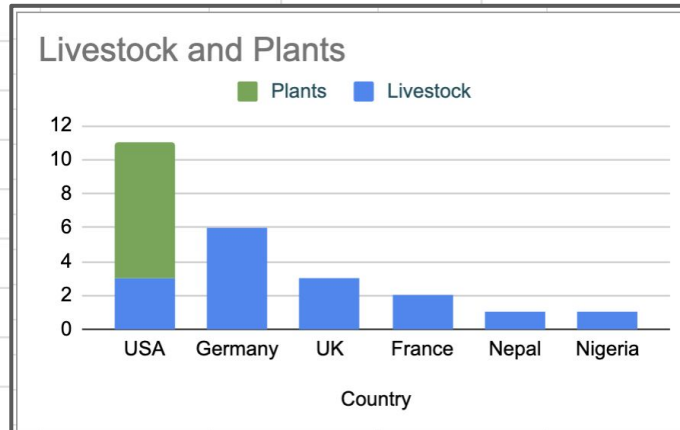# Standardizing scRNAseq metadata of plants and animals to promote FAIRification
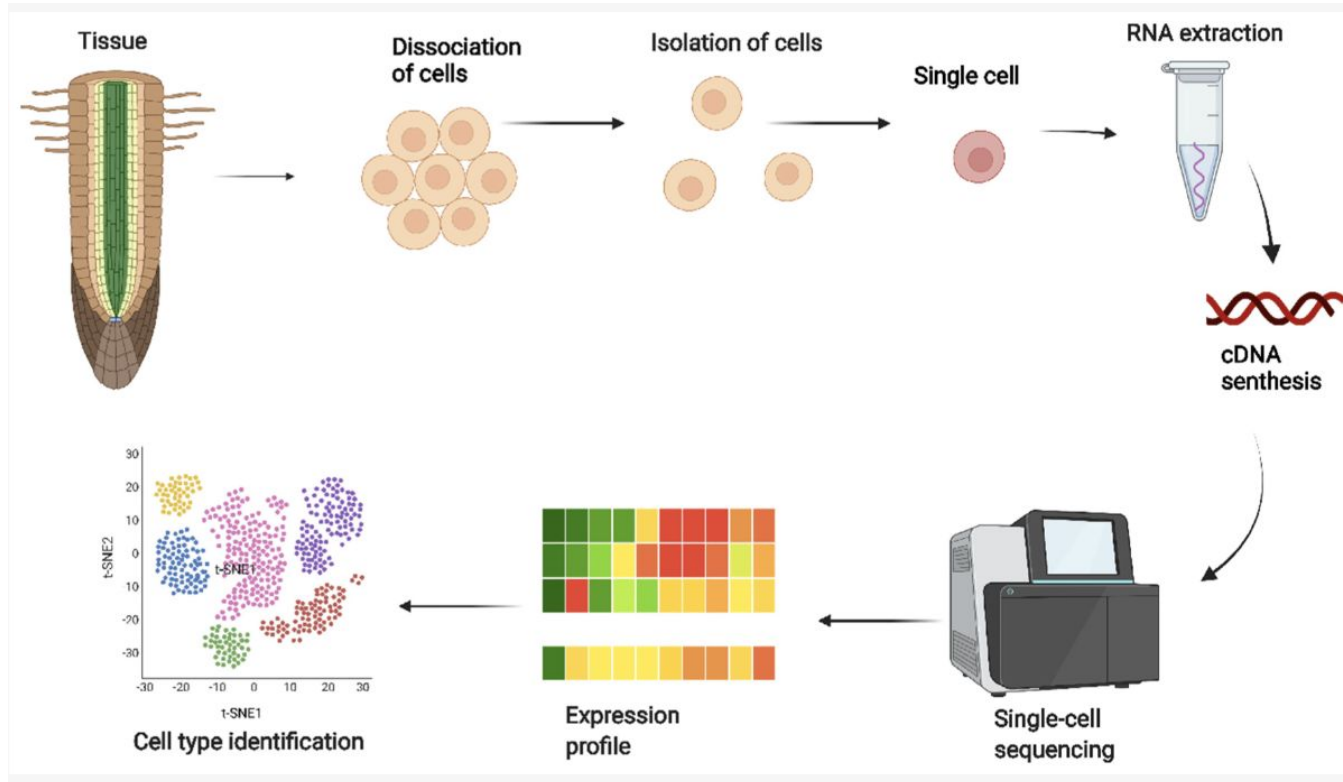
Sunita Kumari, CSHL
AgBioData Community Workshop
April 30, 2024

scRNAseq Biocuration WG Members

Members	Affiliations	Members	Affiliations
Chris Tuggle (Chair)	IowaStateUniv	**Sunita Kumari (Co-Chair)**	CSHL
Muskan Kapoor	IowaStateUniv	**Benjamin Cole (Co-Chair)**	LBL
Mehak Kapoor	IowaStateUniv	Doreen Ware	USDA-ARS
Anca Dorhoi	FLI	Trevor Nolan	Duke
Doreen Becker	FBN	Dae Kwan Ko	MSU
Björn Corleis	FLI	Sushma Naithani	OSU
Virginia Friedrichs	FLI	Rex Nelson	USDA-ARS
Alexander Schäfer	FLI	Farah Saeed	UGA
Praveen Krishna Chitneedi	FBN		
Androniki Psifidi	RVC		
Katie Dubarry	Roslin		
Emily Clark	Roslin		
Elisabetta Giuffra	INRAE		
Giorgia Egidy	INRAE		
Biraj Poudel	Tribhuvan		
Daniel Adediran	Helix Biogen		

PLANT CELL ATLAS

AG2Pi — Agricultural Genome to Phenome Initiative

FAANG — Functional Annotation of Animal Genomes

Livestock and Plants

Plant scRNAseq Workflow

Why scRNAseq is important?

- Cellular heterogeneity: distinct cell populations and states that might be masked in bulk RNA-seq data.

- Cell type identification and characterization: identify and characterize novel or rare cell types, further refine known cell types.

- Developmental biology: study cellular differentiation, developmental trajectories in various organisms, providing insights into the mechanisms governing cellular fate decisions.

- Gene regulatory network analysis: By studying gene expression patterns in individual cells, identify key regulators of cellular processes and cell fate decisions.

- Characterizing tumor microenvironment: valuable for studying the tumor microenvironment, which comprises diverse cell types, including cancer cells, immune cells, and stromal cells. Help in understanding the interactions between these cells, how their states change over time, their roles in cancer progression and response to therapy.

- *Build a better and high‑resolution catalogue of cells in all living organism, commonly known as atlas, which is key resource to better understand growth and development and provide a solution in treating diseases.*

Need on Community wide resources

- Excellent research can be done at a single lab level, with significant effort, but _democratization would benefit agriculture_.

- *SC data and meta-data storage and sharing with community*

- *Utilization of organism-wide data for optimal cell annotation and gene regulatory networks*

- *Improved annotation of reference assemblies (Pan-genome) with SC data*

- *Analytical data pipelines – development and validation*

- *Comparative analysis of SC data across species*

No portal exists to provide access to all public scRNAseq for agricultural species. Very little metadata is available for most such datasets, thus data reuse is very difficult.

Agricultural research output –quantity, quality AND reproducibility could be significantly improved if such _community-wide resources are_ developed!

scRNAseq challenges
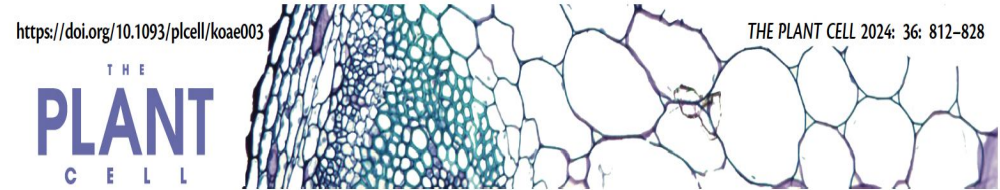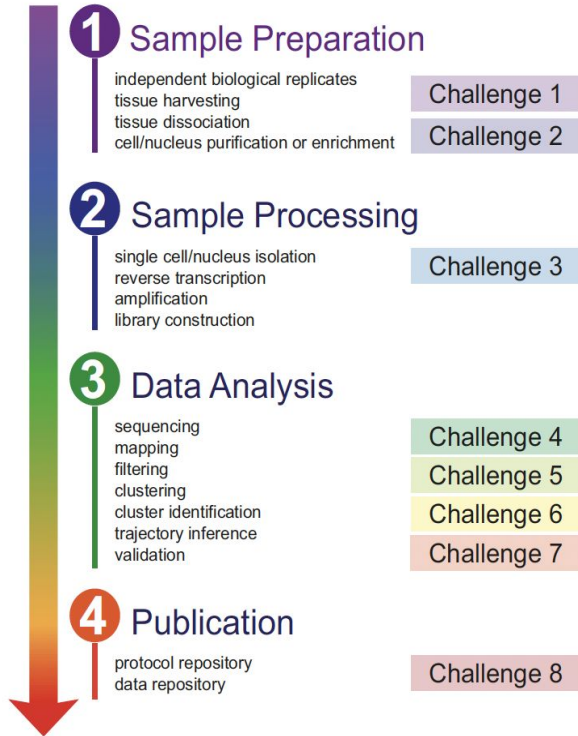
Best practices for the execution, analysis, and data storage of plant single-cell/nucleus transcriptomics

Carolin Grones,[1,2] Thomas Eekhout [1,2,3] Dongbo Shi,[4,5] Manuel Neumann [6] Lea S. Berg [7] Yuji Ke,[1,2] Rachel Shahan [8,9] Kevin L. Cox Jr [10] Fabio Gomez-Cano,[11] Hilde Nelissen [1,2] Jan U. Lohmann [4] Stefania Giacomello,[12] Olivier C. Martin [13] Benjamin Cole [14] Jia-Wei Wang [15] Kerstin Kaufmann [6] Michael T. Raissig [7] Gergo Palfalvi,[16] Thomas Greb [4] Marc Libault [17,*] and Bert De Rybel [1,2,*]

Sample metadata challenges

- Inconsistency in experiment protocol and sample handling - batch effects
- Annotation discrepancies - hinders accurate comparison of scRNAseq data
- Lack of comprehensive agricultural metadata - essential for reproducibility
- Inconsistent sample preparation and sequencing techniques contributes to technical variability which can confound biological signals thus leading to challenges and reproducibility and interpretation of results across different studies

Curate and document metadata adhering to FAIR principles to enhance reusability and reproducibility across studies.

Main Goals Of WG

- Address the challenges of annotating scRNAseq metadata in plants and animals, and establish a set of recommendations for member databases on management of scRNAseq datasets.

- A compilation of minimum standards for metadata describing species, strain, cultivar, developmental stage, cell types, etc. using appropriate ontologies, and details of experimental design. This minimum information is required for storing, analysis and data re-use.

What is our WG working on?

- Survey communities, identify existing tools and approaches for single cell data and meta-data collection and storage.

- Determine additional community needs and potential resources for expanding metadata-enriched single cell data repositories, including genome portals that display cell-type level annotations.

- Review best practice recommendations for integration of member databases data with larger efforts such as public INSDC archives, Plant Cell Atlas and Single Cell Expression Atlas.

What deliverables are we going to produce?

- Define the data and metadata for scRNAseq studies using standardized terms ontologies; Several current challenges addressed by group.

- Database of community and datatypes, selection of dataset for major species that can be used for benchmarking of methods.

- Establish integration routes and processes for connection to larger sc data integration efforts.

- White paper - documenting recommendations for FAIR metadata, datatypes, and recommended methods/pipelines for management of sc data.

How important is our work for the community?

It will facilitate in creating a cohesive community that will provide standards, resources, tools and standardized scRNAseq datasets for FAIR data meta-analysis, leading to initial cell type transcriptome descriptions.

A number of researchers in plant and animal single-cell research have been discussing means to develop such national/global resources

What does the sc ag community need to succeed?

Global repositories and tools to share data, metadata and results

Community annotation of cell types

Apply sc data to biology and breeding phenotypes →G2P and Genetic improvement

Identify additional needed community resources and organization Determine support needed to obtain such resources ($)

One possible funding opportunity is the Agricultural Genome to Phenome Initiative (AG2PI), which funds both conferences and standard research projects to develop community solutions to solving the Genome to Phenome problems. See: https://www.nifa.usda.gov/grants/funding-opportunities/agricultural-genome-phenome-initiative

If you are interested in this initiative, please contact any of the following people:

Chris Tuggle, Iowa State
Ben Cole, DOE- LBL
Wes Warren, U. Missouri
Sunita Kumari/Doreen Ware, CSHL
Peter Harrison, EMBL-EBI
Fiona McCarthy, U. Arizona
Sue Rhee, Michigan State

Questions for Breakout Session

1. How to create a community of researchers interested in scRNAseq
2. How to motivate people for keeping standards for their gene expression submission?
3. Need curators and annotators
4. What we are missing? What can be added to these efforts?
5. Funding opportunities?