AG2PI SEED GRANT - PROJECT FINAL REPORT

PROJECT NAME	Creating a FAIR data ecosystem for incorporating single cell genomics data into agriculture G2P	

PROJECT PRINCIPAL INVESTIGATOR	TODAY'S DATE	PROJECT START DATE	DATE OF COMPLETION
Christopher K. Tuggle/ Muskan Kapoor	10/19/23	05/20/22	08/31/23
TEAM MEMBERS (co-PI, co-I, personnel)		COLLABORATORS	
Peter W. Harrison, Christine Elsik, Nicholas Provart		Tony Burdett, Tim Tickle, Marc Libault, Wes Warren, Ben Cole, James Koltes	

ACCOMPLISHMENTS

Please provide a short summary of the conclusions (both successes and failures) made from your project. Include a description of how this project will provide benefits to the agricultural genome to phenome community and, possibly, to a broader audience. You should include both qualitative and quantitative details, as necessary, to support your conclusions. Include a short accomplishment statement in non-technical language and do not include names.

This is not a technical report. Please keep to no more than 6-8 sentences (e.g., 1-2 sentences per point, above).

Summary:
Our project, aimed at advancing single-cell genomics in the agricultural genome to phenome community, has yielded both successes and identified areas for improvement. We have successfully achieved the ingestion and transfer of pig single-cell data from the FAANG portal to the Human Cell Atlas-Data Coordination Platform (HCA-DCP), as well as its subsequent transfer to the TERRA platform for in-depth analysis. In addition, plant single-cell data was already been integrated into the Single Cell Expression Atlas portal, and we have used their pipeline to ingest the new pig single-cell data into Galaxy for further analysis. We also developed a tool to enable scientists to view results of single-cell analyses in the context of genes on a genome browser.

However, we encountered a couple of challenges along the way. The pig single-cell data is made visible in the HCA browser but resides in the backend of the HCA-DCP data wrangler service. Furthermore, there is currently no direct path from the FAANG portal to the Single Cell Expression Atlas (SCEA) portal, which is an area that needs further attention.

Benefits to the Agricultural Genome to Phenome Community:
As the agricultural genome to phenome community lack metadata standards for describing single-cell data we developed a stramlined anlaytical system for data storage, retrieval, re-use, visualization and comparative annotation across agricultural species. By building this new system, we've made it simpler for scientists to share their data and perform detailed genetic analyses on individual cells. This not only benefits researchers but also has the potential to improve our understanding of agriculture's genetic aspects and have broader applications in genomics research.

Accomplishment Statement (Non-Technical):
We have created an advanced system that makes it easier for scientists in the agricultural genome community to facilitate Single Cell level genomic analysis. Our system allows researchers to submit information about their data, and it can be analyzed using powerful tools in the cloud, without the need to download large datasets or use complex software. This system includes all the tools needed for single-cell analysis, along with helpful examples workbooks of how to use them, as well as a genome visualization tool

to explore gene expression levels for cell types. Our project's main successes are a) the development of a pipeline for efficient transfer of pig single-cell data to platforms like TERRA and Galaxy, as well as b) initiating cross-kingdom discussions that identified the main deficient in single-cell data sharing in agriculture is the lack of tools for efficient single cell dataset annotation.

Products

Please list any products from this project. This may include (but not limited to) publication, concept/white paper, workshop, conference presentation, website, publicly available data or pipelines, etc. Reminder: you are required to make your prodcuts available to the broader stakeholder community using standard USDA practices, open source, FAIR, or other models. Metrics may include number of participants or times accessed, etc. Include links to recordings, DOI, etc. when possible. For presentations and posters, provide authors, date, location and presentation title.

ACTIVITY / PRODUCT	DESCRIPTION (include URL, if applicable)	OUTCOME / METRICS
Review article on the Plant Cell Atlas	Fahlgren, N., M. Kapoor, G. Yordanova, J. Waeseb, B. Cole, P. Harrison, D. Ware, A. Burdett, C. G. Elsik, C. K. Tuggle, N. J. Provart. 2022. Toward a Data Infrastructure for the Plant Cell Atlas. Plant Physiology Oct 6; kiac468 on-line ahead of print. doi:10.1093/plphys/kiac468	We participated in a review article on the Plant Cell Atlas, providing a section on data ingestion and use currently performed at the Human Cell Atlas and how data ingestion and annotation could be streamlined for plant and animal data.
Plant Cell Atlas Symposium (PCA)	M. Kapoor, A. Sokolov, E. S. Ventura, G. Yordanova, N. J. Provart, I. Papatheodorou, N. George, D. Ware, S. Kumari, T. Tickle, B. Cole, T. Burdett, P. Harrison, C. Tuggle. **Date**: 12-13 December 2022 **Location:** virtual conference **Title:** Creating a FAIR data ecosystem for incorporating single cell genomics data into agricultural G2P research	M. Kapoor presented in 2022 Plant cell atlas symposium focusing on Plant data and use of SCEA and Annotare for data submission, storage, visaulization and anlaysis.
American Society of Animal Science (ASAS)	M. Kapoor, A. Sokolov, E. S. Ventura, G. Yordanova, N. J. Provart, I. Papatheodorou, N. George, D. Ware, S. Kumari, T. Tickle, B. Cole, T. Burdett, P. Harrison, C. Tuggle. **Date**: 12-15 March 2023 **Location:** Madison	M. Kapoor presented a poster in the conference focusing on Ingestion of Animal datasets with use of HCA-DCP ingestion service. Developed a Shiny web application which will be an important resource for improved annotation of porcine immune genes and cell types.

	Title: Computational tools and resources for analysis and exploration of single-cell RNAseq data in agriculture	
Interdisciplinary Biological Science Symposium	M. Kapoor, A. Sokolov, E. S. Ventura, G. Yordanova, N. J. Provart, I. Papatheodorou, N. George, D. Ware, S. Kumari, T. Tickle, J. Koltes, B. Cole, M. Libault, C. Elsik, W. Warren, T. Burdett, P. Harrison, C. Tuggle. **Date**: 26-27 April 2023 **Location:** Iowa State University **Title:** Single-cell genomics data incorporation into agricultural G2P research by building a FAIR data ecosystem	M. Kapoor gave a talk on visualization tool for exploring single cell data, shinyPIGGI and how the data goes in different plant for both plant and animal cell atlas ingestion tools- SCEA and HCA respectively. Focused on how the data was ingested in both plant and animal data ingestion pipelines.
AG2PI Conference	M. Kapoor, A. Sokolov, E. S. Ventura, G. Yordanova, N. J. Provart, I. Papatheodorou, N. George, D. Ware, S. Kumari, T. Tickle, J. Koltes, B. Cole, M. Libault, C. Elsik, W. Warren, T. Burdett, P. Harrison, C. Tuggle. **Date**: 15-16 June 2023 **Location:** Kansas City, Missouri **Title:** Creating a FAIR data ecosystem for incorporating single cell genomics data into agricultural G2P research	M. Kapoor presented a poster focusing on animal and plant cell atlas path, how the ingestion was successfully performed, and a tool "JBrowse" to create genome browser tracks that allows visualization of cell type expression of each gene.

Audience

With whom has this work been targeted to and shared? Please describe how this project and its products have been disseminated to a community of interest. Include any outreach activity or information sharing as well as training or professional development opportunities provided in this project.

Researcher interested in the interface between genomics and phenomics, Scientists interested in sharing single-cell RNAseq data with the community. The results from this project have been disseminated to diverse audiences, including readers of plant science journals and attendees at plant and animal science meetings.

CHALLENGES

Changes to team

Have there been any changes to the original team membership (including collaborators) from who was included in the proposal? Please review your proposal then provide an explanation if changes were made.

As the data is also targeted with the plant cell atlas people, during the project discussions with this group identified that there was already a working ingestion pipeline at EBI and CSHL -Single cell Expression Atlas, Gramene collaboration. We ingested both the plant and animal data using their pipeline and showed another route of data ingestion pipeline. The new members include: Irene Papatheodorou, Nancy George, Doreen Ware, Sunita Kumari

Other changes

Were there changes to your project, not including changes to team membership? This may include expansion or reduction in scope. If changes occurred, did these have a significant impact on expenditures? Please explain.

As decribed above we idenitfied that a plant group already had a pipeline for plant data ingestion, although it is low-throughput. Thus we focused on developing a FAANG -orietned pipeline into Human Cell Atlas.

Instaed of using HCA- metadata rules and schema , we reduced the metadata rules to match the FAANG portal schema and then carried out ingestion.

Challenges

Have you experienced any challenges or delays? Please provide the actions you took to resolve them, if possible.

Aim 1:
Challenge based on the difference in the schema of FAANG and HCA:
Action: Used a different schema and schema validator for FAANG JSON files.

Challenge based on different scRNAseq Rules in both portals:
Action: Updated FAANG scRNAseq rules following the HCA portal.

Challenge on the Ingestion side:
Action: Instead of the HCA metadata schema, we pointed out the FAANG metadata schema, which is available on a certain URL, but the data will be ingested in the HCA-DCP ingestion service.

Aim 2:
Challenge in adapting data format for visualization in genome browser:
Action: We developed new code to reformat the data to create a genome browser track.

CONTINUATION OF WORK

Next steps

How do you/your team plan to continue moving this project forward? Include how AG2PI can assist in your forward momentum.

We would like to complete a path to place FAANG data directly into the sc Expression Atlas. We have discussed how to accomplish this with EBI representatives, and we plan to apply for USDA NIFA funding to fund a collabroative effort to work on this.
We are also continuing to discuss such data tools with the plant community. We are preparing a presentation to the AGBioData commuity, in collbaroation with plant biologist Ben Cole of LBNL.

Outreach

In what ways are you able to stay engaged with AG2PI? (check boxes as appropriate)

- ☐ Will present at a field day
- ☐ Will lead a training workshop
- ☒ Would like to participate in any future AG2PI conference
- ☒ Work with AG2PI on a news release on project conclusions
- ☒ Will continue attending AG2PI events
- ☐ Other (please explain)