NASA's GeneLab: An Integrated Omics Data Commons and Workbench

Daniel C. Berrios, MD MPH PhD[1,2], Sylvain V. Costes, PhD[2], Peter B. Tran, MS[2]
[1]USRA, Moffett Field, CA USA; [2]NASA, Moffett Field, CA USA

Purpose

GeneLab (https://genelab.nasa.gov) is a NASA initiative designed to accelerate "open science" biomedical research in support of the human exploration of space and the improvement of life on earth. The GeneLab Data Systems (GLDS) were developed to help investigators corroborate findings from "omics" (genomics, transcriptomics, proteomics, and metabolomics) assays and translate them into systems biology knowledge and eventually therapeutics, including countermeasures to support life in space. Phase I of the project (completed) emphasized developing key capabilities for submission, curation, storage, search, and retrieval of omics data from biomedical research in and of space environments. The development focus for Phase II (completed) was federated data search and retrieval of these kinds of data from other open-access repositories. The last phase of the project (in work) entails developing an omics analysis tool set, and a portal to visualize processed omics data, emphasizing integration with the data repository and search functions developed during the prior phases. The final product will be an open-access system where users can individually or collaboratively publish, search, integrate, analyze, and visualize omics data.

System Design

The GLDS seeks to solve several problems that currently plague the systems biology research community. Omics data are widely distributed across different repositories, many of which contain vast numbers of records, making the discovery of relevant data difficult and time-consuming. Omics data are typically large, with data transport between multiple systems also time-consuming. Each repository employs different metadata schemas, impeding data collation and integration. Finally, many research groups and individuals lack resources to develop and maintain systems providing omics analysis tools and collaboration workspaces, and most publicly available ones serve the different needs of many communities. The GLDS is focused on serving space biomedical researchers; users can choose to search only the data hosted in the GLDS (currently ~20 TB), or expand their searches to other repositories to discover related data. The GLDS employs a metadata warehouse that collects and updates metadata records from: the National Center for Biotechnology Information's Gene Expression Omnibus (GEO), the European Bioinformatics Institute's PRoteomics IDEntifications (PRIDE) repository, and the Argonne National Laboratory's Metagenomics Analysis server (MG-RAST). To bridge differences in metadata the GLDS employs a Common Object Model (COM) using a standard interchange format (e.g., JSON) to which each systems' representation of metadata is mapped. The final components of the GLDS are user workspaces and a suite of omics analysis tools, with integration of each with GLDS data repository, to minimize data transport times. The first two phases of the GLDS have already been used to yield insights into important systems biology research questions for the space biology community[1]. The system at the end of phase II, which included deployment of the repository and user workspace (Figure 1a,1b), had approximately 100 user accounts, and at the end of phase III (with data analysis tools[2], Figure 1c) is targeted to have 1,000 user accounts.

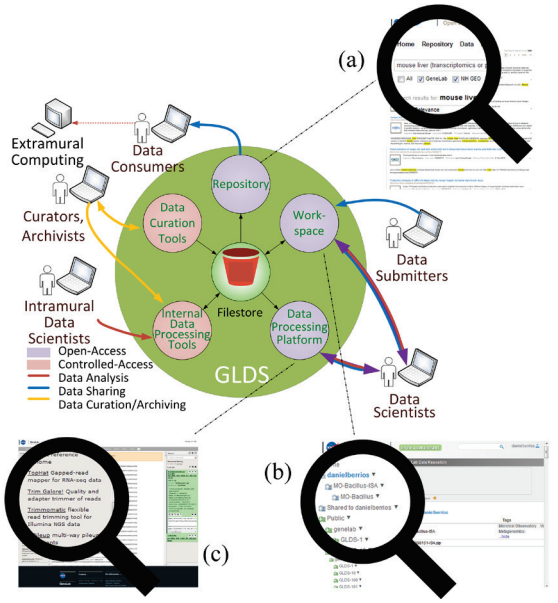

Figure 1. The GeneLab Data Systems (GLDS), including the omics data repository with federated search (a), user workspace (b) and Galaxy[2] server (c).

References

1. Beheshti A, Cekanaviciute E, Smith DJ, Costes SV. Global transcriptomic analysis suggests carbon dioxide as an environmental stressor in spaceflight: a systems biology GeneLab case study. Sci Rep. 2018;8(1):4191.
2. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, et al. Galaxy: a platform for interactive large-scale genome analysis. Genome Res. 2005;15(10):1451-5.