TOPICAL: Optimizing the Accessibility and Usage of Astronaut Omics Data Across the Scientific Community *via* GeneLab

A White Paper Submitted to the National Research Council / National Academy of Sciences Decadal Survey on Biological and Physical Sciences in Space

Principal Author:
Sylvain V. Costes
GeneLab Project Manager and Principal Investigator
Space Biosciences Research Branch
NASA Ames Research Center
Sylvain.V.Costes@nasa.gov
(650) 604-5343

Co-authors:
Beheshti A, Senior Scientist, KBR, NASA Ames Research Center
Galazka JM, GeneLab Project Scientist, NASA Ames Research Center
Paul AM, Assistant Professor, Embry-Riddle Aeronautical University, BMSIS, NASA Ames Research Center
Scott RT, Scientist, KBR, Ames Life Sciences Data Archive, NASA Ames Research Center
David J. Loftus, Medical Officer, NASA Ames Research Center
Sanders LM, Staff Scientist, BMSIS, GeneLab, NASA Ames Research Center

Date:
October 31, 2021

TOPICAL: Optimizing the Accessibility and Usage of Astronaut Omics Data Across the Scientific Community *via* GeneLab

Over the next 10 years, NASA's Science Mission Directorate (SMD) will advance deep space exploration by (1) Discovering secrets of the Universe, (2) Searching for extraterrestrial life, and (3) Protecting and improving life on Earth. GeneLab, NASA's open science data repository for genomics, is positioned to support and enable many of the biological discoveries that will be spearheaded by SMD in the next decade. GeneLab falls under the leadership of the Biological and Physical Sciences Division (BPSD) of SMD. The primary mission of BPSD is to *pioneer scientific discovery in and beyond low Earth orbit, to drive advances in science, technology, and space exploration, to enhance knowledge, education, innovation, and economic vitality.* GeneLab plays a prominent role in this mission. BPSD prioritizes *open science research*, based on the philosophy that an open science approach returns the maximum benefit to the agency. This philosophy aligns well with the core values of GeneLab.

NASA GeneLab Project

NASA's GeneLab (genelab.nasa.gov), also called the GeneLab Project, is an interactive open-access resource where scientists can upload, store, and download omics data from spaceflight experiments and from terrestrial experiments that involve simulating some aspect of the spaceflight environment (microgravity and/or space radiation) (1-2). GeneLab follows ethical practices for sample data collection, curation, and accessibility, and presents the data in an organized and user-friendly platform. Since its inception in 2015, GeneLab has realized its goal of becoming an open-access repository of omics data collected by principal investigators (PIs) working on space biology-funded research projects. The data that have been deposited in GeneLab come from a variety of species including rats, mice, nematode worms, fruit flies, plants, and many different microorganisms. GeneLab also includes data from humans—but from cells and cell lines only. What's conspicuously absent from GeneLab are data from astronauts, which is often confusing users who expecting to find these data. Astronaut privacy has been at the core of this gap and this paper tries to articulate potential solutions.

NASA Open Science Analysis Working Groups (AWG)

Established in 2018, the NASA GeneLab Analysis Working Groups (AWGs) help to optimize the processing of raw omics data to maximize the gain of new knowledge from the GeneLab data sets. Currently, there are 5 GeneLab AWGs, in the following areas: *Plants, Microbes, Animals-Mammals, Animals-Non-mammals,* and cross-cutting areas, *Multi-omics/Systems Biology.* A total of approximately 150 scientists participate in the GeneLab AWGs, from academia (U.S. and international), other federal laboratories besides NASA (e.g., NIH/NCI, Los Alamos National Laboratory, NIST, Lawrence Livermore National Laboratory), prominent research institutes (e.g., Broad Institute) and private companies with relevant expertise (e.g., Illumina). GeneLab AWG members meet monthly and hold an annual workshop hosted at the American Society for Gravitational and Space Research (ASGSR) conference. This scientific consortium has collaborated on multiple projects and published numerous peer-reviewed articles in a very short period (3-12). Both at the 2019 and 2021 AWG workshop, collaborative roundtable discussions and presentations of annual highlights projected the scientific need for regulated access of astronaut data. *A scientific need to compare omics data of different species to astronauts to determine their shared physiological translatability is crucial to plan for future exploratory missions to the lunar surface and Mars, expected to initiate in 2024.* In 2021, as the NASA Ames

Life Science Data Archive (ALSDA) database adopted the GeneLab Data System to create an integrated Open Science Databases System for Space Biology, a new AWG was created to support ALSDA with close to 60 members. As observed for omics data, a similar feedback was collected during the 2021 AWG workshop, with the group highlighting how critical non-omics data from Astronauts are to better understand the relevance of animal models for specific health outcomes in Astronauts.

Benefits of Including Astronaut Data in GeneLab.
As introduced in the previous section, the inclusion of astronaut data in GeneLab will allow the research community to leverage existing animal data to support the interpretation of human data. Clearly, there are many requirements to include astronauts as research subjects, so that their data can be used. The Genetic Information Non-discrimination Act must be followed, as well as the Privacy Act, the Common Rule, and OSHA guidelines. However, these regulations do no prohibit Open Science to collect such information, instead it gives guidelines to be followed. The key step is that astronauts must consent to participate voluntarily, after all relevant information about the studies is disclosed in the consent process. NASA's GeneLab resources include computational modeling of spaceflight and spaceflight-relevant data, providing a platform for data assimilation/management that promotes interagency cooperation and international collaboration of data sharing. *Addition of human data to NASA's GeneLab is critical for shaping the next steps of space exploration.*

Summary plan for NASA GeneLab, according to SMD requirements
Integration of omics and metadata from human crew on previous and future missions will enable science that improves astronaut health and life on Earth, which is a primary goal for SMD. NASA's GeneLab offers a platform that can integrate multiple models of spaceflight omics data, to better predict physiological outcomes affected by deep space spaceflight. This is due to collaboration of non-traditional, yet innovative spaceflight models of research, provided by directorates such as Human Exploration and Operations (HEOMD) and Space Technology (STMD) Mission Directorates. NASA's GeneLab will also assist with SMD requirements, in addition to being an open source repository sharing omics data from national/international partnerships, including private.

Motivations/Relevance for GeneLab integration of human data, in alignment with SMD requirements

- **(1)** Open-access sharing of omics- and meta-datasets that assess crew/human responses to spaceflight/ground-based simulations, respectively.
- **(2)** Cross-disciplinary and national and international collaboration investing in human exploration as high-priority.
  - **a.** improving scientific rigor by expanding mammalian data available for researchers.
  - **b.** providing quantifiable assessment for the relevance of different organismal ground- and space-models for human spaceflight.
- **(3)** Use of omics science to perform human assistance.
  - **a.** mitigating astronaut health risks by facilitating research that leads to biomarker and countermeasure development.
  - **b.** identifying spaceflight-induced long-term outcomes similar to Earth (e.g. "space acting as an aging accelerator")

Rationale to implement a repository of human data to BPS Open Science Databases

(1) *GeneLab AWG researchers seek to make quantitative connections to astronaut physiology following analysis of model organisms.* This is done to improve translational applicability of scientific findings. There is a major gap in understanding how spaceflight affects human physiology, although we have spaceflight data from cell culture and animal models. Although omics data exists for these model organisms, the lack of human omics information precludes comparing and translating findings from model organism data to increase our scientific understanding of spaceflight effects on humans. Translational omics becomes increasingly important as we expand to exploration missions in deep space; therefore the scientific community needs a facile and HIPAA-compliant omics data repository that is publicly accessible and securely houses astronaut data. Archived omics and non-omics astronaut datasets currently exist in sources such as the Life Science Data Archive (LSDA, https://lsda.jsc.nasa.gov/), but these sources typically require implemented funding and/or a long turnover time to attain data access approval.

(2) *GeneLab AWG researchers seek to access original measurements, not summaries from publications.* Currently there does not exist a unified and searchable repository that has access to astronaut omics data, nor connected metadata and physiological data. Despite multiple past and ongoing experiments that include omics measurements, physiological data, and metadata of pre-, during- , and post-flight re-acclimation to ground conditions, the only way to access these data is through peer-reviewed literature citations.

(3) *Currently it is exceptionally difficult to access astronaut physiological and health data.* The release of astronaut data is particularly sensitive due to small sample sizes and mission specific metadata that makes it relatively identifiable. Some of these data were recently accessible to a few AWG members who showed the power of adding them to meta-analysis using omics, revealing for example how mitochondria plays a pivotal role in the response to space stressors (7,10). We therefore encourge the National Research Council (NRC) to provide guidance on how the greater scientific community can access these unique high-throughput data of human subjects, while strictly following Institutional Review Board (IRB) regulations and ethics, ensuring no genetic health information will be identifiable.

Suggestions

How do we protect astronaut data in a way that maintains their right to privacy and protection of their genetic and physiological material? This requires an open discussion on the practical approaches to astronaut data storage in a secure data repository by proposing a few options that would enable generating processed data from raw omics data and metadata, retaining scientific value while neither compromising identity nor personal health information.

We recommend NASA GeneLab's repository maintains ethical standards for distribution of astronaut data, via a case-by-case basis. This would require a formal request of scientific justification from the researcher to both the PI responsible for the datasets and GeneLab. All requests would be reviewed via a pre-approved NASA's GeneLab IRB.

(1) *Adapting the Findability – Accessibility – Interoperability – Reusability (FAIR) principle utilized in modern databases (14).* At a minimum, all astronaut physiological, metadata, and omics data should be easily accessible. The GeneLab AWG consortium would like to

make a strong recommendation to have the NASA GeneLab platform host these secured and de-identifiable astronaut datasets (omics, physiological, and metadata). For this, GeneLab will adapt an IRB to securely assess data requests and to control distribution. One could use approaches taken by other agencies like the Smart IRB (https://smartirb.org/) set up by the National Institute of Health (NIH). Such setup allows one blanket unified IRB to be utilized by all participating institutions. NIH has been extremely successful with such approach, with a large array of registered universities and institutions, simplifying sharing of human data among the registered users under the same IRB. Further, housing data within a repository will enable easy cross-comparison of human data with the wealth of already available animal datasets within the Space Biology Open Science repositories.

(2) *The need to generate an additional secure repository is not required, as all astronaut datasets would be uploaded and curated within the current NASA GeneLab platform.* GeneLab currently houses data within secured database software which encompasses:

a. an advanced submission portal with controlled vocabulary for ontology,
b. a cross-platform document-oriented database for storage of metadata to facilitate data modeling and interpretation,
c. a HIPAA-compliant cloud solution for data storage, and
d. online software for data visualization and omics analysis.

(3) *Data should be as open-source as possible, and as closed-source as necessary.* This can be achieved by taking two steps that are easy to implement to our system. First, removing identifiable information from raw data by processing them into higher order data. For example, raw sequences of RNA are typically used to establish transcriptomics in an individual tissue. As one processes these raw sequences, individual genes are being detected, counted, and eventually a gene expression profile is characterized. During this process, raw data can be used to search for genetic variance of the individual and infer potential risk for disease, which is necessary to protect and deidentify. However, once all sequence data which involve individual genetic variance is deleted and only the gene expression levels are reported, identifying an individual becomes close to impossible.

(4) *Scientific consensus regarding metadata: they can always be open-source*. However, this is not the case for many important metadata related to astronauts (for example, on-board telemetry is not Open Source). Indeed, metadata can be used to identify astronauts. For instance, information regarding mission duration, radiation doses received during a mission, the year of the mission, or the sex of the astronaut can all narrow down the identity of the astronaut being studied. We propose to tackle this issue by removing the year of the associated mission, and aggregating the processed data by group, which would be discretized. For example, the exact duration of a flight is not necessary and could be binned into a larger increment. Environmental data could be summarized in large bins as well (for example, radiation dose increment of 5 mGy provides sufficient resolution biologically, while making it difficult to reconstruct the exact duration of flight). More examples of reducing the personal information available in data or metadata could be provided upon request. Additionally, GeneLab has put in place a thorough process to curate all relevant metadata, showing the importance of capturing and summarizing environmental factors collected on the ISS or STS, in order to interpret such data (6-10). An example of the level of detail afforded by standard GeneLab metadata acquisition is indicated by the space environmental data collection effort (https://genelab.nasa.gov/environmental/) with the summary of CO_2 levels, radiation types and levels, O_2 levels, temperature, vibration,

humidity, and light cycle. For example, high CO_2 levels are correlated with hypoxic molecular signature in mice (12).

(5) *There should be a well-defined path to accessing different levels of astronaut data, depending on sensitivities.* Paths should be linked to various levels of access controls, that can be defined by the GeneLab IRB. For example, aggregated gene or protein expression levels as lowest sensitivity compared to raw sequence data as highest sensitivity.

(6) *Previously collected astronaut data (physiological, omics, and metadata) should be linked and harmonized with GeneLab.* The benefits of this would result in:

a. Harmonizing
    i. metadata frameworks including controlled vocabulary and ontology
    ii. environmental metadata
    iii. processing pipelines when possible
b. Facilitating dataset comparison
c. Leveraging
    i. existing systems
    ii. existing communities
    iii. the possibility of knowledge-based database for humans and risk models

Summary

The GeneLab Analysis Working Group Consortium believes that GeneLab should serve as a repository for astronaut omics data, in addition to data from model organisms. Now is the time to begin discussing strategies for storing and distributing metadata, omics, and physiological data, and integrating data from the Ames Life Sciences Data Archive. Our ultimate goal is to allow the scientific community easy access to these datasets, securely and ethically, to maximize scientific benefit for the agency, and, ultimately, to benefit NASA astronauts.

References

(1) Berrios, D.C., et al., *NASA GeneLab: interfaces for the exploration of space omics data.* Nucleic Acids Res, 2021. **49**(D1): p. D1515-D1522.
(2) Ray, S., et al., *GeneLab: Omics database for spaceflight experiments.* Bioinformatics, 2019. **35**(10): p. 1753-1759.
(3) Overbey, E.G., et al., *NASA GeneLab RNA-seq consensus pipeline: standardized processing of short-read RNA-seq data.* iScience, 2021. **24**(4): p. 102361.
(4) Nelson, C.A., et al., *Knowledge Network Embedding of Transcriptomic Data from Spaceflown Mice Uncovers Signs and Symptoms Associated with Terrestrial Diseases.* Life (Basel), 2021. **11**(1).
(5) Cahill, T., et al., *Mammalian and Invertebrate Models as Complementary Tools for Gaining Mechanistic Insight on Muscle Responses to Spaceflight.* Int J Mol Sci, 2021. **22**(17).
(6) McDonald, J.T., et al., *NASA GeneLab Platform Utilized for Biological Response to Space Radiation in Animal Models.* Cancers (Basel), 2020. **12**(2).
(7) da Silveira, W.A., et al., *Comprehensive Multi-omics Analysis Reveals Mitochondrial Stress as a Central Biological Hub for Spaceflight Impact.* Cell, 2020. **183**(5): p. 1185-1201 e20.
(8) Afshinnekoo, E., et al., *Fundamental Biological Features of Spaceflight: Advancing the Field to Enable Deep-Space Exploration.* Cell, 2020. **183**(5): p. 1162-1184.
(9) Beheshti, A., et al., *GeneLab Database Analyses Suggest Long-Term Impact of Space Radiation on the Cardiovascular System by the Activation of FYN Through Reactive Oxygen Species.* Int J Mol Sci, 2019. **20**(3).
(10) Beheshti, A., et al., *Multi-omics analysis of multiple missions to space reveal a theme of lipid dysregulation in mouse liver.* Sci Rep, 2019. **9**(1): p. 19195.
(11) Beheshti, A., et al., *A microRNA signature and TGF-beta1 response were identified as the key master regulators for spaceflight response.* PLoS One, 2018. **13**(7): p. e0199621.
(12) Beheshti, A., et al., *Global transcriptomic analysis suggests carbon dioxide as an environmental stressor in spaceflight: A systems biology GeneLab case study.* Sci Rep, 2018. **8**(1): p. 4191.
(13) Berrios, D.C., A. Beheshti, and S.V. Costes, *FAIRness and Usability for Open-access Omics Data Systems.* AMIA Annu Symp Proc, 2018. **2018**: p. 232-241.