GeneLab Phase 2: Integrated Search & Data Federation of Space Biology Experiment Data

P.B. Tran[1], D.C. Berrios[1], M.M. Gurram[1], J.C.M. Hashim[1], S. Raghunandan[1], S.Y. Lin[1], T.Q. Le[1], D.M. Heher[1], H.T. Thai[1], J.D. Welch[2], S.M. Caldwell[3], O.G. Stotzky[4], M.G. Skidmore[4]
[1]Intelligent Systems Division, [2]Partnerships Division, [3]Space Biosciences Division, [4]Project Management Division, NASA Ames Research Center, Moffett Field, CA, USA

GeneLab
Open Science for Exploration

The GeneLab project is a **science initiative** to maximize the scientific return of omics data collected from spaceflight and from ground simulations of microgravity and radiation experiments, supported by a **data system** for a public bioinformatics repository and collaborative analysis tools for these data. The mission of GeneLab is to maximize the utilization of the valuable biological research resources aboard the ISS by collecting genomic, transcriptomic, proteomic and metabolomic (so-called "omics") data to enable the exploration of the molecular network responses of terrestrial biology to space environments using a systems biology approach. All GeneLab data are made available to researchers worldwide through its open-access data system.

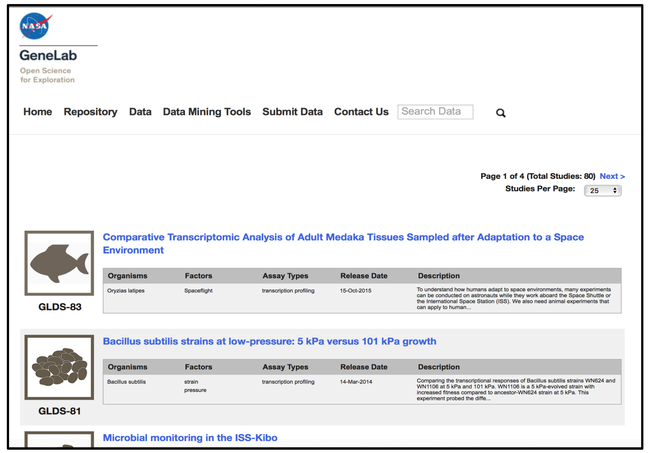
GeneLab is currently being developed by NASA to support worldwide accessibility of biomedical research in order to enable the human exploration of space and improve life on earth. Open access to Phase 1 of the GeneLab Data Systems (GLDS) was implemented in April 2015. Download volumes have grown steadily, mirroring the growth in curated space biology research data sets (80 as of Sept. 2016), now exceeding 10 TB/month, with over 10,000 file downloads since the start of Phase 1. For the period April 2015 to May 2016, most frequently downloaded were data from studies of *Mus musculus* (39%) followed closely by *Arabidopsis thaliana* (30%), with the remaining downloads roughly equally split across 12 other organisms (each < 10% of total downloads). GLDS Phase 2 is focusing on interoperability, supporting data federation, including integrated search capabilities, of GLDS-housed data sets with external data sources, such as gene expression data from NIH/NCBI's Gene Expression Omnibus (GEO), proteomic data from EBI's PRIDE system, and metagenomic data from Argonne National Laboratory's MG-RAST. GEO and MG-RAST employ specifications for investigation metadata that are different from those used by the GLDS and PRIDE (e.g., ISA-Tab). The GLDS Phase 2 system will implement a Google-like, full-text search engine using a Service-Oriented Architecture by utilizing publicly available RESTful web services Application Programming Interfaces (e.g., GEO Entrez Programming Utilities) and a Common Metadata Model (CMM) in order to accommodate the different metadata formats between the heterogeneous bioinformatics databases. GLDS Phase 2 completion with fully implemented capabilities will be made available to the general public in September 2017.

For more information please visit the GeneLab project website at:
http://genelab.nasa.gov

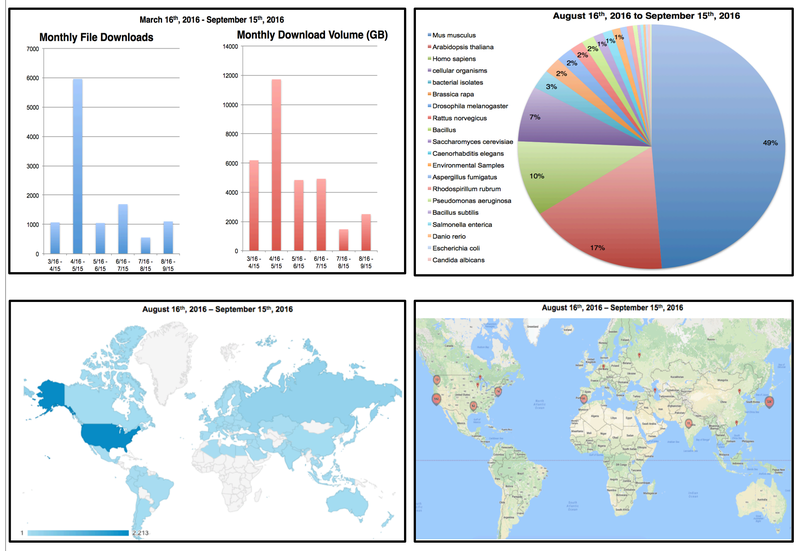
Explore the GeneLab Data System repository at:
https://genelab-data.ndc.nasa.gov/genelab/

ASGSR October 2016

GLDS Phase 1 (aka "C-Gene") Operational System

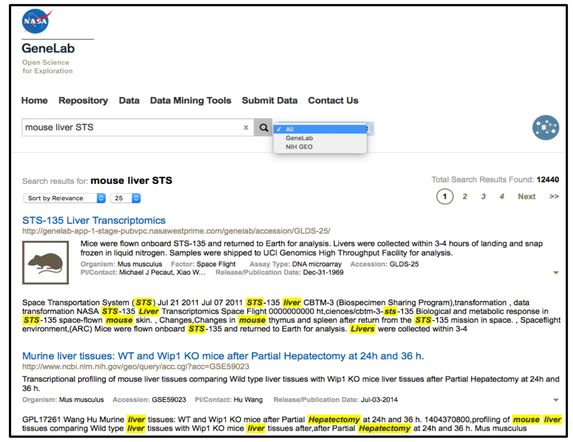


Simple keyword search capability with browsing and navigation

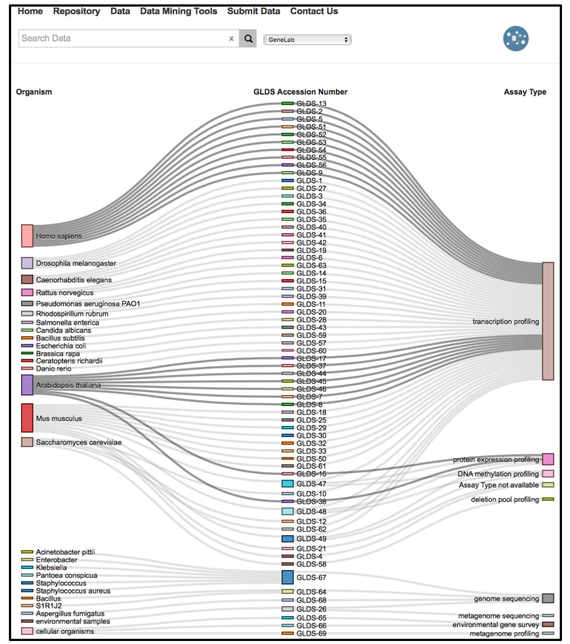


GLDS Metrics (As of 9/15/2016)

GLDS Phase 2 (aka "X-Gene") Work-In-Progress System



Google-like, full-text search engine with data federation/integration to NIH/NCBI's GEO database (over 80K GSE metadata records ingested)



Data Visualization (GLDS Organisms vs. Assay Types)