Single-Cell genomics data incorporation into agricultural G2P research by building a FAIR data ecosystem

Muskan Kapoor

I am a second year PhD student in Dr Tuggle's research group.
My research interests lies in Bioinformatics, Computational Biology and its applications such as tools development and data analysis.

Email: muskan@iastate.edu

Christopher Tuggle

I am a Professor of Molecular Genetics at the Department of Animal Science. I lead the USDA-funded Pig FAANG project, and my interests are in the functional analysis of the pig genome, with an emphasis on immunity and disease resistance.

Email: cktuggle@iastate.edu

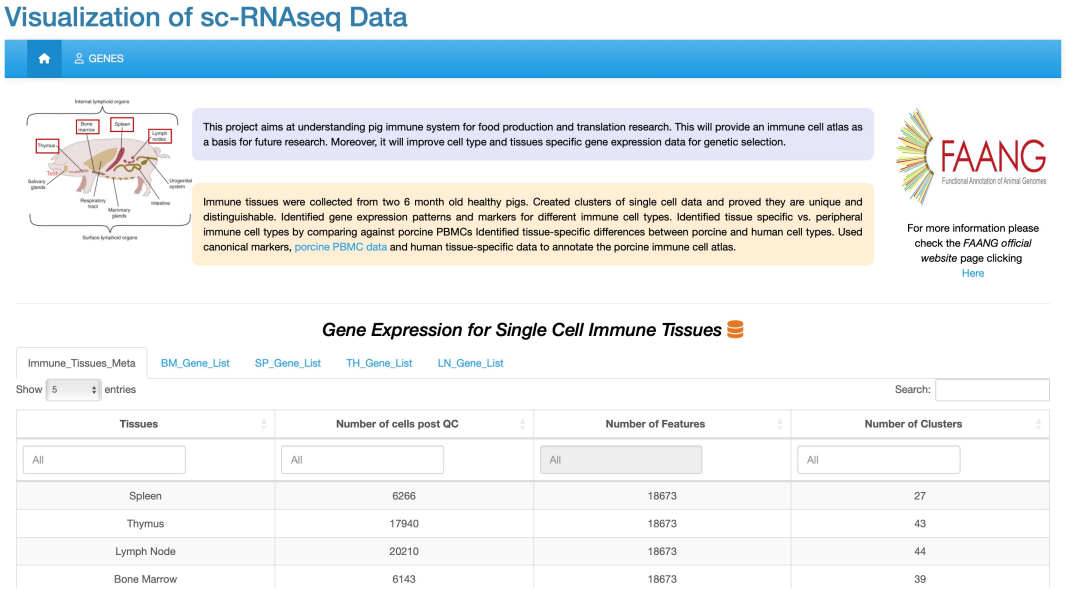
Authors Name

Muskan Kapoor, Christopher Tuggle, Alexey Sokolov, Enrique Ventura, Galabina Yordanova, Nicholas Provart, Irene Papatheodorou, Nancy George, Doreen Ware, Sunita Kumari, Timothy Tickle, Lance Daharsh, James Koltes, Benjamin Cole, Marc Libault, Christine Elsik, Wesley Warren, Tony Burdett, Peter Harrison

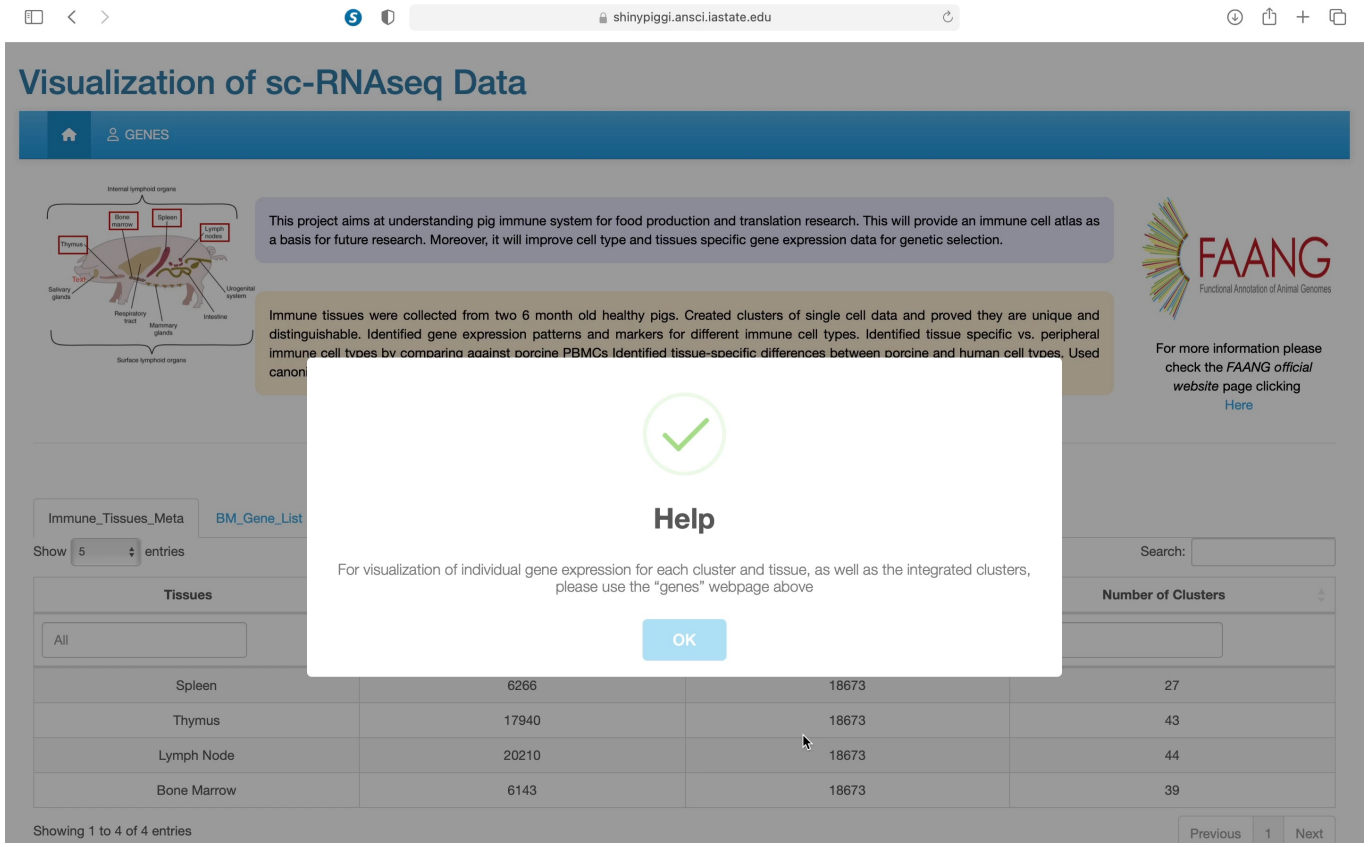
IOWA STATE UNIVERSITY

Visualization Tool for exploring Single-Cell data

- The agriculture genomics community has numerous data analysis tools and standards but limited knowledge describing, visualizing and storing single-cell
  - sc genomics allows transcriptomic profiling of individual cells.
  - Providing good comparative and integrated analysis approach
- A Shiny-based web application, called Shiny-PIGGI,
  - single cell-level transcriptomic study of pig immune tissues and peripheral blood mononuclear cells
  - an important resource for improved annotation of porcine immune genes and cell types



Shiny-PIGGI for visualizing Single-Cell Data

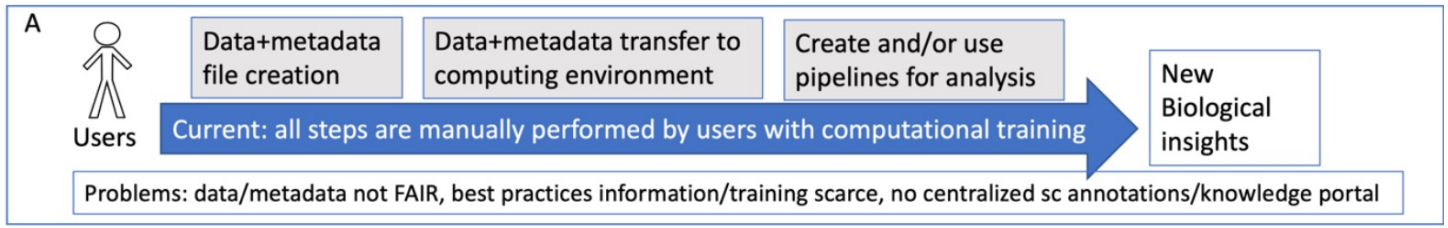


SCAN ME

Fig 1 Created Shiny Tool for visualization of Four tissue mainly Spleen, lymph node, Thymus, Bone marrow, and an integrated object Of four tissues.

The tool is hosted online at : https://shinypiggi.ansci.iastate.edu/

INTRODUCTION



- Single-cell genomics infrastructure efforts, such as the Human Cell Atlas Data Coordination Platform (HCA DCP)
- Resources can benefit our community
- integrated with Terra, a cloud-native workbench for computational biology developed by Broad, Verily, and Microsoft that houses tools for scGenomics analysis
- Pilot scale project for ingestion of scRNAseq with HCA-DCP standards
- Resources (e.g., Terra) can be used to analyze the ingested data.

INTRODUCTION

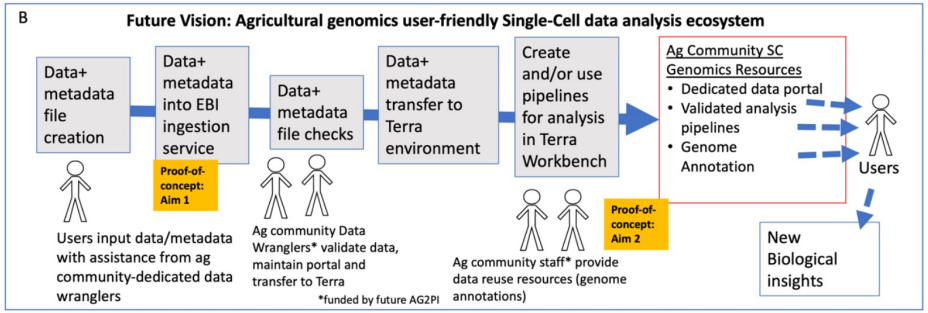
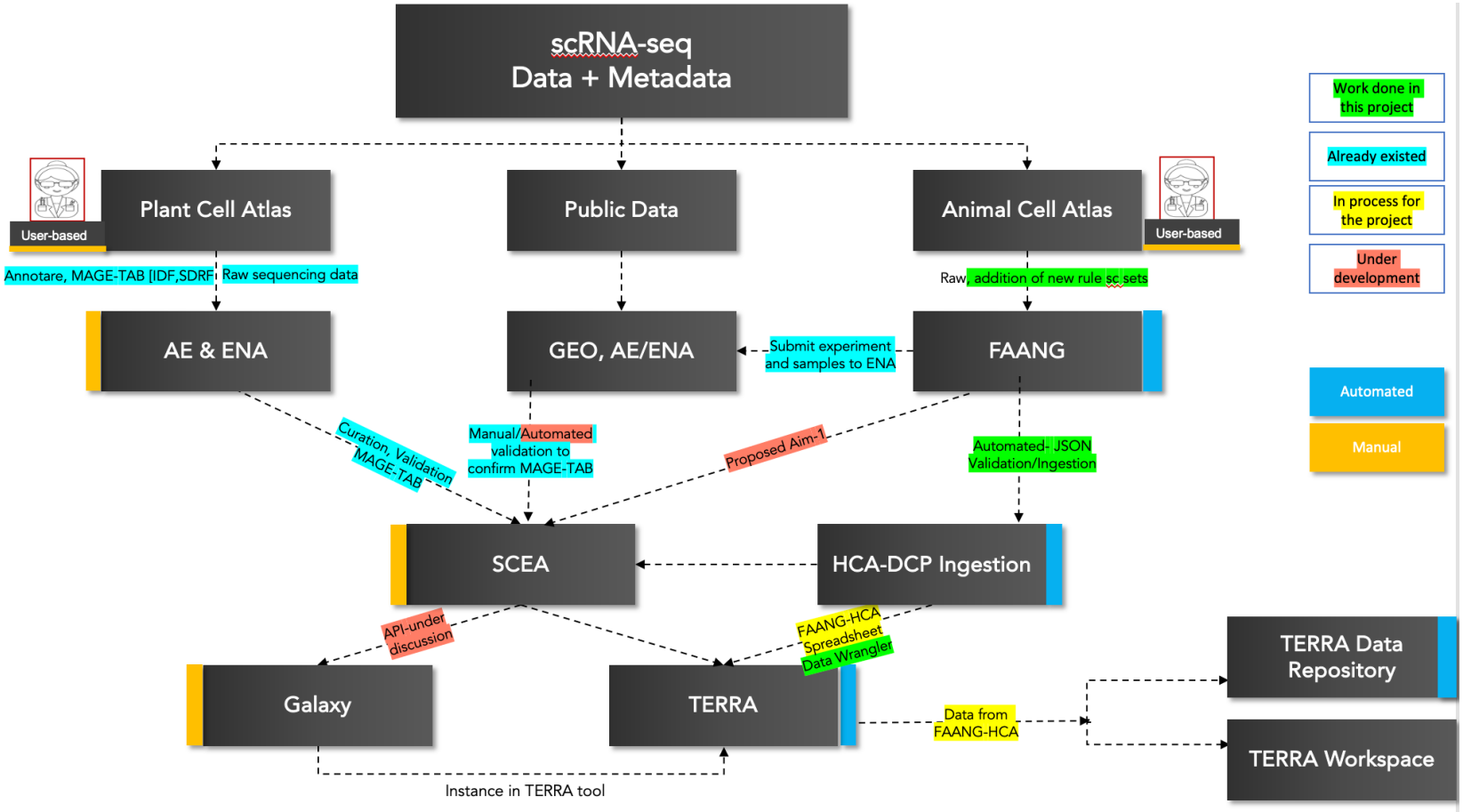


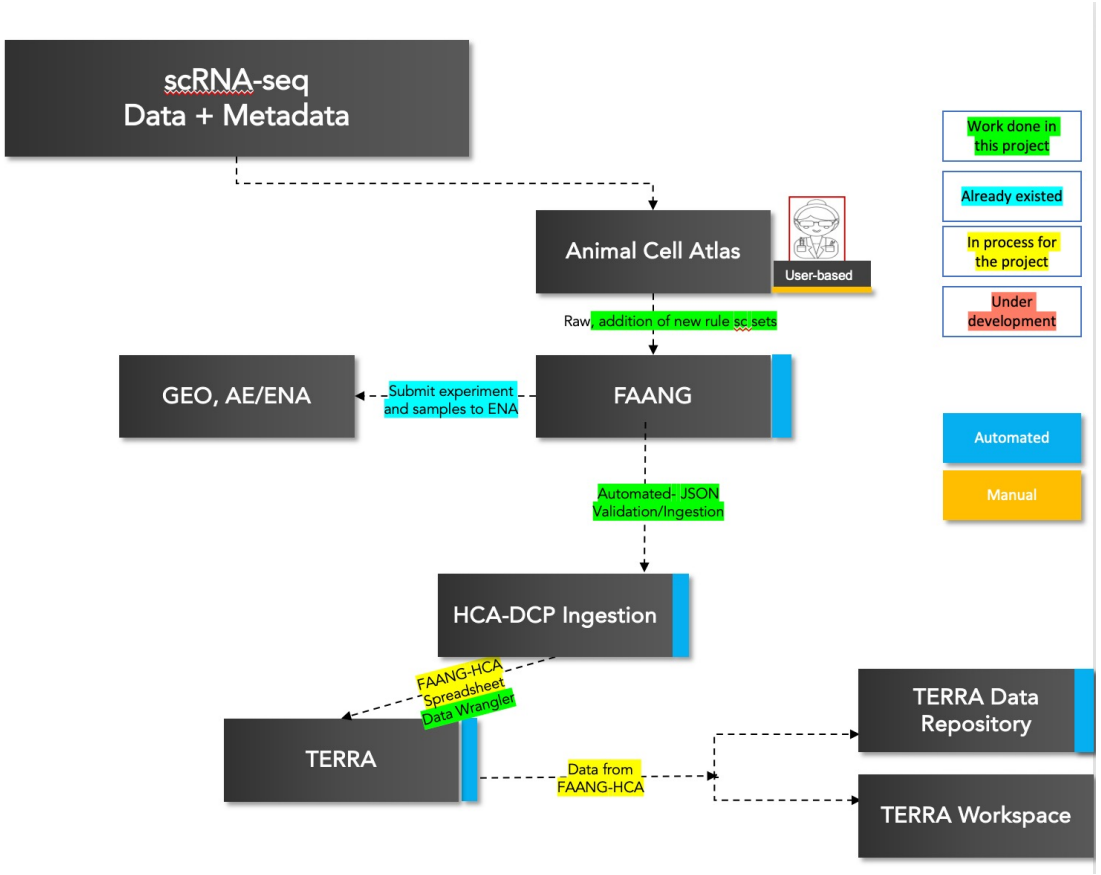
Fig 1. (A) Current Status and (B) Future Vision for Single Cell Data analysis in Agriculture

- Annotare, a data submission tool at EMBL-EBI Currently,
  - the most comprehensive data ingestion portal for high throughput sequencing datasets from plants, fungi, protists, and animals (including humans)
  - ensures that sufficient metadata are collected to enable re-analysis and dissemination via the Single Cell Expression Atlas (SCEA)
- FAANG portal, EMBL-EBI portal are limited to animal datasets
  - provides bulk and scRNAseq data access. Data/metadata can be submitted to the FAANG portal using a semi-automated process
  - files are validated using the HCA DCP metadata and data validation service.
  - and then transferred to Terra for further analysis.

Public data lacking sufficient metadata for efficient reuse

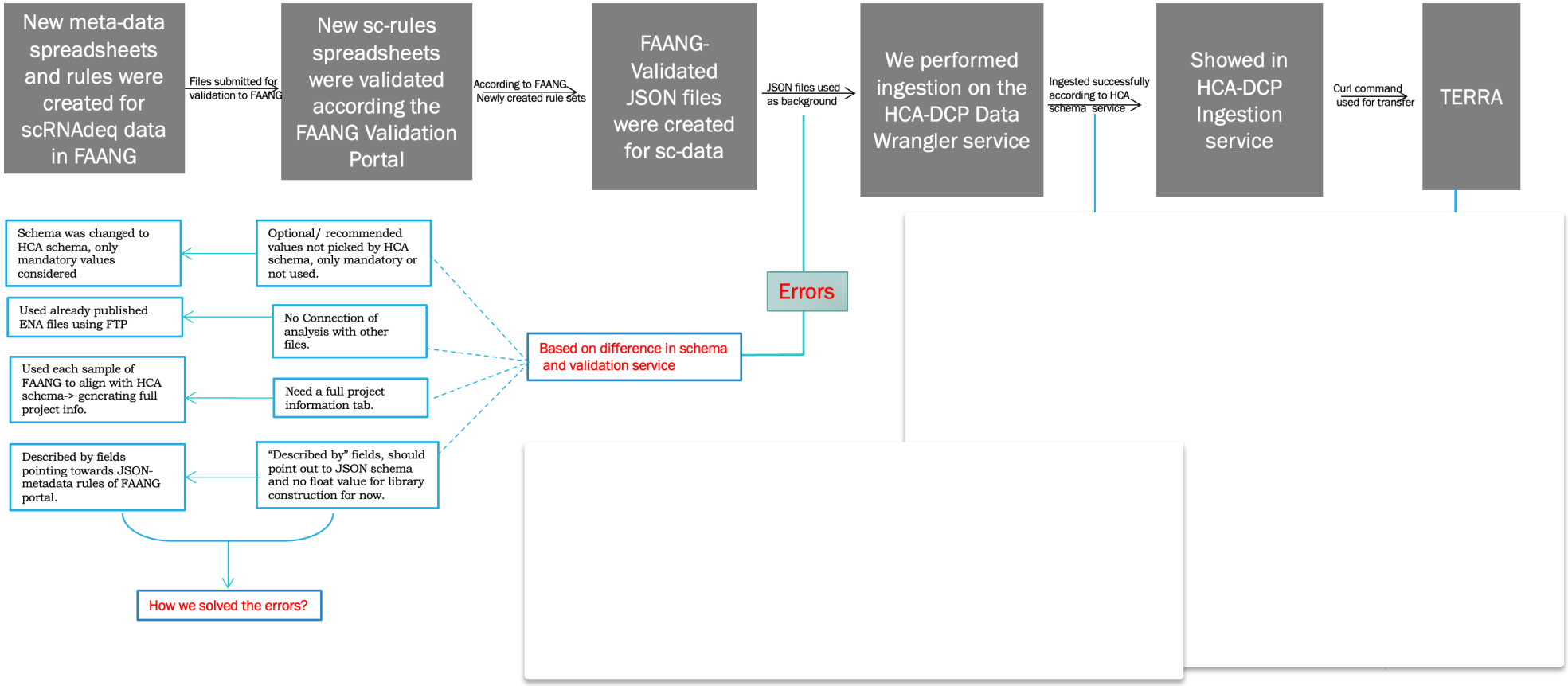


Animal Meta-Data Path

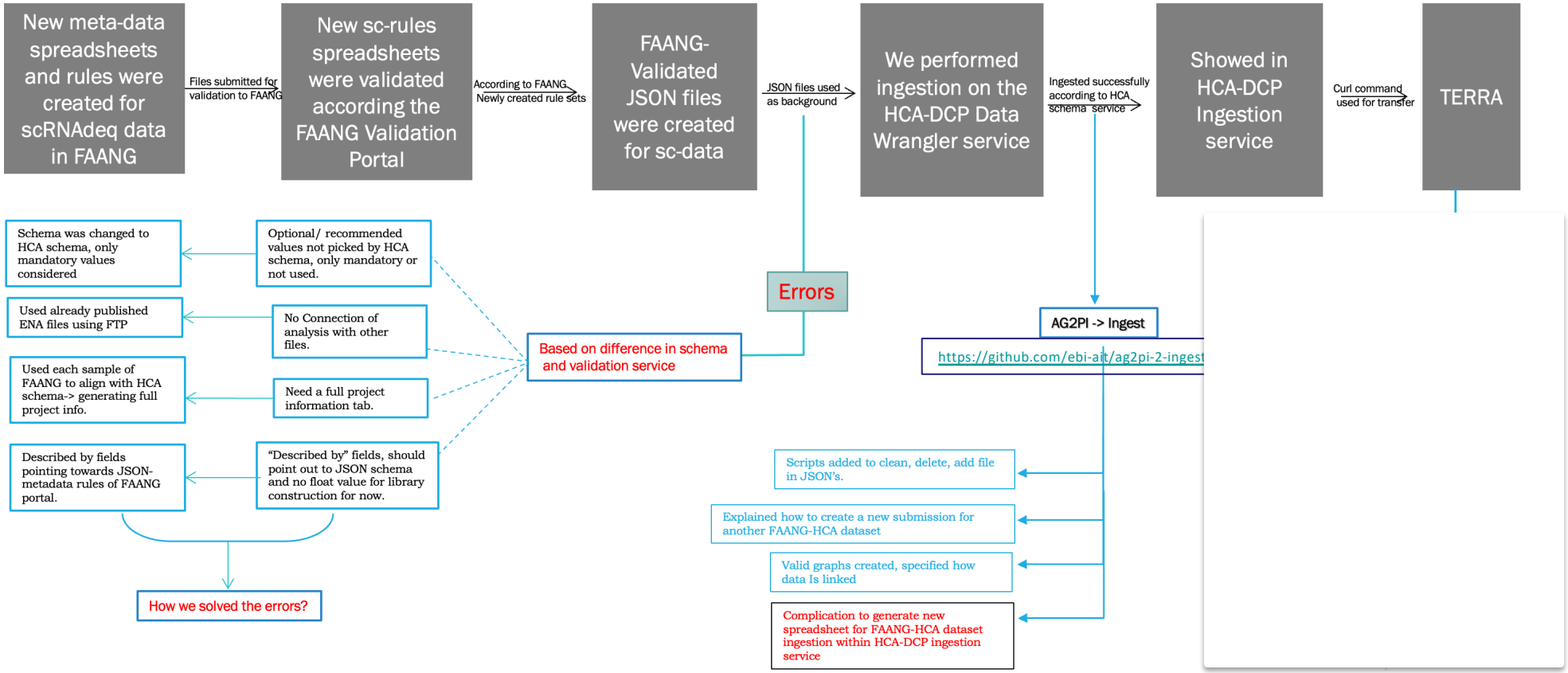


Data Ingestion on Animal Side- Human Cell Atlas

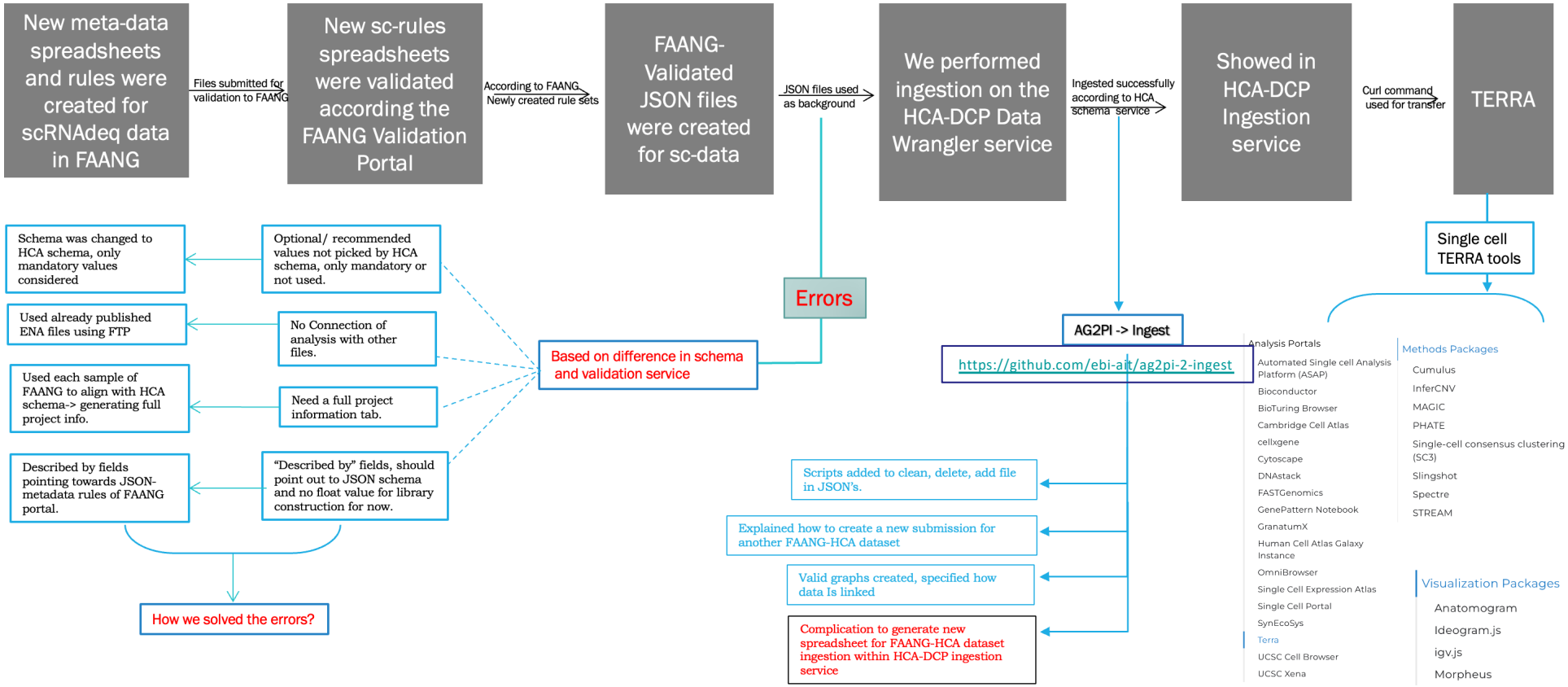
Data Ingestion on Animal Side- Human Cell Atlas



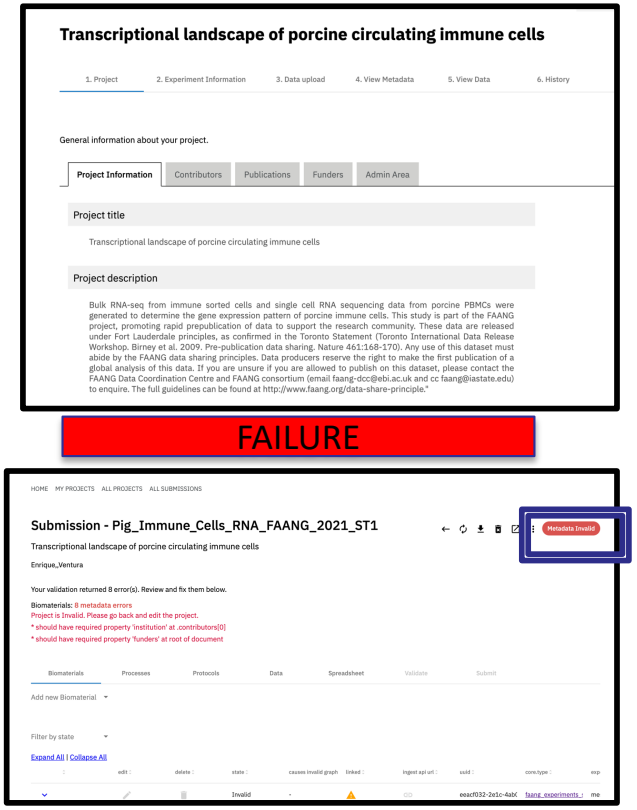
Data Ingestion on Animal Side- Human Cell Atlas



Data Ingestion on Animal Side- Human Cell Atlas



Results on Animal Side- Human Cell Atlas



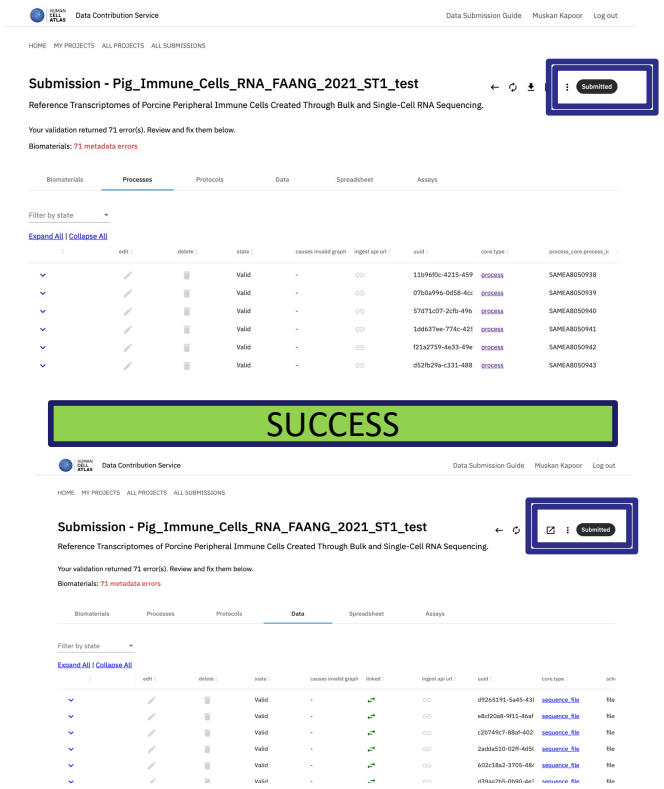
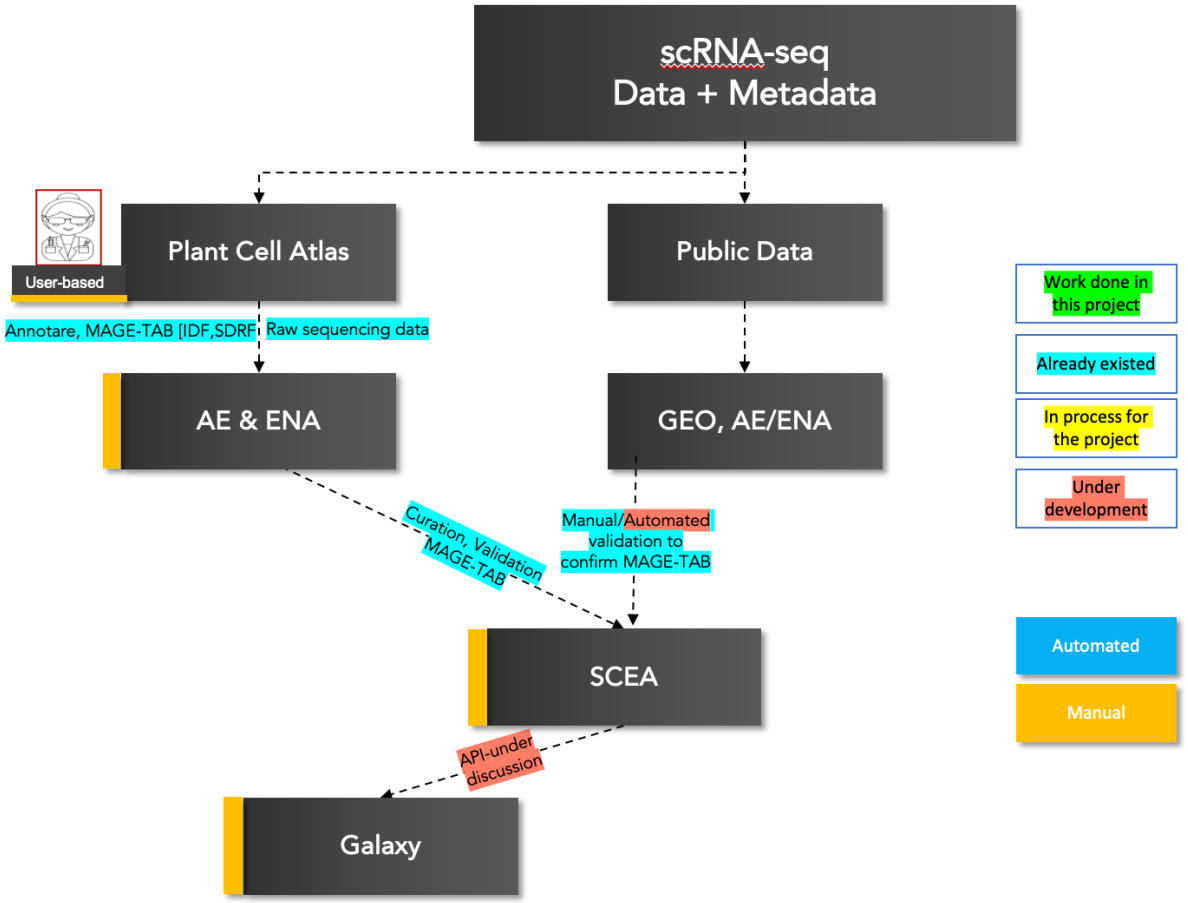


Fig 2 Ingestion in HCA ingestion service according to validated FAANG JSON rules

Fig 3: Ingested Data in HCA –DCP ingestion services

Plant Meta-Data Path



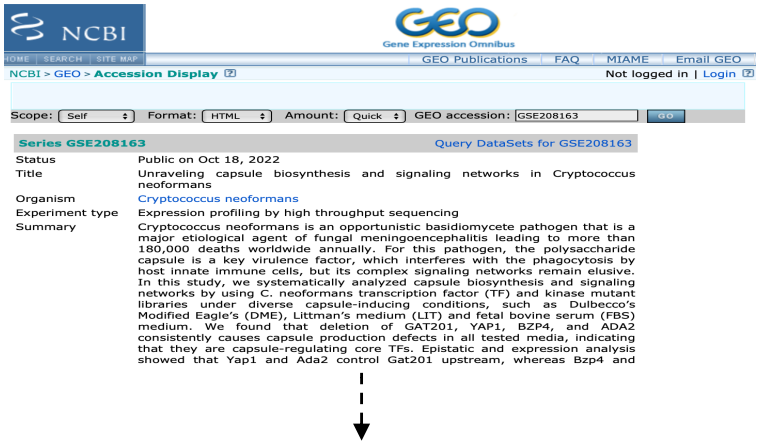
Public Data Ingestion on Animal Side- SCEA

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE208163

Single Cell Pig Data

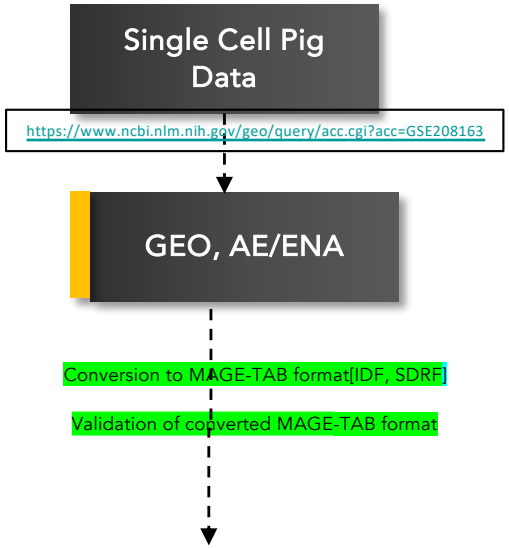
GEO

Conversion to MAGE-TAB format[IDF, SDRF]




```
2023/03/31 18:45:16 INFO Adding age: 7.5 weeks as a characteristic for  GSM6355571
2023/03/31 18:45:16 INFO Adding gender: Female as a characteristic for  GSM6355571
2023/03/31 18:45:16 INFO Skipping download of file NONE
2023/03/31 18:45:16 INFO Adding tissue: Ileum as a characteristic for  GSM6355572
2023/03/31 18:45:16 INFO Adding tissue dissection: Ileum without Peyer's patches as a c
2023/03/31 18:45:16 INFO Adding breed: Mixed-breed as a characteristic for  GSM6355572
2023/03/31 18:45:16 INFO Adding age: 7.5 weeks as a characteristic for  GSM6355572
2023/03/31 18:45:16 INFO Adding gender: Female as a characteristic for  GSM6355572
2023/03/31 18:45:16 INFO Skipping download of file NONE
2023/03/31 18:45:16 INFO Adding tissue: Ileum as a characteristic for  GSM6355573
2023/03/31 18:45:16 INFO Adding tissue dissection: Ileum without Peyer's patches as a c
2023/03/31 18:45:16 INFO Adding breed: Mixed-breed as a characteristic for  GSM6355573
2023/03/31 18:45:16 INFO Adding age: 7.5 weeks as a characteristic for  GSM6355573
2023/03/31 18:45:16 INFO Adding gender: Female as a characteristic for  GSM6355573
```

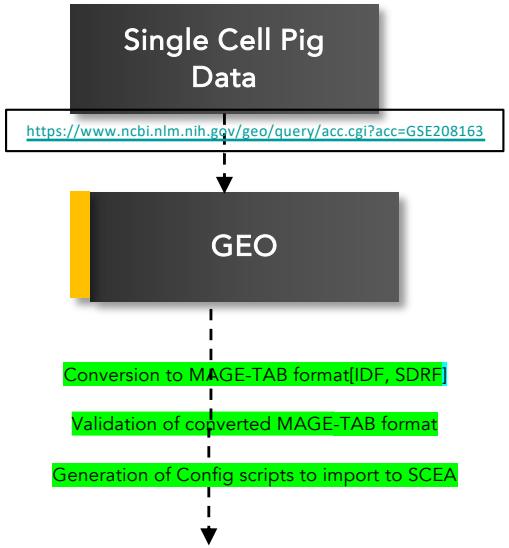
Data Ingestion on Animal Side- SCEA

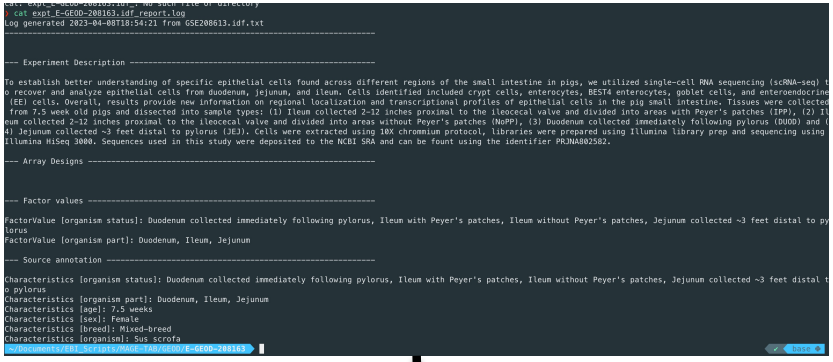



```
2023/03/31 18:45:16 INFO Adding age: 7.5 weeks as a characteristic for  GSM6355571
2023/03/31 18:45:16 INFO Adding gender: Female as a characteristic for  GSM6355571
2023/03/31 18:45:16 INFO Skipping download of file NONE
2023/03/31 18:45:16 INFO Adding tissue: Ileum as a characteristic for  GSM6355572
2023/03/31 18:45:16 INFO Adding tissue dissection: Ileum without Peyer's patches as a c
2023/03/31 18:45:16 INFO Adding breed: Mixed-breed as a characteristic for  GSM6355572
2023/03/31 18:45:16 INFO Adding age: 7.5 weeks as a characteristic for  GSM6355572
2023/03/31 18:45:16 INFO Adding gender: Female as a characteristic for  GSM6355572
2023/03/31 18:45:16 INFO Skipping download of file NONE
2023/03/31 18:45:16 INFO Adding tissue: Ileum as a characteristic for  GSM6355573
2023/03/31 18:45:16 INFO Adding tissue dissection: Ileum without Peyer's patches as a c
2023/03/31 18:45:16 INFO Adding breed: Mixed-breed as a characteristic for  GSM6355573
2023/03/31 18:45:16 INFO Adding age: 7.5 weeks as a characteristic for  GSM6355573
2023/03/31 18:45:16 INFO Adding gender: Female as a characteristic for  GSM6355573
```

```
❯ cat expt_E-GEOD-208163.idf_report.log
Log generated 2023-04-08T18:54:21 from GSE208613.idf.txt
--------------------------------------------------------------------------------


--- Experiment Description ------------------------------------------------------

To establish better understanding of specific epithelial cells found across different regions of the small intestine in pigs, we utilized single-cell RNA sequencing (scRNA-seq) t
o recover and analyze epithelial cells from duodenum, jejunum, and ileum. Cells identified included crypt cells, enterocytes, BEST4 enterocytes, goblet cells, and enteroendocrine
 (EE) cells. Overall, results provide new information on regional localization and transcriptional profiles of epithelial cells in the pig small intestine. Tissues were collected
 from 7.5 week old pigs and dissected into sample types: (1) Ileum collected 2-12 inches proximal to the ileocecal valve and divided into areas with Peyer's patches (IPP), (2) Il
eum collected 2-12 inches proximal to the ileocecal valve and divided into areas without Peyer's patches (NoPP), (3) Duodenum collected immediately following pylorus (DUOD) and (
4) Jejunum collected ~3 feet distal to pylorus (JEJ). Cells were extracted using 10X chrommium protocol, libraries were prepared using Illumina library prep and sequencing using
Illumina HiSeq 3000. Sequences used in this study were deposited to the NCBI SRA and can be fount using the identifier PRJNA802582.

--- Array Designs ---------------------------------------------------------------



--- Factor values ---------------------------------------------------------------

FactorValue [organism status]: Duodenum collected immediately following pylorus, Ileum with Peyer's patches, Ileum without Peyer's patches, Jejunum collected ~3 feet distal to py
lorus
FactorValue [organism part]: Duodenum, Ileum, Jejunum

--- Source annotation -----------------------------------------------------------

Characteristics [organism status]: Duodenum collected immediately following pylorus, Ileum with Peyer's patches, Ileum without Peyer's patches, Jejunum collected ~3 feet distal t
o pylorus
Characteristics [organism part]: Duodenum, Ileum, Jejunum
Characteristics [age]: 7.5 weeks
Characteristics [sex]: Female
Characteristics [breed]: Mixed-breed
Characteristics [organism]: Sus scrofa
~/Documents/EBI_Scripts/MAGE-TAB/GEOD/E-GEOD-208163                                                                                    base
```

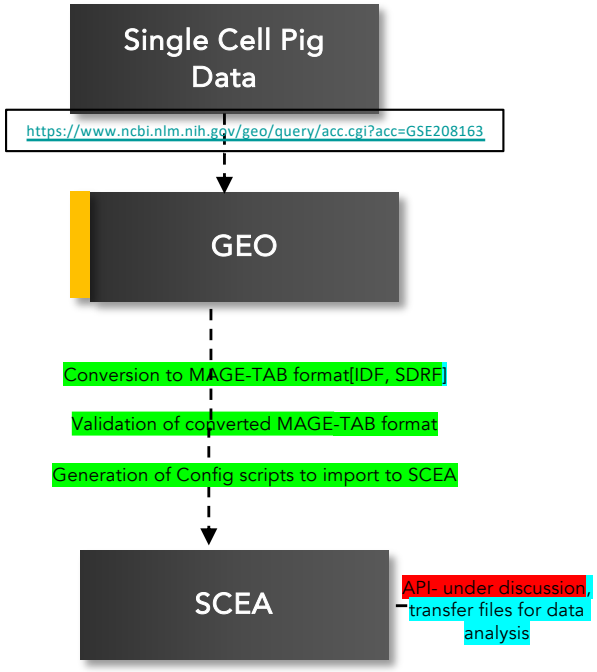
Data Ingestion on Animal Side- SCEA

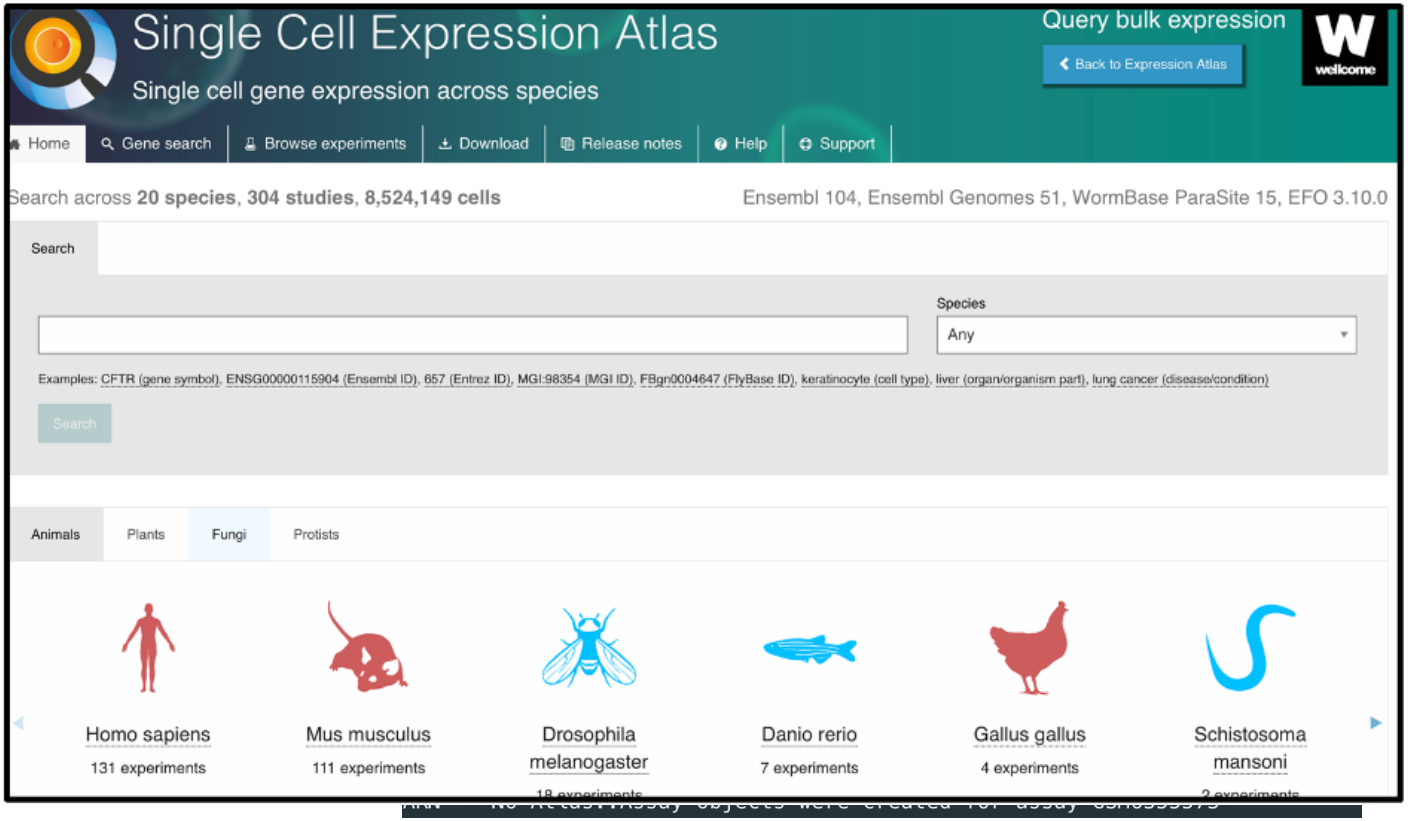




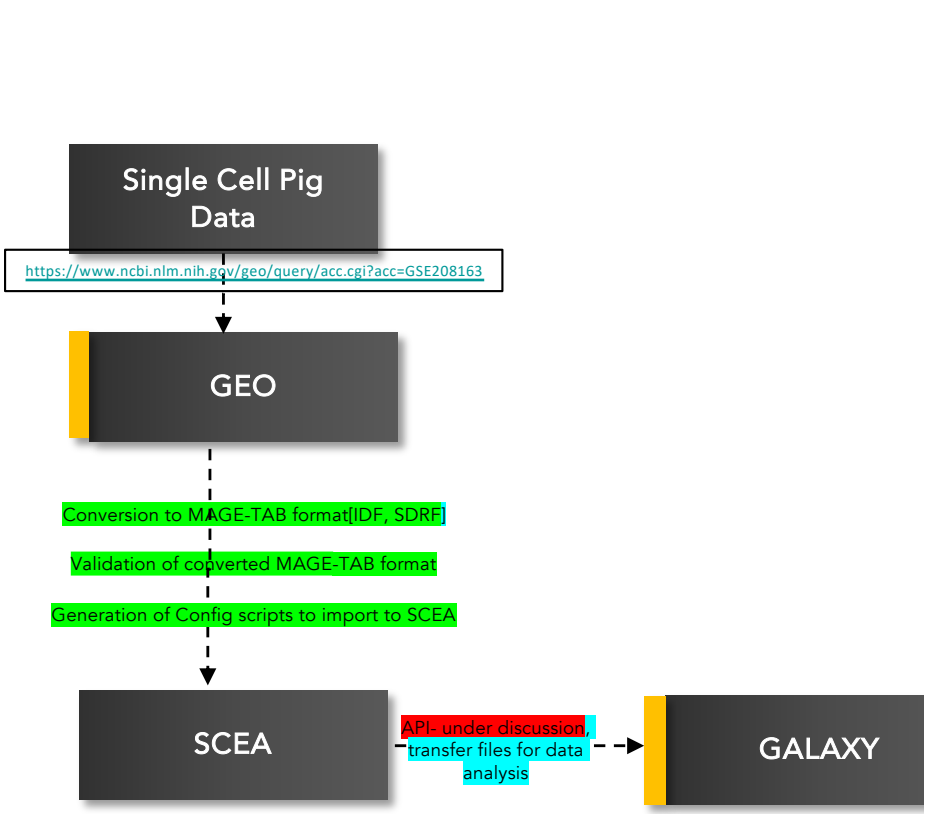

```
 cat atlas_configuration_generation_E-GEOD-208163.log
tlas config generation log for E-GEOD-208163 created at 2023-04-19T23:22:23
--------------------------------------------------------------------------------

NFO  - Reading config for reference factor values and factor types to ignor
ay_group_factor_values.xml
NFO  - Reading MAGE-TAB from /Users/muskankapoor/Documents/EBI_Scripts/MAGE
ARN  - Skipping assay GSM6355560 because it has no Scan node.
ARN  - No Atlas::Assay objects were created for assay GSM6355560
ARN  - No Atlas::Assay objects were created for assay GSM6355560
ARN  - Skipping assay GSM6355572 because it has no Scan node.
ARN  - No Atlas::Assay objects were created for assay GSM6355572
ARN  - No Atlas::Assay objects were created for assay GSM6355572
ARN  - Skipping assay GSM6355573 because it has no Scan node.
ARN  - No Atlas::Assay objects were created for assay GSM6355573
```

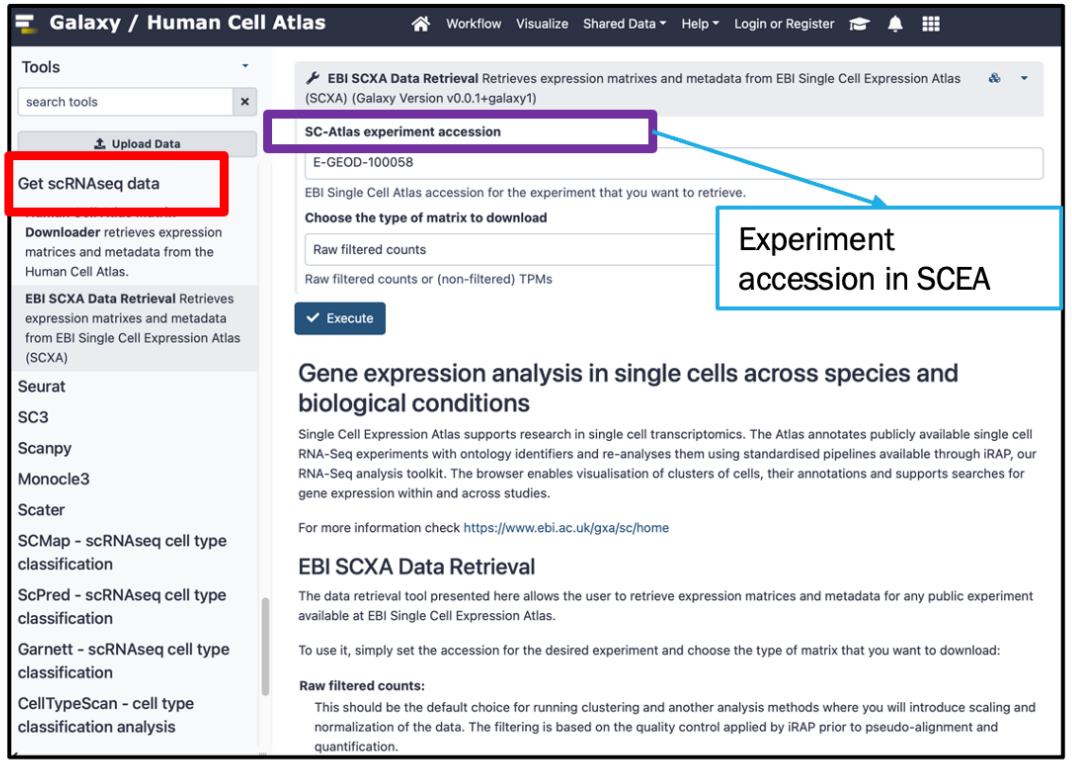
Data Ingestion on Animal Side- SCEA





Data Ingestion on Animal Side- SCEA





Conclusions and Future Scope

We intend to further build upon these existing tools to construct a scientist-friendly data resource and analytical ecosystem to facilitate single cell-level genomic analysis through data ingestion, storage, retrieval, re-use, visualization, and comparative annotation across agricultural species.

- **Animal data**: we will complete the data ingestion into Human Cell Atlas –DCP and test the use of the data in the Terra environment.
- **Plant path improvements:** reducing the manual curation and validation needed to transfer data to FAANG -> $\boldsymbol{SCEA}$
- **Shiny-PIGGI:** we will add the Cluster annotation functionality to the tool and ask for user feedback.

Acknowledgements

Muskan Kapoor[1], Alexey Sokolov[2], Enrique Sapena Ventura[2], Galabina Yordanova[2], Nicholas J. Provart[3], Irene Papatheodorou[2], Nancy George[2], Doreen Ware[4,5], Sunita Kumari[4], Timothy Tickle[6], James Koltes,[1] Benjamin Cole[7], Marc Libault[8], Christine Elsik[9], Wesley Warren[10], Tony Burdett[2], Peter Harrison[2], and Christopher Tuggle[1]

[1]Bioinformatics and Computational Biology Program, Department of Animal Science, Iowa State University, Ames, IA 50011, U.S.A.
[2]EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, 12 U.K.
[3]Department of Cell and Systems Biology/Centre for the Analysis of Genome Evolution and Function, University of Toronto, Toronto, ON M5S 3B2, Canada.
[4]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 11724, USA.
[5]U.S. Department of Agriculture, Agricultural Research Service, NEA Robert W. Holley Center for Agriculture and Health, Cornell University, Ithaca, NY 14853, USA.
[6]Data Sciences Platform, The Broad Institute of MIT and Harvard, 415 Main Street, 21 Cambridge, MA 02142, U.S.A.
[7]DOE-Joint Genome Institute, Lawrence Berkeley National Laboratory, 1, Cyclotron Road, 16 Berkeley CA 94720, U.S.A.
[8]Department of Agronomy and Horticulture, Beadle Center N305, University of Nebraska-Lincoln, Lincoln NE 68588-0660, U.S.A.
[9]Division of Animal Science and Division of Plant Science and Technology, S134D Animal Science Research Center, University of Missouri-Columbia, Columbia, MO 65211
[10]Division of Animal Science, 440G/446 Life Sciences Center, University of Missouri-Columbia, Columbia, MO 65211

AG2PI
Agricultural Genome to Phenome Initiative

USDA

FAANG
Functional Annotation of Animal Genomes

Thank You!