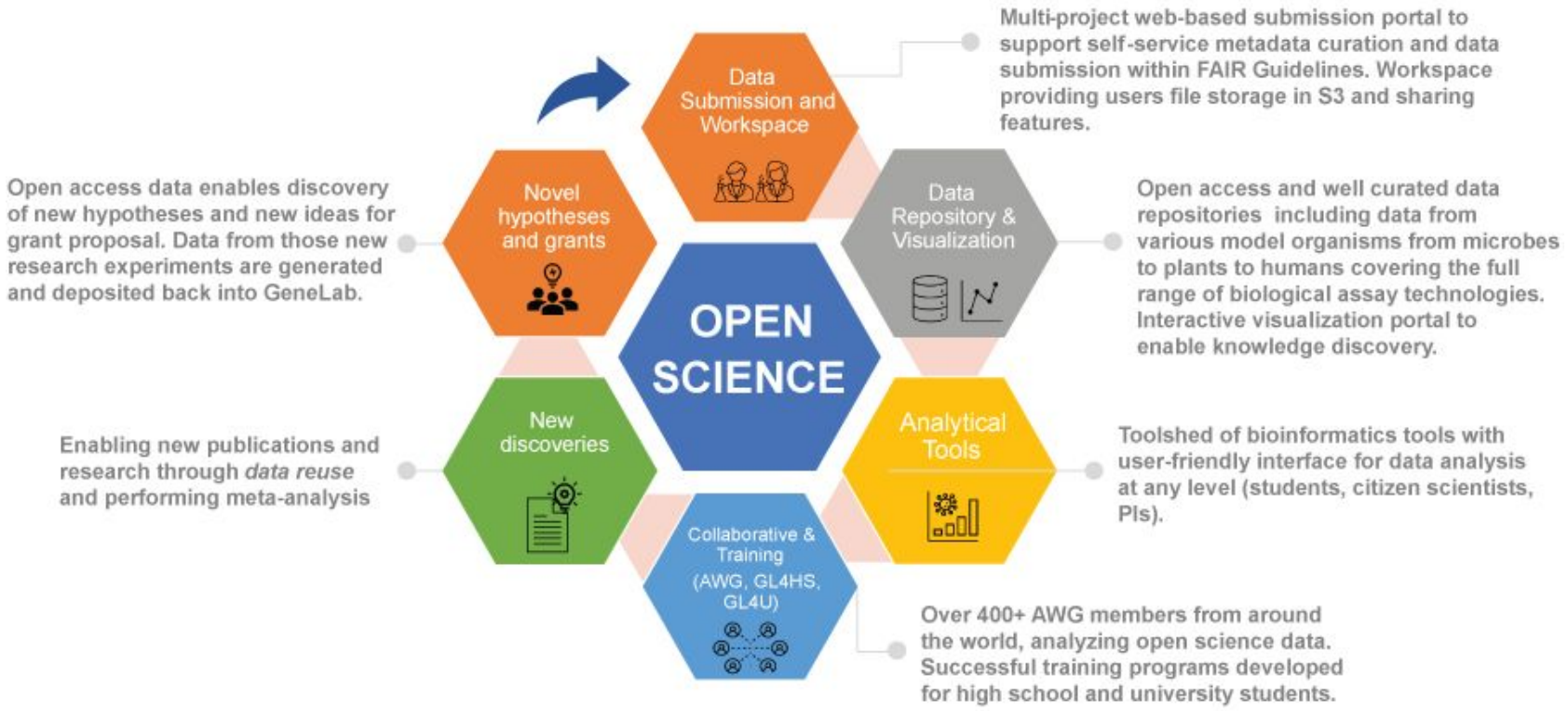
Unlocking the Secrets of Space Biology with NASA's GeneLab and the OSDR Dataset

Anzor Gozalishvili, Kristine Eliosidze, Revaz Revazashvili, Amiran Gozalishvili, Dea Gejadze, Salome Javashvili

Space Colonization

- How does space affect living organisms
- Identify challenges by the change of environment
- Strategies of sustaining life and obtaining health
- etc

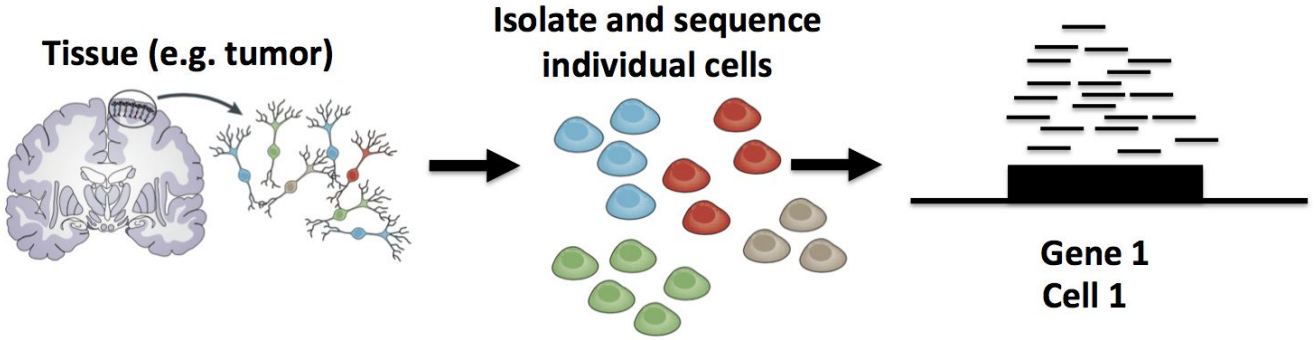
NASA Open Science for Life in Space



Objective: Building the Space Biology “Model Zoo”

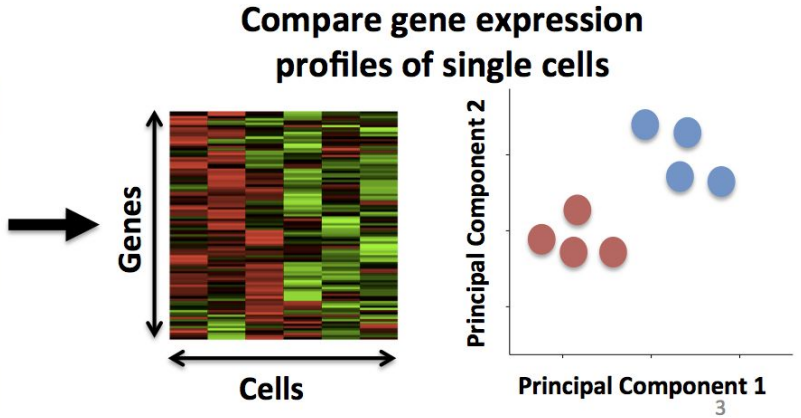
1. to design a comprehensive database of publicly available biomedical datasets that could be used to pretrain different models for a “model zoo”
2. to determine relevant publicly available space biology datasets that could then be used to refine the models to investigate specific space biology questions.

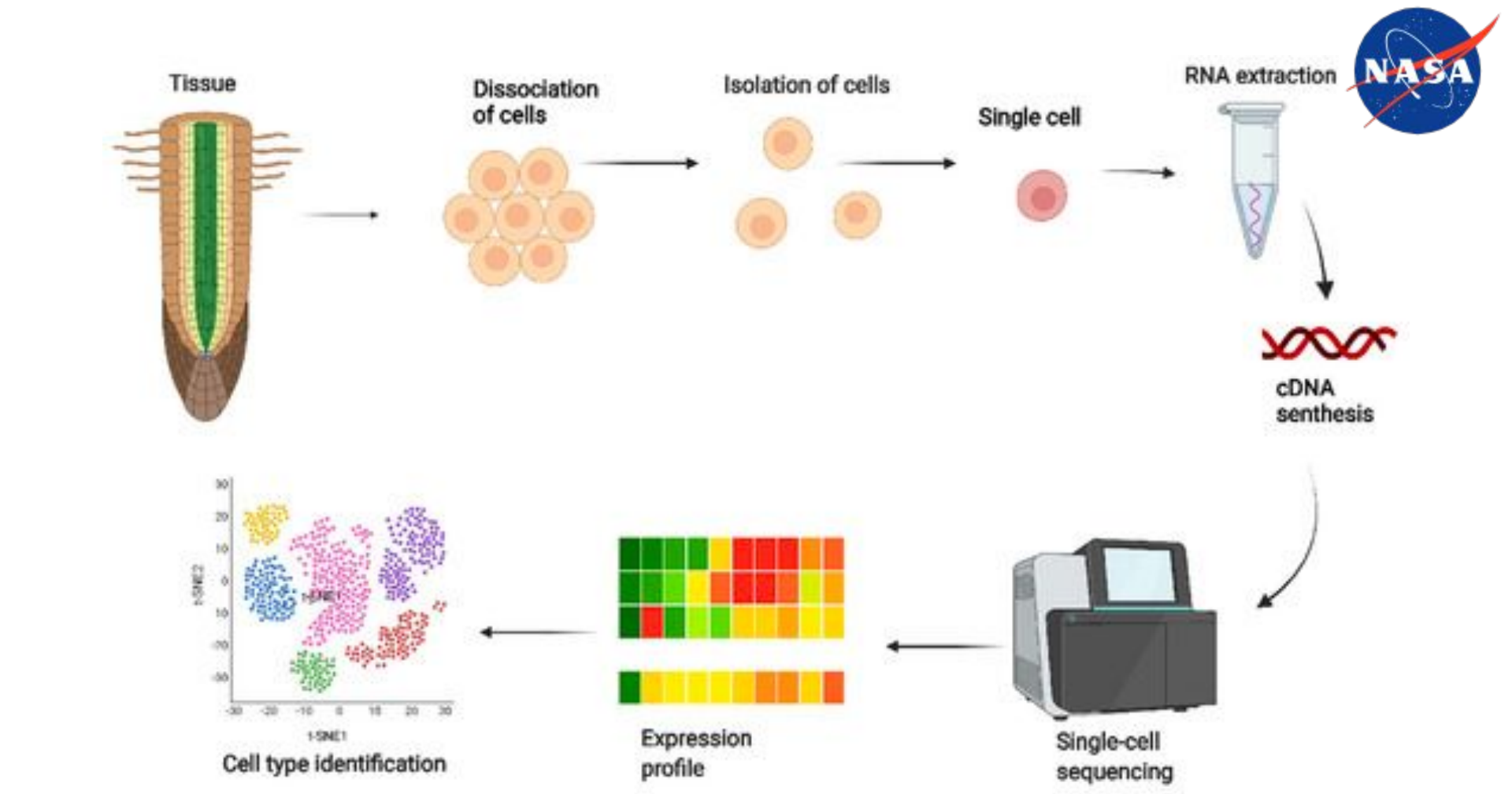
Single-cell RNA-Seq (scRNA-Seq)



Read Counts

	Cell 1	Cell 2	…
Gene 1	18	0	
Gene 2	1010	506	
Gene 3	0	49	
Gene 4	22	0	
…			




Tissue
Dissociation of cells
Isolation of cells
Single cell
RNA extraction
cDNA senthesis
Single-cell sequencing
Expression profile
Cell type identification
t-SNE1
t-SNE2

General Search Filters

Data Source

- [x] GeneLab
- [x] ALSDA
- [ ] NIH GEO
- [ ] EBI PRIDE
- [ ] ANL MG-RAST

Data Type

- [x] Study
- [ ] Experiment
- [ ] Subject
- [ ] Biospecimen
- [ ] Payload
- [ ] Mission
- [ ] Hardware
- [ ] Vehicle

Study Search Filters

Project Type

- [ ] Ground
- [ ] Spaceflight
- [ ] High Altitude

Open Science Data Repository Search

Search Datasets

Sort By: Release Date

Items per page: 25 | 1 – 25 of 456

Study

OSD-654

Toward countering muscle and bone loss with spaceflight: GSK3 as a potential target (Tibia, RR9 and HLU, MicroCT and DXA Scanning)

Organisms	Factors	Assay Types	Release Date	Description
Mus musculus	Spaceflight Space Mission Genotype	Bone Microstructure Bone Microstructure	04-Oct-2023	We examined the effects of ~30 days of spaceflight on glycogen synthase kinase 3 (GSK3) content and inhibitory serine phosphorylation in murine muscle and bone samples from four separate missions (BIO...

Highlights: Image Scan Acquisition Image Processing And Reconstruction *ALSDA* Transformed Data DEXA Scan... with a reusable template, which was created through feedback provided by subject matter experts in the *ALSDA... alsda*

Study

OSD-660

Toward countering muscle and bone loss with spaceflight: GSK3 as a potential target (Tibialis Anterior, RR9, Western Blot)

Organisms	Factors	Assay Types	Release Date	Description
Mus musculus	Spaceflight	protein quantification	26-Sep-2023	We examined the effects of ~30 days of spaceflight on glycogen synthase kinase 3 (GSK3) content and inhibitory serine phosphorylation in murine muscle and bone samples from four separate missions (BIO...

Highlights: Blocking Protocol Labeling Information Western Blot Imaging Western Blot Quantification *ALSDA*... with a reusable template, which was created through feedback provided by subject matter experts in the *ALSDA... alsda*

Assays

Assay Name: Bone Microstructure

Technology Platform: SkyScan 1176 V1 build 12

Technology Type: Micro-Computed Tomography

Select Export Columns

Sample Name	Protocol REF	Parameter Value: Scanner	Parameter Value: Volume Of Interest Location	Parameter Value: Scan Medium	Parameter Value: Contrast Stain Applied	Parameter Value: X-ray Intensity	Parameter Value: Voxel Size	Parameter Val Integration Tir Exposure
FViv16	Image Scan Acquisition	SkyScan 1176 V.1.1 build 12, Bruker microCT, Belgium	proximal tibia and tibia midpoint	air	No	9 micrometer	45 kilovolt	850 millisec
FViv17	Image Scan Acquisition	SkyScan 1176 V.1.1 build 12, Bruker microCT, Belgium	proximal tibia and tibia midpoint	air	No	9 micrometer	45 kilovolt	850 millisec
FViv18	Image Scan Acquisition	SkyScan 1176 V.1.1 build 12, Bruker microCT, Belgium	proximal tibia and tibia midpoint	air	No	9 micrometer	45 kilovolt	850 millisec
FViv19	Image Scan Acquisition	SkyScan 1176 V.1.1 build 12, Bruker microCT, Belgium	proximal tibia and tibia midpoint	air	No	9 micrometer	45 kilovolt	850 millisec
FViv20	Image Scan Acquisition	SkyScan 1176 V.1.1 build 12, Bruker	proximal tibia and tibia midpoint	air	No	9 micrometer	45 kilovolt	850 millisec

Our Dataset (NASA_OSDR)

- Huggingface dataset
- Easy to access (2 lines of code)
- Easy to explore (Huggingface UI)
- Reproducible and shareable work
- Potential usage for model pretraining from 100s of NLP models on huggingface
- Multi-Modal dataset for exploring general patterns
- Can initiate open competition while maintaining leaderboard (papers with code)

Data Accessibility issues in GeneLab

- Experiments with different goals (500 experiments)
  Single experiment level analysis
- Different formats
- Inconvenient structure
- Broken Urls
- Impossible to perform global visualizations in UI
- Not ready for pre-training Models
- No downstream tasks defined

Google Colab

Open in Colab

In [ ]:
```
!pip install datasets
```

In [3]:
```
import datasets
```

In [4]:
```
from datasets import load_dataset
```

In [9]:
```
dataset = load_dataset('anz2/NASA_OSDR')
```

In [17]:
```
dataset['train'][10]
```

Out[17]:
```
{'Sample Name': 'GSM2684068',
 'Protocol REF': 'Nucleic Acid Extraction',
 'Parameter Value: DNA Fragmentation': 'sonication',
 'Parameter Value: DNA Fragment Size': '200-300 base pair',
 'Extract Name': 'GSM2684068',
 'Protocol REF.1': 'Library Construction',
 'Parameter Value: Library Strategy': 'BisPCR2',
 'Parameter Value: Library Selection': 'other',
 'Parameter Value: Library Layout': 'PAIRED',
 'Protocol REF.2': 'Nucleic Acid Sequencing',
 'Parameter Value: Sequencing Instrument': 'Illumina MiSeq',
 'Assay Name': 'BisPCR2',
 'Parameter Value: Read Length': '150 base pair',
 'Raw Data File': 'GLDS-524_wgbs_GSM2684068_R1_raw.fastq.gz, GLDS-524_wgbs_GSM2684068_R2_raw.fastq.gz',
 'Protocol REF.3': 'GeneLab raw data processing protocol',
 'Parameter Value: Read Depth': '199027 read',
 'Parameter Value: MultiQC File Names': 'GLDS-524_Gwgbs_raw_multiqc_report.zip'}
```

Datasets: anz2/NASA_OSDR like 0

License: apache-2.0

Dataset card | Files and versions | Community | Settings

Dataset Viewer

Auto-converted to Parquet | API | Go to dataset viewer

Split

train (25 rows)

Search this dataset

Sample Name string · classes 25 values	Protocol REF string · classes 1 value	Parameter Value: DNA Fragmentation string · classes 1 value	Parameter Value: DNA Fragment Size string · classes 1 value	Extract Name string · classes 25 values	Protocol REF.1 string · classes 1 value	Parameter Value: Library Strategy string · classes 2 values
GSM2684058	Nucleic Acid Extraction	sonication	200-300 base pair	GSM2684058	Library Construction	Bisulfite-Seq
GSM2684059	Nucleic Acid Extraction	sonication	200-300 base pair	GSM2684059	Library Construction	Bisulfite-Seq
GSM2684060	Nucleic Acid Extraction	sonication	200-300 base pair	GSM2684060	Library Construction	Bisulfite-Seq
GSM2684061	Nucleic Acid Extraction	sonication	200-300 base pair	GSM2684061	Library Construction	Bisulfite-Seq
GSM2684062	Nucleic Acid Extraction	sonication	200-300 base pair	GSM2684062	Library Construction	Bisulfite-Seq

Potential Outcomes & Findings?

- Better foundational models for RNA representation learning and contribution to “Model Zoo”
- Find “space-born” RNA’s

Thank You!