GeneLab: Visualization and VWG - Introduction

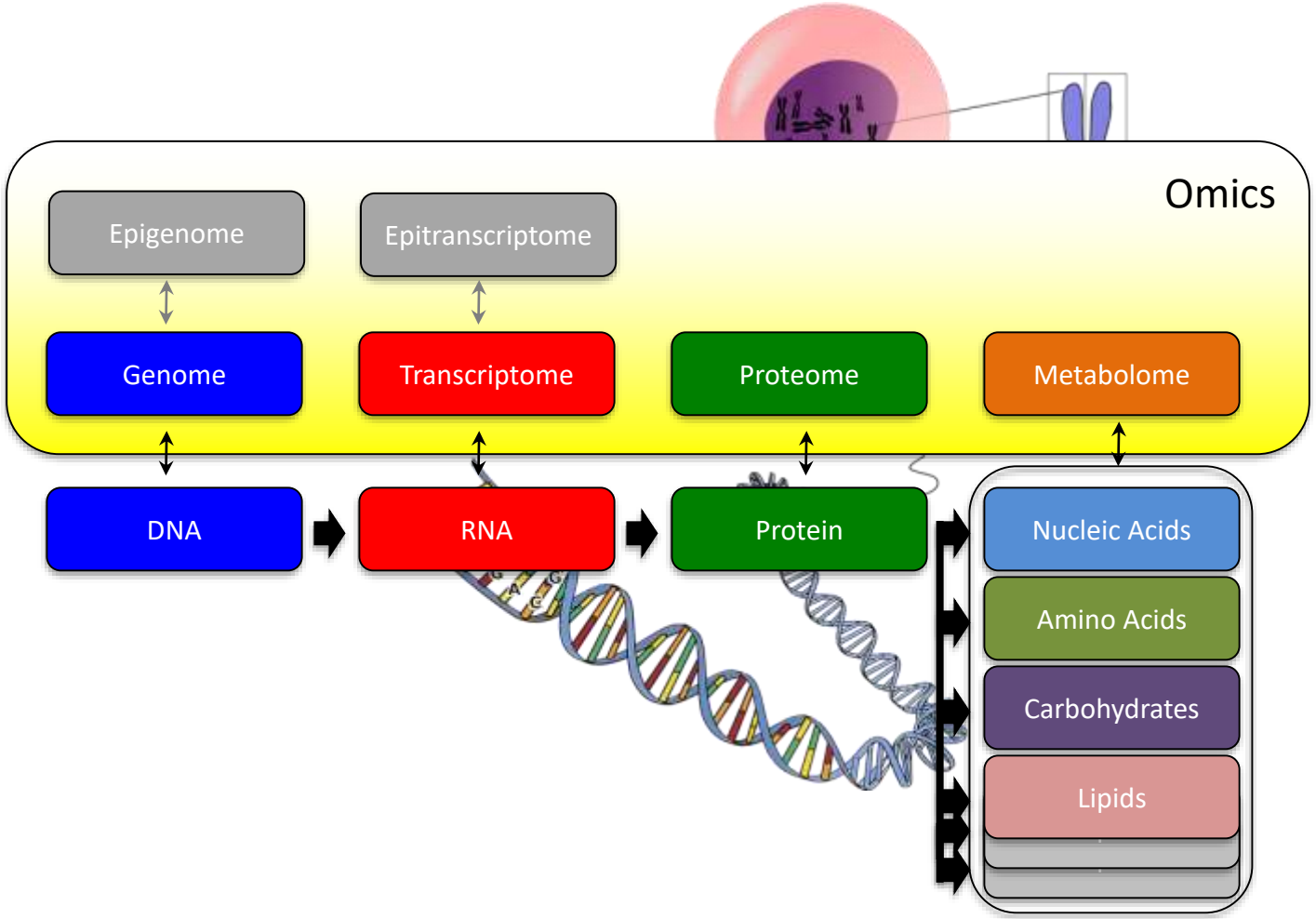
National Aeronautics and Space Administration

Disseminate and reuse data, tools, and samples post-project

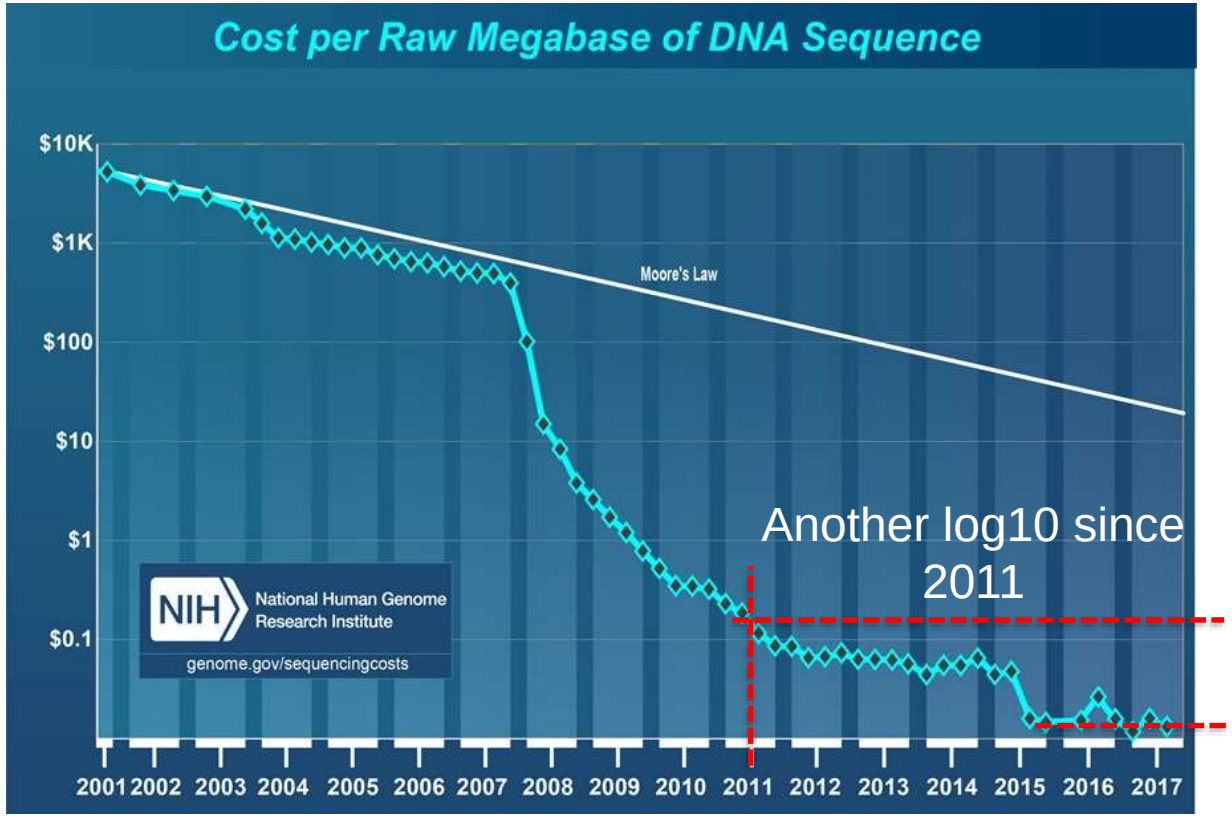
Sylvain V. Costes, PhD
GeneLab Project Manager

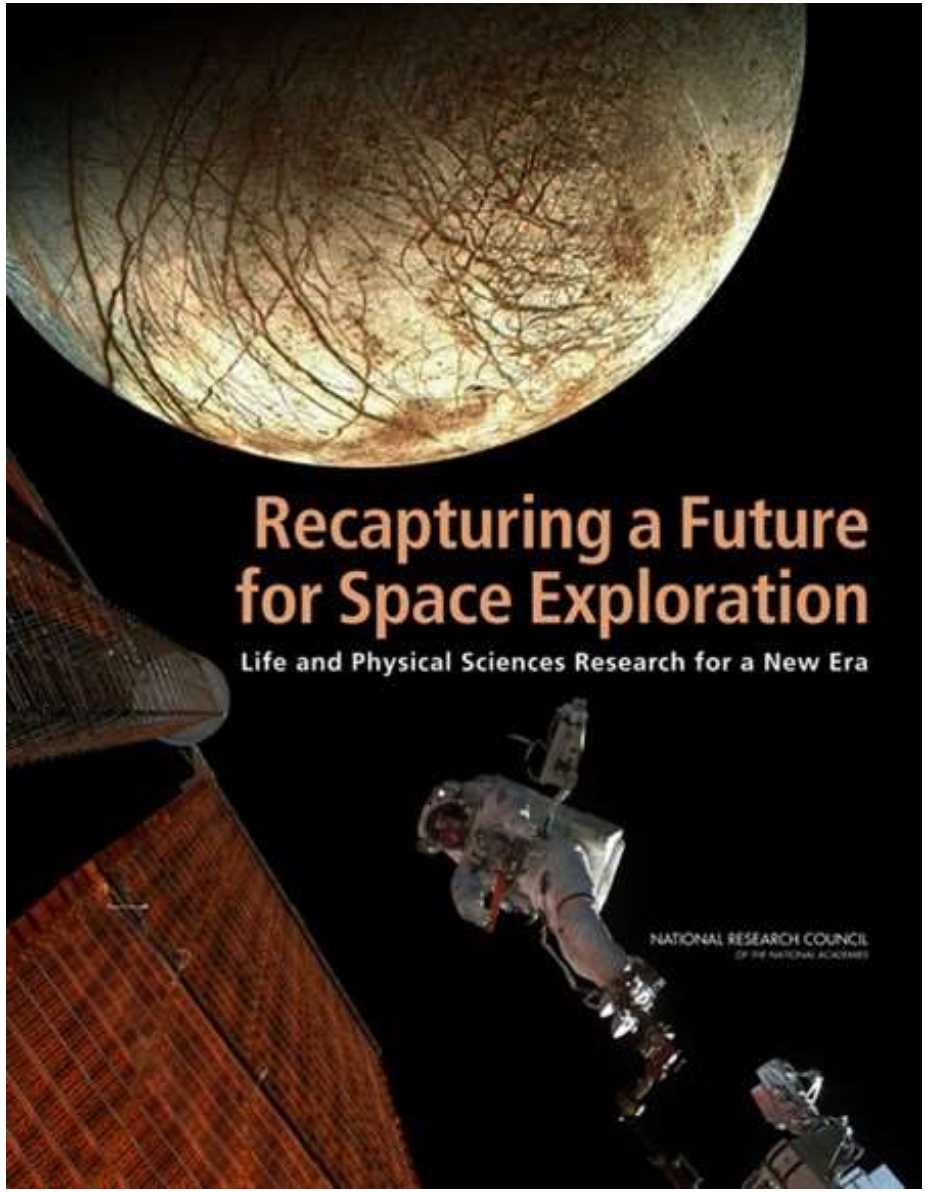
Afshin Beheshti, PhD
GeneLab Scientist and VWG Lead

What is Omics?



2011 NRC Decadal Survey and the Sequencing Paradigm Shift



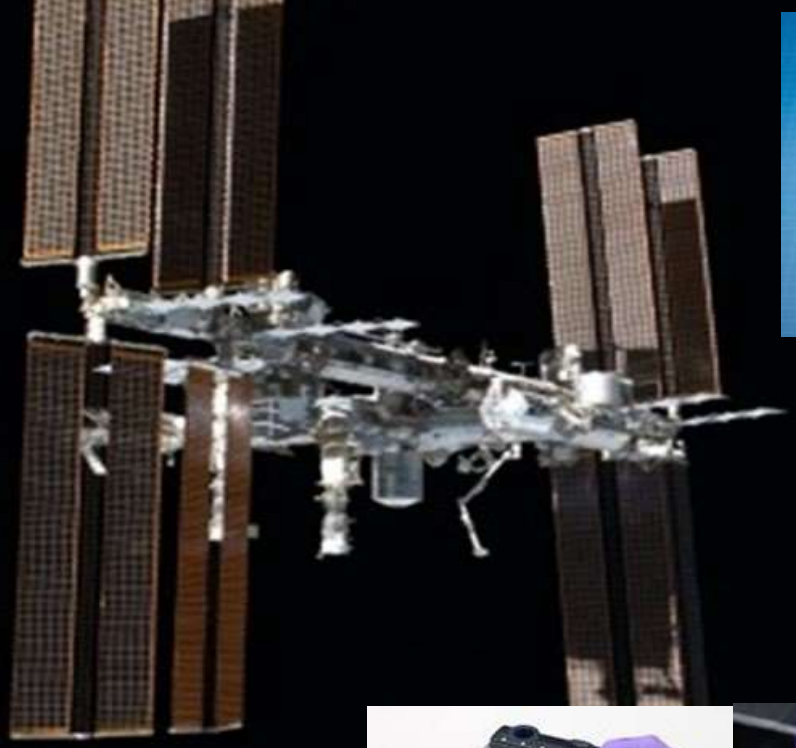


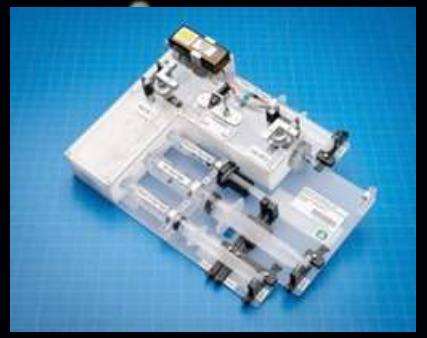
*"…**genomics, transcriptomics, proteomics, and metabolomics** offer an immense opportunity to understand the effects of spaceflight on biological systems…"*

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

Omics Acquisition in Space is Now a Reality

This is truly an exciting time for cellular and molecular biology, omics and biomedicine research on ISS with these amazing additions to the suite of ISS Laboratory capabilities.



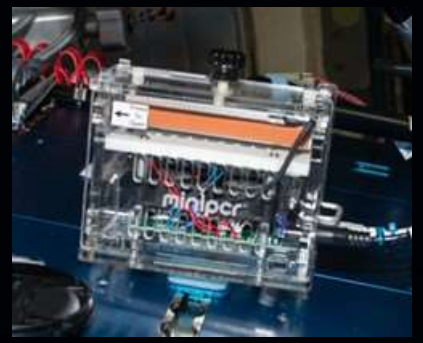


Sample Preparation Module

Oxford Nanopore MinION Gene Sequencer

Reaction tube containing lyophilized chemical assay bead (proprietary)

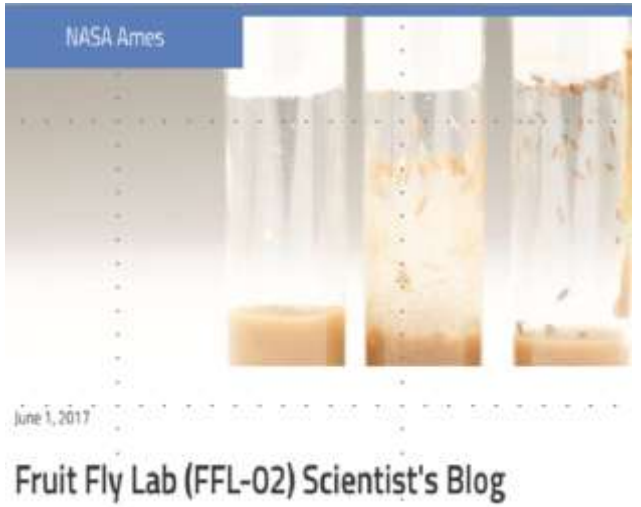
Cepheid Smart Cycler qRT-PCR

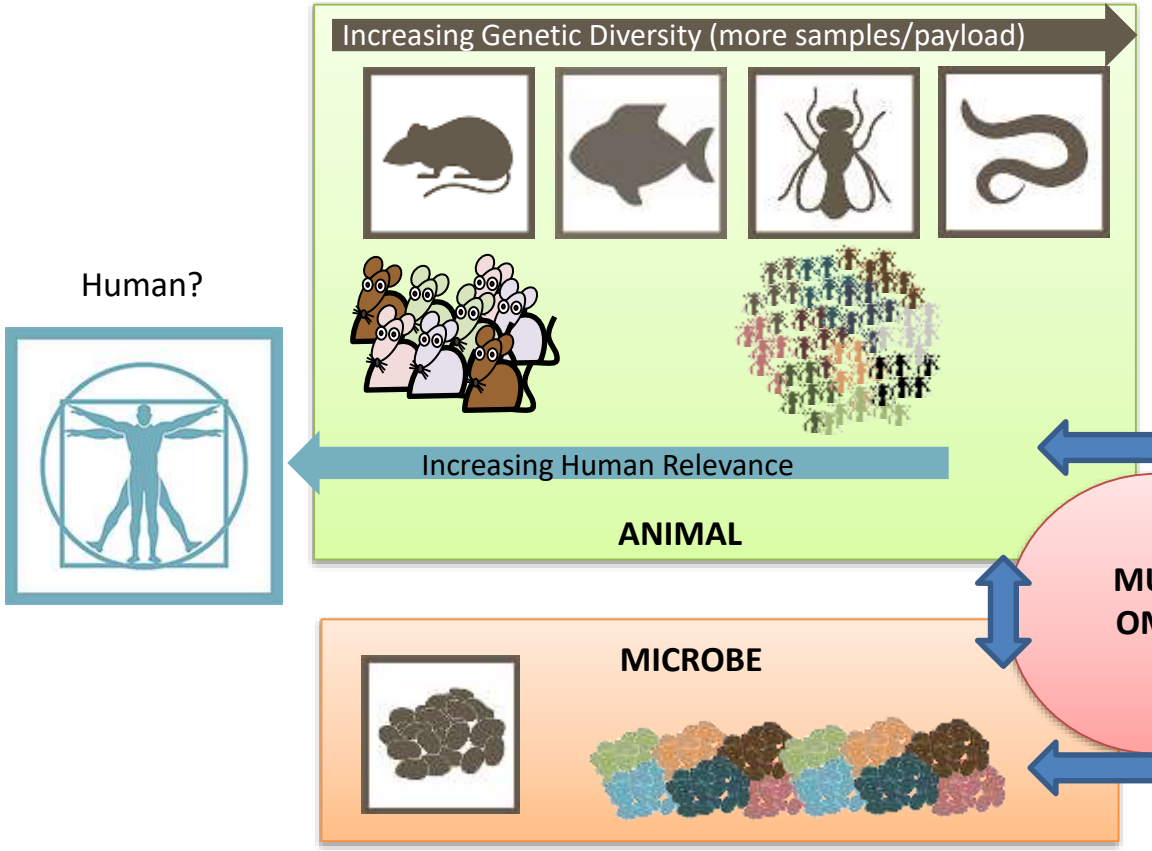


Mini-PCR

GeneLab ecosystem: maximizing knowledge by bringing experiments together as a system

- **Sequencing on ISS is still limited in the amount of data generated**
  - Most of the work needs to happen on earth
- **Measurements on human cannot be too invasive and limited in numbers**
  - Usage of animals





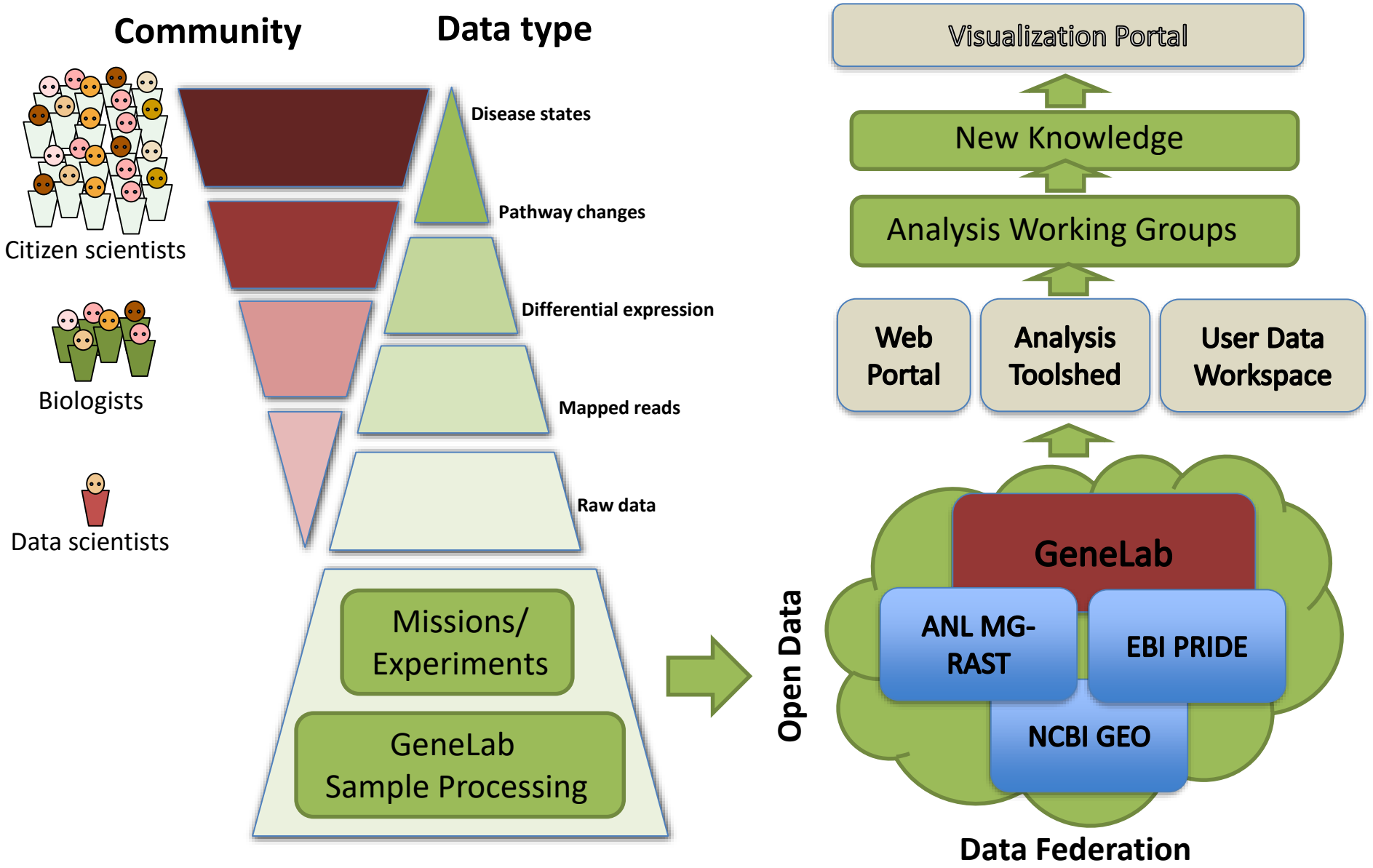
Identify Shared Processes/ Molecular Signatures

- Hypoxic Response ?
- Oxidative Stress
- Common Tissue (e.g. muscle, liver, heart, eyes, brain,…)

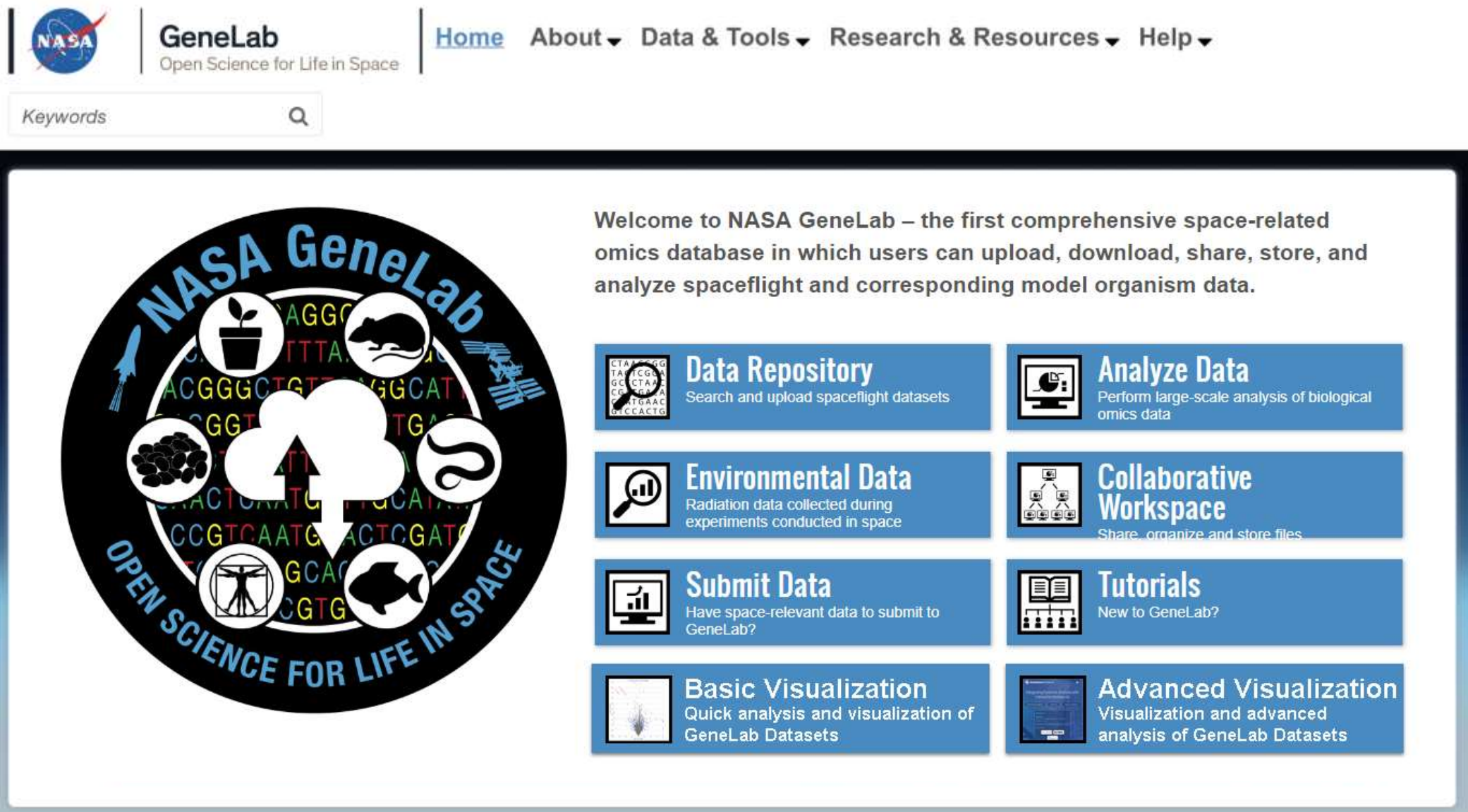
For Spaceflight

- High "n" number – statistically significant data
- Genetically identical animals
- Low resource requirements
- Short life cycle - multiple generations
- Measure response of a whole multicellular animal
- **Flies used as a model for humans for innate immunity, circadian rhythm, oxidative stress, neurobehavior, development, genetics, GWAS, "omics" studies etc.**

GeneLab Data Democratization



GeneLab Webpage: genelab.nasa.gov

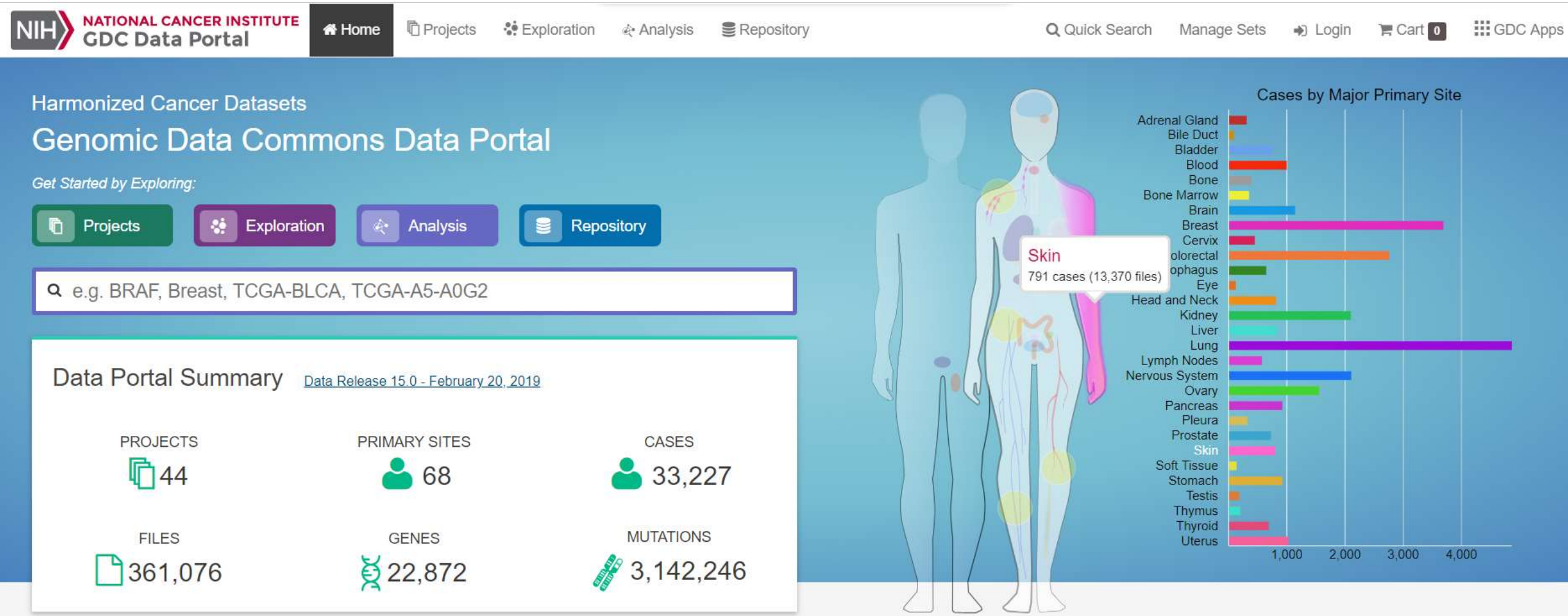


Master schedule

GeneLab Visualization	2018			2019									2019-2020
	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sept	
Phase 1: Formed VWG													
Phase 2: Prototype Development-Visualization on GeneLab													
Phase 2: Prototype Development-Notebook based visualization													
Phase 3: Integration & Testing of Visualization with GeneLab													
Phase 4: Maintenance & Future Collaborations													

What is Visualization and How it can be utilized

TCGA
https://portal.gdc.cancer.gov/



What is Visualization and How it can be utilized

NATIONAL CANCER INSTITUTE GDC Data Portal

Home | Projects | Exploration | Analysis | Repository | Quick Search | Manage Sets | Login | Cart 0 | GDC Apps

TP53

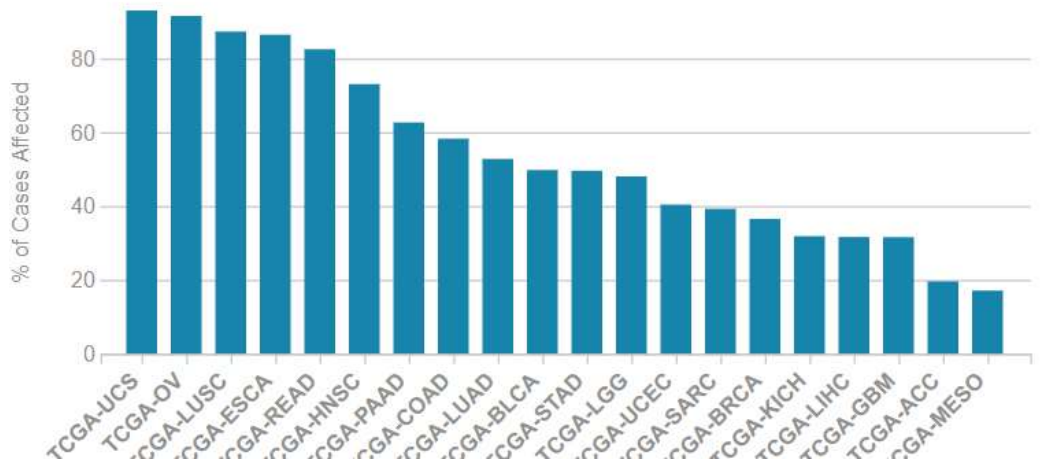
Type	protein_coding
Location	chr17:7661779-7687550 (GRCh38)
Strand	–
Description	This gene encodes a tumor suppressor protein containing transcriptional activation, DNA binding, and oligomerization domains. The encoded protein responds to diverse cellular stresses to regulate expression of target genes, thereby inducing cell cycl… ▾ more
Annotation	Cancer Gene Census

HGNC	HGNC:11998
OMIM	191170
Ensembl	ENSG00000141510

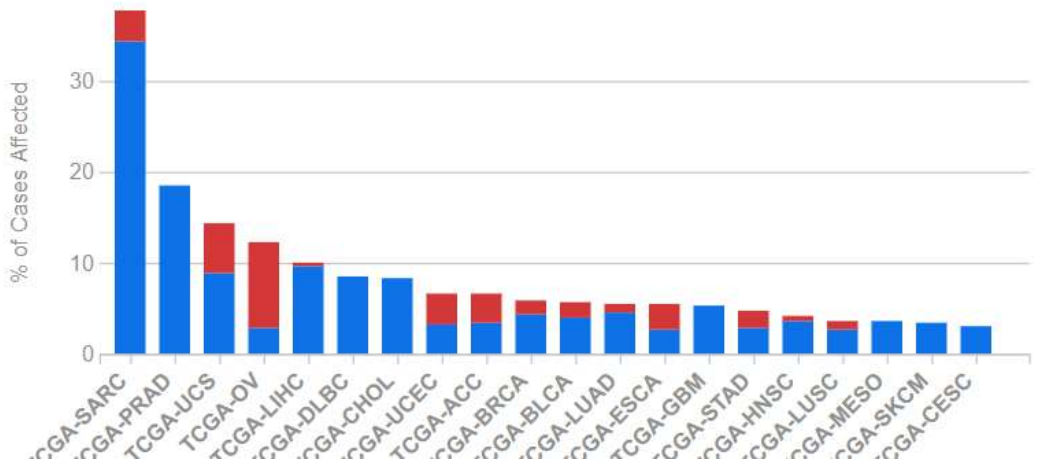
Cancer Distribution

Open in Exploration

4,008 CASES AFFECTED BY 1,291 MUTATIONS ACROSS 32 PROJECTS



4,251 CASES AFFECTED BY 638 CNV EVENTS ACROSS 31 PROJECTS



What is Visualization and How it can be utilized

NIH NATIONAL CANCER INSTITUTE GDC Data Portal

Home | Projects | Exploration | Analysis | Repository | Quick Search | Manage Sets | Login | Cart 0 | GDC Apps

Set Operations
3/4/2019 x

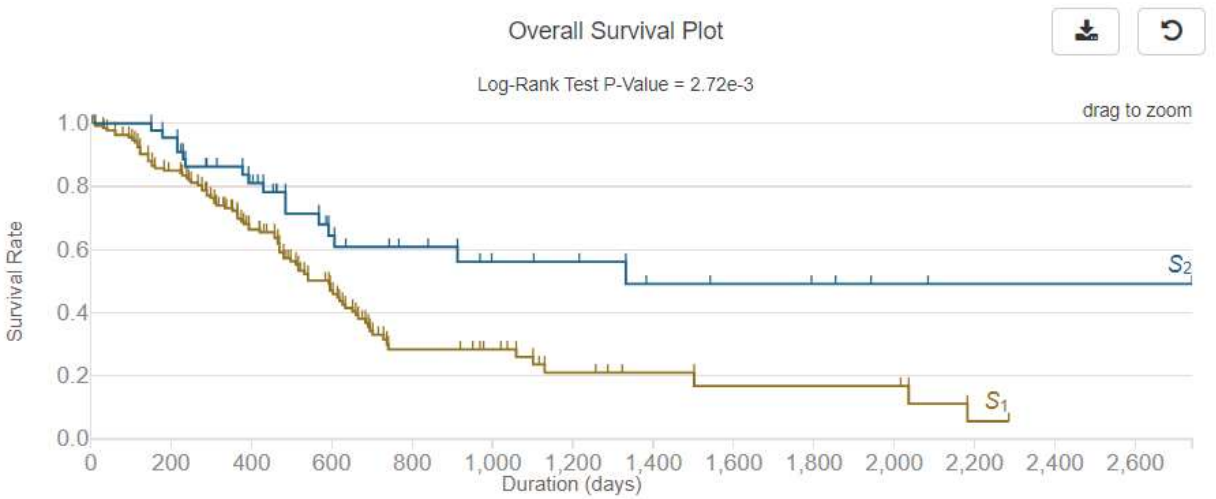
Cohort Comparison
3/4/2019 x

Delete All

Cohort Comparison

Demo showing cases with pancreatic cancer with and without mutations in the gene KRAS.

Survival Analysis



Cases included in Analysis	# Cases S_1	%	# Cases S_2	%
Overall Survival Analysis	139	99%	45	6%

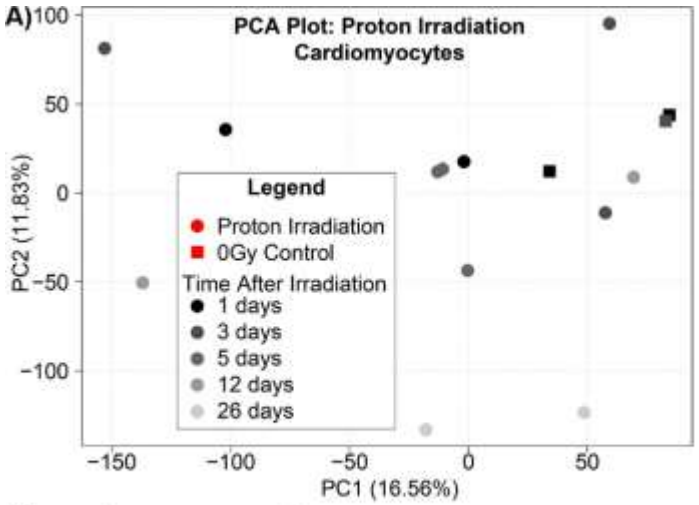
Cohort	# Cases
S_1 : Pancreas - KRAS mutated	140
S_2 : Pancreas - KRAS not mutated	774

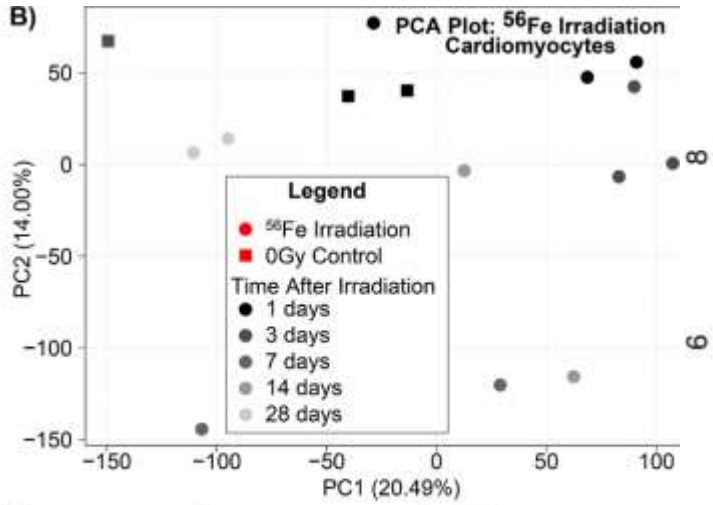
Open Venn diagram in new tab

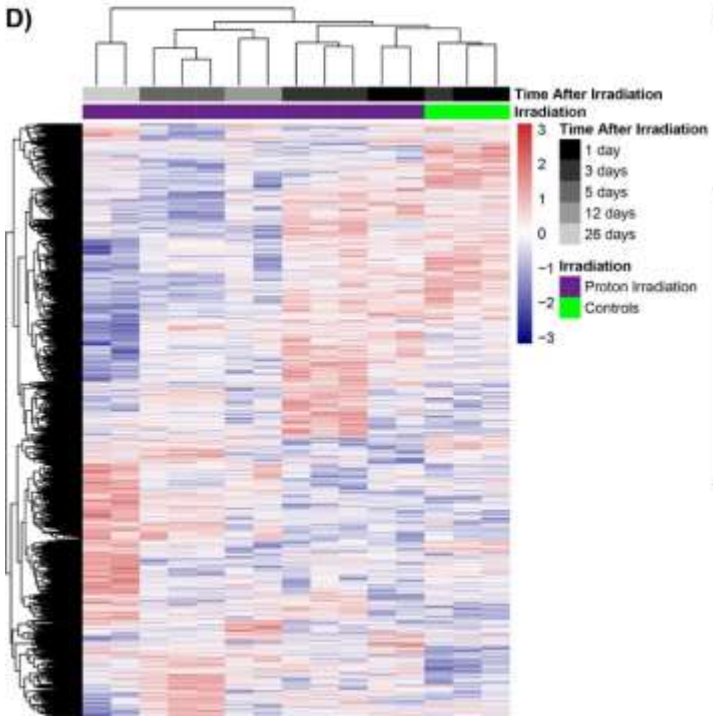
- [x] Survival
- [ ] Ethnicity
- [x] Gender
- [x] Vital Status
- [ ] Race

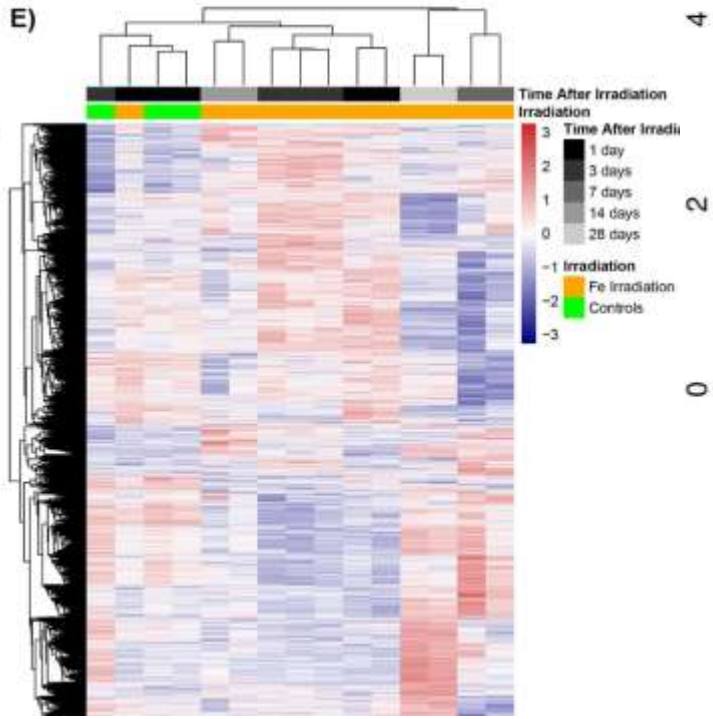
Gender

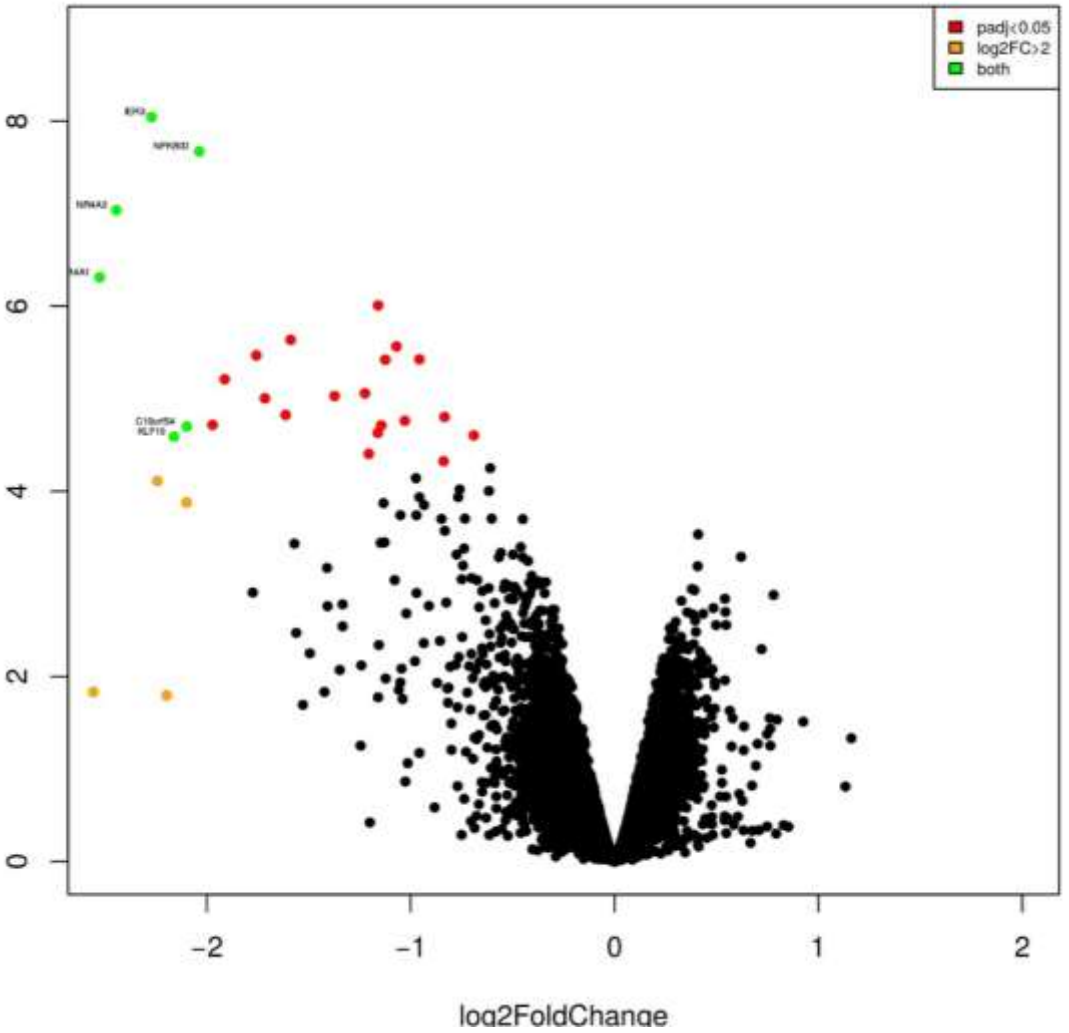
What is Visualization and How it can be utilized











GeneLab Visualization Roadmap

Basic visualization

- Fixed Interaction
- Limited Functionality

Advanced Visualization

- Customizable
- Ad Hoc

Phase 1: Visualization Working Group (VWG)

- **Two Components:**
  - One team developing the Visualization Tools
  - One team assisting with API development
- **Both teams are involved on a <u>voluntary basis</u> and VWG designed to be assistance from the global scientific community**
- **Established on 10/2018**

Leading the VWG

Afshin Beheshti

NASA Ames Research Center

Advanced Visualization Team

Michael Reich

Ted Liefeld

UNIVERSITY of CALIFORNIA, SAN DIEGO
SCHOOL OF MEDICINE

GenePattern Notebook

Nirav Merchant

Upendra Kumar Devisetty

RESEARCH, DISCOVERY & INNOVATION
Data Science Institute
THE UNIVERSITY OF ARIZONA

CYVERSE

SCIENCE Through Data-Driven Discovery

Simon Gilroy

Richard Barker

WISCONSIN
UNIVERSITY OF WISCONSIN-MADISON

Chris Mason

Kirill Grigorev

Cornell University

Basic Visualization Team

Daniel Couch

MUSC Health

Rien Dijkstra

Jess Alfredsen

Tobias Lindberg

Sebastian Steensson

Phase 2: Prototype Development: Basic Visualization

	Work Already Done	Future Work
Rien Dijkstra (Astrophysicists/Software Engineer in Amsterdam) • Interactive PCA plots • Interactive Volcano Plots • Interactive Comparison Plots • Interactive Dendrogram	1st Prototype: 12/21/18 2nd Prototype: 1/11/19 3rd Prototype: 2/8/19	1st Scheduled Finalized Prototype for GeneLab to Test: 4/01/19 UAT: 9/2019
DAMVAD (Software Engineers and Big Data Analysts) • Interactive Heat map • Interactive Mouse Diagram integrated with the omics data on GeneLab	1st Prototype: 12/24/18	1st Scheduled Finalized Prototype for GeneLab to Test: 6/01/19 UAT: 9/2019
Daniel Couch (Systems Programmer) • Interactive Network tool: • Adapt his tool GAIL for GeneLab	Pending Prototype (Possible date: 4/1/19)	1st Scheduled Finalized Prototype for GeneLab to Test: 7/01/19 UAT: 9/2019

Phase 2: Prototype Development: Advanced Visualization

GeneLab API Visualization Development

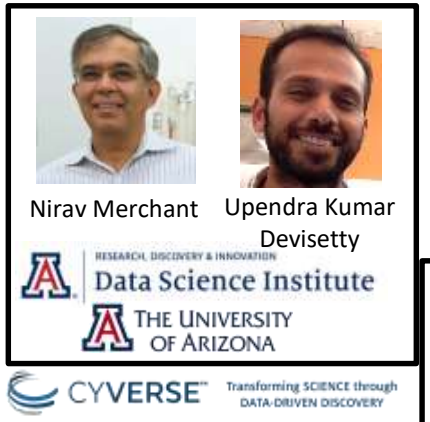
API wrapper development and integration with GenePattern Notebook

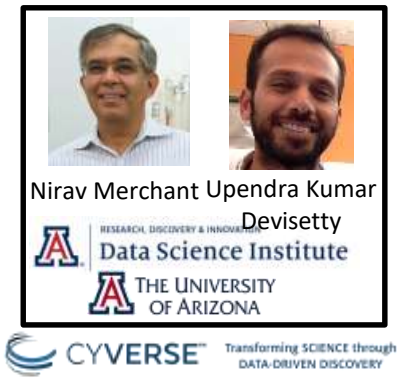
Incorporation of GenePattern/GeneLab API into Cyverse

Cyverse to implement a basic GenePattern Notebook/GeneLab API interface

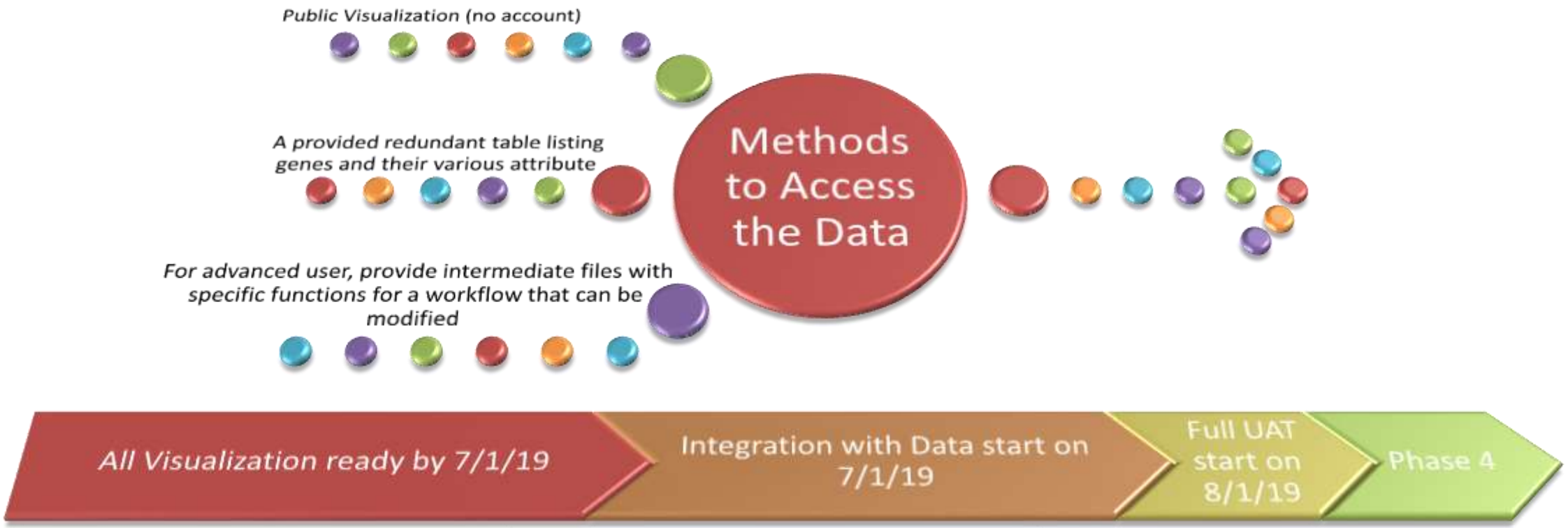
Visualization and Advanced Omics Analysis

Utilizing both GenePattern Notebook and Cyverse with the GeneLab API tool, GeneLab pipelines will be made available for advanced omics analysis



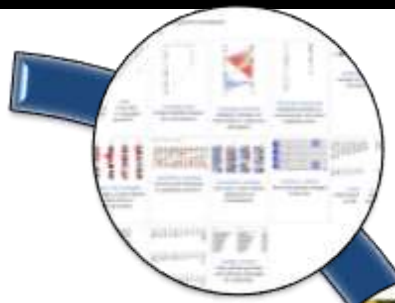


Phase 3: Integration & Testing of Visualization with GeneLab



Phase 4: Maintenance & Future Collaborations

GeneLab Visualization Collaboration for 2020



Additional Development: More visualization from Broad

External Visualization: MetaSub

External Visualization: TOAST/MANGO

Continued evolution of visualization on GeneLab

Kirill's API GeneLab API Development

setup.py Removed tqdm 6 days ago

README.md

GeneFab

Installation

GeneFab has been tested under Python 3.5+.
It is recommended to use a user installed Conda environment.
Currently GeneFab can be installed via **pip** directly from github:

```
pip install -e git://github.com/LankyCyril/genefab.git#egg=genefab
```

For development purposes, it can also be cloned and used like this:

```
$ git clone https://github.com/LankyCyril/genefab
$ cd genefab
$ conda env create --name genefab --file environment-linux.yaml
$ conda activate genefab
$ python
>>> from genefab import get_datasets, GLDS
    ...
```

Demo

https://github.com/LankyCyril/genefab/blob/master/genefab-demo.ipynb

Description

The current iteration of GeneFab supports GeneLab datasets processed according to API version 2.1. As of April 2, 2019, these are only the datasets GLDS-4, GLDS-30, and GLDS-42.

GeneLabDataSet(accession, storage_prefix=".genelab", index_by="Sample Name", verbose=False)

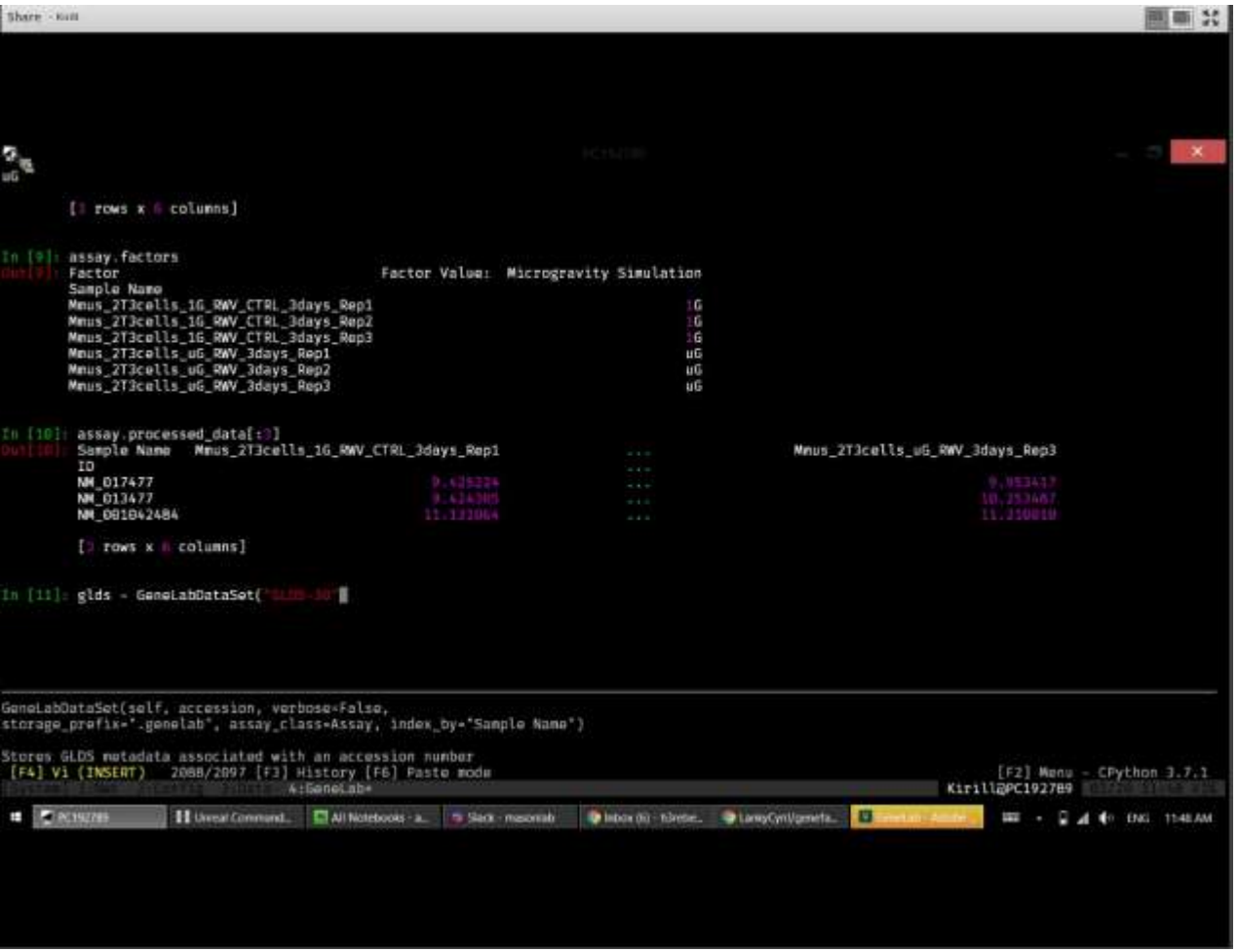
Initializes a GeneLabDataSet instance corresponding to `accession`.
Each GeneLabDataSet contains references to assays performed in the study; they are stored under the field `assays` as a list of `Assay` instances.

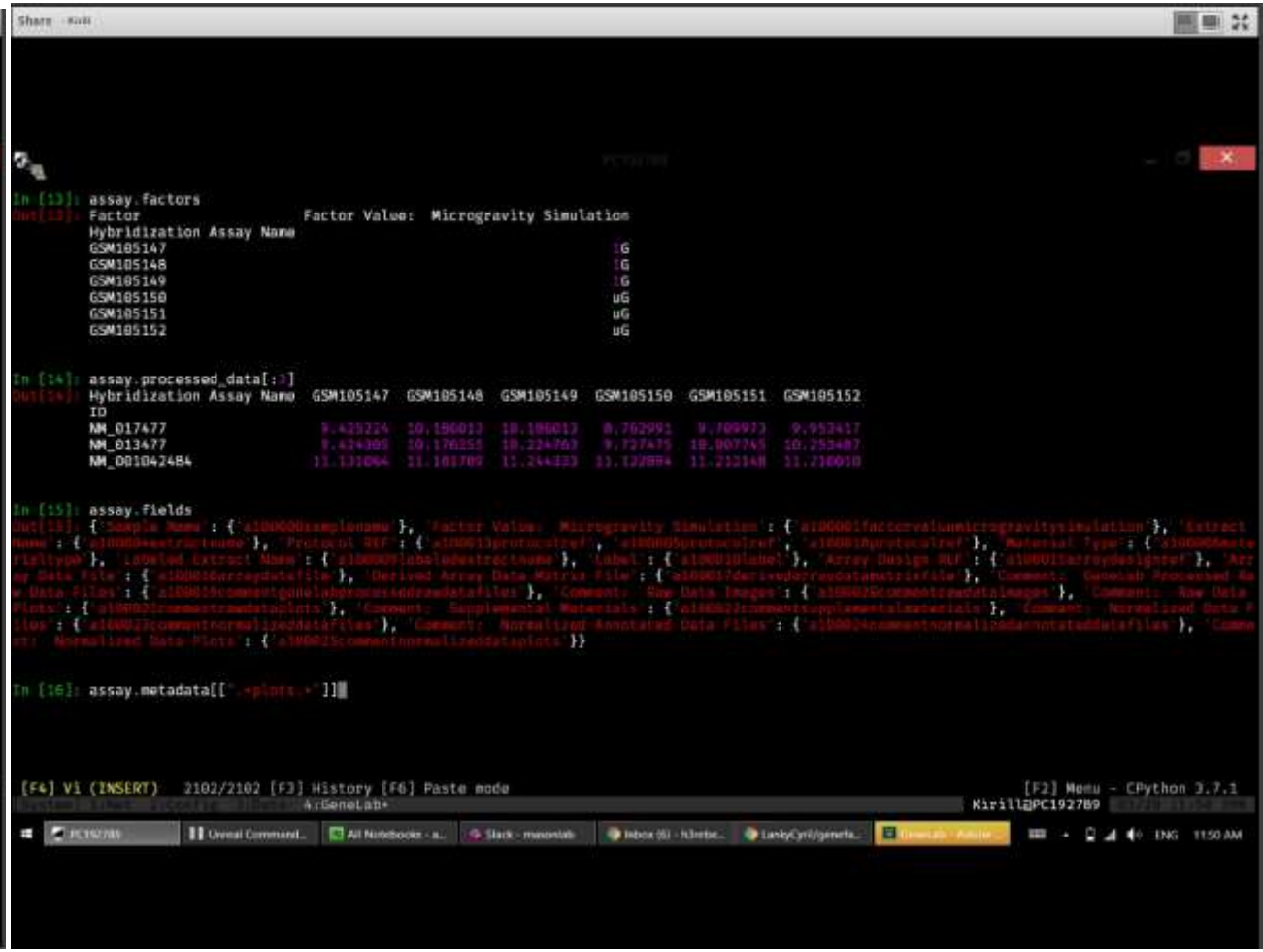
Chris Mason

Kirill Grigorev

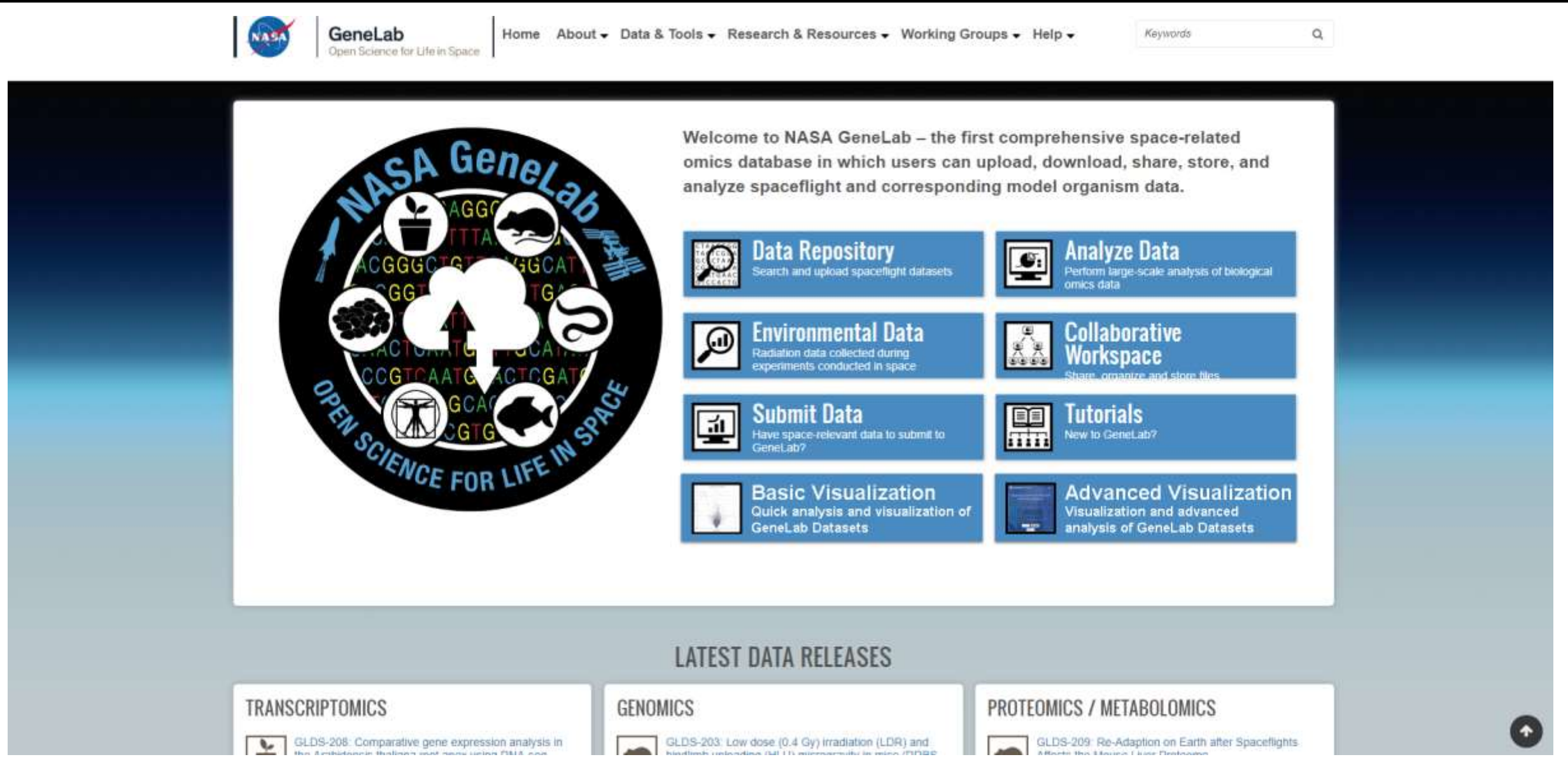
Cornell University

Kirill's API GeneLab API Development





Mock Up of potential visualization on GeneLab



Mock Up of potential visualization on GeneLab: Mock Basic Visualization webpage

GeneLab
Open Science for Life in Space

Home GLDS Data Repository Environmental Data Tools Submit Data Tutorials Help

Search for specifc molecular factor (e.g. p53)

Search for single or multiple GLDS (e.g. GLDS-4)

Search for Organs (e.g. liver)

Search for Species (e.g. plants)

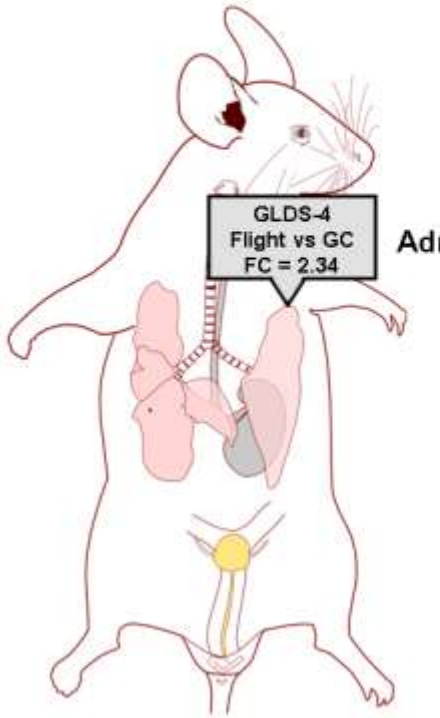
Volcano Plots

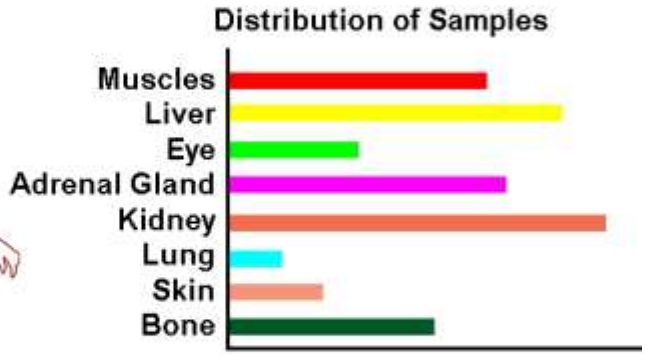
Heatmaps

PCA Plots

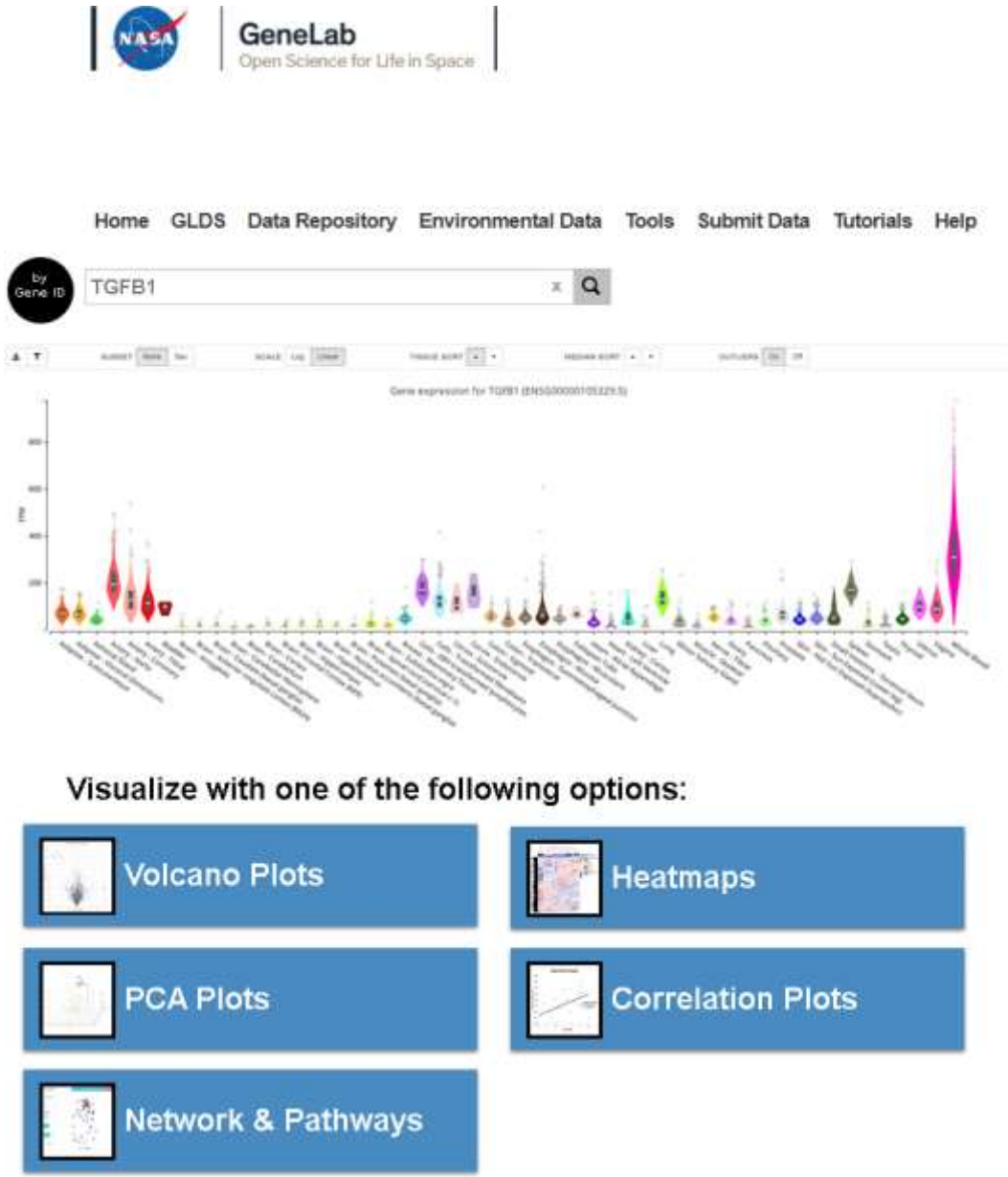
Correlation Plots

Network & Pathways

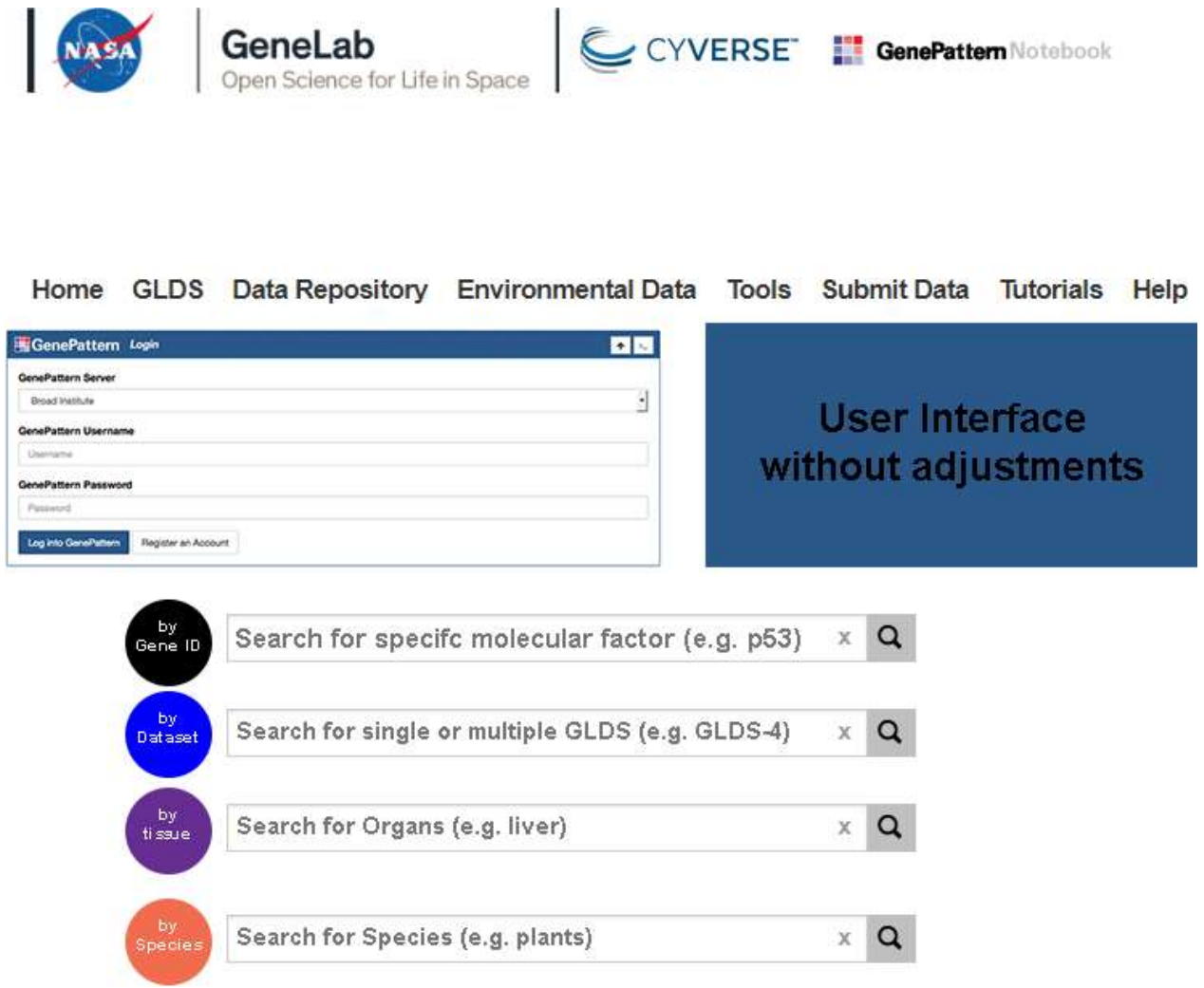




Mock Up of potential visualization on GeneLab: Mock Basic Visualization webpage



Mock Up of potential visualization on GeneLab: Mock Advanced Visualization webpage



Mock Up of potential visualization on GeneLab: Mock Advanced Visualization webpage

Visualization in Development: Basic Visualization examples

Query data | Pair plots | Volcano plots | Principle Component Analysis | Dendrograms

GeneLab Study

Organism

Tissue Type

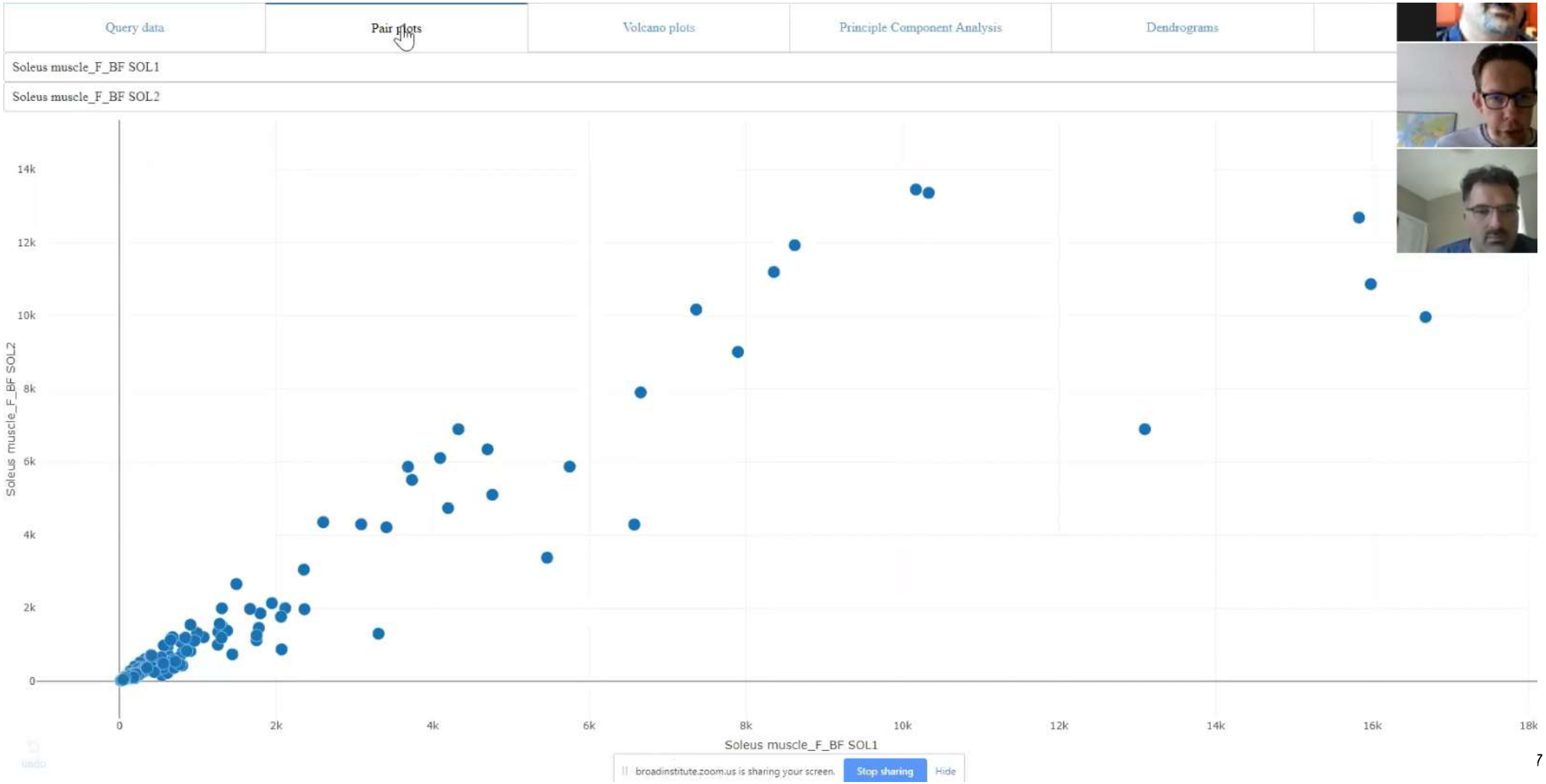
Gene Name

Comparison Type

GeneLab Study	Organism	Tissue Type	Gene Name	Comparison Type	Fold Change (Log2)
GLDS-111	mouse	Extensor digitorum muscle	0610042C05RIK	F vs GC	-0.4887913788371339
GLDS-111	mouse	Extensor digitorum muscle	0610042C05RIK	F vs SMG	-0.5816786559450451
GLDS-111	mouse	Extensor digitorum muscle	0610042C05RIK	SMG vs GC	0.09288727710791107
GLDS-111	mouse	Extensor digitorum muscle	1110013L07RIK	F vs GC	-0.321165860205428
GLDS-111	mouse	Extensor digitorum muscle	1110013L07RIK	F vs SMG	-0.12584214734109933
GLDS-111	mouse	Extensor digitorum muscle	1110013L07RIK	SMG vs GC	-0.1953237128643288
GLDS-111	mouse	Extensor digitorum muscle	1110019N10RIK	F vs GC	-0.4581764992642475
GLDS-111	mouse	Extensor digitorum muscle	1110019N10RIK	F vs SMG	-0.2940955606236057
GLDS-111	mouse	Extensor digitorum muscle	1110019N10RIK	SMG vs GC	-0.16408093864064185
GLDS-111	mouse	Extensor digitorum muscle	1110058L19RIK	F vs GC	0.0009118229516197414
GLDS-111	mouse	Extensor digitorum muscle	1110058L19RIK	F vs SMG	-0.016582993188499186
GLDS-111	mouse	Extensor digitorum muscle	1110058L19RIK	SMG vs GC	0.01749481614011892
GLDS-111	mouse	Extensor digitorum muscle	1810015C04RIK	F vs GC	1.112762287749198
GLDS-111	mouse	Extensor digitorum muscle	1810015C04RIK	F vs SMG	0.948963506381427
GLDS-111	mouse	Extensor digitorum muscle	1810015C04RIK	SMG vs GC	0.1637987813677712
GLDS-111	mouse	Extensor digitorum muscle	2010315L10RIK	F vs GC	0.074744107876769
GLDS-111	mouse	Extensor digitorum muscle	2010315L10RIK	F vs SMG	0.2166768235598912
GLDS-111	mouse	Extensor digitorum muscle	2010315L10RIK	SMG vs GC	-0.14193271568312243
GLDS-111	mouse	Extensor digitorum muscle	2400001E08RIK	F vs GC	-0.42270257656608934
GLDS-111	mouse	Extensor digitorum muscle	2400001E08RIK	F vs SMG	-0.23755516143965874
GLDS-111	mouse	Extensor digitorum muscle	2400001E08RIK	SMG vs GC	-0.1851474151264307
GLDS-111	mouse	Extensor digitorum muscle	2400001E08RIK | LOC433216	F vs GC	-0.08367188308301321
GLDS-111	mouse	Extensor digitorum muscle	2400001E08RIK | LOC433216	F vs SMG	-0.16508960464339414
GLDS-111	mouse	Extensor digitorum muscle	2400001E08RIK | LOC433216	SMG vs GC	0.08141772156038099
GLDS-111	mouse	Extensor digitorum muscle	2610018G03RIK	F vs GC	-0.49107700944892374
GLDS-111	mouse	Extensor digitorum muscle	2610018G03RIK	F vs SMG	0.06138994651388862
GLDS-111	mouse	Extensor digitorum muscle	2610018G03RIK	SMG vs GC	-0.5524669559628123
GLDS-111	mouse	Extensor digitorum muscle	2610208M17RIK	F vs GC	0.16867090574058852
GLDS-111	mouse	Extensor digitorum muscle	2610208M17RIK	F vs SMG	0.10444698472176682
GLDS-111	mouse	Extensor digitorum muscle	2610208M17RIK	SMG vs GC	0.06423292191882157
GLDS-111	mouse	Extensor digitorum muscle	2700049H19RIK	F vs GC	-0.1

broadinstitute.zoom.us is sharing your screen. Stop sharing Hide

Visualization in Development: Basic Visualization examples



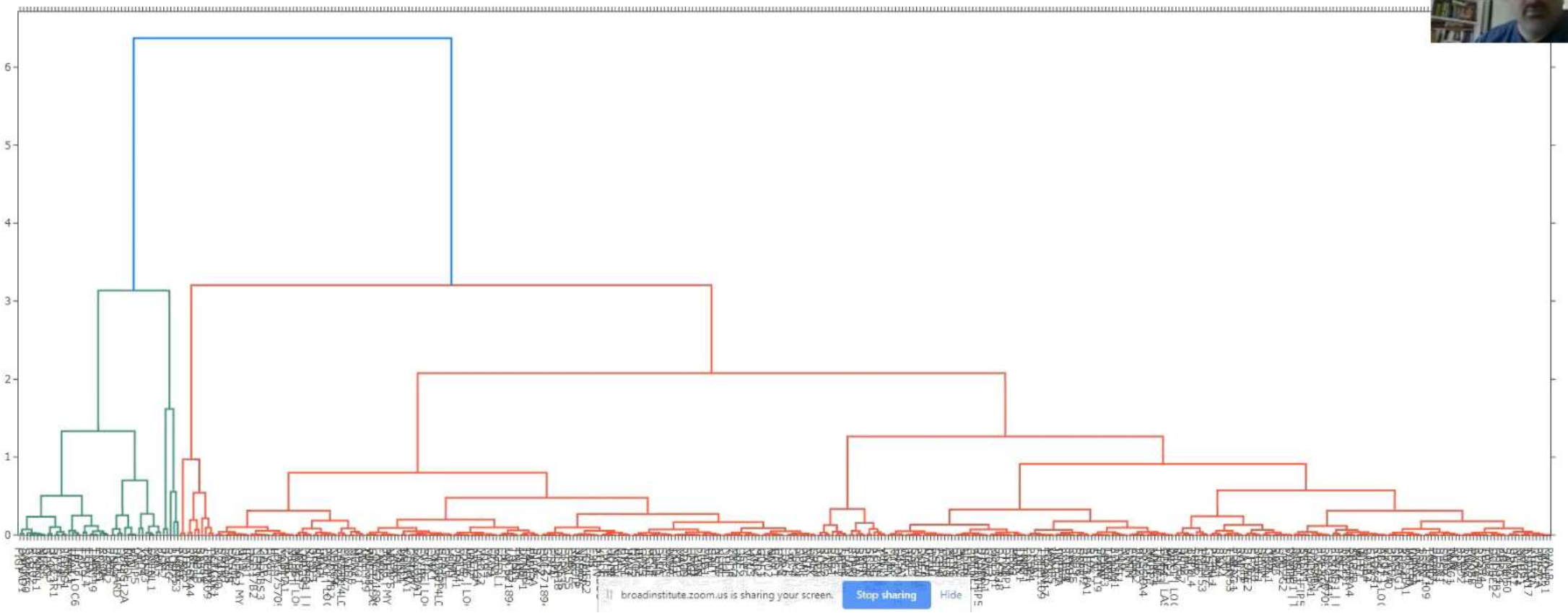
Visualization in Development: Basic Visualization examples

Query data | Pair plots | Volcano plots | Principle Component Analysis | Dendrograms

F vs GC

Bottom dendrogram

Hide heatmap



broadinstitute.zoom.us is sharing your screen. Stop sharing Hide

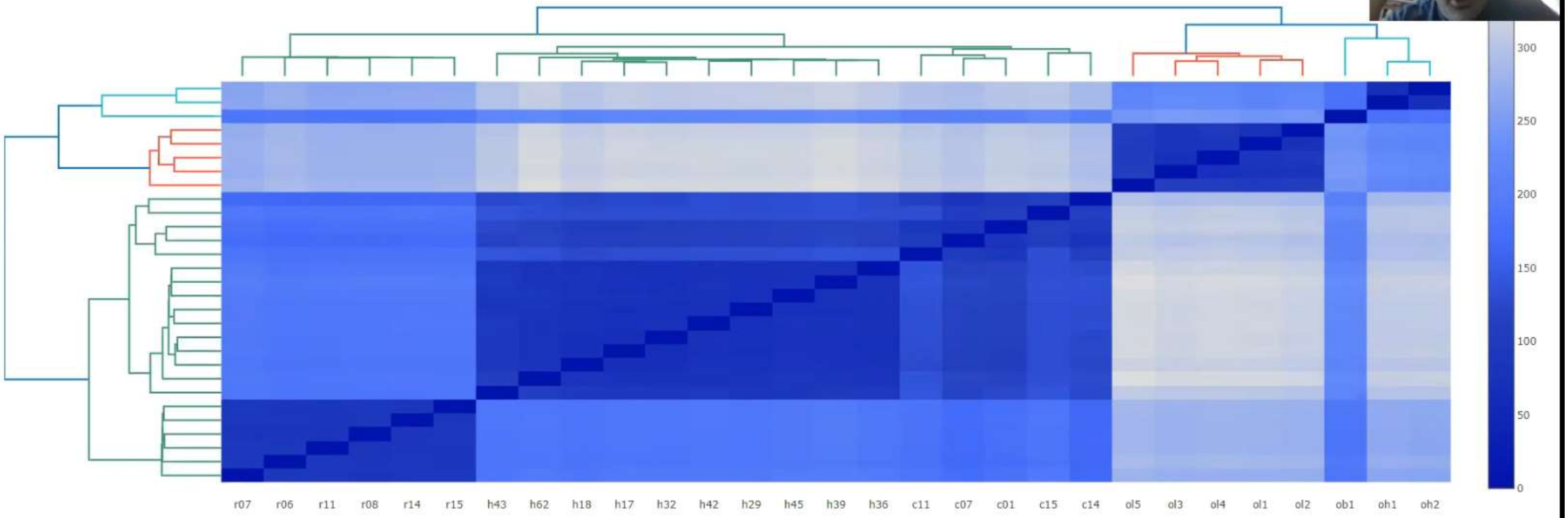
Visualization in Development: Basic Visualization examples

Query data | Pair plots | Volcano plots | Principle Component Analysis | Dendrograms

F vs GC

Left dendrogram

Hide heatmap



broadinstitute.zoom.us is sharing your screen. Stop sharing Hide

VWG Workshop

Location: Boston Marriott Cambridge, 50 Broadway, Cambridge, MA 02142

Purpose: This is a workshop with all members who are involved with the VWG and are assisting GeneLab with developing our visualization capabilities. This workshop will allow all members to present what they are working on, what stage the visualization is at, and future development of the visualization on GeneLab. Lastly, this workshop will allow for communication between all VWG members and allow hands-on demos of all tools and development of how to fully integrate all visualization components with the GeneLab for the next phase of development.

Each presentation from VWG members should be interactive and should include the following in their presentations:

1. Brief overview of the tool and platform. Should include in your presentation the architecture of your platforms, demo on how to use it, and any technical hurdles that we can discuss.
2. Demo on the tool. If possible please provide a way that the VWG members might be able to use the tool on their own computers.
3. The code for your tool made available earlier or during the presentation. If you want you can post your tool on our GitHub page (https://github.com/abehesht/GeneLab-Visualization)
4. Also allow for time during your presentation to have people interact with you and allow for brainstorming on how your tool can easily be integrated into the overall visualization flow that we will implement on GeneLab.

Agenda for Monday, 4/22/19:

8:00am – 8:40am	Introduction & GeneLab visualization: Sylvain and Afshin
8:40am – 9:10am	GeneLab Data Systems framework and deployment: Dan Berrios
9:10am – 9:30am	GeneLab Metadata & Processed data: Sam Gebre
9:30am – 10:30am	Cyverse: Sarah Roberts and Nirav Merchant
10:30am – 10:45am	Coffee break
10:45am – 11:45am	API Presentation and Discussion (GeneLab API wrapper: Kirill Grigorev)
11:45am – 12:00pm	Framework summary
12:00pm – 1:30pm	Lunch (on your own)
1:30pm – 2:15pm	GenePattern Notebook: Ted Lefield
2:15pm – 2:45pm	Interactive Plots for GeneLab: Rien Dijkstra
2:45pm – 3:30pm	Interactive Visualization Tools for GeneLab: Jess Alfredsen, Sebastian Steensson
3:30pm – 3:45pm	Coffee break
3:45pm – 4:45pm	Network tool: Daniel Couch
4:45pm – 5:00pm	Discussion and wrap up for the day

VWG Workshop

Agenda for Tuesday, 4/23/19:

8:00am – 8:10am	Other visualization tools Introduction: Afshin Beheshti
8:10am – 9:00am	TOAST/MANGO: Richard Barker, Simon Gilroy
9:00am – 10:30am	Interactive discussion on how to integrate the API with Basic Visualization tools
10:30am – 10:45am	Coffee break
10:45am – 12:30pm	Interactive discussion on how to integrate all visualization tools together in one package
12:30pm – 2:00pm	Lunch (on your own)
2:00pm – 5:00 pm	Working Session .
5:00pm – 5:50pm	Break before going to Broad Institute
5:50pm – 5:55pm	Meet on Hotel lobby and walk over to Broad

Tuesday Evening meetup with the Broad Institute at the Broad Institute:

Brief Presentations from GeneLab and VWG:

6:00pm – 6:10pm	Afshin Beheshti: Brief introduction of who we are at GeneLab and a very brief presentation of what we do and why are holding this meetup
6:10pm – 6:15pm	Rien Dijkstra
6:15pm – 6:20pm	Jess Alfredsen, Sebastian Steensson
6:20pm – 6:25pm	Cyverse: Sarah Roberts and Nirav Merchant
6:25pm – 6:30pm	TOAST/MANGO: Richard Barker

Brief presentations from people at Broad Institute for potential future visualization collaborations:

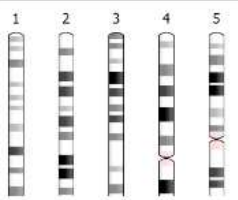
6:30pm – 6:40pm	Eric Weitz, Broad Institute: Ideogram
6:50pm – 7:00pm	Jen Lapan, Broad Institute: Data Donation Platform
7:00 pm – 7:30pm	Social interaction between everyone

Broad Visualization to add to GeneLab: Ideogram

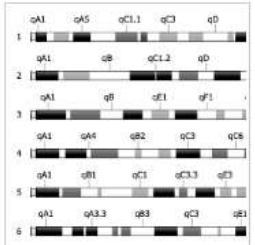
Overview | Ideogram

Ideogram.js is a JavaScript library for chromosome visualization. Ideogram supports drawing and animating genome-wide datasets.

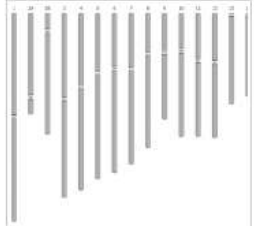
In the examples below, genomic data is fetched from servers and rendered on the client using D3.js.



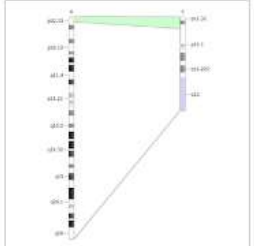
Human
Chromosomes 1-22, X and Y rendered vertically.



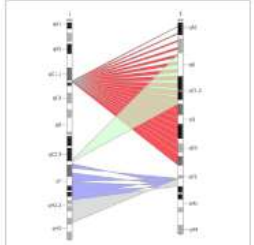
Mouse
Chromosomes 1-19, X and Y rendered horizontally, with labeled bands.



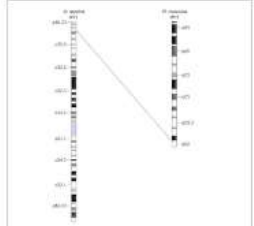
Eukaryotes
Genomes for many organisms, using data fetched via API.



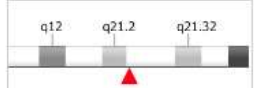
Homology, basic
Simple relations between two chromosomes.



Homology, advanced
Relations between two chromosomes, using color and opacity.

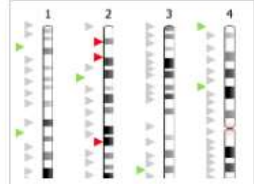


Homology, interspecies
Relations between two chromosomes, each from a different taxon.

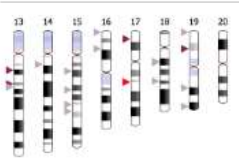


Annotations, basic
Location of a gene on a chromosome.

Annotations, overlaid
Overlaid features on chromosomes.

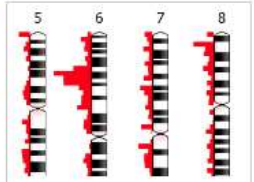


Annotations, tracks
Stacked features on chromosomes.

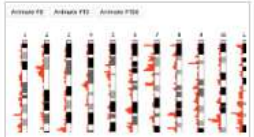


Annotations, external data
Data fetched from a third party API, annotated by shape and color.

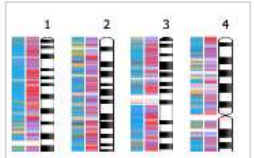
Annotations, file URL
Features parsed from standard file formats, e.g. BED, loaded via URL.



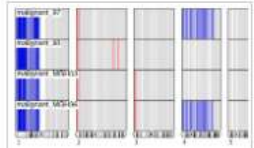
Annotations, histogram
Use shape to show feature distributions on chromosomes.



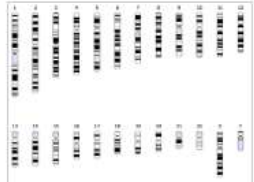
Annotations, animated
Genome-wide transitions in expression over time.



Annotations, heatmap.
Use color to show feature distributions on chromosomes.

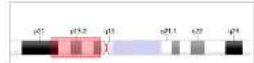


Geometry, collinear
All chromosomes arranged in one line.

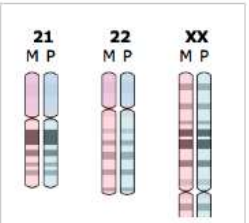


Layout, small
Chromosomes surrounded by other content.

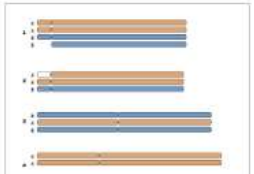
Layout, tabs
Tab containers and scrolling for large ideograms.



Brush
Move and resize a slider-like widget to change genomic coordinates in focus.

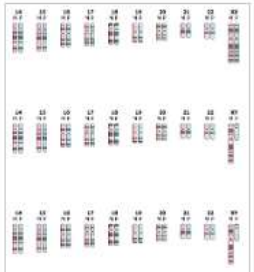


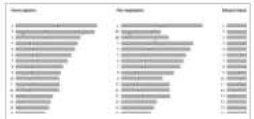
Ploidy, basic
Diploid genome of a



Ploidy, rearrangements
Loss and gain of cytogenetic arms and chromatids.

Ploidy, recombination
Gene conversion and double chromosome crossover among





Multiple, primates
Three primate genomes, with orthologs annotated for comparison.

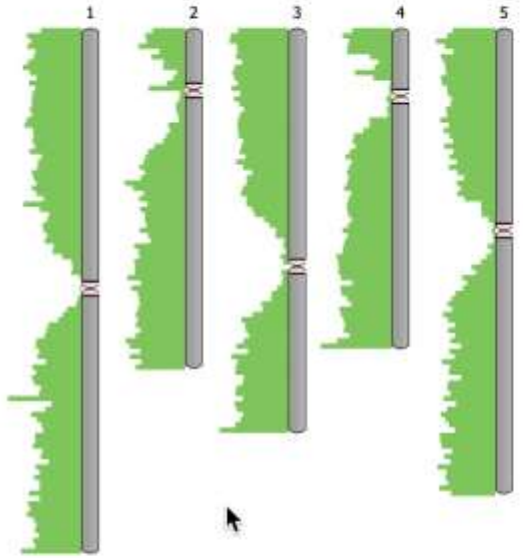
Eric Weitz, PhD

Ideogram in Context of Space Biology

Plant gene expression in space | Ideogram

Overview | Previous | Next | Source

This ideogram shows the distribution of all *Arabidopsis thaliana* genes throughout the genome. Explore how this plant's genes changed expression while in outer space by applying filters below.



Gene type
- Protein coding
- ncRNA
- lncRNA
- tRNA
- rRNA
- snRNA
- snoRNA

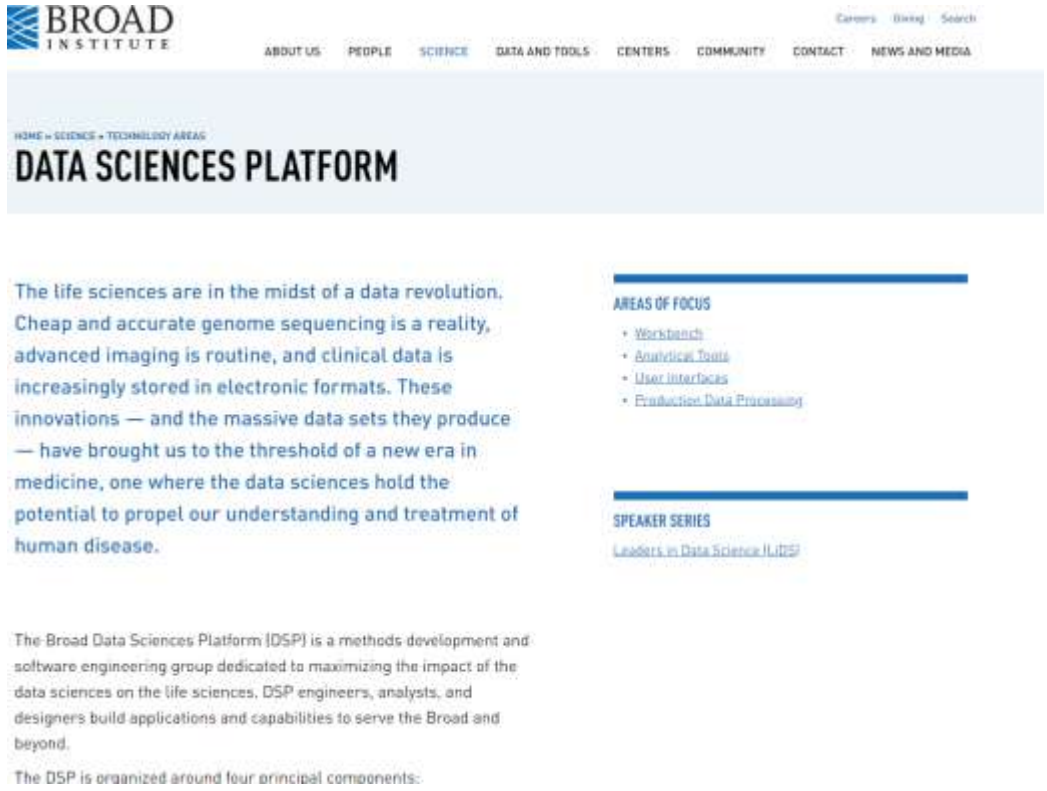
Significant expression change
- Yes
- No

Expression data from NASA GeneLab, in collaboration with Afshin Behesthi and Richard Barker. Ideogram is developed by Eric Weitz.

Broad Visualization to add to GeneLab: Data Donation Platform

Broad's engineering effort to advance ubiquitous, global, direct-to-participant genomic research. DDP is in a learning-and-prep mode now, anticipating a broad range of user demands to visualize ingested D2P data and eventual, analytical outputs

The Data Donation Platform (DDP): A software stack that enables direct participant engagement, including consent and recontact, via intuitive web and mobile interfaces. DDP provides the underlying infrastructure for disease-specific registries such as the Angiosarcoma Project, the Rare Genomes Project, and the Global A-T Family Data Platform.



Jen Lapan, PhD