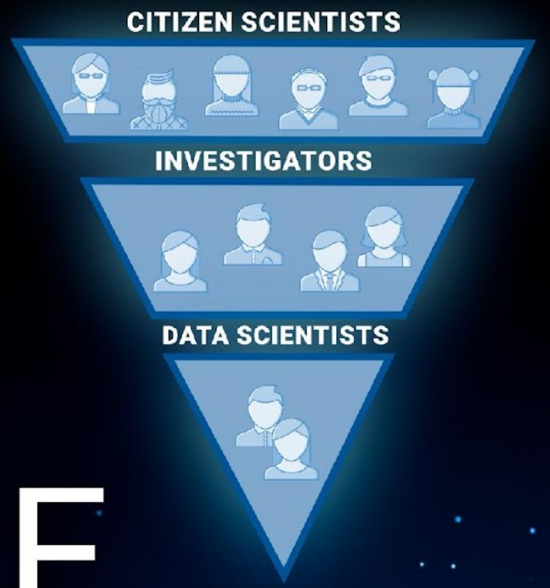
TISSUE REPOSITORY

SCIENTIFIC COMMUNITY

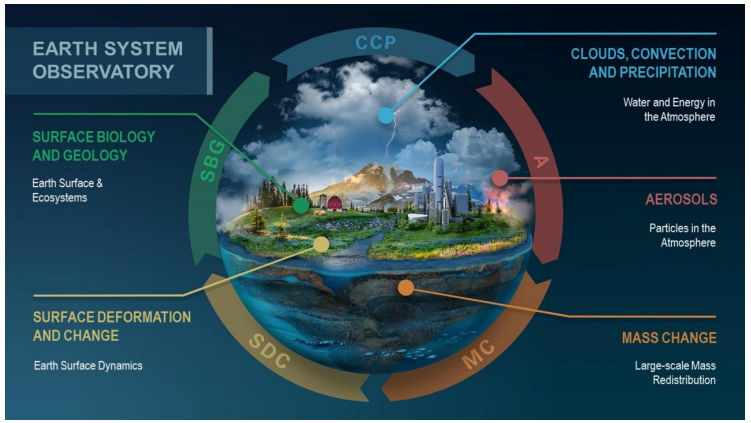


OPEN SCIENCE

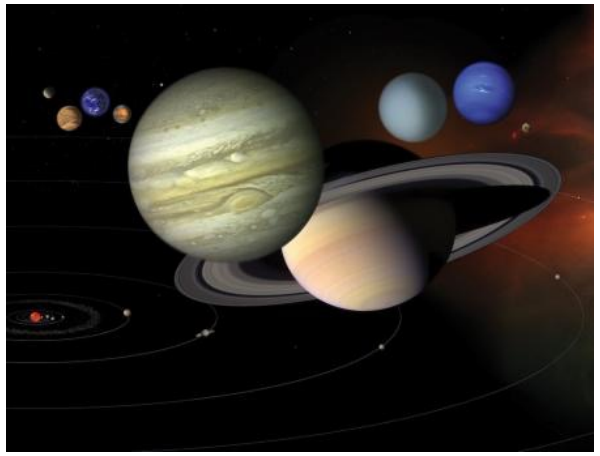
DATA REPOSITORY

GENELAB OMICS + **ALSDA** PHENOTYPIC

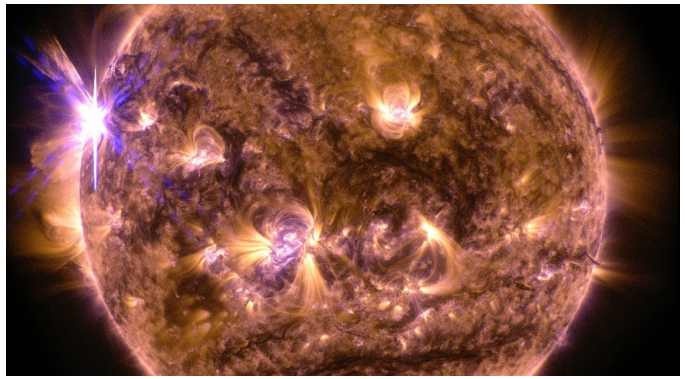
NASA Science Mission Directorate (SMD)



Earth Science
The study of planet Earth

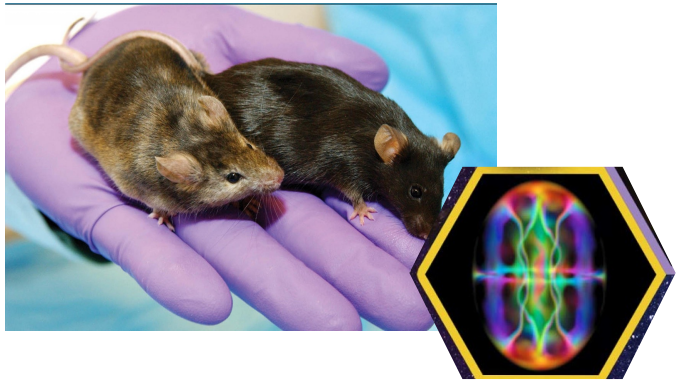


Planetary Science
The study of the origin and history of the solar system and the potential for extraterrestrial life



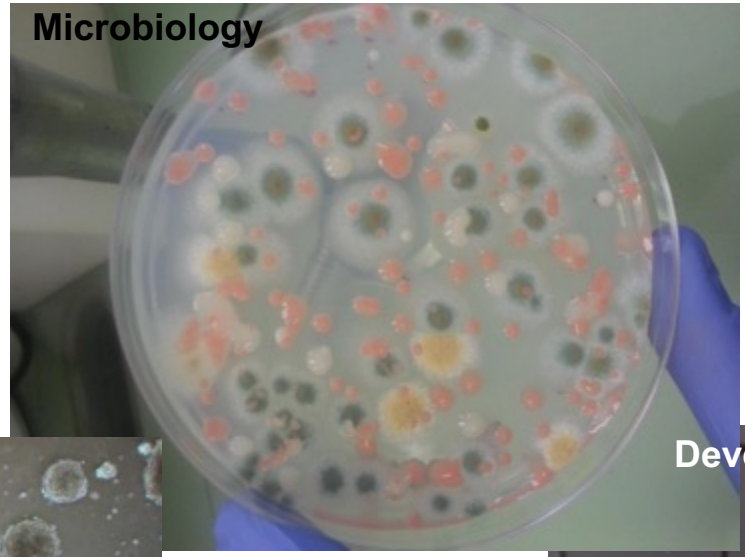
Heliophysics
The study of the sun and its effects on the solar system

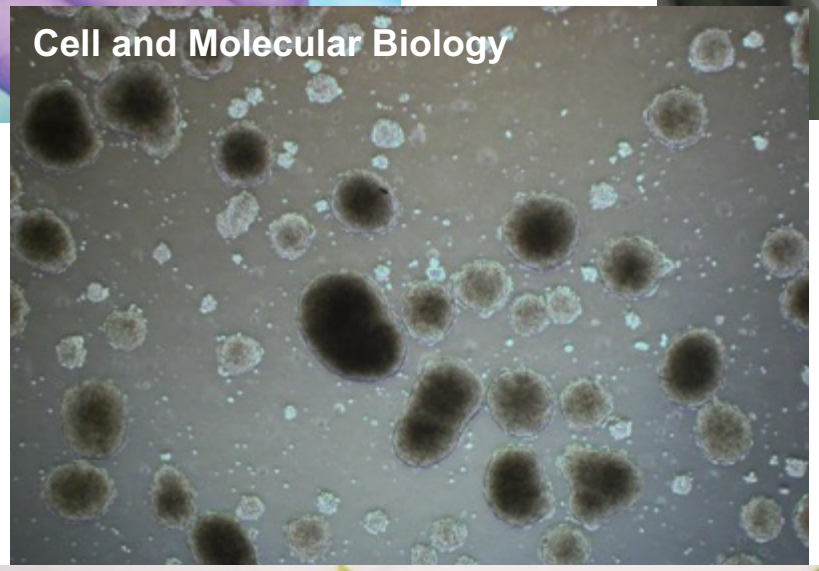
Astrophysics
The study of the origin, structure, evolution and destiny of the universe

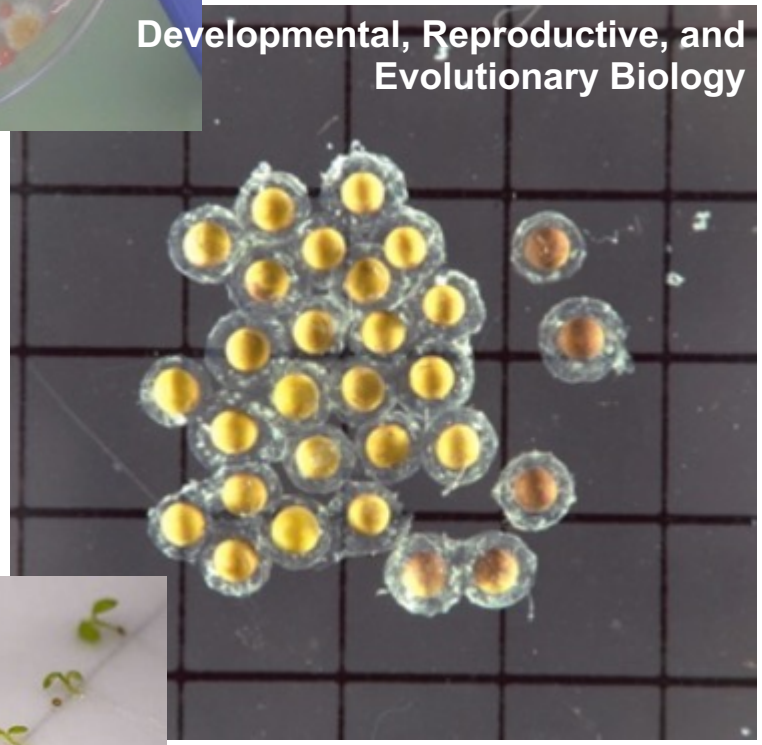


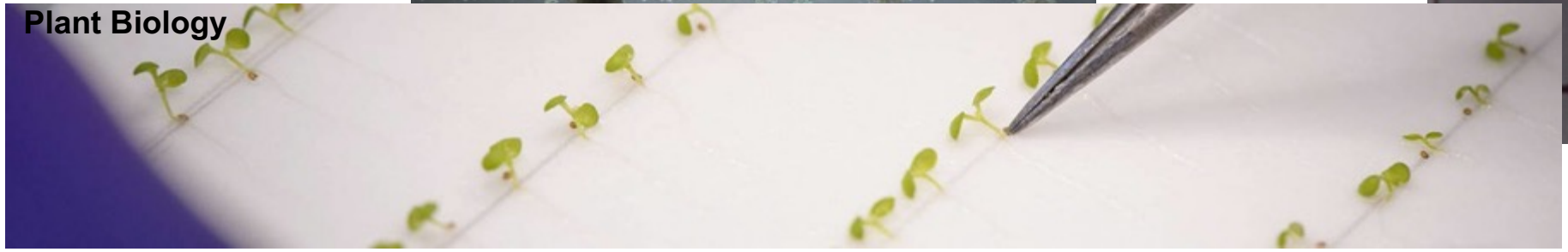
Biological and Physical Sciences
The study of how spaceflight environment effects biological and physical systems

BPS: Space Biology Program





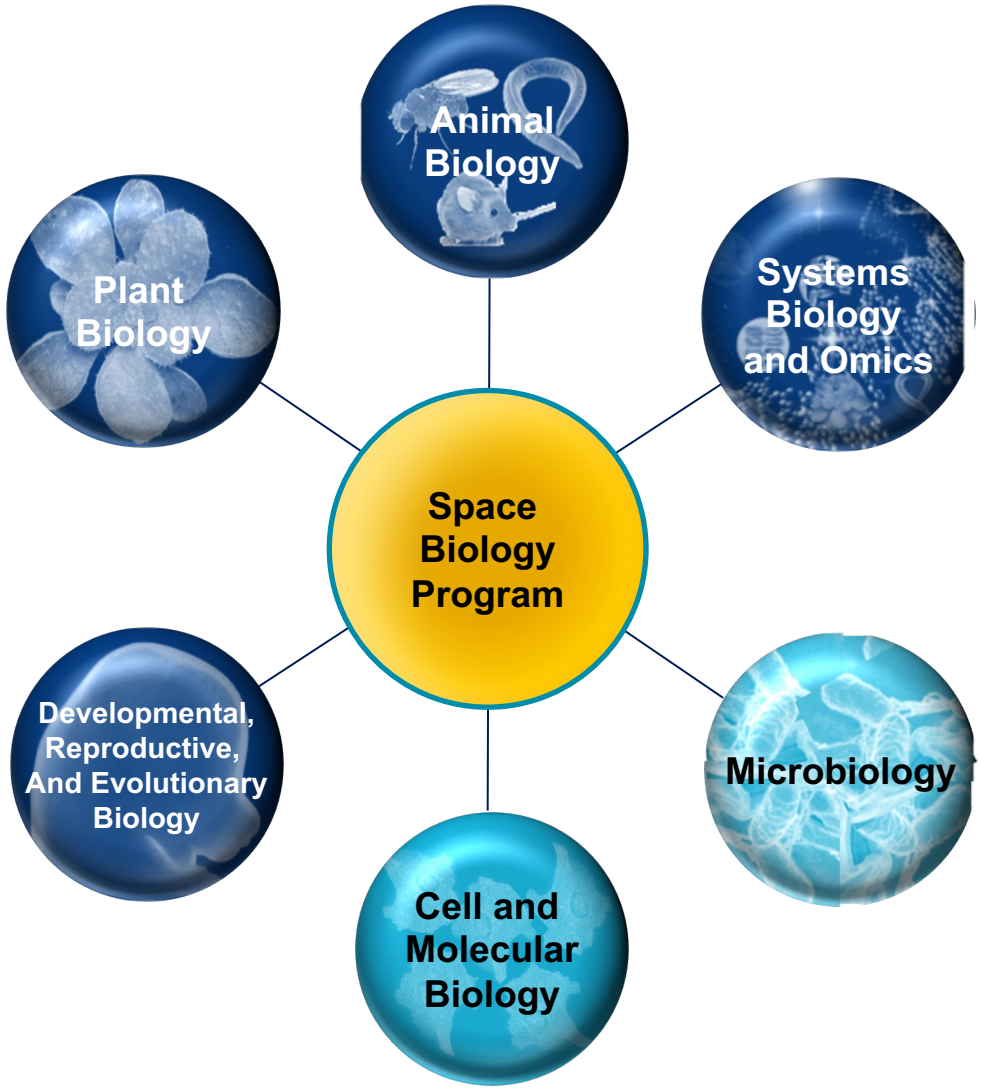




https://science.nasa.gov/biological-physical/programs/space-biology

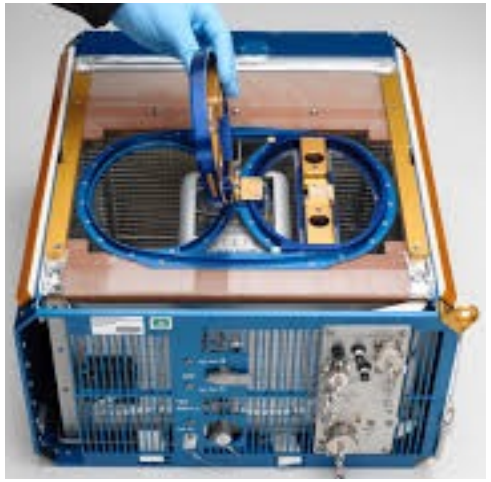
Space Biology Objectives

- Discover how biological systems respond to the space environment
- Identify the underlying mechanisms and develop models for biological systems in space
- Develop technologies to enable spaceflight research
- Promote open science through the GeneLab Data System and Ames Life Sciences Data Archive
- Provide mechanistic understanding to support human health in space
- Transfer the knowledge and technology of space-based research to benefit life on Earth



Flight Hardware To Support Space Biology Research

NASA: Bioculture System

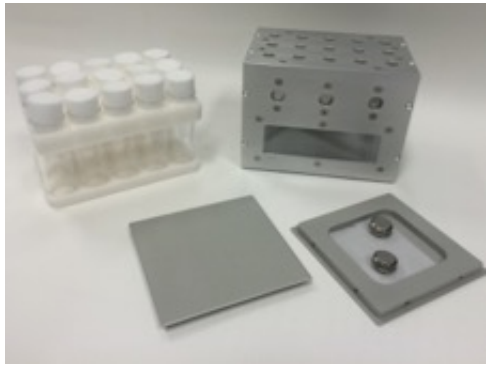


NASA: Rodent Habitat

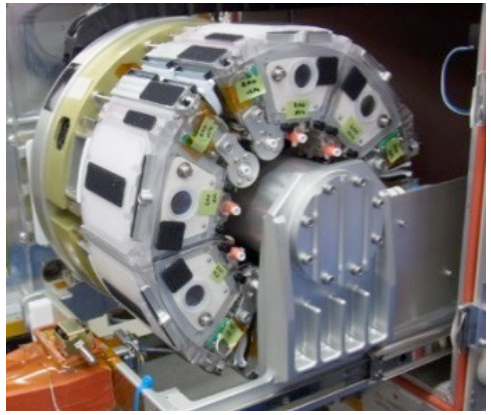
NASA-JAXA: Life Sciences Glovebox

BioServe: BioCell and Fluid Processing Apparatus

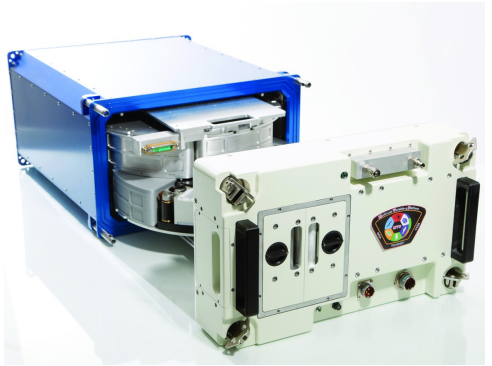
NASA: WetLab-2 (qPCR)



NASA: Vented Fly Box



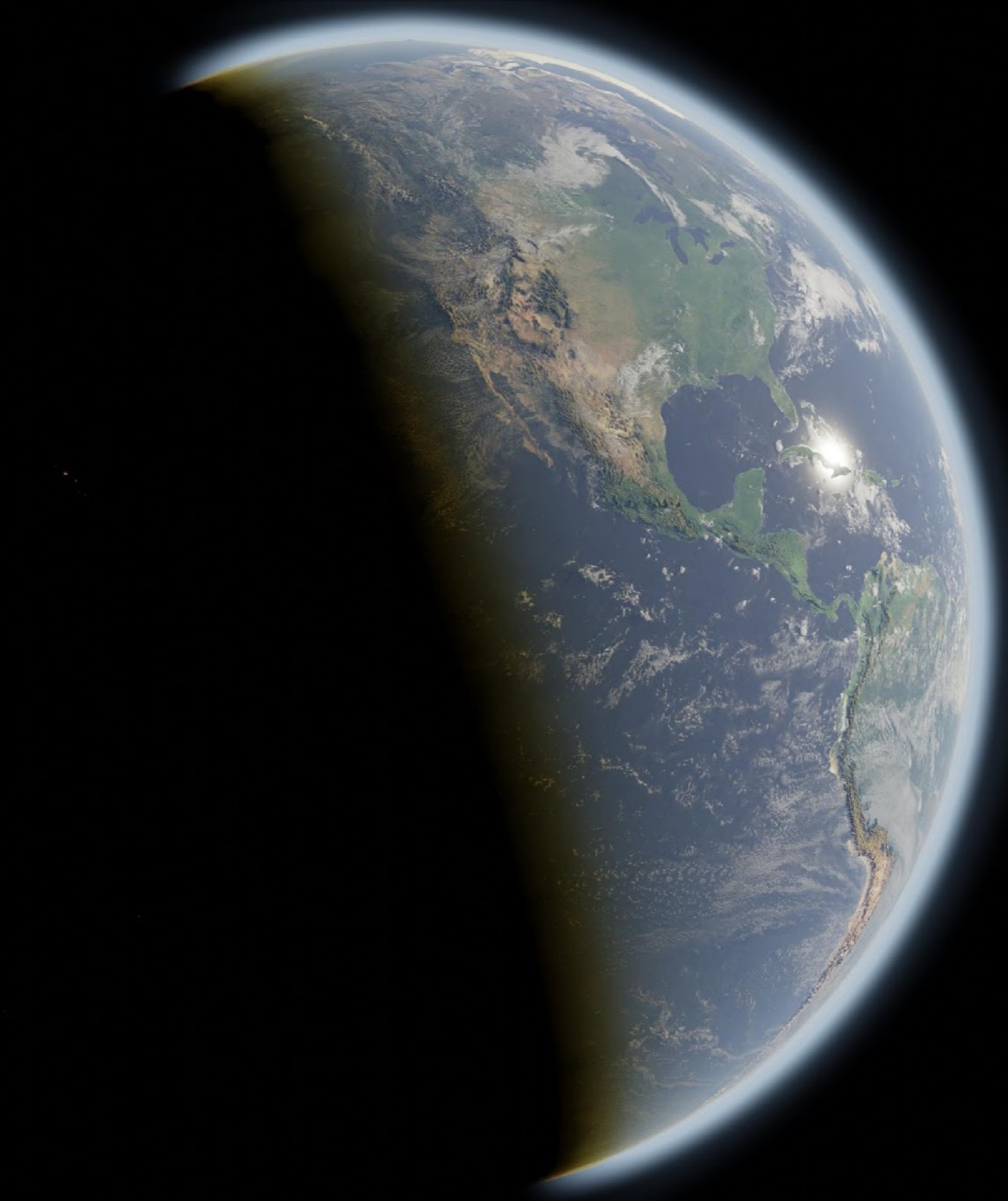
JAXA: Mouse Habitat Unit



Redwire: Multi-Use Variable-G Platform

Open Science

"We define open science as a collaborative culture enabled by technology that empowers the **open sharing of data, information, and knowledge** within the scientific community and the wider public to accelerate scientific research and understanding."



Ramachandran, R., Bugbee, K., & Murphy, K. J. **Moving from Open Data to Open Science**. Earth and Space Science, Wiley Publication. https://doi.org/10.1029/2020EA001562

OSDR Mission and Vision

Mission

- To enable biological discovery and health resilience for space exploration through open science.

Vision

- Design and deploy a unique repository, the Open Science Data Repository (OSDR), housing standardized metadata and data from spaceflight or spaceflight-relevant samples, following FAIR (Findable, Accessible, Interoperable, and Reusable) principles
- Engage with the scientific community by prioritizing the needs of the Open Science Analysis Working Group (AWG) members, and supporting open source, open science, citizen science, and education initiatives such as GeneLab for High Schools (GL4HS) and GeneLab for Colleges and Universities (GL4U)
- Partner with spaceflight-relevant projects through sample sharing or augmentation of experimental samples to expand analyses
- Process spaceflight-relevant samples answering knowledge gaps, establishing best-practices for sample processing and providing common omics processing platforms for the space biology community, leading to a more cohesive sets of independent datasets
- Curate, beyond best-practice, spaceflight-relevant datasets and make processed data publicly available as expediently as possible
- Provide open-access to computing resources and visualization tools for raw and processed data, democratizing the access to spaceflight-relevant data and disseminating knowledge of how life responds to the space environment. This platform will be an essential tool to discover new countermeasures for human exploration of space, and will inevitably benefit life on Earth as well.

NASA Biological Open Science Resources

Biospecimen Sharing Program (BSP)

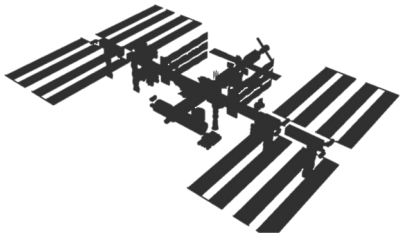
Dissection and preservation of rodent tissues from Flight and Ground investigations. Coordination of internal tissue sharing

NASA Biological Institutional Scientific Collection (NBISC)

Collection of non-human specimens and space microbial culture

Ames Life Sciences Data Archive (ALSDA)

Collection and curation of mission, project, and imaging data



NASA GeneLab (GL)

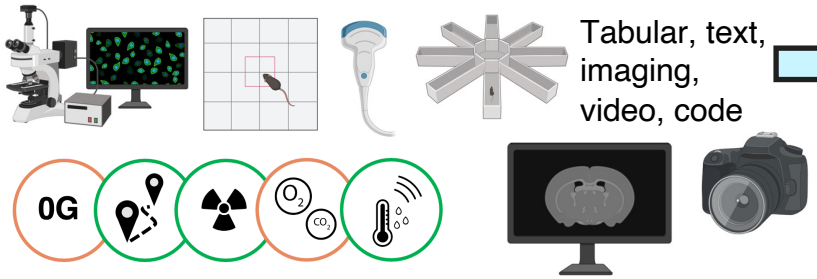
Collection and curation of omics data

NASA Internal Program

Open-Source Science Programs — Available Globally

Integrating Biological Data Repositories

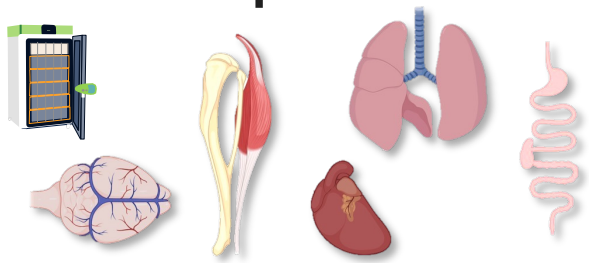
Physiological/Phenotypic/Imaging/ Environmental Telemetry Data



Tabular, text, imaging, video, code

Molecular/Omics Data

Biospecimens



ALSDA

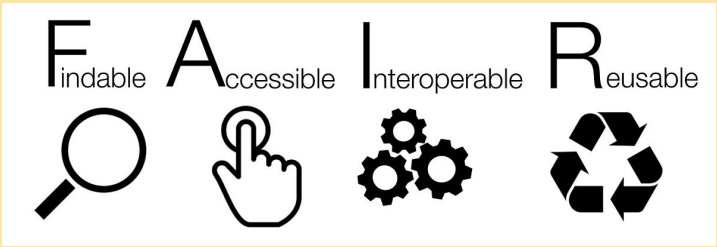
GeneLab

NBISC

NASA Open Science Data Repository (OSDR)

osdr.nasa.gov/bio

- Single Submission Portal (BDME)
- User Interface/Website Tool for RDSAs (Research Data Submission Agreements)
- Maximally Open Access with Necessary Controls for Sensitive Data
- Data Maximally FAIR



Slide Credit: Ryan Scott, ALSDA

Open Science Data Repository (osdr.nasa.gov)

- User-submitted data
- GeneLab-generated data
- Ingested data

Open Science Projects

Open Science Projects primary goals aim to increase collaborative scientific data sharing, analysis and more rapid scientific advancement.

GeneLab

GeneLab, an open science multi-omics repository, covering transcriptomics, metagenomics, epigenomics, proteomics, and metabolomics. Studies comprise of data from model organisms including microbes, plants, fruit flies, rodents and humans.

Learn more GeneLab

BSP

The NASA Space Biology Biospecimen Sharing Program (BSP) collects biospecimens to maximize the scientific return from biological spaceflight and associated ground investigations and to encourage and broaden participation from the scientific community in space biology-related research.

Learn more about BSP

ALSDA

Ames Life Sciences Data Archive (ALSDA) collects, curates, and makes available space-relevant higher-order phenotypic datasets. Datasets that enable scientists to perform retrospective analysis across missions, experiments, life science disciplines, research subjects, and species.

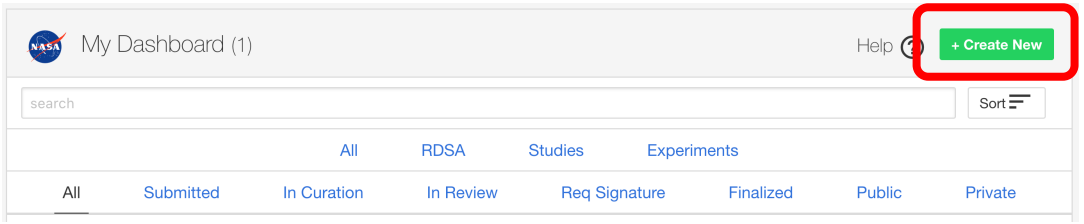
Learn more about ALSDA

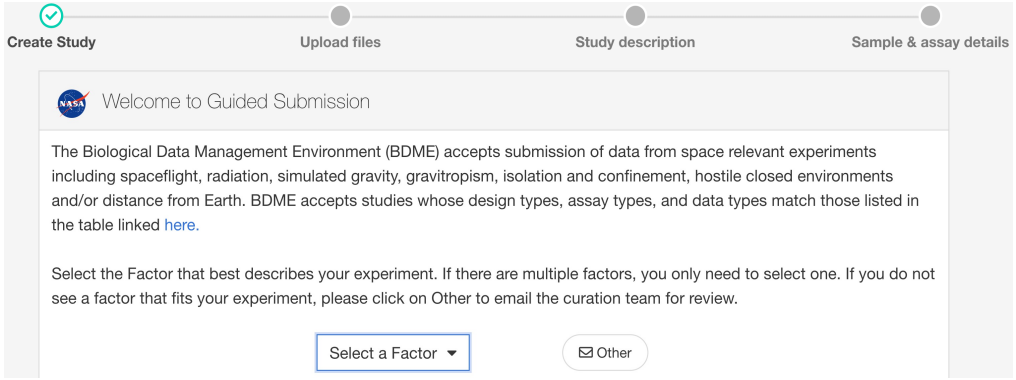
NBISC

NASA Biological Institutional Scientific Collection (NBISC) is a biorepository of non-human samples collected from NASA-funded spaceflight investigations and correlative ground studies. The purpose of NBISC is to receive, store, document, preserve, and make the collection available to the scientific community.

Learn more about NBISC

Submission Portal – BDME (https://osdr.nasa.gov/bio/submission-sso-login.html)





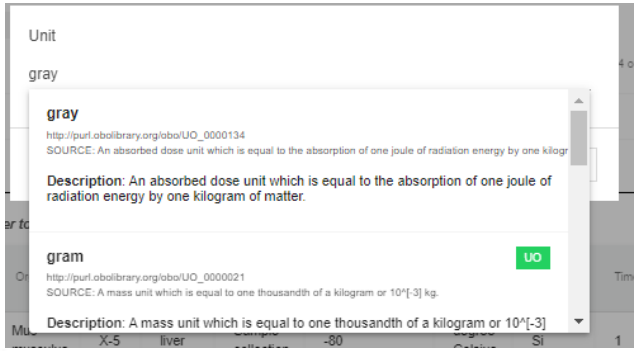
Sample level data

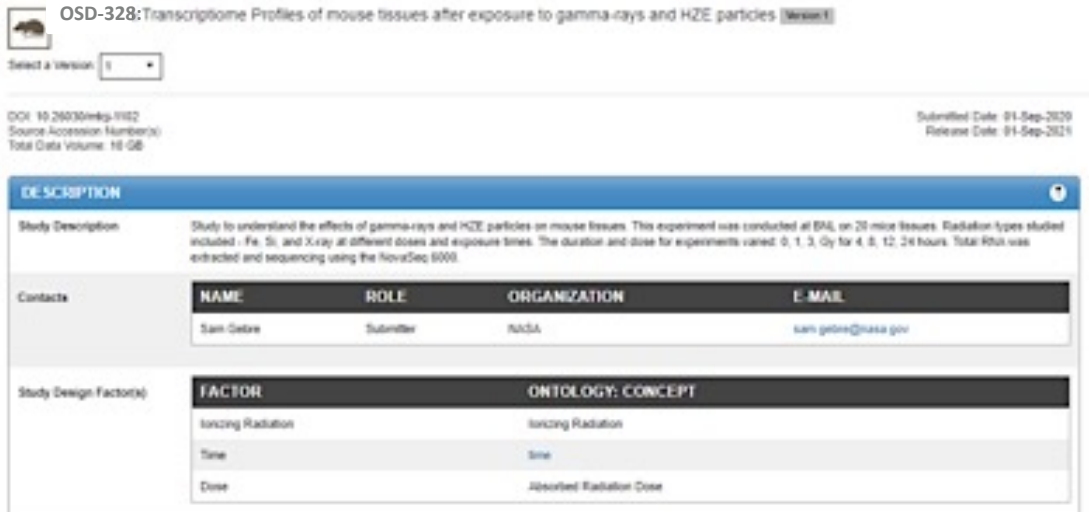
Sample Name	Organism	Strain	Material Type	Protocol REF	Parameter Value - Storage Temperature Setting	Unit	Ionizing Radiation	Time	Unit	Dose	Unit
Sample1	Mus musculus	X-5	liver	Sample collection	-80	degree Celsius	Silicon	1	hour	1	gray
Sample2	Mus musculus	X-5	liver	Sample collection	-80	degree Celsius	Sham-irradiated	0	hour	0	gray

Assay level data with assay-specific fields

Sample Name	Protocol REF	Parameter Value - QA Instrument	Parameter Value - QA Assay	Parameter Value - QA Score	Unit	Extract Name	Protocol REF
Sample1	Nucleic Acid Extraction	BioAnalyzer	NanoChip	8	RIN	Extract 1	Library Construction
Sample2	Nucleic Acid Extraction	BioAnalyzer	NanoChip	9	RIN	Extract 2	Library Construction
Sample3	Nucleic Acid Extraction	BioAnalyzer	NanoChip	9.1	RIN	Extract 3	Library Construction

Terms are connected to a controlled ontology





- Space biology data are submitted through BDME.
  - ✓ Phenotypic, Behavioral, Physiological Data
  - ✓ Omics Data
- Allows efficient input of sample and assay level metadata alongside data files.
- Tailored to specialized needs of spaceflight samples.

PIs are presented with a study preview and a private link for sharing with reviewers and collaborators.

Open Science Data Repository (osdr.nasa.gov)

Open Science Projects

Open Science Projects primary goals aim to increase collaborative scientific data sharing, analysis and more rapid scientific advancement.

GeneLab

GeneLab, an open science multi-omics repository, covering transcriptomics, metagenomics, epigenomics, proteomics, and metabolomics. Studies comprise of data from model organisms including microbes, plants, fruit flies, rodents and humans.

Learn more GeneLab

BSP

The NASA Space Biology Biospecimen Sharing Program (BSP) collects biospecimens to maximize the scientific return from biological spaceflight and associated ground investigations and to encourage and broaden participation from the scientific community in space biology-related research.

Learn more about BSP

ALSDA

Ames Life Sciences Data Archive (ALSDA) collects, curates, and makes available space-relevant higher-order phenotypic datasets. Datasets that enable scientists to perform retrospective analysis across missions, experiments, life science disciplines, research subjects, and species.

Learn more about ALSDA

NBISC

NASA Biological Institutional Scientific Collection (NBISC) is a biorepository of non-human samples collected from NASA-funded spaceflight investigations and correlative ground studies. The purpose of NBISC is to receive, store, document, preserve, and make the collection available to the scientific community.

Learn more about NBISC

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*

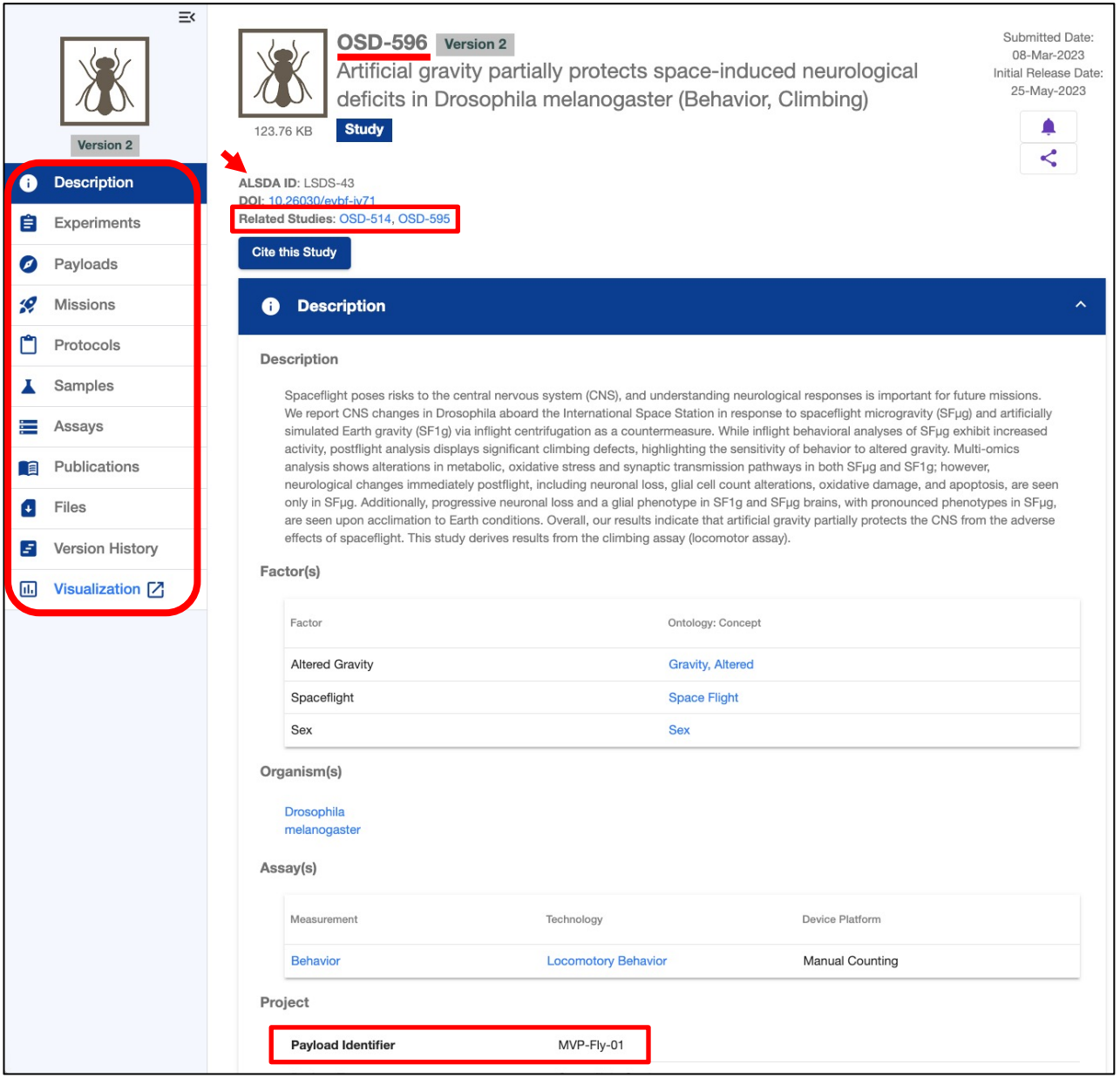
Study

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 ...

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



Version 2

- Description
- Experiments
- Payloads
- Missions
- Protocols
- Samples
- Assays
- Publications
- Files
- Version History
- Visualization

OSD-596 Version 2

Artificial gravity partially protects space-induced neurological deficits in Drosophila melanogaster (Behavior, Climbing)

123.76 KB Study

Submitted Date: 08-Mar-2023
Initial Release Date: 25-May-2023

ALSDA ID: LSDS-43
DOI: 10.26030/eybf-jv71
Related Studies: OSD-514, OSD-595

Cite this Study

Description

Description

Spaceflight poses risks to the central nervous system (CNS), and understanding neurological responses is important for future missions. We report CNS changes in Drosophila aboard the International Space Station in response to spaceflight microgravity (SFμg) and artificially simulated Earth gravity (SF1g) via inflight centrifugation as a countermeasure. While inflight behavioral analyses of SFμg exhibit increased activity, postflight analysis displays significant climbing defects, highlighting the sensitivity of behavior to altered gravity. Multi-omics analysis shows alterations in metabolic, oxidative stress and synaptic transmission pathways in both SFμg and SF1g; however, neurological changes immediately postflight, including neuronal loss, glial cell count alterations, oxidative damage, and apoptosis, are seen only in SFμg. Additionally, progressive neuronal loss and a glial phenotype in SF1g and SFμg brains, with pronounced phenotypes in SFμg, are seen upon acclimation to Earth conditions. Overall, our results indicate that artificial gravity partially protects the CNS from the adverse effects of spaceflight. This study derives results from the climbing assay (locomotor assay).

Factor(s)

Factor	Ontology: Concept
Altered Gravity	Gravity, Altered
Spaceflight	Space Flight
Sex	Sex

Organism(s)

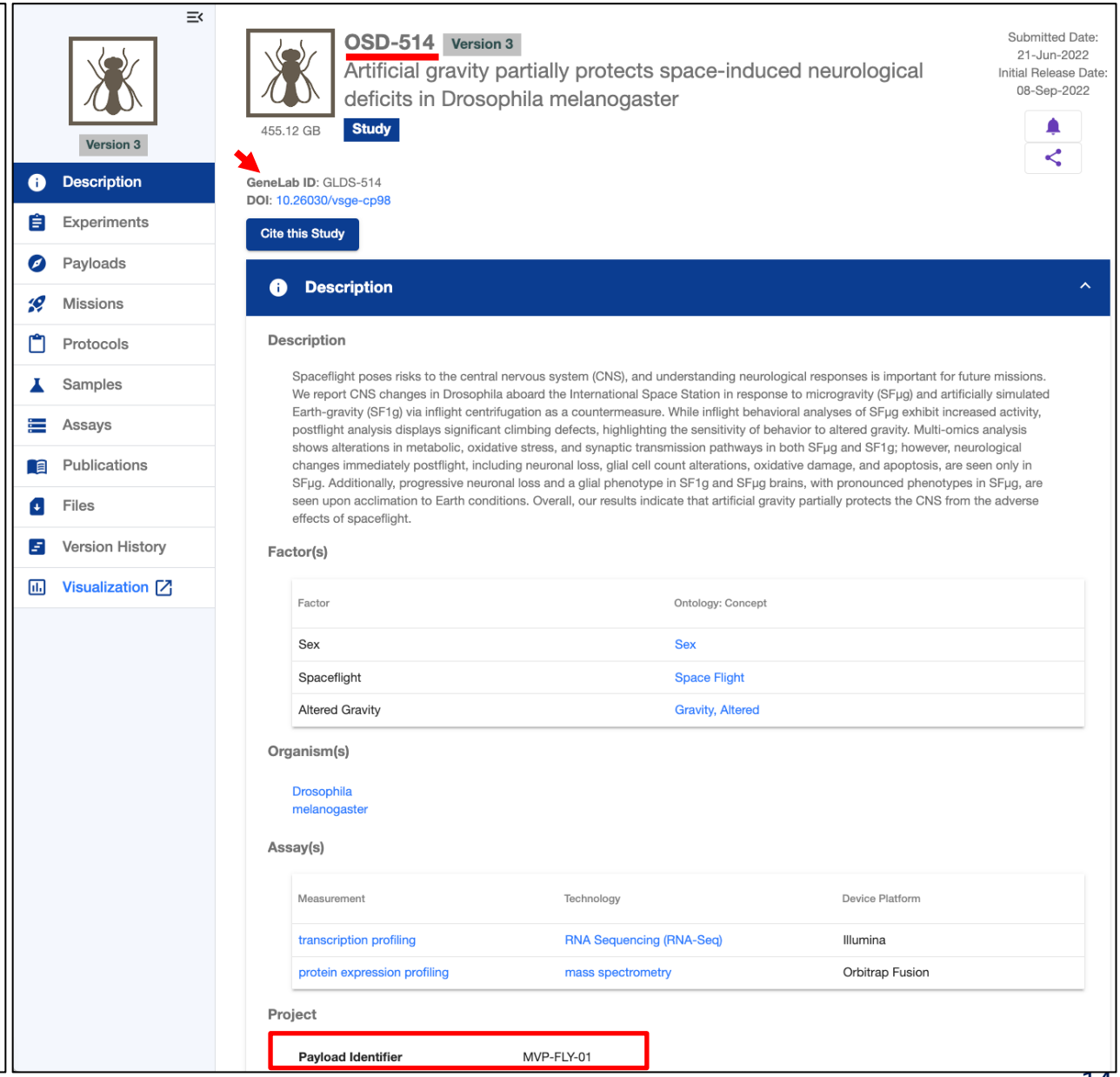
Drosophila melanogaster

Assay(s)

Measurement	Technology	Device Platform
Behavior	Locomotory Behavior	Manual Counting

Project

Payload Identifier	MVP-Fly-01



Version 3

- Description
- Experiments
- Payloads
- Missions
- Protocols
- Samples
- Assays
- Publications
- Files
- Version History
- Visualization

OSD-514 Version 3

Artificial gravity partially protects space-induced neurological deficits in Drosophila melanogaster

455.12 GB Study

Submitted Date: 21-Jun-2022
Initial Release Date: 08-Sep-2022

GeneLab ID: GLDS-514
DOI: 10.26030/vsge-cp98

Cite this Study

Description

Description

Spaceflight poses risks to the central nervous system (CNS), and understanding neurological responses is important for future missions. We report CNS changes in Drosophila aboard the International Space Station in response to microgravity (SFμg) and artificially simulated Earth-gravity (SF1g) via inflight centrifugation as a countermeasure. While inflight behavioral analyses of SFμg exhibit increased activity, postflight analysis displays significant climbing defects, highlighting the sensitivity of behavior to altered gravity. Multi-omics analysis shows alterations in metabolic, oxidative stress, and synaptic transmission pathways in both SFμg and SF1g; however, neurological changes immediately postflight, including neuronal loss, glial cell count alterations, oxidative damage, and apoptosis, are seen only in SFμg. Additionally, progressive neuronal loss and a glial phenotype in SF1g and SFμg brains, with pronounced phenotypes in SFμg, are seen upon acclimation to Earth conditions. Overall, our results indicate that artificial gravity partially protects the CNS from the adverse effects of spaceflight.

Factor(s)

Factor	Ontology: Concept
Sex	Sex
Spaceflight	Space Flight
Altered Gravity	Gravity, Altered

Organism(s)

Drosophila melanogaster

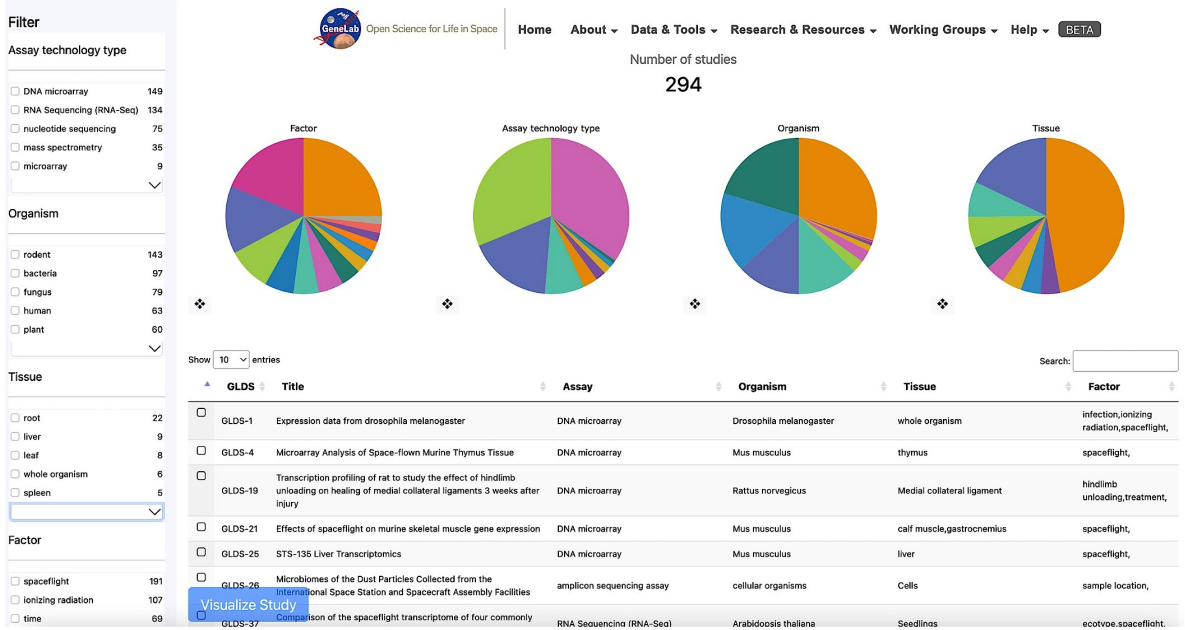
Assay(s)

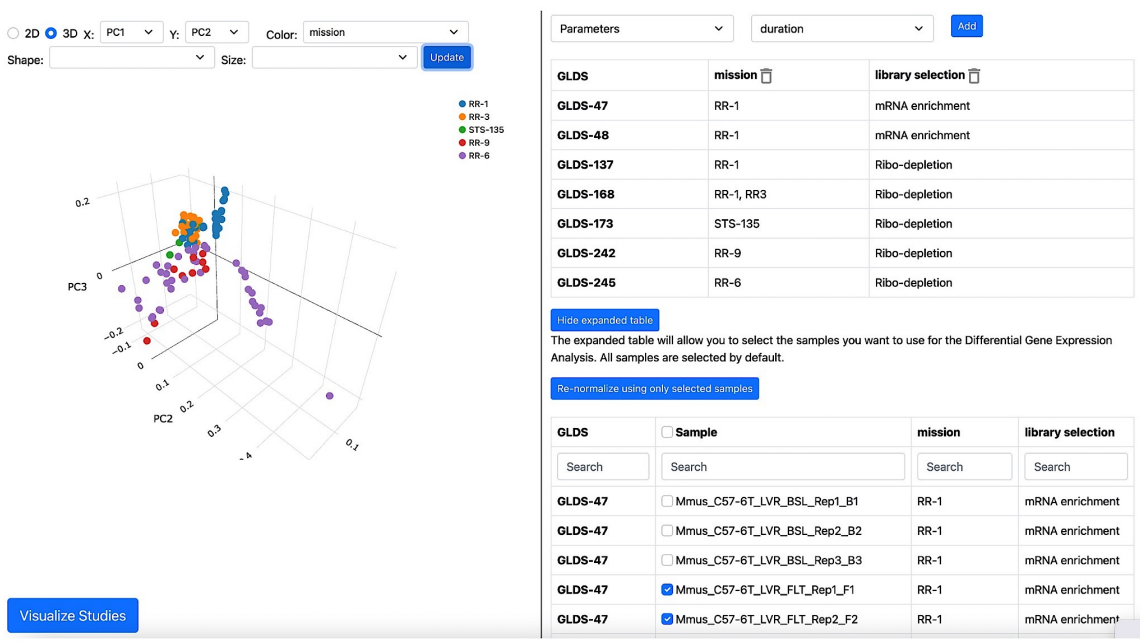
Measurement	Technology	Device Platform
transcription profiling	RNA Sequencing (RNA-Seq)	Illumina
protein expression profiling	mass spectrometry	Orbitrap Fusion

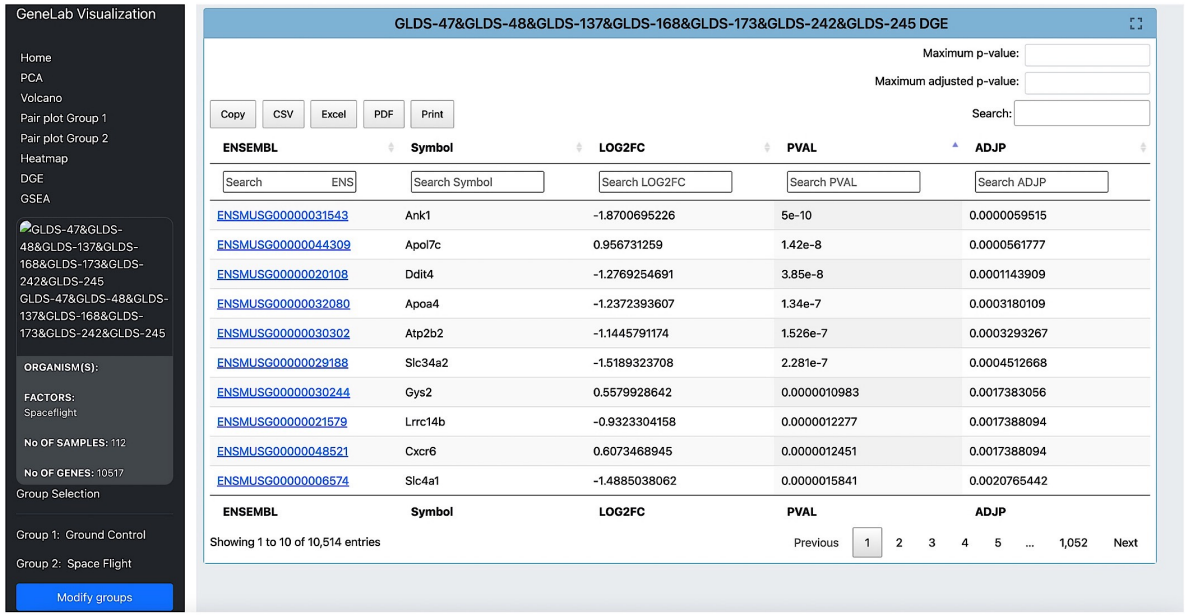
Project

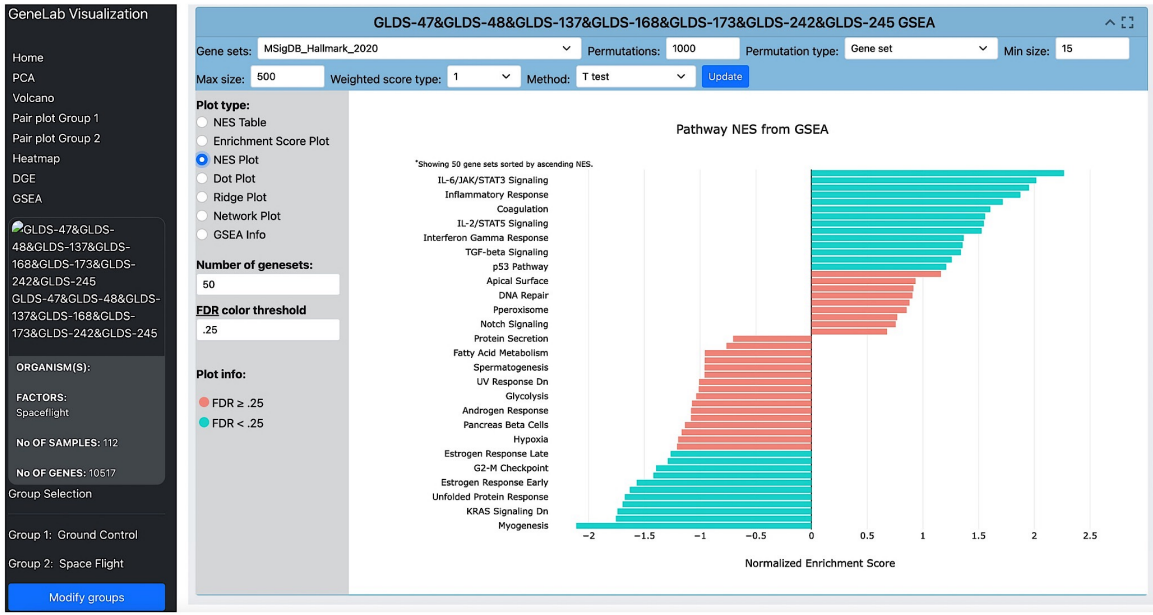
Payload Identifier	MVP-FLY-01

GeneLab Data Visualization Portal (https://visualization.genelab.nasa.gov/data/)









OSDR Database (GeneLab and ALSDA)

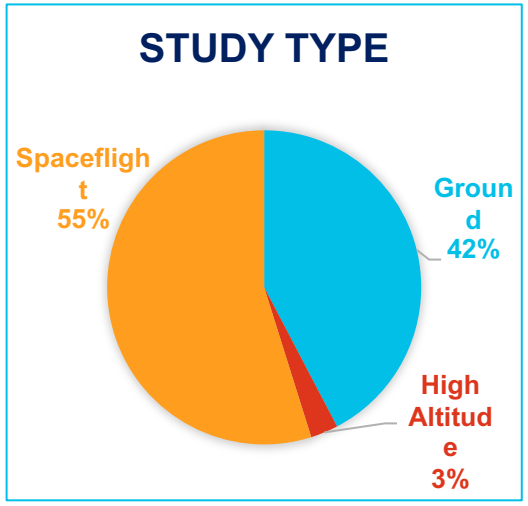
464
Studies

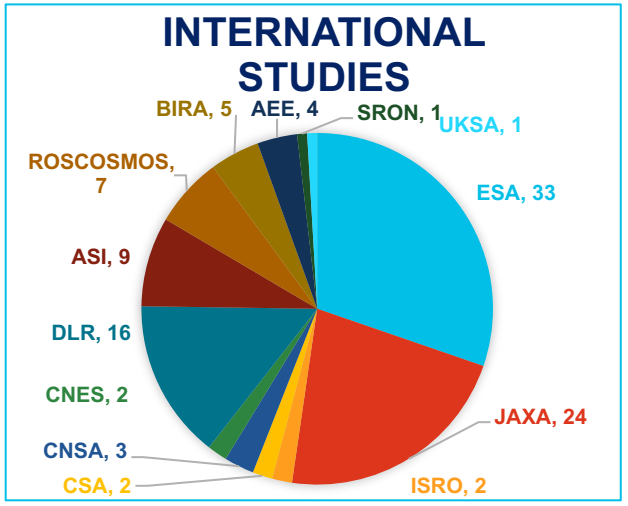
893
Datasets

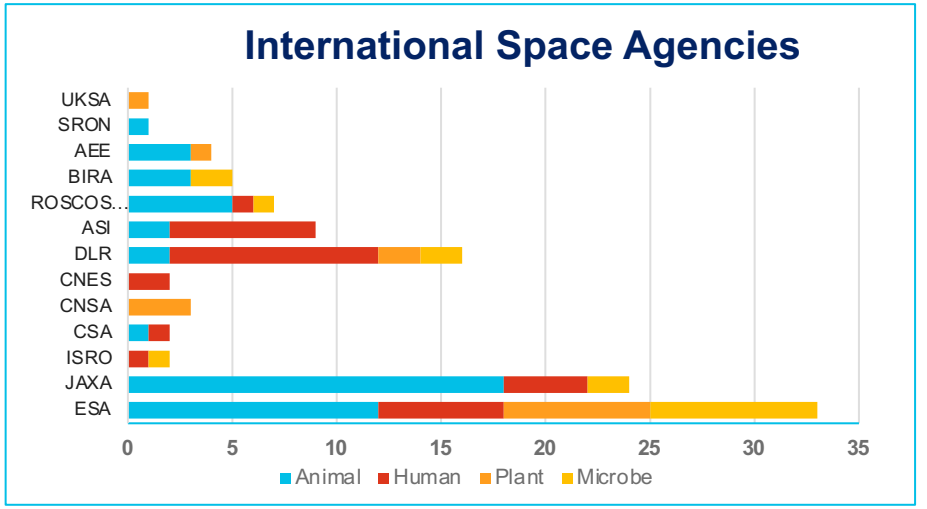
45
Species

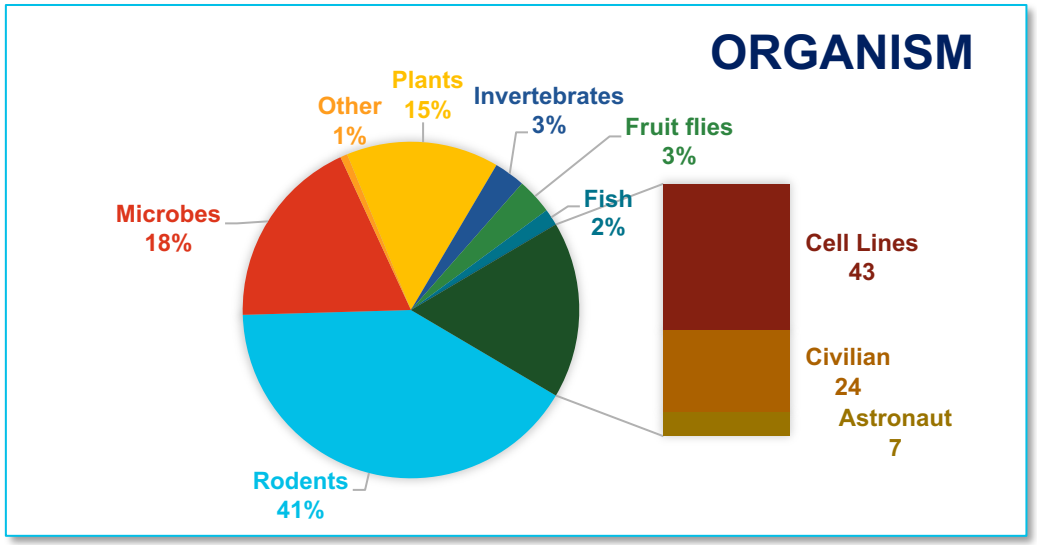
>60
Assays

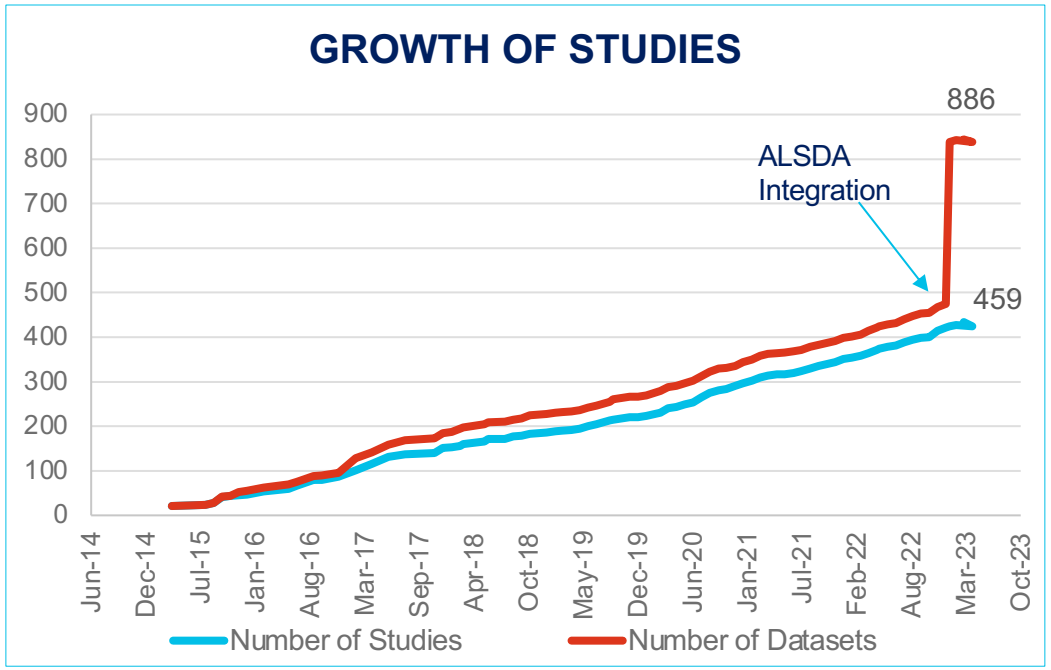
>150TB
Data









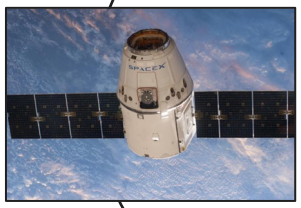


Civilian and Astronaut	Bed Rest, Spaceflight, Mars simulation
Cell Lines	Radiation (Ground), Simulated uG, Spaceflight, Parabolic Flight

Environmental Data Collected From Rodent Research Missions

Developing a data processing pipeline for Mission, Vehicle, and Hardware data. A pipeline that allows data repositories to easily collect, process, and store mission data in both CSV and JSON formats.

In Flight Research: ISS cabin and habitat module environments



Ascent and Return: Payload habitat environment

Return: Splashdown to Lab transport environment

Ground Control Operations: ISS cabin analog, habitat module, and lab to pad transport environments.

Data Collection Points

Relative Humidity Rodent Transporter Rodent Habitat ISS Cabin ISSES Cabin LAR Sensor Package	Temperature Rodent Transporter Rodent Habitat ISS Cabin ISSES Cabin LAR Sensor Package
Carbon dioxide ISS Cabin LAR Sensor Package	Acceleration Rodent Transporter LAR Sensor Package
Oxygen LAR Sensor Package	Radiation ISS Cabin

Environmental Data Visualization Portal (https://visualization.osdr.nasa.gov/eda/)

Environmental Data App

Mission Dashboard

RR-1

- Mission Info
- Telemetry Data
- Radiation Data

Mission Comparison

Data Tables

Environmental Data App

The Environmental Data App (EDA) is a portal where users can visualize and compare ISS (International Space Station) cabin environmental telemetry data and radiation data gathered from spaceflight missions. On the Mission Dashboard the users can visualize Temperature, Carbon dioxide (CO2), and Relative Humidity measurements and Radiation doses recorded in the ISS cabin and ground control for the duration of the mission. The Mission Comparison feature can be used to visualize data across multiple missions. Raw and summary data is available for download on the Data Tables tab through the Data Tables section.

[Rodent Research 1 (SpaceX-4)]

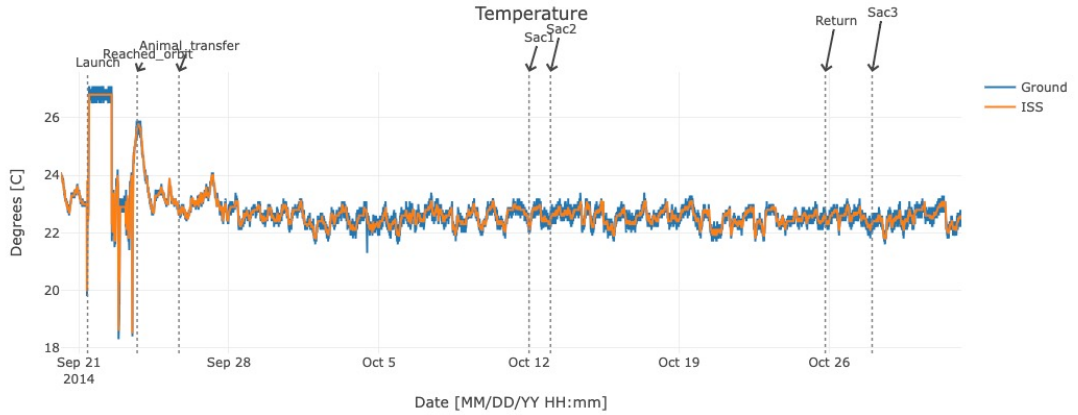
NASA's Rodent Research Hardware System provides a research platform aboard the International Space Station (ISS) for long-duration rodent experiments in space. Scientists and engineers at Ames Research Center (ARC) developed the new system for the space station based on the Animal Enclosure Module (AEM) that flew aboard 27 space shuttle missions between 1983 and 2011 and supported studies ranging from four to 18 days. Rodent spaceflight experiments have contributed significantly to our understanding of the effects of microgravity on biological processes that are directly relevant to humans in space. The maiden voyage of the system, Rodent Research-1 (RR1), launched on SpaceX-4 CRS-4 on September 21, 2014 and returned October 25, 2014 , during ISS expeditions 41/42. Lasting 37 days, RR1 was the longest duration spaceflight rodent study to date conducted in a NASA facility. RR1 was the first mission to transport rodents aboard an unmanned commercial vehicle. The primary goal is for NASA to validate hardware and demonstrate critical research operations, while supporting the Center for the Advancement of Science in Space (CASIS) in sponsoring the first commercial research study.

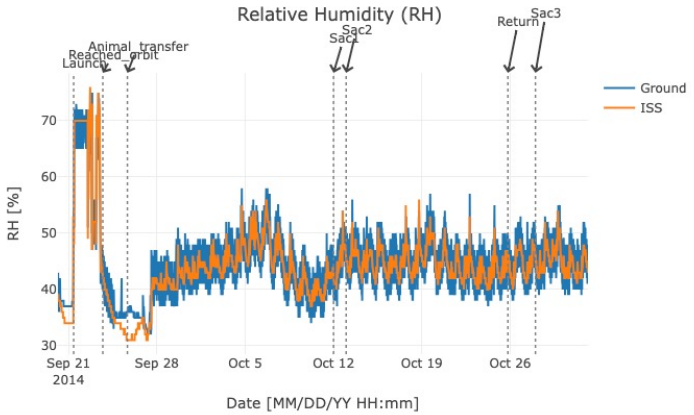
Mission Milestone Dates

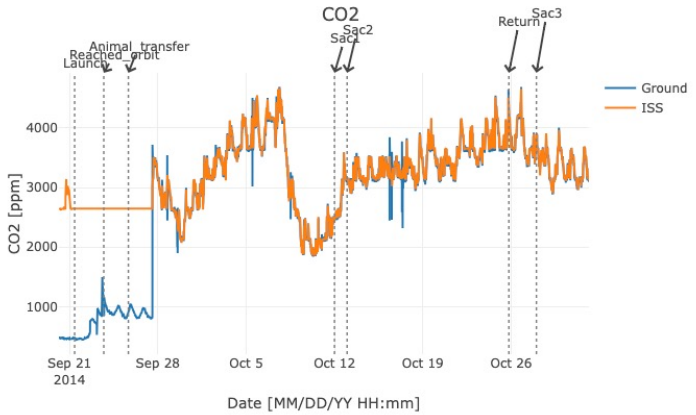
Display Mission milestones

- Launch: 2014-09-21 09:49
- Reached orbit: 2014-09-23 17:19
- Animal transfer: 2014-09-25 16:05
- Sac1: 2014-10-12 00:00
- Sac2: 2014-10-13 00:00
- Return: 2014-10-25 19:36
- Sac3: 2014-10-28 00:01

Telemetry data







Environmental Data Visualization Portal (https://visualization.osdr.nasa.gov/eda/)

Environmental Data App

Mission Dashboard

RR-1

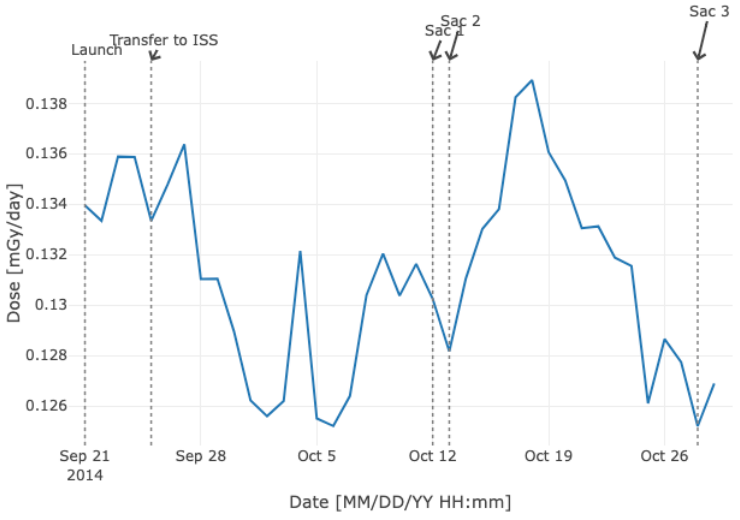
- Mission Info
- Telemetry Data
- Radiation Data

Mission Comparison

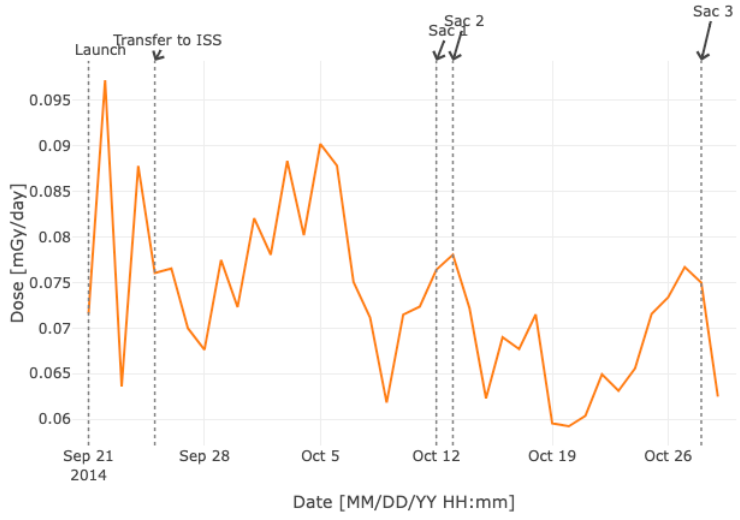
Data Tables

Radiation data

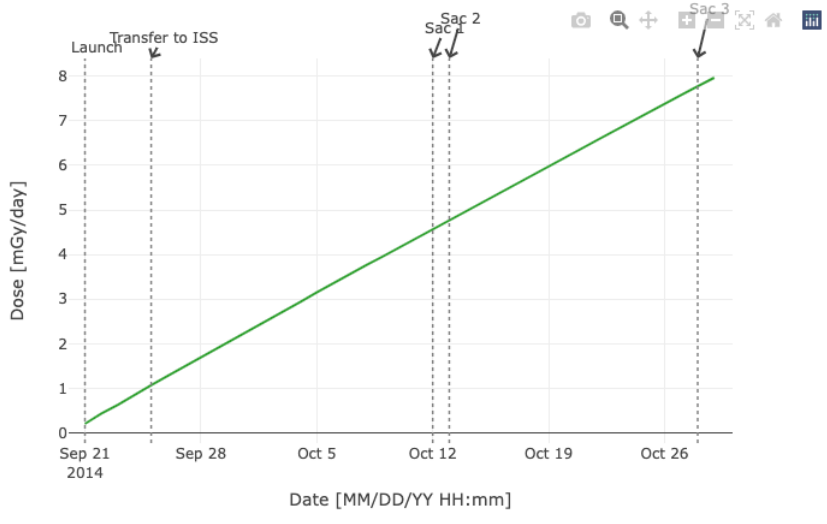
Galactic Cosmic Ray (GCR)



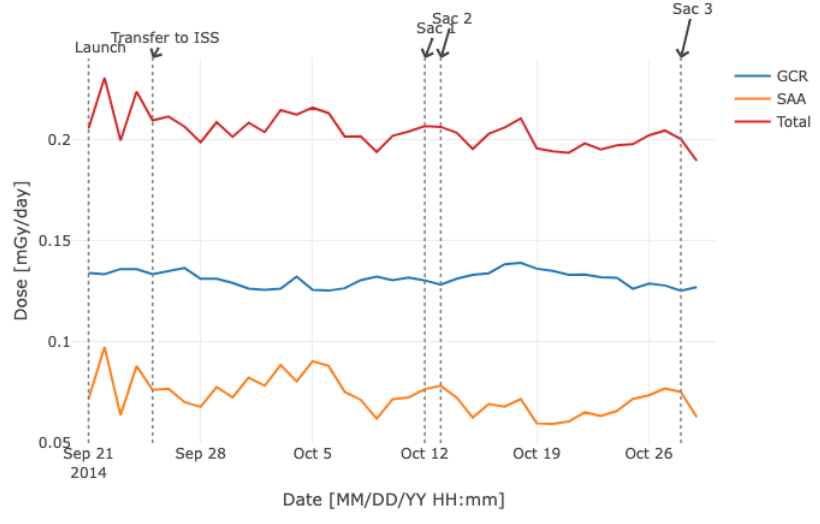
South Atlantic Anomaly (SAA)



Accumulated Radiation Dose



Total Radiation Dose



Open Science Analysis Working Groups (AWGs)

We invite you to join - https://osdr.nasa.gov/bio/awg/join.html

ANIMAL

128 members

Facilitates the use of omics in understanding basic mechanisms by which animals and constituent tissues and cells adapt to the spaceflight environment.

MULTI-OMICS

342 members

Interactions between the different omics to provide complete understanding of the entire system begin studied.

MICROBES

130 members

Focuses on analyzing microbial datasets within GeneLab that includes gene-expression, proteomic, metabolomic and environmental metagenomic datasets.

PLANTS

121 members

Share and discuss the latest developments in Astrobotany – the discipline of botany concerned with interactions between plant biology and space environment.

AI/ML

138 members

Focuses on developing data AI-readiness guidelines, algorithm and automation development, and developing ethical guidelines to increase trust and explainability surrounding AI in space biology.

ALSDA

249 members

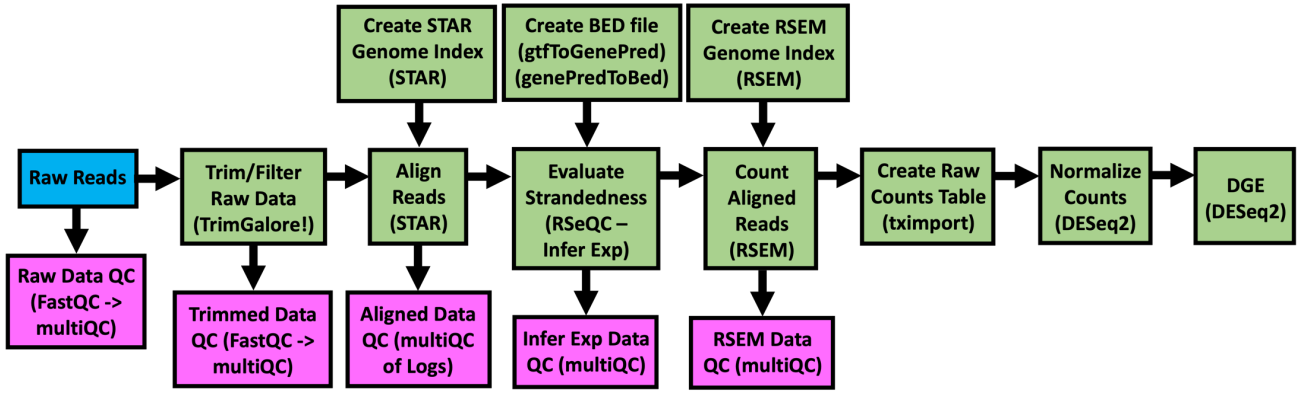
Feedback on science data and metadata standards for physiological, phenotypic, and behavioral datasets to be reusable. Datasets span from raw to processed-results data, and across tabular, bioimaging, and video formats.

*Consist of **500+ scientists** from multiple space agencies, international institutions, and industry. Scientists meet monthly with each group to provide feedback, develop standards, and analyze data.*

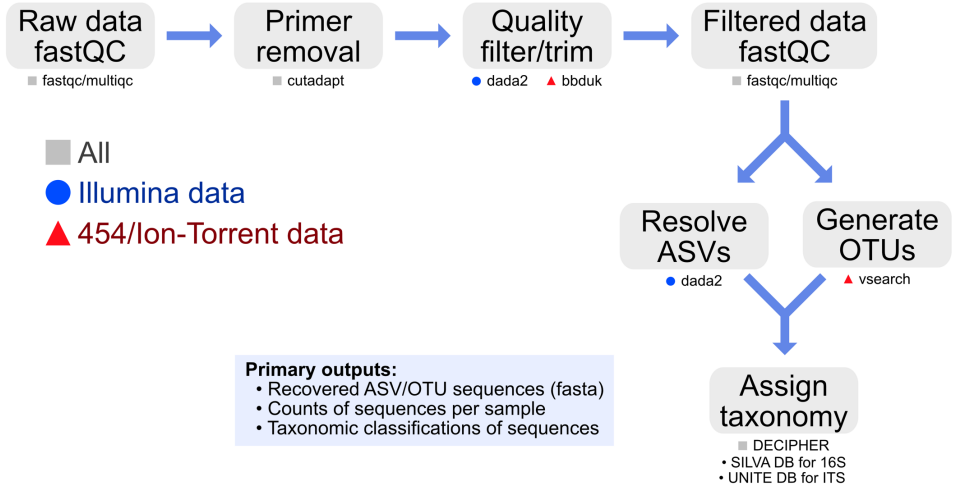
GeneLab Data Processing Pipelines

Build consensus data processing pipelines with the scientific community

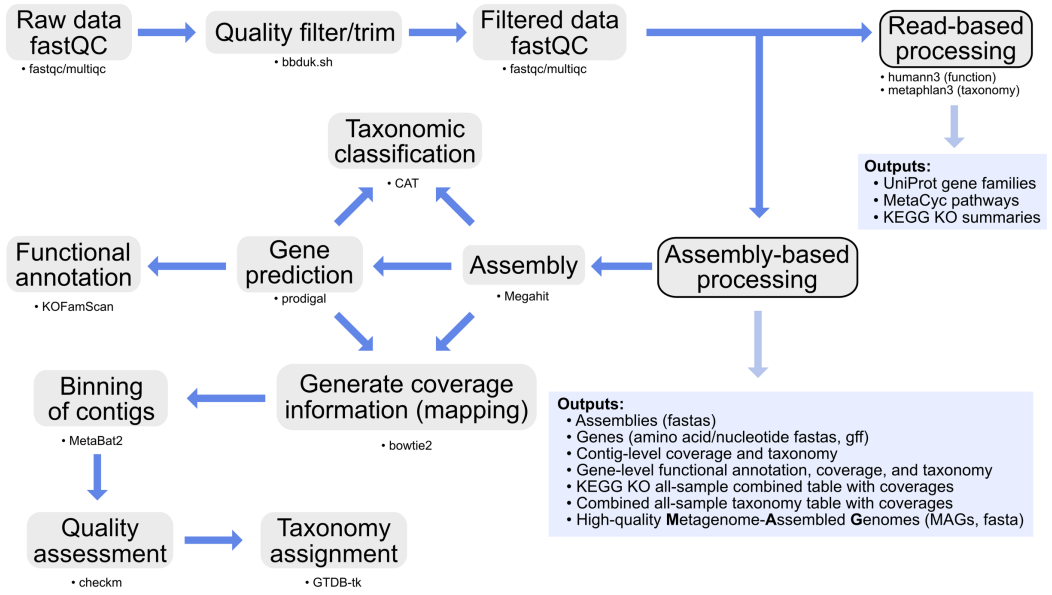
RNA Sequencing Data



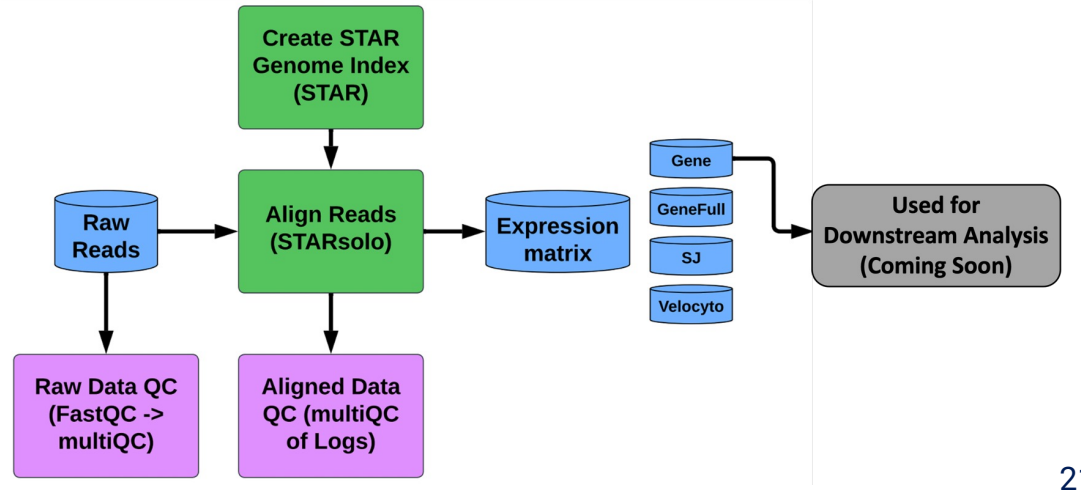
Amplicon Sequencing Data



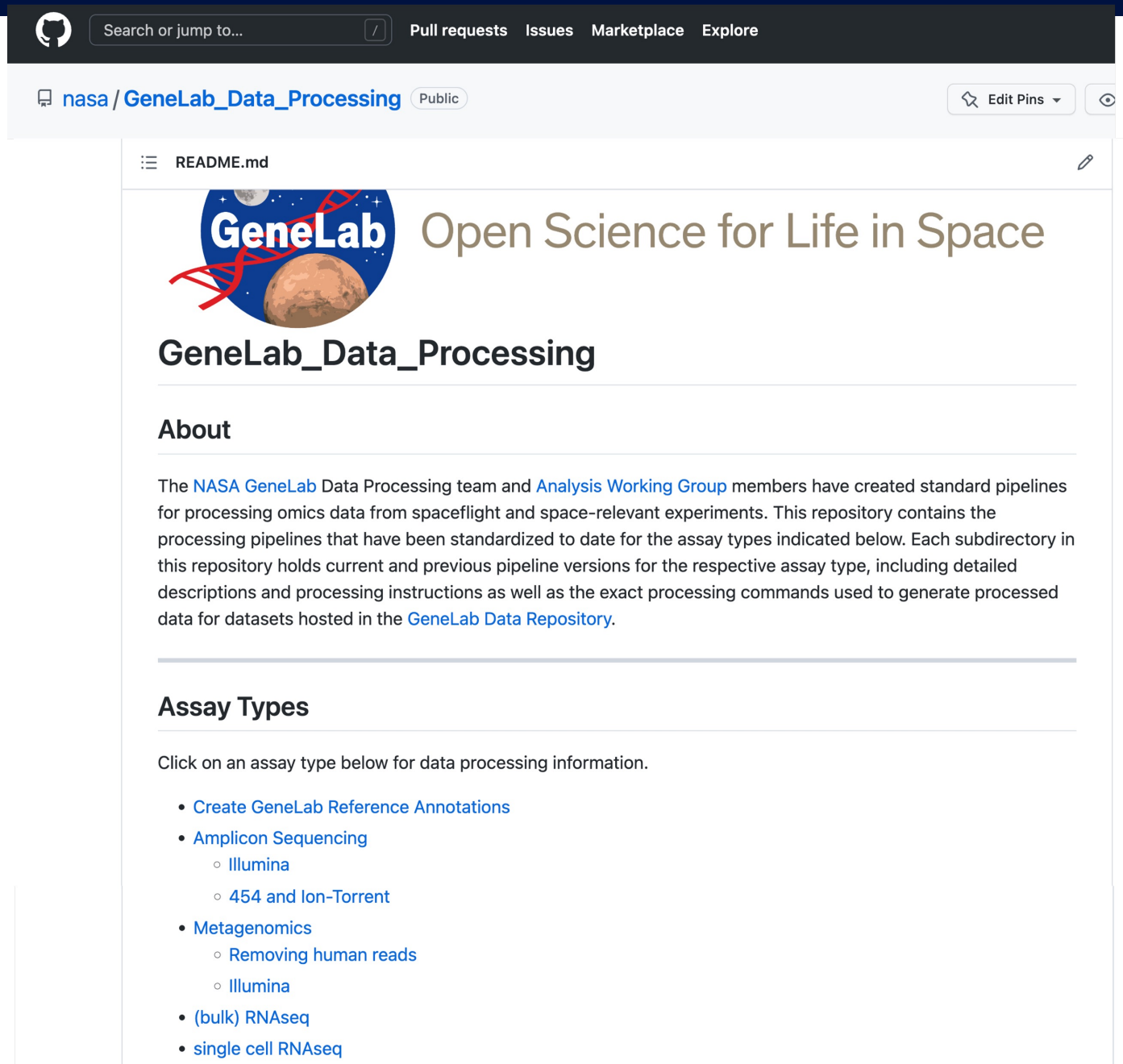
Metagenomics Data



Single Cell RNA Sequencing Data



GeneLab Data Processing GitHub Repo



https://github.com/nasa/GeneLab_Data_Processing

Data Reuse Enables New Discoveries

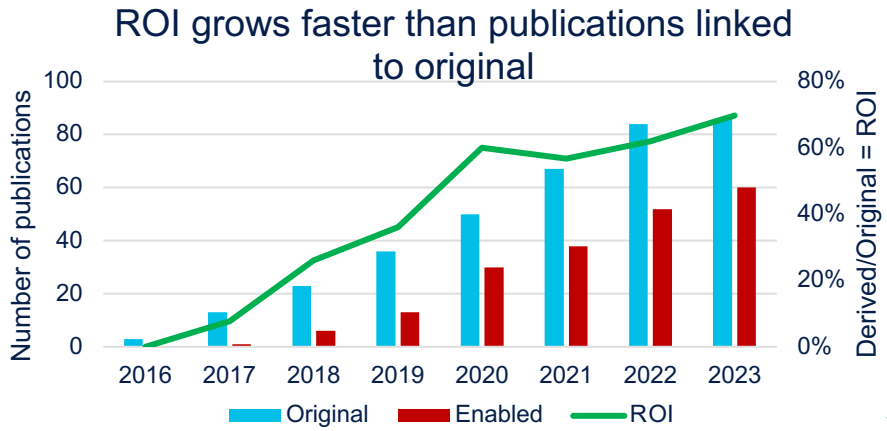
Return on Investment

90
Original Publication linked to OSDR

60
Enabled Publication linked to OSDR

80+
Presentations linked to OSDR

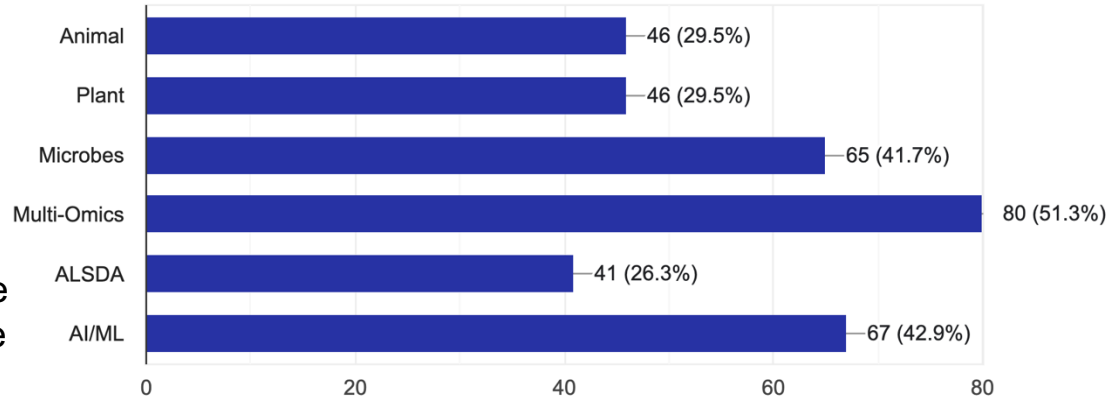
142+
Datasets used in enabled publications



Analysis Working Groups

156
New AWG Members

Members have the option to join more than one AWG



Latest Publications

- Microbiomes AWG: Multidrug-resistant Acinetobacter pittii is adapting to and exhibiting potential succession aboard the International Space Station
- Microbiomes AWG: Machine learning algorithm to characterize antimicrobial resistance associated with the International Space Station surface microbiome
- Plant AWG: Meta-analysis of the space flight and microgravity response of the Arabidopsis plant transcriptome
- AI/ML AWG: Research Prospects of Biomonitoring and Precision Health in Deep Space Supported by Artificial Intelligence
- AI/ML AWG: Recommendations on Biological Research and Self-driving Labs in Deep Space Supported by Artificial Intelligence
- GeneLab: Batch effect correction methods for NASA GeneLab transcriptomic datasets

INTERNATIONAL STATUS

ESA OMICS data arriving in OSDR: ~318 TB
Federation of all non-omics data with ESA

Training The Next Generation Of Scientists

GeneLab for High Schools (GL4HS):
A four-week intensive training summer program for rising high school juniors and seniors to learn bioinformatics and computational biology methods and techniques to analyze space omics data.
Learn more and apply at: https://www.nasa.gov/ames/genelab-for-high-schools

GeneLab for Colleges/Universities (GL4U):
For educators and students to learn how analyze omics data using GeneLab standard pipelines and space-relevant data

Access to course materials: https://github.com/nasa/GeneLab-Training/tree/main/GL4U

Space Life Sciences Training Program (SLSTP):
Provides undergraduate students entering their junior or senior years, and entering graduate students, with professional experience in space life science disciplines.

Learn more and apply at: https://www.nasa.gov/ames/research/space-life-sciences-training-program

Spaceflight Technology, Applications, and Research (STAR):
Annual course for PIs, senior research scientists, and postdoctoral scholars and aims to facilitate their entry to space biology and preparation for conducting spaceflight experiments using NASA and commercial platforms.

NASA Postdoctoral Program:
If you're an early-career or senior scientist, apply to the NASA Postdoctoral Program to help pursue NASA's mission and experience the world's most diverse technology and expertise.

Complete NASA's open science curriculum!

Open Science 101:

A community-developed introduction to **core open science skills**

- Know how to write a NASA open science and data management plan
- Learn about tools and best practices
- Increase the impact & visibility of your science
- Earn your digital NASA open science badge

All 5 modules now available through a self-paced online course and through in-person and virtual instructor-led workshops.

Take OS101!

https://go.nasa.gov/40pPQMx

TOPS @ AGU 2023

Sessions:

NASA's Transform to Open Science (TOPS) initiative
- Thursday, Dec 14, 2:12 - 2:22 p.m.
- 2014 - West (Level 2, West)

U51A - Science for All - Organizing Agile Openness
- Friday, Dec. 15, 8:30 - 10 a.m.
- *303-304 - South (Level 3, South)*

Voices from the Global Open Science Community: Showcasing Progress, Challenges, and Lessons Learned
- Friday, Dec. 15, 8:30 A.M. PST
- *301-302 - South (Level 3, South)*

Town Hall:

NASA Earth Science Division
- Tues, Dec. 12, 6:30 - 8:30 p.m.
- 2020 - West (Level 2, West)

Exhibit Hall:

NASA Exhibit Booth #531
Hyperwall Talks - Tuesday, Dec. 12
2:45 - 3:00 PM - Open Source Science Initiative
3:00 - 3:15 PM - Transforming to Open Science

Dec. 11-15, 2023

Moscone Center
San Francisco, CA

Accessing OSDR 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*

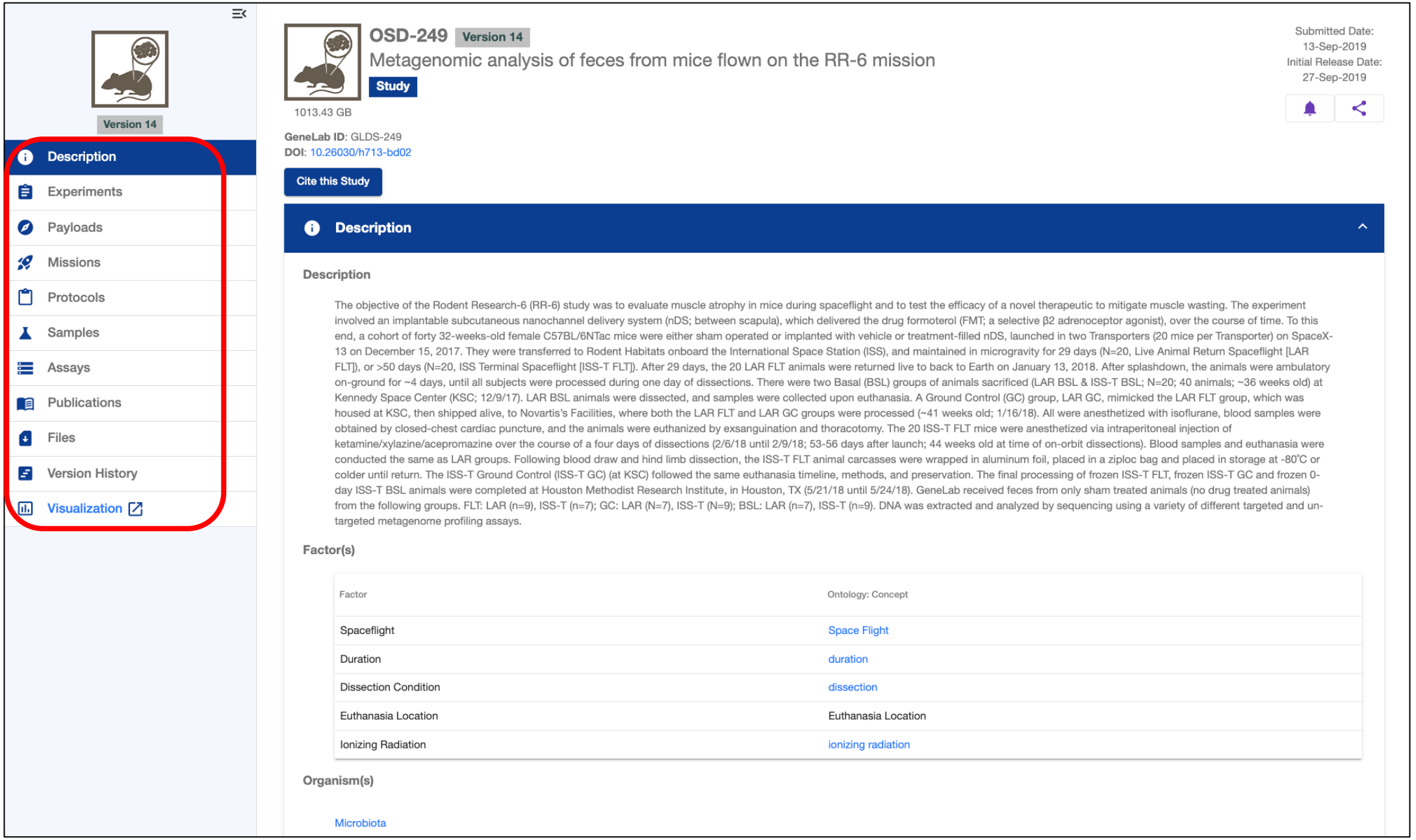
Study

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 ...

OSD-249 (https://osdr.nasa.gov/bio/repo/data/studies/OSD-249)



Version 14

- Description
- Experiments
- Payloads
- Missions
- Protocols
- Samples
- Assays
- Publications
- Files
- Version History
- Visualization

OSD-249 Version 14

Metagenomic analysis of feces from mice flown on the RR-6 mission

Study

1013.43 GB

Submitted Date: 13-Sep-2019
Initial Release Date: 27-Sep-2019

GeneLab ID: GLDS-249
DOI: 10.26030/h713-bd02

Cite this Study

Description

Description

The objective of the Rodent Research-6 (RR-6) study was to evaluate muscle atrophy in mice during spaceflight and to test the efficacy of a novel therapeutic to mitigate muscle wasting. The experiment involved an implantable subcutaneous nanochannel delivery system (nDS; between scapula), which delivered the drug formoterol (FMT; a selective β2 adrenoceptor agonist), over the course of time. To this end, a cohort of forty 32-weeks-old female C57BL/6NTac mice were either sham operated or implanted with vehicle or treatment-filled nDS, launched in two Transporters (20 mice per Transporter) on SpaceX-13 on December 15, 2017. They were transferred to Rodent Habitats onboard the International Space Station (ISS), and maintained in microgravity for 29 days (N=20, Live Animal Return Spaceflight [LAR FLT]), or >50 days (N=20, ISS Terminal Spaceflight [ISS-T FLT]). After 29 days, the 20 LAR FLT animals were returned live to back to Earth on January 13, 2018. After splashdown, the animals were ambulatory on-ground for ~4 days, until all subjects were processed during one day of dissections. There were two Basal (BSL) groups of animals sacrificed (LAR BSL & ISS-T BSL; N=20; 40 animals; ~36 weeks old) at Kennedy Space Center (KSC; 12/9/17). LAR BSL animals were dissected, and samples were collected upon euthanasia. A Ground Control (GC) group, LAR GC, mimicked the LAR FLT group, which was housed at KSC, then shipped alive, to Novartis's Facilities, where both the LAR FLT and LAR GC groups were processed (~41 weeks old; 1/16/18). All were anesthetized with isoflurane, blood samples were obtained by closed-chest cardiac puncture, and the animals were euthanized by exsanguination and thoracotomy. The 20 ISS-T FLT mice were anesthetized via intraperitoneal injection of ketamine/xylazine/acepromazine over the course of a four days of dissections (2/6/18 until 2/9/18; 53-56 days after launch; 44 weeks old at time of on-orbit dissections). Blood samples and euthanasia were conducted the same as LAR groups. Following blood draw and hind limb dissection, the ISS-T FLT animal carcasses were wrapped in aluminum foil, placed in a ziploc bag and placed in storage at -80˚C or colder until return. The ISS-T Ground Control (ISS-T GC) (at KSC) followed the same euthanasia timeline, methods, and preservation. The final processing of frozen ISS-T FLT, frozen ISS-T GC and frozen 0-day ISS-T BSL animals were completed at Houston Methodist Research Institute, in Houston, TX (5/21/18 until 5/24/18). GeneLab received feces from only sham treated animals (no drug treated animals) from the following groups. FLT: LAR (n=9), ISS-T (n=7); GC: LAR (N=7), ISS-T (N=9); BSL: LAR (n=7), ISS-T (n=9). DNA was extracted and analyzed by sequencing using a variety of different targeted and un-targeted metagenome profiling assays.

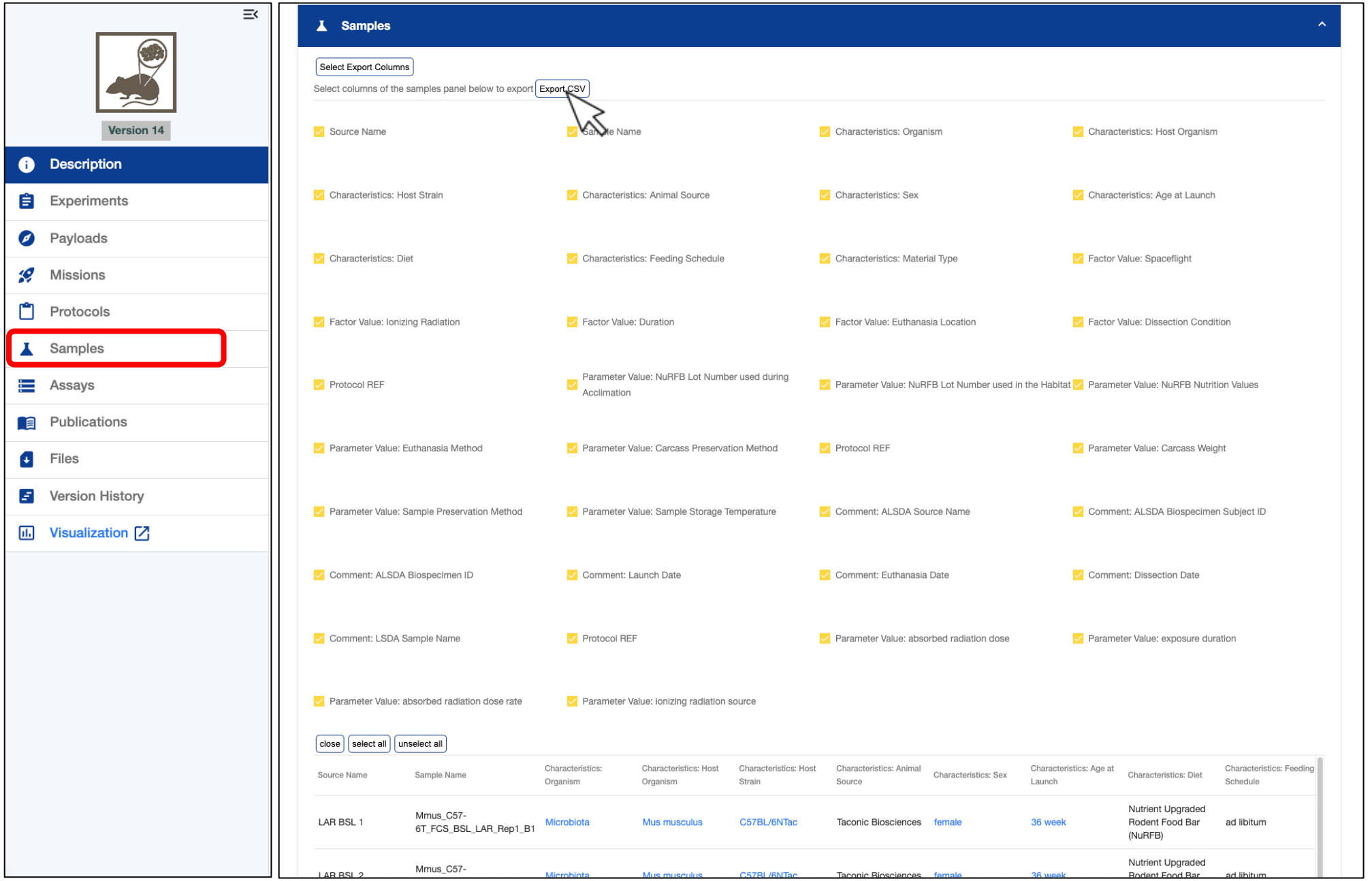
Factor(s)

Factor	Ontology: Concept
Spaceflight	Space Flight
Duration	duration
Dissection Condition	dissection
Euthanasia Location	Euthanasia Location
Ionizing Radiation	ionizing radiation

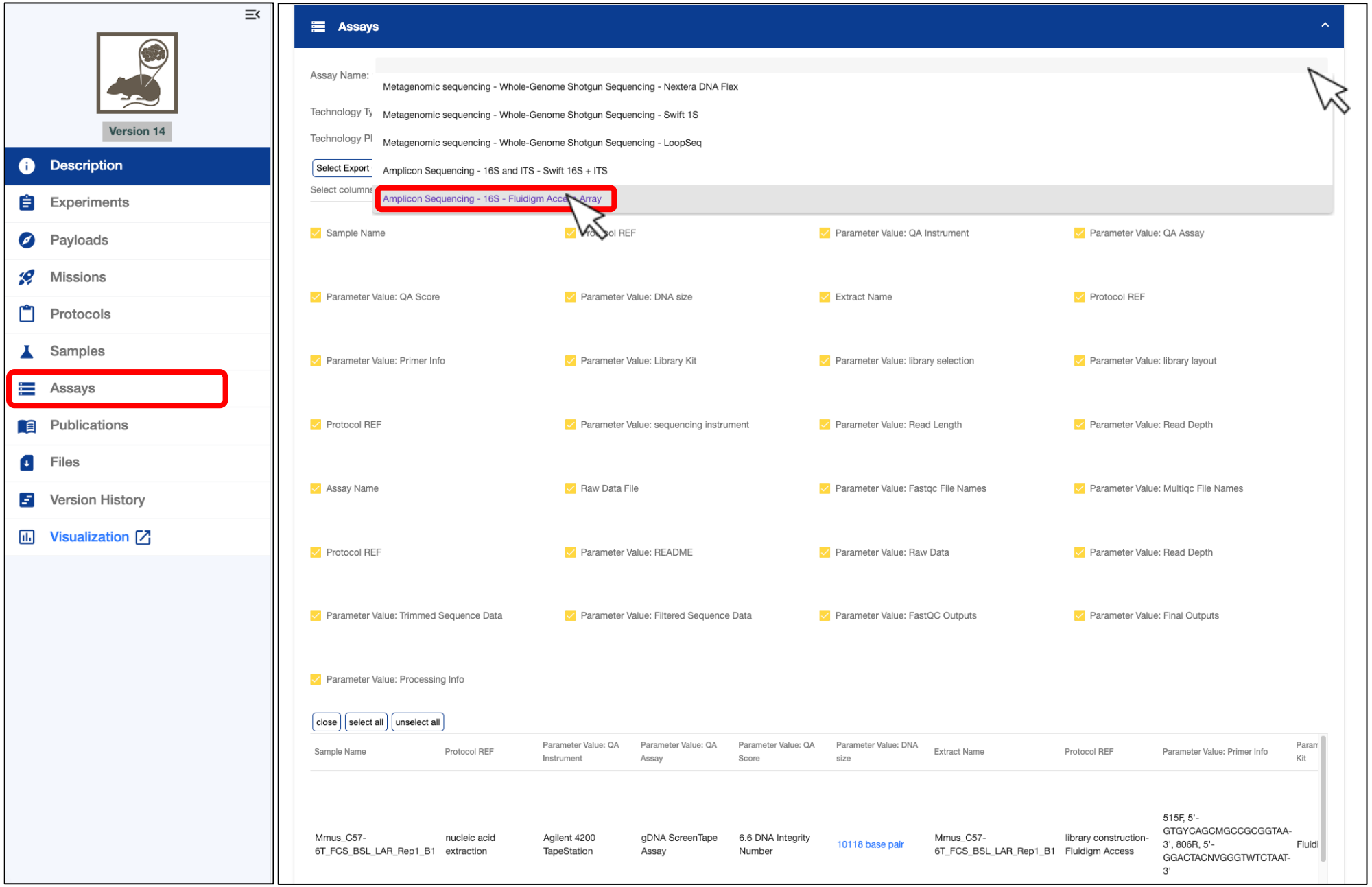
Organism(s)

Microbiota

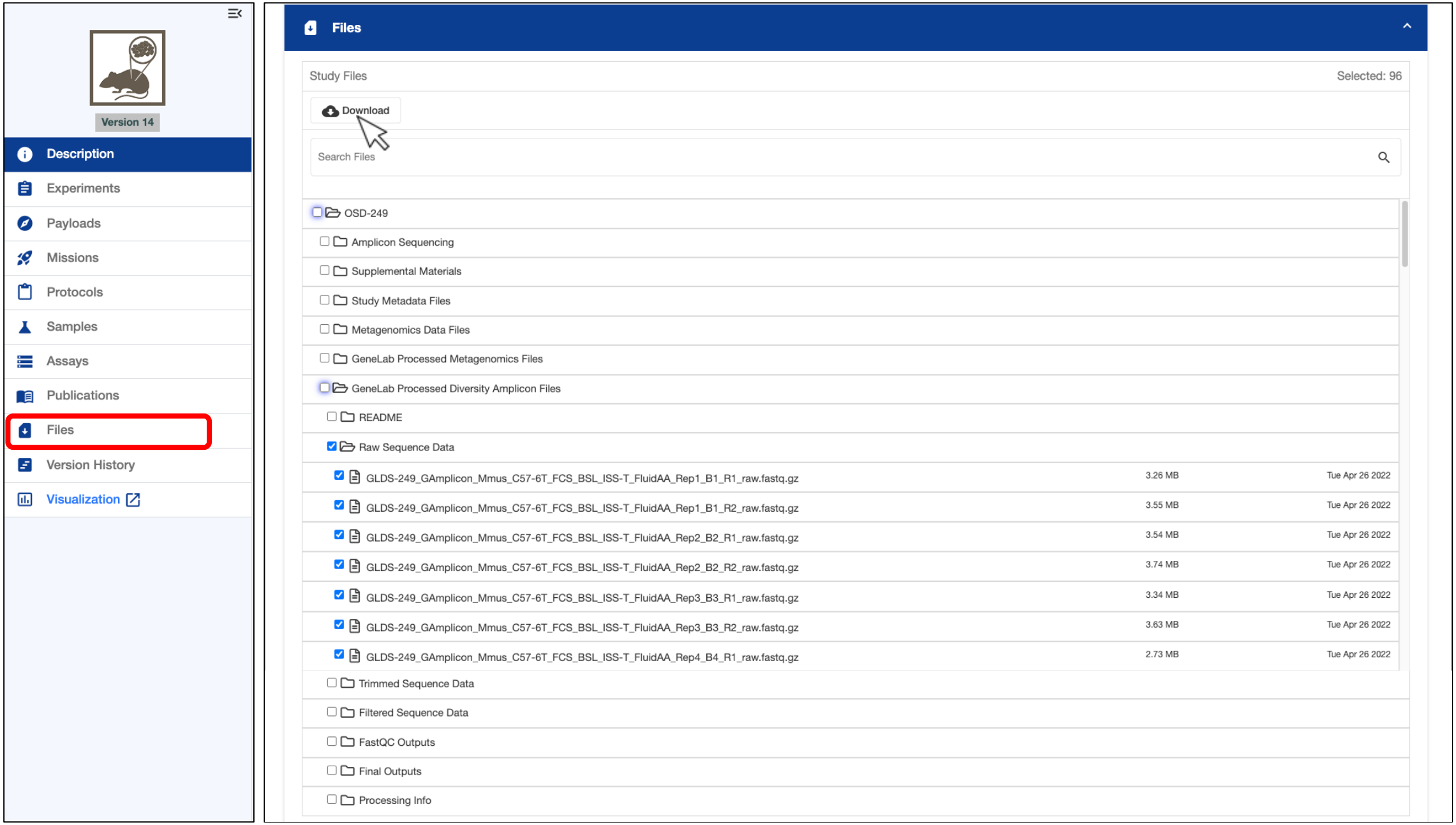
OSD-249: Export Specific Sample Metadata Table Columns



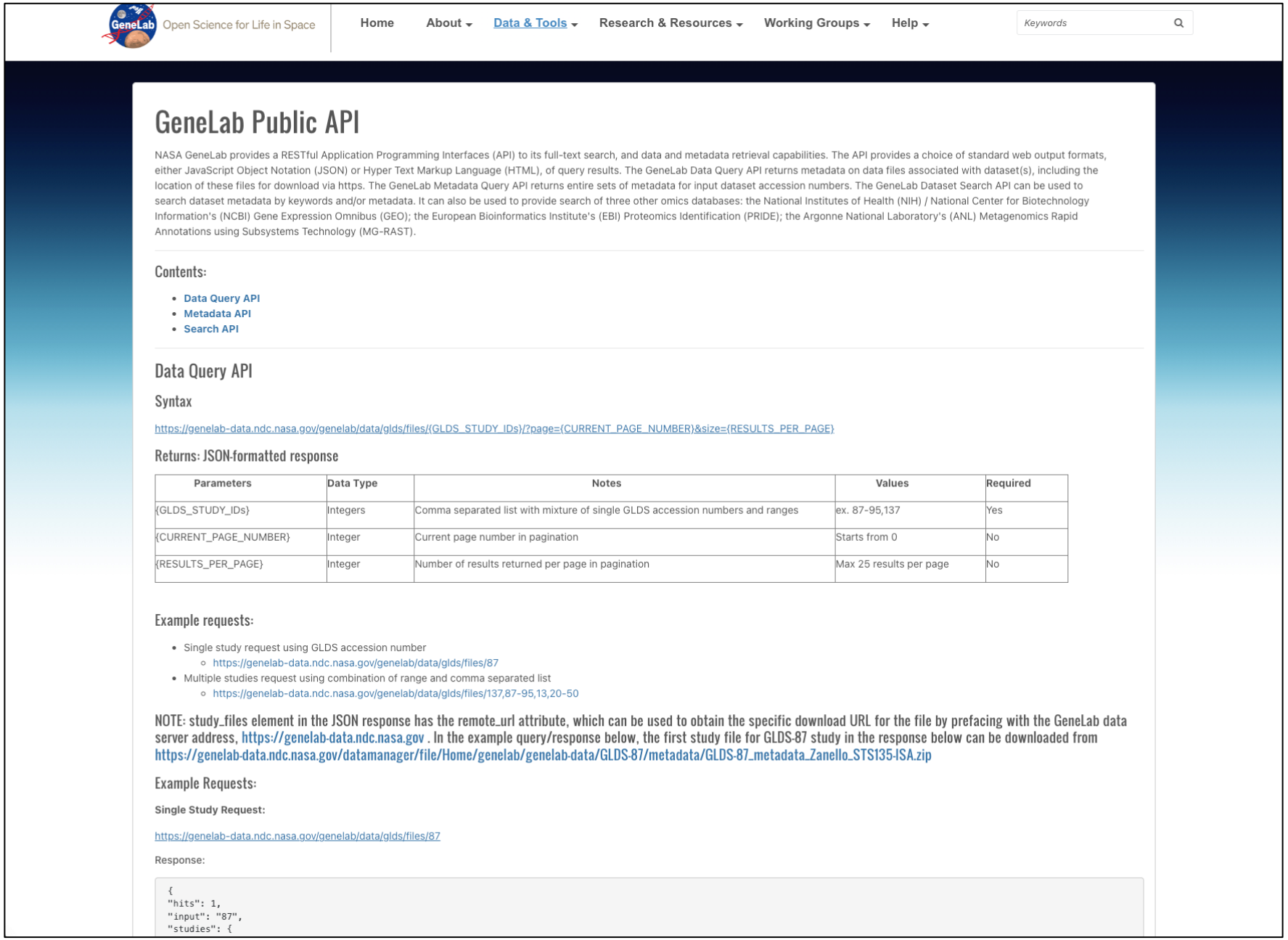
OSD-249: Export Specific Assay Metadata Table Columns



OSD-249: Download Files



Accessing OSDR Data – Public API (https://genelab.nasa.gov/genelabAPIs)



Open Science for Life in Space

Home | About | Data & Tools | Research & Resources | Working Groups | Help | Keywords

GeneLab Public API

NASA GeneLab provides a RESTful Application Programming Interfaces (API) to its full-text search, and data and metadata retrieval capabilities. The API provides a choice of standard web output formats, either JavaScript Object Notation (JSON) or Hyper Text Markup Language (HTML), of query results. The GeneLab Data Query API returns metadata on data files associated with dataset(s), including the location of these files for download via https. The GeneLab Metadata Query API returns entire sets of metadata for input dataset accession numbers. The GeneLab Dataset Search API can be used to search dataset metadata by keywords and/or metadata. It can also be used to provide search of three other omics databases: the National Institutes of Health (NIH) / National Center for Biotechnology Information's (NCBI) Gene Expression Omnibus (GEO); the European Bioinformatics Institute's (EBI) Proteomics Identification (PRIDE); the Argonne National Laboratory's (ANL) Metagenomics Rapid Annotations using Subsystems Technology (MG-RAST).

Contents:

- Data Query API
- Metadata API
- Search API

Data Query API

Syntax

https://genelab-data.ndc.nasa.gov/genelab/data/glds/files/{GLDS_STUDY_IDs}/?page={CURRENT_PAGE_NUMBER}&size={RESULTS_PER_PAGE}

Returns: JSON-formatted response

Parameters	Data Type	Notes	Values	Required
{GLDS_STUDY_IDs}	Integers	Comma separated list with mixture of single GLDS accession numbers and ranges	ex. 87-95,137	Yes
{CURRENT_PAGE_NUMBER}	Integer	Current page number in pagination	Starts from 0	No
{RESULTS_PER_PAGE}	Integer	Number of results returned per page in pagination	Max 25 results per page	No

Example requests:

- Single study request using GLDS accession number
  - https://genelab-data.ndc.nasa.gov/genelab/data/glds/files/87
- Multiple studies request using combination of range and comma separated list
  - https://genelab-data.ndc.nasa.gov/genelab/data/glds/files/137,87-95,13,20-50

NOTE: study_files element in the JSON response has the remote_url attribute, which can be used to obtain the specific download URL for the file by prefacing with the GeneLab data server address, https://genelab-data.ndc.nasa.gov . In the example query/response below, the first study file for GLDS-87 study in the response below can be downloaded from https://genelab-data.ndc.nasa.gov/datamanager/file/Home/genelab/genelab-data/GLDS-87/metadata/GLDS-87_metadata_Zanello_STS135-ISA.zip

Example Requests:

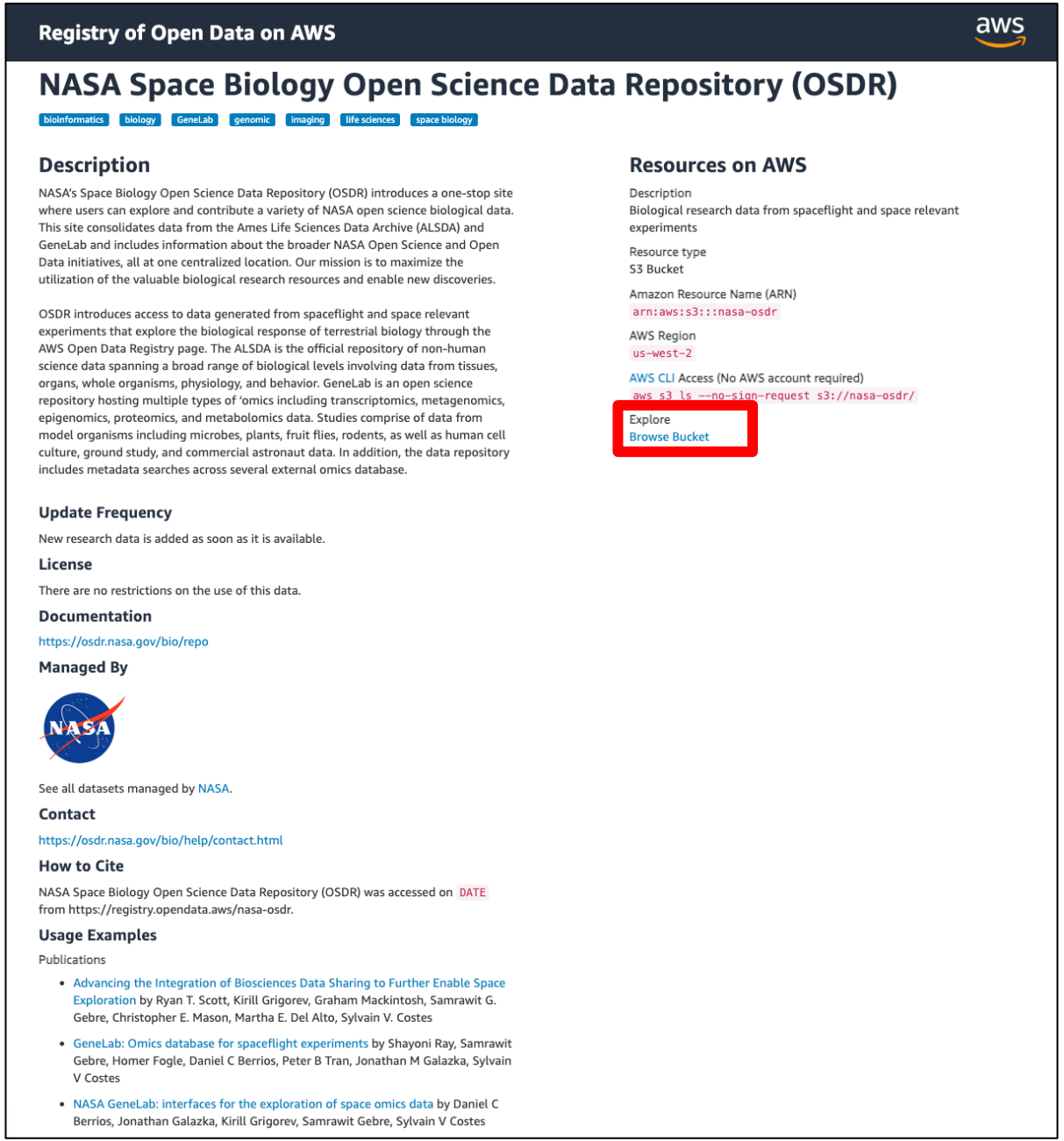
Single Study Request:

https://genelab-data.ndc.nasa.gov/genelab/data/glds/files/87

Response:

```
{
"hits": 1,
"input": "87",
"studies": {
```

Accessing OSDR Data Files – Public s3 Bucket (https://registry.opendata.aws/nasa-osdr/)



Registry of Open Data on AWS

aws

NASA Space Biology Open Science Data Repository (OSDR)

bioinformatics | biology | GeneLab | genomic | imaging | life sciences | space biology

Description

NASA's Space Biology Open Science Data Repository (OSDR) introduces a one-stop site where users can explore and contribute a variety of NASA open science biological data. This site consolidates data from the Ames Life Sciences Data Archive (ALSDA) and GeneLab and includes information about the broader NASA Open Science and Open Data initiatives, all at one centralized location. Our mission is to maximize the utilization of the valuable biological research resources and enable new discoveries.

OSDR introduces access to data generated from spaceflight and space relevant experiments that explore the biological response of terrestrial biology through the AWS Open Data Registry page. The ALSDA is the official repository of non-human science data spanning a broad range of biological levels involving data from tissues, organs, whole organisms, physiology, and behavior. GeneLab is an open science repository hosting multiple types of 'omics including transcriptomics, metagenomics, epigenomics, proteomics, and metabolomics data. Studies comprise of data from model organisms including microbes, plants, fruit flies, rodents, as well as human cell culture, ground study, and commercial astronaut data. In addition, the data repository includes metadata searches across several external omics database.

Update Frequency

New research data is added as soon as it is available.

License

There are no restrictions on the use of this data.

Documentation

https://osdr.nasa.gov/bio/repo

Managed By

See all datasets managed by NASA.

Contact

https://osdr.nasa.gov/bio/help/contact.html

How to Cite

NASA Space Biology Open Science Data Repository (OSDR) was accessed on `DATE` from https://registry.opendata.aws/nasa-osdr.

Usage Examples

Publications

- Advancing the Integration of Biosciences Data Sharing to Further Enable Space Exploration by Ryan T. Scott, Kirill Grigorev, Graham Mackintosh, Samrawit G. Gebre, Christopher E. Mason, Martha E. Del Alto, Sylvain V. Costes
- GeneLab: Omics database for spaceflight experiments by Shayoni Ray, Samrawit Gebre, Homer Fogle, Daniel C Berrios, Peter B Tran, Jonathan M Galazka, Sylvain V Costes
- NASA GeneLab: interfaces for the exploration of space omics data by Daniel C Berrios, Jonathan Galazka, Kirill Grigorev, Samrawit Gebre, Sylvain V Costes

Resources on AWS

Description
Biological research data from spaceflight and space relevant experiments

Resource type
S3 Bucket

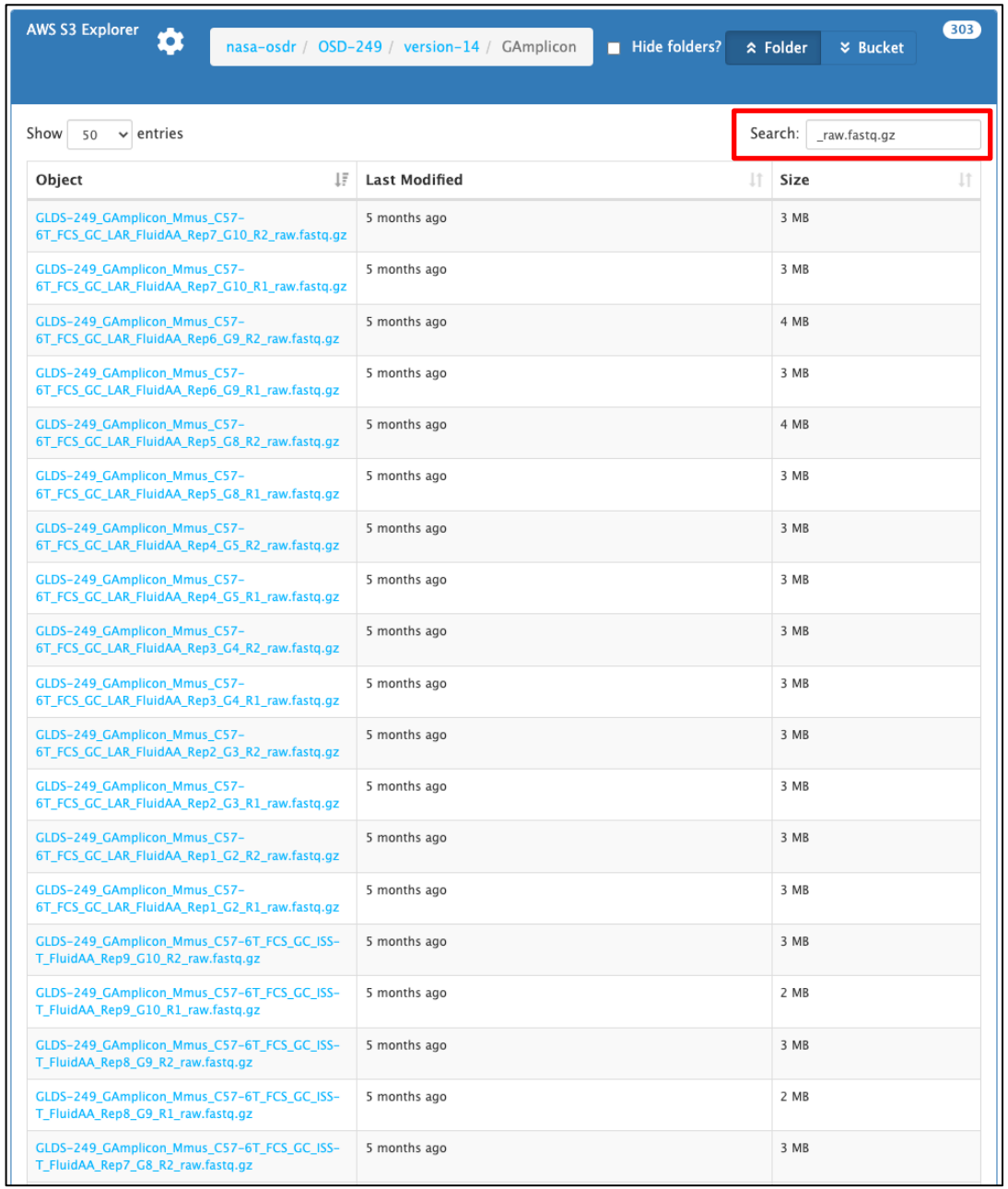
Amazon Resource Name (ARN)
`arn:aws:s3:::nasa-osdr`

AWS Region
`us-west-2`

AWS CLI Access (No AWS account required)
`aws s3 ls --no-sign-request s3://nasa-osdr/`

Explore
Browse Bucket

Accessing OSDR Data Files – Public s3 Bucket (https://registry.opendata.aws/nasa-osdr/)



Amplicon (Illumina) Sequencing Overview

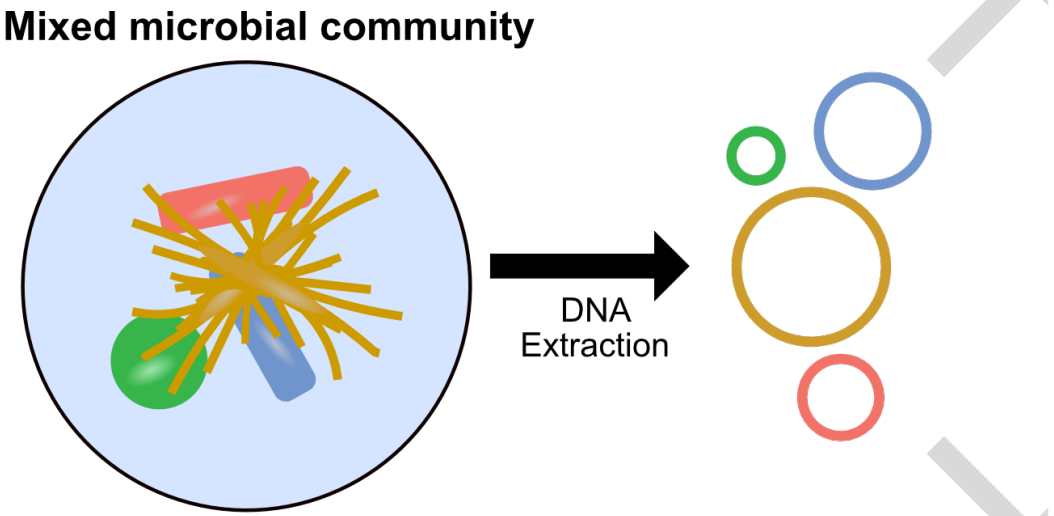
Common Sequencing Approaches in Microbial Ecology

Amplicon sequencing involves the use of “primers” – short sequences that match highly conserved stretches of DNA

These primers bind to and help amplify DNA from microbes in our sample, usually targeting a portion of a single gene

We then use these recovered sequences from our sample to try to get information about the microbial community that is present (often trying to figure out “who” is there)

Metagenomics sequencing involves attempting to sequence “all” the DNA present in a system

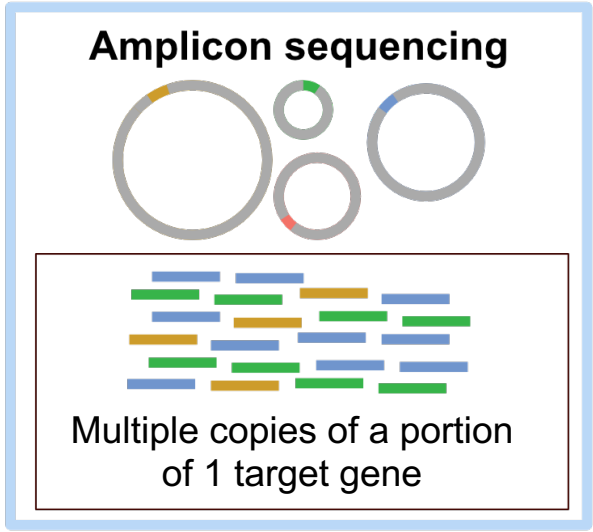


We also then try to use these recovered DNA sequences to try to get information about the microbial community present in our sample

With metagenomics sequencing, in addition to getting insight into which microbes are present, we will often also try to understand “what” they might be doing functionally

Amplicon Sequencing Overview: Select the Target Region

Amplicon sequencing involves the use of “primers” – short sequences that match highly conserved stretches of DNA



These primers most often target a gene that is universally present across all known life, while also being useful for delineating between different organisms. These target genes are often referred to as “marker-genes”.

The most common marker-gene target is part of the “ribosome”, specifically the small subunit ribosomal RNA gene (SSU rRNA), known as the “16S” in bacteria and archaea and the “18S” in eukarya

The “ribosome” is an assemblage of multiple proteins and RNA molecules, and it is essential for protein synthesis in all known bacteria, archaea, and eukarya

Because many molecules of proteins and RNA (including the SSU rRNA) need to interact together to function properly as a full ribosome, there is a lot of evolutionary pressure functionally constraining them, which results in the DNA that encodes for these molecules having relatively conserved sequences

The SSU rRNA itself is particularly suited for amplicon analysis because it has both:

1. regions that are highly conserved sequence-wise (making them good targets for the primers to find them and amplify them)
2. regions that are highly variable sequence-wise (making them good for delineating between different organisms)

Amplicon Sequencing Overview: Amplifying the Target Region

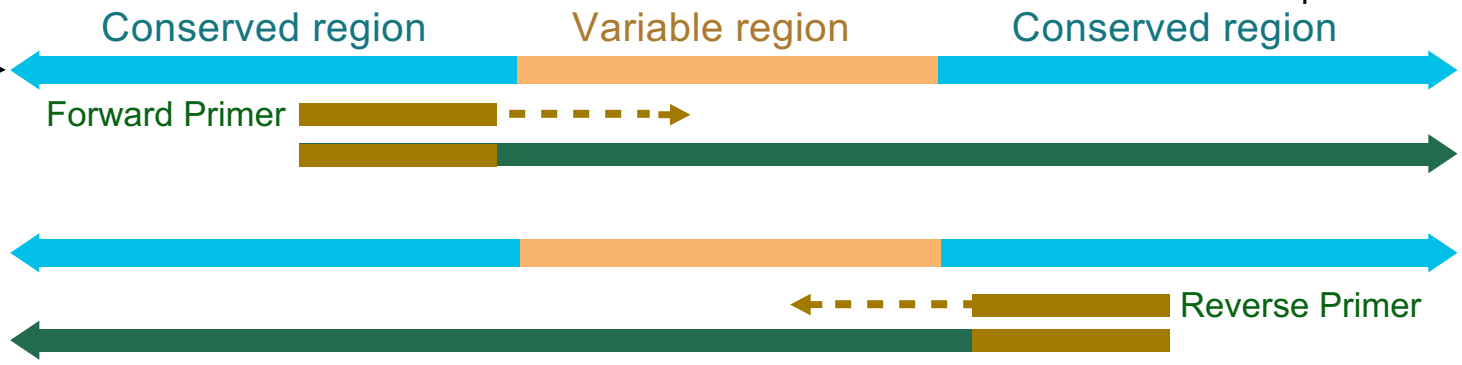
A typical example of amplifying an SSU rRNA gene via PCR:

This represents a portion of the common "16S" marker-gene

It starts out as part of a longer portion of DNA that came from an organism's genome

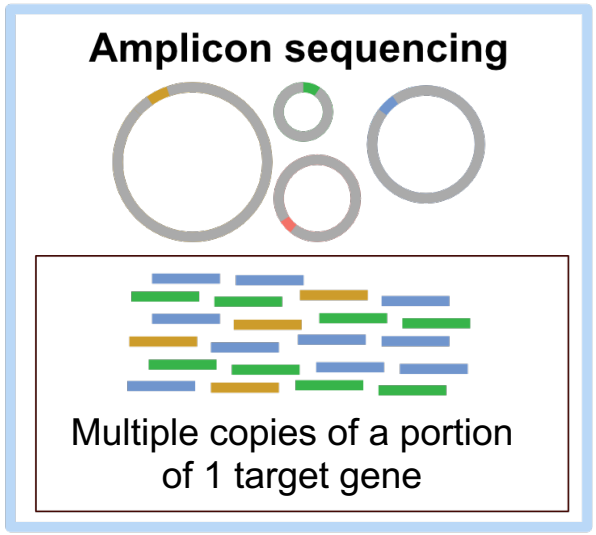
The primers we put in target the region we want to amplify

1st round of amplification



At the start of the amplification process, the forward primers will bind to some of the original molecules of DNA, enabling DNA polymerase to start building a copy

Simultaneously, the reverse primers will do the same with different original molecules of DNA



In the next round of amplification, there are going to be many copies that were initiated by either the forward or reverse primer

One initiated by the forward primer will start specifically where that primer starts

2nd round of amplification

This represents our target "amplicon"

When a copy like this is then amplified via the "other" primer, the reverse primer in this case, we will start getting copies of just our target "amplicon"

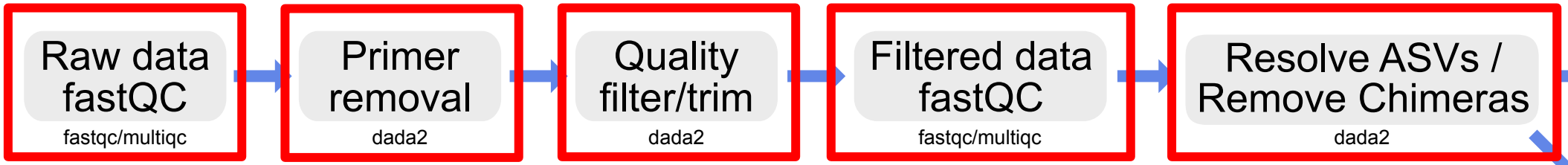
After multiple rounds of amplification, there will be many copies of just our target amplicon, which is what we then actually sequence

They are all colored the same here, but the initial DNA molecules that were first amplified came from different initial biological sequences, so there are actually many copies of different initial DNA sequences here

After multiple rounds of amplification

GeneLab Amplicon (Illumina) Sequencing Data Processing Pipeline

GeneLab Amplicon (Illumina) Sequencing Data Processing Pipeline



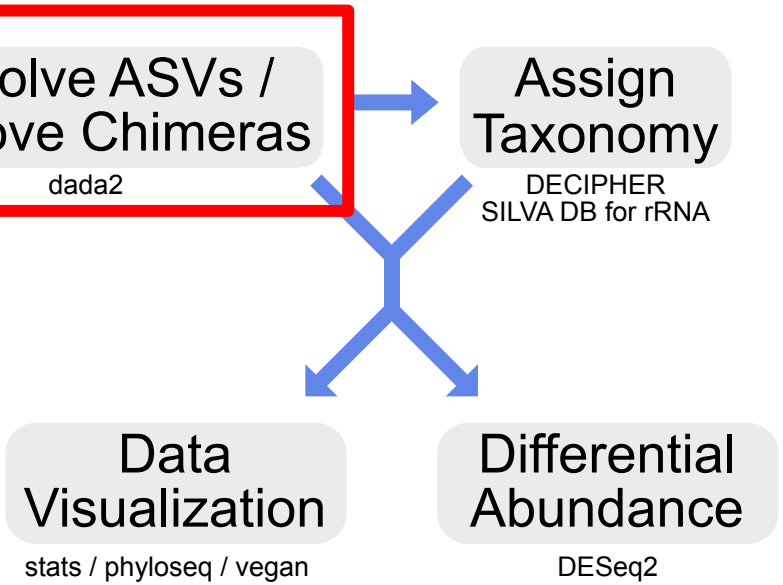
Primary outputs:
- Recovered ASV sequences (fasta)
- Counts of sequences per sample
- Taxonomic classifications of sequences

Data Visualizations:
- Beta and alpha diversity plots
- Taxonomic bar plots
- Differential abundance analysis and volcano plots

```
>ASV_1
AGTTTGATCA...
>ASV_2
TTTATGGAGA...
...
```

	Sample_A	Sample_B	...
ASV_1	975	537	...
ASV_2	40	480	...
...	...	...	...

	domain	phylum	...
ASV_1	Bacteria	Proteobacteria	...
ASV_2	Bacteria	Firmicutes	...
...	...	...	...



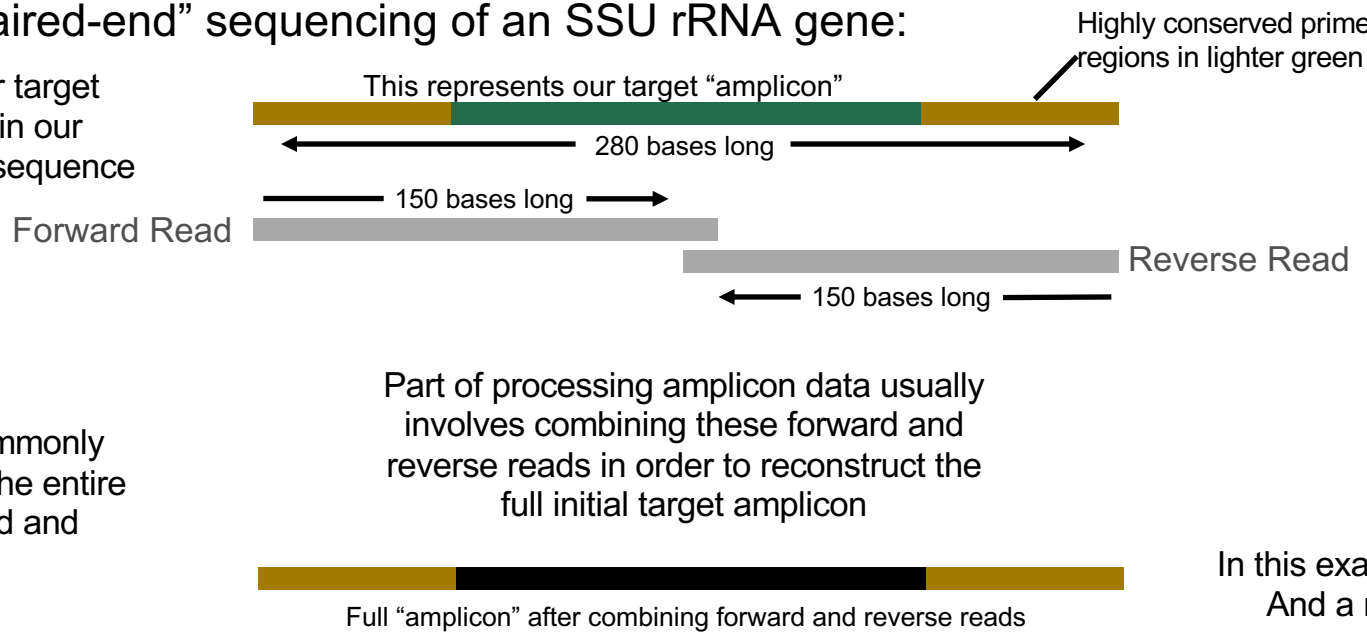
Processing AmpSeq Data: Resolve and Quantify ASVs

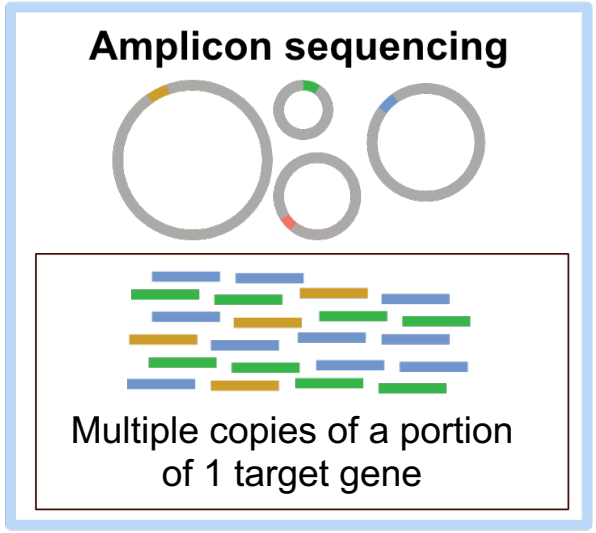
A typical example of “paired-end” sequencing of an SSU rRNA gene:

It’s all of these copies of our target amplicon, from organisms in our original sample, that we then sequence

Often, the target amplicon is longer than the “reads” the sequencer is generating

“Paired-end” sequencing is commonly employed to enable us to span the entire target amplicon with a forward and reverse read





In this example, we have 2x150 reads = 300 bases
And a nominal target amplicon length of 280
300 – 280 = 20
So we have ~20 bases of expected overlap
(there is no “one” cutoff, but a minimum of 8 is reasonable)

In order to be able to combine our reads, we need to be sure that together they span the full length of the target amplicon

Following primer-region removal, these recovered unique amplicon sequences (known as **A**mplicon **S**equence **V**ariants, ASVs) would typically then be counted to see how many times each unique sequence appears in each sample, creating a table such as this:

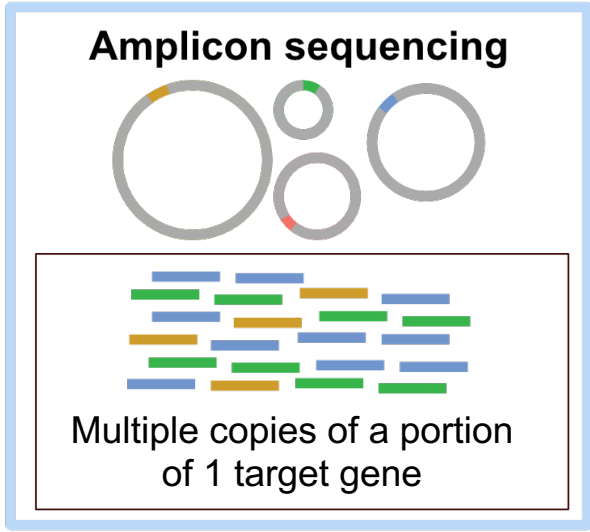
	Sample_A	Sample_B	…
Seq_1	975	537	…
Seq_2	40	480	…
…	…	…	…

These counts we get from processing amplicon data represent *counts of recovered gene-copies*.

They do not represent counts of organisms or counts of genomes.

What Are Chimeras?

"Chimeras" in the sequencing world are artificial sequences that are created during the amplification process when 2 or more biological sequences join together



As discussed earlier, a primer would be what enables these sequences to be amplified via PCR:

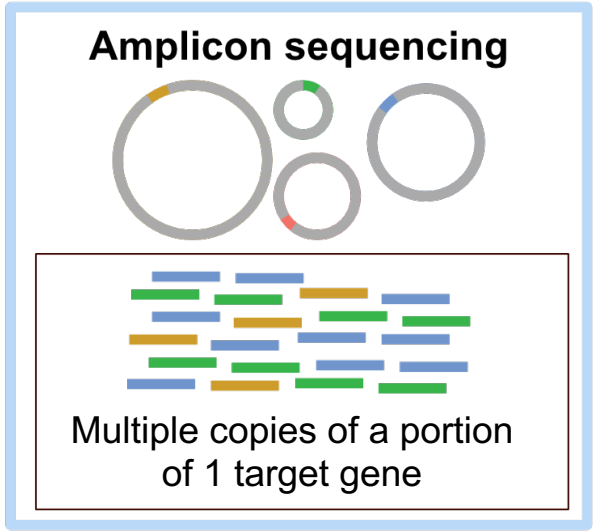
But sometime that amplification process can prematurely terminate:

That partial sequence may then act as a "primer" by attaching to a different sequence, even with some mismatches between them, and then being extended based on this different sequence:

Ultimately yielding a chimeric sequence, a mixture of both, that then continues to be amplified:

Processing AmpSeq Data: Remove Chimeras

"Chimeras" in the sequencing world are artificial sequences that are created during the amplification process when 2 or more biological sequences join together



Chimeras are often very common in amplicon datasets in terms of the *unique* sequences recovered, but not usually in terms of the *total* sequences recovered.

Remember our example count table from earlier:

	Sample_A	Sample_B	…
Seq_1	975	537	…
Seq_2	40	480	…
…	…	…	…

The total number of *unique* sequences would be however many rows are in this table (2 showing)

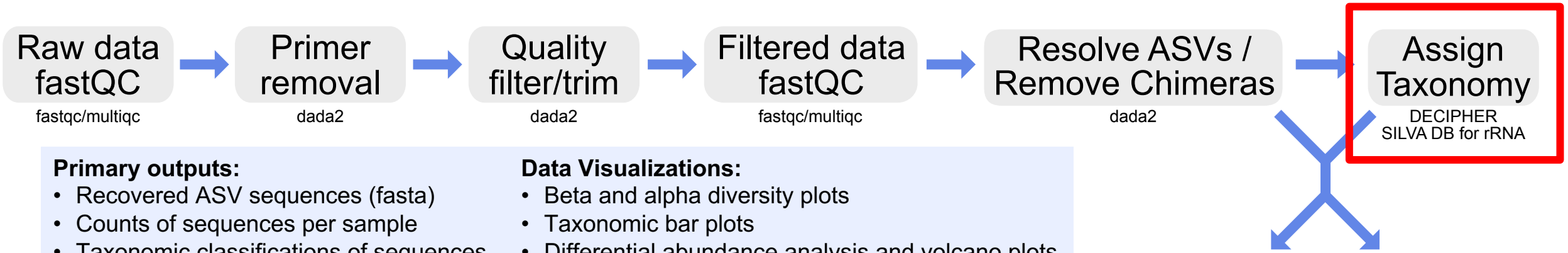
The total count of *all* sequences recovered would be the sum of all the values in the cells (2,032 showing)

Chimeras may frequently make up a large proportion of *unique* sequences recovered, but those unique chimeric sequences are usually not seen that many times each, and therefore make up only a small fraction of the *total* sequences recovered.

It is important to perform a step that tries to identify and remove chimeras, as they otherwise can drastically, artificially increase the apparent diversity of sequences that appears to exist in a sample.

Automated methods of chimera detection typically involve checking if a sequence can be made exactly by mixing together other sequences.

GeneLab Amplicon (Illumina) Sequencing Data Processing Pipeline



Primary outputs:
- Recovered ASV sequences (fasta)
- Counts of sequences per sample
- Taxonomic classifications of sequences

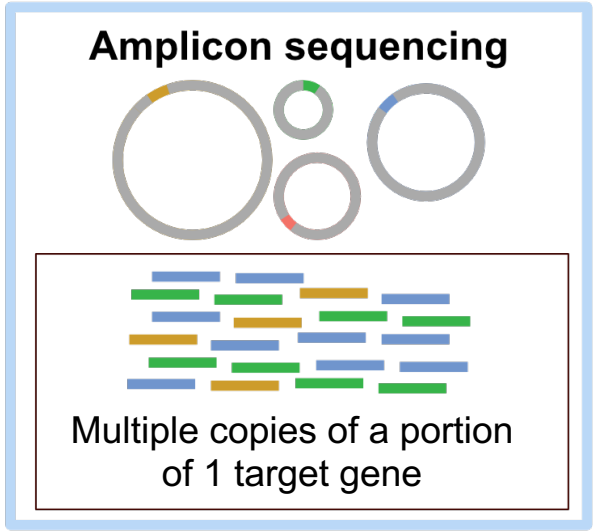
Data Visualizations:
- Beta and alpha diversity plots
- Taxonomic bar plots
- Differential abundance analysis and volcano plots

```
>ASV_1
AGTTTGATCA...
>ASV_2
TTTATGGAGA...
...
```

	Sample_A	Sample_B	...
ASV_1	975	537	...
ASV_2	40	480	...
...	...	...	...

	domain	phylum	...
ASV_1	Bacteria	Proteobacteria	...
ASV_2	Bacteria	Firmicutes	...
...	...	...	...

Processing AmpSeq Data: Assigning Taxonomy



Once we have some set of unique sequences recovered from our samples that we want to move forward with, it is common to try to assign a taxonomy to each of them, so we can get an idea of which microbes are present in our system

This is typically done by comparing our recovered sequences to a reference-sequence database in an effort to identify the taxonomy of the source organism each sequence likely came from

Like all things that use a reference database and any method, the results of a given taxonomic classification are entirely dependent on the reference database and method used

Moreover, taxonomic classification of a short portion of a single gene is difficult, and can often even be impossible assign a taxonomy at the species-level because there are cases where the portion we sequenced could be identical between different organisms

It's also worth remembering that taxonomic delineations are arbitrary; they are not fundamental units of biology that actually exist. Biological taxonomy is an ongoing process of trying to define things and attach agreed-upon labels to them.

That doesn't negate its utility, of course!
While there can be exceptions everywhere, and we don't really know the functions of some organism based on its taxonomy alone, taxonomic labels still provide a requisite means of communication and can help guide hypotheses.

GeneLab Amplicon (Illumina) Sequencing Data Processing Pipeline

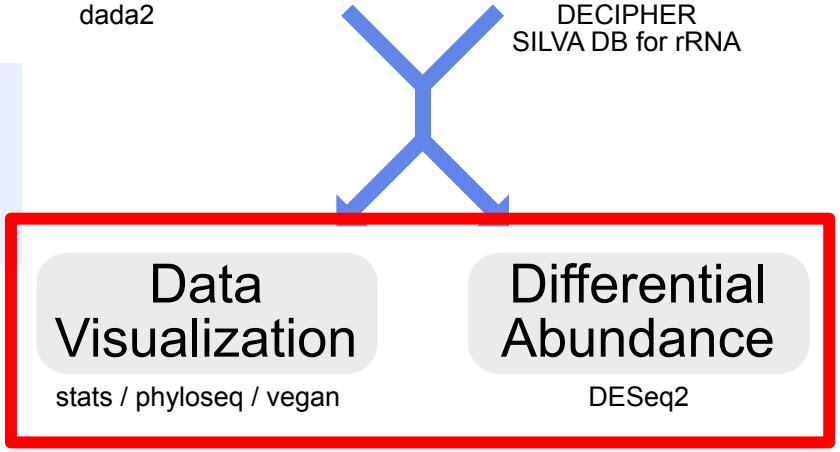
Primary outputs:
- Recovered ASV sequences (fasta)
- Counts of sequences per sample
- Taxonomic classifications of sequences

Data Visualizations:
- Beta and alpha diversity plots
- Taxonomic bar plots
- Differential abundance analysis and volcano plots

```
>ASV_1
AGTTTGATCA...
>ASV_2
TTTATGGAGA...
...
```

	Sample_A	Sample_B	...
ASV_1	975	537	...
ASV_2	40	480	...
...	...	...	...

	domain	phylum	...
ASV_1	Bacteria	Proteobacteria	...
ASV_2	Bacteria	Firmicutes	...
...	...	...	...



Processing AmpSeq Data: Data Visualization – Beta Diversity

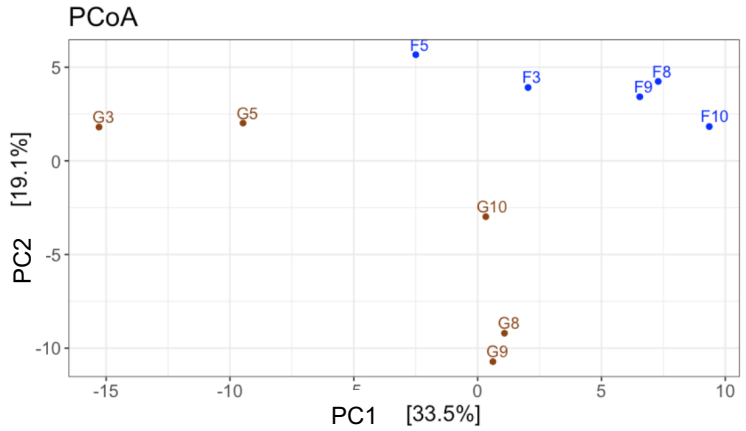
- Beta diversity involves calculating metrics such as distances or dissimilarities based on how samples relate to each other
- Important decisions to consider:
  - ➢ How the data are normalized (to account for differences in sample read-depth)
  - ➢ What is used for calculating distance/dissimilarity
  - ➢ How the data are ordinated and/or clustered

Normalization

- We use DESeq2's variance stabilizing transformation (VST) to normalize ASV counts

Ordination Method

- Ordinations provide visualizations of sample-relatedness based on dimension reduction. In our case, the dimensions are ASV counts.
- We perform **P**rinciple **Co**ordinates **A**nalysis (PCoA) on the normalized ASV counts using the *phyloseq* package

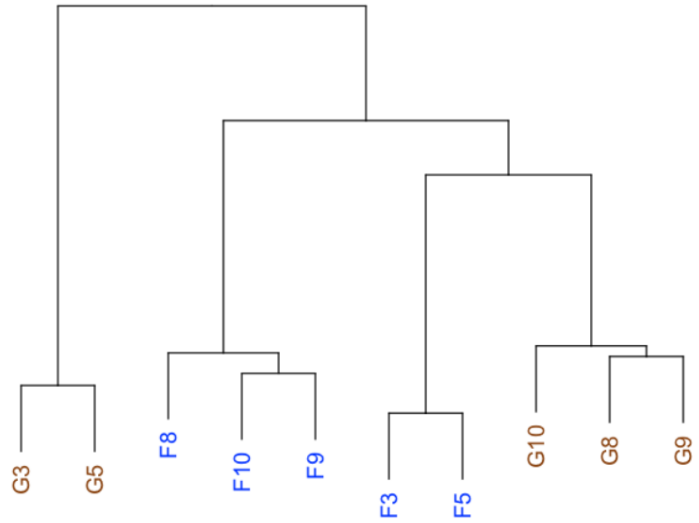


Distance/Dissimilarity Metric

- We calculate a Euclidean distance matrix of our samples using the normalized ASV counts

Clustering Method

- We perform hierarchical clustering, using Ward.D2, to group similar samples based on the Euclidean distances

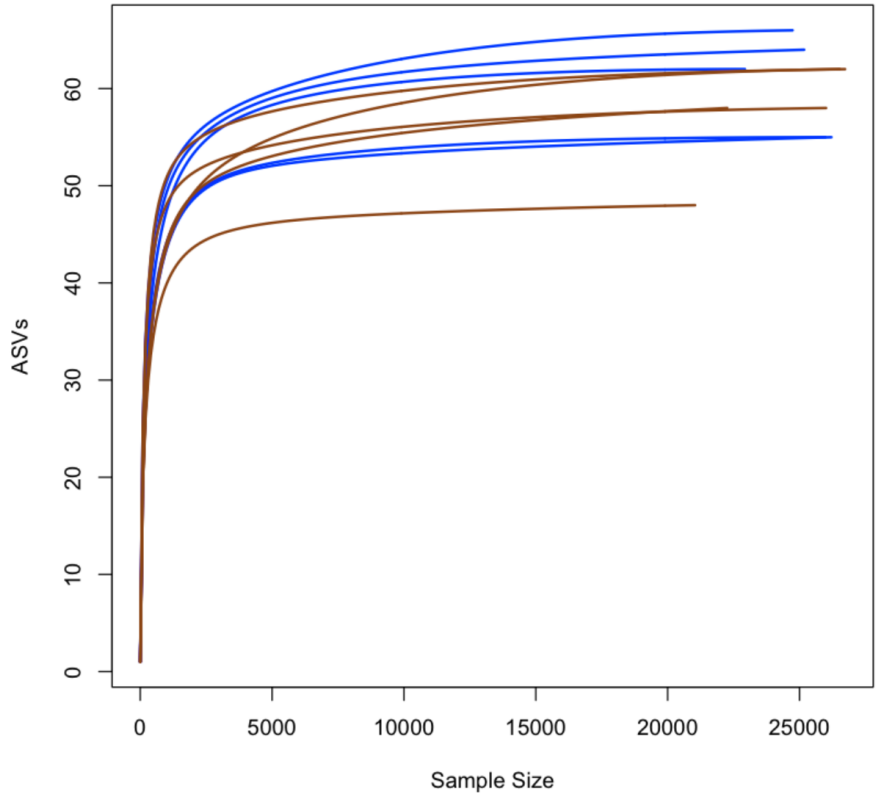


Processing AmpSeq Data: Data Visualization – Alpha Diversity

- Alpha diversity involves summary metrics that describe individual samples
  - Richness – the total number of distinct units in a sample (number of unique ASVs)
  - Evenness – how close in number each distinct unit is observed in a sample
  - Diversity – a combination of richness and evenness

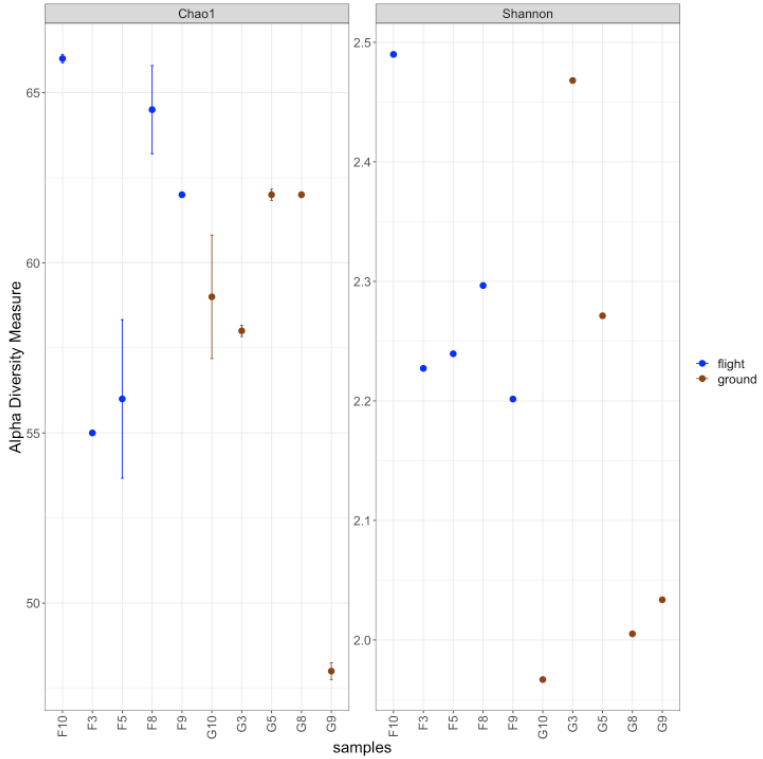
Rarefaction Curves

- A visual representation of the diversity that exists in samples, we use the rarecurve() function



Richness and Diversity Estimates

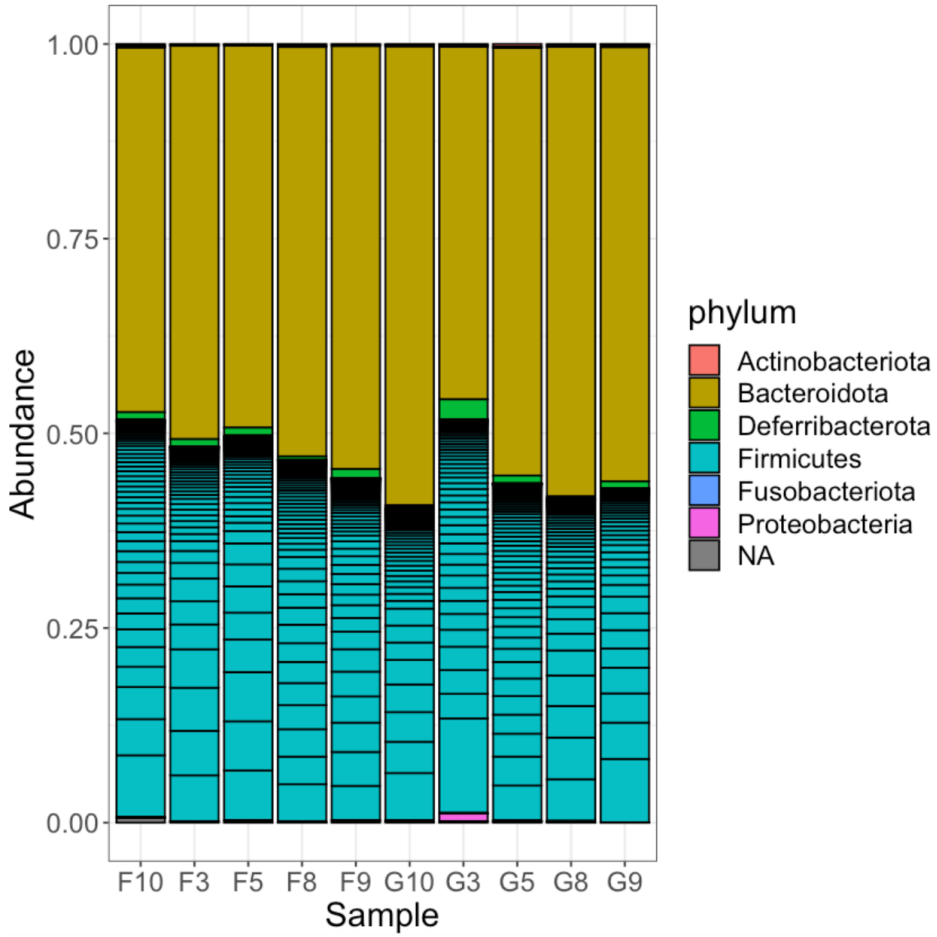
- *Chao1* – A richness estimator
- *Shannon's diversity index* – A metric of diversity that includes richness and evenness
- We use the plot_richness() function of phyloseq



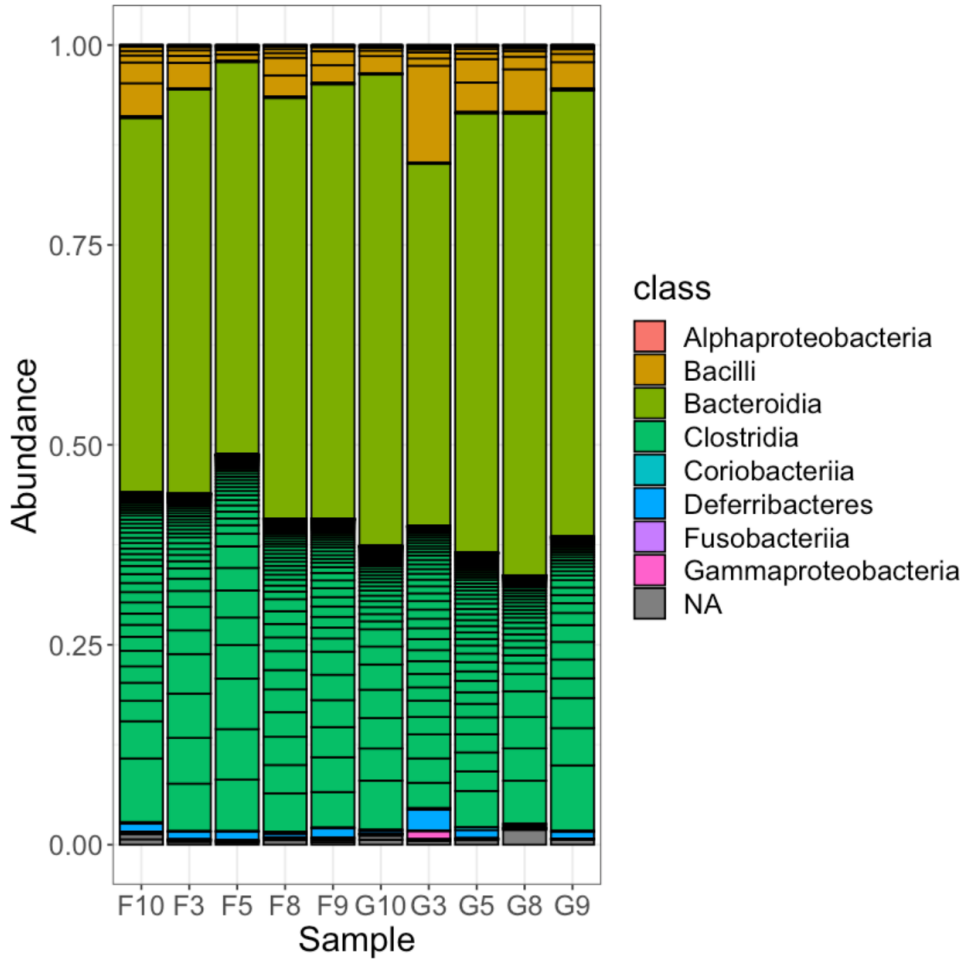
Processing AmpSeq Data: Data Visualization – Taxonomy

- Taxonomic figures can be used as a visual summary of the data
- These can be generated by normalizing all counts as a proportion of the total number of counts per sample, using the transform_sample_counts() function of the phyloseq package, then plotting those data with plot_bar()

By Phyla



By Class



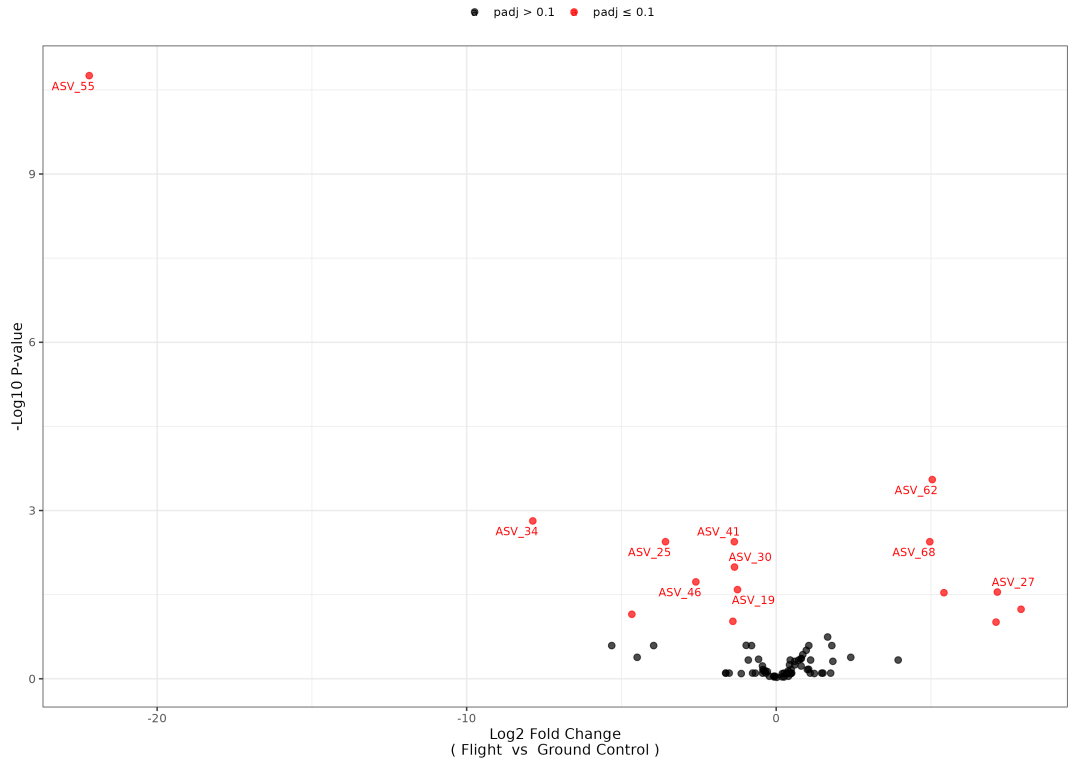
Processing AmpSeq Data: Data Visualization – Differential Abundance Analysis

- Differential abundance analysis (DAA) is used to test for differences in abundances of sequences (e.g. ASVs) or taxa recovered between groups
- We use DESeq2 to perform DAA, which runs 3 consecutive steps:
  1. normalizing for sample read-depth and composition (using the median of ratios method to estimate size factors)
  2. transforming the data (estimates dispersion and performs data normalization)
  3. testing for differential abundance between groups (using the negative binomial GLM fitting and Wald statistics by default)

DAA Table

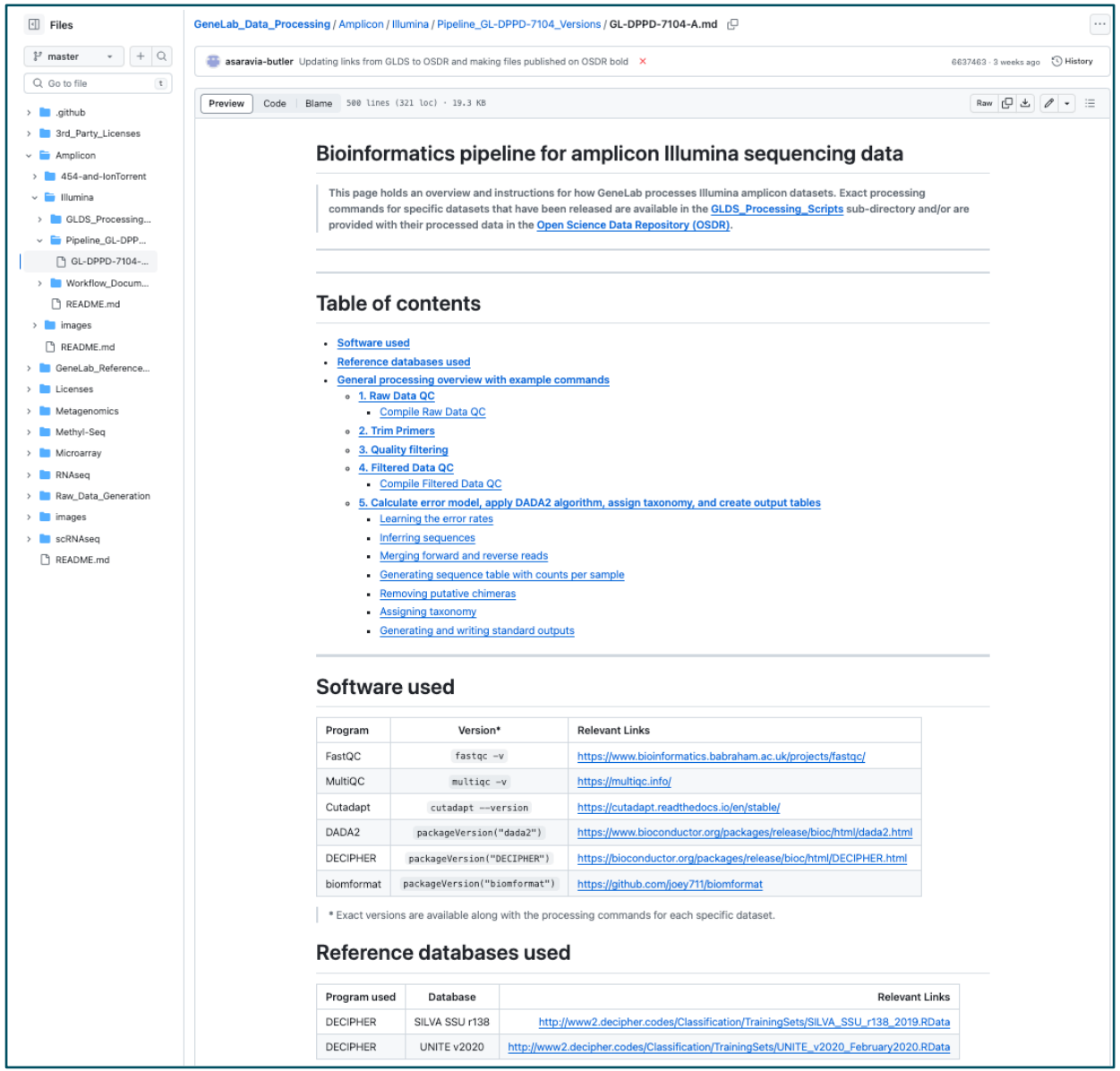
	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
ASV_3	860.738994	-1.399241	0.5890130	-2.375569	1.752192e-02	9.932675e-02
ASV_19	150.237667	-1.249008	0.4272325	-2.923485	3.461365e-03	2.769092e-02
ASV_25	114.485574	-3.569227	0.9612061	-3.713280	2.045903e-04	4.091806e-03
ASV_27	97.950896	7.150070	2.4068052	2.970772	2.970522e-03	2.640464e-02
ASV_30	86.741651	-1.342332	0.4048856	-3.315337	9.153265e-04	1.046087e-02
ASV_33	71.259216	-7.870632	1.9236449	-4.091520	4.285545e-05	1.142812e-03
ASV_41	49.164451	-1.361315	0.3811661	-3.571450	3.550109e-04	4.733479e-03
ASV_46	43.039718	-2.612064	0.8335157	-3.133791	1.725639e-03	1.725639e-02
ASV_52	30.142809	-4.543276	1.8465438	-2.460421	1.387740e-02	8.539939e-02
ASV_54	26.325446	-22.201389	3.0274634	-7.333330	2.245029e-13	1.796023e-11

Volcano Plot



GeneLab Amplicon (Illumina) Sequencing Data Processing Pipeline

GeneLab's Amplicon Seq pipeline details are available on GitHub:
https://github.com/nasa/GeneLab_Data_Processing/blob/master/Amplicon/Illumina/Pipeline_GL-DPPD-7104_Versions/GL-DPPD-7104-A.md



Files

master

Go to file

- .github
- 3rd_Party_Licenses
- Amplicon
  - 454-and-IonTorrent
  - Illumina
    - GLDS_Processing...
    - Pipeline_GL-DPP...
      - GL-DPPD-7104-...
    - Workflow_Docum...
    - README.md
  - images
  - README.md
- GeneLab_Reference...
- Licenses
- Metagenomics
- Methyl-Seq
- Microarray
- RNAseq
- Raw_Data_Generation
- images
- scRNAseq
- README.md

GeneLab_Data_Processing / Amplicon / Illumina / Pipeline_GL-DPPD-7104_Versions / GL-DPPD-7104-A.md

asaravia-butler Updating links from GLDS to OSDR and making files published on OSDR bold — 6637463 · 3 weeks ago — History

Preview | Code | Blame | 500 lines (321 loc) · 19.3 KB | Raw

Bioinformatics pipeline for amplicon Illumina sequencing data

This page holds an overview and instructions for how GeneLab processes Illumina amplicon datasets. Exact processing commands for specific datasets that have been released are available in the GLDS_Processing_Scripts sub-directory and/or are provided with their processed data in the Open Science Data Repository (OSDR).

Table of contents

Software used

Program	Version*	Relevant Links
FastQC	`fastqc -v`	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
MultiQC	`multiqc -v`	https://multiqc.info/
Cutadapt	`cutadapt --version`	https://cutadapt.readthedocs.io/en/stable/
DADA2	`packageVersion("dada2")`	https://www.bioconductor.org/packages/release/bioc/html/dada2.html
DECIPHER	`packageVersion("DECIPHER")`	https://bioconductor.org/packages/release/bioc/html/DECIPHER.html
biomformat	`packageVersion("biomformat")`	https://github.com/joey711/biomformat

* Exact versions are available along with the processing commands for each specific dataset.

Reference databases used

Program used	Database	Relevant Links
DECIPHER	SILVA SSU r138	http://www2.decipher.codes/Classification/TrainingSets/SILVA_SSU_r138_2019.RData
DECIPHER	UNITE v2020	http://www2.decipher.codes/Classification/TrainingSets/UNITE_v2020_February2020.RData

2. Trim Primers

The location and orientation of primers in the data is important to understand in deciding how to do this step. `cutadapt` has many options for primer identification and removal. They are described in detail on their documentation page here: https://cutadapt.readthedocs.io/en/stable/guide.html#adapter-types

The following example commands show how it was done for some samples of GLDS-200, which was 2x250 sequencing of the 16S gene using these primers:

- forward: 5′-GTGCCAGCMGCCGCGGTAA-3′
- reverse: 5′-GGACTACVSGGGTATCTAAT-3′

Due to the size of the target amplicon and the type of sequencing done here, both forward and reverse primers are expected to be on each of the forward and reverse reads. It therefore takes "linked" primers as input for forward and reverse reads, specified above by the `...` between them. It also expects that the primers start at the first position of the reads ("anchored"), specified with the leading `^` characters.

The following website is useful for reverse complementing primers and dealing with degenerate bases appropriately: http://arep.med.harvard.edu/labgc/adnan/projects/Utilities/revcomp.html

```
cutadapt -a ^GTGCCAGCMGCCGCGGTAA...ATTAGATACCCSBGTAGTCC -A ^GGACTACVSGGGTATCTAAT...TTACCGCGGCKGCTGGCAC \
         ## Define what B represents; and define what K represents ##
         -o Primer-trimmed-R1.fq.gz -p Primer-trimmed-R2.fq.gz Input_R1_raw.fastq.gz Input_R2_raw.fastq.gz \
         --discard-untrimmed
```

Parameter Definitions:

- `-a` – specifies the primers and orientations expected on the forward reads (when primers are linked as noted above)
- `-A` – specifies the primers and orientations expected on the reverse reads (when primers are linked as noted above)
- `-o` – specifies output of forward, primer-trimmed reads
- `-p` – specifies output of reverse, primer-trimmed reads
- `Input_R1_raw.fastq.gz` – this and following "R2" version are positional arguments specifying the forward and reverse reads, respectively, for input
- `--discard-untrimmed` – this filters out those reads? where the primers were not found as expected

Input Data:

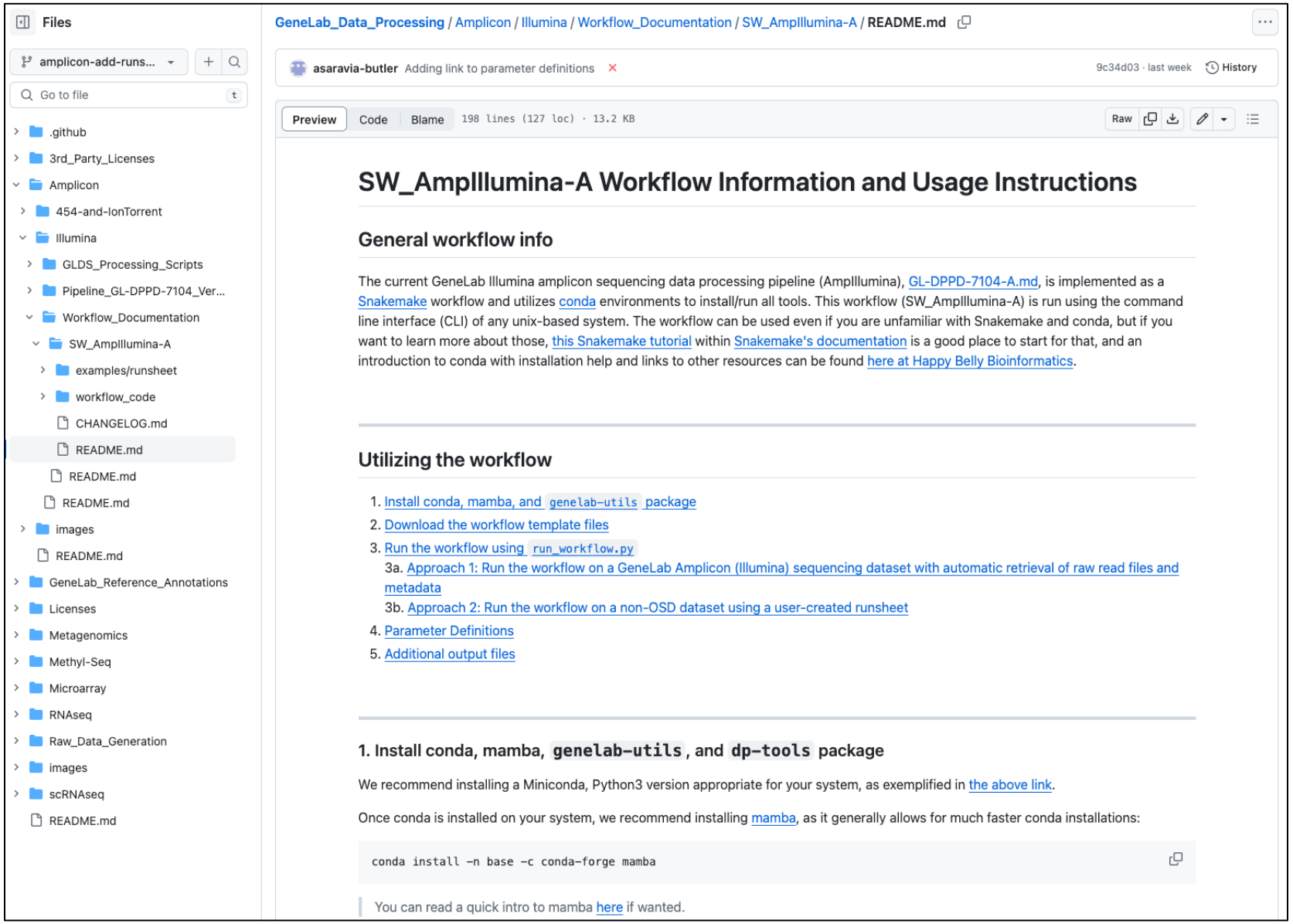
- fastq, compressed or uncompressed (original reads)

Output Data:

- **trimmed.fastq.gz**, compressed or uncompressed (trimmed reads)
- **trimmed-read-counts.tsv** (per sample read counts before and after trimming)
- **cutadapt.log** (log file of standard output and error from cutadapt)

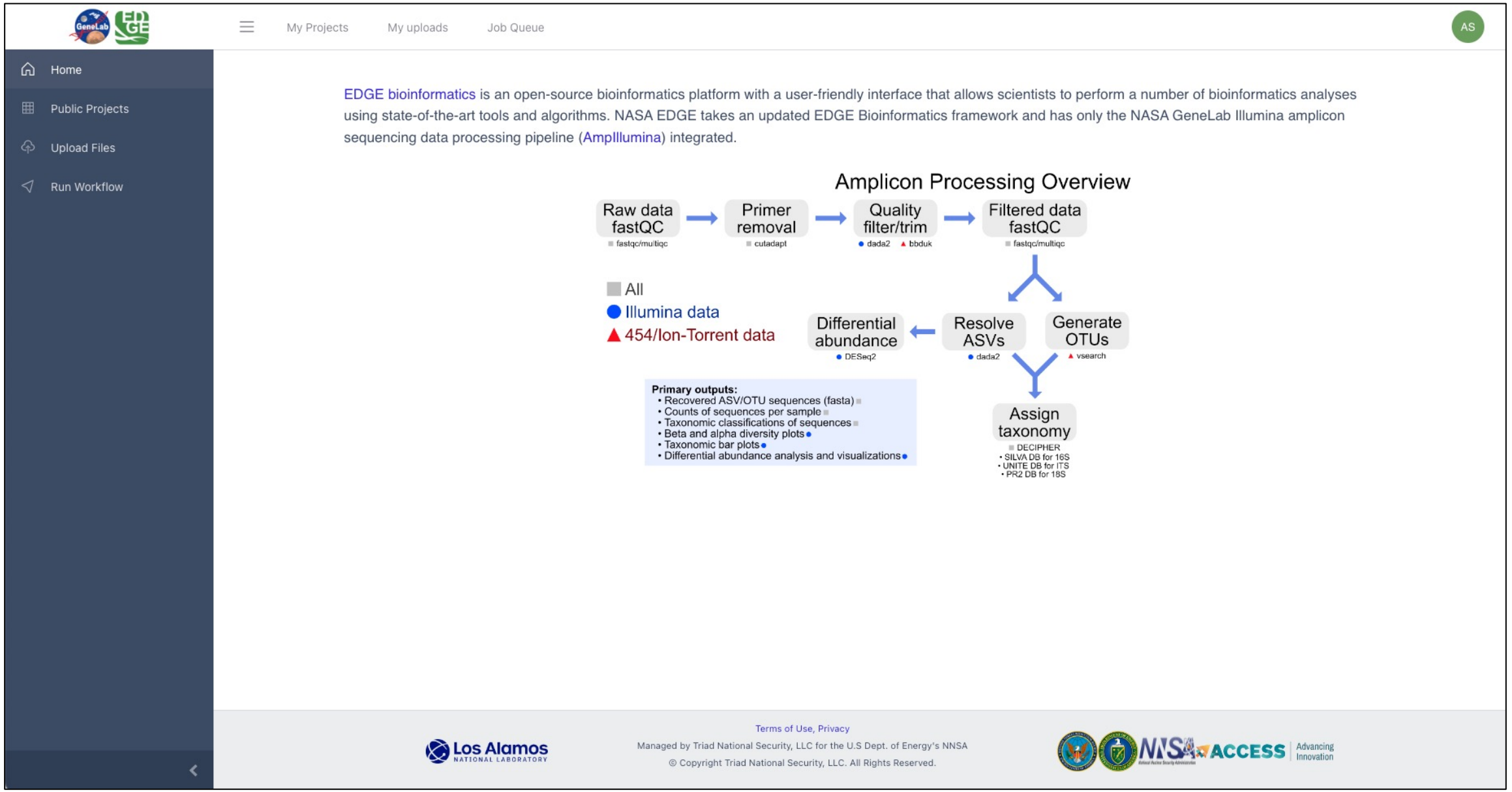
GeneLab Amplicon (Illumina) Sequencing Data Processing Workflow

GeneLab's Amplicon Seq workflow details are available on GitHub:
https://github.com/nasa/GeneLab_Data_Processing/blob/amplicon-add-runsheet-visualizations/Amplicon/Illumina/Workflow_Documentation/SW_AmpIllumina-A



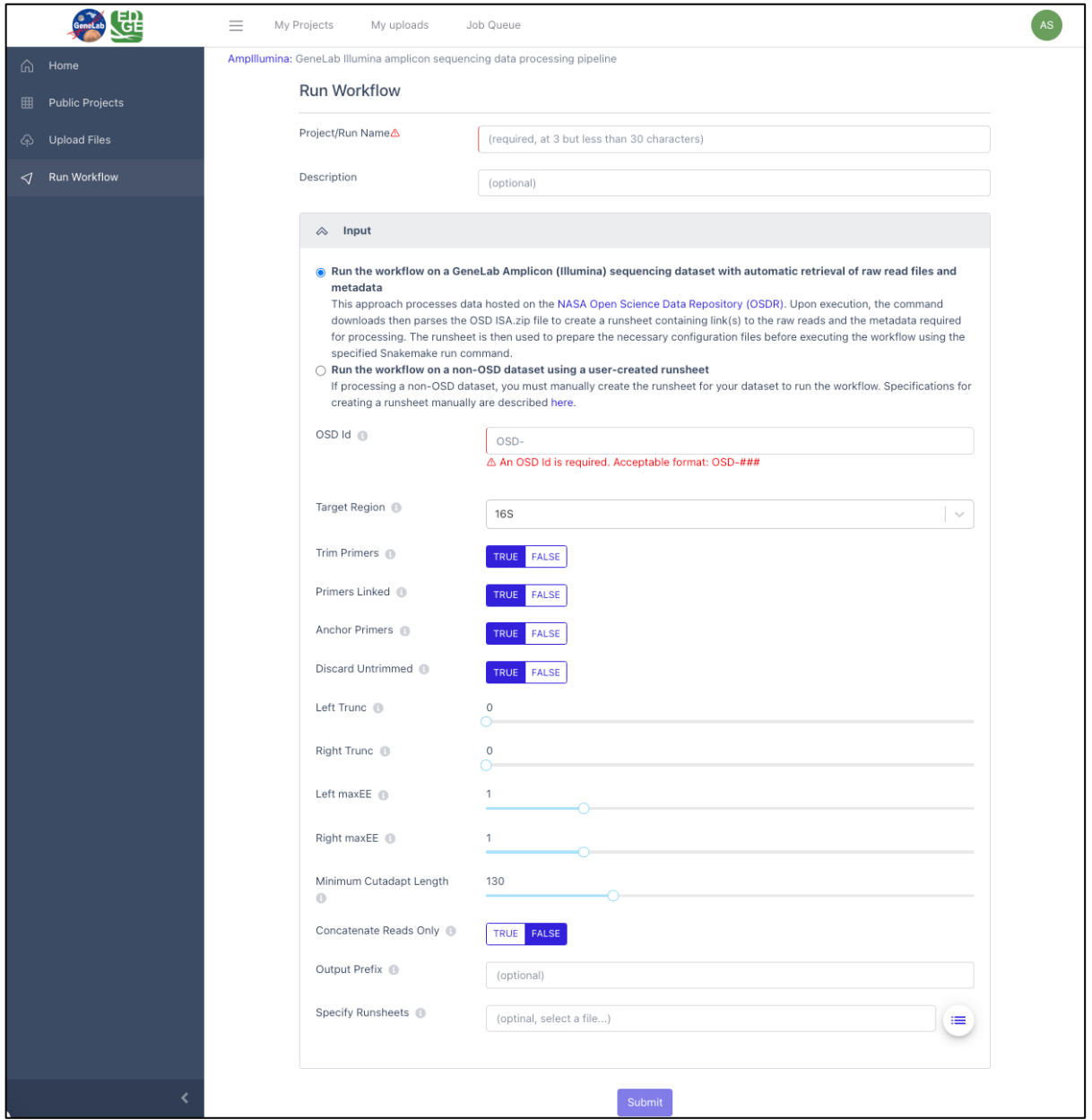
NASA EDGE GeneLab AmpSeq Workflow

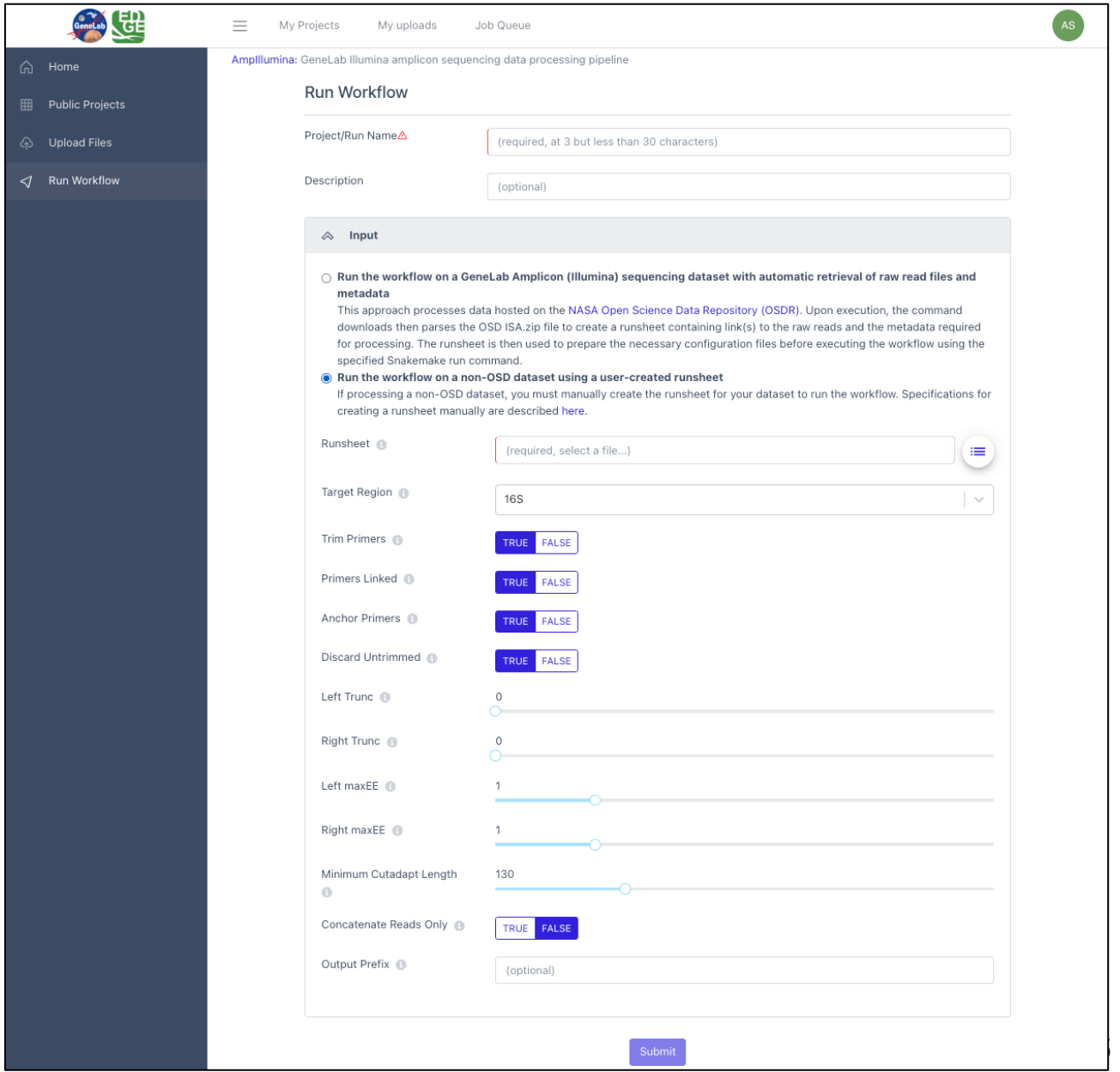
https://nasa.edgebioinformatics.org/home



NASA EDGE GeneLab AmpSeq Workflow

https://nasa.edgebioinformatics.org/amplllumina





Acknowledgements

- **NASA OSDR**
  - PM: Sylvain Costes, PhD
  - PS: Lauren Sanders, PhD
  - OSDR Team
- **NASA GeneLab**
  - PM: Sylvain Costes, PhD
  - DPM: Samrawit Gebre
  - PS/DP: Amanda Saravia-Butler, PhD
  - DP: Mike Lee, PhD
  - GeneLab Team
- **NASA ALSDA**
  - PM: Sylvain Costes, PhD
  - DPM: Danielle Lopez
  - PS: Ryan Scott
  - ALSDA Team

Funding

- GeneLab is funded by the NASA Space Biology program within the NASA Science Mission Directorate's (SMD) Biological and Physical Sciences (BPS) Division

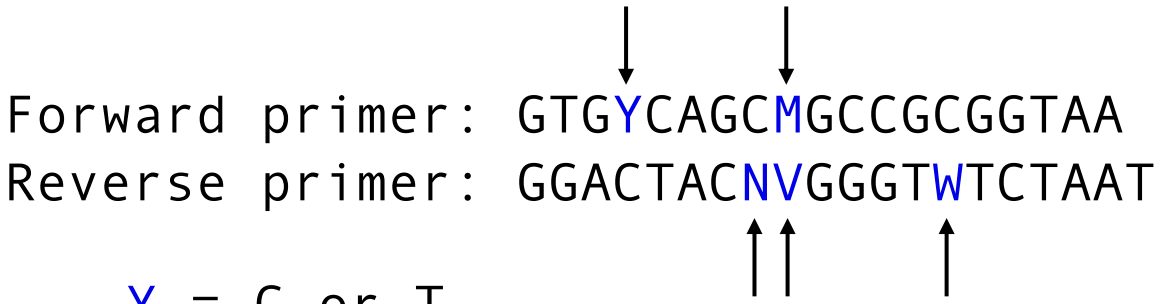
GeneLab Amplicon (Illumina) Sequencing Data Processing Pipeline EXTRA SLIDES

Why It's Important To Remove Primers

Remember from our sketch earlier that the part of the DNA our primers bind to is included in our final amplicon that we sequence

Why It's Important To Remove Primers

Primers often have "degenerate" bases,
which are positions that can be multiple bases



Y = C or T
M = A or C
N = A, T, G, or C
V = A, C, or G
W = A or T

These varied primers will bind to a biological sequence with some frequency even if they are not identical

A primer written like this:

GTGYCAGCMGCCGCGGTAA

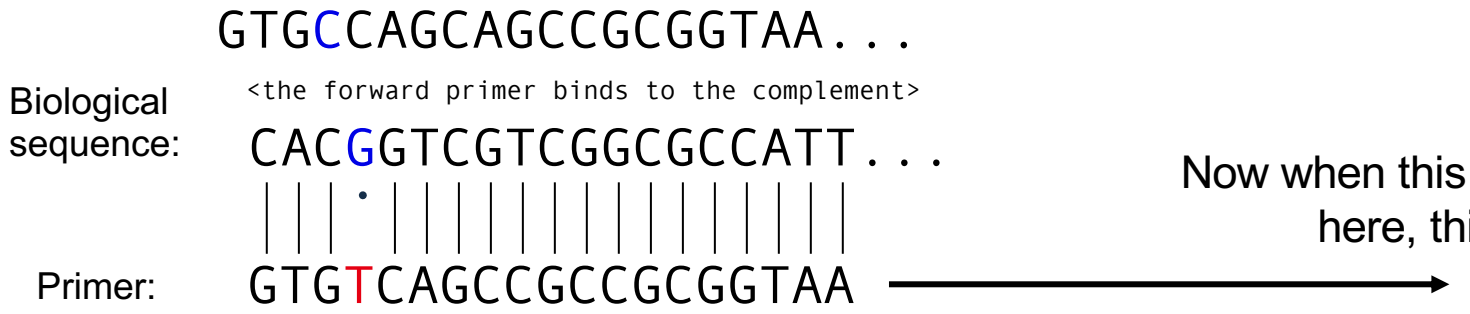
Actually means these are the sequences used during PCR amplification:

GTGCCAGCAGCCGCGGTAA

GTGCCAGCCGCCGCGGTAA

GTGTCAGCAGCCGCGGTAA

GTGTCAGCCGCCGCGGTAA



Now when this sequence is built, extending from the primer here, this molecule will have a 'T' in the 4th position, even though the biological source had a 'C'

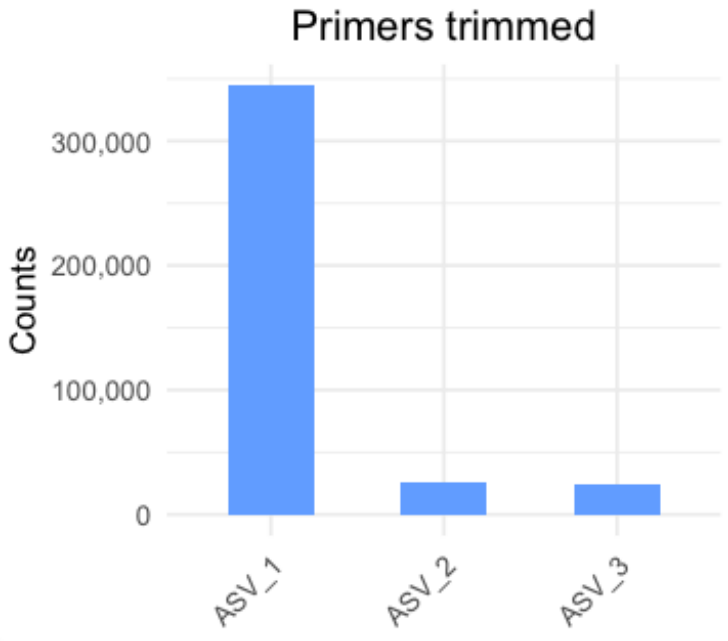
That sequence will get amplified further and show up in our data – presenting the true biological sequence between the primers, but with technically introduced variations in the primer regions.

This is why it is critical we remove the primers during processing, so that we are only working with true biological sequences.

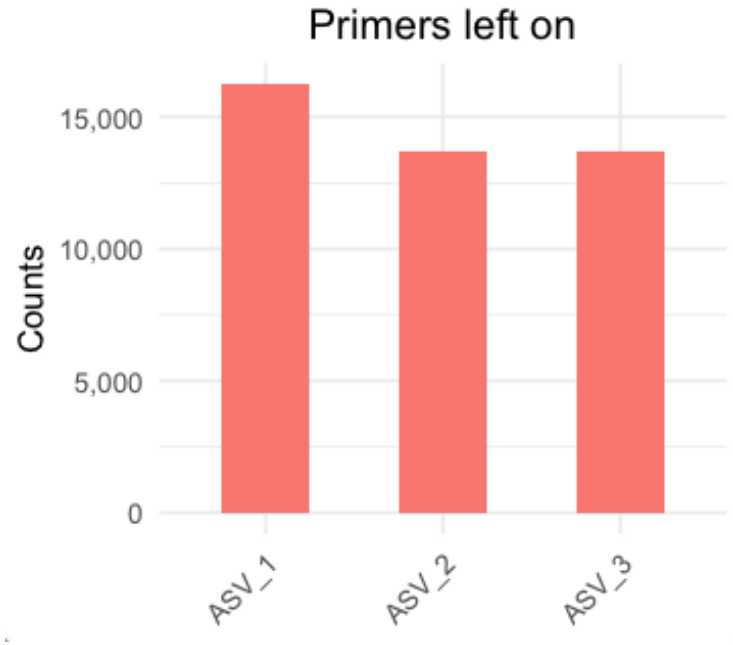
Why It's Important To Remove Primers

Example with real data

When trimming the primers off at the start of processing, the most abundant ASV (Amplicon Sequence Variant) recovered was sequenced over 300,000 times.



When processed *without* trimming the primers, the most abundant ASV recovered was sequenced just over 15,000 times.



Assigned taxonomy

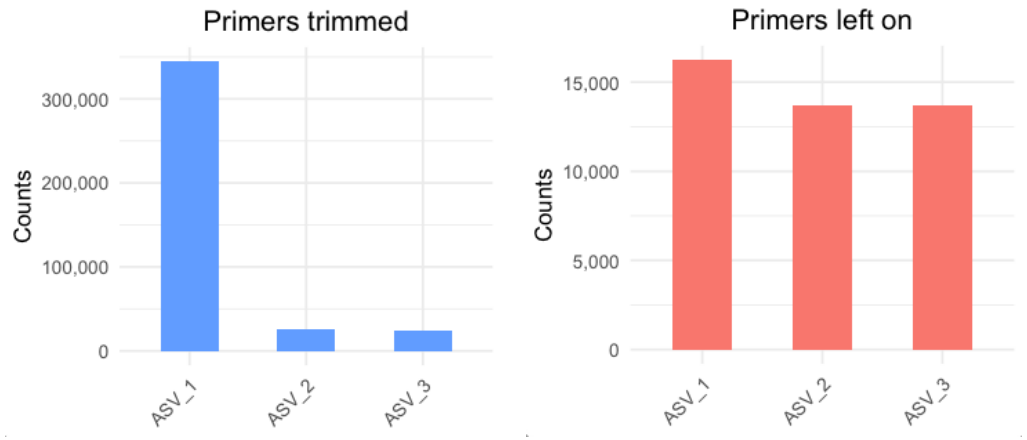
ASV_1: phylum Bacteroidota; genus *Parabacteroides*
ASV_2: phylum Firmicutes; genus *Erysipelatoclostridium*
ASV_3: phylum Firmicutes; genus *Ruminococcus*

ASV_1: phylum Bacteroidota; genus *Parabacteroides*
ASV_2: phylum Bacteroidota; genus *Parabacteroides*
ASV_3: phylum Bacteroidota; genus *Parabacteroides*

***Parabacteroides* definitely dominates the sample in either case, but nothing else shows up in the top 3 when we left the primers on.**

Why It's Important To Remove Primers

We can look at these sequences and see exactly what is going on.



This is the super-abundant ASV_1 from when the primers were trimmed:

```
>ASV_1
TACGGAGGATGCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGTGCGTAGGTGGTTAATTAAGTCAGCGGTGAAAGTTTGTGGCTCAACCATAAAATTGCCGTTGAAACTGGTTGACTTGAGTATATTTGAGGTAGGCGGAATGCGTG
GTGTAGCGGTGAAATGCATAGATATCACGCAGAACTCCGATTGCGAAGGCAGCTTACTAAACTATAACTGACACTGAAGCACGAAAGCGTGGGGATCAAACAGG
```

These are the sequences for the first 3 ASVs from when the primers were left on:

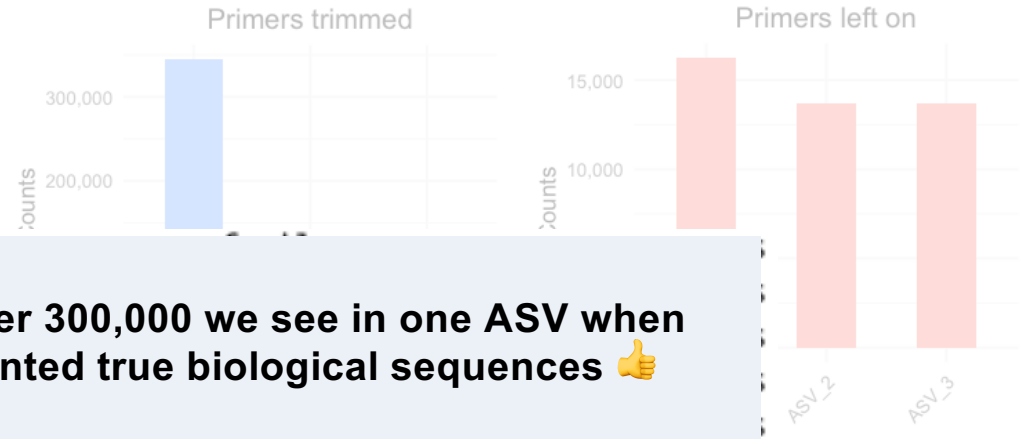
```
>ASV_1
GTGCCAGCAGCCGCGGTAATACGGAGGATGCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGTGCGTAGGTGGTTAATTAAGTCAGCGGTGAAAGTTTGTGGCTCAACCATAAAATTGCCGTTGAAACTGGTTGACTTGAGTATATTT
GAGGTAGGCGGAATGCGTGGTGTAGCGGTGAAATGCATAGATATCACGCAGAACTCCGATTGCGAAGGCAGCTTACTAAACTATAACTGACACTGAAGCACGAAAGCGTGGGGATCAAACAGGATTAGAAACCCTAGTAGTCC
>ASV_2
GTGCCAGCAGCCGCGGTAATACGGAGGATGCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGTGCGTAGGTGGTTAATTAAGTCAGCGGTGAAAGTTTGTGGCTCAACCATAAAATTGCCGTTGAAACTGGTTGACTTGAGTATATTT
GAGGTAGGCGGAATGCGTGGTGTAGCGGTGAAATGCATAGATATCACGCAGAACTCCGATTGCGAAGGCAGCTTACTAAACTATAACTGACACTGAAGCACGAAAGCGTGGGGATCAAACAGGATTAGAAACCCTGGTAGTCC
>ASV_3
GTGCCAGCCGCCGCGGTAATACGGAGGATGCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGTGCGTAGGTGGTTAATTAAGTCAGCGGTGAAAGTTTGTGGCTCAACCATAAAATTGCCGTTGAAACTGGTTGACTTGAGTATATTT
GAGGTAGGCGGAATGCGTGGTGTAGCGGTGAAATGCATAGATATCACGCAGAACTCCGATTGCGAAGGCAGCTTACTAAACTATAACTGACACTGAAGCACGAAAGCGTGGGGATCAAACAGGATTAGAAACCCTAGTAGTCC
```

The highlighted portion for all 3 is identical to the above sequence recovered when the primers were trimmed.

The only difference in each is 1 base that was degenerate in the primers.
All of these variants are technical, not biological.

Why It's Important To Remove Primers

In fact, when we failed to remove the primers, the top 36 most abundant ASVs were all the same biological sequence, just with single-nucleotide variations introduced by the degenerate primers.



If we sum these 36 variants, we get back to the over 300,000 we see in one ASV when we removed the primers, where our data represented true biological sequences 👍

If we didn't properly remove the primers, it would appear as though there were dozens more unique variants for *Parabacteroides* than there actually were.

This is why it is critical we remove the primers during processing, so that we are only working with true biological sequences.

This is the s

>ASV_1
TACGGAGGATGC
GTGTAGCGGTGA

ASV_1 Bacte
ASV_2 Bacte
ASV_3 Bacte
ASV_4 Bacte
ASV_5 Bacteria Bacteroidota Bacteroidia Bacteroidales Tannerellaceae Parabacteroides
ASV_6 Bacte
ASV_7 Bacte
ASV_8 Bacte
ASV_9 Bacte
ASV_10 Bacteria Bacteroidota Bacteroidia Bacteroidales Tannerellaceae Parabacteroides
ASV_11 Bacte
ASV_12 Bacte
ASV_13 Bacte
ASV_14 Bacte
ASV_15 Bacte
ASV_16 Bacteria Bacteroidota Bacteroidia Bacteroidales Tannerellaceae Parabacteroides

AGGCGGAATGCGTG

>ASV_1
GTGCCAGCAGCC
GAGGTAGGCGGA

ACTTGAGTATATTT
AGTAGTCC

>ASV_2
GTGCCAGCAGCC
GAGGTAGGCGGA

ACTTGAGTATATTT
GGTAGTCC

>ASV_3
GTGCCAGCCGCCGCGGTAATACGGAGGATGCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGTGCGTAGGTGGTTAATTAAGTCAGCGGTGAAAGTTTGTGGCTCAACCATAAAATTGCCGTTGAAACTGGTTGACTTGAGTATATTT
GAGGTAGGCGGAATGCGTGGTGTAGCGGTGAAATGCATAGATATCACGCAGAACTCCGATTGCGAAGGCAGCTTACTAAACTATAACTGACACTGAAGCACGAAAGCGTGGGGATCAAACAGGATTAGAAACCCTAGTAGTCC

The highlighted portion for all 3 is identical to the above sequence recovered when the primers were trimmed.

The only difference in each is 1 base that was degenerate in the primers.
All of these variants are technical, not biological.