Topical: ***Omics and Open Science: A Platform and Approach for the Future for Space Biology***

Porterfield, D. M.[1], Tulodziecki, D.[2]; Costes, S. V.[3-4]; Beheshti, A.[5,6,7], Sanders, L. M.[3]
[1]Department of Agricultural and Biological Engineering, Purdue University, West Lafayette, IN, (765)494-1162, porterf@purdue.edu; [2]Department of Philosophy, Purdue University, West Lafayette, IN; [3]Genelab, NASA Ames Research Center, Mountain View CA; [4]Biosciences Division, NASA Ames Research Center, Mountain View, CA; [5]KBR, Space Biosciences Division, NASA Ames Research Center, Moffett Field, CA; [6]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, 02142, USA; [7]COVID-19 International Research Team.

Submitted to the Decadal Survey on Biological and Physical Sciences Research in Space 2023-2032 (BPS2023) conducted by The National Academies of Sciences, Engineering and Medicine.

Abstract

Funding organizations around the world are adopting open science policies, resulting in a pressing need for open science programs. In response to the 2011 decadal survey, NASA sought to expand and accelerate omics research, releasing its GeneLab Strategic Plan in 2014. GeneLab is an open science data repository and analysis portal for spaceflight and space-relevant omics data. GeneLab's output has been outstanding, but its full potential as a way to transform space biology has not yet been achieved. NASA should pursue the development of GeneLab as an open science platform in earnest.

Introduction

During the era of completion of the International Space Station (ISS) and the retirement of the Shuttle program, NASA effectively decimated its basic programs for the support of life and physical sciences. The effects of these cuts on research programs and communities were drastic, delaying, if not sacrificing the capacity to develop NASA-mission-critical science for future human exploration. As a result, the previous 2011 decadal survey [1] highlighted the importance of re-establishing these basic programs. Further, it called for the expanded use of omics in space biology due to its potential for advancing both the quality and throughput of basic research. In response, NASA sought to rebuild the science community in a way that would expand and accelerate research and development beyond its traditional base, in particular with respect to omics research. NASA released its GeneLab Open Science Strategic Plan in 2014 [2] and since then there have been substantial investments in the establishment of GeneLab, an open science data repository and analysis portal for spaceflight and space-relevant omics data. While GeneLab's output in terms of both quality and quantity has been outstanding, its full potential as a platform to transform space biology has not yet been achieved. Adequate investment in the development of GeneLab as an open science platform should be pursued by the agency in earnest.

Open Science

'Open science' is a broad term encompassing a diverse set of practices aimed at facilitating the making available of scientific research both to other members of the scientific community and the public. In this vein, the *National Academies of Sciences, Engineering, and Medicine*, for example, speak of open science as "aim[ing] to ensure the free availability and usability of scholarly publications, the data that result from scholarly research, and the methodologies, including code or algorithms that were used to generate those data" [3]. More generally, the open science movement covers many areas, ranging from open access of research outputs and open data to open peer review, citizen science, and the generation of a variety of open infrastructures and tools. It has been recognized that open science has the potential to transform science: broader access to scientific research outputs is becoming more widespread and both public and private funding organizations around the world are adopting open science policies. A number of European projects are geared toward facilitating the implementation of open science principles, such as the *Amsterdam Call for Action on Open Science*, the *Berlin*

Declaration, and the *Budapest Open Access Initiative*. As recently as 2018 the *European Open Science Cloud* was formed as an international initiative to provide open science infrastructure, and just this year *Plan S* took effect: an initiative for open access publishing supported by *cOAlitionS*, an international consortium of research and funding agencies from a number of European countries.

Intergovernmental organizations around the world have also recognized the social and economic benefits of open science and jumped on the open science bandwagon, including the European Commission, the European Parliament, the European Council, the Organization for Economic Cooperation and Development (OECD), and even the World Bank. UNESCO's new *Recommendation on Open Science for 2021* is set to be adopted by the General Conference at its 41st session in November 2021, after having been tasked in the previous session by 193 member states "with the development of an international standard-setting instrument on Open Science" [4]. In the US, the White House Office of Science and Technology Policy has been pursuing similar goals: on 22 February 2013, it mandated federal science agencies to increase access to unclassified research obtained through federal funding (including data and metadata) [5] and as recently as 2020, it convened a series of meetings on open science and public access. Also in the US, the National Academies of Sciences, Engineering, and Medicine's, "Roundtable on Aligning Incentives for Open Science is convening critical stakeholders to discuss the effectiveness of current incentives for adopting open science practices, current barriers of all types, and ways to move forward to optimally align reward structures and institutional values" [6].

Within the scientific research community open data has already proven to be highly significant in the Space Telescope Institute, as well as in the earth sciences. In the life sciences the shift toward open data has been most noticeable with the advent of readily available genomics databases. After the sequencing of the first human genome became complete in 2003, several national and international collaborations were established to analyze, annotate, and archive bioinformatic data. Examples of these database projects include the *HapMap Project*, the *Allen Brain Atlas*, *The Cancer Genome Atlas,* and *1000 Genomes*. U.S. government agencies have also taken actions to increase the availability of genomic data, particularly for bioinformatic analysis. The National Center for Biotechnology Information, in conjunction with the European Bioinformatics Institute and the DNA Databank in Japan, established the *Sequence Read Archive* for whole genome sequencing data [7]. Collectively, these systems have resulted in the publication of thousands of peer-reviewed articles within the past twenty years, and have often driven collaborations between multiple principal investigators [8, 9]. Most recently, we have seen the power of open (omics) data as a driving factor behind the rapidity of progress with respect to COVID-19 research. The broad and open availability of this data has also made highly collaborative papers the norm in this area, with many prominent papers having 50+ authors (see, for example, [10]). It is thus clear that open science has gained momentum and that there is a pressing need for open science programs and policies at national, international, and institutional levels.

GeneLab: Open Science for Life on Earth

The GeneLab system functions partially as an open data repository for NASA investigations, but it is actually built around an instrumentation and software platform that analyzes and integrates omics at the level of systems biology. Collectively, omics are a group of technologies that all deliver high content data profiles of distinct classes of biological molecules. Genomics (and, in fact, the genome era) started with the human genome project and the expansion and improvement of basic DNA sequencing technology. DNA is the template for gene expression in the form of mRNA translation, and transcriptomics allows for measurement and cataloging of all gene expression within the system based on nucleotide biotechnology. Proteomics captures information about protein profiles, while metabolomics measures small molecules and biochemical pathways, with both being based mostly on the derivatization of mass spectrometry. Together these high-content technologies allow for the study of the flow of biological information into activity at a systems level. The resulting systems-level data sets are massive and can be used to drive open science in a unique way in biology. Given its investment in the GeneLab platform, NASA should provide the support needed to fully engage and incentivize the expanding community of scientists joining the space biology community.

Limitations to Research can be Overcome

Research output and laboratory operation in space face unique challenges. One of these is that the ISS-laboratory has a limited lifetime in orbit. This is partially a political issue, but also one of hardware, since the estimated structural and operational lifetime of the ISS is practically limited. As a result, total ISS research output is a function both of its lifetime and its throughput. Throughput in turn is a function of the number of investigations conducted on the ISS, and of the significance of the scientific outcomes resulting from these investigations. Therefore, NASA should fly as many high-quality scientific investigations as possible between now and the end of ISS operations in orbit. The number of such investigations is limited by several bottlenecks, including crew time, spaceflight up-mass, scientific hardware, payloads processing, and the supplies sent up to the ISS. Ever since the shuttle was retired this has been severely limiting, and with commercial space providers coming to the level for full utilization only recently, recovery has been slow. With respect to scientific sample return, down-mass is also an issue, since stowage on the Soyuz crew vehicle is severely limited. Jaxa's HTV, the European Space Agency's Ariane, and the Russian Progress vehicles only deliver cargo and are burned up on re-entry. Of the US commercial providers, the only current return capability is via SpaceX's Dragon.

Omics Enables Open Space Biology

Advances in computing and information technology have changed the way we learn, explore, socialize, and do business. In the life sciences, the dawn of the genomics and bioinformatics era promised advances in biology, agriculture, the environment, and medicine. The flow of information from genome into phenome involves gene transcription, protein translation, and protein-mediated metabolism. By exploring multidimensional data across different omics domains (genomics, epigenomics, transcriptomics, proteomics, metabolomics, lipidomics, miRNomics), the emergence of omics technologies now offers the potential to adopt a "system of systems" approach in biology. High content gene-expression and fast and inexpensive DNA

sequencing techniques are in common use by now and the data sets produced by these new technologies will become even more valuable as the tools to analyze them continue to advance. Importantly, these large data sets can be leveraged to create more opportunities for research if we adopt a reference-experiment approach for generating open science data. Contrasting this new approach with existing models of single-PI flight experiments and single-hypothesis science management only serves to highlight the many inadequacies and limitations of these traditional models, especially in the unique research environment for space biology.

The GeneLab Open Science Strategic Plan described such an approach: community-designed reference experiments are used to generate big-data based on integrated omics (open research), which will then be deposited into open data and informatics systems (open data) in order to engage and support a large and diverse research community of PIs (open competition). PI funding for these translational programs will come from NASA grant opportunities. Through targeted translational NASA ground research announcements, the community is encouraged to use the data generated by these reference experiments to develop hundreds, if not thousands, of individual, hypothesis-driven investigations on the ground from just a single flight opportunity. Because the data is open to all, we expect that competition will encourage rapid innovation by both NASA and non-NASA community users and stakeholders. In particular, this will increase both the diversity and numbers of young investigators entering the field.

Summary and Recommendations

The 2014 GeneLab Open Science Strategic Plan established the goal of enabling an integrated omics platform for systems level biological data and described a new open research approach for implementing community reference experiments. This platform has been highly successful and established NASA as a leading agency in promoting open science. In the next decade, open science is set to emerge both globally and nationally as a priority for federal funding programs. With GeneLab, NASA and SMD are well positioned to emerge as a leading innovator at the federal level with respect to science policy. Through fully engaging with – and executing – the open science model, they have the opportunity to set an example for other agencies and their attempts to implement open science policies for science program management. The specific recommendations to achieve this include:

1. REFERENCE EXPERIMENTS. Establish a plan to conduct open science reference experiment campaigns for all major biological systems. These plans should follow the description for the execution of full open science reference experiments described in the GeneLab strategic plan.
2. INCENTIVIZATION AND SUPPORT. NASA should include a funding opportunity for ground based translational research: no less than $10M should be set-aside for forty awards. These translational awards would seek to identify new lines of research to further develop new knowledge derived from the spaceflight experiment.

Cited References:

[1] National Research Council 2011. *Recapturing a Future for Space Exploration: Life and Physical Sciences Research for a New Era*. Washington, DC: The National Academies Press. https://doi.org/10.17226/13048

[2] NASA. 2018. GeneLab Strategic Plan. https://www.nasa.gov/ames/research/space-biosciences/technical-reports or https://www.spacestationresearch.com/wp-content/uploads/GeneLabStrategicPlan_Baseline_2014.pdf [last accessed 31 October 2021]

[3] National Academies of Sciences, Engineering, and Medicine. 2018. Open Science by Design: Realizing a Vision for 21st Century Research. Washington, DC: The National Academies Press. https://doi.org/10.17226/25116

[4] https://en.unesco.org/science-sustainable-future/open-science [last accessed 31 October 2021]

[5] Holdren, J.P. 2013. Memorandum for the Heads of Executive Departments and Agencies: Increasing Access to the Results of Federally Funded Scientific Research. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf [last accessed 31 October 2021]

[6] https://www.nationalacademies.org/our-work/roundtable-on-aligning-incentives-for-open-science [last accessed 31 October 2021]

[7] Kodama, Y., Shumway, M. and Leinonen, R., 2012. The sequence read archive: explosive growth of sequencing data. *Nucleic acids research*, *40*(D1), pp. D54-D56. https://doi.org/10.1093/nar/gkr854

[8] Pautasso, M., 2012. Publication growth in biological sub-fields: patterns, predictability and sustainability. *Sustainability*, *4*(12), pp.3234-3247. https://doi.org/10.3390/su4123234

[9] Piwowar, H.A., Day, R.S. and Fridsma, D.B., 2007. Sharing detailed research data is associated with increased citation rate. *PloS one*, *2*(3), p.e308. https://doi.org/10.1371/journal.pone.0000308

[10] McDonald, J.T., Enguita, F.J., Taylor, D., Griffin, R.J., Priebe, W., Emmett, M.R., Sajadi, M.M., Harris, A.D., Clement, J., Dybas, J.M. and Aykin-Burns, N., 2021. Role of miR-2392 in driving SARS-CoV-2 infection. *Cell Reports*, p.109839. https://doi.org/10.1016/j.celrep.2021.109839