search by name, and search for GL4U-amplicon-2023, and choose "Create Instance" on the "GL4U-amplicon-2023" image
  - give the instance a name, like "Amplicon-Course"
  - select "m3.medium"
  - move the slide to create as many instances as needed, if you need more than the max that can be created at one time, do this process in as many steps that are needed
  - click "Advanced Options", and click to "Assign a public IP address to this instance"
  - at the bottom is a "Boot Script", select the entire text and delete it, then replace it with the following:

```
#cloud-config
users:
  - default
  - name: exouser
    shell: /bin/bash
    groups: sudo, admin
    sudo: ['ALL=(ALL) NOPASSWD:ALL']{ssh-authorized-keys}
  - name: gl4u
    shell: /bin/bash
    groups: users
    lock_passwd: false
    passwd: $1$V9O.SGtD$LqHZP91jT/Sjhax8kWSQF1
ssh_pwauth: true
package_update: true
package_upgrade: {install-os-updates}
packages:
  - git{write-files}
```

Creating links for participants to access their instances

Each instance has its own IP address, and that IP address can be used to provide a link to the participants to access their own cloud-computing environment through a web-browser. The below examples are with the mock IP address "XXX.XXX.XXX.XX", so you would need to alter that for each individual IP, but this is what the links would look like:

A link structured like this would take you to the base Jupyter lab environment: `http://XXX.XXX.XXX.XX:8000`
A link structured like this would take you to the opening overview notebook page:
`http://XXX.XXX.XXX.XX:8000/lab/tree/00-overview.ipynb`

Each user has the same user name (gl4u) and password (gl4u2023). **These instances are ephemeral and are not meant to hold anything that needs to be secure.**

Deleting instances

When finished, you can select multiple instances on the Instance page of Jetstream2 for the appropriate allocation, and choose to delete the instances.

Detailed Jetstream2 documentation can be found here, and feel free to reach out to Mike.Lee@nasa.gov 🙂

Questions?

https://github.com/nasa/GeneLab-Training/tree/main/GL4U

Sign up for the GL4U mailing list

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

Acknowledgements

- **NASA GeneLab**
  - PM: Sylvain Costes, PhD
  - DPM: Samrawit Gebre
  - DP: Lauren Sanders, PhD
  - DP: Amanda Saravia-Butler, PhD
  - DP: Mike Lee, PhD
  - GeneLab Team
- **JPL Planetary Protection**
  - PM: Alvin L. Smith II, PhD
  - Engineer: Lisa Guan
  - Scientist: Arman Seuylemezian
  - JPL PP Team
- **Universities Space Research Association**
  - Saba Hussain
  - Tristyn Acasio
  - Rachel Gilbert
- **NASA**
  - BPS Dir: Lisa Carnell
  - SB ADC: Parag Vaishampayan, PhD
- **San Jose State University**
  - Philip Heller, PhD
  - Steven Boring
- **California State University, Los Angeles**
  - Gustavo Ramirez, PhD
- **SMCE System Administrators**

Funding

- Compute resources were made available through the NASA Science Managed Cloud Environment (SMCE) funded by SMD AWS Space Act Agreement
- GeneLab is funded by the NASA Space Biology program within the NASA Science Mission Directorate's (SMD) Biological and Physical Sciences (BPS) Division
- JPL Engineering and Science Directorate (ESD) HBCU/MSI Internal Funding Award

Extra Slides

Teaching GL4U Content At Home: Educator Feedback

Dr. Wei-Jen Lin
Professor, Dept. of Biological Sciences
California State Polytechnic University, Pomona

CPP-GL4U RNAseq Independent Study Course, Fall 2023

- Dr. Lin taught GL4U RNAseq content as a semester-long independent study course for 11 junior & senior CPP undergraduate students
- *Compute resources*: NASA NCCS SMCE
- Dr. Lin completed the survey questions below to provide feedback about her experience teaching the GL4U RNAseq bootcamp content (Intro and RNAseq modules) at her home institution