

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 lie 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 Provar, 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

Visualization of sc-RNAseq Data

Gene Expression for Single Cell Immune Tissues

Immune_Tissues_Meta BM_Gene_List SP_Gene_List TH_Gene_List LN_Gene_List

Show 5 entries Search:

Tissues	Number of cells post QC	Number of Features	Number of Clusters
All	All	All	All
Spleen	6266	18673	27
Thymus	17940	18673	43
Lymph Node	20210	18673	44
Bone Marrow	6143	18673	39

Shiny-PIGGI for visualizing Single-Cell Data

Visualization of sc-RNAseq Data

GENES

Internal lymphoid organs: Thymus, Bone marrow, Spleen, Lymph nodes

Surface lymphoid organs: Salivary glands, Respiratory tract, Mucosal glands, Intestine, Urinogenital system

This project aims at understanding pig immune system for food production and translation research. This will provide an immune cell atlas as a basis for future research. Moreover, it will improve cell type and tissues specific gene expression data for genetic selection.

Immune tissues were collected from two 6 month old healthy pigs. Created clusters of single cell data and proved they are unique and distinguishable. Identified gene expression patterns and markers for different immune cell types. Identified tissue specific vs. peripheral immune cell types by comparing against porcine PBMCs. Identified tissue-specific differences between porcine and human cell types. Used canonical gene sets for cell type identification.

FAANG: Functional Annotation of Animal Genomes

For more information please check the FAANG official website page clicking [Here](#)

Help

For visualization of individual gene expression for each cluster and tissue, as well as the integrated clusters, please use the "genes" webpage above

OK

Immune_Tissues_Meta | BM_Gene_List

Show 5 entries

Tissues				
All				
Spleen	6266	18673		27
Thymus	17940	18673		43
Lymph Node	20210	18673		44
Bone Marrow	6143	18673		39

Showing 1 to 4 of 4 entries

Search: []

Number of Clusters []

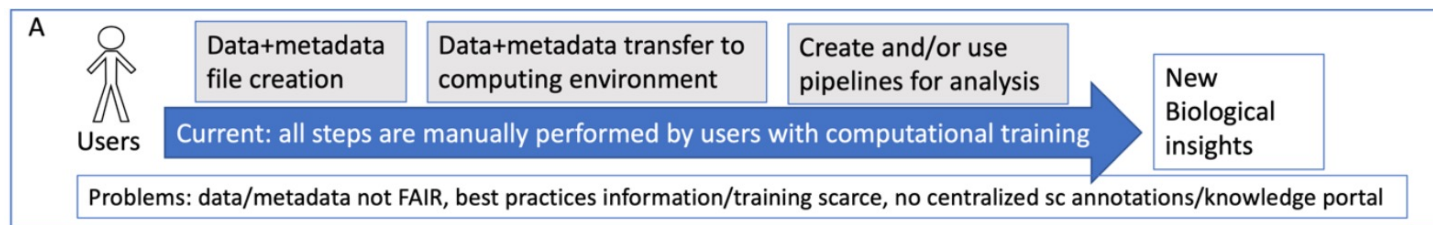
Previous 1 Next



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

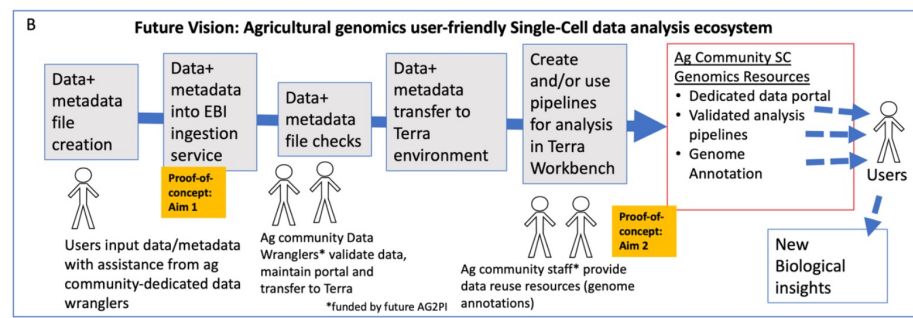
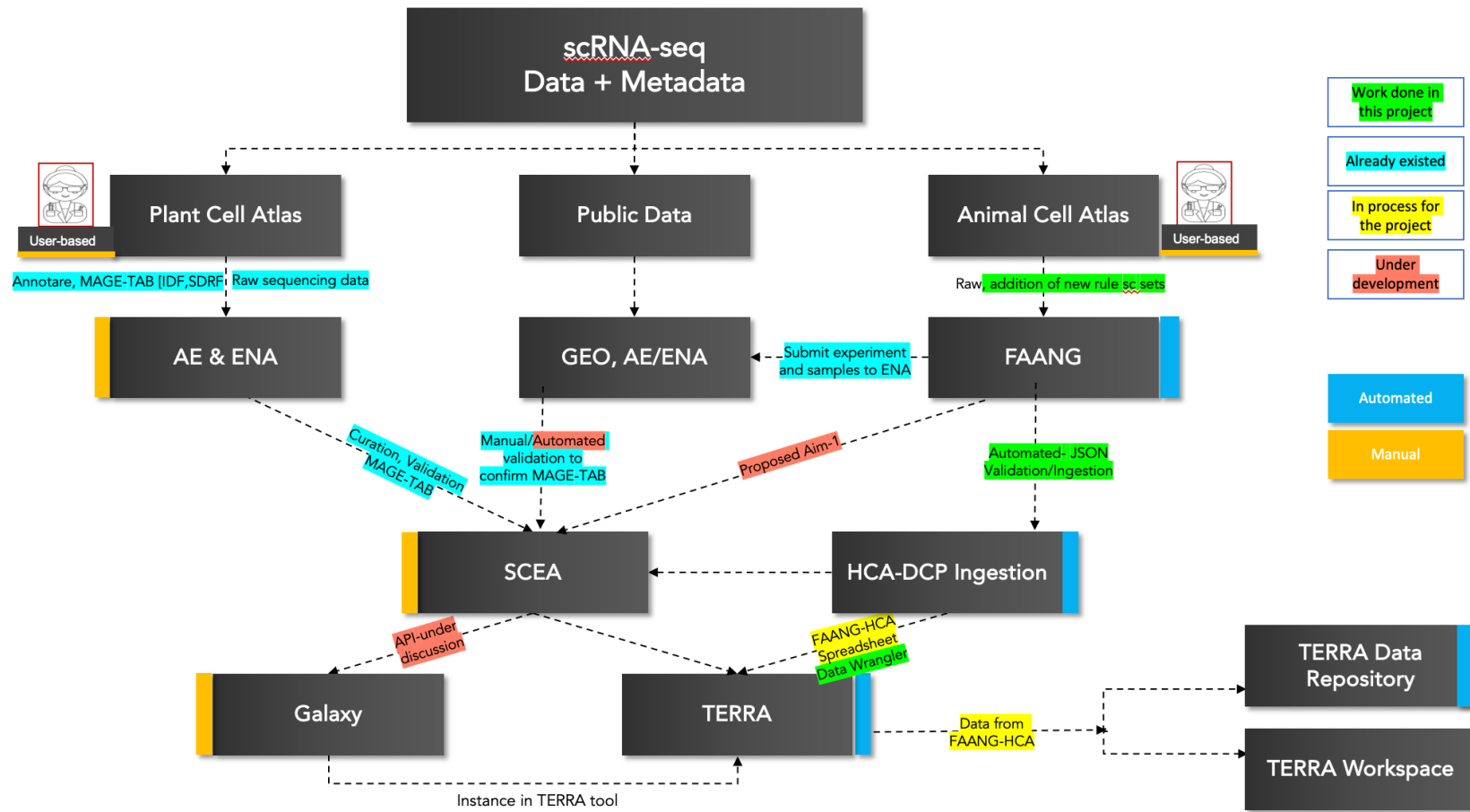


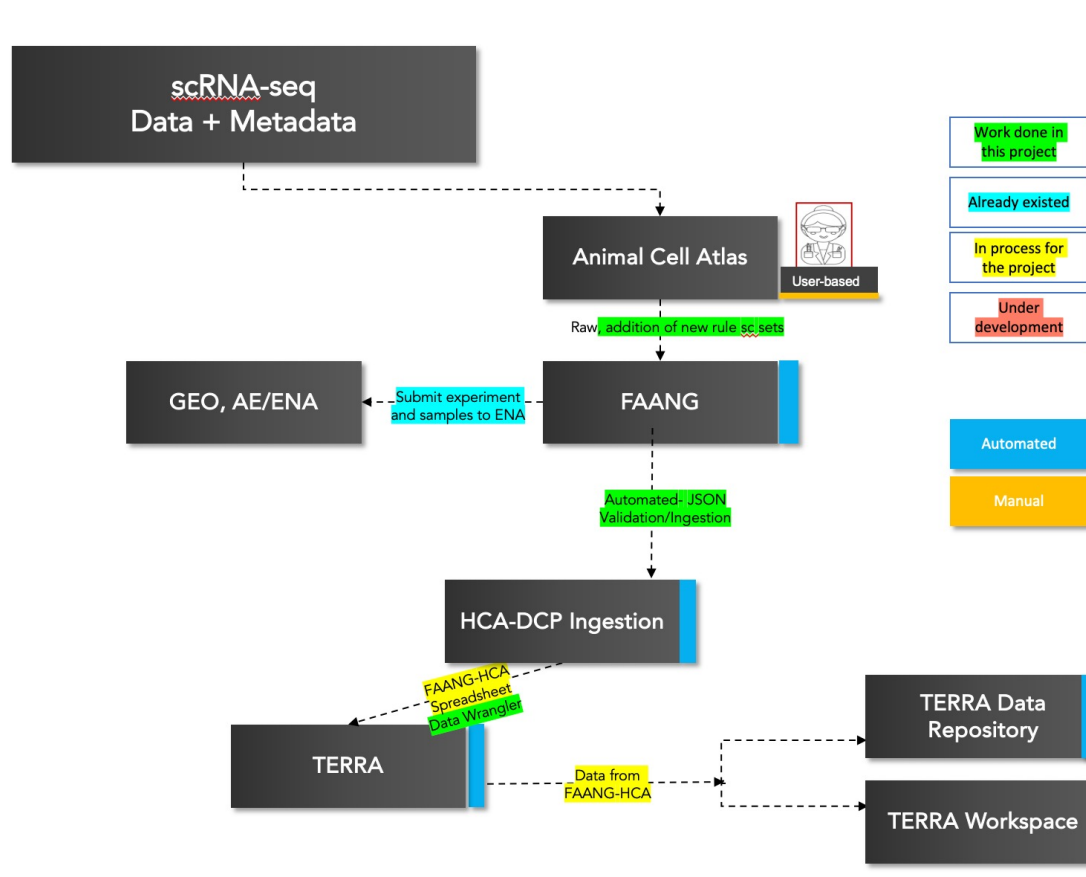
Fig 1. (A) Current Status and (B) Future Vision for Single Cell Data analysis in Agriculture

- 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.

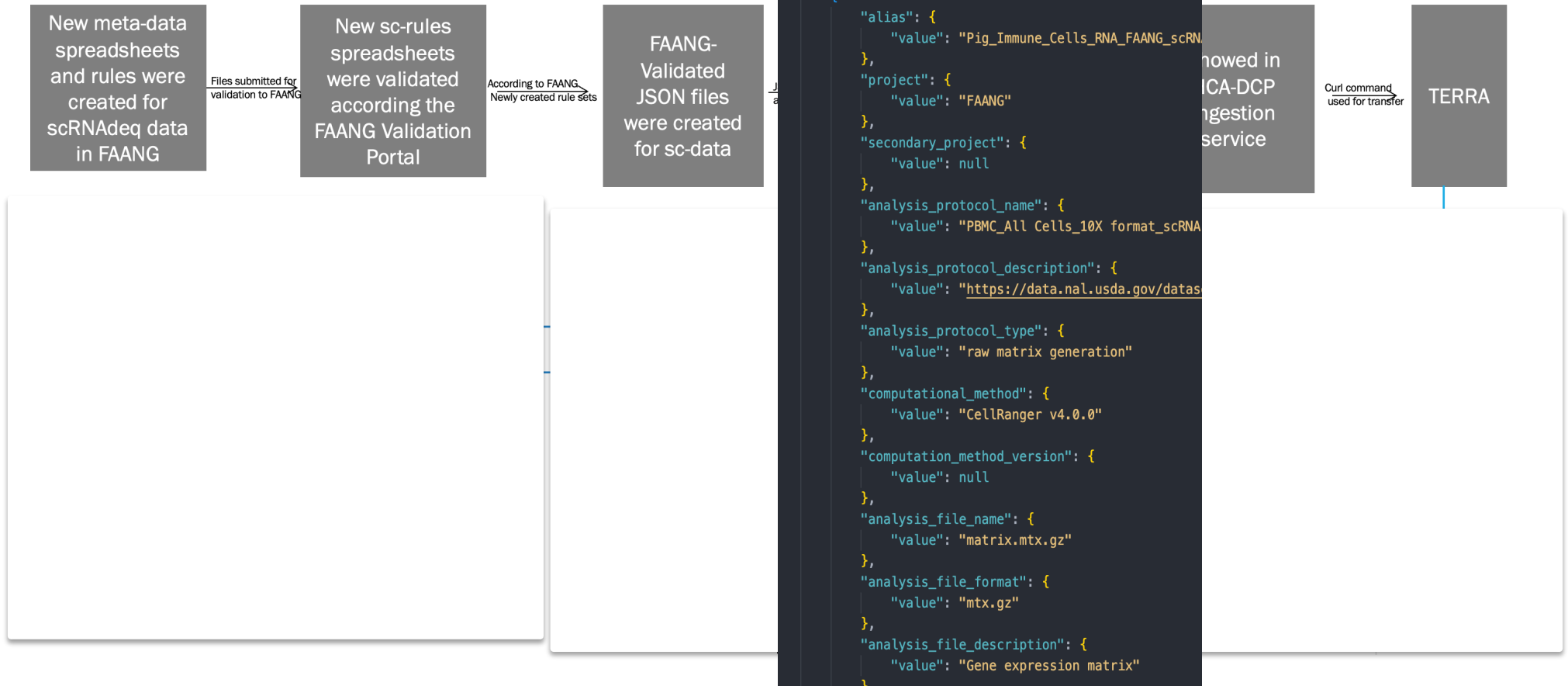
Public data lacking sufficient metadata for efficient reuse



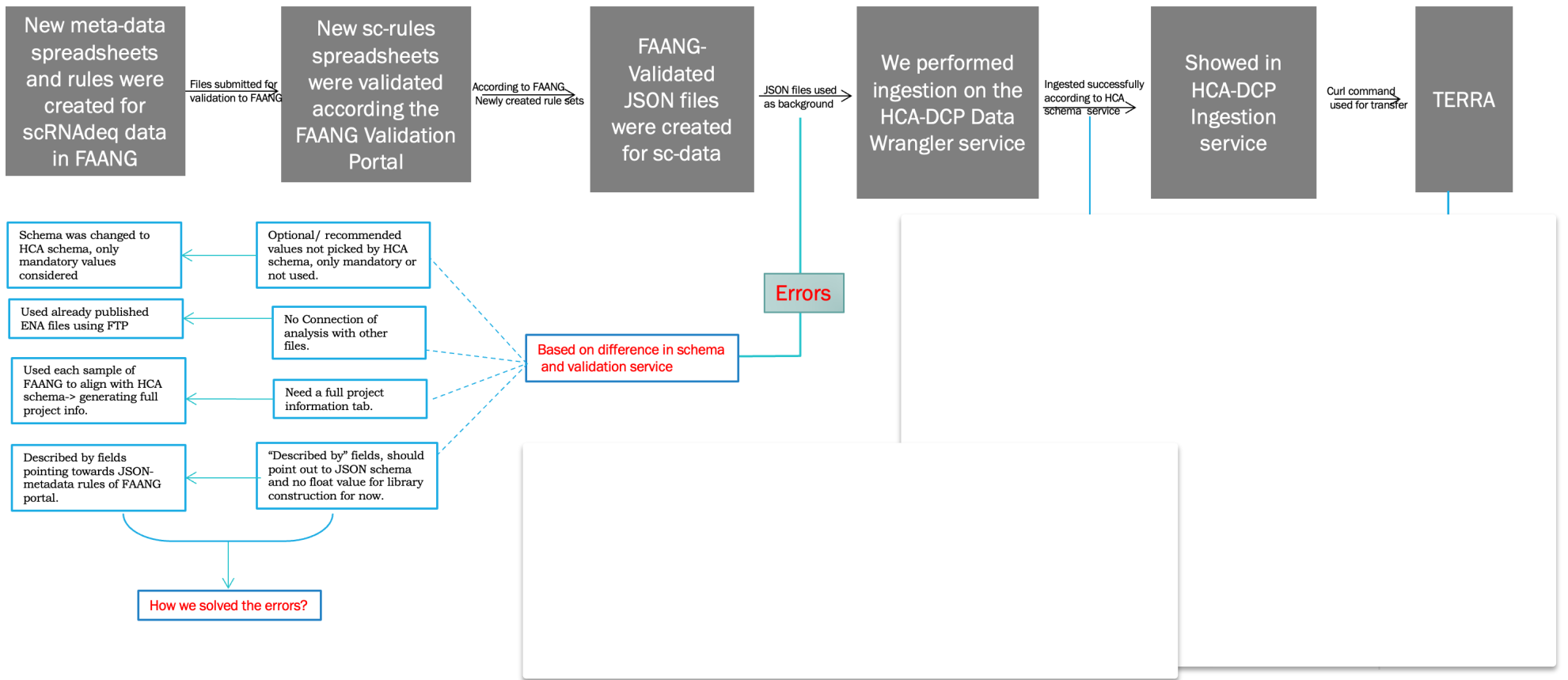
Animal Meta-Data Path



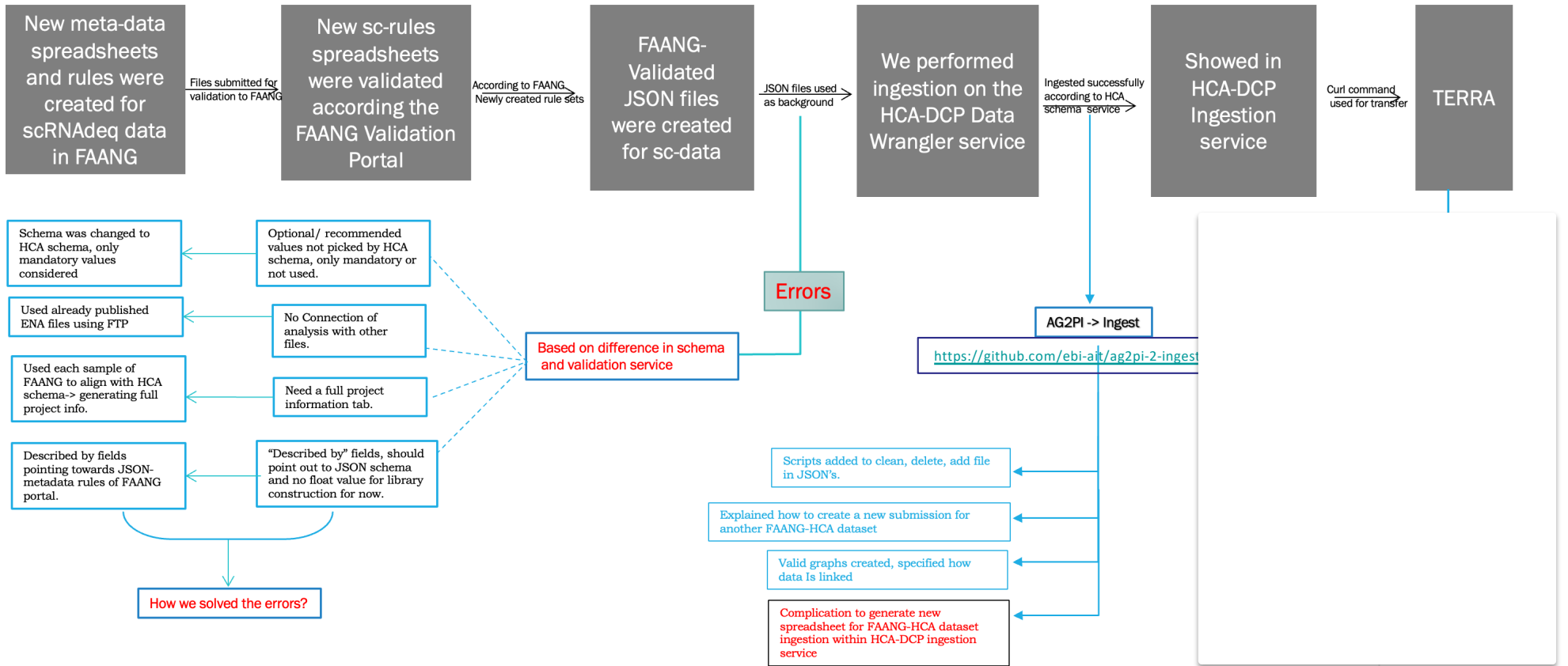
Data Ingestion on Animal Side- Human Cell Atlas



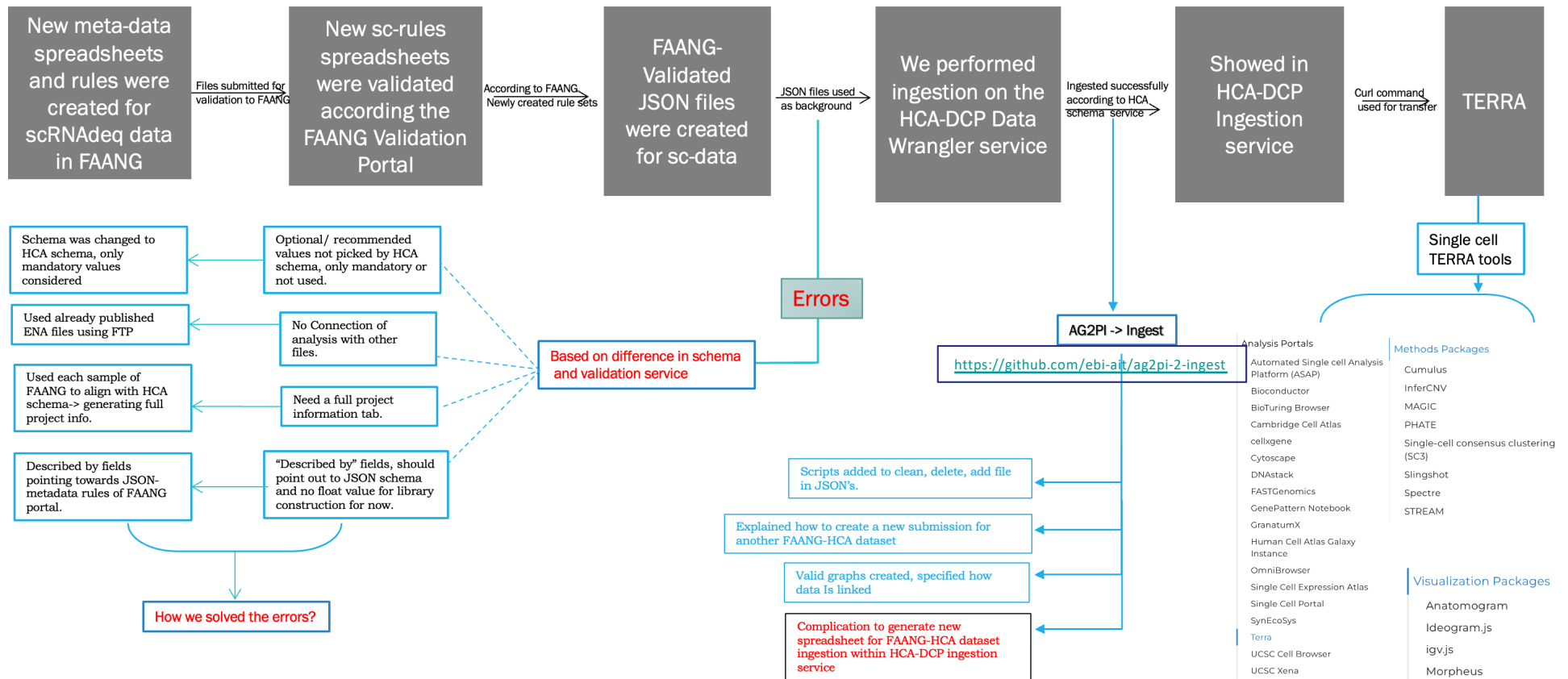
Data Ingestion on Animal Side- Human Cell Atlas



Data Ingestion on Animal Side- Human Cell Atlas



Data Ingestion on Animal Side- Human Cell Atlas



Results on Animal Side- Human Cell Atlas

Transcriptional landscape of porcine circulating immune cells

1. Project 2. Experiment Information 3. Data upload 4. View Metadata 5. View Data 6. History

General information about your project.

Project Information Contributors Publications Funders Admin Area

Project title
Transcriptional landscape of porcine circulating immune cells

Project description
Bulk RNA-seq from immune sorted cells and single cell RNA sequencing data from porcine PBMCs were generated to determine the gene expression pattern of porcine immune cells. This study is part of the FAANG project, promoting rapid prepublication of data to support the research community. These data are released under Fort Lauderdale principles, as confirmed in the Toronto Statement (Toronto International Data Release Workshop, Birney et al. 2009. Pre-publication data sharing. Nature 461:168-170). Any use of this dataset must abide by the FAANG data sharing principles. Data producers reserve the right to make the first publication of a global analysis of this data. If you are unsure if you are allowed to publish on this dataset, please contact the FAANG Data Coordination Centre and FAANG consortium (email: faang-dcc@ebi.ac.uk and cc_faang@iastate.edu) to enquire. The full guidelines can be found at <http://www.faang.org/data-share-principles>.

FAILURE

HOME MY PROJECTS ALL PROJECTS ALL SUBMISSIONS

Submission - Pig_Immune_Cells_RNA_FAANG_2021_ST1 Metadata Invalid

Transcriptional landscape of porcine circulating immune cells
Enrique_Ventura

Your validation returned 8 error(s). Review and fix them below.

Biomaterials: 8 metadata errors
Project is invalid. Please go back and edit the project.
* should have required property 'institution' at .contributors[0]
* should have required property 'funders' at root of document

Biomaterials Processes Protocols Data Spreadsheet Validation Submit

Add new Biomaterial

Filter by state

Expand All | Collapse All

	edit	delete	state	causes invalid graph	linked	ingest api url	valid	core type	exp
invalid									xxxx12-242-442 faang_experiment_016

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



HOME MY PROJECTS ALL PROJECTS ALL SUBMISSIONS

Submission - Pig_Immune_Cells_RNA_FAANG_2021_ST1 Submitted

Reference Transcripts of Porcine Peripheral Immune Cells Created Through Bulk and Single-Cell RNA Sequencing.

Your validation returned 71 error(s). Review and fix them below.

Biomaterials: 71 metadata errors

Biomaterials Processes Protocols Data Spreadsheet Assays

Filter by state

Expand All | Collapse All

	edit	delete	state	causes invalid graph	linked	ingest api url	valid	core type	process_core_process_
Valid									1298500-4215-459 000038 SAMEAS09038
Valid									0765d996-0d58-4cc 000038 SAMEAS09039
Valid									57871d07-2d7b-4796 000038 SAMEAS09040
Valid									148687ee-774c-421 000038 SAMEAS09041
Valid									121d2759-4e33-479e 000038 SAMEAS09042
Valid									d928279c-c311-488 000038 SAMEAS09043

SUCCESS

HOME MY PROJECTS ALL PROJECTS ALL SUBMISSIONS

Submission - Pig_Immune_Cells_RNA_FAANG_2021_ST1 Submitted

Reference Transcripts of Porcine Peripheral Immune Cells Created Through Bulk and Single-Cell RNA Sequencing.

Your validation returned 71 error(s). Review and fix them below.

Biomaterials: 71 metadata errors

Biomaterials Processes Protocols Data Spreadsheet Assays

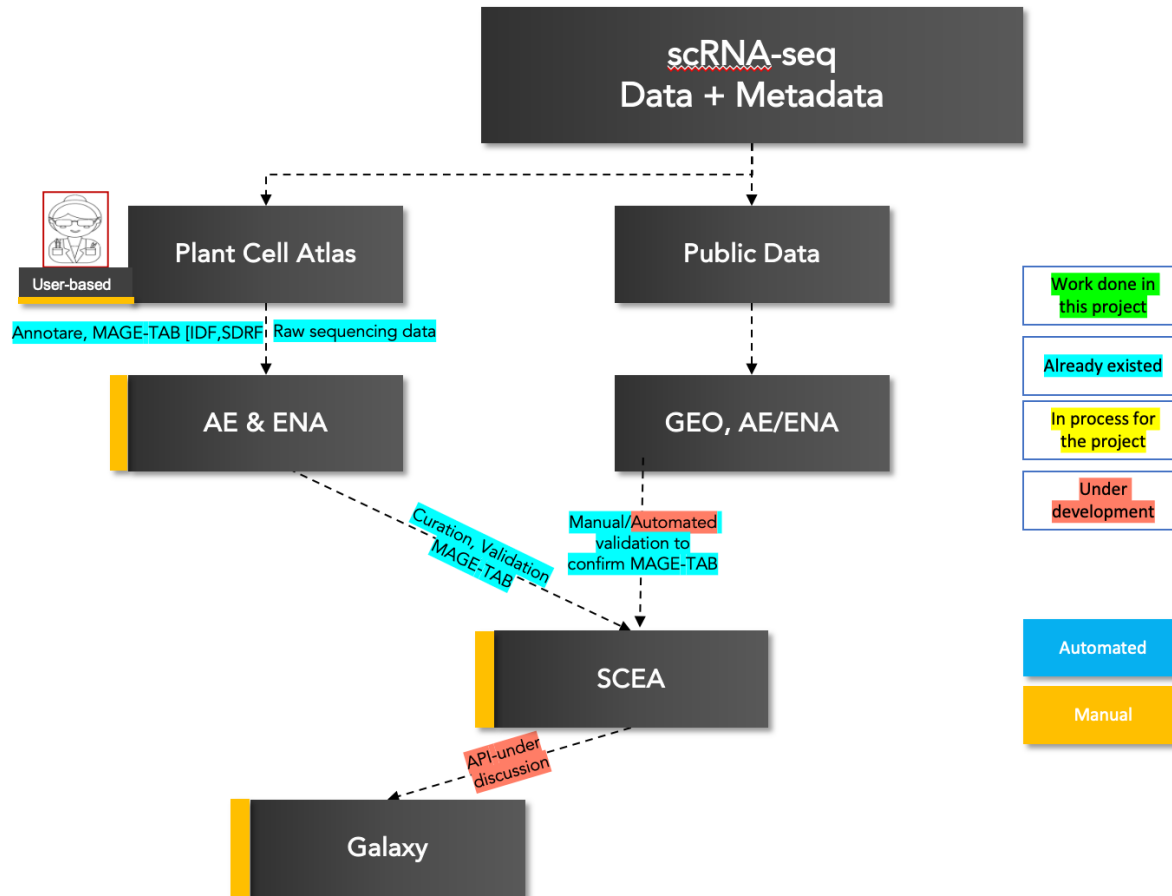
Filter by state

Expand All | Collapse All

	edit	delete	state	causes invalid graph	linked	ingest api url	valid	core type	sch
Valid									#926191-5a45-431 000006_316 No
Valid									eb1220ab-9111-46af 000006_316 No
Valid									c307447c-88bf-402 000006_316 No
Valid									2448a610-028f-485c 000006_316 No
Valid									60218a2-3705-486 000006_316 No
Valid									r19aa76c7b76b-dc1 000006_316 No

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



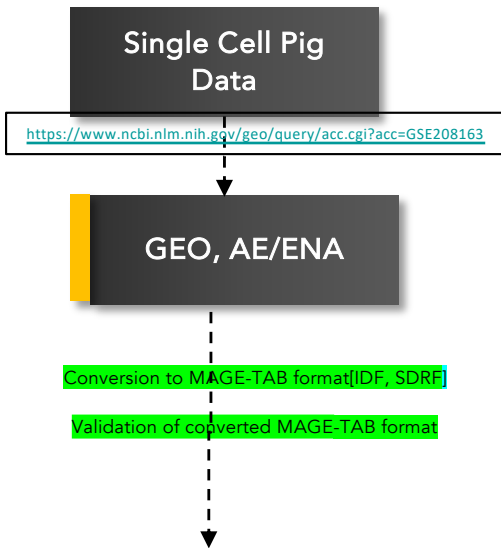
GEO

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
2023/03/31 18:45:16 INFO Skipping download of file NONE
```

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 characteristic for GSM6355572
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 characteristic for GSM6355573
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
2023/03/31 18:45:16 INFO Skipping download of file NONE
```

```
cat: expt_E-GEOD-208163.idf: No such file or directory
> cat expt_E-GEOD-208163.idf_report.log
Log generated 2023-04-08T18:54:21 from GSE208613.idf.txt

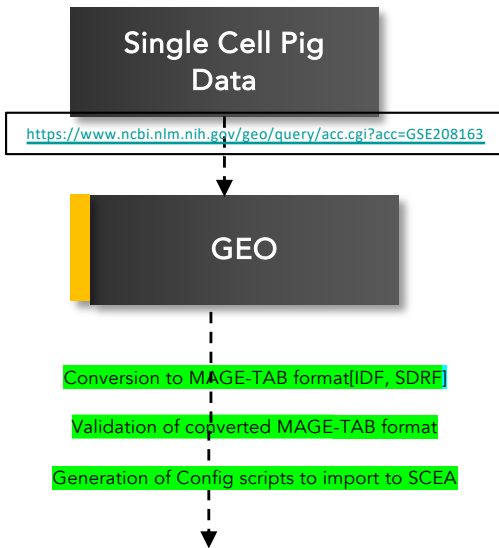
-----
--- Experiment Description -----
To establish better understanding of specific epithelial cells found across different regions of the small intestine in pigs, we utilized single-cell RNA sequencing (scRNA-seq) to recover and analyze epithelial cells from duodenum, jejunum, and ileum. Cells identified 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) Ileum 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 chromium 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 found 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 pylorus
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 to 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
```

Data Ingestion on Animal Side- SCEA



```
cat expl_E-GE00-208163.idf_report.log
Log generated 2023-04-08T18:54:21 from GSE208163.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) to 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 (IPP1), (2) Ileum collected 2-12 inches proximal to the ileocecal valve and divided into areas without Peyer's patches (NOPP), (3) Duodenum collected immediately following pylorus (DU00) and (4) Jejunum collected ~3 feet distal to pylorus (J03). Cells were extracted using 10X chromium 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 found using the identifier PRJNA802502.

----- 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 pylorus
FactorValue [organism part]: Duodenum, Ileum, Jejunum

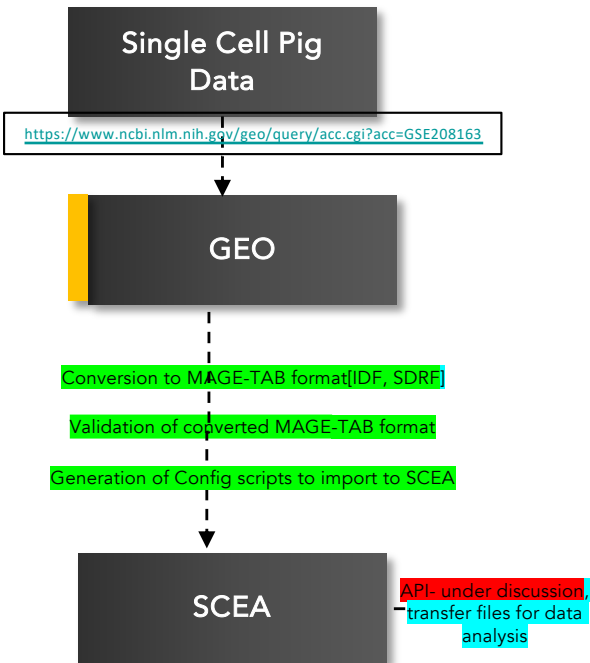
----- Source annotation -----
Characteristics [organism status]: Duodenum collected immediately following pylorus, Ileum with Peyer's patches, Ileum without Peyer's patches, Jejunum collected ~3 feet distal to pylorus
Characteristics [organism part]: Duodenum, Ileum, Jejunum
Characteristics [age]: 7.5 weeks
Characteristics [sex]: Female
Characteristics [breed]: Mixed-breed
Characteristics [organism]: Sus scrofa
Characteristics [accession]: GSE208163
```

```
cat atlas_configuration_generation_E-GE00-208163.log
tlas config generation log for E-GE00-208163 created at 2023-04-19T23:22:23

-----
NFO - Reading config for reference factor values and factor types to ignore
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

```
microarray_data_processing: No such file or directory
cat: cat: E:GEO-208163.idf_report.log
Log generated 2023-04-08T18:54:21 from GSE208613.idf.txt
```



Single Cell Expression Atlas

Single cell gene expression across species

Query bulk expression

← Back to Expression Atlas

Home Gene search Browse experiments Download Release notes Help Support

Search across 20 species, 304 studies, 8,524,149 cells

Ensembl 104, Ensembl Genomes 51, WormBase ParaSite 15, EFO 3.10.0

Search

Species: Any

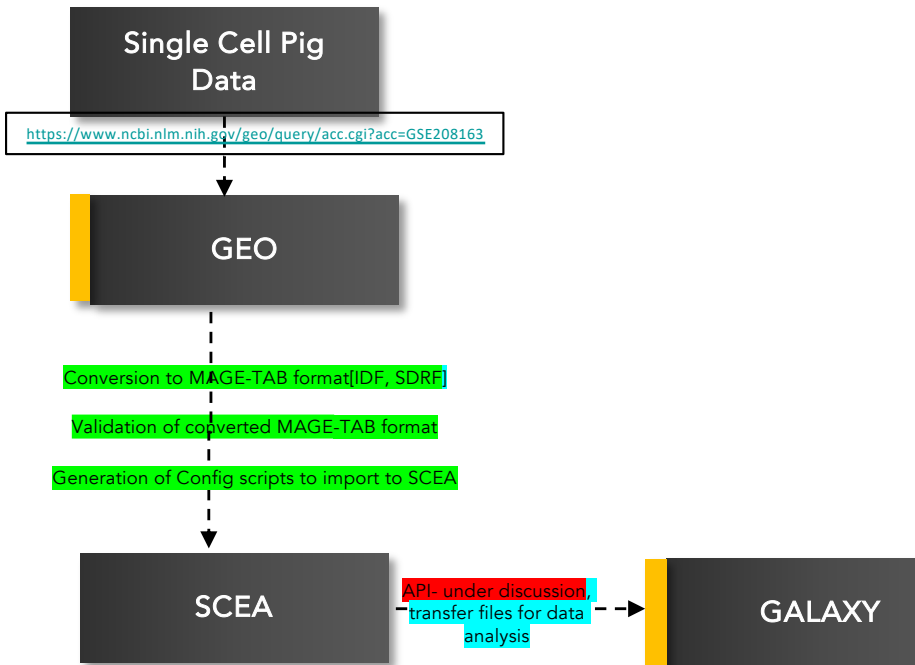
Examples: CFTR (gene symbol), ENSG00000115904 (Ensembl ID), 657 (Entrez ID), MGI:98354 (MGI ID), FBgr0004647 (FlyBase ID), keratinocyte (cell type), liver (organ/organism part), lung cancer (disease/condition)

Search

Animals Plants Fungi Protists

Species	Experiments
Homo sapiens	131 experiments
Mus musculus	111 experiments
Drosophila melanogaster	18 experiments
Danio rerio	7 experiments
Gallus gallus	4 experiments
Schistosoma mansoni	2 experiments

Data Ingestion on Animal Side- SCEA



The screenshot shows the Galaxy Human Cell Atlas interface. The 'Tools' panel on the left lists various tools, with 'Get scRNAseq data' highlighted in a red box. The main panel displays the configuration for the 'EBI SCXA Data Retrieval' tool. The 'SC-Atlas experiment accession' field is highlighted in a purple box and contains the value 'E-GEO-100058'. A blue arrow points from this field to a blue box labeled 'Experiment accession in SCEA'. The 'Choose the type of matrix to download' section has 'Raw filtered counts' selected. The 'Execute' button is visible at the bottom of the tool configuration.

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 -> *SCEA*
- **Shiny-PIGGI:** we will add the Cluster annotation functionality to the tool and ask for user feedback.

Acknowledgements

Muskan Kapoor¹, Alexey Sokolov², Enrique Sapena Ventura², Galabina Yordanova², Nicholas J. Provard³, Irene Papatheodorou², Nancy George², Doreen Ware^{4,5}, Sunita Kumari⁴, Timothy Tickle⁶, James Koltes,¹ Benjamin Cole⁷, Marc Libault⁸, Christine Elsik⁹, Wesley Warren¹⁰, Tony Burdett², Peter Harrison², and Christopher Tuggle¹

¹Bioinformatics and Computational Biology Program, Department of Animal Science, Iowa State University, Ames, IA 50011, U.S.A.

²EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, 12 U.K.

³Department of Cell and Systems Biology/Centre for the Analysis of Genome Evolution and Function, University of Toronto, Toronto, ON M5S 3B2, Canada.

⁴Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 11724, USA.

⁵U.S. Department of Agriculture, Agricultural Research Service, NEA Robert W. Holley Center for Agriculture and Health, Cornell University, Ithaca, NY 14853, USA.

⁶Data Sciences Platform, The Broad Institute of MIT and Harvard, 415 Main Street, 21 Cambridge, MA 02142, U.S.A.

⁷DOE-Joint Genome Institute, Lawrence Berkeley National Laboratory, 1, Cyclotron Road, 16 Berkeley CA 94720, U.S.A.

⁸Department of Agronomy and Horticulture, Beadle Center N305, University of Nebraska-Lincoln, Lincoln NE 68588-0660, U.S.A.

⁹Division of Animal Science and Division of Plant Science and Technology, S134D Animal Science Research Center, University of Missouri-Columbia, Columbia, MO 65211

¹⁰Division of Animal Science, 440G/446 Life Sciences Center, University of Missouri-Columbia, Columbia, MO 65211



Thank You!
