National Aeronautics and Space Administration

# Introduction to NASA, GeneLab, PP, and the Command Line

Biological and Physical Sciences

Amanda M. Saravia-Butler, Ph.D.
GeneLab Data Processing Lead
NASA Ames Research Center
Contractor: KBR

# 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

- NASA Space Biology
  - BPS PS: Lisa Carnell
  - PM: Robert Vik
  - PS: Parag Vaishampayan, PhD

- San Jose State University
  - Philip Heller, PhD
  - Steven Boring

**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

# Introduction to the National Aeronautics and Space Administration (NASA)

# A Brief History of NACA

https://www.nasa.gov/ames/the-national-advisory-committee-for-aeronautics

# A Brief History of NASA

**1957**

Soviet Union launches Sputnik 1

**1958**

*'An Act to provide for research into the problems of flight within and outside the Earth's atmosphere, and for other purposes.'*

**1958**

NASA established and absorbs itself into NACA

US launches it's first Earth satellite

**1961-3**

**Mercury**'s single astronaut program: Can humans survive in space?

**1965-6**

**Project Gemini**: Two astronauts practice space operations

**1968-72**

**Project Apollo**: Explore the Moon

**1981**

Space Shuttle's first mission: STS-1

**1984**

Congress authorized NASA to build the Space Station

**2011**

Last Space Shuttle Mission: STS-135

# NASA's Vision

**We reach for new heights and reveal the unknown for the benefit of humankind**

# NASA's Mission

**Drive advances in science, technology, aeronautics, and space exploration to enhance knowledge, education, innovation, economic vitality and stewardship of Earth**

https://www.nasa.gov/content/the-crawlers

Cell Science Bioculture System

NASA astronaut Christina Koch

https://www.nasa.gov/cygnss/overview

# How does NASA achieve its Mission?

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

  - Human Exploration and Operations (HEOMD)

  - Science (SMD)

  - Space Technology (STMD)

https://www.nasa.gov/offices/olia/overview/index.html

# NASA Organization

Note: Administrator may delegate direct reports to Deputy Administrator at his/her discretion.
* Deputy Administrator designated as Chief Acquisition Officer (CAO).
** NOJMO oversees the Jet Propulsion Laboratory contract.
*** Programmatic reporting to the Science Mission Directorate Associate Administrator.
JPL is a Federally Funded Research and Development Center (FFRDC).

www.nasa.gov

Feb 2022

https://www.nasa.gov/about/org_index.html

# NASA Leadership

## Administrator: Bill Nelson

Sen. Bill Nelson was sworn in as the 14th NASA administrator on May 3, 2021, tasked with carrying out the Biden-Harris administration's vision for the agency. Nelson served in the U.S. Senate for 18 years from Florida and as a payload specialist on space shuttle mission 61-C in 1986. Bill Nelson's full bio: https://www.nasa.gov/feature/nasa-administrator-bill-nelson/

## Deputy Administrator: Pamela Melroy

One of only two women to command a space shuttle, Melroy logged more than 38 days in space. All three of her missions were assembly missions to build the International Space Station. After serving more than two decades in the Air Force and as a NASA astronaut, Melroy took on a number of leadership roles, including at Lockheed Martin, the Federal Aviation Administration, the Defense Advanced Research Projects Agency, Nova Systems Pty, Australia, and as an advisor to the Australian Space Agency. She currently is an independent consultant and a member of the National Space Council's Users Advisory Group. Pam Melroy's full bio: https://www.nasa.gov/sites/default/files/atoms/files/melroy_pamela.pdf

# NASA Centers and Facilities

# NASA Centers and Facilities

NASA HQ

Ames Research Center

Armstrong Flight Research Center

Glenn Research Center

Goddard Space Flight Center

Jet Propulsion Lab

Johnson Space Center

Kennedy Space Center

Langley Research Center

Marshall Space Flight Center

Stennis Space Center

White Sands Test Facility

https://history.nasa.gov/

# What has NASA done for us?

## A Workforce Innovating To Go Farther

**Space Launch System/Multi-Purpose Crew Vehicle**
NASA is innovating to continue American leadership in an ambitious space program that takes us farther than ever before. The Agency is supporting technology innovation and commercial partnerships that fuel the American economy. With the development of the newest generation of space transportation, the Space Launch System (SLS) and the Orion Multi-Purpose Crew Vehicle, NASA is developing capabilities to access deep space destinations like the Moon, an asteroid, and Mars.

This capability is being built right here in the United States. Thousands of workers in over 30 states are currently building our Nation's next generation of space travel.

**An exploration flight test of Orion will take place in 2014 with a follow-on integrated Orion/SLS uncrewed flight in 2017.**

A crewed flight of the integrated deep space system will occur as early as 2021.

## Game-Changing Innovation

**International Space Station—A Laboratory for Life Improvement Today**
The United States has an active astronaut program. In fact, American human space flight occurs daily on board the International Space Station (ISS).

NASA is committed to maintaining American excellence in science, technology, engineering, and mathematics (STEM) while sustaining a long-term human presence in space. The ISS provides important technology and innovation development, not only for future long-term human space missions, but to enhance life on Earth today.

Outstanding advances in microbial vaccine development, unprecedented cancer research involving treatment delivery, and improved water purification technologies are just some of the technological advances that research on board the ISS has provided to life here on Earth.

**Science**
NASA is committed to groundbreaking science missions today and in the future. Our missions focus on providing critical data about our home planet, unraveling the mysteries of our universe, understanding solar variability and the impacts of space weather, and exploring the diverse planetary bodies of our solar system.

**Aeronautics**
Aeronautics research is leading the Nation to less congested airways and a more environmentally friendly aviation industry.

**Investment in American Technology**
By pushing the boundaries of technology, NASA's programs keep American business and workers on the cutting edge. Investing in technology development today enables future missions of NASA while growing the economy by creating new industries, jobs, products, and services. To ensure access to the most innovative tech development, NASA partners with American companies throughout the Nation in all 50 states.

**$12.5 Billion in FY 2011 Obligated to American Industry**

## Inspiring the Next Generation

Over the past year, NASA's K–12 education projects reached more than 1 million students through STEM programs and initiatives. The Summer of Innovation Program engaged over 45,000 students and over 5,500 educators through camps and activities in 46 states, the District of Columbia, and Puerto Rico.

**$941 Million in FY 2011 Obligated to Educational Institutions Throughout the Country**

## Investing in a New Aerospace Economy

**Commercial Partners**
NASA is committed to developing a new aerospace economy. The Agency is currently partnering with American industry to achieve safe and reliable access to low-Earth orbit and the ISS.

In fact, private industry began test flights in 2010, and these will continue through 2012. This progress will culminate with human access to the ISS by 2017.

Building a strong commercial space industry enables the development of a capability that will assure the Nation's future in space. Doing so will increase national revenue while positioning American companies as competitors in the global marketplace in the space economy.

## NASA Technologies Americans Use Daily

**TV Satellite Dish**
NASA developed ways to correct errors in the signals coming from spacecraft. This technology is used to reduce noise in TV signals coming from satellites.

**Medical Imaging**
NASA developed ways to process signals from spacecraft to produce clearer images. This technology also makes photolike images of the insides of our bodies possible.

**Vision Screening System**
This technology uses techniques developed for processing space images to examine the eyes of children and find out quickly if they have any vision problems. The child doesn't have to say a word!

**Ear Thermometer**
Instead of measuring temperature using a column of mercury (which expands as it heats up), this thermometer has a lens like a camera and detects infrared energy, which we feel as heat. The warmer something is (like your body), the more infrared energy it puts out. This technology was originally developed to detect the birth of stars.

**Firefighter Equipment**
Firefighters wear suits made of fire-resistant fabric developed for use in spacesuits.

**Smoke Detector**
These devices were first used in the Earth-orbiting space station called Skylab (launched back in 1973) to help detect toxic vapors. Now they are used in most homes and other buildings to warn people of fire.

**Automobile Design Tools**
A computer program developed by NASA to analyze a spacecraft or airplane design and predict how parts will perform is now used to help design automobiles.

**Thermal Gloves and Boots**
Thermal gloves and boots have heating elements that run on rechargeable batteries worn on the inside wrist of the glove or embedded in the sole of the ski boot. This technology was adapted from a spacesuit design for the Apollo astronauts.

**Advanced Plastics**
Spacecraft and other electronics need very special, low-cost materials as the base for printed circuits (like those inside your computer). Some of these "liquid crystal polymers" have turned out to be very good, low-cost materials for making containers for food and beverages.


National Aeronautics and Space Administration
**Headquarters**
Office of Legislative and Intergovernmental Affairs
300 E Street SW
Washington, DC 20546-0001
Phone: 202-358-1055
Fax: 202-358-4340
*http://www.nasa.gov/offices/olia*

# What's in store for NASA?

1. Extend and sustain human activities across the solar system.

2. Expand scientific understanding of the Earth and the universe in which we live.

3. Create the innovative new space technologies for our exploration, science, and economic future.

4. Advance aeronautics research for societal benefit.

5. Enable program and institutional capabilities to conduct NASA's aeronautics and space activities.

6. Share NASA with the public, educators, and students to provide opportunities to participate in our mission, foster innovation and contribute to a strong national economy.

# NASA SMD Science by the Numbers

https://science.nasa.gov/about-us

# NASA Science

**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

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

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

# 5+ Decades of NASA Life Sciences Research

# Exposure to Spaceflight Changes our Physiological Systems

**The Musculoskeletal System**

Osteoporosis
Osteoarthritis
Muscle atrophy

**The Cardiovascular System**

CVD
Atherosclerosis
Anemia

**The Lymphatics**

Lymphatic pump disorders

**The Central Nervous System**

Memory loss
Retinal edema

**The Reproductive System**

Sperm motility
Infertility
Hormonal imbalances

**The Digestive System**

Liver fibrosis
NAFLD
IBS
Microbiome

Astronaut image retrieved from http://marinscienceseminar.com/space-travel-how-does-outer-space-affect-your-body/

# Exposure to Spaceflight Changes Plant Physiology Too

# Space Biology Program

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

# Spaceflight Hardware: Cell Culture Systems

**Bioculture System**

**CellCult**

# Spaceflight Hardware: Microbiology

**Biological Research in Canisters (BRIC)**

**Fig. 1** BRIC-60 and BRIC-60M with petri dish configuration.

**Fig. 2** BRIC-60 containing

**Fig. 3** BRIC-60 Canisters as flown

# Spaceflight Hardware: Small Animals

**Multi-use Variable gravity Platform (MVP)**

# Spaceflight Hardware: Plants

**Advanced Plant Habitat**

APH Flight Unit in Ground Support Equipment

Dwarf Wheat Growing in APH

# Spaceflight Hardware: Plants

**Veggie Habitat**

**Veggie Configured for Growth of Lettuce**

Veggie LED light bank.

Veggie plant pillow with quick disconnect fitting for adding water.

Veggie with 6 plant pillows contained within adjustable bellows.

Astronaut Steve Swanson harvesting lettuce grown on ISS.

Astronauts Kjell Lindgren (left) and Scott Kelly (right) enjoying freshly harvested lettuce grown on ISS.

Astronaut Steve Swanson basking in the glow of Veggie.

# Spaceflight Hardware: Rodent Research

**Animal Access Unit (AAU)**

**RR Habitat**

# Space Biology Program

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

# The Future of Space Exploration

https://www.nasa.gov/feature/nasa-looks-ahead-in-2022-artemis-x-59-lunar-exploration-and-more | https://www.youtube.com/watch?v=el7RFWBlFDU

# Introduction to GeneLab

# 2011 NRC Decadal Survey and the Sequencing Paradigm Shift

*"…**genomics, transcriptomics, proteomics, and metabolomics** offer an immense opportunity to understand the effects of spaceflight on biological systems…"*

*"…Such techniques generate considerable amounts of **data that can be mined and analyzed** for information by multiple researchers…"*

# What are ‘omics?

# Beyond the Central Dogma

Gonzalez-Pardo and Alvarez, 2013

# 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**

# GeneLab: Open Science for Life in Space (https://genelab.nasa.gov)

## Data Portal

- First multi-omics space related database
- 371 publicly available studies
- High-order data for transcriptomics and metagenomics studies
- Spaceflight environmental data associated with sample metadata (radiation dosimetry, temperature, humidity)
- Interactive **visualization portal** to explore changes in gene expression

## Tools

- User friendly submission portal to submit and publish data
- Analysis platform to tools and workflows to analyze your own data or data from the GeneLab repository.
- Tutorials and online resources to learn how to analyze RNA-Seq data
- Workspace to store, share, and organize data files

## Sequencing

- State-of-the-art sequencing facility to process spaceflight samples
- Optimized SOPs with standard processing workflow
- Provide sequencing service for NASA funded PI without any university overhead
- Generate high quality data from shared tissues for open science access

## Community

- Analysis Working Groups comprised of over 250+ scientists worldwide collaborating and analyze space omics data.
- Education Working Group focused on providing resources to educators and students to learn about space biology and bioinformatics

# GeneLab: Open Science for Life in Space (https://genelab.nasa.gov)

GeneLab Open Science for Life in Space

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

Keywords

Welcome to NASA GeneLab - the first comprehensive space-related omics database; users can upload, download, share, store, and analyze spaceflight and spaceflight-relevant data from experiments using model organisms.

**Data Repository**
Search and upload spaceflight datasets

**Analyze Data**
Perform large-scale analysis of biological omics data

**Environmental Data**
Radiation data collected during experiments conducted in space

**Collaborative Workspace**
Share, organize and store files

**Submit Data**
Have space-relevant data to submit to GeneLab?

**Visualize Data**
Interact with GeneLab processed data

# Maximizing knowledge by bringing experiments together

**Identify Shared Processes/ Molecular Signatures**

- Hypoxic Response ?
- Oxidative Stress
- Common Tissue (e.g. muscle, liver, heart, eyes, brain,…)

**For Spaceflight**

- High "n" number – statistically significant data
- Genetically identical animals
- Low resource requirements
- Short life cycle - multiple generations
- Measure response of a whole multicellular animal
- **Flies used as a model for humans for innate immunity, circadian rhythm, oxidative stress, neurobehavior, development, genetics, GWAS, "omics" studies etc**

# NASA Biological Open Science Resources

### Biospecimen Sharing Program (BSP)

- *Dissects and preserves* rodent tissues from Flight and Ground investigations
- *Coordinates* internal tissue sharing

### NASA Biological Institutional Scientific Collection (NBISC)

- *Collection* of non-human specimens and space microbial culture

### Ames Life Sciences Data Archive (ALSDA)

- *Collects and curates physiological,* mission, project and imaging data

### GeneLab (GL)

- *Collects and curates* omics data

# GeneLab Data Systems

- Open access data
- **FAIR (Findable, Accessible, Interoperable, Reusable)**
- Controlled access tools
- API - internal and external
- User Friendly Interface
- Tutorials
- Self-service Submission Portal

- Federated search – GEO, PRIDE, MG-RAST, **ALSDA**
- Database & Cloud – Scalable, easy access, fast
- Web apps
  - Data Access & Management
  - Security
  - Operation
  - Governance and Integration
- Open Source software – *no maintenance cost for software*

- Multiple data sources
  - Standard metadata organization
  - Open file formats

# GeneLab: Submission Portal – GEODE (https://genelab.nasa.gov)

# GeneLab Environment for Online Data Entry (GEODE)

| 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 |
| Sample3 | Mus musculus | X-5 | liver | Sample collection | -80 | degree Celsius | Iron-56 Ion Irradiation | 24 | hour | 2 | gray |
| Sample4 | Mus musculus | X-5 | liver | Sample collection | -80 | degree Celsius | Silicon | 12 | hour | 3 | gray |
| Sample5 | Mus musculus | X-5 | liver | Sample collection | -80 | degree Celsius | Sham-irradiated | 0 | hour | 0 | gray |

Sample level data

Terms are connected to a controlled ontology

- Space Biology PIs can now submit data through GEODE.
- Allows efficient input of sample and assay level metadata alongside data files.
- Tailored to specialized needs of spaceflight samples.
- Now exploring extending this portal for use on ALSDA submissions.

| 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 |

Assay level data with fields specific to that assay

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

# GeneLab: Data Repository (https://genelab.nasa.gov)

GeneLab Open Science for Life in Space

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

Search Data

All | GeneLab | NIH GEO | EBI PRIDE | ANL MG-RAST

**Search Filters (GeneLab Only)**

Project Type | Factors | Organisms | Assay Type | Clear

**Show Only:**

Studies With Visualizations

**Page 1 of 12 (Total Studies: 278)** Next >

**Studies Per Page:** 25

**GLDS-322**

### Comparative RNA-Seq transcriptome analyses reveal dynamic time dependent effects of 56Fe, 16O, and 28Si irradiation on the induction of murine hepatocellular carcinoma

| Organisms | Factors | Assay Types | Release Date | Description |
|---|---|---|---|---|
| Mus musculus | Ionizing Radiation<br>Time | transcription profiling | 27-Aug-2020 | One of the health risks posed to astronauts during deep space flights is exposure to high charge, high-energy (HZE) ions (Z>13) which can lead to induction of hepatocellular carcinoma (HCC). We perfor... |

**GLDS-320**

### Gamma radiation and HZE treatment of seedlings in Arabidopsis

| Organisms | Factors | Assay Types | Release Date | Description |
|---|---|---|---|---|
| Arabidopsis thaliana | Genotype<br>Time of sample collection after treatment<br>Ionizing Radiation | transcription profiling | 18-Sep-2014 | Plants exhibit a robust transcriptional response to gamma radiation which includes the induction of transcripts required for homologous recombination and the suppression of transcripts that promote ce... |

# GeneLab: Data Repository

GLDS-137

Version 4

- DESCRIPTION
- PROTOCOLS
- SAMPLES
- ASSAYS
- PUBLICATIONS
- STUDY FILES
- VISUALIZATION

Search Data

**GLDS-137:** Rodent Research-3-CASIS: Mouse liver transcriptomic, proteomic, and epigenomic data Version 4

Select a Version: 4

DOI: 10.26030/9k6w-4c28
Source Accession Number
Total Data Volume: 2.5 TB

Submitted Date: 04-Aug-2017
Release Date: 28-Aug-2017

## DESCRIPTION

**Study Description**

The Rodent Research-3 (RR-3) mission was sponsored by the pharmaceutical company Eli Lilly and Co. and the Center for the Advancement of Science in Space to study the effectiveness of a potential countermeasure for the loss of muscle and bone mass that occurs during spaceflight. Twenty BALB/c, 18-weeks old female mice (ten controls and ten treated) were flown to the ISS and housed in the Rodent Habitat for 39-42 days. Twenty mice of similar age, sex and strain were used for ground controls housed in identical hardware and matching ISS environmental conditions. Basal controls were housed in standard vivarium cages. Spaceflight, ground controls and basal groups had blood collected, then were euthanized, had one hind limb removed, and finally whole carcasses were stored at -80 C until dissection. All mice in this data set received only the control/sham injection.

**Contacts**

| NAME | ROLE | ORGANIZATION | E-MAIL |
|---|---|---|---|
| Ruth Globus | Project Scientist | NASA Ames Research Center | ruth.k.globus@nasa.gov |
| Jonathan Galazka | Project Scientist | NASA GeneLab | jonathan.m.galazka@nasa.gov |
| Rosamund Smith | Principal Investigator | Eli Lilly and Company | |

GLDS-137

Version 4

DESCRIPTION

PROTOCOLS

**Mission**

| MISSION NAME | START DATE | END DATE |
|---|---|---|
| SpaceX-8 | 08-Apr-2016 | 11-May-2016 |

**Acknowledgments**

Funding for sample processing and sequencing was provided to the GeneLab project by the NASA Space Biology Program Office, Space Life and Physical Sciences Research and Applications Division. Samples from the RR-3 experiment were provided to GeneLab through the Biospecimen Sharing Plan of the NASA Life Sciences Data Archive.

**Data Citation**

Download Citation (BibTex)

Download Citation (RIS)

Smith R, Cramer M, Globus R, Galazka J. "Rodent Research-3-CASIS: Mouse liver transcriptomic, proteomic, and epigenomic data", GeneLab, Version 4, http://doi.org/10.26030/9k6w-4c28

GLDS-137

Version 4

DESCRIPTION

PROTOCOLS

SAMPLES

ASSAYS

PUBLICATIONS

STUDY FILES

VISUALIZATION

Select Export Columns

Select columns of the samples panel below to export

Export CSV

- Source Name
- Sample Name
- Protocol REF
- Comment: Source description
- Characteristics: Organism
- Characteristics: Strain
- Comment: Animal Source
- Characteristics: sex
- Characteristics: Age at Launch
- Characteristics: Diet
- Comment: Feeding Schedule
- Parameter Value: Habitat
- Factor Value: Spaceflight
- Parameter Value: treatment
- Parameter Value: Duration
- Parameter Value: Euthanasia Chemical
- Parameter Value: Carcass Weight
- Parameter Value: Source Storage Temperature
- Comment: Initiation Date
- Comment: Completion Date
- Comment: Dissection date
- Protocol REF
- Comment: Source description
- Material Type
- Comment: Original Submitted Sample Name
- Comment: Sample Name Description
- Comment: LSDA Biospecimen Source ID
- Comment: LSDA Biospecimen Subject ID
- Parameter Value: Sample Preservation Method
- Parameter Value: Sample Storage Temperature

close select all unselect all

**SAMPLES**

| Source Name | Sample Name | Protocol REF | Comment: Source description | Characteristics: Organism | Characteristics: Strain | Comment: Animal Source |
|---|---|---|---|---|---|---|
| CB1 | Mmus_BAL-TAL_LVR_BSL_Rep1_B1 | mouse habitation | previously dissected carcass | Mus musculus | BALB/c | Taconic Biosciences |
| CB2 | Mmus_BAL-TAL_LVR_BSL_Rep2_B2 | mouse habitation | previously dissected carcass | Mus musculus | BALB/c | Taconic Biosciences |

# GeneLab: Data Repository

STUDY FILES

VISUALIZATION

## STUDY FILES

To view files, click on the folder of interest.

4 files selected

Download Selected Files

- All Files
  - RNA-Seq Data
  - Proteomics Data Files
  - Whole Genome Bisulfite Sequencing Data
  - GeneLab Processed RNA-Seq Files
    - Aligned sequence data
    - Trimmed sequence data/FastQC reports
    - Merged sequence data/FastQC reports
    - Aligned sequence data/Alignment logs
    - Raw counts data
    - Trimmed sequence data/MultiQC report
    - Differential expression analysis data
    - Normalized counts data
    - Merged sequence data/MultiQC report
    - Trimmed sequence data
    - Merged sequence data
    - Trimmed sequence data/Trimming reports
  - Study Metadata Files

| | FILES | FILE SIZE | RESOURCE CATEGORY | RESOURCE DESCRIPTION |
|---|---|---|---|---|
| ☑ | GLDS-137_rna_seq_Mmus_BAL-TAL_LVR_FLT_Rep5_F5_SJ.out.tab | 7.26 MB | Aligned sequence data | Data processed by the GeneLab Team, providing users high order data using standardized methods. |
| ☑ | GLDS-137_rna_seq_Mmus_BAL-TAL_LVR_BSL_Rep4_B4_R1_trimmed_fastqc.zip | 433.13 KB | Trimmed sequence data/FastQC reports | Data processed by the GeneLab Team, providing users high order data using standardized methods. |
| ☐ | GLDS-137_rna_seq_Mmus_BAL-TAL_LVR_GC_Rep4_G5_R2_raw_fastqc.zip | 456.51 KB | Merged sequence data/FastQC reports | Data processed by the GeneLab Team, providing users high order data using standardized methods. |
| ☑ | GLDS-137_rna_seq_Mmus_BAL-TAL_LVR_FLT_Rep2_F2_R2_raw_fastqc.html | 636.29 KB | Merged sequence data/FastQC reports | Data processed by the GeneLab Team, providing users high order data using standardized methods. |
| ☑ | GLDS-137_rna_seq_Mmus_BAL-TAL_LVR_GC_Rep3_G3_R1_trimmed_fastqc.zip | 434.84 KB | Trimmed sequence data/FastQC reports | Data processed by the GeneLab Team, providing users high order data using standardized methods. |
| ☐ | GLDS-137_rna_seq_Mmus_BAL-TAL_LVR_GC_Rep1_G1_R1_raw_fastqc.zip | 449.76 KB | Merged sequence data/FastQC reports | Data processed by the GeneLab Team, providing users high order data using standardized methods. |

# GeneLab: Data Repository

371
Studies

422
Datasets

45
Species

>10
Assays

>135TB
Data

# GeneLab: Environmental Data (https://genelab.nasa.gov)

GeneLab Open Science for Life in Space

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

Keywords

Welcome to NASA GeneLab - the first comprehensive space-related omics database; users can upload, download, share, store, and analyze spaceflight and spaceflight-relevant data from experiments using model organisms.

**Data Repository**
Search and upload spaceflight datasets

**Analyze Data**
Perform large-scale analysis of biological omics data

**Environmental Data**
Radiation data collected during experiment conducted in space

**Collaborative Workspace**
Share, organize and store files

**Submit Data**
Have space-relevant data to submit to GeneLab?

**Visualize Data**
Interact with GeneLab processed data

# GeneLab: Environmental Data

Users can view environmental data relevant to a dataset.

This includes temperature, relative humidity, CO2, radiation exposure.

Users can compare conditions of different datasets.

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

GeneLab Open Science for Life in Space

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

Keywords

Welcome to NASA GeneLab - the first comprehensive space-related omics database; users can upload, download, share, store, and analyze spaceflight and spaceflight-relevant data from experiments using model organisms.

**Data Repository**
Search and upload spaceflight datasets

**Analyze Data**
Perform large-scale analysis of biological omics data

**Environmental Data**
Radiation data collected during experiments conducted in space

**Collaborative Workspace**
Share, organize and store files

**Submit Data**
Have space-relevant data to submit to GeneLab?

**Visualize Data**
Interact with GeneLab processed data

# GeneLab: Data Visualization Portal

# GeneLab Data Democratization

# Re-use of Data and Enabling New Discoveries

**40** enabled publications (**10 publications produced by the AWGs**) using data available in GeneLab.

**Mammalian and Invertebrate Models as Complementary Tools for Gaining Mechanistic Insight on Muscle Responses to Spaceflight**

by Thomas Cahill [1], Henry Cope [2], Joseph J. Bass [3], Eliah G. Overbey [4], Rachel Gilbert [5,6], Willian Abraham da Silveira [1,7], Amber M. Paul [5,8,9], Tejaswini Mishra [10], Raúl Herranz [11], Sigrid S. Reinsch [5], Sylvain V. Costes [5], Gary Hardiman [1,12], Nathaniel J. Szewczyk [3,13] and Candice G. T. Tahimic [5,14,*]

**An Integrative Network Science and Artificial Intelligence Drug Repurposing Approach for Muscle Atrophy in Spaceflight Microgravity**

Vidya Manian*, Jairo Orozco-Sandoval and Victor Diaz-Martinez

*life*

*Article*

**Knowledge Network Embedding of Transcriptomic Data from Spaceflown Mice Uncovers Signs and Symptoms Associated with Terrestrial Diseases**

Charlotte A. Nelson [1], Ana Uriarte Acuna [2,3], Amber M. Paul [2,4], Ryan T. Scott [2,3], Atul J. Butte [5,6], Egle Cekanaviciute [2], Sergio E. Baranzini [1,5,7,*] and Sylvain V. Costes [2,*]

ELSEVIER

Computational and Structural Biotechnology Journal

Volume 19, 2021, Pages 2223-2235

**Reanalysis of the Mars500 experiment reveals common gut microbiome alterations in astronauts induced by long-duration confinement**

N.J.B. Brereton [a,1], F.E. Pitre [a,1], E. Gonzalez [b,c,1]

**Rad-Bio-App: a discovery environment for biologists to explore spaceflight-related radiation exposures**

Richard Barker, Sylvain V. Costes, Jack Miller, Samrawit G. Gebre, Jonathan Lombardino & Simon Gilroy

*npj Microgravity* **7**, Article number: 15 (2021) | Cite this article

***Cell: The biology of spaceflight package***

A coordinated package of 29 scientific papers published in five *Cell Press* journals featuring **9** papers utilizing data or resources in GeneLab. *High impact research highlights:*

- Comparative Transcriptomics Identifies Neuronal and Metabolic Adaptations to Hyper gravity and Microgravity in Caenorhabditis elegans, *iScience*
- Comprehensive Multi-omics Analysis Reveals Mitochondrial Stress as a Central Biological Hub for Spaceflight Impact, *Cell*
- NASA GeneLab RNA-seq consensus pipeline: standardized processing of short-read RNA-seq data, *iScience*

# GeneLab Power Users

*GeneLab Analysis Working Groups (AWGs) consist of **250+ scientists** from multiple space agencies, international institutions, and industry. Scientists meet monthly with each group to analyze data in the GeneLab repository.*
***We invite you to join - https://genelab.nasa.gov/awg/join!***

## ANIMAL

***68 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

***128 members***

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

## PLANTS

***58 members***

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

## MICROBES

***50 members***

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

# Analysis Working Groups (AWGs) are Driving Data Reuse

# Accelerating the pace of scientific discovery through open data

- GeneLab has established the AWGs with the goal of maximizing the generation of new knowledge from these rare and complex spaceflight datasets. With over 250+ scientists and experts worldwide the four working groups focus Animal, Plants, Microbe and Multi-omics strategies.
- Since 2018, the AWGs have published 10 peer-reviewed publication re-using data in GeneLab:

- Mammalian and Invertebrate Models as Complementary Tools for Gaining Mechanistic Insight on Muscle Responses to Spaceflight, *International Journal of Molecular Sciences*
- Immunological and hematological outcomes following protracted low dose/low dose rate ionizing radiation and simulated microgravity, *Scientific Reports*
- NASA GeneLab RNA-seq consensus pipeline: standardized processing of short-read RNA-seq data, *iScience*
- NASA GeneLab RNA-seq consensus pipeline: standardized processing of short-read RNA-seq data, *Cell System*
- Comparative Transcriptomics Identifies Neuronal and Metabolic Adaptations to Hypergravity and Microgravity in Caenorhabditis elegans, *iScience*
- Comprehensive Multi-omics Analysis Reveals Mitochondrial Stress as a Central Biological Hub for Spaceflight Impact, *Cell*
- Fundamental Biological Features of Spaceflight: Advancing the Field to Enable Deep-Space Exploration, *Cell*
- Multi-omics analysis of multiple missions to space reveal a theme of lipid dysregulation in mouse liver, *Science Reports*
- NASA GeneLab Platform Utilized for Biological Response to Space Radiation in Animal Models, *Cancers*
- Mice Exposed to Combined Chronic Low-Dose Irradiation and Modeled Microgravity Develop Long-Term Neurological Sequelae, *International Journal of Molecular Sciences*

# GeneLab Enabled Publications – 40+

genes — MDPI

Article

**Detection of Genes in *Arabidopsis thaliana* L. Responding to DNA Damage from Radiation and Other Stressors in Spaceflight**

Vidya Manian [1,2,*], Jairo Orozco-Sandoval [1] and Victor Diaz-Martinez [1]

International Journal of Molecular Sciences — MDPI

Article

**Mammalian and Invertebrate Models as Complementary Tools for Gaining Mechanistic Insight on Muscle Responses to Spaceflight**

Thomas Cahill [1], Henry Cope [2], Joseph J. Bass [3], Eliah G. Overbey [4], Rachel Gilbert [5,6], Willian Abraham da Silveira [1,7], Amber M. Paul [5,8,9], Tejaswini Mishra [10], Raúl Herranz [11], Sigrid S. Reinsch [5], Sylvain V. Costes [5], Gary Hardiman [1,12], Nathaniel J. Szewczyk [3,13] and Candice G. T. Tahimic [5,14,*]

iScience — CellPress OPEN ACCESS

Article

**Comparative Transcriptomics Identifies Neuronal and Metabolic Adaptations to Hypergravity and Microgravity in *Caenorhabditis elegans***

Craig R.G. Willis,[1] Nathaniel J. Szewczyk,[2,3] Sylvain V. Costes,[4] Ingrid A. Udranszky,[5] Sigrid S. Reinsch,[4] Timothy Etheridge,[1,7,*] and Catharine A. Conley[6]

npj Microgravity — www.nature.com/npjmgrav

ARTICLE OPEN

**Comparison of *Bacillus subtilis* transcriptome profiles from two separate missions to the International Space Station**

Michael D. Morrison[1], Patricia Fajardo-Cavazos[1] and Wayne L. Nicholson[1]

SCIENTIFIC REPORTS — nature research

Corrected: Author Correction

OPEN

**Multi-omics analysis of multiple missions to space reveal a theme of lipid dysregulation in mouse liver**

Afshin Beheshti[1,8*], Kaushik Chakravarty[2,8], Homer Fogle[1], Hossein Fazelinia[3], Willian A. da Silveira[4], Valery Boyko[1], San-Huei Lai Polo[1], Amanda M. Saravia-Butler[5], Gary Hardiman[4], Deanne Taylor[6], Jonathan M. Galazka[7] & Sylvain V. Costes[7*]

npj Microgravity — www.nature.com/npjmgrav

COMMENT OPEN

**Rad-Bio-App: a discovery environment for biologists to explore spaceflight-related radiation exposures**

Richard Barker[1✉], Sylvain V. Costes[2], Jack Miller[3,4], Samrawit G. Gebre[4], Jonathan Lombardino[5] and Simon Gilroy[1✉]

Cell — CellPress

Article

**Comprehensive Multi-omics Analysis Reveals Mitochondrial Stress as a Central Biological Hub for Spaceflight Impact**

Willian A. da Silveira,[1,23] Hossein Fazelinia,[2,23] Sara Brin Rosenthal,[3,23] Evagelia C. Laiakis,[4,23] Man S. Kim,[2,23]

# Cell Bundle - Highlights

**Highlights**

- Multi-omics analysis and techniques with NASA's GeneLab platform.
- Mitochondrial dysregulation driving spaceflight health risks.
- NASA Twin Study data validates mitochondrial dysfunction during space missions.

**Largest amount of astronaut data in one paper!**

**Highlights**

- Comparative transcriptomics in *C. elegans* exposed to hypergravity and spaceflight
- Bioinformatics identifies novel putative regulators of altered gravitational load
- Candidate molecules infer a close gravity > *daf-16*/FOXO > neuronal link

**Highlights**

- Data integration between GeneLab and ALSDA furthers the ability to utilize technologies such as artificial intelligence (AI) and machine learning (ML) methodologies to develop new hypotheses.
- Database standards and accessibility such as FAIR principles and APIs enables open-access science promoting collaborative efforts to interpret spaceflight effects by integrating omics and physiological data to the systems level.

**Highlights**

- Experimentation is necessary to understand how organisms respond to space
- RNAseq datasets are impacted by preservation protocols used on the ISS
- Impacts can be alleviated with improved carcass preservation protocols

- ISSOP is an international consortium of scientists
- Optimizing the conditions for scientists and the general public to derive valid hypotheses from these precious space omics data by reducing confounding factors and increasing interoperability at the global level.

https://www.cell.com/c/the-biology-of-spaceflight

# GeneLab Teams

# Integrating data beyond *omics*

**In 2021, the Ames Life Sciences Data Archive (ALSDA) transitioned to use the GeneLab Data Systems. The goal of this integration was to develop a multi-project submission portal to reduce the burden on PIs submitting data and enabling discovery of both omics and phenotypic data.**

(Images from Ronca et al., 2019 https://doi.org/10.1038/s41598-019-40789-y; Cells, Bone, RR Habitat NASA-ARC)

**Open Science Data Repositories**

- Phenotypic-Physiological-Behavioral-Imaging-Level Data
- Mission Telemetry data from spacecraft and hardware associated to experimental studies
- Federated with GeneLab (Molecular Omics-Level Data)

# ALSDA Science Community-Driven Curation Standards

- Last year, ALSDA transitioned their legacy database to a new infrastructure to comply to **FAIR and Open data.**
- **AWG** is a community of PIs and other SMEs, who contribute their expertise to establish a consensus on standards for the assay, sample, and study design metadata which must be collected from PIs to **ensure data reusability.**

ALSDA Analysis Working Group:
~100 members
Hosted 5 monthly meetings

Email to learn more and join!
arc-dl-alsda@mail.nasa.gov

(Credit: Amazing ALSDA AWG Members)

| Assay Metadata Configurations | Drafted | AWG Covered | Portal | Dataset Creation |
|---|---|---|---|---|
| Bone Microstructure (MicroCT) | ☑ | 09/08/21 | ☑ | ☑ |
| Behavior (Novel Object Recognition) | ☑ | 09/08/21 | ☑ | ☑ |
| Molecular/Cellular Imaging (Lt/Fluor Microscopy) | ☑ | 09/08/21 | ☑ | |
| Flow Cytometry (Flow Cytometry) | ☑ | 09/08/21 | ☑ | ☑ |
| Behavior (Elevated Plus Maze) | ☑ | 12/08/21 | ☑ | ☑ |
| Behavior (Open Field) | ☑ | 12/08/21 | | |
| Behavior (Gait) | ☑ | 12/08/21 | | |
| Protein Analysis (Western Blot/Immunoblot) | ☑ | 12/08/21 | ☑ | ☑ |
| Calcium Uptake (Spectrofluorometry) | ☑ | 12/08/21 | ☑ | ☑ |
| Behavior (Barnes Maze) | ☑ | 12/08/21 | | |
| Osteocellular (Histomorphometry) | ☑ | 02/16/22 | | ☑ |
| Mechanical Test (3pt-Bend) | ☑ | 02/16/22 | | |
| Intra-Ocular Pressure (Rebound Tonometry) | ☑ | | | |
| Blood Analysis (Serum Metabolites) | | | | |
| Echocardiogram (Ultasonography) | ☑ | 02/16/22 | | |
| Protein Analysis (ELISA) | ☑ | | | |
| Behavior (Radial Arm; Water Maze) | | | | |
| Behavior (Morris; Water Maze) | | | | |
| Muscle Physiology (Muscle Fiber Typing) | | | | |
| Mineral Density and Composition (DXA) | | | | |
| Neurophysiology (Electrophysiological Recordings) | | | | |
| Physiology (Blood Pressue) | | | | |
| Physiology (Pulse Oximetry) | | | | |
| Metabolites (HPLC with ECD) | | | | |
| Cardiovascular Imaging (MRI) | | | | |
| Neuroimaging (MRI) | | | | |
| Eye Microstructure (microCT) | | | | |
| Grip Force (Force Transducer) | | | | |
| Electron Microscopy (Scanning) | | | | |
| Electron Microscopy (Transmission) | | | | |
| Behavior (Spatial Novelty Y Maze) | | | | |
| Behavior (Tail Suspension Test) | | | | |
| Behavior (Startle Test) | | | | |
| Behavior (Prepulse Inhibition) | | | | |
| Behavior (Contextual Fear Conditioning) | | | | |
| Behavior (Rotarod) | | | | |
| Behavior (Water Maze) | | | | |
| Behavior (Wire Hang) | | | | |
| Behavior (Escher Staircase Track) | | | | |

With the support of ALSDA AWG:
Completed 10 New Assay metadata standards for the Submission Portal

# From multiple databases to a knowledge-based system

# Introduction to Planetary Protection

# What is JPL

- -Founded in 1930s, by a few Caltech doctoral students who were building alcohol fueled rocket motors with professor Theodore Von Karman
- -Federally funded, governed by Caltech
- -5000 employees, over 1000 students.
- -10+ active missions: Juno, Cassini, MRO, MSL, Mars 2020, Dawn, Europa Clipper, InSight.

# What is Planetary Protection?

**Planetary Protection (PP) aims to:**

- Protect other worlds from microorganisms found on Earth.
- Prevent forward and backward contamination.

**Why?**

- How can we look for life on other planets if we bring our own?
- Meet NASA Planetary Protection requirements
- Comply with international treaty obligations

# Bioburden Requirements for a Mars-Bound Biologically Sensitive Mission

- PP requirements as per NASA Procedural Requirements 8020.12D
  - $\leq 5\times10^5$ total spores at launch
  - $\leq 3\times10^5$ total spores on planned landing hardware
  - $\leq 300$ spores/$m^2$ for exposed surfaces
- For a biologically sensitive mission, each mission develops a biological contamination control campaign to include: hardware design, material selection, material processing, manufacturing, assembly, testing, launch operations, space flight operations, and surface operations.
- Microbial bioburden reduction and monitoring is conducted throughout the mission life cycle.
- Monitoring and/or verification of biological cleanliness is conducted using direct wipe and swab samples extracted and processed using a cultivation based technique (NASA standard assay) which measures heat-tolerant organisms.

# Planetary Protection Implementation Process Example

# Planetary Protection Biological Cleanliness Verification Sampling

Clean hardware (and table or bag) with an Iso-Propyl alcohol wipe prior to sample or install

Assay Swab or Wipe
(water is used as solvent)

Process swabs and wipes,
~3 hours required post-assay

Count Plates – 24h, 48h, and 72h

Jet Propulsion Laboratory
California Institute of Technology

# NASA Standard Spore Assay

- Does detect:
  - –spores
  - –viable, cultivable, aerobic
  - –32°C, 3 days
- Does not detect:
  - –non-cultivable, vegetative cells grown under various conditions
  - –NASA standard assay underestimates the total microbial burden.

# Bioburden Accounting for Mars Landed Spacecraft
# Raw Spore Counts to Calculated Bioburden

## Sum of the Means Approach

The total effective area sampled was represented by:

$$A_s = \sum_{i=1}^{n_s} a_{s_i} f_{s_i} e_{s_i}$$

The total number of spores counted was:

$$N_{tot} = \sum_{i=1}^{n_s} N_{s_i}$$

The bioburden density, B, was given by:

$$B = N_{tot} / A_s$$

The estimate of the total bioburden, N, was given by:

$$N = B\, A_0$$

| Variable | Definition |
|---|---|
| $A_0$ | the total area represented by a group or sample set, $m^2$ |
| $n_s$, $n_w$ | the total number of swabs or wipes |
| $n_{tot}$ | the total number of samples |
| $a_{si}$, $a_{wj}$ | is the area sampled by the $i^{th}$ swab and the $j^{th}$ wipe, ($m^2$) |
| $f_s$, $f_w$ | the pour fractions for swabs and wipes |
| $e_s$, $e_w$ | the recovery efficiencies for swabs and wipes |
| $A_s$ | the total effective area, ($m^2$) |
| $N_{si}$ | the number of CFU counted in the $i^{th}$ swab sample |
| $N_{wj}$ | the number of CFU counted in the $j^{th}$ wipe sample |
| $N_{tot}$ | the total number of CFU in a group, (spores/$m^2$) |
| B | |

## Poisson and Gaussian Statistics

**Examples equations utilized:**

$$\sigma = 1 \div (A_0 A_s)$$

$$B_{max} = 1/A_s + 3\sigma$$

$$\sigma = 1 \div A_s$$

$$B_{max} = 1/A_s + 3\sigma$$

$$\sigma = \sqrt{(N/A_s)}$$

$$B_{max} = N/A_s + 3\sigma$$

Jet Propulsion Laboratory
California Institute of Technology

# Bacterial Identification

- 16s rRNA is evolutionarily a highly conserved gene.
- It is used for phylogenetic classification of bacterial isolates at species level.
- 97.5% sequence similarity yields and accurate species level identification
- 95% yields an accurate genus level identification

# Overall Workflow

Surface samples of spacecraft hardware and associated surfaces are taken using sterile wipes and swabs

Hardy microorganisms are isolated via NASA standard spore assay (heat shock at 80°C for 15 minutes)

Microbial Identification via MALDI-TOF MS and 16S rRNA Sequencing



This document has been reviewed and determined not to contain export controlled technical data.

# JPL Microbial Culture Collection

- Viking (1975): 1454 strains
- Mars Pathfinder (1996): 62 strains
- Mars Odyssey (2001): 97 strains
- Mars Exploration Rovers (2003): 870 strains
- Phoenix (2007): 824 strains
- Mars Science Laboratory (2012): 1437 strains
- InSight (2018): 1000 strains
- M2020 (2020): 3685 strains

- In total, JPL microbial archive includes 9,429 **spacecraft related cultures** that span NASA's robotic exploration of Mars ranging from 1975-2017.



This document has been reviewed and determined not to contain export controlled technical data.

# What is MALDI-TOF MS

- Matrix Assisted Laser Desorption Ionization Time of Flight Mass Spectrometry
- Generates ions from large molecules while minimizing fragmentation
- Identification based on a protein profile

For required markings, please visit https://mh.jpl.nasa.gov

# Identification Techniques (MALDI-TOF MS)

## Formic Acid Tube Extraction

Inoculate two loops of biomass into 1 ml of 75% ethanol

Vortex for 15 minutes

Pellet isolate and decant ethanol

Add 100 µL 70% formic acid and vortex again for 15 minutes

Add 100 µL 100% acetonitrile and centrifuge x2 minutes

Add 1 µL supernatant to target and overlay with matrix



This document has been reviewed and determined not to contain export controlled technical data.

# MALDI-TOF Background

Goal : Rapid species- level identification of bacterial colonies

Slide provided by: Akemi Hinzer

# MALDI-TOF Scoring Strategy

| Match Type | Score Met | ≥ 3 Scores | Interpretation |
|---|---|---|---|
| Perfect ID | 2.2 | ✓ | Confident in identification |
| Redo Species | 2.2 | | Identified without confidence |
| Genus Match | 2.0 | ✓ | Confident in genus identification |
| Redo Genus | 2.0 | | Genus identified without confidence |
| Taxonomic Group | 2.0 | ✓ | Genus identified with mismatched species |
| Not Reliable | - | | All scores were below 2.0 |
| No Peaks | | | Experiment produced no peaks to match |

Slide provided by: Akemi Hinzer

# Data Analysis

- Mass spectra of unknown isolates are compared against
  - The Bruker reference database containing:
    - 8540 bacterial strains
  - The JPL in-house curated database containing:
    - 509 bacterial strains

This document has been reviewed and determined not to contain export controlled technical data.

# Preliminary Findings

Of the 16,839 samples collected, 1,686 plates produced colonies

From these positive plates, 4,232 colonies were subcultured and archived

Of the 3,101 isolates analyzed so far by

Slide provided by: Akemi Hinzer

# Preliminary Findings

| Genus | Frequency | Genus | Frequency |
|---|---|---|---|
| *aerius* | 2 | *lentus* | 2 |
| *amyloliquefaciens* | 1 | *manliponensis* | 1 |
| *anthracis* | 5 | *methylotrophicus* | 1 |
| *asahii* | 1 | *mojavensis* | 6 |
| *atrophaeus* | 7 | *nakamurai* | 4 |
| *badius* | 1 | *nealsonii* | 3 |
| *circulans* | 12 | *niabensis* | 4 |
| *clausii* | 4 | *oceani* | 10 |
| *cytotoxicus* | 1 | *paralicheniformis* | 1 |
| *deserti* | 14 | *pumilus* | 3 |
| *drentensis* | 1 | *rhizospaerae* | 6 |
| *endophyticus* | 4 | *ruris* | 1 |
| *flexus* | 10 | *siamensis* | 4 |
| *foraminis* | 7 | *simplex* | 1 |
| *galactosidilyticus* | 3 | *siralis* | 3 |
| *ginsengihumi* | 7 | *stratospericus* | 11 |
| *halosaccharovorans* | 2 | *subterraneus* | 13 |
| *herbersteinensis* | 1 | *tequilensis* | 1 |
| *horneckiae* | 1 | *thermoamylovorans* | 2 |
| *indicus* | 1 | *thuringiensis* | 1 |
| *infantis* | 9 | *toyonensis* | 8 |
| *invictus* | 1 | *vallismortis* | 2 |
| *kochii* | 1 | *velezensis* | 1 |
| *koreensis* | 2 | *vietnamensis* | 2 |
| *kyonggiensis* | 2 | *zhangzhouensis* | 2 |

Slide provided by: Akemi Hinzer

# What if there are no matches to known sequences


For required markings, please visit https://mh.jpl.nasa.gov

# Novel Species

INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY Volume 70, Issue 1

Research Article | Free

## *Bacillus glennii* sp. nov. and *Bacillus saganii* sp. nov., isolated from the vehicle assembly building at Kennedy Space Center where the Viking spacecraft were assembled

Arman Seuylemezian[1,†], Logan Ott[1,‡], Sarah Wolf[1], Jomel Fragante[1], Oscar Yip[1], Rüdiger Pukall[2], Peter Schumann[2], Parag Vaishampayan[1]

View Affiliations

First Published: 03 February 2020 | https://doi.org/10.1099/ijsem.0.003714

Info | Sections | Side by side view | PDF | Tools | Share

ABSTRACT

Two Gram-stain-positive, motile, endospore-forming, aerobic strains, d
from environmental air sampling at the vehicle assembly building at Ca
were assembled. Growth was observed at pH 7–9 (optimum, pH 9) for s
23a$^T$. Both strains displayed growth in 0–5% NaCl with an optimum at 1
Strains V44-8$^T$ and V47-23a$^T$ grew optimally at 32 °C, (15–32 °C) and 25
strains contained meso-diaminopimelic acid as the diagnostic diamino
phosphatidylethanolamine and diphosphatidylglycerol. The predominar
and iso-$C_{15:0}$. Strain V47.23a$^T$ shared its highest 16S rRNA sequence si
96.9%, and V44.8$^T$ with *Bacillus zeae* DSM-103964$^T$ at 96.6%. Based on
position inferred from 16S rRNA gene sequence analyses, the isolates
*Bacillus* that forms a separate clade when compared to close relatives.
acid identity values between strains V44-8$^T$ and DSM-103964$^T$ were 72.
62.4% and 69.1%, respectively. Based on the phenotypic, genomic and b
represent two novel species in the genus *Bacillus* for which the names
BAA-2860$^T$ =DSM 105192$^T$)), and *Bacillus saganii* sp. nov. [V47-23a$^T$ (=A

Keywords

planetary protection, *Bacillus*, Viking, Kennedy Space Center, cleanroo

## *Paenibacillus xerothermodurans* sp. nov., an extremely dry heat resistant spore forming bacterium isolated from the soil of Cape Canaveral, Florida

Navjot Kaur[1,†], Arman Seuylemezian[2,†], Prashant P. Patil[3], Prabhu Patil[3], Srinivasan Krishnamurti[1], Joseph Varelas[4], David J. Smith[5], Shanmugam Mayilraj[1], Parag Vaishampayan[2]

View Affiliations

First Published: 21 August 2018 | https://doi.org/10.1099/ijsem.0.002967

Info | Sections | Side by side view | PDF | Tools | Share

ABSTRACT

le, endospore-producing, facultative anaerobic bacterial strain, designated ATCC 27380$^T$, was
soil of Cape Canaveral, Florida, USA. Growth was observed at 20–42 °C (optimum, 37 °C), at
0) and in the presence of 0.5–3% NaCl (optimum 0.5%). The cell wall contained *meso*-
diagnostic amino acid and the isoprenoid quinone was MK-7. The polar lipids present were
hatidylethanolamine, diphosphatidylglycerol and one unknown phospholipid. The main fatty
*teiso*-$C_{15:0}$. Phylogenetic analysis based on 16S rRNA gene sequencing affiliated strain ATCC
*acillus*, and showed the highest sequence similarity to *Paenibacillus rigui* JCM 16352$^T$ (97.0
f type strains exhibited 16S rRNA gene sequence similarity values below 95.9%. The draft
l a size of 4,361,187 bases, with a G+C content of 51.0%. The average nucleotide identity and
tion values between strain ATCC 27380$^T$ and *P. rigui* JCM 16352$^T$ were 72.5% and 18.5%,
low the threshold suggested for species differentiation (96% and 70%, respectively). The
between strain ATCC 27380$^T$ and *P. rigui* JCM 16352$^T$ was 68.72%, which was above the
rcation of 65%. Based on phenotypic, genotypic and chemotaxonomic data, strain ATCC
species in the genus *Paenibacillus*, for which the name *Paenibacillusxerothermodurans* sp.
-1629$^T$=ATCC 27380$^T$) is proposed.

> Int J Syst Evol Microbiol. 2010 May;60(Pt 5):1031-1037. doi: 10.1099/ijs.0.008979-0. Epub 2009 Aug 7.

## Bacillus horneckiae sp. nov., isolated from a spacecraft-assembly clean room

Parag Vaishampayan [1], Alexander Probst [1], Srinivasan Krishnamurthi [2], Sudeshna Ghosh [1], Shariff Osman [1], Alasdair McDowall [3], Arunachalam Ruckmani [2], Shanmugam Mayilraj [2], Kasthuri Venkateswaran [1]

Affiliations + expand

PMID: 19666815 DOI: 10.1099/ijs.0.008979-0

### Abstract

Five Gram-stain-positive, motile, aerobic strains were isolated from a clean room of the Kennedy Space Center where the Phoenix spacecraft was assembled. All strains are rod-shaped, spore-forming bacteria, whose spores were resistant to UV radiation up to 1000 J m(-2). The spores were subterminally positioned and produced an external layer. A polyphasic taxonomic study including traditional biochemical tests, fatty acid analysis, cell-wall typing, lipid analyses, 16S rRNA gene sequencing and DNA-DNA hybridization studies was performed to characterize these novel strains. 16S rRNA gene sequencing and lipid analyses convincingly grouped these novel strains within the genus Bacillus as a cluster separate from already described species. The similarity of 16S rRNA gene sequences among the novel strains was >99 %, but the similarity was only about 97 % with their nearest neighbours Bacillus pocheonensis, Bacillus firmus and Bacillus bataviensis. DNA-DNA hybridization dissociation values were <24 % to the closest related type strains. The novel strains had a G+C content 35.6+/-0.5 mol% and could liquefy gelatin but did not utilize or produce acids from any of the carbon substrates tested. The major fatty acids were iso-C(15 : 0) and anteiso-C(15 : 0) and the cell-wall diamino acid was meso-diaminopimelic acid. Based on phylogenetic and phenotypic results, it is concluded that these strains represent a novel species of the genus Bacillus, for which the name Bacillus horneckiae sp. nov. is proposed. The type strain is 1P01SC(T) (=NRRL B-59162(T) =MTCC 9535(T)).

# Novel Genera

## Description of Tersicoccus phoenicis gen. nov., sp. nov. isolated from spacecraft assembly clean room environments

Parag Vaishampayan [1], Christine Moissl-Eichinger [2], Rüdiger Pukall [3], Peter Schumann [3], Cathrin Spröer [3], Angela Augustus [4], Anne Hayden Roberts [4], Greg Namba [4], Jessica Cisneros [4], Tina Salmassi [4], Kasthuri Venkateswaran [1]

Affiliations + expand

PMID: 23223813 DOI: 10.1099/ijs.0.047134-0

### Abstract

Two strains of aerobic, non-motile, Gram-reaction-positive cocci were independently isolated from geographically distinct spacecraft assembly clean room facilities (Kennedy Space Center, Florida, USA and Centre Spatial Guyanais, Kourou, French Guiana). A polyphasic study was carried out to delineate the taxonomic identity of these two isolates (1P05MA(T) and KO_PS43). The 16S rRNA gene sequences exhibited a high similarity when compared to each other (100 %) and lower than 96.7 % relatedness with Arthrobacter crystallopoietes ATCC 15481(T), Arthrobacter luteolus ATCC BAA-272(T), Arthrobacter tumbae DSM 16406(T) and Arthrobacter subterraneus DSM 17585(T). In contrast with previously described Arthrobacter species, the novel isolates maintained their coccidal morphology throughout their growth and did not exhibit the rod-coccus life cycle typically observed in nearly all Arthrobacter species, except A. agilis. The distinct taxonomic identity of the novel isolates was confirmed based on their unique cell-wall peptidoglycan type (A.11.20; Lys-Ser-Ala2) and polar lipid profile (presence of phosphatidylglycerol, diphosphatidylglycerol, phosphatidylinositol, an unknown phospholipid and two unknown glycolipids). The G+C content of the genomic DNA was 70.6 mol%. The novel strains revealed MK-9(H2) and MK-8(H2) as dominant menaquinones and exhibited fatty acid profiles consisting of major amounts of anteiso-C15 : 0 and anteiso-C17 : 0 and moderate amounts of iso-C15 : 0 discriminating them again from closely related Arthrobacter species. Based on these observations, the authors propose that strains 1P05MA(T) and KO_PS43 be assigned into a separate genus Tersicoccus gen. nov. For this new taxon, comprising strains 1P05MA(T) and KO_PS43, we propose the name Tersicoccus phoenicis gen. nov., sp. nov. (the type species of Tersicoccus), represented by the type strain Tersicoccus phoenicis 1P05MA(T) ( = NRRL B-59547(T) = DSM 30849(T)).

# Introduction to the Command Line

# What is a Terminal?

- A terminal window is a command-line interface used to interact with your computer (without graphics)
- The terminal window displays a blank text window, or “command line”, where you can type commands to communicate with your computer
- A terminal uses a specific language interpreter, or “shell”, to interpret commands typed into the command line
- Unix is a computer operating system written in the C programming language
  - By default, terminals on Linux and Mac computers run a Unix shell
  - Unix shells come in a variety of flavors including zsh, csh, tcsh, and bash, each having slightly different features and functionality
  - We will be using a Unix shell simulator, bash, during this bootcamp

# Command Syntax

- General command line syntax: command arguments
- Depending on the command used, arguments can be optional or required
  - Note: Arguments are also referred to as “options”, “parameters”, or “flags”
- All available arguments for a command can be found by looking at the command’s manual or help page, which can be opened in the terminal.
  - To view a command’s manual, we use the `man` function in Unix followed by the command of interest
    - Example: man <command_of_interest>
  - To view a command’s help page, we use the `--help` or `-h` flag for the command of interest
    - Example: <command_of_interest> --help
- Below is the manual of a common Unix command, mkdir, which was displayed in a terminal by running the command: man mkdir

```
asaravia — less • man mkdir — 192×27

MKDIR(1)                BSD General Commands Manual                MKDIR(1)

NAME
     mkdir -- make directories

SYNOPSIS
     mkdir [-pv] [-m mode] directory name ...

DESCRIPTION
     The mkdir utility creates the directories named as operands, in the order specified, using mode rwxrwxrwx (0777) as modified by the current umask(2).

     The options are as follows:

     -m mode
             Set the file permission bits of the final created directory to the specified mode.  The mode argument can be in any of the formats specified to the chmod(1)
             command.  If a symbolic mode is specified, the operation characters ``+'' and ``-'' are interpreted relative to an initial mode of ``a=rwx''.

     -p      Create intermediate directories as required.  If this option is not specified, the full path prefix of each operand must already exist.  On the other hand,
             with this option specified, no error will be reported if a directory given as an operand already exists.  Intermediate directories are created with permission
             bits of rwxrwxrwx (0777) as modified by the current umask, plus write and search permission for the owner.

     -v      Be verbose when creating directories, listing them as they are created.

     The user must have write permission in the parent directory.

:
```

# Running Unix Commands

- The SYNOPSIS section of the mkdir manual shows how the command should be run:

```
SYNOPSIS
     mkdir [-pv] [-m mode] directory name ...
```

- For the mkdir command, `mkdir` is written first, followed by the flags indicated in the brackets, and ending with the directory name.
  - Note: Anything in brackets [] is optional and any part of the command not in brackets is required.
  - For the mkdir command, which parts are optional and which are required?
- Do you notice how the different parts of the command are separated? **With spaces!**
- On the command line, spaces are special
  - Spaces are used to know how to properly separate distinct parts of a command
  - For this reason, it's not ideal for file or directory (aka folder) names to contain spaces. Thus, it's better to use dashes (-) or underscores (_) instead of spaces when naming files and directories.
    - Example: "example_test1.txt" is preferred over "example test1.txt"

# Running Unix Commands

- Before we run the mkdir command to make a new directory, we first need to decide where to make the new directory.
- On computers, files are organized into directories (aka folders).
- When you open a terminal, it opens to a particular directory on your system, referred to as your “working (aka current) directory”.
- You can see what your working directory is by executing the `pwd` command, which stands for “print working directory”
- In the example below, you’ll see the output of the `pwd` command:

```
(base) asaravia@ARLAL0121011036 ~ % pwd
/Users/asaravia
```

- What is the current (aka working) directory?
- You can list what’s in your working directory with the `ls` command as shown on the left below (note: `ls` stands for “list”) :

```
(base) asaravia@ARLAL0121011036 ~ % ls
Accellion       Documents       Music           igv
Applications    Downloads       Pictures        temp_files
COV-IRT         GLDS-121        Public
COV-IRT_tmp     Library         anaconda3
Desktop         Movies          gsea_home
```

- We can also navigate to the /Users/asaravia directory using the Finder window as shown on the right above.
  - Do you notice any similarities between the files shown in /Users/asaravia on the command line versus the Finder window?

# Running Unix Commands

- Similar to how you can navigate to a different directory (aka folder) within your current directory by double clicking on the folder icon in the Finder window, you can use the `cd` command to change directories in your terminal (note: `cd` stands for “change directory”).
- Below, we will navigate to the Desktop using the command line (left) and in Finder (right):

```
(base) asaravia@ARLAL0121011036 ~ % cd Desktop
(base) asaravia@ARLAL0121011036 Desktop % pwd
/Users/asaravia/Desktop
```

- What is the working directory now?
- How are directories distinguished on the command line?

- Let’s see what’s on the Desktop using the `ls` command in the terminal below:

```
(base) asaravia@ARLAL0121011036 Desktop % ls
GLDS-104                        Miscellaneous
Google_Drive_Overload           Sat Feb 20 22-19-12 2021
IACUC_RR-10_correspondence      annotation_DBI_packages
Job_related_CS_challanges       druva
Logyx_NASA                      openEDITS12.fmp12
Microarray_Datasets             stuff_to_transfer
```

- Compare the content on the Desktop in the terminal with what is shown in Finder

# Running Unix Commands

- Now that we've set the Desktop as our working directory, let's use the `mkdir` command to create a new folder on the Desktop.
- Recall the SYNOPSIS section of the mkdir manual, which shows how the command should be run:

```
SYNOPSIS
     mkdir [-pv] [-m mode] directory name ...
```

- In the example below, we'll make a directory titled "myDirectory" on the Desktop using the command line then view what's on the Desktop using both the command line and Finder:

```
(base) asaravia@ARLAL0121011036 Desktop % mkdir myDirectory
```

```
(base) asaravia@ARLAL0121011036 Desktop % ls
GLDS-104                        Sat Feb 20 22-19-12 2021
Google_Drive_Overload           annotation_DBI_packages
IACUC_RR-10_correspondence      druva
Job_related_CS_challanges       myDirectory
Logyx_NASA                      openEDI1512.tmp12
Microarray_Datasets             stuff_to_transfer
Miscellaneous
```

*Note*: The `mkdir` command is the equivalent of creating a new folder in Finder.

# Creating Variables

- There will be times when we need to set our working directory to a folder that is deep within a directory structure. Rather than navigating one directory at a time, like you would in Finder, you can navigate directly to your folder of interest with a single `cd` command in the terminal by writing the explicit path to that directory as shown:

```
(base) asaravia@ARLAL0121011036 ~ % pwd
/Users/asaravia
(base) asaravia@ARLAL0121011036 ~ % cd /Users/asaravia/Desktop/myDirectory
(base) asaravia@ARLAL0121011036 myDirectory % pwd
/Users/asaravia/Desktop/myDirectory
```

- If there is a location that we need to visit often, rather than writing out the explicit path every time, we can create a variable containing the explicit path. In Unix, you can do this using an equal sign to assign the path to a variable name as shown:

```
(base) asaravia@ARLAL0121011036 ~ % pwd
/Users/asaravia
(base) asaravia@ARLAL0121011036 ~ % myDir=/Users/asaravia/Desktop/myDirectory
(base) asaravia@ARLAL0121011036 ~ % echo $myDir
/Users/asaravia/Desktop/myDirectory
(base) asaravia@ARLAL0121011036 ~ % cd $myDir
(base) asaravia@ARLAL0121011036 myDirectory % pwd
/Users/asaravia/Desktop/myDirectory
```

- What command was introduced in the example above?
- What was added to the variable name when using the variable in Unix?

# Creating Variable Names

- When creating variable names there are a few import things to keep in mind:
  - Do not name your variable after a common function or a variable built-in to the coding language being used
  - Avoid using spaces and any characters other than letters, numbers, dashes, and underscores – Why?
  - Choose informative, memorable, and short variable names

- Example: You want to create a variable to store gene expression data from Drosophila melanogaster (aka fruit flies) that were grown at 10°C, 20°C, or 30°C. Which of the names below would be good variable names to use?

  1. GeneExpressionData_Drosophilamelanogaster_DifferentTemperatures
  2. Dmel_GE_dT
  3. FruitFlies_GE_dT
  4. FruitFlies_GE_10°C_ 20°C_ 30°C
  5. echo
  6. myData
  7. Dmel$GE$dT
  8. Dmelanogaster GE Data

# Some Common Unix Commands

**date**

- prints out the date and time

**head**

- prints the first 10 lines by default of a file

**pwd**

- stands for "print working directory", will print the path to your current working directory

**ls**

- stands for ”list", lists all files in the current working directory, or in the path provided as a positional argument at the end of the command

**mkdir**

- creates a new folder (aka directory)

**cd**

- stands for "change directory”, allows you to navigate to different directories on the command line

**cp**

- stands for "copy”, allows you to copy files from one location to another (or in the same location as the original)

**mv**

- stands for "move”, used to either rename a file/folder and/or move a file/folder to another location

**cat**

- stands for "concatenate”, allows you to view the entire contents of a file and combine files

# Some Common R Functions

**R command format:** In R, you pass arguments to functions by putting them between parentheses

**help()**

- provides a summary of how to use the R function provided inside the parenthesis

**read.csv()**

- allows you to read information in your .CSV file into R

**head()**

- prints the first 6 lines by default of a data object

**dim()**

- prints the dimensions of a data object loaded in R

**summary()**

- prints a summary of the data object provided inside the parenthesis

**list.files()**

- lists all the files in the location indicated in the parenthesis (if no location is indicated, it will list the files in the current or working directory)

**getwd()**

- prints the current working directory

**setwd()**

- changes the current working directory to the location provided by a file path string

**write.csv(data, 'filename.csv')**

- creates a csv file containing the data held in the variable indicated to the left of the comma with the file name indicated to the right of the comma

# Some Common Data Frame Manipulations in R

**DF + 1**

- perform an arithmetic operation to all cells of a data frame (represented by DF above)

**log(DF)**

- calculates the natural log for all cells of a data frame

**ceiling(DF)**

- rounds decimal values in each cell of a data frame up to the nearest integer

**DF['column1']**

- extracts only the indicated column from a data frame

**DF['row1',]**

- extracts only the indicated row from a data frame

**DF[ rowSums(DF)>0 , ]**

- filters the data in a data frame to only contain the rows whose sum is greater than zero (the mathematical operator and filtering value can be changed as needed)

**DF['column4'] <- c(1,2,3,4,5,6,7,8,9,10)**

- adds a new column (column 4) to the data frame with the values indicated (note: the number of values must match the number of rows in the data frame the column is added to)

**cbind(DF, DF2)**

- combines the columns in data frame, DF, with the columns in another data frame, DF2

# Jupyter Notebook

- Throughout this bootcamp, we will be running several of the commands we just discussed. However, since working directly on the command line takes time to get used to, we will be using Jupyter Notebooks to run our commands.
- Jupyter Notebook (JN) was created as part of Project Jupyter (https://jupyter.org/) and is a web application that allows you to create documents that have both narrative text and live code; meaning you can run commands through a web browser (instead of a terminal).
- Additional advantages of JNs:
  - ➢ Facilitates sharing code and results between groups
  - ➢ Ease of publishing your code
  - ➢ Allows for reproducibility

## 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 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.

You can read about `EnhancedVolcano()` and see some examples by clicking here

```
# 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',
    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, height = 8.5, dpi = 300)
```

# Compute Cluster

- When we execute commands in our JNs, they will be run on a simulated compute cluster.
- A compute cluster is a set of servers, which are like computers, that are all connected, and each server is used for a specific purpose. Below is an example of a relatively simple cluster containing 4 servers (aka nodes) – 1 head node and 3 compute nodes:

- **Head node**: The server that allows you to connect to a cluster (aka the login server). You should not run any commands that require a lot of computational resources on the head node. Why do you think that is?
- **Compute nodes**: Servers that are used to run commands, programs, and scripts (aka jobs). Compute nodes usually contain more computational resources than the head node.

# Compute Cluster

- There are often multiple people logged into a cluster at the same time, and they all want to run jobs. This could be a problem for the servers in the cluster. Why?
  - Hint: Think about what happens when you try to run a bunch of applications on your personal computer. Now imagine multiple people each running several applications on a server…
- To avoid multiple people running jobs all at the same time and risking freezing the servers, compute clusters are equipped with a job scheduler.
- Rather than executing jobs directly, users submit their jobs to the job scheduler, which will check each compute node to see if there are enough resources available to run the job. If there are, the job will begin running on one of the compute nodes; if not, the job will get queued and will not start until enough resources are available.
- During this bootcamp, we will be “submitting” all computationally expensive jobs to simulated cluster job scheduler, SLURM.

# SLURM Array Script

```
#!/bin/bash
#
#SBATCH --job-name="GLDS-104_raw_fastqc"
#SBATCH --output=raw_fastqc_out_logs/GLDS-104_raw_fastqc_%A_%a.out
#SBATCH --partition=priority
#SBATCH --mem=6000
#SBATCH --array=1-12
#
#SBATCH --mail-user=user@nasa.gov
#SBATCH --mail-type=END

. ~/.profile

echo "GLDS-104_raw_fastqc"
echo ""

start=$(date +%s)
echo "start time: $start"
echo $HOSTNAME
echo "My SLURM_ARRAY_TASK_ID: " $SLURM_ARRAY_TASK_ID

source activate /miniconda38_admin/envs/fastqc_09-2021

in_dir=/GLDS-104/00-RawData/Fastq
out_dir=/GLDS-104/00-RawData/FastQC_Reports

sample=$(cat samples.txt | sed -n ${SLURM_ARRAY_TASK_ID}p)

echo ""
echo "SAMPLE: ${sample}"

echo ""
echo "FastQC version: "
fastqc --version
echo ""

call="fastqc -o $out_dir $in_dir/${sample}_R*_raw.fastq.gz"

echo $call
echo ""
eval $call


echo ""
end=$(date +%s)
echo "end time: $end"
runtime_s=$(echo $(( end - start )))
echo "total run time(s): $runtime_s"
sec_per_min=60
sec_per_hr=3600
runtime_m=$(echo "scale=2; $runtime_s / $sec_per_min;" | bc)
echo "total run time(m): $runtime_m"
runtime_h=$(echo "scale=2; $runtime_s / $sec_per_hr;" | bc)
```

# SLURM Array Script – SLURM commands

```
#!/bin/bash
#
#SBATCH --job-name="GLDS-104_raw_fastqc"
#SBATCH --output=raw_fastqc_out_logs/GLDS-104_raw_fastqc_%A_%a.out
#SBATCH --partition=priority
#SBATCH --mem=6000
#SBATCH --array=1-12
#
#SBATCH --mail-user=user@nasa.gov
#SBATCH --mail-type=END
```

- `#!/bin/bash` – Specifies that the Unix terminal is running bash
- `#SBATCH --job-name="GLDS-104_raw_fastqc"` – Specifies the job name
- `#SBATCH --output=raw_fastqc_out_logs/GLDS-104_raw_fastqc_%A_%a.out` – Instructs SLURM where to write standard output
- `#SBATCH --partition=priority` – Indicates which queue to add the job to
- `#SBATCH --mem=6000` – Specifies the required RAM in Mb
- `#SBATCH --array=1-12` – Indicates the job numbers to run
- `#SBATCH --mail-user=user@nasa.gov` – Specifies the e-mail address to send SLURM notifications
- `#SBATCH --mail-type=END` – Instructions SLURM to send an e-mail when jobs complete

# SLURM Array Script – nonSLURM commands

```
. ~/.profile
```

Instructions bash to use the user profile settings

```
echo "GLDS-104_raw_fastqc"
echo ""
```

Prints the job name followed by blank line to the standard output

```
start=$(date +%s)
echo "start time: $start"
```

Records then prints the job start date in seconds to the standard output

```
echo $HOSTNAME
```

Prints the name of the server running the job to standard out

```
echo "My SLURM_ARRAY_TASK_ID: " $SLURM_ARRAY_TASK_ID
```

Prints the SLURM job ID number to standard out

```
source activate /miniconda38_admin/envs/fastqc_09-2021
```

Activates the required conda environment

```
in_dir=/GLDS-104/00-RawData/Fastq
out_dir=/GLDS-104/00-RawData/FastQC_Reports
```

Specifies the directory holding the input files and where to print the output files

```
sample=$(cat samples.txt | sed -n ${SLURM_ARRAY_TASK_ID}p)
```

Links each array job number to a sample name

```
echo ""
echo "SAMPLE: ${sample}"
```

Prints a blank line followed by the sample name to standard out

```
echo ""
echo "FastQC version: "
fastqc --version
echo ""
```

Prints a blank line followed by the FastQC tool version number and then another blank line to standard out

# SLURM Array Script – nonSLURM commands

```
call="fastqc -o $out_dir $in_dir/${sample}_R*_raw.fastq.gz"

echo $call
echo ""
eval $call

echo ""
end=$(date +%s)
echo "end time: $end"
runtime_s=$(echo $(( end - start )))
echo "total run time(s): $runtime_s"
sec_per_min=60
sec_per_hr=3600
runtime_m=$(echo "scale=2; $runtime_s / $sec_per_min;" | bc)
echo "total run time(m): $runtime_m"
runtime_h=$(echo "scale=2; $runtime_s / $sec_per_hr;" | bc)
```

Stores the fastqc command as a variable

Prints the exact fastqc command to standard out followed by a blank line then executes the fastqc command

Prints a blank line followed by the job end date in seconds then calculates and prints the job run time in seconds, minutes, and hours to standard out

# Questions?