Orthologous marker groups reveal broad cell identity conservation across plant single-cell transcriptomes

Song Li

Associate Professor
School of Plant and Environmental Sciences
Virginia Tech

The OMG website: https://orthomarkergenes.org/

Lab website: https://github.com/LiLabAtVT/
Email: songli@vt.edu

Acknowledgements

Tran (Nina) Chau
Graduate Student

Sai Pavan Bathala
MS in computer science
(currently at Bloomburg)

EAGER-Tools for Cells
PGRP

Prakash (PR) Timilsina, Ph.D.
Postdoc Associate
Currently at Boston Children's Hospital

Sanchari Kundu
Graduate Student

Bas Bargmann, Ph.D.
Virginia Tech

Link to project website: https://github.com/LiLabAtVT/OrthoMarkerGeneGroups
The OMG website: https://orthomarkergenes.org/

Email: songli@vt.edu

Presentation Outline

- Background and rationale
- The OMG (orthologous marker gene groups) framework
- Other single cell tools for plant biology
  - Co-expression
  - Gene function prediction
  - 3UTR mapping
- Summer workshop

GOAL: Determine cell types in scRNA-seq data for non-ATH species

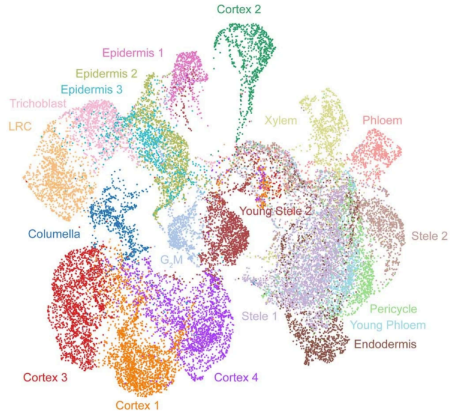
0. Data integration (for closely related species)

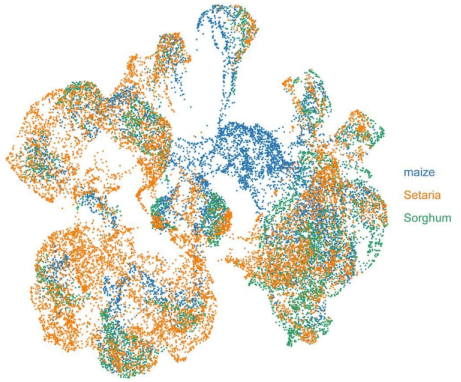
1. Marker gene-based approaches
- Identify clusters
- Find cluster markers
- Find ATH orthologous markers

2. Correlation. scRNA-seq vs bulk RNA-seq data

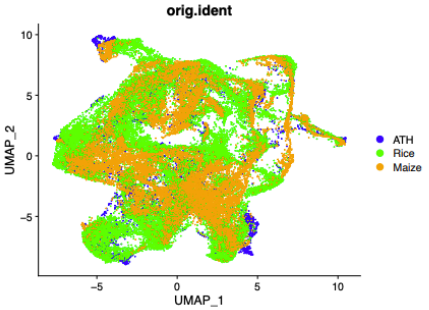
3. Index of Cell Identity Method (ICIM)
- Expression level
- Enrichment in clusters
- Selected markers

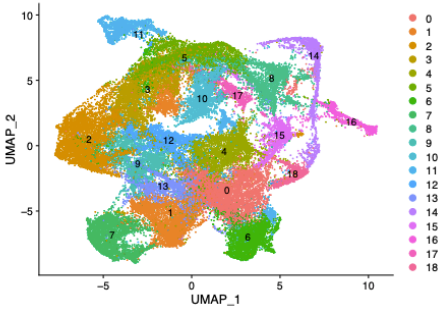
4. Using GO functions of marker genes





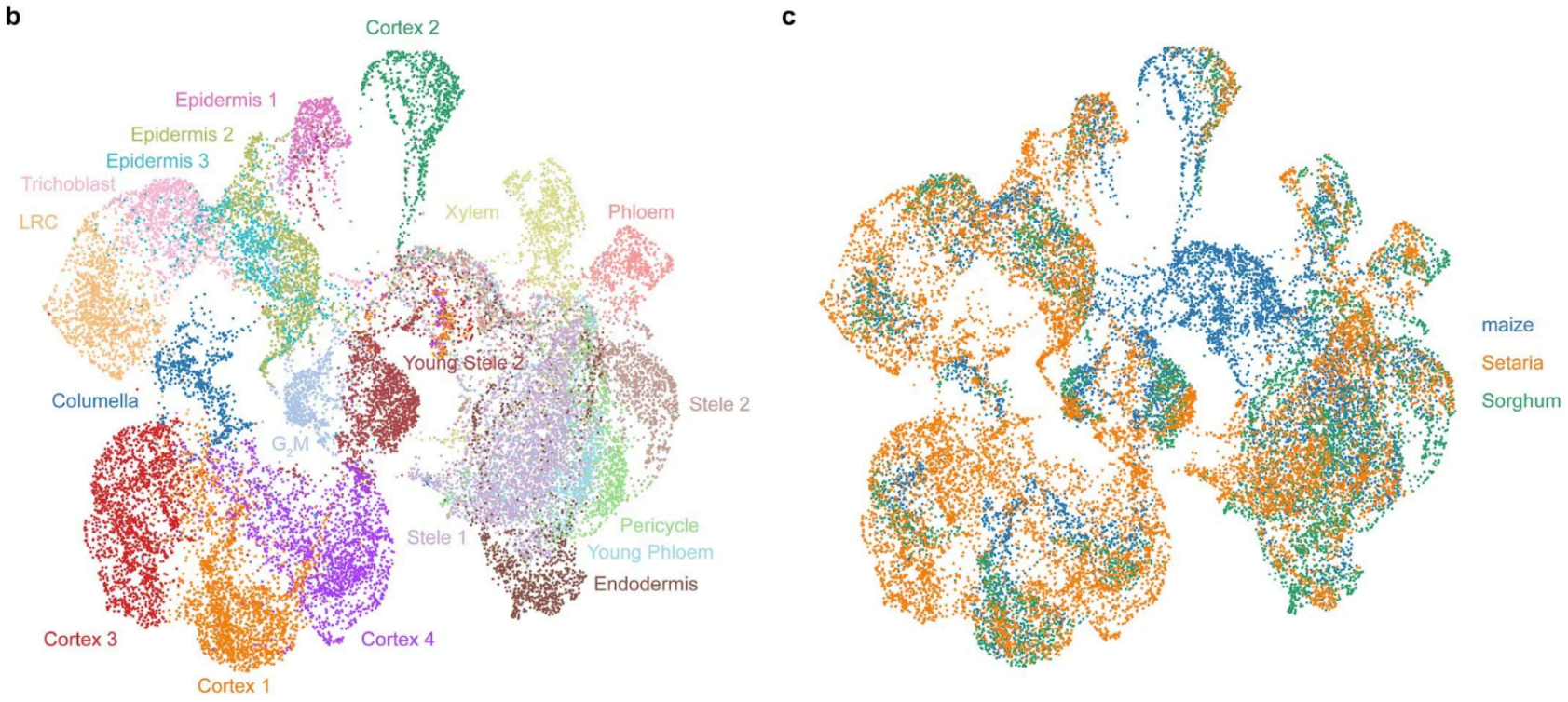
Guillotin et al., 2023, Nature





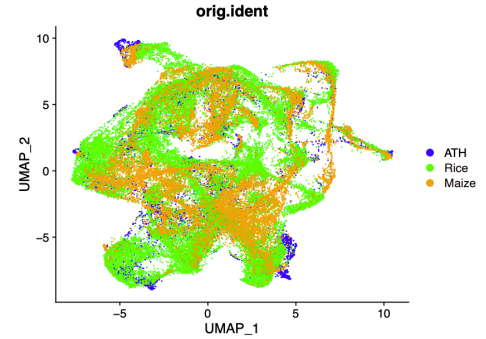
Chau et al., 2025, Nature Communication

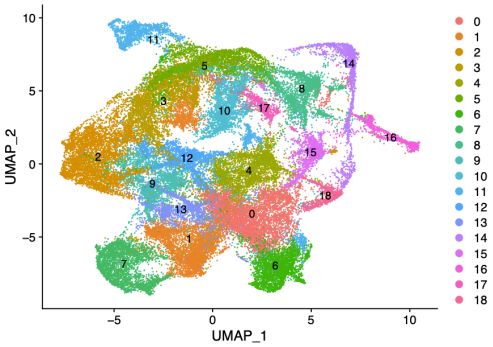
GOAL: Determine cell types in scRNA-seq data for non-ATH species

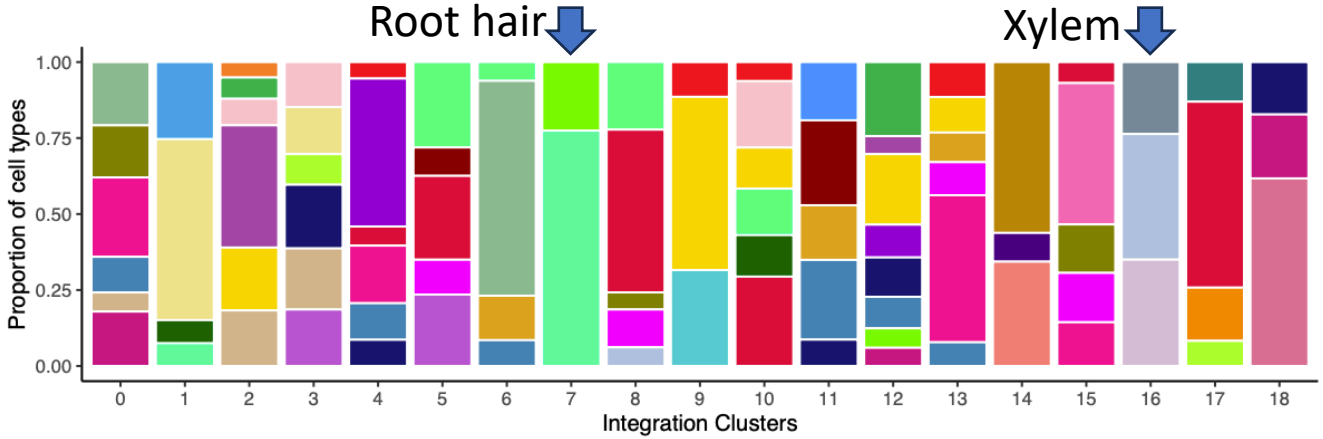


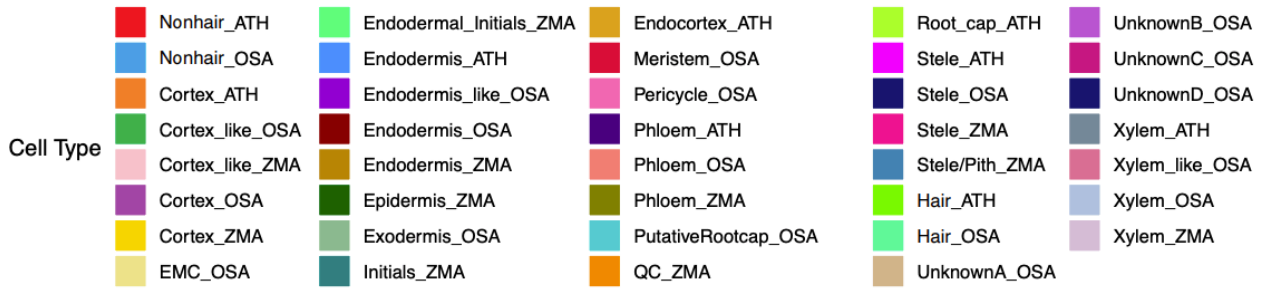
Guillotin et al., 2023, Nature

An example of unsuccessful integration across diverse species





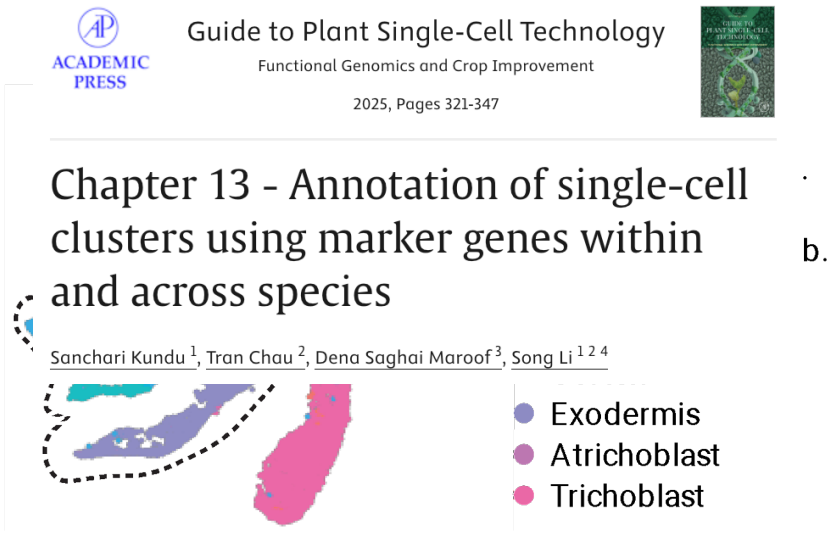




Chau et al., 2025, Nature Communication

GOAL: Determine cell types in scRNA-seq data for non-ATH species

1. **Marker gene-based** approaches
   - Identify clusters
   - Find cluster markers
   - Find ATH orthologous markers
2. **Correlation.** scRNA-seq vs bulk RNA-seq data
3. **Index of Cell Identity Method** (ICIM)
   - Expression level
   - Enrichment in clusters
   - Selected markers
4. **Using GO functions** of marker genes



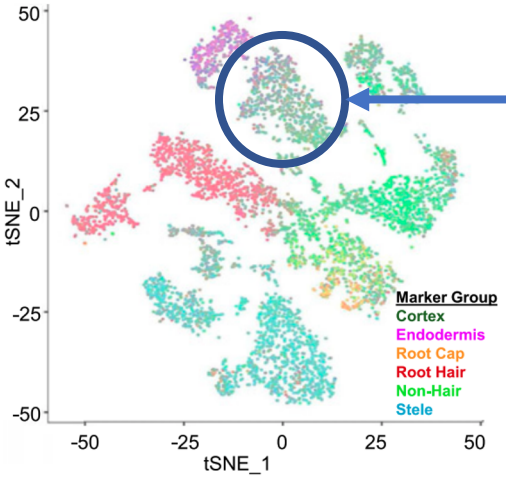
Canto et al., Nature Plant, 2024

Bulk RNA-seq from
Kajala et al., Cell, 2021,
doi: 10.1016/j.cell.2021.04.024

Prakash Timilsena

Use machine learning to identify marker genes for cross species comparison

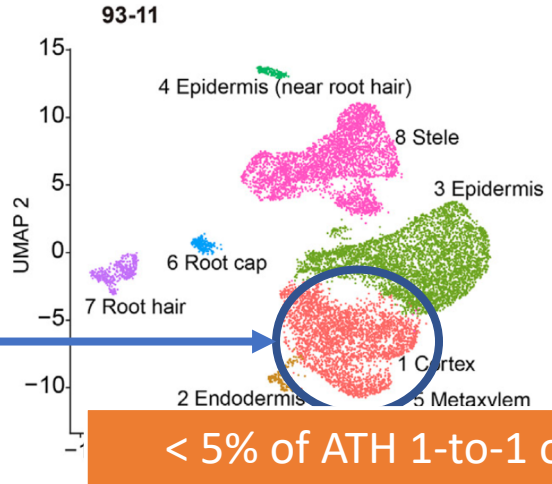
Arabidopsis scRNAseq



Ryu et al.., 2019, Plant Phys.

Cortex

Rice scRNAseq



Haidong Yan

< 5% of ATH 1-to-1 orthologous markers are conserved in rice

Table 1 Comparison of number of overlapped rice makrers in three cell types among six marker types.

markertype	celltype	marker_num	rice_marker_num	overlap_marker_num	overlap_ratio	binomial_testing_pval
SHAP	Cortex	2118	674	93	0.137982196	5.74E-11
SVMM	Cortex	1286	674	54	0.080118694	0.00000491
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
ICIM	Cortex	42	674	0	0	1
KNOW	Cortex	304	674	10	0.014836795	0.128534626
BULR	Cortex	62	674	0	0	1

Machine Learning Markers

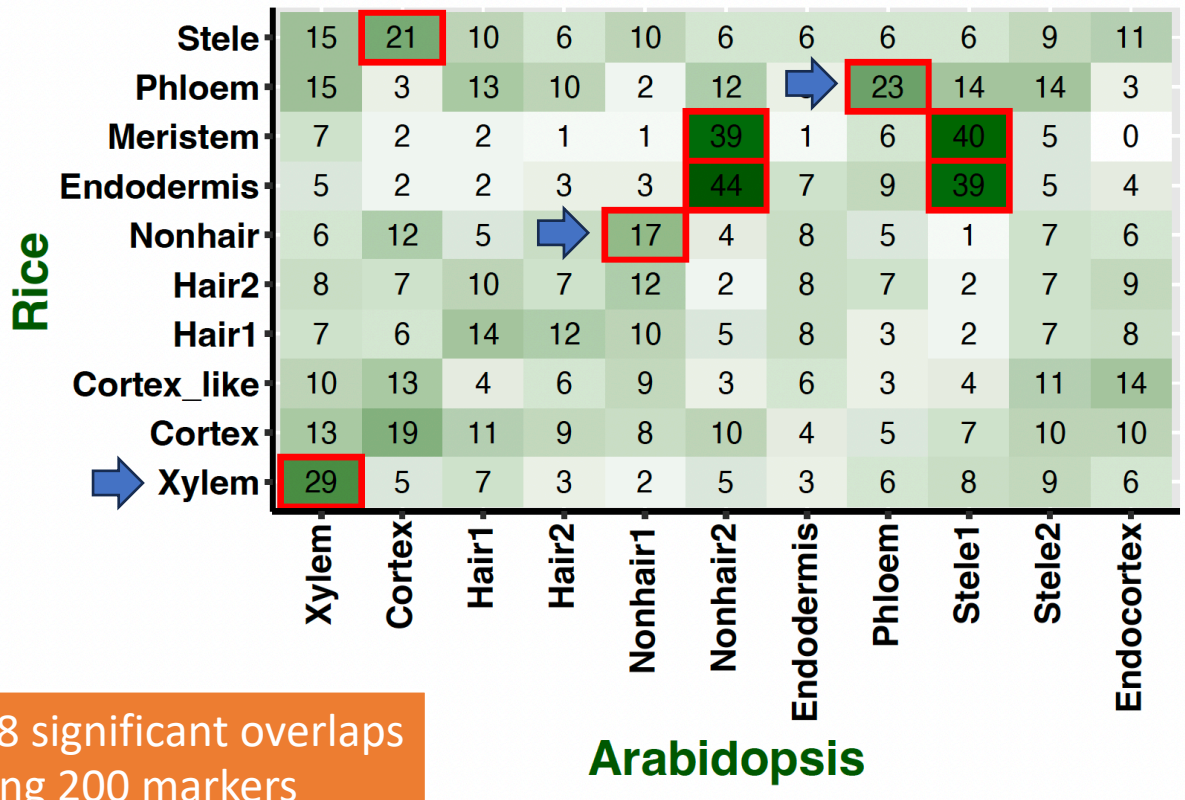
Existing Markers

Arabidopsis Rice

Yan et al., New Phytologist. 2022

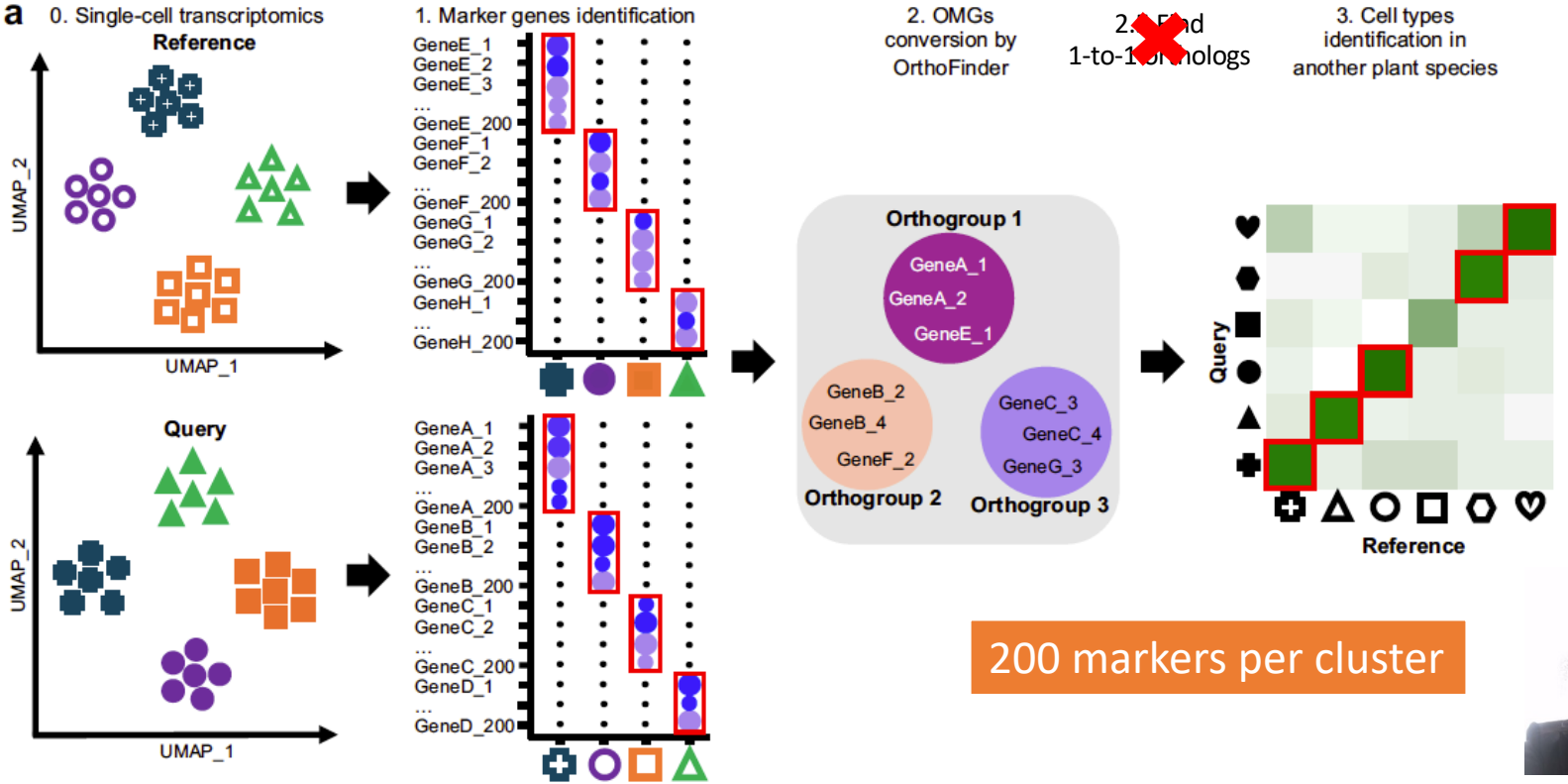
Github: https://github.com/LiLabAtVT/SPMarker

Rice and Arabidopsis one-to-one (FDR<0.01)

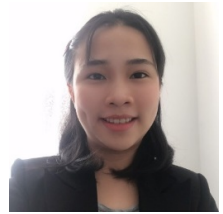


3 out of 8 significant overlaps
Using 200 markers

The OMG (Ortho Marker Groups) Pipeline

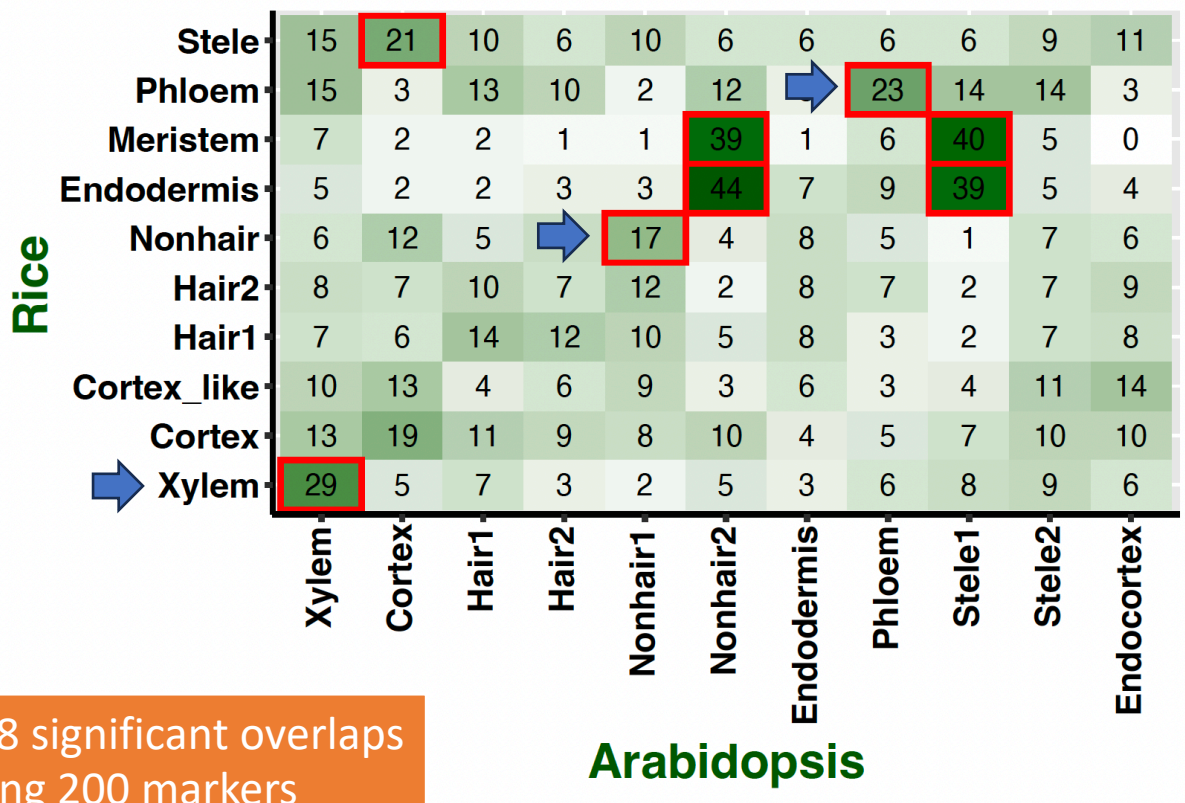


200 markers per cluster



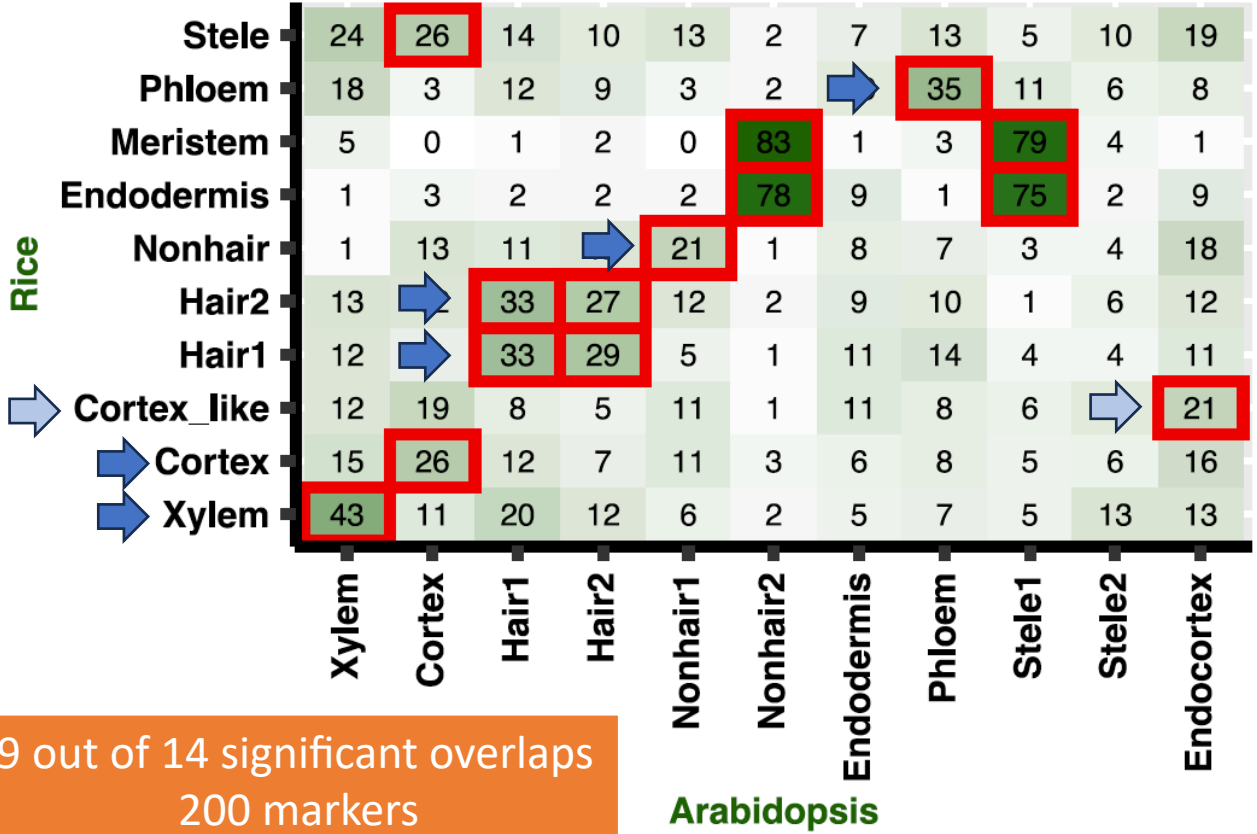
Tran (Nina) Chau
Graduate Student

Rice and Arabidopsis one-to-one (FDR<0.01)



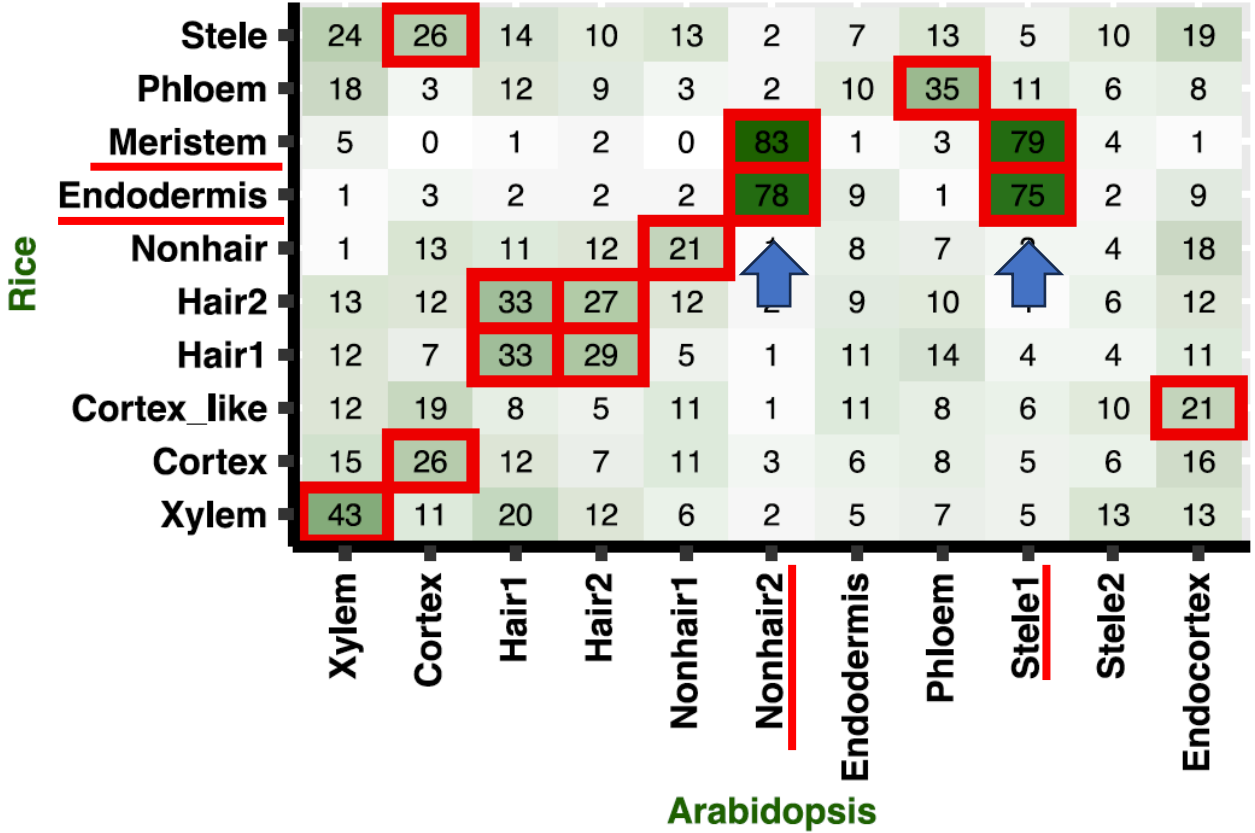
3 out of 8 significant overlaps
Using 200 markers

Rice and Arabidopsis OMG method (FDR < 0.01)

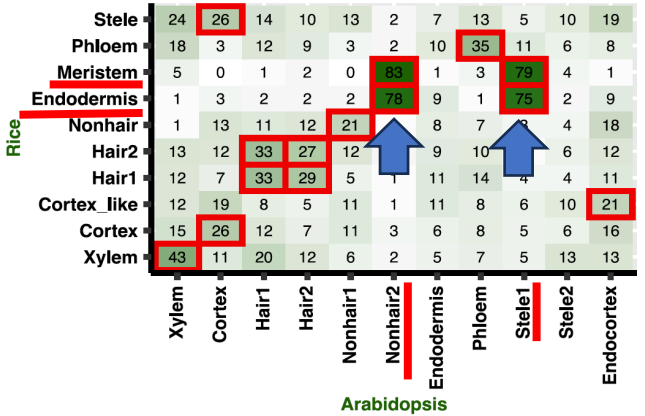


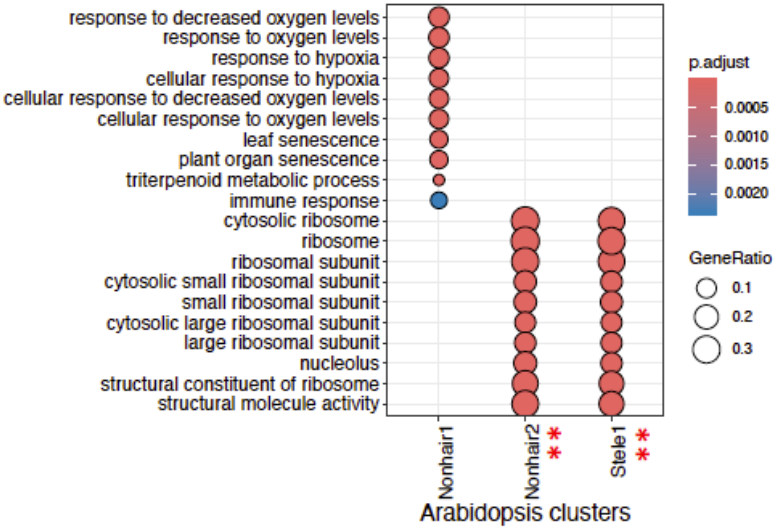
9 out of 14 significant overlaps
200 markers

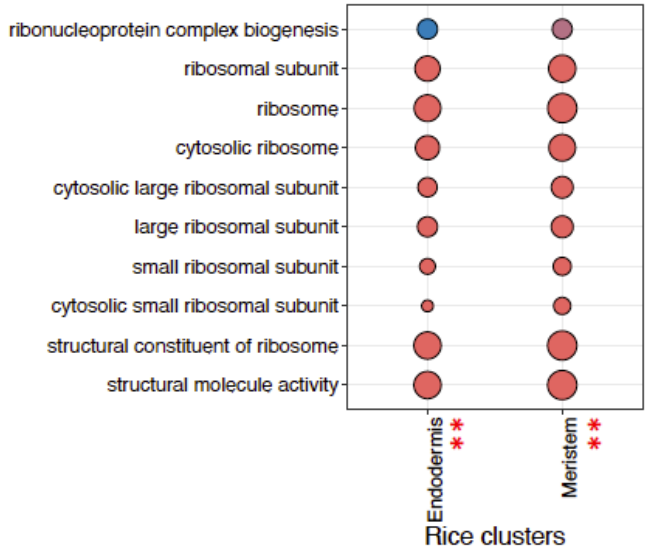
Rice and Arabidopsis: wrongly labeled cell types



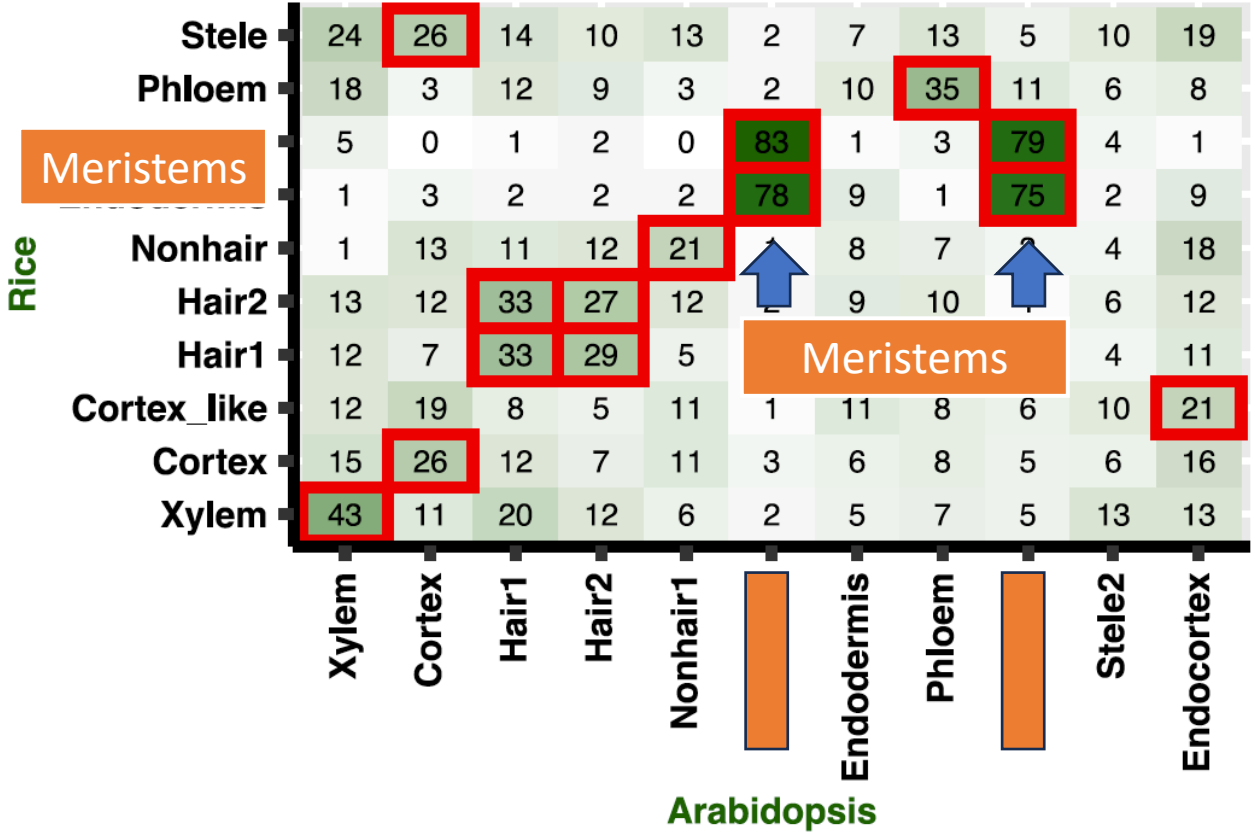
Rice and Arabidopsis: wrongly labeled cell types







Rice and Arabidopsis: wrongly labeled cell types



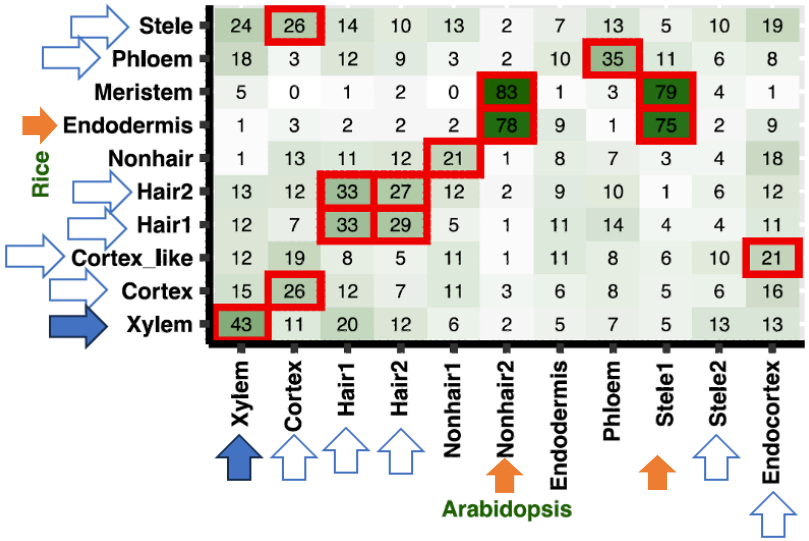
low-quality cells?

protoplast artifacts?

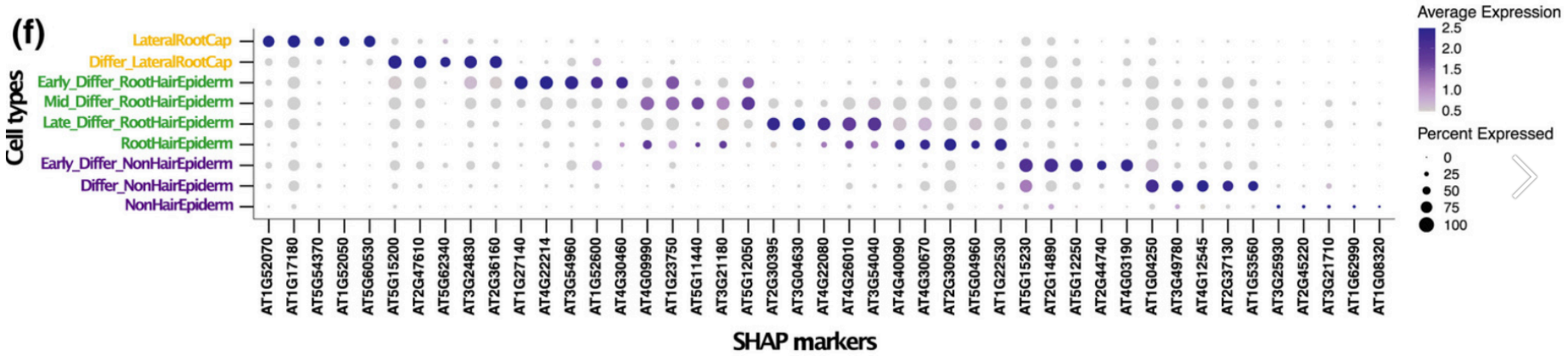
meristematic nonhair cells?

Summary

- Assign cell identities to single cell clusters in plants using OMG
  - Enables cross species comparison in plants
  - Marker based method == scalable comparison across many datasets
  - Use Seurat output
  - Importance of using a statistical test
  - OMG websites and R package
- Other tools:
  - Co-expression analysis
  - 3UTR annotation
  - GO function prediction (published)

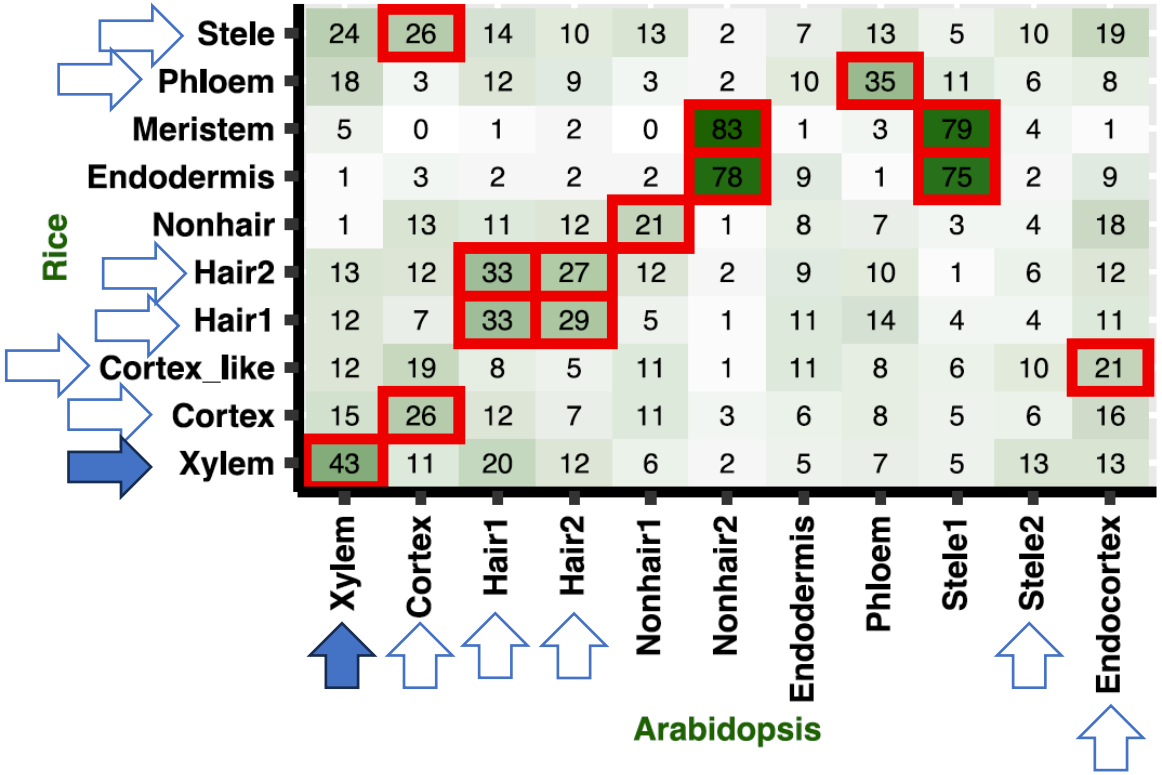


Problem with using a few markers and the importance of using a statistical test

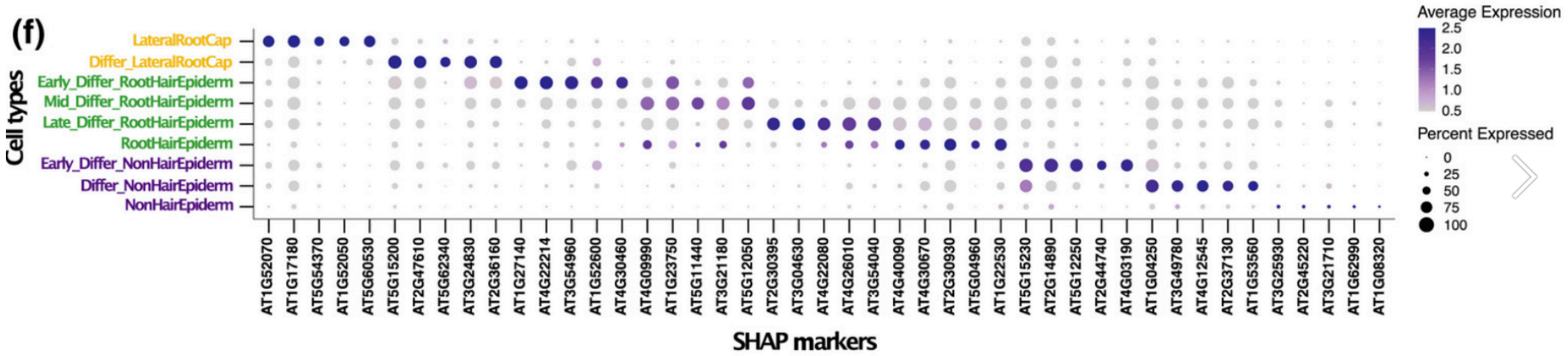


Yan et al., 2022

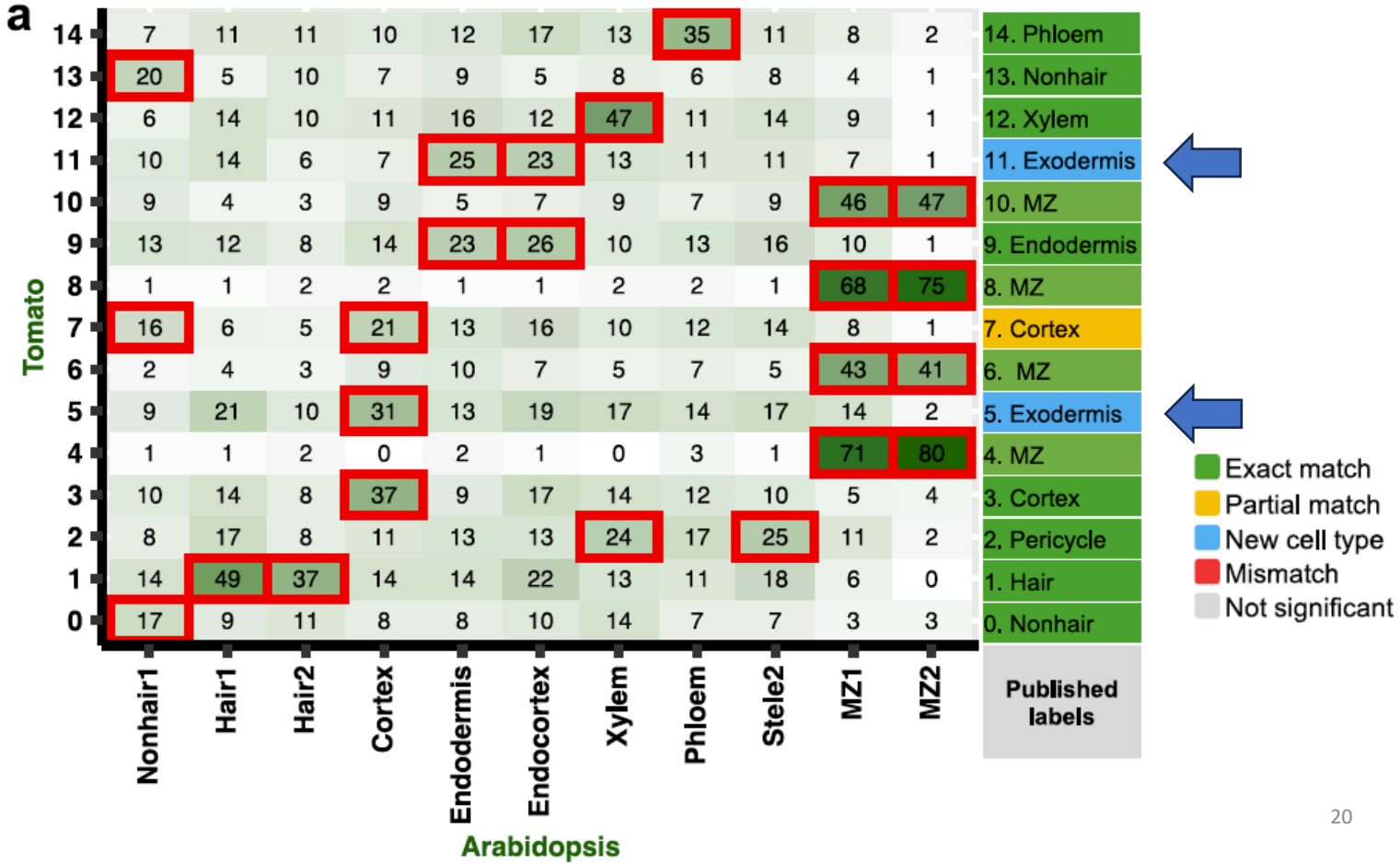
Problem with using a few markers and the importance of using a statistical test



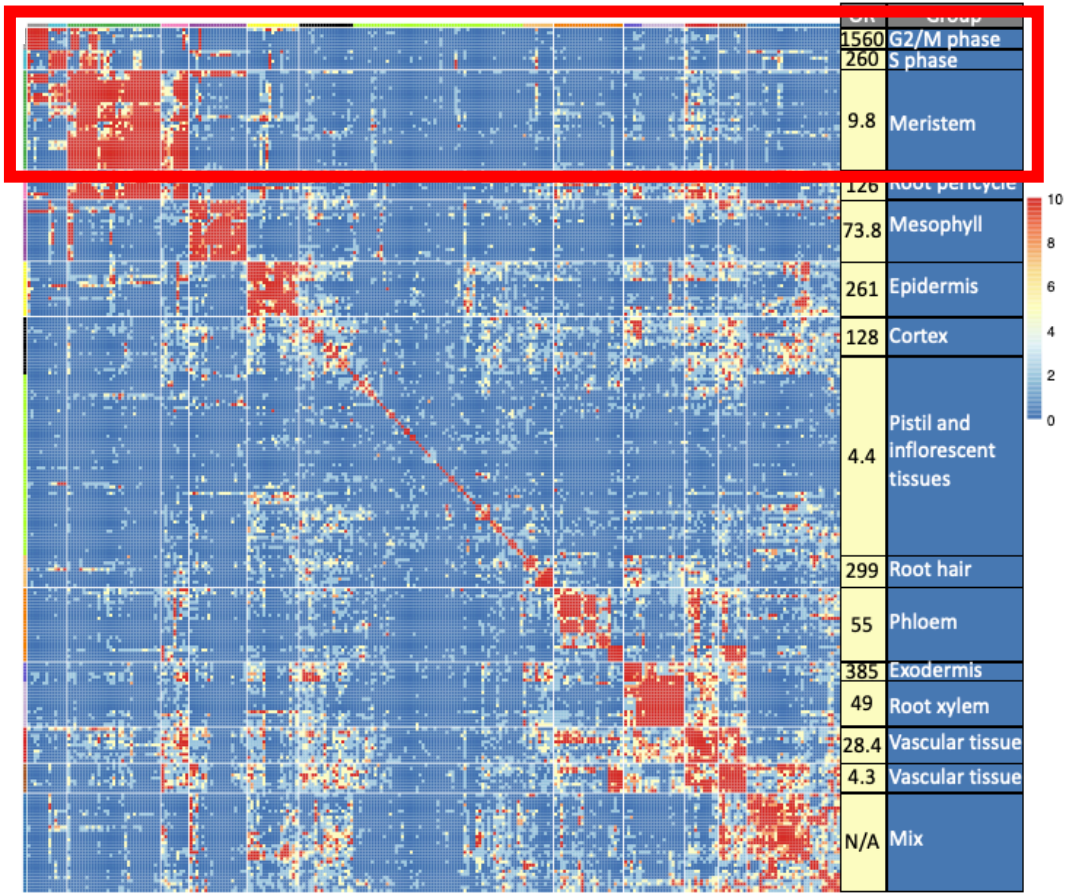
Problem with using a few markers and the importance of using a statistical test



Arabidopsis and tomato: new cell type



Mapping of single cell clusters from 15 species



- Gene expression in each individual cell
  - Root
  - Shoot apex
  - Leaf
  - Inflorescence
  - Whole plant

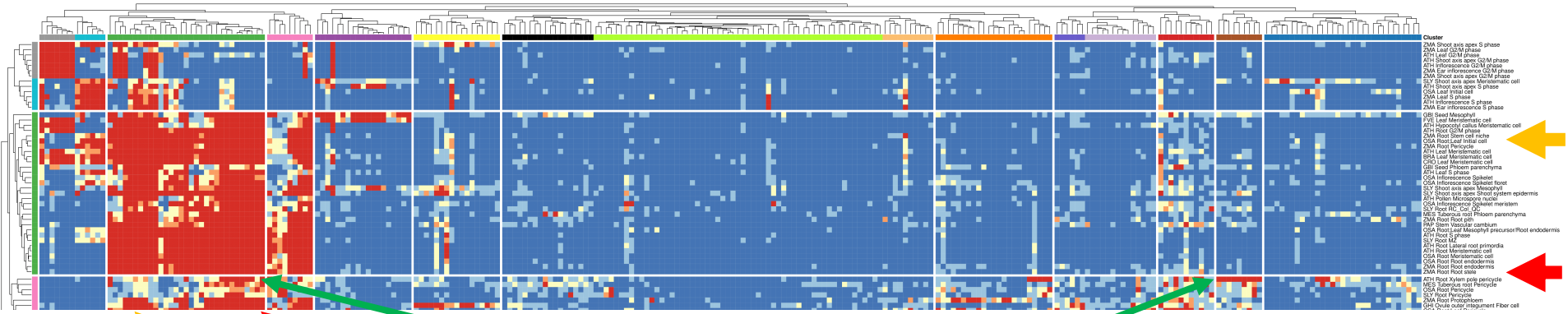
Nina (Tran Chau)

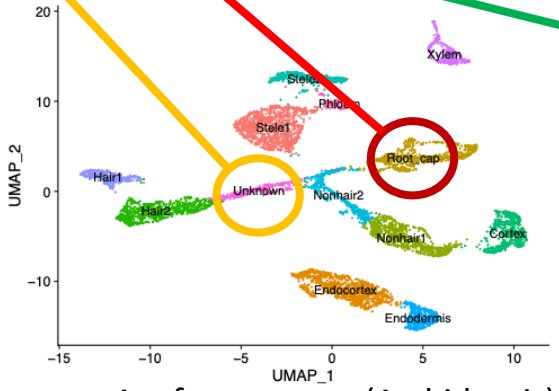
- 15 species, ~1 million cells, and 268 clusters

- Orthologous marker groups enable cell type mapping between monocots and dicots.

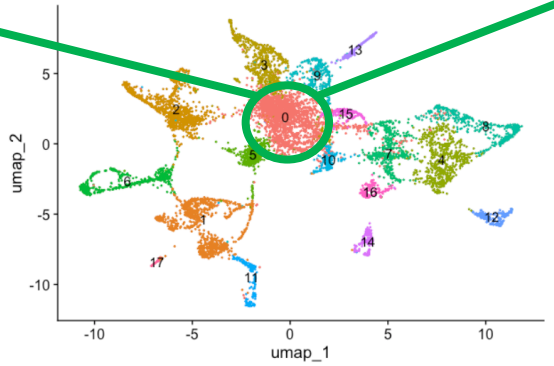
Chau et al., 2025. Nature Communications

Using OMG to annotate plant single cell maps





A reference map (Arabidopsis)

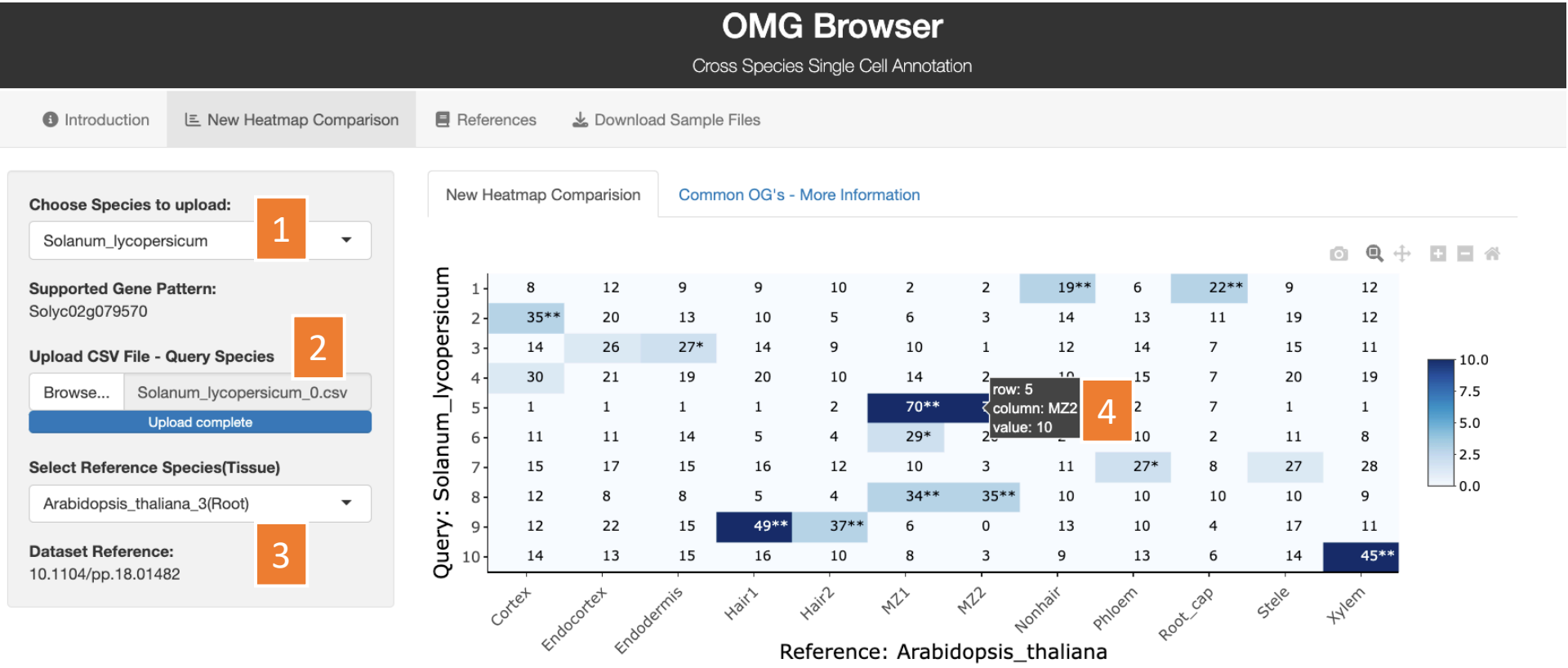


Query single cell map

- Integration free mapping across diverse plant species
- Allows statistical testing regarding cell identities

GitHub

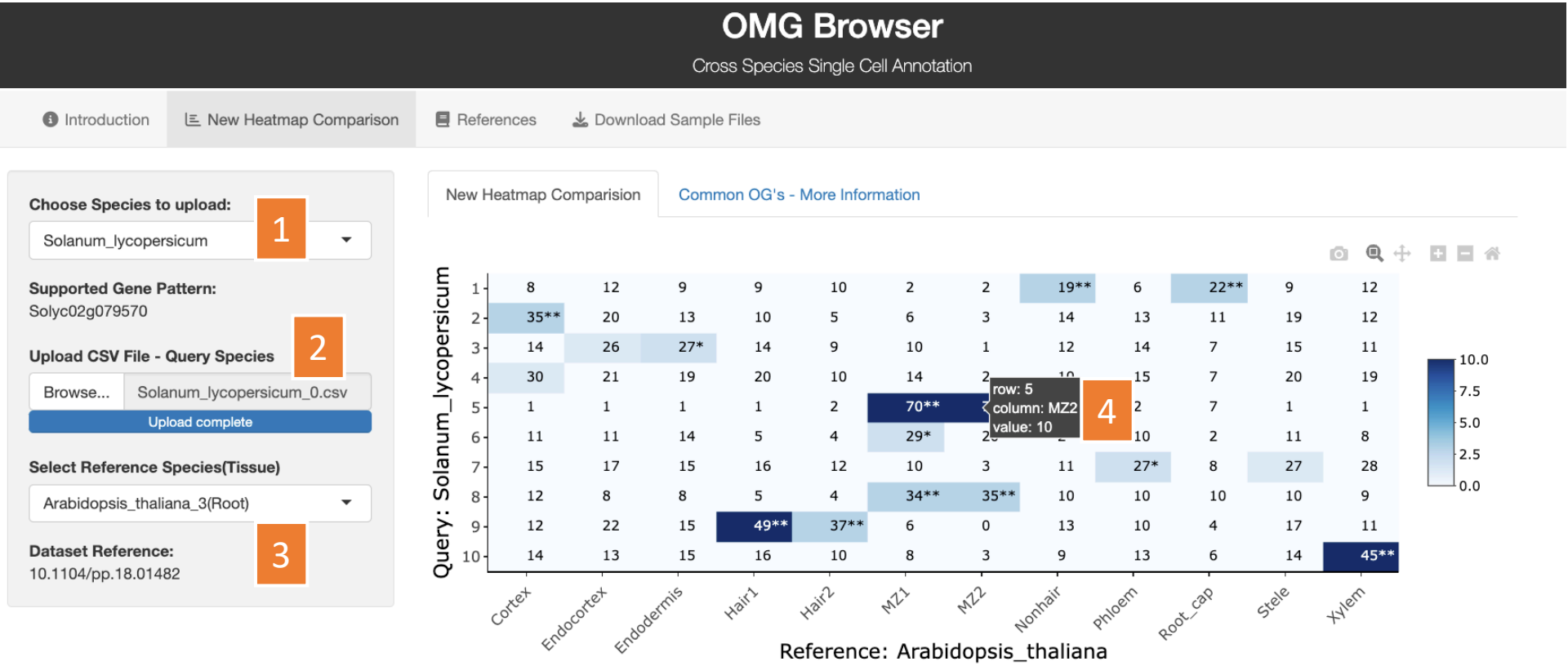
The OMG website



Input data: Seurat marker list file with fold change

name	p_val	p_val_adj	pct_1	pct_2	pct_diff	avg_log2FC	clusterName
Solyc04g011390	0	0	0.789	0.334	0.455	2.11288019	5
Solyc01g074000	0	0	0.859	0.464	0.395	2.10869367	5
Solyc01g086820	0	0	0.812	0.362	0.45	2.074077	5
Solyc01g099410	0	0	0.904	0.653	0.251	1.98499678	5
Solyc09g082710	0	0	0.856	0.405	0.451	1.83647942	5
Solyc01g079110	0	0	0.725	0.276	0.449	1.82659547	5
Solyc06g064630	0	0	0.987	0.777	0.21	1.73490661	5
Solyc06g083820	0	0	0.978	0.708	0.27	1.72068817	5
Solyc11g065190	0	0	0.637	0.167	0.47	1.6700392	5
Solyc02g077480	0	0	0.796	0.416	0.38	1.64235017	5

The OMG website



Output: marker name in ref/query species

Upload and Select a cell to know more

Show 10 entries

Search:

Cell Data

	gene	Orthogroup	clusterName	avg_log2FC	Species
1	Solyc04g011390	OG0000127	5	2.11288019470838	Query
2	Solyc09g082710	OG0002013	5	1.83647942297174	Query
3	Solyc06g083820	OG0002863	5	1.72068817262044	Query
4	Solyc03g078290	OG0003259	5	1.63410092596824	Query
5	Solyc06g007220	OG0019891	5	1.61401336840312	Query
6	Solyc03g120780	OG0003269	5	1.59451193967745	Query
7	Solyc03g080160	OG0001029	5	1.57602543221737	Query
8	Solyc05g054610	OG0000127	5	1.57592060718733	Query
9	Solyc01g096580	OG0001953	5	1.57479654374213	Query
10	Solyc08g006900	OG0001840	5	1.54722293086819	Query

Showing 1 to 10 of 289 entries

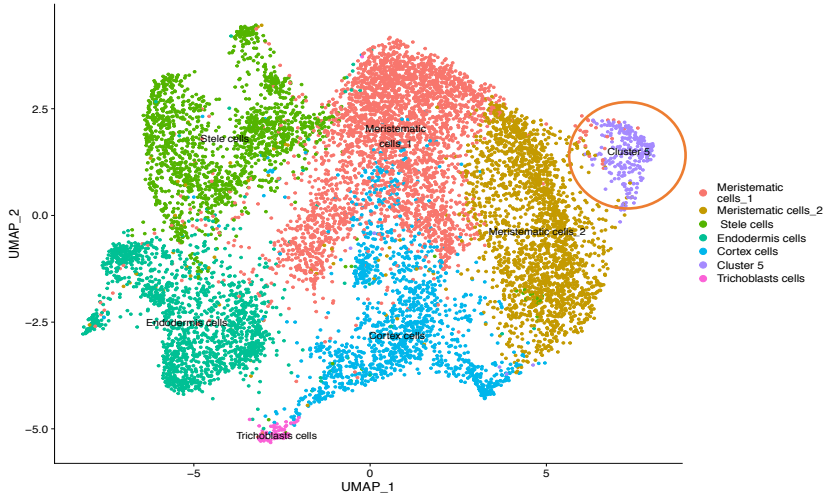
Previous 1 2 3 4 5 … 29 Next

Download Table

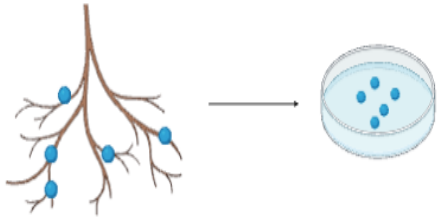
How to use OMG?

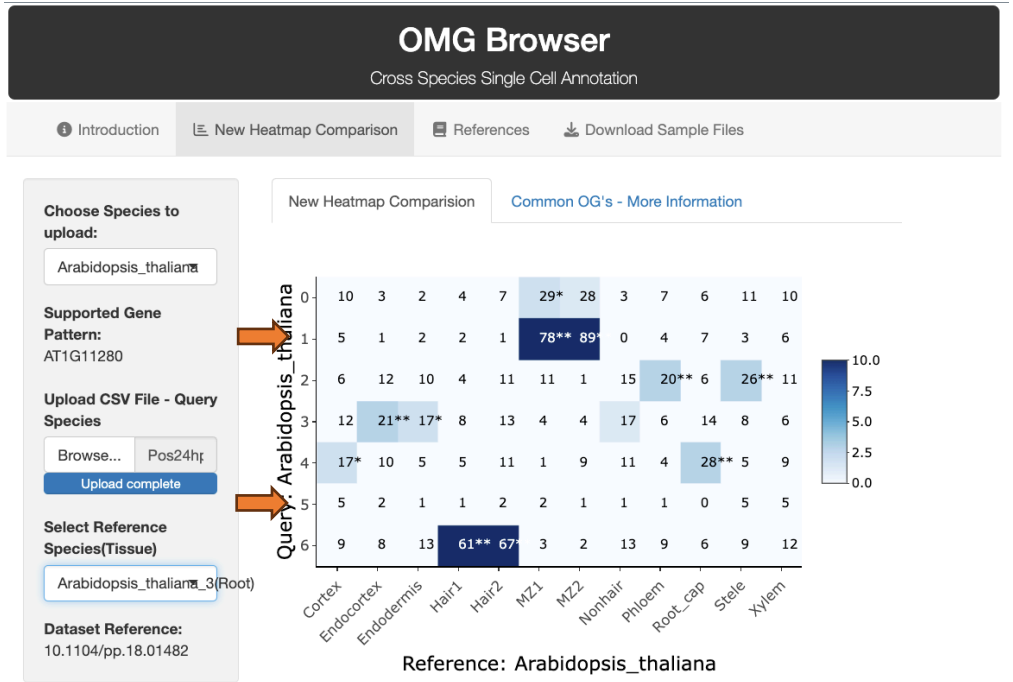
- Annotate single cell clusters for plant species
  - For 15 species in the current version
    - Example using Arabidopsis + pathogen infection
  - For species not in the current version
    - BLAST2OMG
    - LLM2OMG
- Connect plant single cell data with other domains of research and applications
  - GWAS/GP → Plant breeding
  - Genetic engineering, synthetic biology
- Apply in non-plant systems?
  - Interested in testing this concept in other systems.

How to use OMG? Arabidopsis pathogen interaction

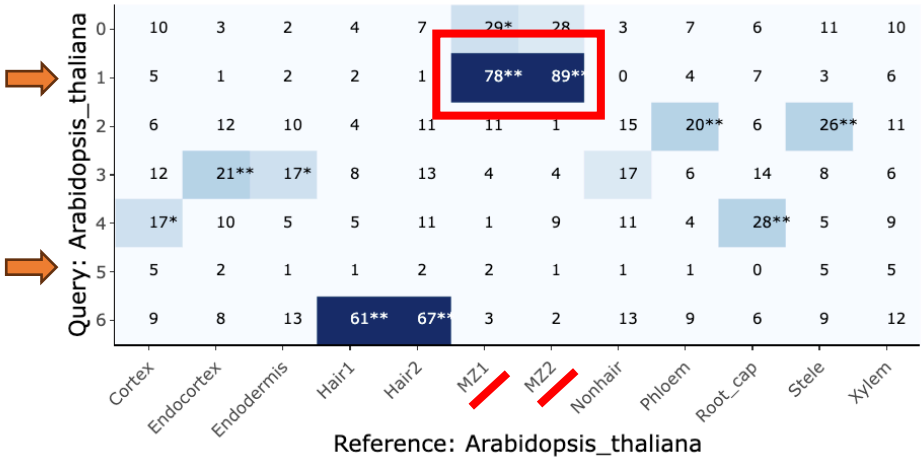


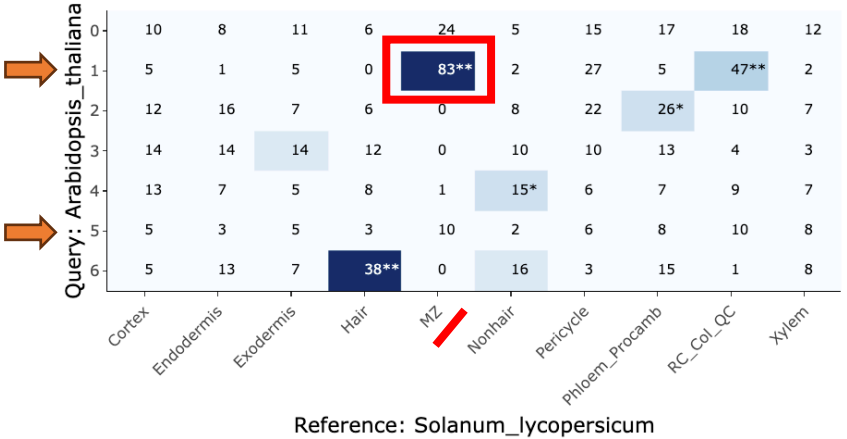
Infected roots **Nuclei Isolation**

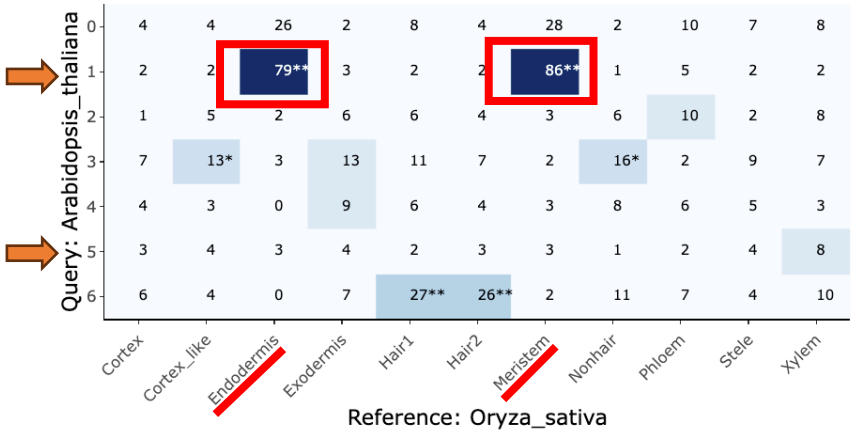




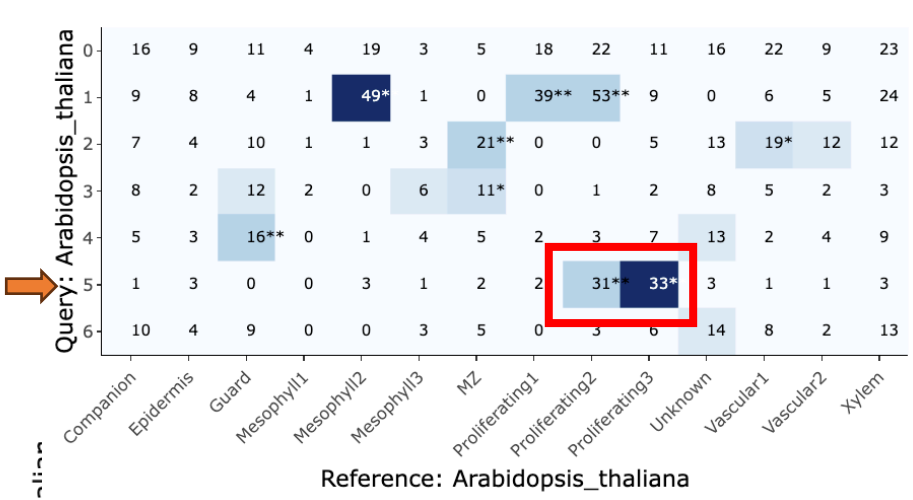
Arabidopsis root + pathogen vs other roots

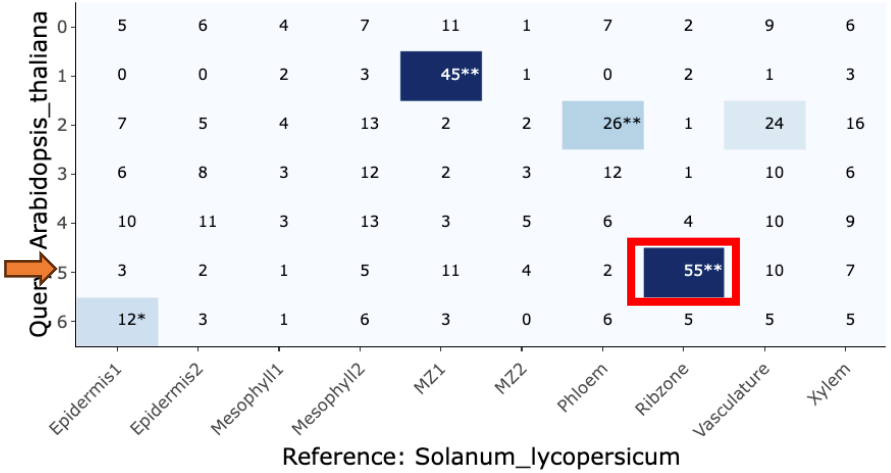


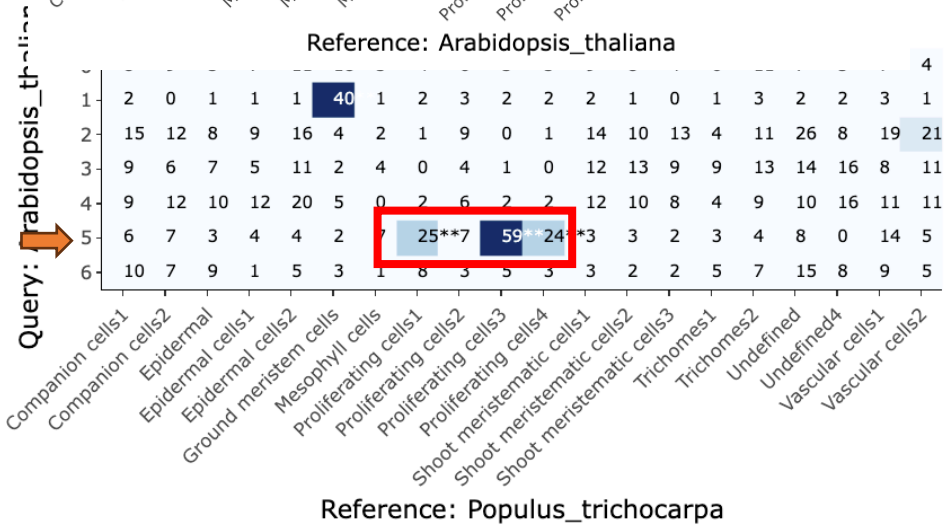


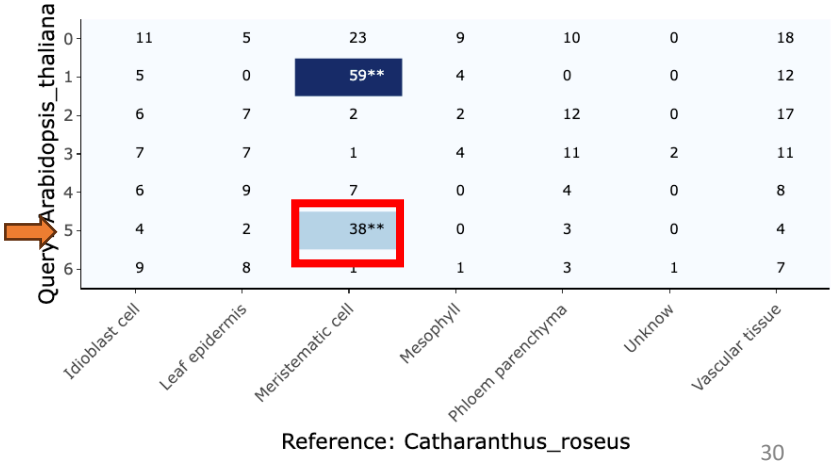


Arabidopsis root + pathogen vs shoot apex/leaf







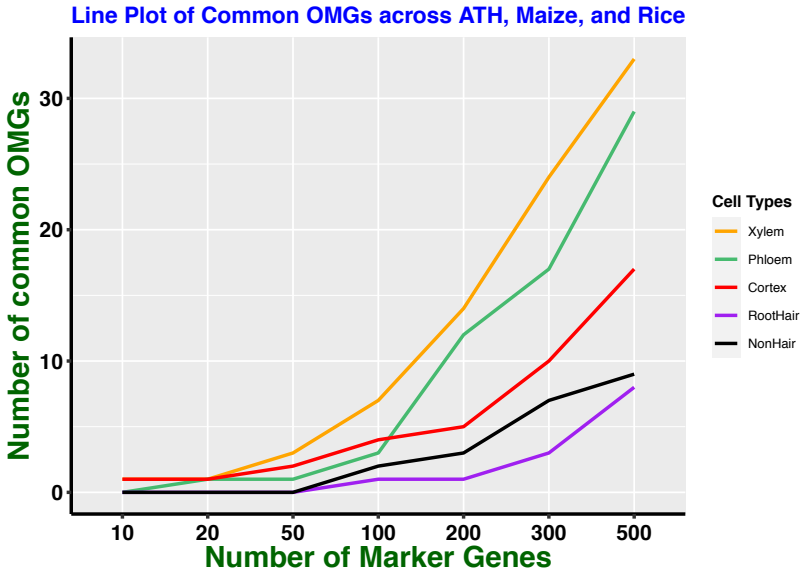


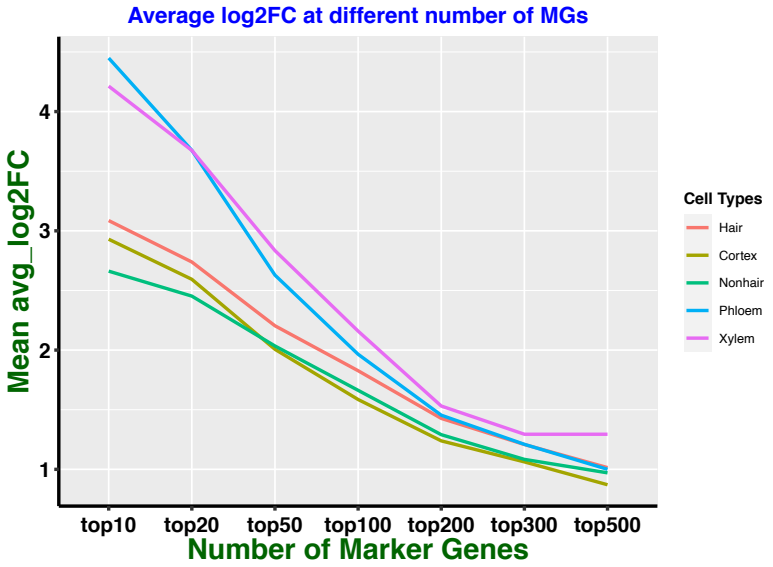
How to use OMG?

- Annotate single cell clusters for plant species
  - For 15 species in the current version
    - Example using Arabidopsis + pathogen infection
  - For species not in the current version
    - BLAST2OMG
    - LLM2OMG
- Connect plant single cell data with other domains of research and applications
  - GWAS/GP → Plant breeding
  - Genetic engineering, synthetic biology
- Apply in non-plant systems?
  - Interested in testing this concept in other systems.

Limitations

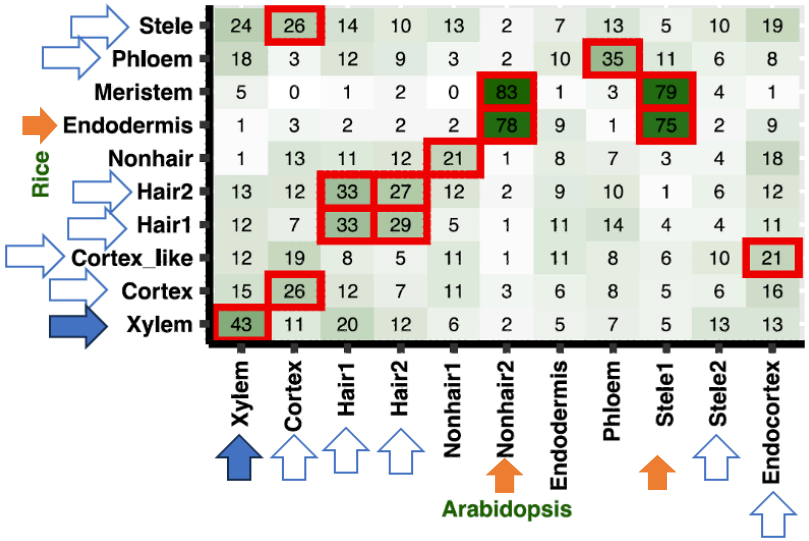
- Number of clusters
- Number of markers
- Inconsistent naming
- Reproduce the same published UMAP clusters



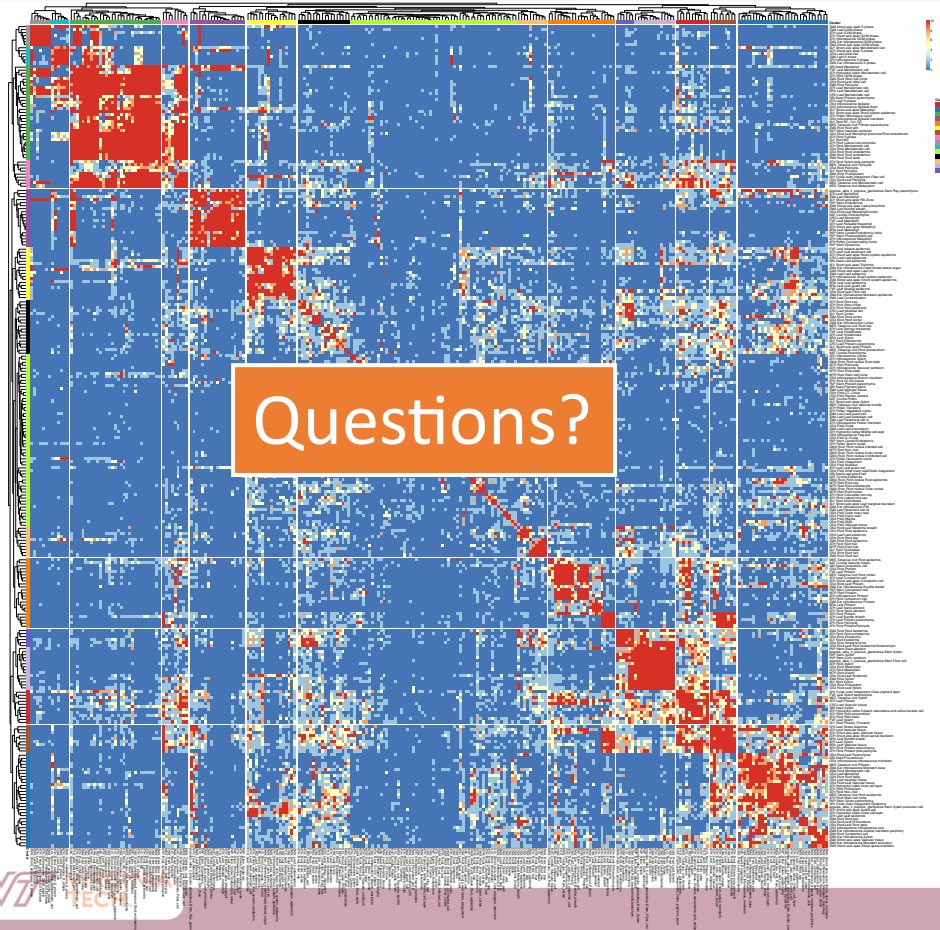


Summary

- Assign cell identities to single cell clusters in plants using OMG
  - Enables cross species comparison in plants
  - Marker based method == scalable comparison across many datasets
  - Use Seurat output
  - Importance of using a statistical test
  - OMG websites and R package
- Other tools:
  - Co-expression analysis
  - 3UTR annotation
  - GO function prediction (published)



Summer camp for incorporating your data to the OMG framework



Plant single-cell atlas:

- 15 species
- 268 cell clusters
- ~1 million cells
- 53,600 marker genes
- 200 highly specific marker per cluster
- 14 “functional” clusters

- NSF funded summer training camp for single cell analysis at Virginia Tech (email songli@vt.edu)

Chau et al., 2025, Nature Communications