Creating a FAIR data ecosystem for incorporating single-cell genomics data into agricultural G2P research

Muskan Kapoor, Enrique Sapena Ventura, Alexey Sokolov, Galabina Yordanova, Nicholas J. 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, and Christopher Tuggle

ABSTRACT

The agriculture genomics community has numerous data submission standards but limited knowledge describing and storing single-cell (e.g., scRNAseq) data. Other single-cell genomics infrastructure efforts, such as the Human Cell Atlas Data Coordination Platform (HCA DCP), have resources that could benefit our community. For example, the HCA DCP is integrated with Terra, a cloud-native workbench for computational biology developed by Broad, Verily, and Microsoft that houses tools for scGenomics analysis. In Aim-1 we describe a pilot-scale project to determine if our current metadata standards for livestock and crops can be used to ingest scRNAseq datasets in a manner consistent with HCA DCP standards and if established resources (e.g., Terra) can be used to analyze the ingested data. Currently, the most comprehensive data ingestion portal for high throughput sequencing datasets from plants, fungi, protists, and animals (including humans) at the EMBL-European Bioinformatics Institute, Annotare, ensures that sufficient metadata are collected to enable re-analysis and dissemination via the Single Cell Expression Atlas (SCEA). Another EMBL-EBI portal limited to animal datasets, the FAANG portal, provides bulk and scRNAseq data access which uses a semi-automated process to submit and validate files using the HCA DCP metadata and data validation service. Once incorporated, datasets are used to augment the DCP resource for the scientific community. These files are also incorporated using EMBL-EBI's HCA DCP ingestion service and then transferred to Terra for further analysis. In Aim-2 we test and develop prototype tools to visualize the output of scRNAseq analyses on genome browsers and comparing across tissues and cell populations.

GOAL: IMPROVING SC DATA ANALYSIS INFRASTRUCTURE

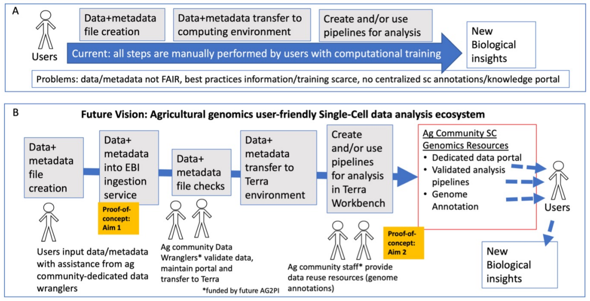

A
Users
Data+metadata file creation
Data+metadata transfer to computing environment
Create and/or use pipelines for analysis
Current: all steps are manually performed by users with computational training
New Biological insights
Problems: data/metadata not FAIR, best practices information/training scarce, no centralized sc annotations/knowledge portal
B
Future Vision: Agricultural genomics user-friendly Single-Cell data analysis ecosystem
Data+ metadata file creation
Data+ metadata into EBI ingestion service
Proof-of-concept: Aim 1
Data+ metadata file checks
Data+ metadata transfer to Terra environment
Create and/or use pipelines for analysis in Terra Workbench
Proof-of-concept: Aim 2
Ag Community SC Genomics Resources
• Dedicated data portal
• Validated analysis pipelines
• Genome Annotation
Users
New Biological insights
Users input data/metadata with assistance from ag community-dedicated data wranglers
Ag community Data Wranglers* validate data, maintain portal and transfer to Terra
*funded by future AG2PI
Ag community staff* provide data reuse resources (genome annotations)

Fig 1. (A) Current Status describing the pilot scale project when the data and meta-data file is created and transferred to a computing environment without the help of data wranglers or curators
(B) Future Vision: a more detailed explanation for transferring data to a computing environment that can further be utilized by the agricultural community.

RESULTS- AIM 1

GENERAL WORKFLOW

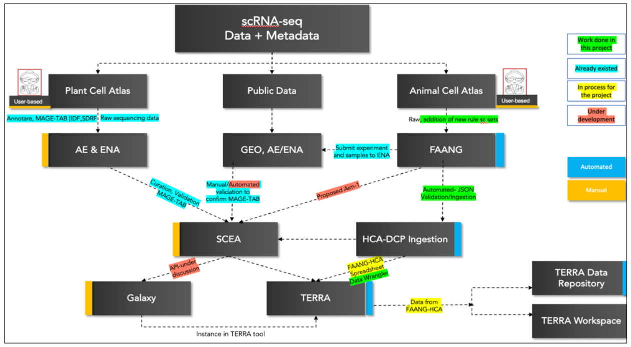

scRNA-seq Data + Metadata
Plant Cell Atlas
User-based
Public Data
Animal Cell Atlas
User-based
Annotare, MAGE-TAB [IDF,SDRF]
Raw sequencing data
Raw, addition of new rule sc sets
AE & ENA
GEO, AE/ENA
FAANG
Submit experiment and samples to ENA
Curation, Validation MAGE-TAB
Manual/Automated validation to confirm MAGE-TAB
Proposed Aim-1
Automated- JSON Validation/Ingestion
SCEA
HCA-DCP Ingestion
API-under discussion
FAANG-HCA Spreadsheet Data Wrangler
Galaxy
TERRA
Data from FAANG-HCA
TERRA Data Repository
TERRA Workspace
Instance in TERRA tool
Work done in this project
Already existed
In process for the project
Under development
Automated
Manual

Fig 2. The general route of meta-data flow in Animal, Plant, and Public Cell Atlas.

ANIMAL SIDE WORKFLOW

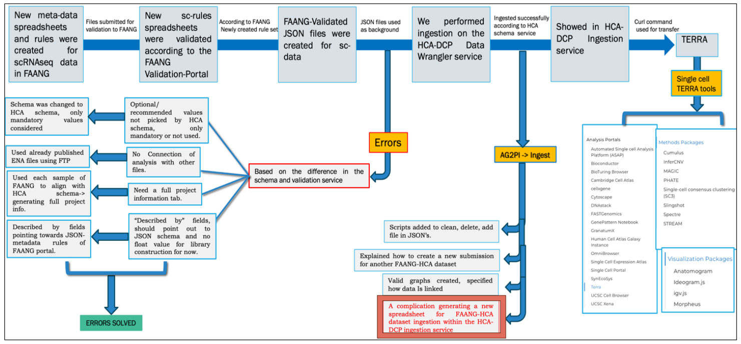
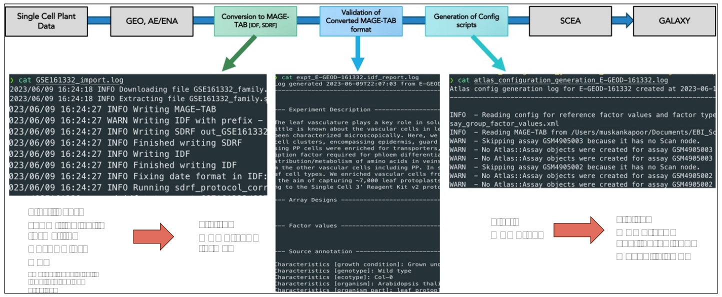

New meta-data spreadsheets and rules were created for scRNAseq data in FAANG
Files submitted for validation to FAANG
New sc-rules spreadsheets were validated according to the FAANG Validation-Portal
According to FAANG Newly created rule set
FAANG-Validated JSON files were created for sc-data
JSON files used as background
We performed ingestion on the HCA-DCP Data Wrangler service
Ingested successfully according to HCA schema service
Showed in HCA-DCP Ingestion service
Curl command used for transfer
TERRA
Single cell TERRA tools
Schema was changed to HCA schema, only mandatory values considered
Optional/ recommended values not picked by HCA schema, only mandatory or not used.
Used already published ENA files using FTP
No Connection of analysis with other files.
Used each sample of FAANG to align with HCA schema-> generating full project info.
Need a full project information tab.
Described by fields pointing towards JSON-metadata rules of FAANG portal.
"Described by" fields, should point out to JSON schema and no float value for library construction for now.
Based on the difference in the schema and validation service
Errors
ERRORS SOLVED
AG2PI -> Ingest
Scripts added to clean, delete, add file in JSON's.
Explained how to create a new submission for another FAANG-HCA dataset
Valid graphs created, specified how data Is linked
A complication generating a new spreadsheet for FAANG-HCA dataset ingestion within the HCA-DCP ingestion service
Analysis Portals
Automated Single cell Analysis Platform (ASAP)
Bioconductor
BioTuring Browser
Cambridge Cell Atlas
cellxgene
Cytoscape
DNAstack
FASTGenomics
GenePattern Notebook
GranatumX
Human Cell Atlas Galaxy Instance
OmniBrowser
Single Cell Expression Atlas
Single Cell Portal
SynEcoSys
Terra
UCSC Cell Browser
UCSC Xena
Methods Packages
Cumulus
InferCNV
MAGIC
PHATE
Single-cell consensus clustering (SC3)
Slingshot
Spectre
STREAM
Visualization Packages
Anatomogram
Ideogram.js
igv.js
Morpheus

Fig 4A. Ingestion of Animal Side workflow, the meta-data from the FAANG data portal was validated and transferred to HCA- DCP ingestion service. The red box indicates that we are generating a new spreadsheet because of the schema difference in both portals.

PLANT SIDE WORKFLOW

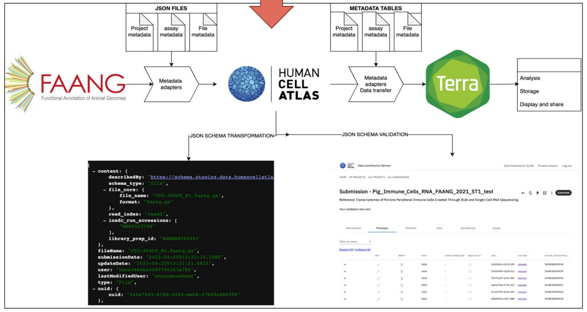

Single Cell Plant Data
GEO, AE/ENA
Conversion to MAGE-TAB [IDF, SDRF]
Validation of Converted MAGE-TAB format
Generation of Config scripts
SCEA
GALAXY

```
> cat GSE161332_import.log
2023/06/09 16:24:18 INFO Downloading file GSE161332_family.
2023/06/09 16:24:18 INFO Extracting file GSE161332_family.s
023/06/09 16:24:27 INFO Writing MAGE-TAB
023/06/09 16:24:27 WARN Writing IDF with prefix –
023/06/09 16:24:27 INFO Writing SDRF out_GSE161332
023/06/09 16:24:27 INFO Finished writing SDRF
023/06/09 16:24:27 INFO Writing IDF
023/06/09 16:24:27 INFO Finished writing IDF
023/06/09 16:24:27 INFO Fixing date format in IDF:
023/06/09 16:24:27 INFO Running sdrf_protocol_corr
```

```
> cat expt_E-GEOD-161332.idf_report.log
Log generated 2023-06-09T22:07:03 from E-GEOD
-------------------------------------------------

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

The leaf vasculature plays a key role in solu
ittle is known about the vascular cells in le
een characterized microscopically. Here, we
ell clusters, encompassing epidermis, guard
sing PP cells were enriched for transporters,
iption factor required for phloem differentia
tribution/metabolism of amino acids in veins
 the other vascular cells including PP. In s
f cell types. We enriched vascular cells fro
 the aim of capturing ~7,000 leaf protoplasts
g to the Single Cell 3' Reagent Kit v2 proto

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


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


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

haracteristics [growth condition]: Grown unc
haracteristics [genotype]: Wild type
haracteristics [ecotype]: Col-0
haracteristics [organism]: Arabidopsis thali
haracteristics [organism part]: leaf protopl
```

```
> cat atlas_configuration_generation_E-GEOD-161332.log
Atlas config generation log for E-GEOD-161332 created at 2023-06-1
----------------------------------------------------------------------

INFO  - Reading config for reference factor values and factor type
say_group_factor_values.xml
INFO  - Reading MAGE-TAB from /Users/muskankapoor/Documents/EBI_Sc
WARN  - Skipping assay GSM4905003 because it has no Scan node.
WARN  - No Atlas::Assay objects were created for assay GSM4905003
WARN  - No Atlas::Assay objects were created for assay GSM4905003
WARN  - Skipping assay GSM4905002 because it has no Scan node.
WARN  - No Atlas::Assay objects were created for assay GSM4905002
WARN  - No Atlas::Assay objects were created for assay GSM4905002
```

Fig 3. Ingestion of Plant Side, the Single-cell data from public archives follows a route through three scripts. The data can be visualized in the SCEA portal itself and analyzed in the GALAXY through web API retrieval.



```
{
 - content: {
      describedBy: "https://schema.staging.data.humancellatla
      schema_type: "file",
    - file_core: {
         file_name: "CT2-300CT_R1.fastq.gz",
         format: "fastq.gz"
      },
      read_index: "read2",
    - insdc_run_accessions: [
         "ERR5523708"
      ],
      library_prep_id: "SAMEA8050953"
   },
   fileName: "CT2-300CT_R1.fastq.gz",
   submissionDate: "2023-04-25T13:31:15.168Z",
   updateDate: "2023-04-25T13:31:21.691Z",
   user: "5ece3464ec0680746267e784",
   lastModifiedUser: "anonymousUser",
   type: "File",
 - uuid: {
      uuid: "f24e7043-818d-4309-aeb8-57b00a6b0350"
   },
```

Submission - Pig_Immune_Cells_RNA_FAANG_2021_ST1_test

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

Your validation returned

Fig 4B. Ingestion in HCA-DCP; First result is one of the entities in the API, showing the transformation in the format. The second result is an image of the UI of validated entities

RESULTS- AIM 2

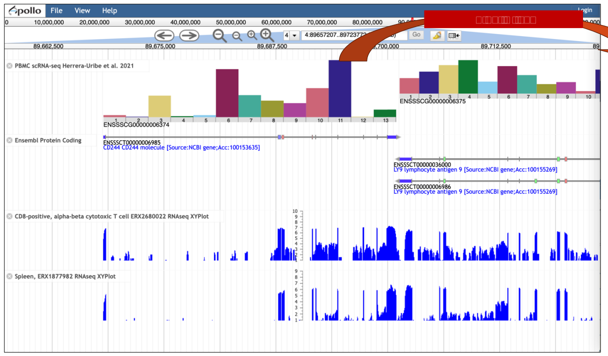
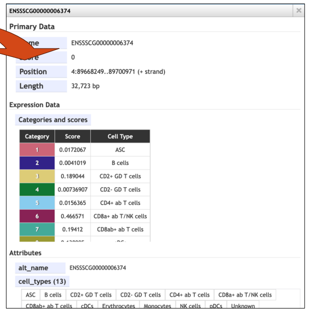

Apollo
File
View
Help
Login
10,000,000
20,000,000
30,000,000
40,000,000
50,000,000
60,000,000
70,000,000
80,000,000
4:89657207..89723772
Go
89,662,500
89,675,000
89,687,500
89,700,000
89,712,500
PBMC scRNA-seq Herrera-Uribe et al. 2021
ENSSSCG00000006374
ENSSSCG00000006375
Ensembl Protein Coding
ENSSSCT00000006985
CD244 CD244 molecule [Source:NCBI gene;Acc:100153635]
ENSSSCT00000036000
LY9 lymphocyte antigen 9 [Source:NCBI gene;Acc:100155269]
ENSSSCT00000006986
LY9 lymphocyte antigen 9 [Source:NCBI gene;Acc:100155269]
CD8-positive, alpha-beta cytotoxic T cell ERX2680022 RNAseq XYPlot
Spleen, ERX1877982 RNAseq XYPlot

Fig 5. JBrowse view showing tracks for PBMC scRNA-seq, Ensembl protein-coding genes, and two bulk RNA-seq experiments. The scRNA-seq track shows a bar chart for each gene, depicting gene expression levels for each cell type.


ENSSSCG00000006374

Primary Data

Name	ENSSSCG00000006374
Score	0
Position	4:89668249..89700971 (+ strand)
Length	32,723 bp

Expression Data

Categories and scores

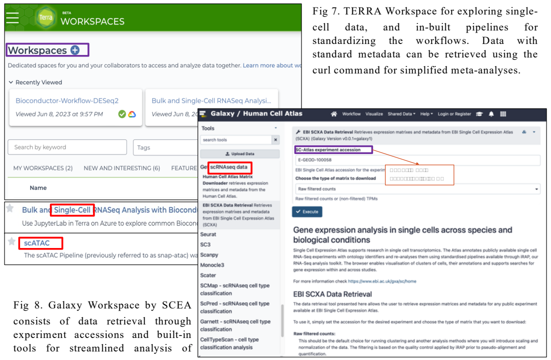
Category	Score	Cell Type
1	0.0172067	ASC
2	0.0041019	B cells
3	0.189044	CD2+ GD T cells
4	0.00736907	CD2- GD T cells
5	0.0156365	CD4+ ab T cells
6	0.466571	CD8a+ ab T/NK cells
7	0.19412	CD8ab+ ab T cells

Attributes

alt_name	ENSSSCG00000006374
cell_types (13)	ASC, B cells, CD2+ GD T cells, CD2- GD T cells, CD4+ ab T cells, CD8a+ ab T/NK cells, CD8ab+ ab T cells, cDCs, Erythrocytes, Monocytes, NK cells, pDCs, Unknown

Fig 6. A detailed panel, available by right-clicking a bar chart, provides a table called 'Categories and Scores' showing the gene expression level (called 'Score') for each cell type.

SINGLE CELL CLOUD-BASED ANALYSIS TOOLS- FUTURE VISION



Dedicated spaces for you and your collaborators to access and analyze data together. Learn more about wo

Bioconductor-Workflow-DESeq2
Viewed Jun 8, 2023 at 9:57 PM

Bulk and Single-Cell RNASeq Analysi...
Viewed Jun 8, 20

Bulk and Single-Cell RNASeq Analysis with Biocond
Use JupyterLab in Terra on Azure to explore common Biocon

scATAC
The scATAC Pipeline (previously referred to as snap-atac) wa

Get scRNAseq data

Human Cell Atlas Matrix Downloader retrieves expression matrices and metadata from the Human Cell Atlas.

EBI SCXA Data Retrieval Retrieves expression matrices and metadata from EBI Single Cell Expression Atlas (SCXA)

EBI SCXA Data Retrieval Retrieves expression matrixes and metadata from EBI Single Cell Expression Atlas (SCXA) (Galaxy Version v0.0.1+galaxy1)

SC-Atlas experiment accession: E-GEOD-100058

EBI Single Cell Atlas accession for the experim

Choose the type of matrix to download: Raw filtered counts

Raw filtered counts or (non-filtered) TPMs

Gene expression analysis in single cells across species and biological conditions

Single Cell Expression Atlas supports research in single cell transcriptomics. The Atlas annotates publicly available single cell RNA-Seq experiments with ontology identifiers and re-analyses them using standardised pipelines available through iRAP, our RNA-Seq analysis toolkit. The browser enables visualisation of clusters of cells, their annotations and supports searches for gene expression within and across studies.

For more information check https://www.ebi.ac.uk/gxa/sc/home

EBI SCXA Data Retrieval

The data retrieval tool presented here allows the user to retrieve expression matrices and metadata for any public experiment available at EBI Single Cell Expression Atlas.

To use it, simply set the accession for the desired experiment and choose the type of matrix that you want to download:

Raw filtered counts:

This should be the default choice for running clustering and another analysis methods where you will introduce scaling and normalization of the data. The filtering is based on the quality control applied by iRAP prior to pseudo-alignment and quantification.

Fig 7. TERRA Workspace for exploring single-cell data, and in-built pipelines for standardizing the workflows. Data with standard metadata can be retrieved using the curl command for simplified meta-analyses.

Fig 8. Galaxy Workspace by SCEA consists of data retrieval through experiment accessions and built-in tools for streamlined analysis of single-cell data.

CONCLUSIONS & FUTURE SCOPE

- We are building upon existing tools to develop a scientist-friendly data resource and analytical ecosystem and facilitate single cell-level genomic analysis across agricultural species. This is being accomplished through data ingestion, storage, retrieval, re-use, visualization, and comparative annotation.
- On the animal side, we intend to develop an automated pipeline to transfer the meta-data from the FAANG portal to the SCEA portal.
- On the Plant side, we intend to make the pipeline to SCEA more automated, in a similar fashion to the FAANG data portal functionality.

CONTACT INFORMATION

Muskan Kapoor

Graduate Research Assistant

Bioinformatics & Computational Biology, Animal Science

Email: muskan@iastate.edu ; cktuggle@iastate.edu

Iowa State University
Bioinformatics & Computational Biology Graduate Program

Acknowledgments: This research was supported by the U.S. Department of Agriculture, National Institute of Food and Agriculture grants 2022-70412-38454, 2021-70412-35233, and 2020-70412-32615