GeneLab: Open Science for Life in Space

National Aeronautics and Space Administration

Jonathan M. Galazka
GeneLab Project Scientist

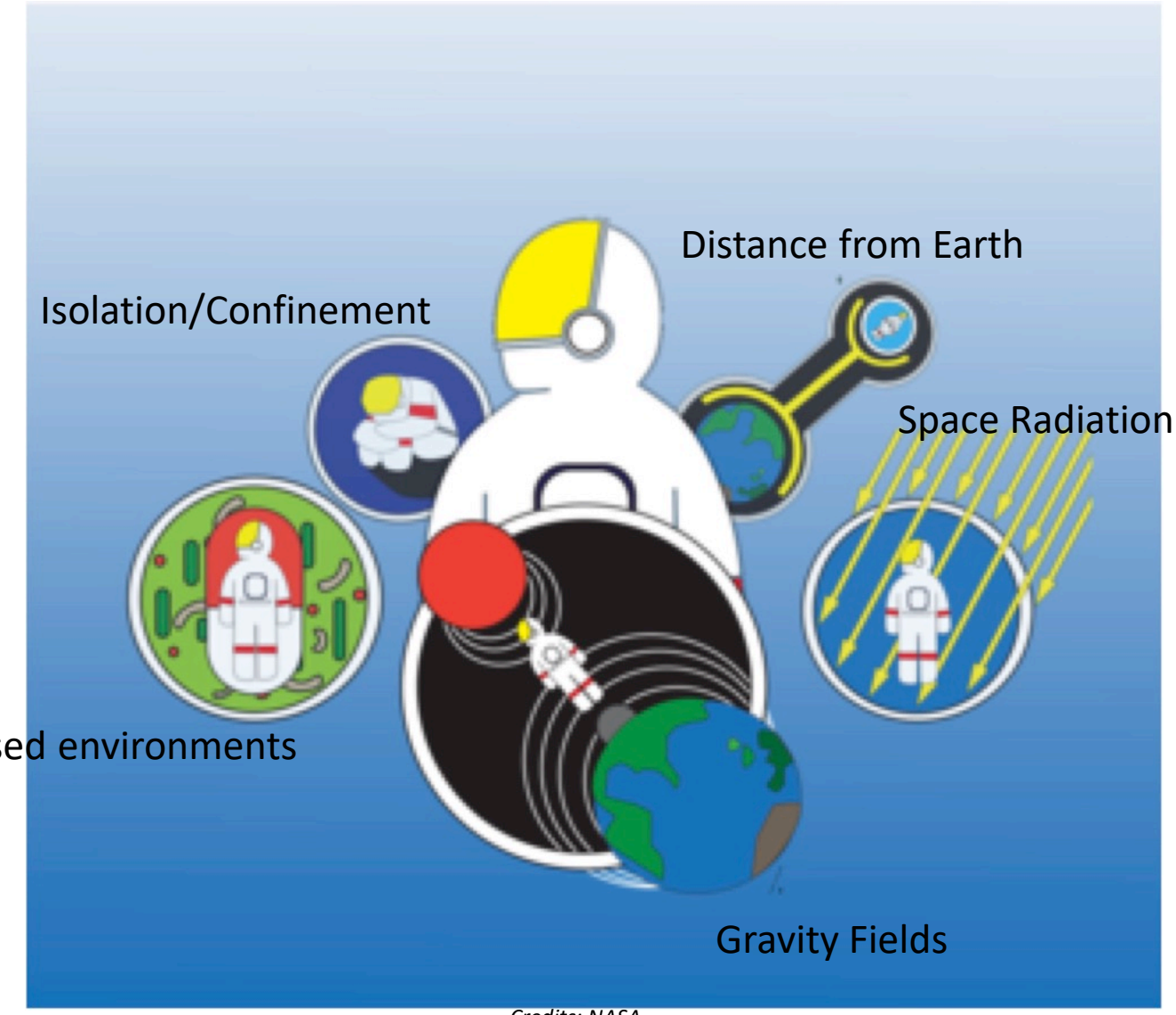
Talk outline

- **GeneLab background and motivation**
- **GeneLab Data System (GLDS) and datasets**
  - Data system architecture (AWS, search, federation,)
  - Datasets, dataset metadata, and curation
  - Collaborative workspace
  - Cloud analysis platform
- **GeneLab community**
  - Analysis Working Groups (AWGs)
- **Increasing the scientific return on Space Biology investments**
  - Leveraging biomedical research
  - Aggregating Space Biology research
  - Increasing data volume from Space Biology research

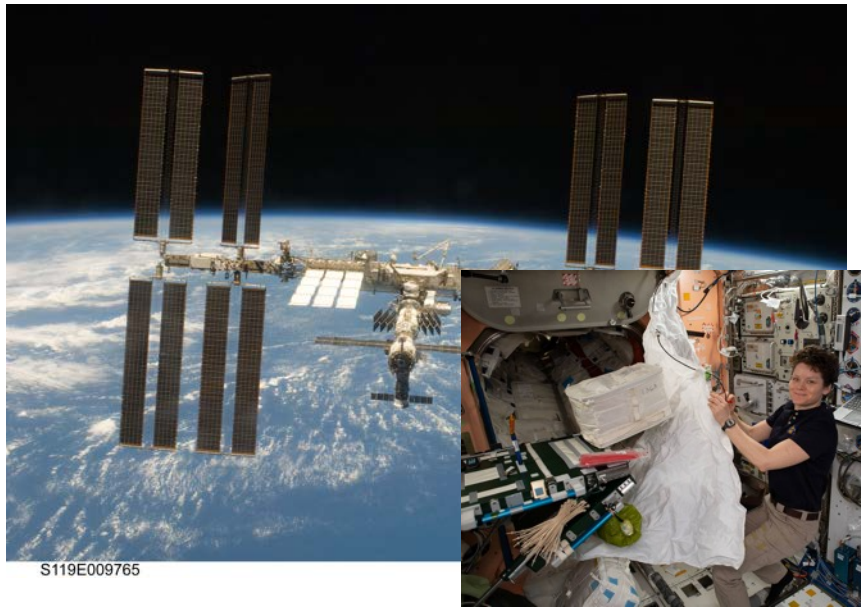
Talk outline

- **GeneLab background and motivation**
- **GeneLab Data System (GLDS) and datasets**
  - Data system architecture (AWS, search, federation,)
  - Datasets, dataset metadata, and curation
  - Collaborative workspace
  - Cloud analysis platform
- **GeneLab community**
  - Analysis Working Groups (AWGs)
- **Increasing the scientific return on Space Biology investments**
  - Leveraging biomedical research
  - Aggregating Space Biology research
  - Increasing data volume from Space Biology research

Challenges of Spaceflight

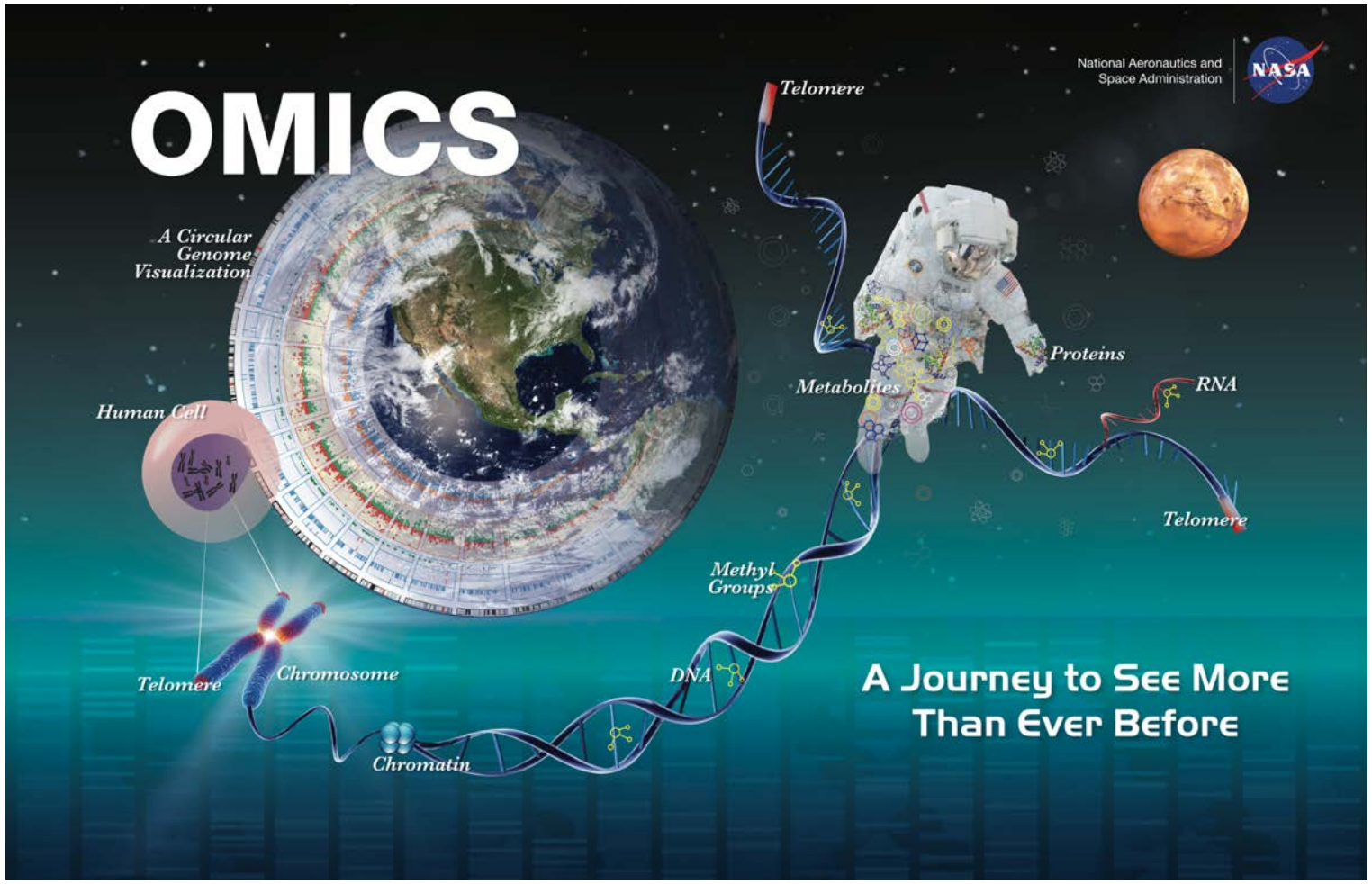


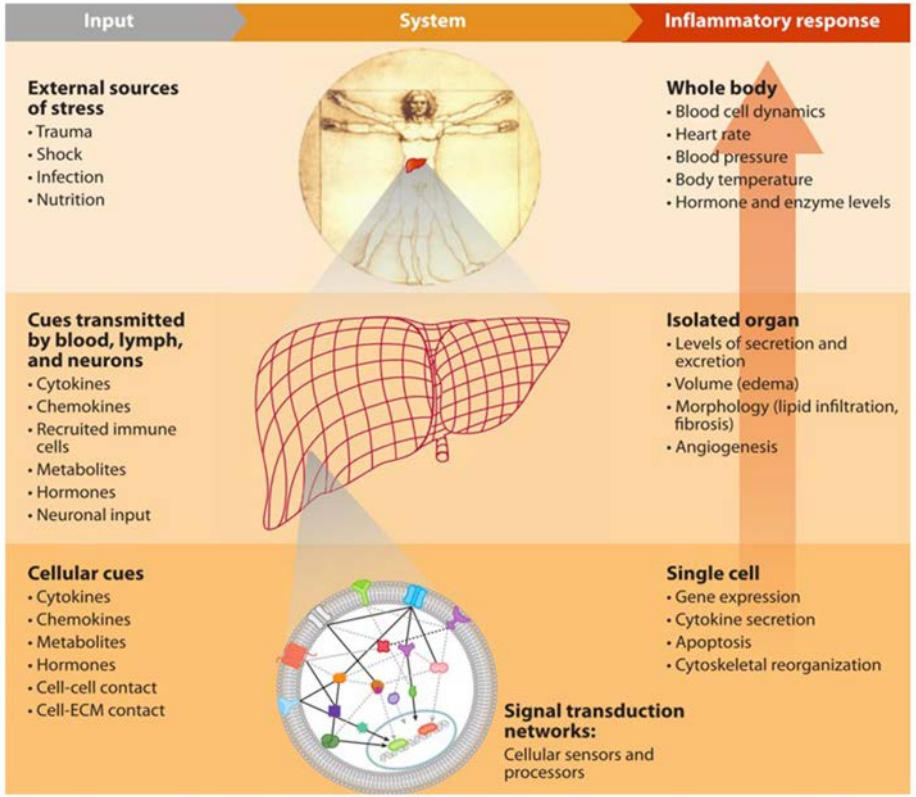
Credits: NASA



Biological systems are complex hierarchical systems

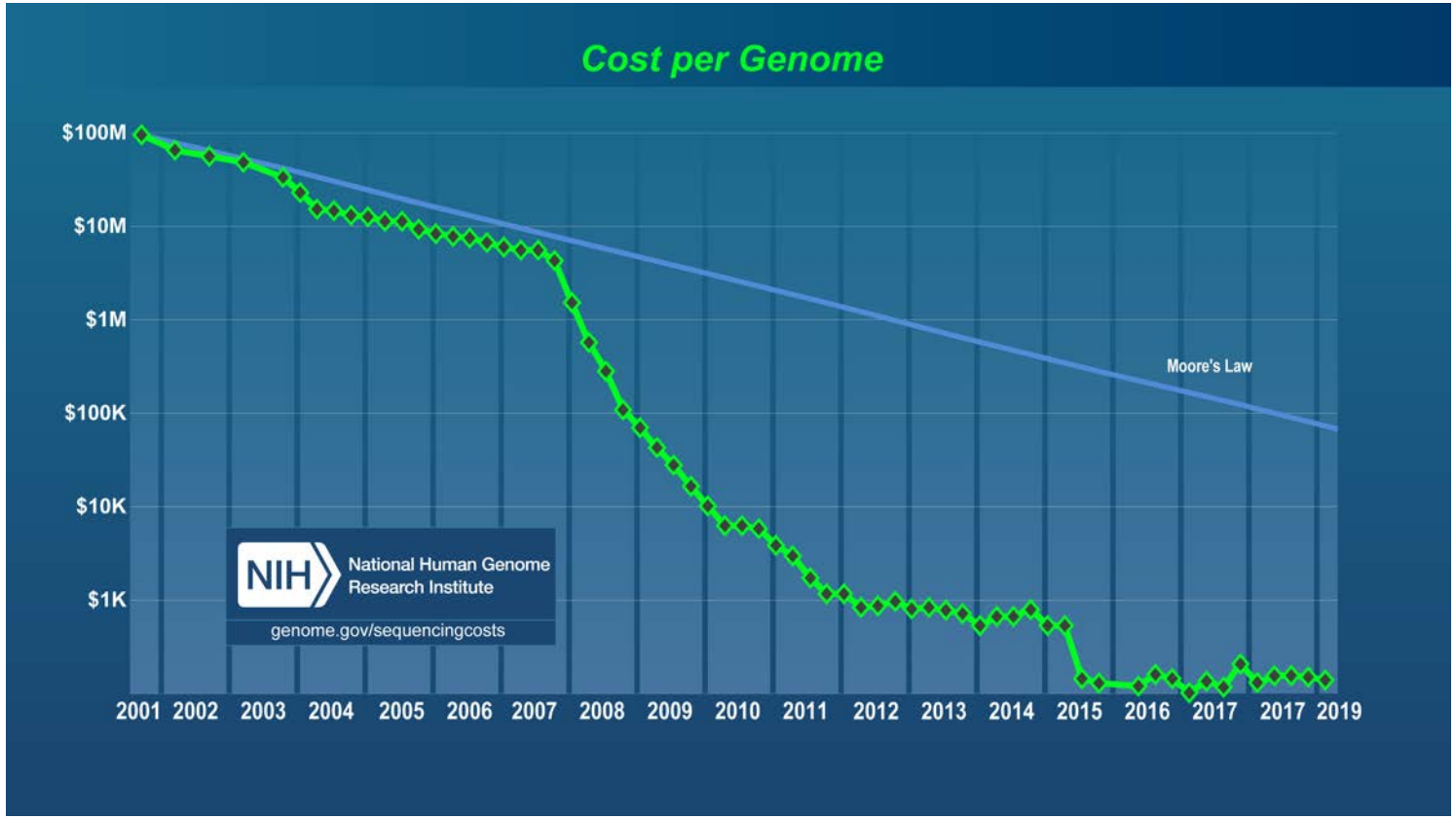
Humans body: 79 organs, 200 cell types, 20,000 protein coding genes, Many more mRNA splice isoforms, Many more protein isoforms, 16,000 endogenous metabolites (22,000 from food), Plus…DNA, RNA, and protein modifications, histone modifications, protein-protein interaction, etc., etc., etc., We're complicated!





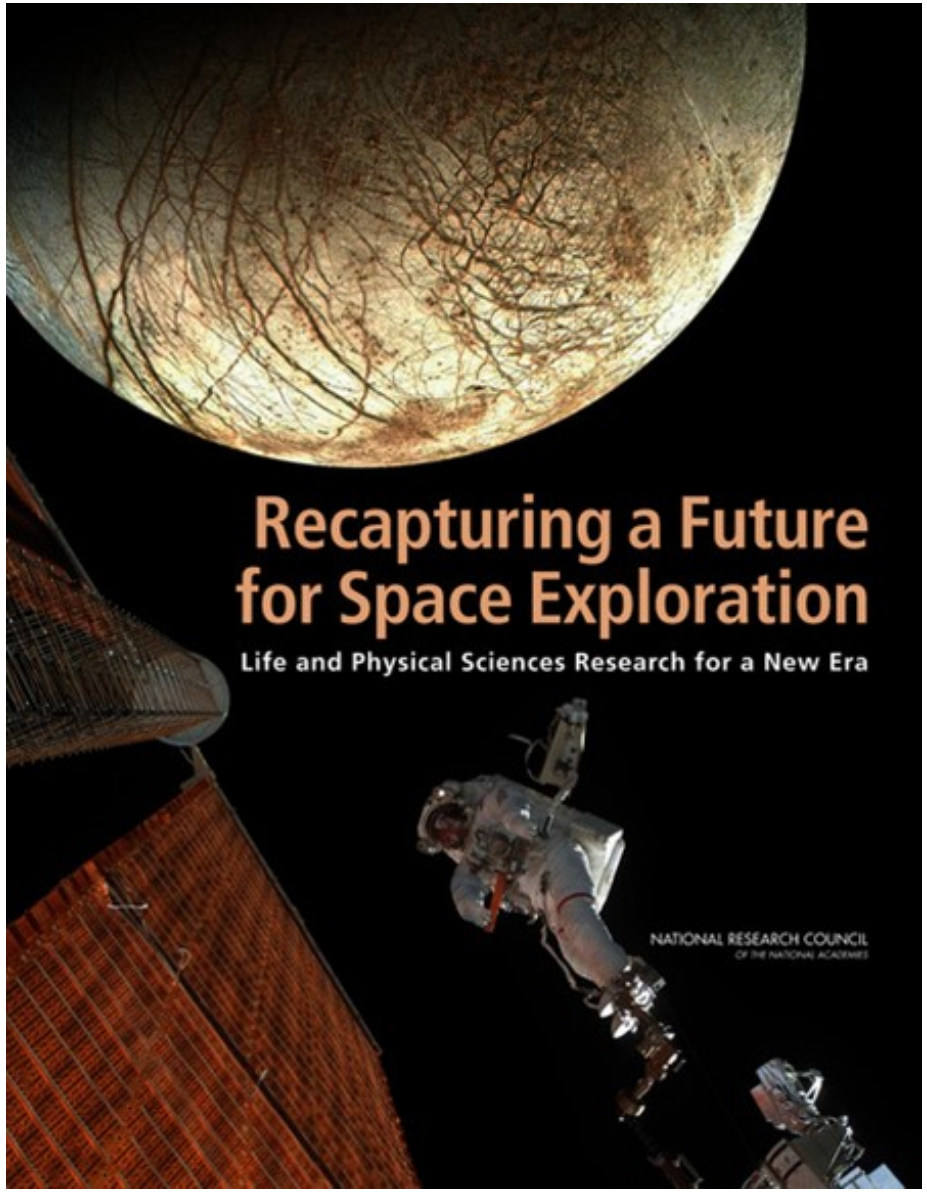
Kamisoglu et al. Front. Pharmacol., 27 February 2017

2011 NRC Decadal Survey and the Sequencing Paradigm Shift

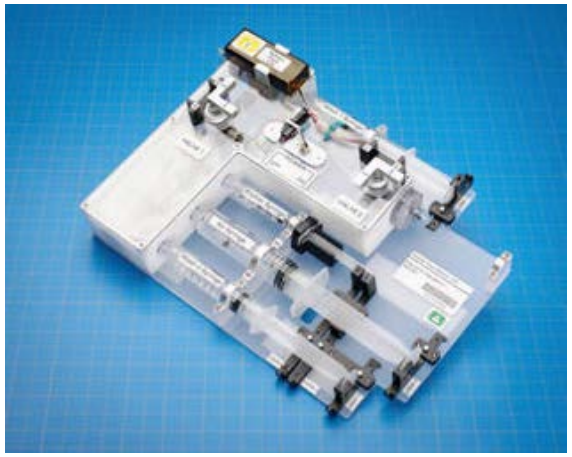


*"…**genomics, transcriptomics, proteomics, and metabolomics** offer an immense opportunity to understand the effects of spaceflight on biological systems…"*

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



Omics Acquisition in Space



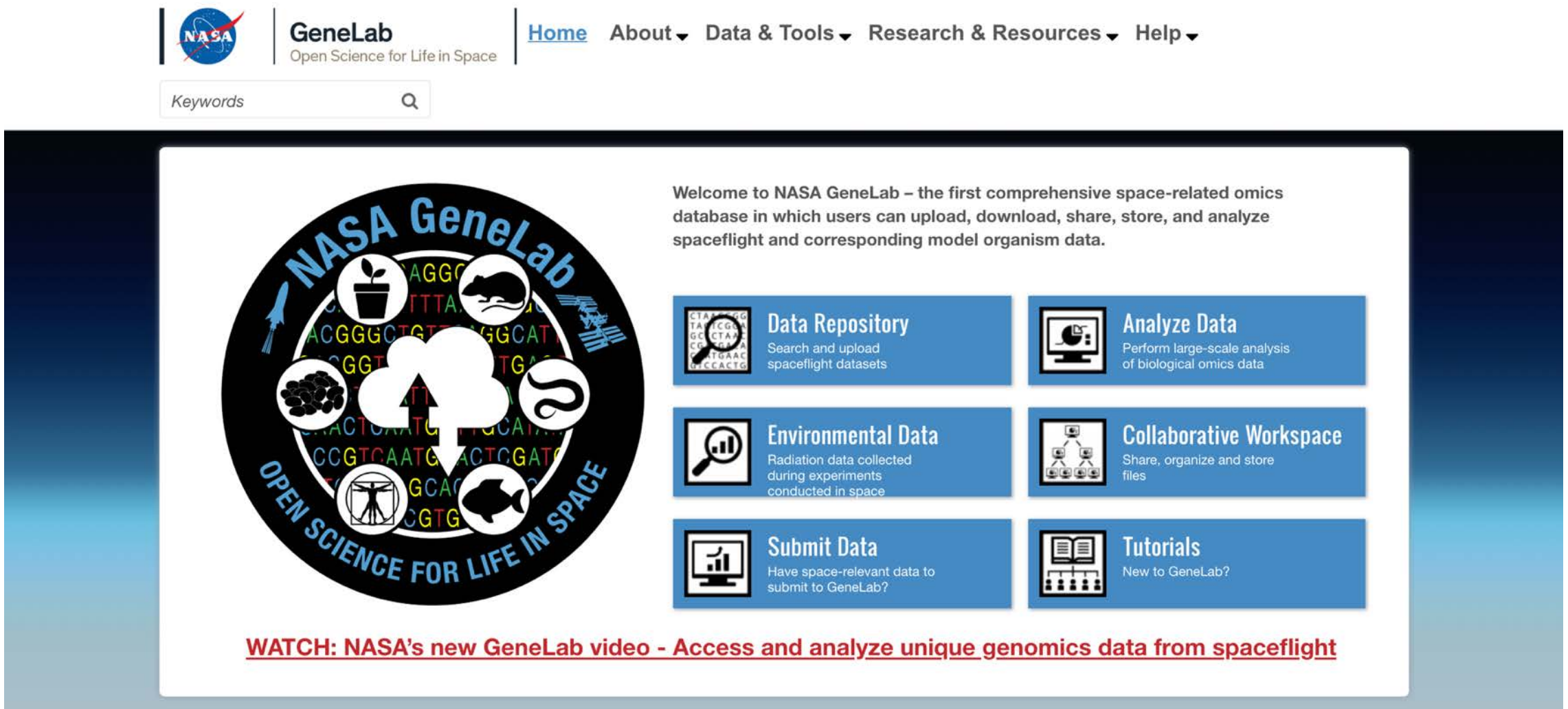
Sample Preparation Module

Oxford Nanopore MinION Gene Sequencer

Cepheid Smart Cycler qRT-PCR

GeneLab

- Mission: To enable scientific discovery and space exploration through multi-omics data-driven research.
- Currently funded by SLPSRA Space Biology program. Previously received funding from the ISS program.

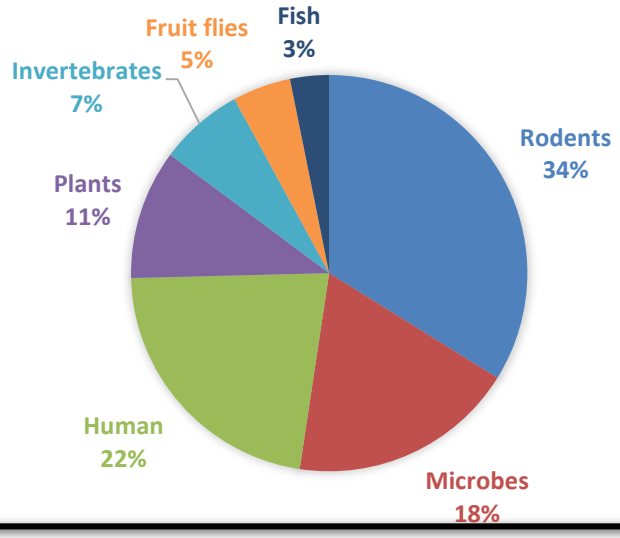


Overview: Database content

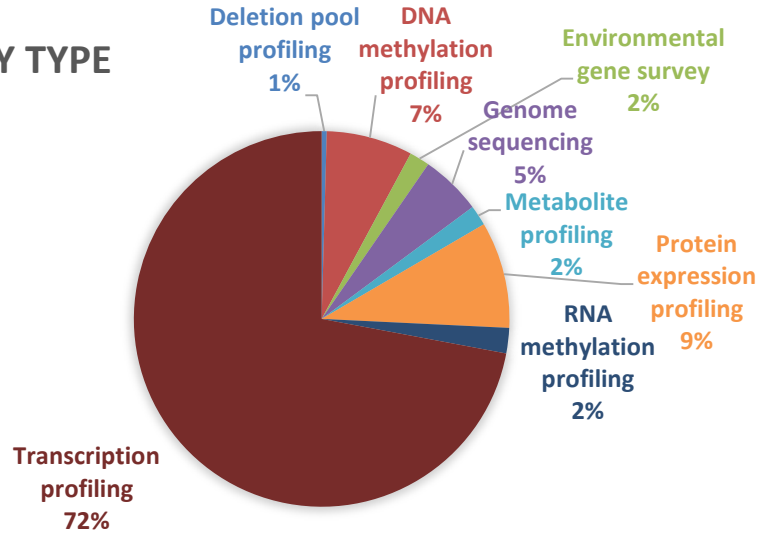
GeneLab contains decades of Space Biology omics data

STUDY TYPE

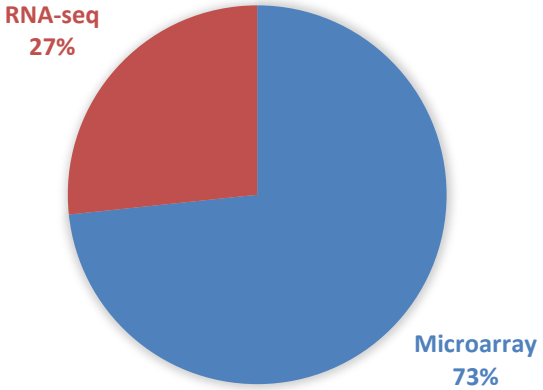
ORGANISM



ASSAY TYPE



TRANSCRIPTION PROFILING

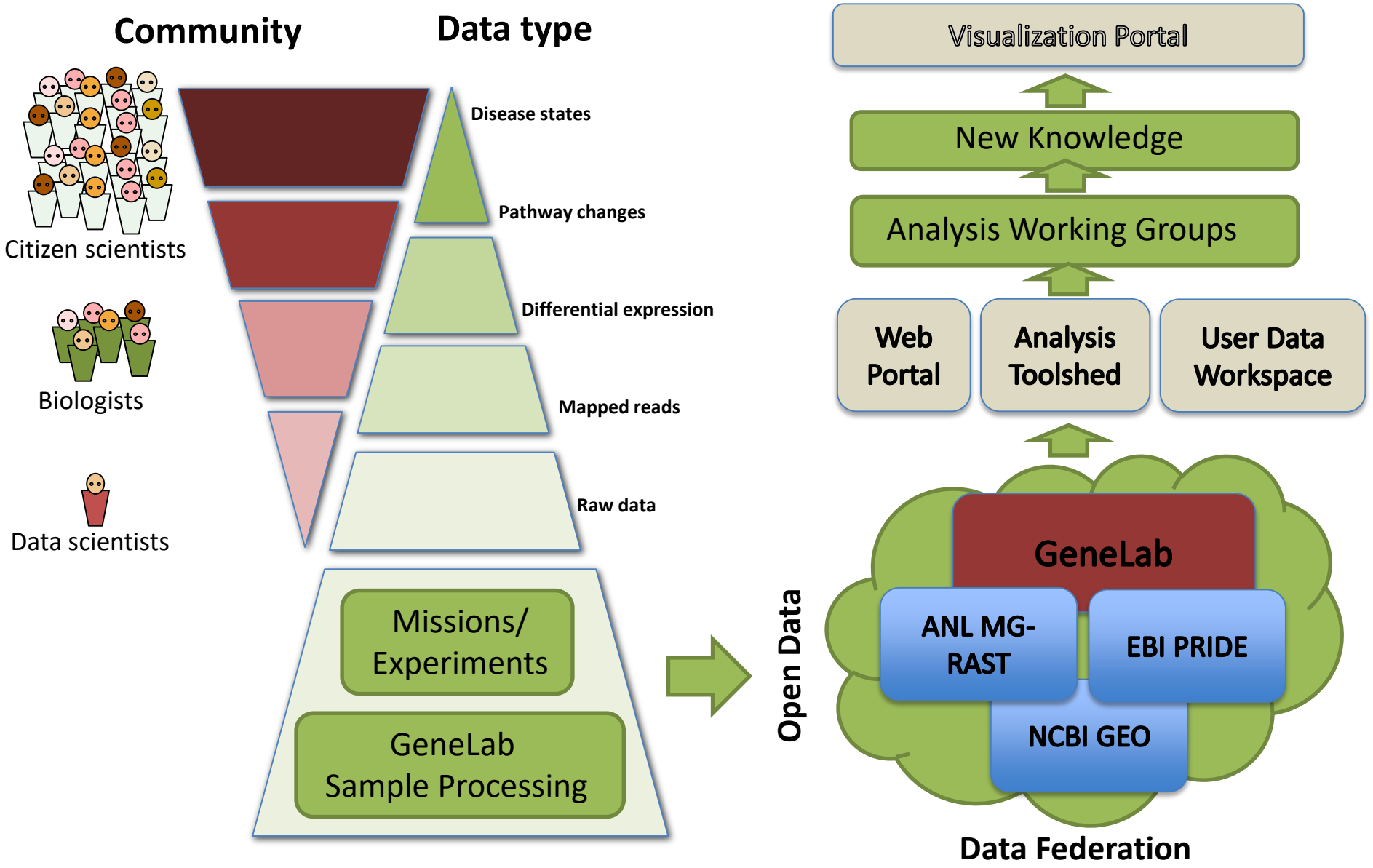


Total # of studies: 193 (6-11-2019)

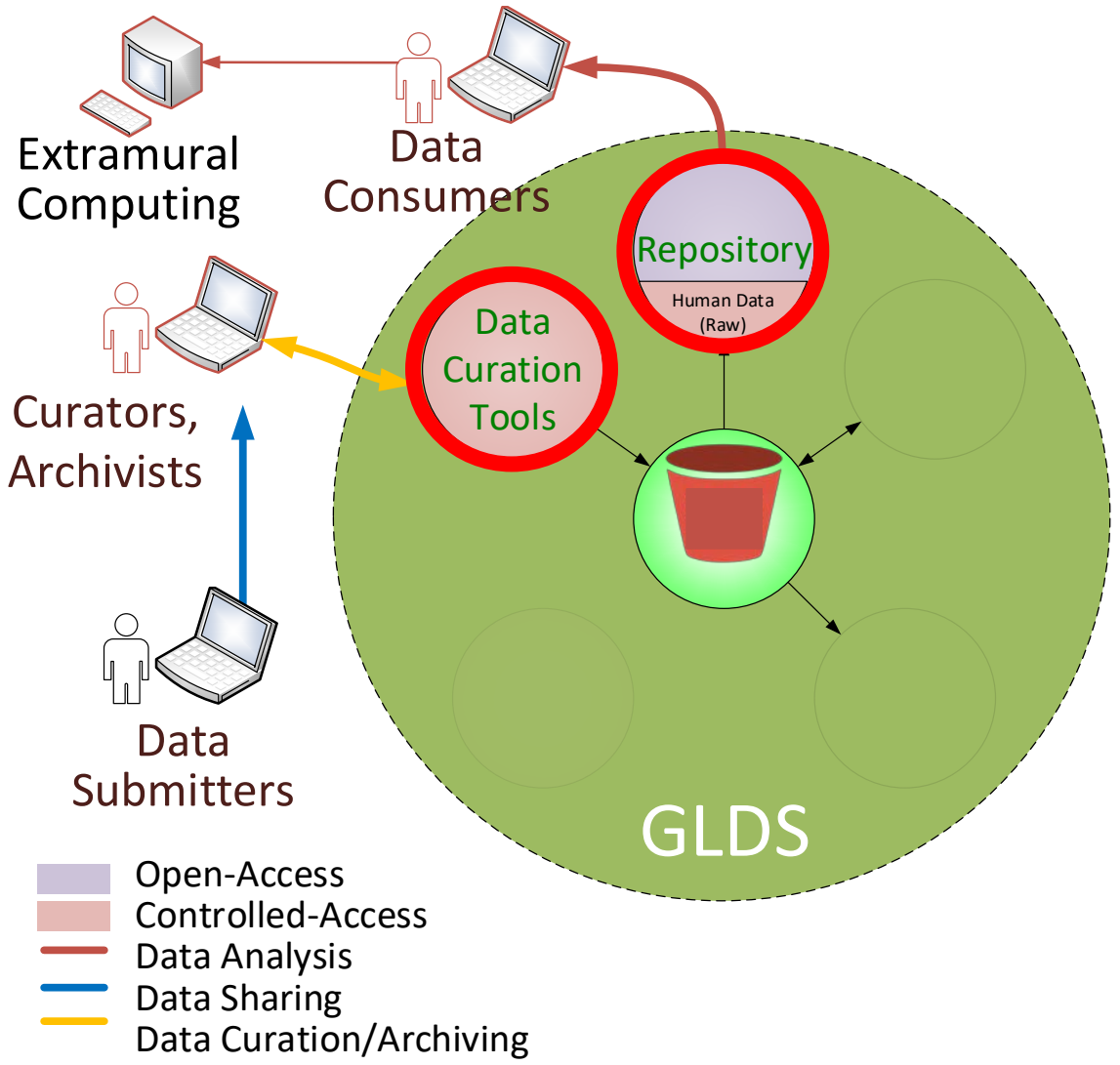
Talk outline

- **GeneLab background and motivation**
- **GeneLab Data System (GLDS) and datasets**
  - Data system architecture (AWS, search, federation)
  - Datasets, dataset metadata, and curation
  - Collaborative workspace
  - Cloud analysis platform
- **GeneLab community**
  - Analysis Working Groups (AWGs)
- **Increasing the scientific return on Space Biology investments**
  - Leveraging biomedical research
  - Aggregating Space Biology research
  - Increasing data volume from Space Biology research

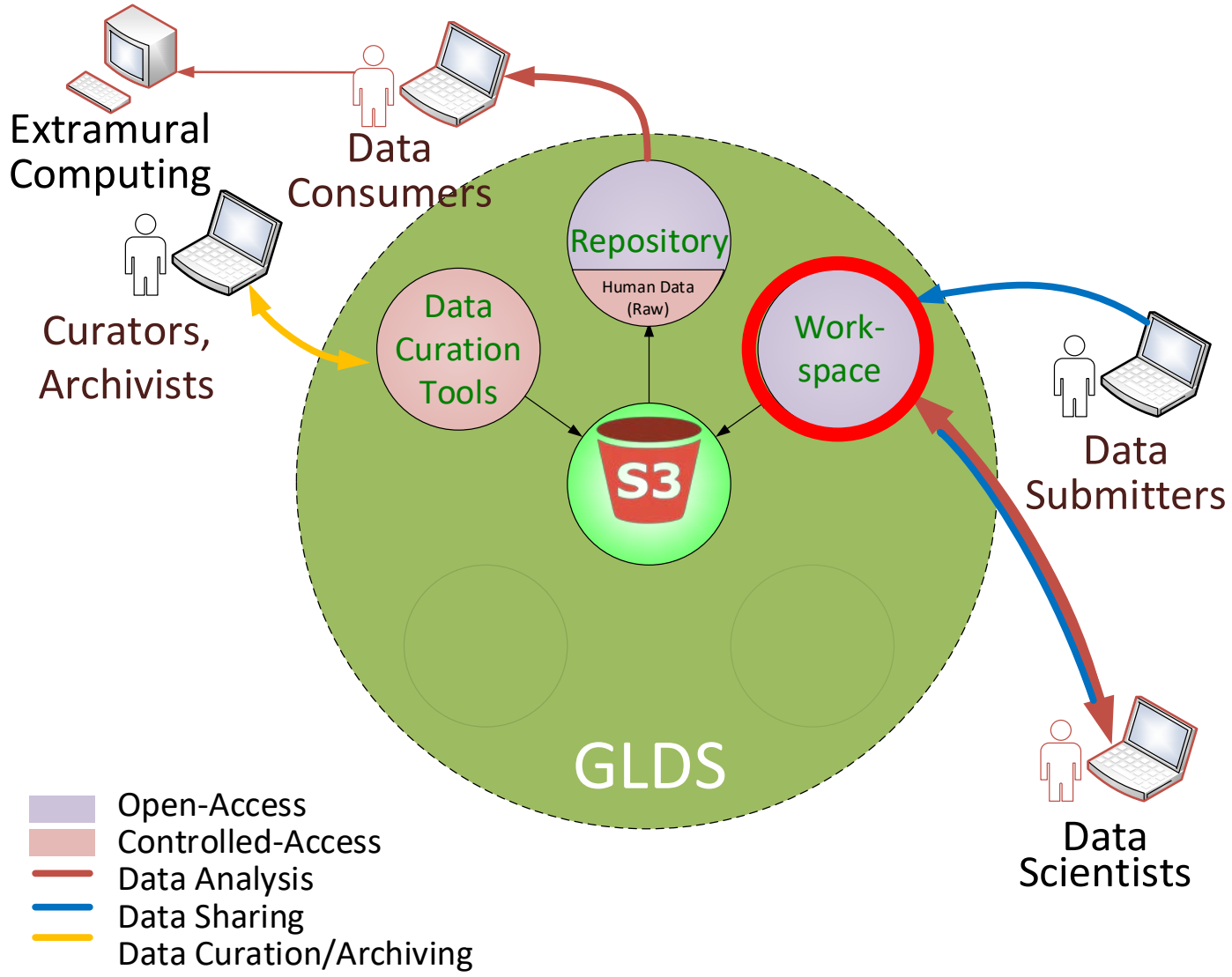
GeneLab Data Democratization



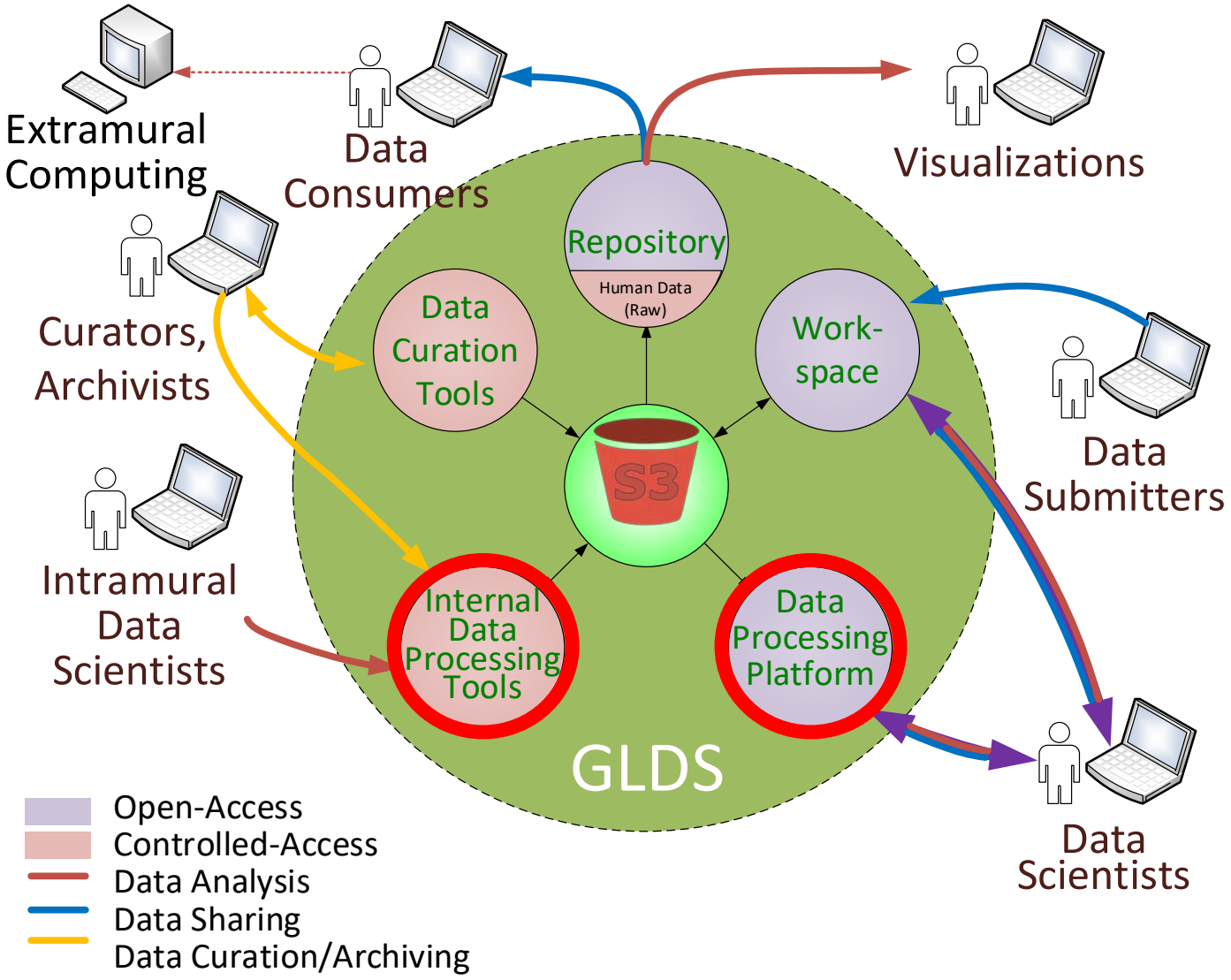
GLDS Systems Architecture Phase I



GLDS Systems Architecture
Phase II



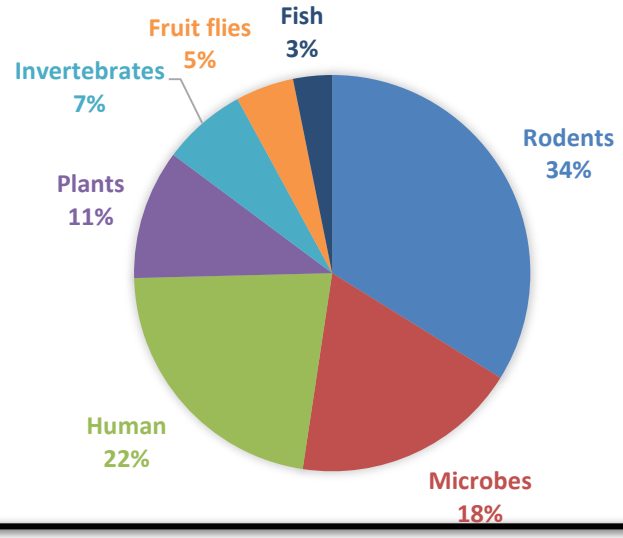
GLDS Systems Architecture Phase III



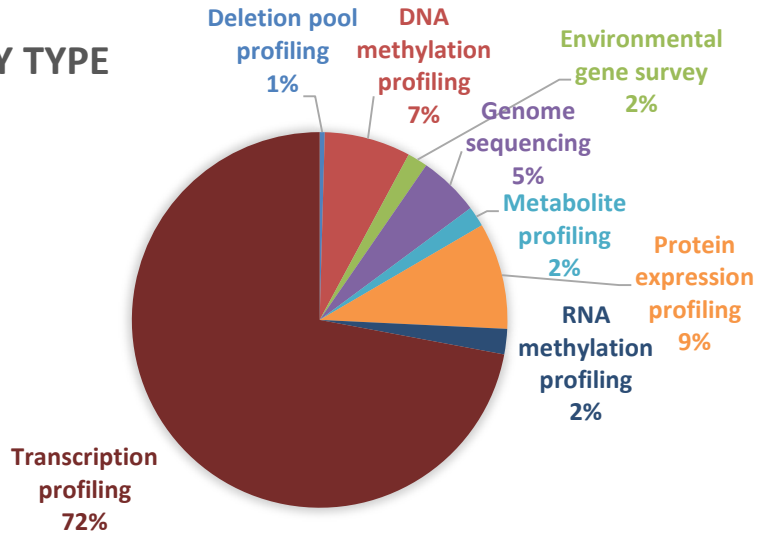
Overview: Database content

STUDY TYPE

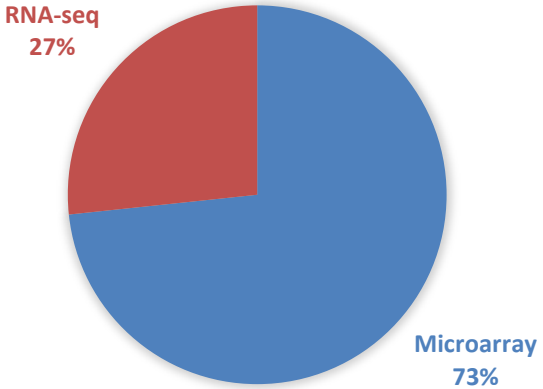
ORGANISM



ASSAY TYPE

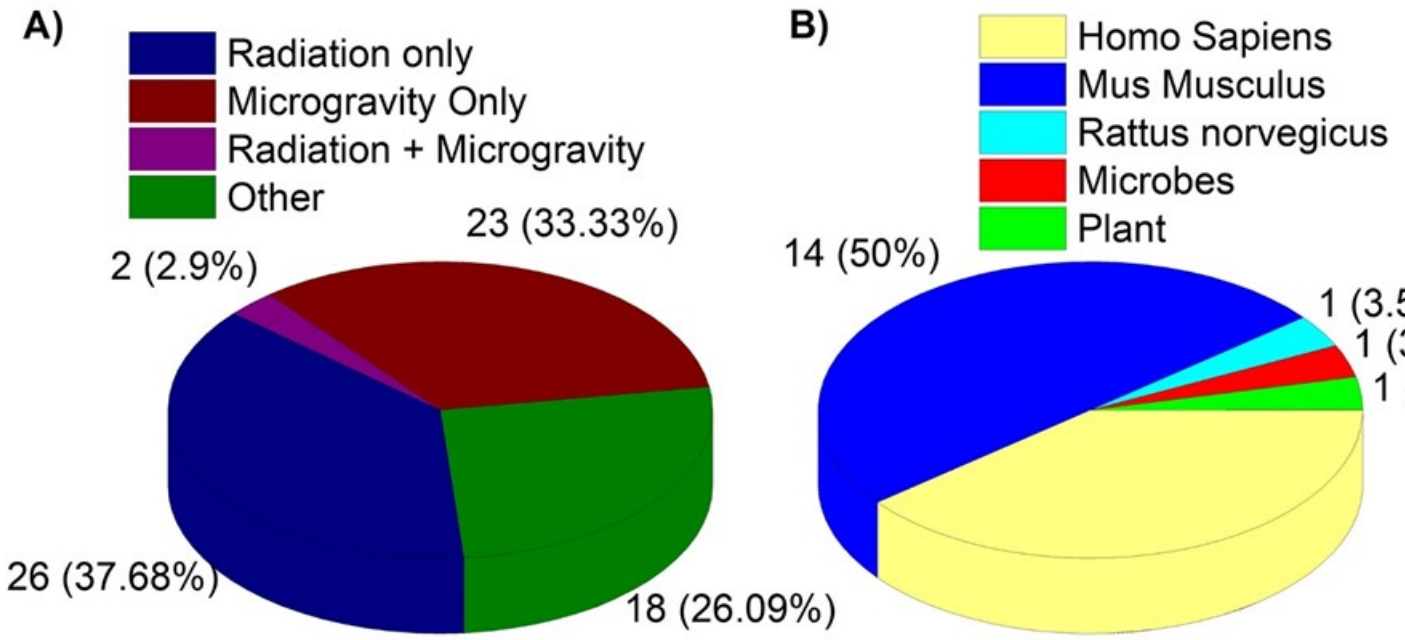


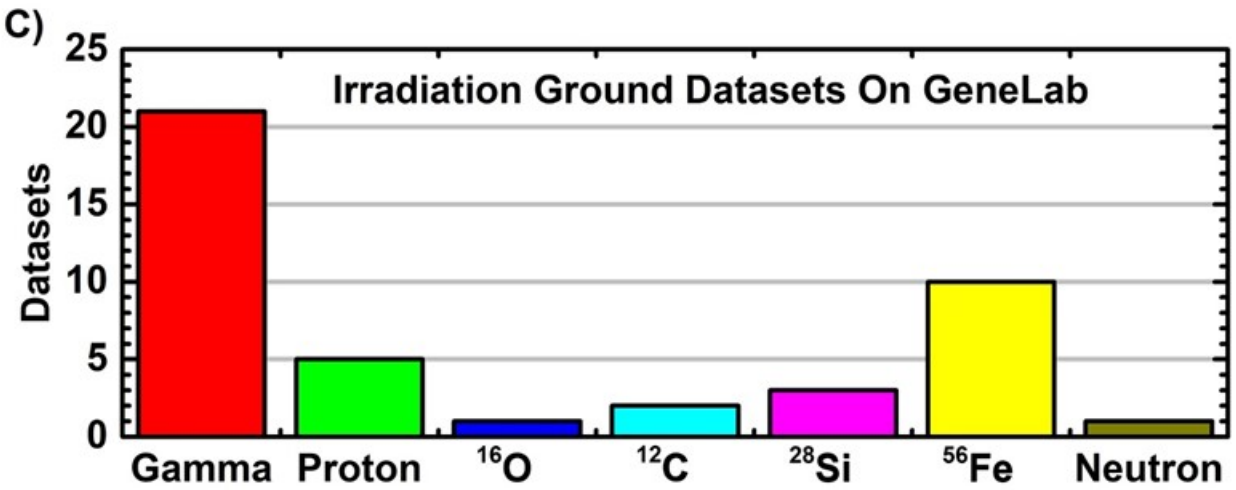
TRANSCRIPTION PROFILING

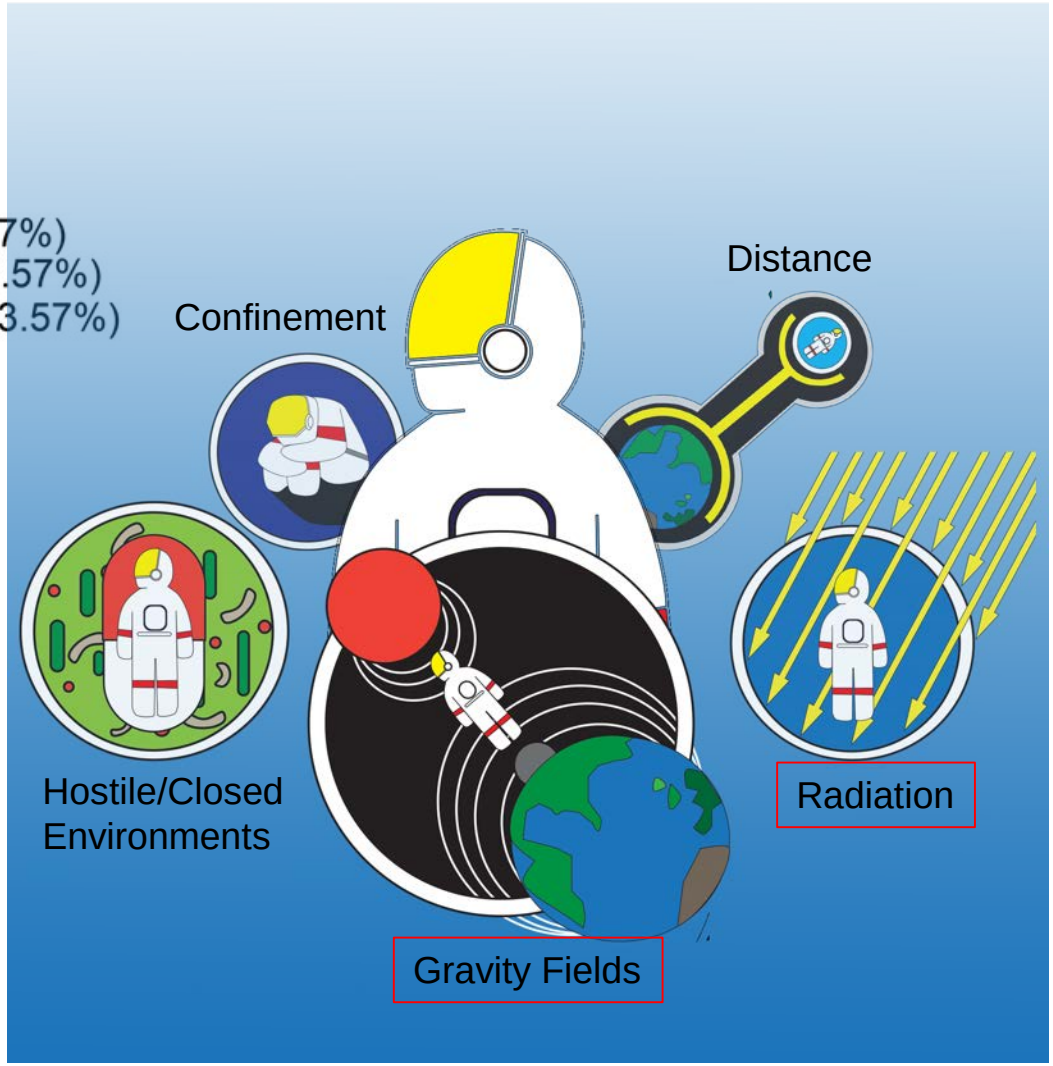


Total # of studies: 193 (6-11-2019)

69 Ground Data Sets: Radiation and simulated microgravity

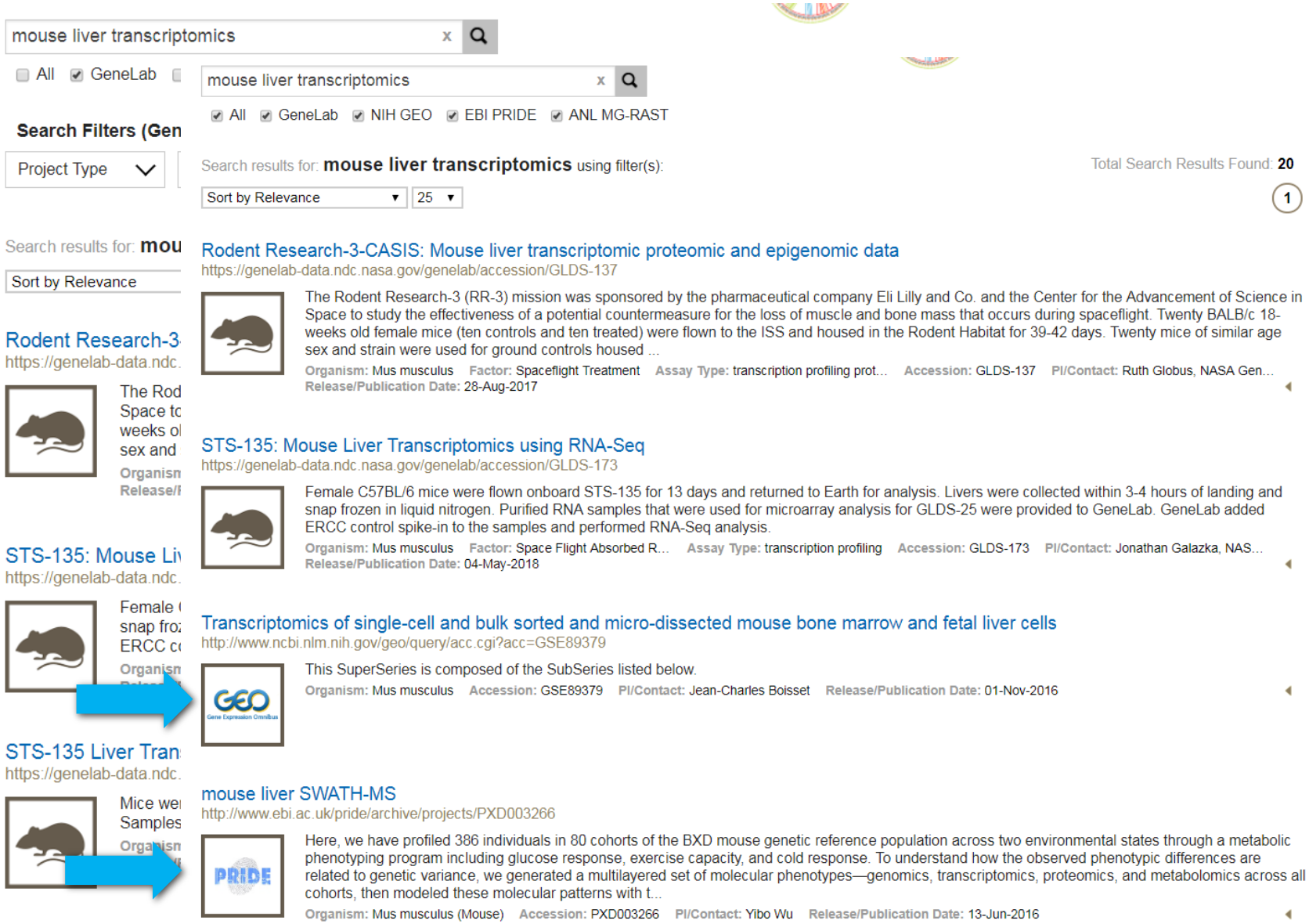






Beheshti et al., Radiation Research 2018

Searching the Repository



Environmental Metadata

GeneLab recognizes the need to include environmental data to spaceflight experiments. These include but not limited to:

- Radiation dosimetry (High and low LET)
- Temperature
- Humidity
- CO_2 levels
- Payload specific sensors: temperature, light cycle, humidity, CO_2 levels
  - We recommend radiation specific sensors for payloads
- Events in space – planned or unplanned*
- Launch and Return environment*

*Not currently in scope but will add if provided or requested by PI

This data
spacefli
- Min

Other environmental data:

LSDA →  → GLDS Repository

GLDS Phase 2 Repository

GeneLab
Open Science for Exploration

Home | Repository | Data | Tools | Environmental Data | Submit Data | Help | Workspace

Search Data

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

Search Filters (GeneLab Only)

Project Type | Factors | Organisms | Assay Type | Clear

NASA GeneLab Data System

NASA GeneLab expands scientists' access to experiments onboard the International Space Station that explore the molecular response of terrestrial biology to spaceflight environments. Our mission is to maximize the utilization of the valuable biological research resources aboard the International Space Station by collecting genomic, transcriptomic, proteomic, and metabolomic data known as "omics".

Data Repository

The data repository hosts space biology and space-related datasets funded by multiple space agencies around the world. Use the search options above to filter through the GeneLab repository or to search across several databases.

Collaborative Workspace

GeneLab has customized a workspace for file sharing and access to data analysis tools. To access the workspace, users will need to create an account. Data analysis tool integration is ongoing and more tools will be available in future releases.

https://genelab-data.ndc.nasa.gov/genelab/

Environmental data cont.

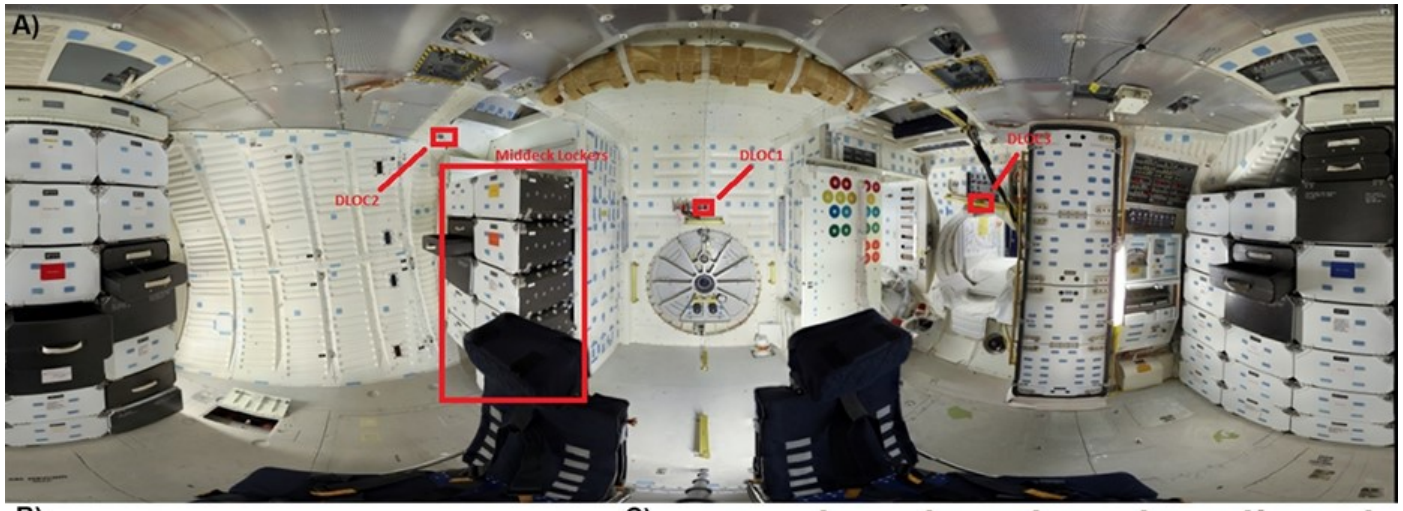
Environmental Metadata – currently only have radiation data available for STS and BION-M1 mission datasets

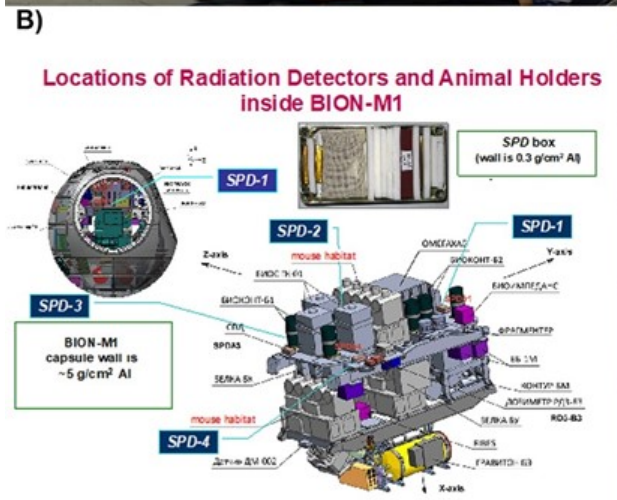
Radiation Metadata for GLDS Studies

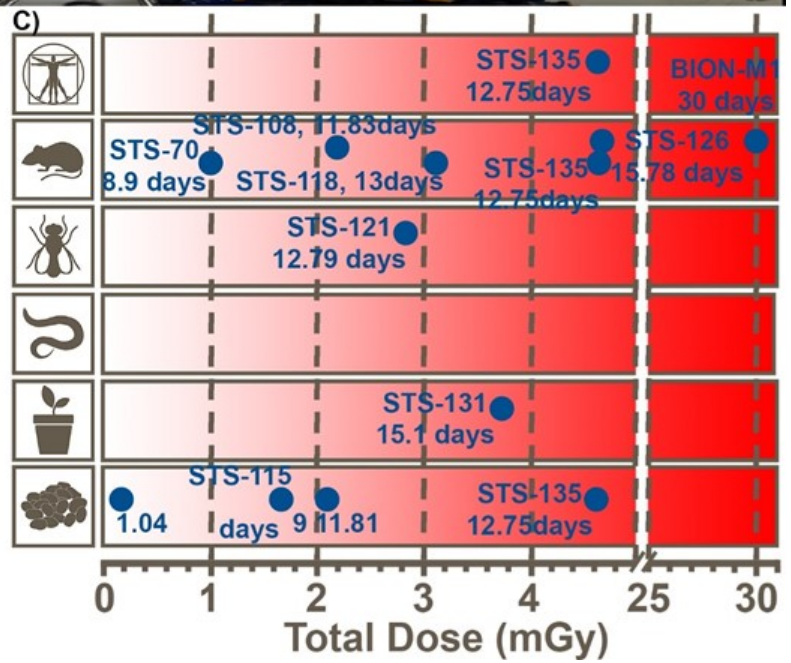
DESCRIPTION | PROTOCOLS | SAMPLES | ASSAYS | PUBLICATIONS | STUDY FILES

Source Name	Sample Name	Exposure Duration	Maximum Total Absorbed Dose	Minimum Total Absorbed Dose	Com
Mouse Tissue	AEM Control 18	12.76 day			
Mouse Tissue	AEM Control 28	12.76 day			
Mouse Tissue	AEM Control 32	12.76 day			
Mouse Tissue	AEM Control 36	12.76 day			
Mouse Tissue	AEM Control 44	12.76 day			
Mouse Tissue	Flight 52	12.76 day	5.46 mGy	3.26 mGy	
Mouse Tissue	Flight 58	12.76 day	5.46 mGy	3.26 mGy	
Mouse Tissue	Flight 60	12.76 day	5.46 mGy	3.26 mGy	
Mouse Tissue	Flight 64	12.76 day	5.46 mGy	3.26 mGy	
Mouse Tissue	Flight 66	12.76 day	5.46 mGy	3.26 mGy	

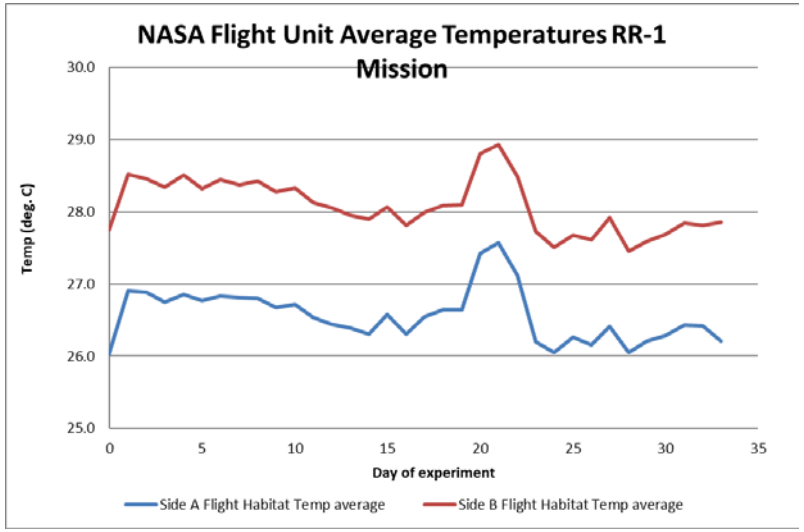
Radiation Dosimetry for STS samples (ISS to follow)

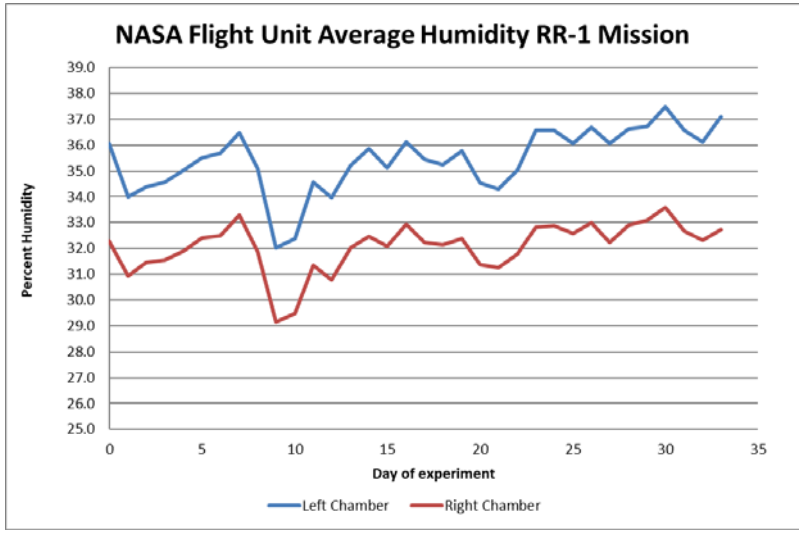






Beheshti et al., Radiation Research 2018



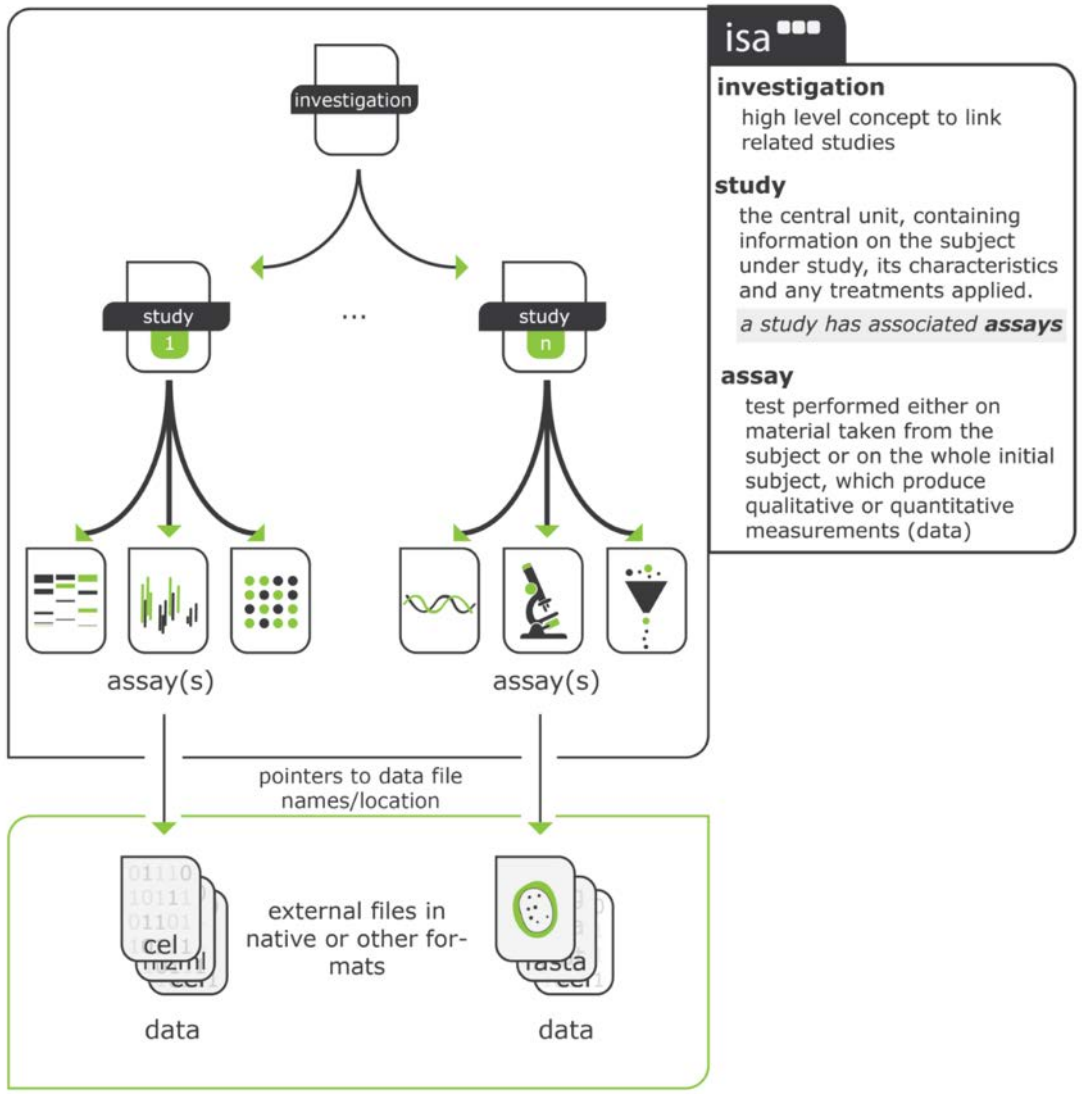


STS = Space Transportation System (Shuttle Program)

GeneLab Data and Metadata model

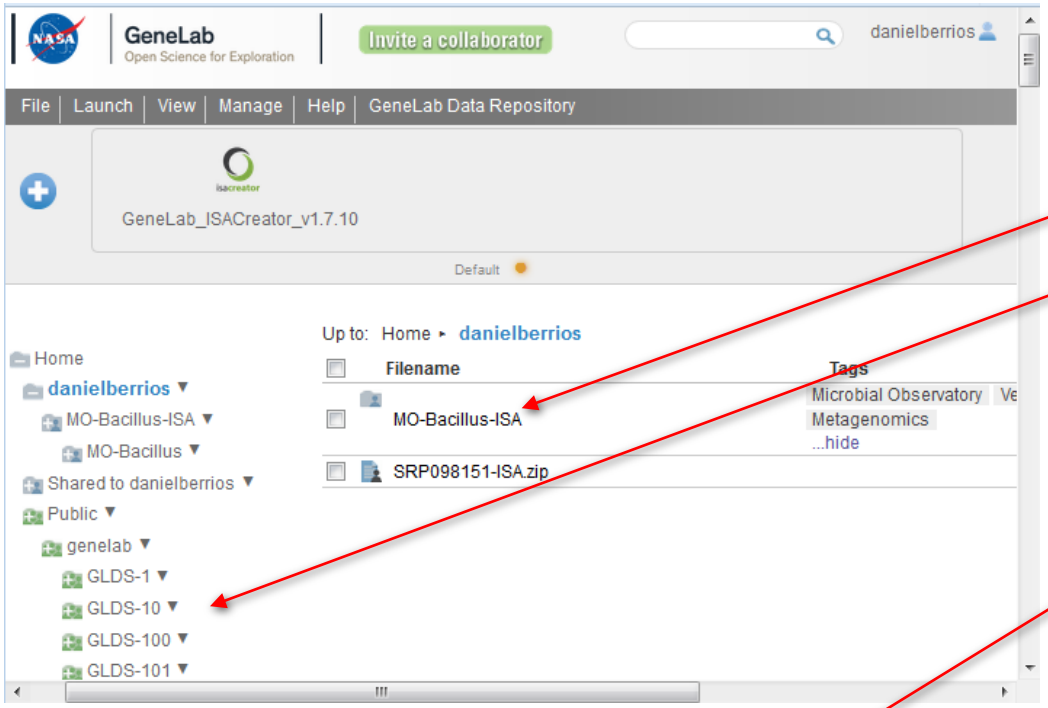
- **GeneLab utilizes the Investigation Study Assay (ISA) model for organizing metadata**

- **GeneLab participates in ontology groups to help standardize spaceflight ontology**



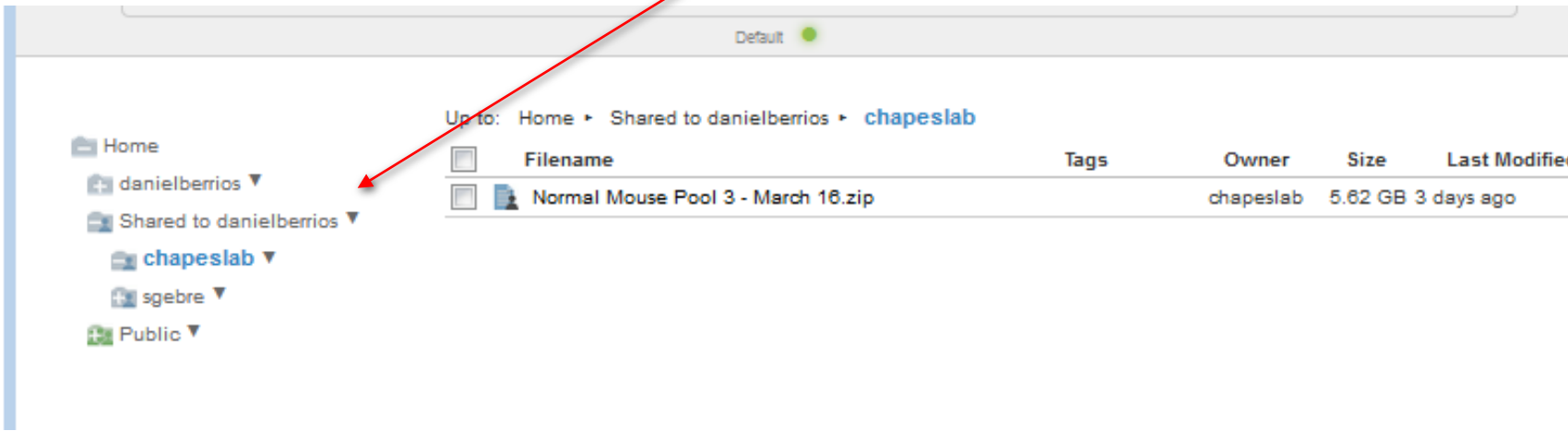
GeneLab Workspace

- Sequencing data can be massive
- Terabytes
- Moving data requires special considerations
- The GeneLab allows user to upload, download and share raw and processed data

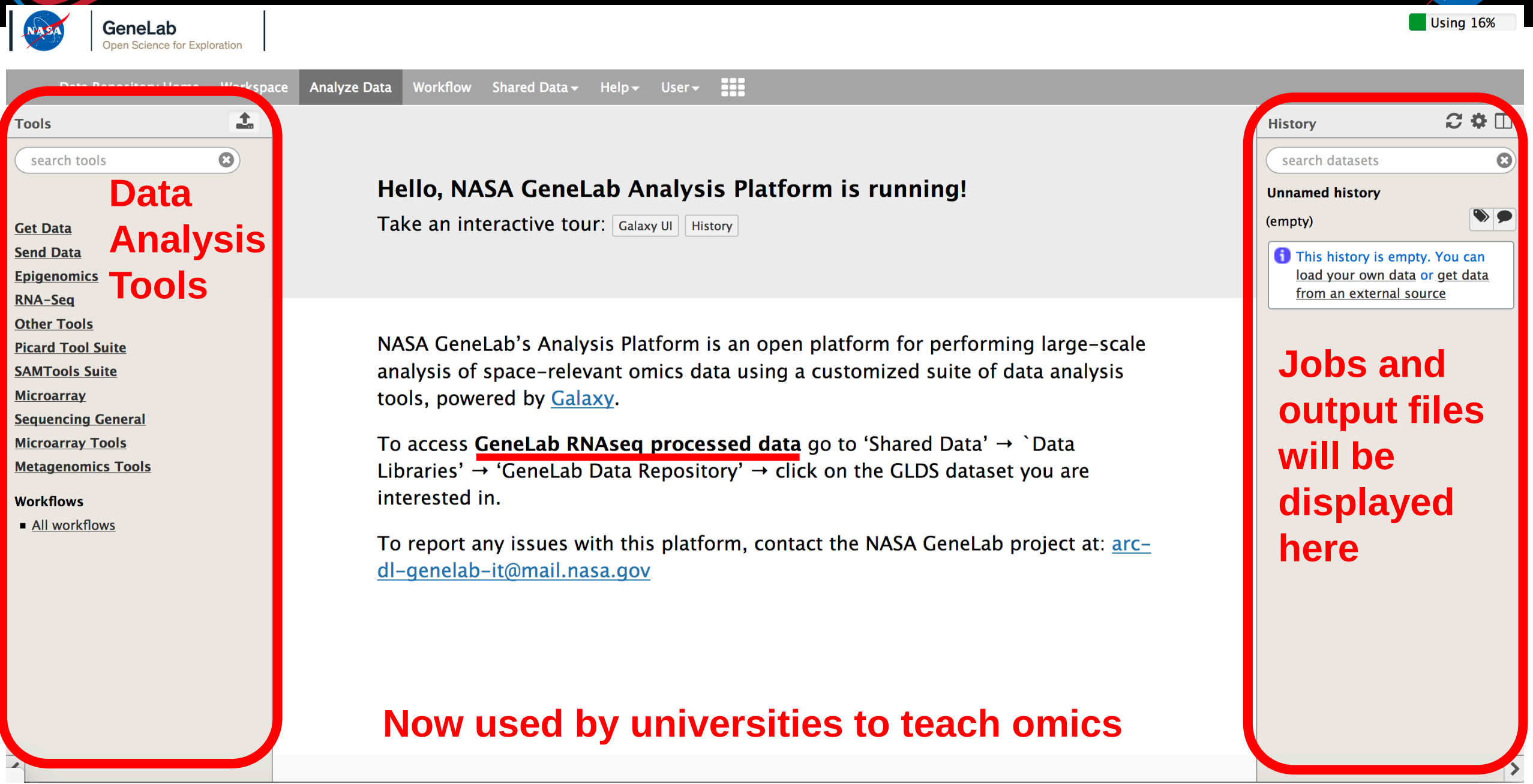


- User files
- Public repository files

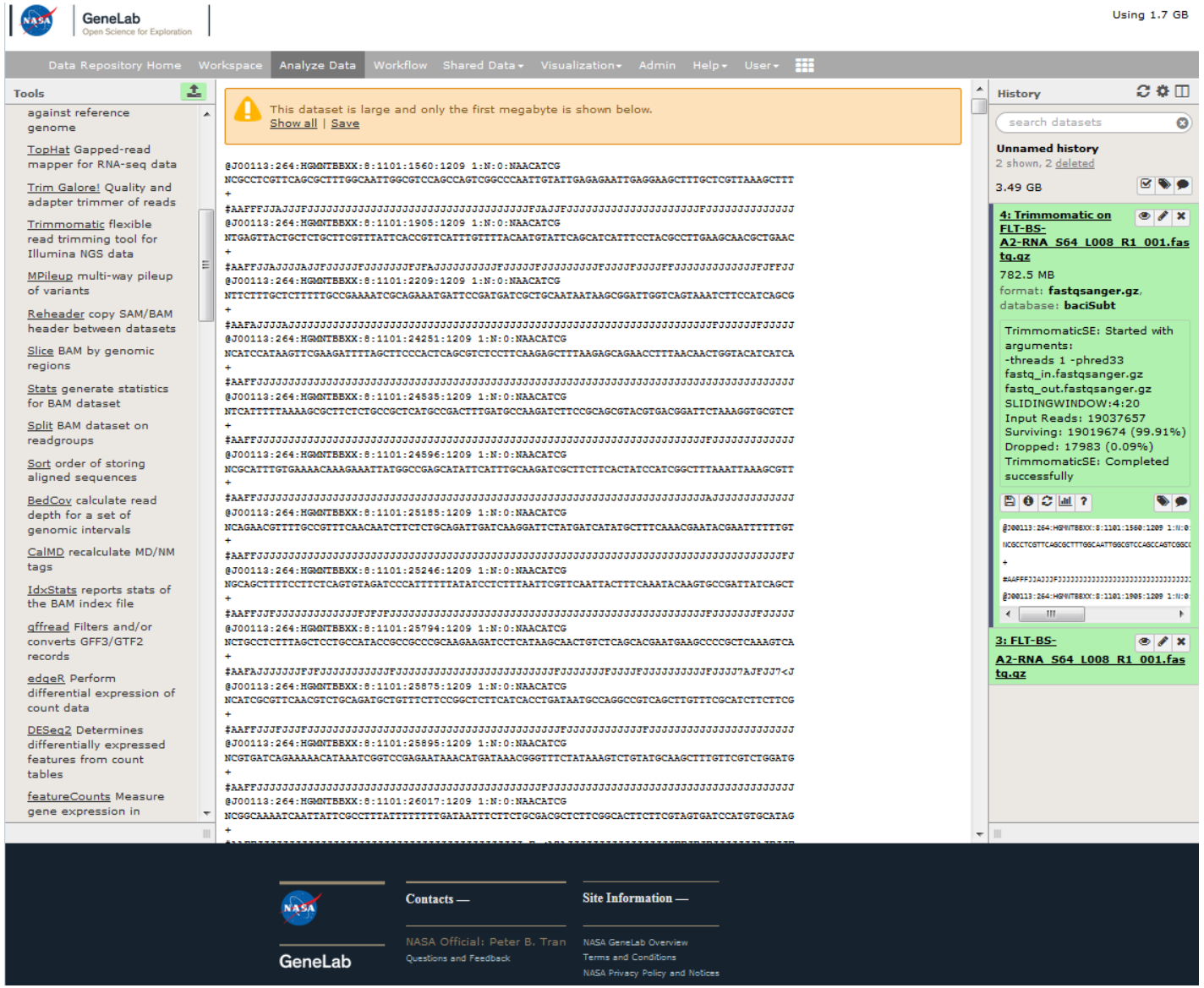
- Privately Shared files
- Files submitted for Public Sharing (Repository)



GeneLab Analysis Platform

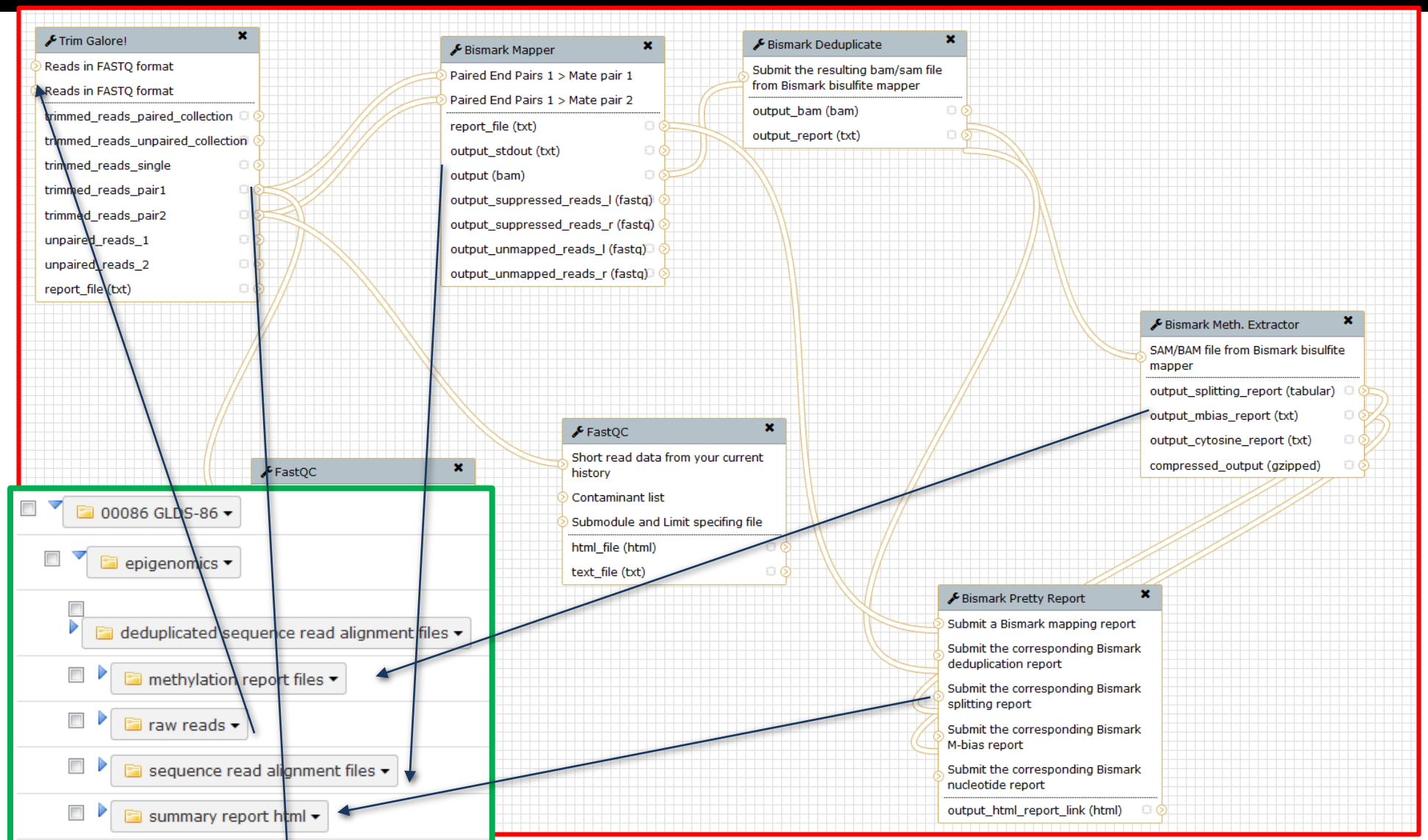


GLDS Public Data Processing Platform



- Galaxy Server
- Galaxy Toolshed (subset)

GLDS-Galaxy Data Processing Workflows & Data Libraries



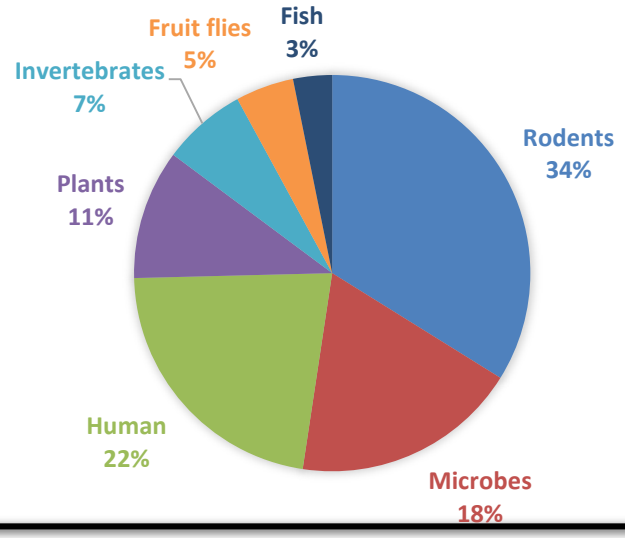
Talk outline

- **GeneLab background and motivation**
- **GeneLab Data System (GLDS) and datasets**
  - Data system architecture (AWS, search, federation,)
  - Datasets, dataset metadata, and curation
  - Collaborative workspace
  - Cloud analysis platform
- **GeneLab community**
  - Analysis Working Groups (AWGs)
- **Increasing the scientific return on Space Biology investments**
  - Leveraging biomedical research
  - Aggregating Space Biology research
  - Increasing data volume from Space Biology research
- **Future efforts and conclusion**

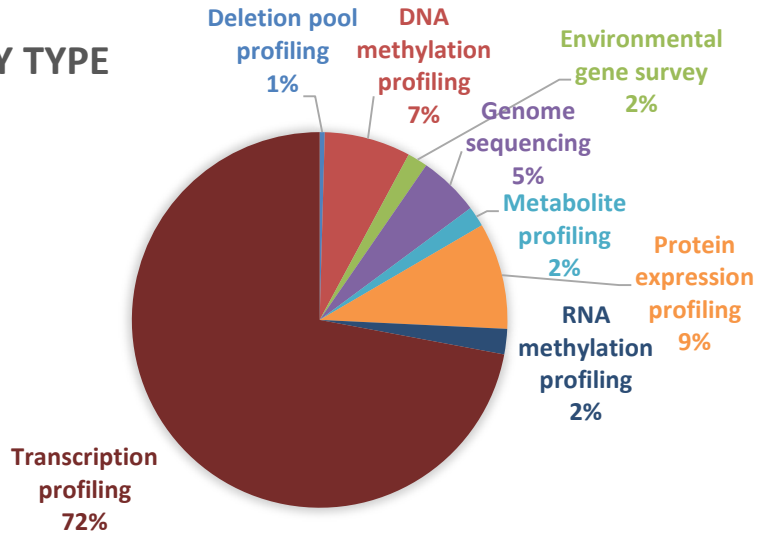
How can we increase user engagement and knowledge generation?

STUDY TYPE

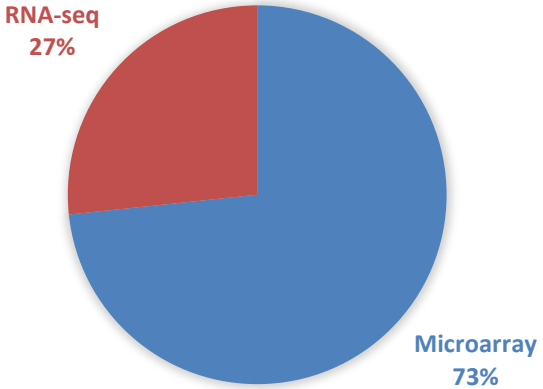
ORGANISM



ASSAY TYPE



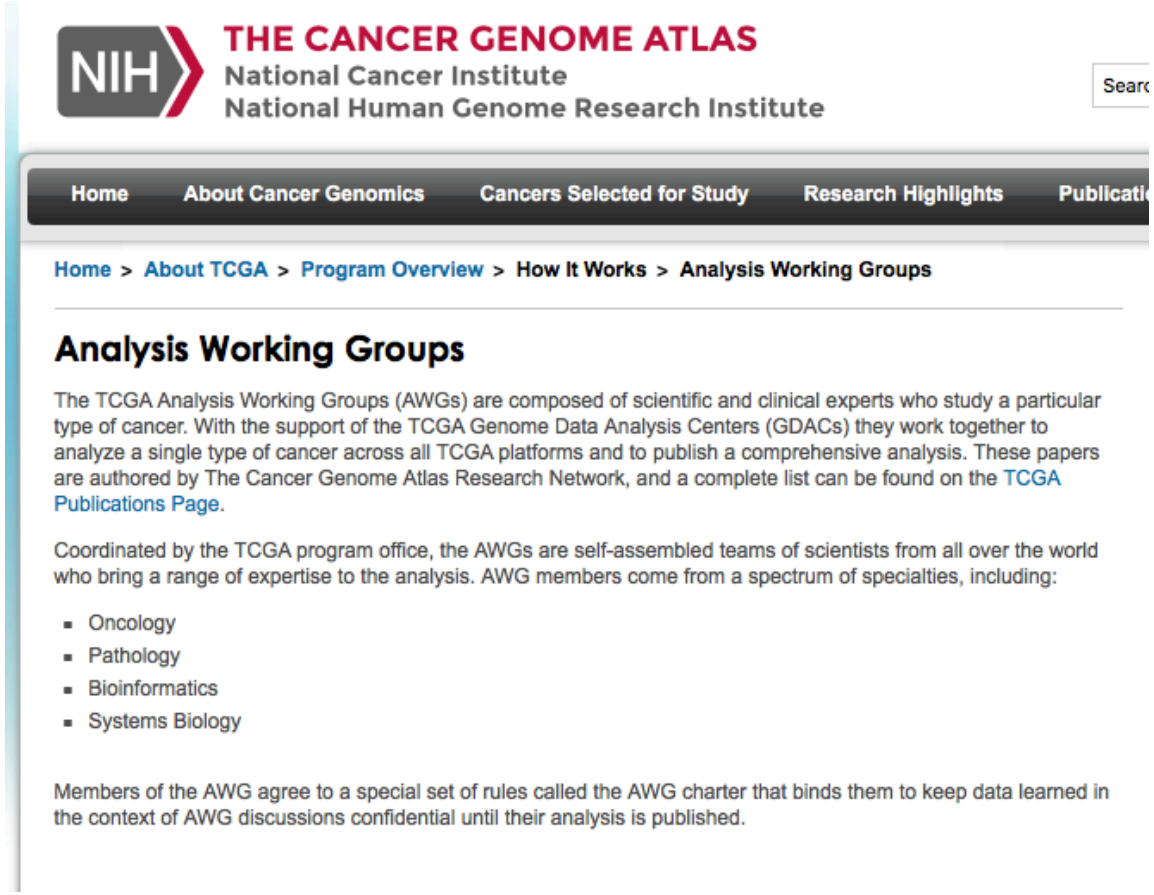
TRANSCRIPTION PROFILING



Total # of studies: 193 (6-11-2019)

GeneLab Analysis Working Groups (AWGs)

- **Modeled on the NIH TCGA Analysis Working Groups**
- **Analyze a single type of cancer across all TCGA platforms and publish comprehensive analyses**
- **Have generated very high profile publications**



THE CANCER GENOME ATLAS
National Cancer Institute
National Human Genome Research Institute

Home | About Cancer Genomics | Cancers Selected for Study | Research Highlights | Publicati

Home > About TCGA > Program Overview > How It Works > Analysis Working Groups

Analysis Working Groups

The TCGA Analysis Working Groups (AWGs) are composed of scientific and clinical experts who study a particular type of cancer. With the support of the TCGA Genome Data Analysis Centers (GDACs) they work together to analyze a single type of cancer across all TCGA platforms and to publish a comprehensive analysis. These papers are authored by The Cancer Genome Atlas Research Network, and a complete list can be found on the TCGA Publications Page.

Coordinated by the TCGA program office, the AWGs are self-assembled teams of scientists from all over the world who bring a range of expertise to the analysis. AWG members come from a spectrum of specialties, including:

- Oncology
- Pathology
- Bioinformatics
- Systems Biology

Members of the AWG agree to a special set of rules called the AWG charter that binds them to keep data learned in the context of AWG discussions confidential until their analysis is published.

An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics
Cell. **Volume 173** Issue 2: p400-416 Read the full article

A Pan-Cancer Analysis of Enhancer Expression in Nearly 9000 Patient Samples
Cell. **Volume 173** Issue 2: p386-399 Read the full article

Comprehensive Characterization of Cancer Driver Genes and Mutations
Cell. **Volume 173** Issue 2: p371-385 Read the full article

Pathogenic Germline Variants in 10,389 Adult Cancers
Cell. **Volume 173** (2): p355-370 Read the full article

Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation
Cell. **Volume 173** Issue 2: p338-354 Read the full article

GeneLab AWG: Charter

The purpose of the GeneLab Analysis Working Groups (AWGs) is to optimize the processing of raw omics data from the GeneLab repository in order to maximize the gain of new knowledge from such complex datasets. The AWGs will also assess and improve the effectiveness of the GeneLab Data System (GLDS) through intensive utilization of the analytics to be deployed therein.

The AWGs aspire to scientific excellence, and participation in AWGs is strictly on a volunteer basis. The primary activity of each AWG is to establish analytical processes to generate higher order data from data housed in the GLDS with relevance to one or more specific application areas. For example, a mammalian AWG would analyze all data relevant to mammalian systems (human, mouse, rat, etc.). The GeneLab AWGs aspire to yield:

Canonical data analysis pipelines. The interpretation of analysis results is too often confounded by variability in data-processing protocols. This can make comparison of results between studies even more difficult. By developing canonical analysis pipelines, the AWGs will help speed the harmonization of results across space biology experiments

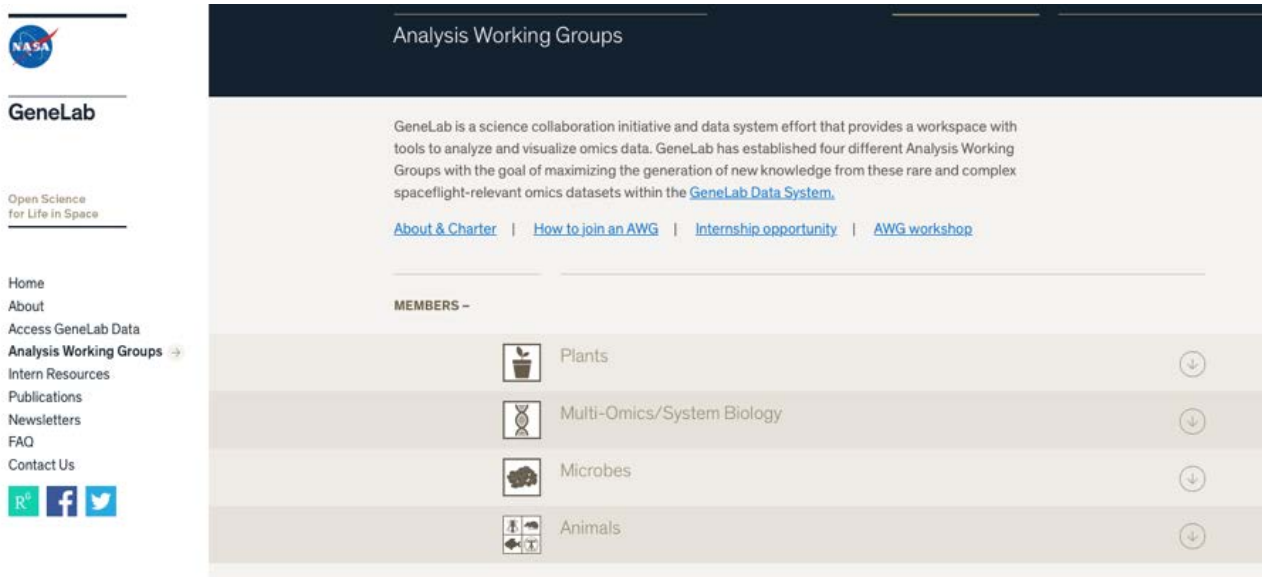
Processed ("higher-order") space biology data. The AWGs will work with GeneLab staff to generate processed data using the canonical analysis pipelines. This higher-order data will be made available to the broader scientific community through the GLDS. This data will serve as a reference to the community on use of the pipelines, and will be fully reproducible using the GLDS and referenced pipelines. Optimal strategies regarding the display of processed data to the scientific community will also be discussed.

Critiques of the GLDS. The value of the GLDS to the scientific community is best judged through extensive use of the systems. As the AWGs will be using the GLDS intensively (see 1 and 2 above), they will be poised to give valuable feedback to GeneLab on its strengths and weaknesses.

Scientific communication of available analyses and interpretations*.* The AWGs are expected to communicate the development and application of the canonical data analysis pipelines with the broader scientific community. Such communications could take the form of self-published white papers, newsletters, or peer-review publications and conference presentations. These communications will also serve to catalyze the generation of new hypotheses for further experimentation.

AWG Recruitment

- Reached out to members of Space Biology community, scientist at different NASA centers, outside domain experts
- Sent open invitation through GeneLab newsletter
- Posted open invitation on GeneLab website (still active)
- Continue to recruit and accept new members

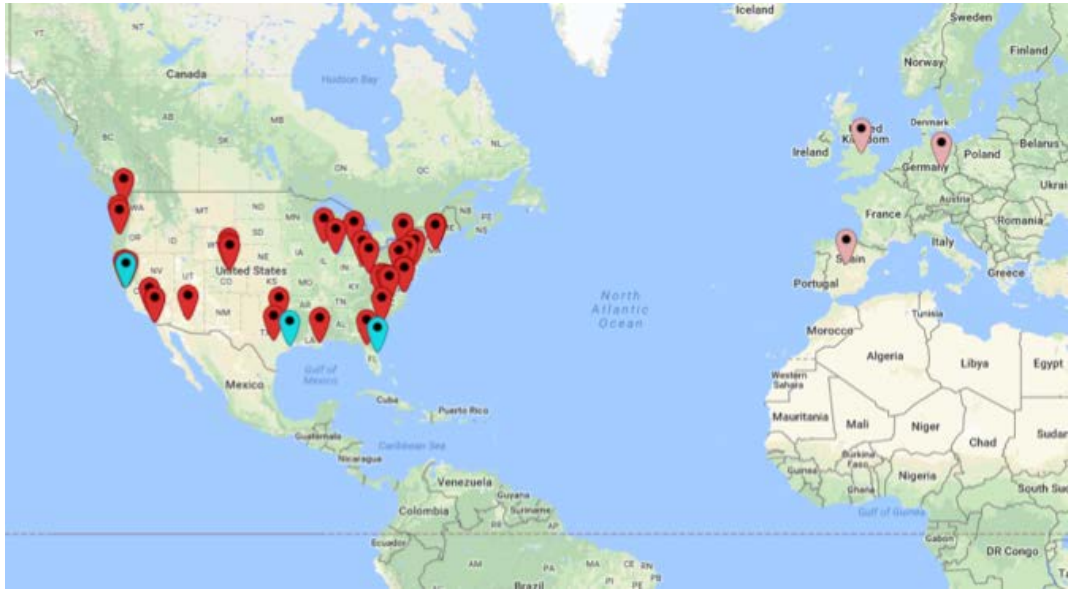


GeneLab is recruiting members for Analysis Working Groups to analyze NASA omics data!

NASA's GeneLab Project is recruiting investigators, bioinformaticians, graduate students, and postdocs to participate in Analysis Working Groups (AWGs) that will investigate specific subsets of omics data from experiments conducted onboard the International Space Station, the Space Shuttle, as well as ground-based research with relevance to spaceflight (e.g. radiation or unloading/weightlessness).

What are GeneLab AWGs?

- **>100 individuals**
- **4 Groups: Plants, Microbes, Animals, Multi-omics**
- **Monthly meetings**
- **Deliverables:**
  - Consensus pipelines for primary analysis of data (Microarray, RNASeq, Bisulfite sequencing, Proteomics, 16S metagenomics, Whole genome metagenomics)
  - Recommendations for visualization of data



Talk outline

- **GeneLab background and motivation**
- **GeneLab Data System (GLDS) and datasets**
  - Data system architecture (AWS, search, federation,)
  - Datasets, dataset metadata, and curation
  - Collaborative workspace
  - Cloud analysis platform
- **GeneLab community**
  - Analysis Working Groups (AWGs)
- **Increasing the scientific return on Space Biology investments**
  - Leveraging biomedical research
  - Aggregating Space Biology research
  - Increasing data volume from Space Biology research

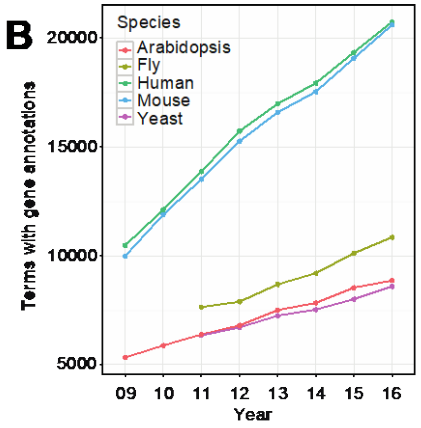
GeneLab contains “legacy” data back to the Space Shuttle era

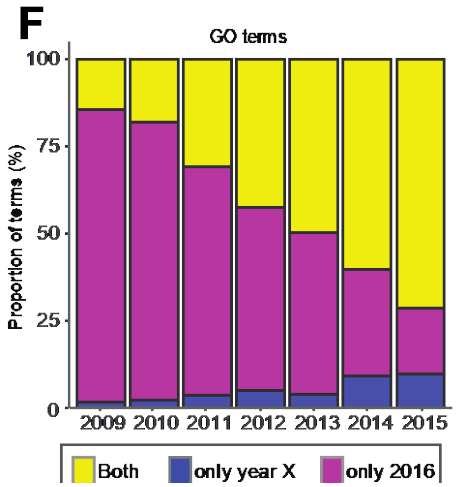
GLDS	Release date	Flight	Flight date	Title	Organism	Platform	Publication
GLDS-21	2008	STS-108	2001	Effects of spaceflight on murine skeletal muscle gene expression	Mouse	Affymetrix	Effects of spaceflight on murine skeltal muscle gene expression. Allen et al. 2008. J Appl Physiol.
GLDS-25	2015	STS-135	2011	STS-135 Liver Transcriptomics	Mouse	Affymetrix	Spaceflight activates lipotoxic pathways in mouse liver. Jonscher et al. 2016. Plos One.
GLDS-87	2016	STS-135	2011	Spaceflight effects on the mouse retina: Histological, gene expression and epigenetic changes after flight on STS-135	Mouse	Affymetrix	Molecular effects of spaceflight in the mouse eye after space shuttle missions STS-135. Theriot & Zanello. 2014. Gravitational and Space Research.
GLDS-4	2010	STS-118	2007	Microarray Analysis of Space-flown Murine Thymus Tissue	Mouse	Affymetrix	Microarray analysis of spaceflown murine thymus tissue reveals changes in gene expression regulating stress and glucocorticoid receptors. Lebsack et al. 2010. J Cell Biochem.

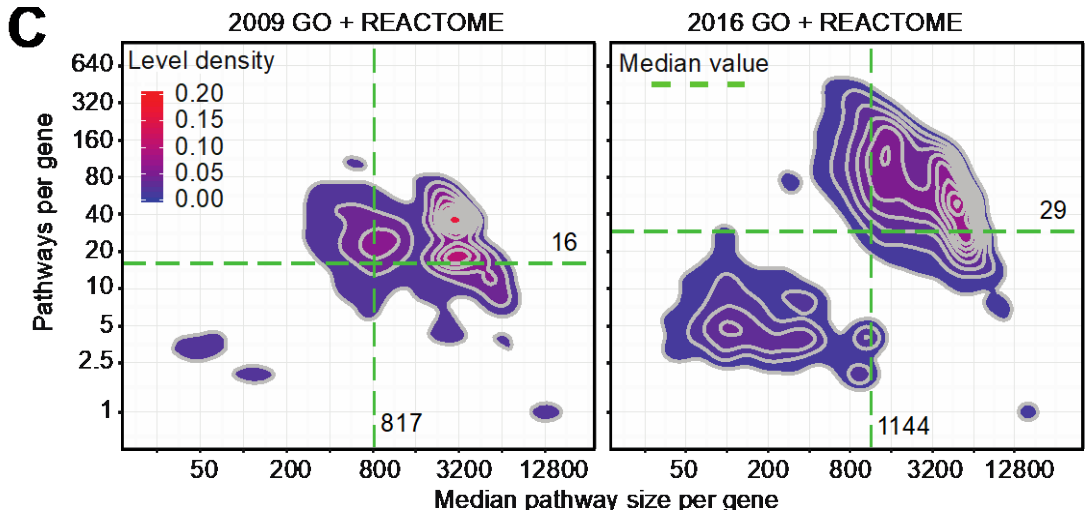
- Data has been analyzed and published.
- What’s the value in reanalyzing?

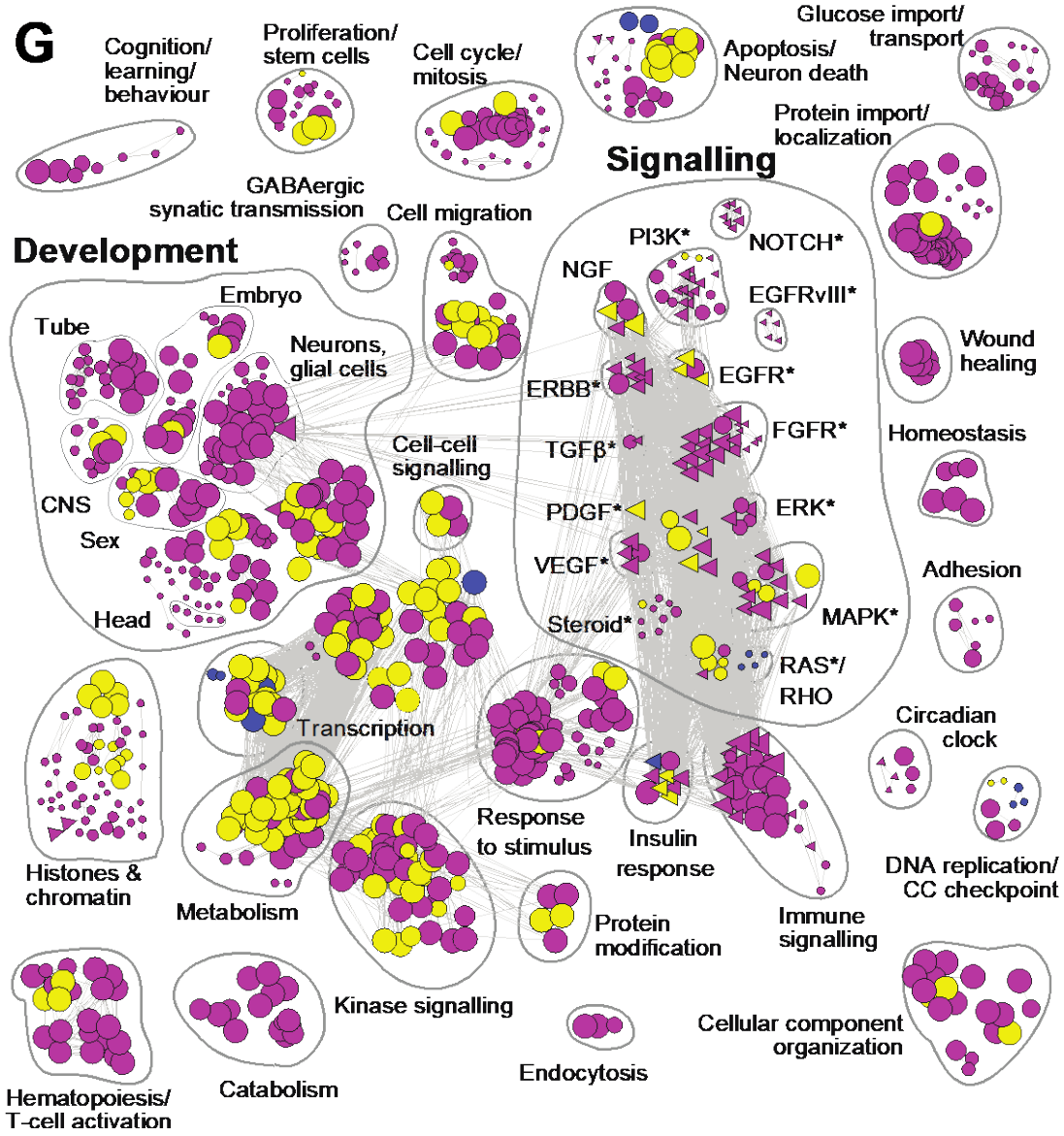
Value of Data Reanalysis

Impact of knowledge accumulation on pathway enrichment analysis. Wadi L, Meyer M, Weiser J, Stein L, Reimand J. Nature Methods. 2016









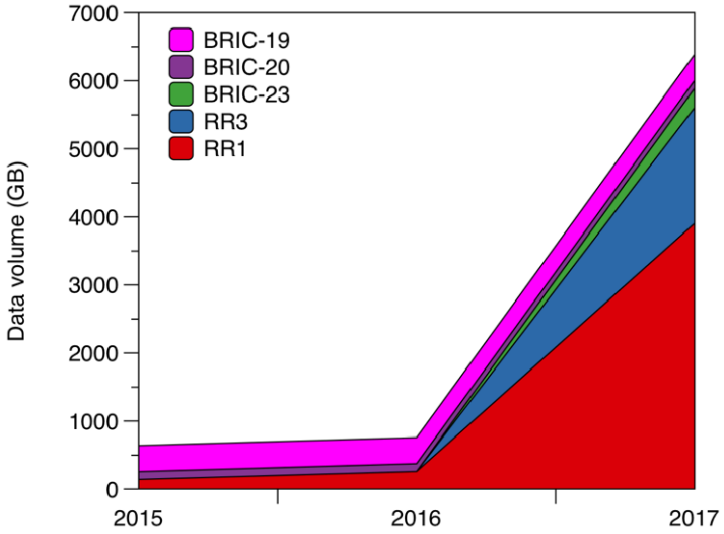
Talk outline

- **GeneLab background and motivation**
- **GeneLab Data System (GLDS) and datasets**
  - Data system architecture (AWS, search, federation,)
  - Datasets, dataset metadata, and curation
  - Collaborative workspace
  - Cloud analysis platform
- **GeneLab community**
  - Analysis Working Groups (AWGs)
- **Increasing the scientific return on Space Biology investments**
  - Leveraging biomedical research
  - Aggregating Space Biology research
  - Increasing data volume from Space Biology research

GeneLab Sample Processing Laboratory (SPL)

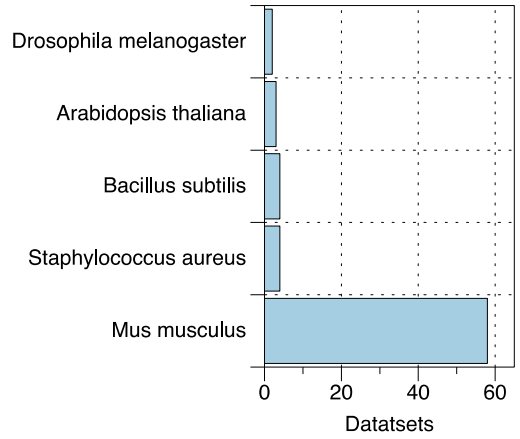
- Expertise:
  - DNA/RNA/protein extraction
  - Animal work

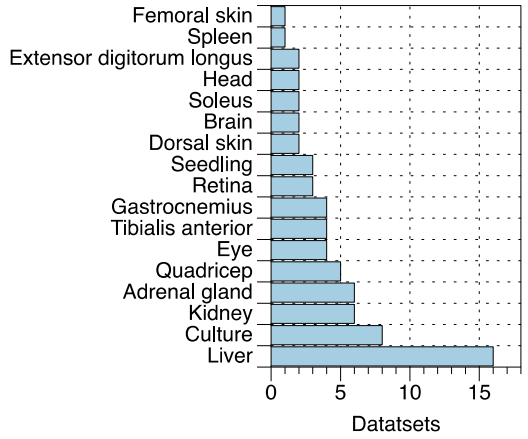
- Develop standards for sample processing (species dependent)
- Responsible for ~50% of GeneLab data by volume

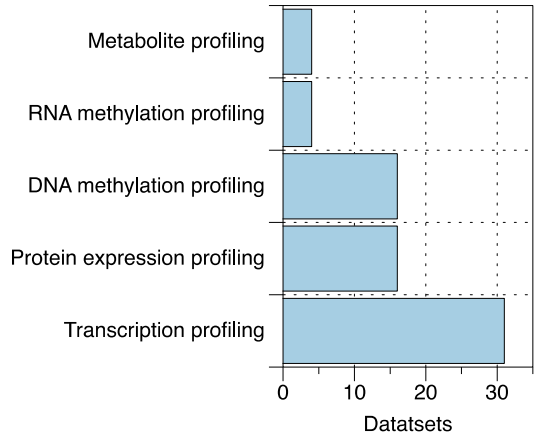


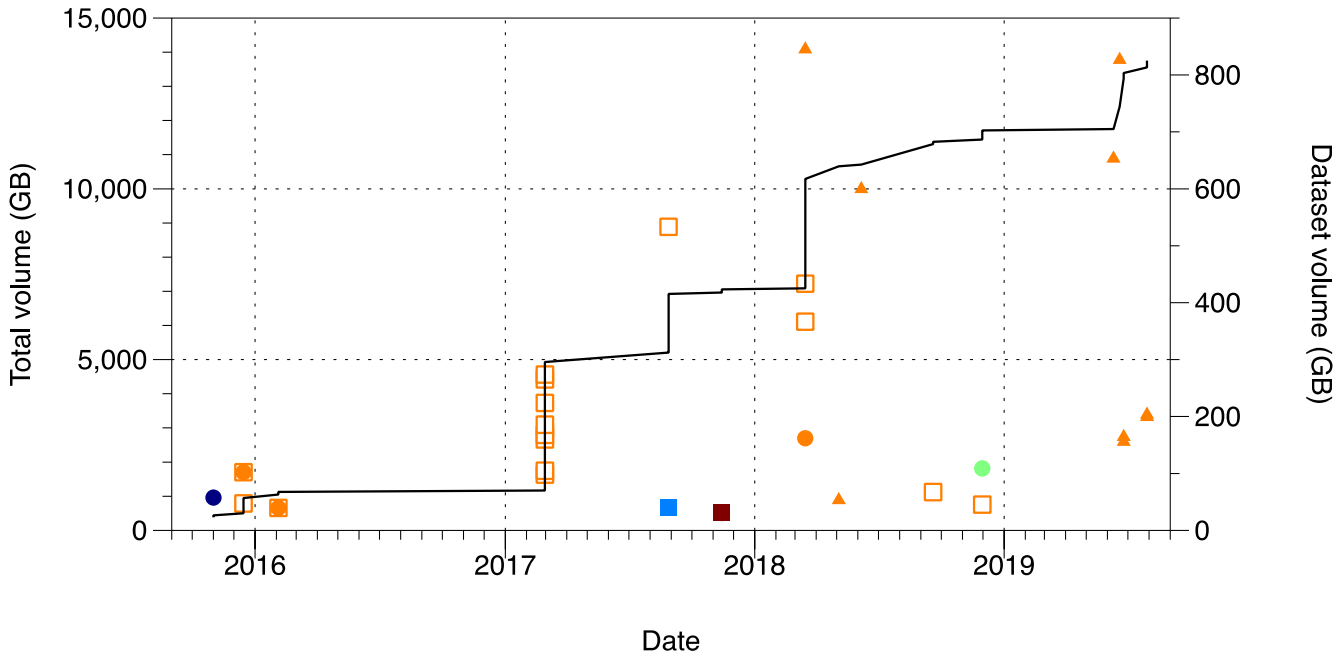
GeneLab Data Generation Summary

Since 2015, GeneLab has helped generate 70 published datasets equaling ~14 TB of data





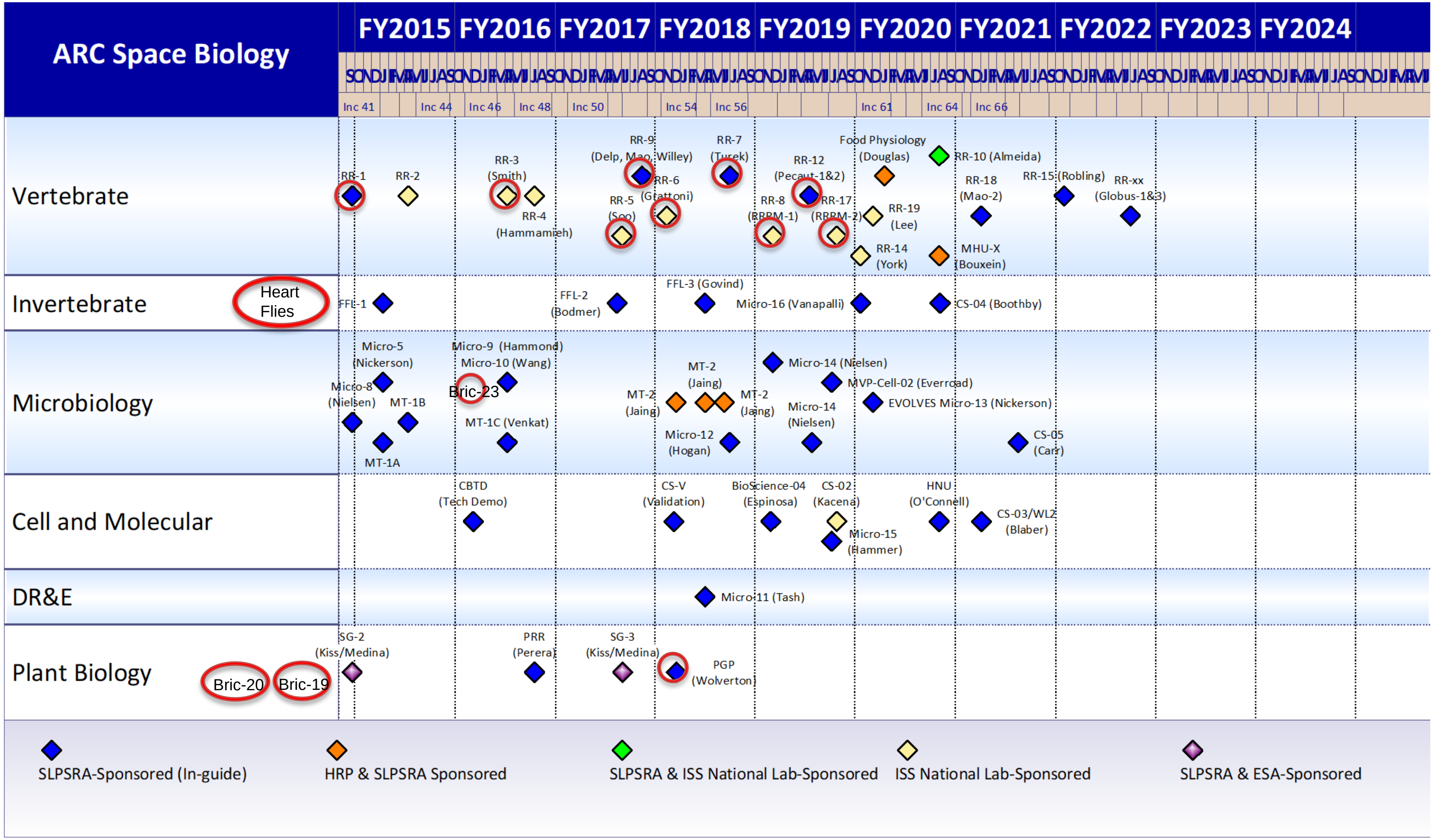




Shapes: Assay type
Triangle: Transcriptional profiling
Open square: DNA methylation profiling
Closed square: Metabolite profiling
Circle: Protein expression profiling
Diamond: RNA methylation profiling

Color: Species
Dark blue: *Arabidopsis thaliana*
Light blue: *Bacillus subtilis*
Green: *Drosophila melanogaster*
Orange: *Mus musculus*
Red: *Staphylococcus aureus*

Ames Space Biology "Fly-off Schedule" 2015 – 2022



○ = GeneLab data generation (actual and planned)

GeneLab Recommended Sequencing Standards

GeneLab suggests the following sequencing parameters to best capture important signals in RNA-sequencing data.

	Representative genome	Transcriptome complexity	Ribodepleted RNA-seq	Poly-A enriched RNA-seq
Human	*H. sapiens* GRCh38.p12	High	60 M clusters, 150 bp PE	40 M clusters, 150 bp PE
Rodents (mouse and rat)	*M. musculus* GRCm38.p6	High	60 M clusters, 150 bp PE	40 M clusters, 150 bp PE
Fruit fly	*D. melanogaster* Release 6 plus ISO1 MT	Medium	40 M clusters, 150 bp PE	30 M clusters, 150 bp PE
Worms	*C. elegans* WBcel235	Medium	40 M clusters, 150 bp PE	30 M clusters, 150 bp PE
Plants	*A. thaliana* TAIR10.1	High	60 M clusters, 150 bp PE	40 M clusters, 150 bp PE
Fungi	*S. cerevisiae* S288C R64	Low	20 M clusters, 150 bp PE	13 M clusters, 150 bp PE
Bacteria	*E. coli* str. K-12 substr. MG1655	Low	20 M clusters, 150 bp PE	13 M clusters, 150 bp PE

Spike-in Controls. At a minimum, GeneLab recommends that all sequencing samples include ERCC Spike-in Mix 1 in all samples. This will allow assessment of the dynamic range within each sample. In addition, in experiments focused upon comparing 2 levels of a given factor (spaceflight vs. ground control), GeneLab recommends including Mix 1 in the level A samples (spaceflight) and Mix 2 in the level B samples (ground control). This allows direct assessment of the available power in the data. In all cases, these Mixes should be added to samples before library preparation and sequencing. These Spike-in standards are available from Thermo Fisher: Mix 1 (https://www.thermofisher.com/order/catalog/product/4456740?SID=srch-srp-4456740), Mix 1 and 2 (https://www.thermofisher.com/order/catalog/product/4456739?SID=srch-srp-4456739).

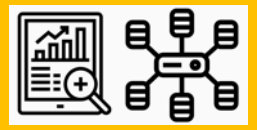
Library preparation and sequencing standard. GeneLab recommends that all sample pools to be sequenced include at least one replicate of the SEQC Universal Reference RNA available from Agilent (https://www.agilent.com/en/product/gene-expression-microarray-platform/gene-expression-microarray-kits-reagents/gene-expression-universal-reference-rnas-228491). This RNA should be included as a sample during each library preparation batch.

Ames Life Science Data Archive & NASA ARC Institutional Scientific Collection

Ames Life Sciences Data Archive (ALSDA)

- Data and Descriptions from 837 Experiments
  - Flight and Ground-based Experiments (non-human life sciences)
  - Space Biology and Human Research Program funded
  - Data Submission Agreements
- Information is curated and made available (when approved) on public LSDA website

Data Archived

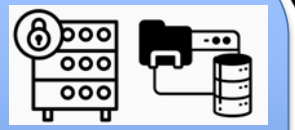


Biospecimen metadata
Payload
Mission
Hardware
Personnel
Research Subject
Images
Publication citations
PI and Project Data Sets

NASA ARC Institutional Scientific Collection (ISC)

- 32,037 Non-Human Biospecimens
  - Spaceflight investigations and related ground controls
  - Tissue request process managed by ALSDA
  - Tissue request proposal reviewed by appropriate funding organization (SB or HRP)
  - Housed in ARC Biospecimen Storage Facility (BSF)
  - Curated by ALSDA

Collection Management



Secure database
Metadata: Conditions, Tissue type, Descriptions, Species, Fixations, Chain of Custody Record, Telemetry
Storage: -80ºC, -20ºC, +4ºC, Ambient
Features: Power backup, Alarm system for power failure (24-7 staff on-call)

NASA ARC Institutional Scientific Collection (ISC)

Biospecimens Available for Research

		2016	2014-17	2017	2018	2017	2017	2017	2017	2016	2014	2011	2010	2007	2001	1996	1995	1993	1993	1991	1985	1974
		ISS Biological Research in Canister23	Hindimb Unloaded	ISS/JAXA Mouse Habitat Unit-2	ISS Rodent Research 7	ISS Rodent Research 9	ISS Rodent Research 6	ISS Rodent Research 5	ISS Rodent Research 4	ISS Rodent Research 3	ISS Rodent Research 1	STS-133 Immune2	STS-131 Immune	STS-118 CBTM2	STS-108 CBTM	STS-72 NIH.R3	STS-70 NIH.R2	STS-56 PARE.03	STS-58 SLS-2	STS-40 SLS-1	STS-51B SL-3	Cosmos 1129
System	**Specimen Types**	**Bacteria**	**Rat**	**Mouse**	**Mouse**	**Mouse**	**Mouse**	**Mouse**	**Mouse**	**Mouse**	**Mouse**	**Mouse**	**Mouse**	**Mouse**	**Quail**	**Rat**	**Rat**	**Rat**	**Rat**	**Rat**	**Rat**	**Rat**
Circulatory	Blood, Heart, Lymph nodes, Spleen		X			X					X			X	X			X	X	X		
Digestive	Cecum, Colon, Duodenum, GI Tract, Ileum, Intestine, Jejunum, Pancreas, Stomach			X	X	X	X		X	X	X	X		X	X	X		X	X	X		
Endocrine	Adrenal glands, Liver, Salivary glands, Thymus, Thyroid		X		X	X	X			X	X	X	X	X		X	X	X	X	X		
Excretory	Bladder, Fecal pellets, Kidney		X			X	X					X	X	X		X		X	X	X		
Integumentary	Adipose, Skin, Whiskers		X	X		X	X	X	X	X	X			X	X	X		X				
Muscular	Adductor longus, Diaphragm, EDL, Gastroc, Plantaris, Soleus		X		X	X				X	X					X		X	X	X	X	
Neurosensory	Brain, Eyes, Hypothalamus, Pituitary									X	X			X		X		X	X		X	
Reproductive	Ovary, Uterus, Testes		X	X		X	X			X				X		X		X	X	X		
Respiratory	Esophagus, Lung, Trachea		X	X	X	X	X			X			X	X		X		X	X	X		
Skeletal	Calvaria, Femur, Humerus, Jawbone, Parietal, Pelvis, Ribs, Tail, Tibia, Vertebrae			X	X	X		X	X	X				X	X	X			X			X
DNA and Protein extracts	Adrenal, Kidney, Liver, TA, EDL, Eye, Gastroc, Quadriceps, Soleus					X	X				X											
DNA and Protein extracts	Bacillus subtilis, Staphylococcus aureus	X																				

Next page
For details

GeneLab Team