Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: 09 January 2023
Manuscript Category

OXFORD

Subject Section

Single-cell analysis of Arabidopsis roots.

Lotte Van de Vreken

Master of Science in Biochemistry and Biotechnology - Ghent University
C002703A - Data Mining
Prof. Dr. Yvan Saeys

Abstract

Motivation: *Arabidopsis thaliana* is an important model species in plant molecular sciences. It serves as the basis for all fundamental research and crop improvement programs. Plant roots are important for water uptake and nutrient acquisition. Despite their importance, they are often neglected in research due to their growth in soil which makes it more difficult to study them. That is why they are often described as the hidden half. Single-cell transcriptomics can give more insights in root development and identify key regulatory genes, which can form the basis of new plant breeding approaches.
Results: In this study, two single cell datasets of Arabidopsis roots were analyzed. The data was preprocessed, clustered and cell types were assigned to each cluster. The detected amount of cells were compared between both datasets for each cell type.
Contact: lotte.vandevreken@ugent.be
Supplementary information: Jupyter notebooks and original data are available at my personal Github page: https://github.com/lotte-vandevreken/DM_sc_Arabidopsis_root.git

1 Introduction

Arabidopsis thaliana is a small, annual rosette plant that belongs to the Brassicaceae (Krämer, 2015). Many well-known vegetables such as brussel sprouts, cabbage and cauliflower also belong to this family. The small size and short life-cycle of *Arabidopsis* made it very popular as a model plant for angiosperms in the 1980s. Up until today, this plant is being used over the whole world as the first step in fundamental plant research after which knowledge can be transferred to crop species. *Arabidopsis* is an attractive model plant because of its short generation time. In 6 weeks, a seed can grow out to a mature plant and generate its own seeds (Meyerowitz, 1987). Another factor that makes this plant really interesting is the ability to cross- and self-pollinate. Because of these nice characteristics, research institutes all over the world started doing research on this little plant, resulting in a large community where knowledge is shared and great progress is being made. Research on this plant forms the basis of modern crop improvement and modern breeding techniques.

Plant roots have a radial pattern with three fundamental types of tissue: the protoderm (e.g. the epidermis), the ground tissue (e.g. the cortex), and the vascular tissue (Schiefelbein *et al.*, 1997). In the Arabidopsis primary root, there are single cell-thick rings of epidermis, cortex, endodermis, and pericycle tissues with a constant number of eight cells per ring for the cortex and endodermis layers surrounding the central stele (Figure 1). Root systems of terrestrial plants have two main functions: acquisition of soil-based resources and anchorage (Fitter, 2002). Additionally, they

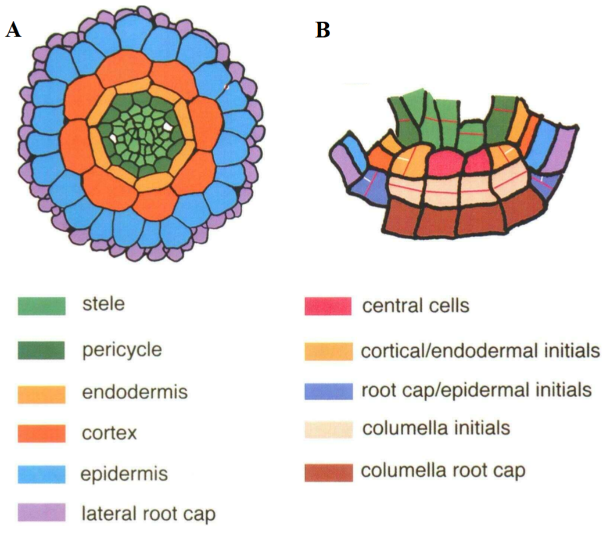

Fig. 1. Cellular Organization of the Arabidopsis Root. (A) Colorized drawing of the transverse section of an Arabidopsis root from the late meristamatic region. A single layer of lateral root cap cells surrounds the epidermis. (B) Colorized drawing of the promeristem region in an Arabidopsis root apex. The red lines indicate the planes of division that occur in the initials; white lines indicate the secondary divisions that occur in the cortical/endodermal and the lateral root cap/epidermal initials (Schiefelbein et al., 1997).

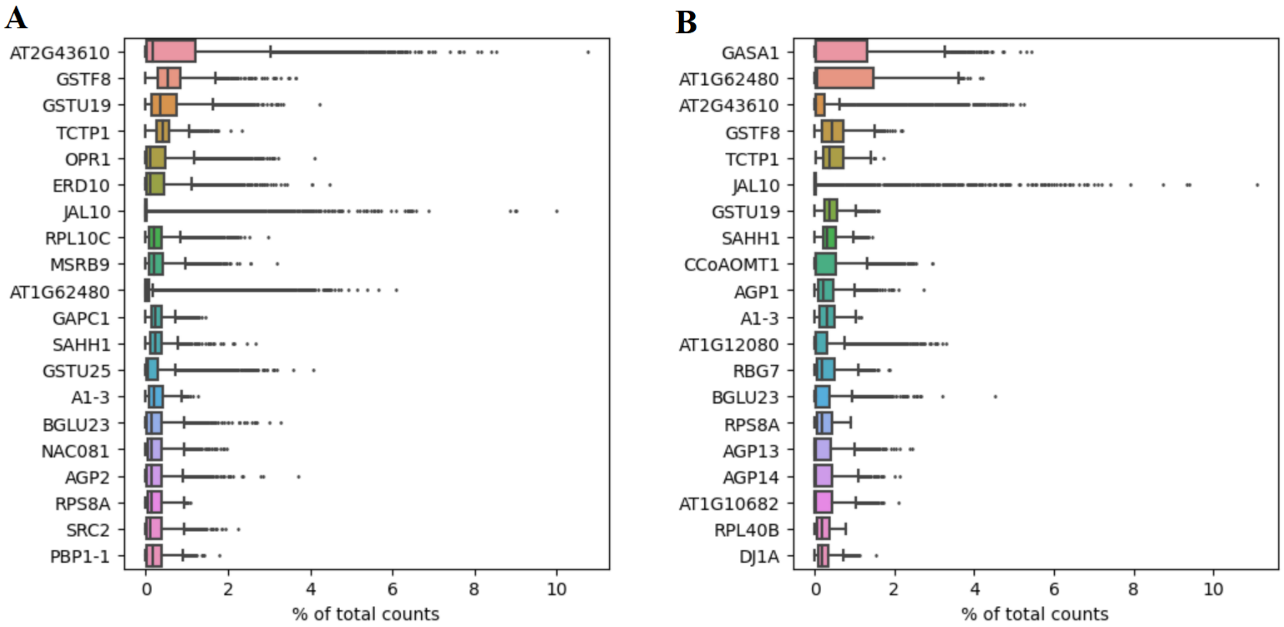

Fig. 2. Fraction of counts assigned to each gene over all cells. A Top 20 genes of Dataset 1. (B) Top 20 genes of Dataset 2.

have some other secondary functions such as storage, synthesis of growth regulators, defense, propagation and dispersal. Although that roots have a very important role in plant development and versatility, they are often overlooked in research. Researchers tended to only look into the visible parts of the plant, including the shoot and the leafs, and ignore the important role of the root system. Slowly, the importance of roots is being recognized and root system architecture, root gene expression and root signaling is being studied more and more. In this research field, single-cell technologies could result in a better understanding of root development and root responses to several stresses, which will have major beneficial effects in agriculture.

2 Material and Methods

2.1 Datasets

In this study, two single-cell RNA sequencing datasets of Arabidopsis roots were analyzed. For both datasets, 6-day old *Arabidopsis thaliana* seedlings were used. Different sequencing platforms were used: Denyer *et al.* (2019) (Dataset 1) generated the data with the NextSeq 550 (GSE123818), while Wendrich *et al.* (2020) (Dataset 2) used the Illumina HiSeq 4000 (GSE141730). Both original datasets were retrieved from the Gene Expression Omnibus (Edgar *et al.*, 2002).

2.2 Preprocessing

Both datasets were first preprocessed to allow faster analysis with higher accuracy. First, all protoplasting-induced genes were removed from the dataset. For the single-cell analysis of plant cells, protoplasting is necessary to achieve individual cells. However, due to this enzymatic treatment on the living cells, certain genes are induced (Birnbaum *et al.*, 2003). These genes are not relevant for the analysis and were removed from the dataset. Dataset 1 only contained the gene stable IDs for each gene. These gene IDs were replaced with their corresponding gene name using Ensembl BioMarts (Kinsella *et al.*, 2011).

Using the Python *scanpy* package, the data was further preprocessed (Wolf *et al.*, 2018): cells with less then 200 genes and genes that occurred in less than 3 cells were removed. The presence of mitochondrial genes, which indicates low quality, was checked. Cells with a high amount of different genes were removed (>12000 for Dataset 1; >13000 for Dataset 2). Finally, the data was library size corrected and log transformed.

2.3 Clustering

To reduce the dimensions but keep as much of the information as possible, a principle component analysis (PCA) was performed. Then, the neighborhood graph was calculated with 30 neighbors and 50 principle components (PCs). The data was clustered with Louvain clustering with a resolution of 1.0. Data visualization was performed with UMAP (McInnes *et al.*, 2018).

2.4 Cell type annotation

Using t-tests, highly differential genes of each cluster were identified and ranked. The top 500 differentially expressed genes for each cluster were compared with a list of known marker genes derived from literature. For the clusters that could not be annotated in this way, the expression patterns of the top 5 differentially expressed genes was observed using the BAR eFP browser (Winter *et al.*, 2007). If these expression patterns were clear and similar for the top 5 differentially expressed genes, the cluster cell type could be assigned based on that. To determine whether the data was overclustered resulting in clusters containing the same cell type, a correlation matrix was made.

3 Results

3.1 Preprocessing

The data of Dataset 1 and Dataset 2 was preprocessed as described above. To gain some insights in the data, the fraction of counts assigned to each gene over all cells was visualized. The 20 genes with the highest means are shown in Figure 2.

Dataset 1 contained 4727 cells with 27629 genes. After preprocessing, 4720 cells with 21678 measured genes were remained in the dataset.

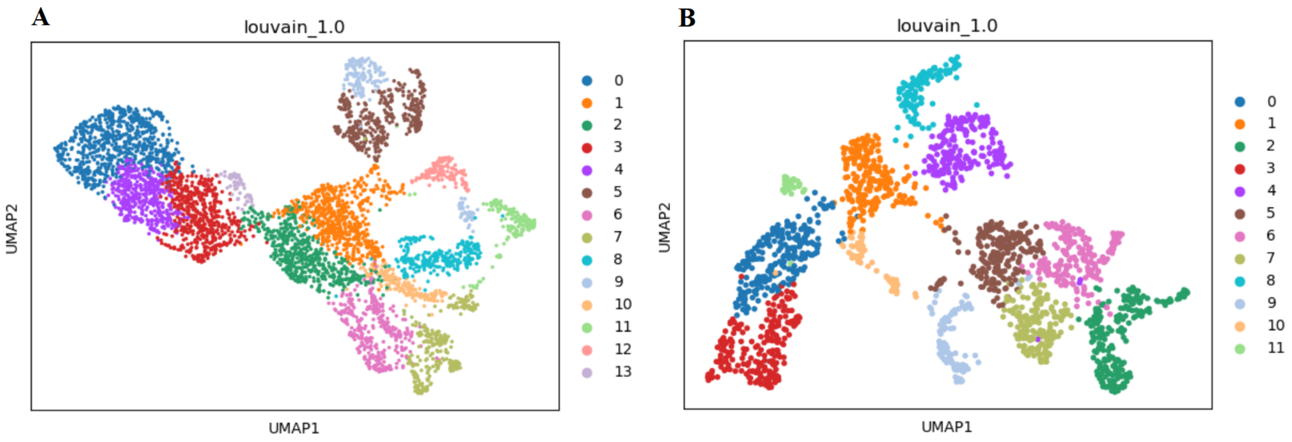

Fig. 3. Louvain clustering. The clustering was performed with the sc.tl.louvain function from the scanpy package with a resolution of 1.0. (A) Louvain clustering of Dataset 1. (B) Louvain clustering of Dataset 2.

Dataset 2 contained 1975 cells with 34262 measured genes. After preprocessing, 1973 cells with 22275 measured genes were remained in the dataset.

3.2 Clustering

On both datasets, PCA was performed to reduce the amount of dimensions while remaining maximal variance. The neighborhood graph was computed using 30 neighbors and 50 principle components. With Louvain clustering, the data was clustered. For Dataset 1, the clustering resulted in 14 clusters (Figure 3A). Dataset 2 was clustered in 12 clusters (Figure 3B).

3.3 Cell type annotation

Genes that are highly expressed in one cluster but lowly expressed in all other clusters were identified by means of a t-test. To be able to assign a cell type to each cluster, known marker genes were searched in the top 500 differentially expressed genes of each cluster.

3.3.1 Known marker genes

For Dataset 1, known marker genes were found for cluster 0, 3, 4, 5, 7, 8, 9, 10, 11, 12, and 13. In cluster 0, 3 and 4, marker genes for root cap cells were found. However, a distinction between the clusters could not be made because some markers appeared in multiple of these clusters: SMB and BRN2 were detected in cluster 0 and cluster 4; ACR4 was detected in cluster 3 and cluster 4. This indicates that either these known marker genes are not specific enough to be able to distinguish between specific root cap cell types or that the data was overclustered meaning that cluster 0, 3, and 4 actually contain the same cell type. For cluster 5 and 9, markers for the stele were detected. DAG1 was detected in both clusters, implicating again that DAG1 expression does not allow us to distinguish between two cell types or that there is overclustering. Cluster 7 had clear markers for trichoblasts, cluster 8 for atrichoblasts. In cluster 10, only one marker was found, specific for atrichoblasts. Because only one marker was detected in cluster 10 and two were detected in cluster 8, it was assumed that cluster 8 contains true atrichoblast cells. Cluster 10 is located next to cluster 8, which could mean that both clusters contain the same cell type (Figure 3). Cluster 11 had markers for the cortex; cluster 12 for the endodermis. In cluster 13, marker genes for the quiescent centre (QC) and for the root cap were found. No markers were found for clusters 1, 2, and 6.

For Dataset 2, known marker genes were found for cluster 0, 2, 3, 5, 6, 7, 8, 9, and 11. In cluster 0 and cluster 3, marker genes for the root cap were found. In both clusters, SMB was detected. The other markers were different between both clusters but have overlapping expression patterns in columella and the lateral root cap. Therefore, it was not possible to assign distinct cell types to these clusters. In cluster 2, two markers were found that are being expressed in the procambium and the xylem. Additionally, a third marker was found which is specific for the stele. The stele is an overarching term for all cell types in the central part of the root derived from the procambium. Therefore, the more specific cell types procambium and xylem were being preferred. Cluster 5 had markers for the stele, pericycle and xylem pole pericycle (XPP). Both pericycle and XPP belong to the stele, but because two markers of XPP were found and only one of pericycle, this cluster probably contained XPP cells. Cluster 6 contained multiple markers for the phloem. Cluster 7 contained two markers for the stele and one marker for the pericycle. Just as in cluster 2, the more specific cell type pericycle is preferred. Cluster 8 contained markers that had expression patterns in both the cortex and the endodermis. These markers are not specific enough to distinguish between these two cell types. Cluster 9 contained markers of the endodermis. Therefore, it could be concluded that cluster 8 most probably contained cells from the cortex and cluster 9 from the endodermis. In cluster 11, markers for multiple cell types including QC, columella, lateral root cap and procambium were found. Because of the variety of reported cell types and the presence of the QC in this list, and additionally that this cluster was small in size, cluster 11 most probably contained cells from the QC. No markers were found in clusters 1, 4, and 10.

3.3.2 Correlation between clusters

In both datasets, root cap cell types could not be distinguished. Additionally, some clusters did not contain any differentially expressed

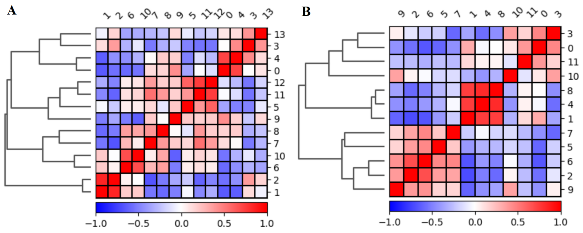

Fig. 4. Correlation matrix of clusters detected by Louvain clustering. (A) Dataset 1. (B) Dataset 2.

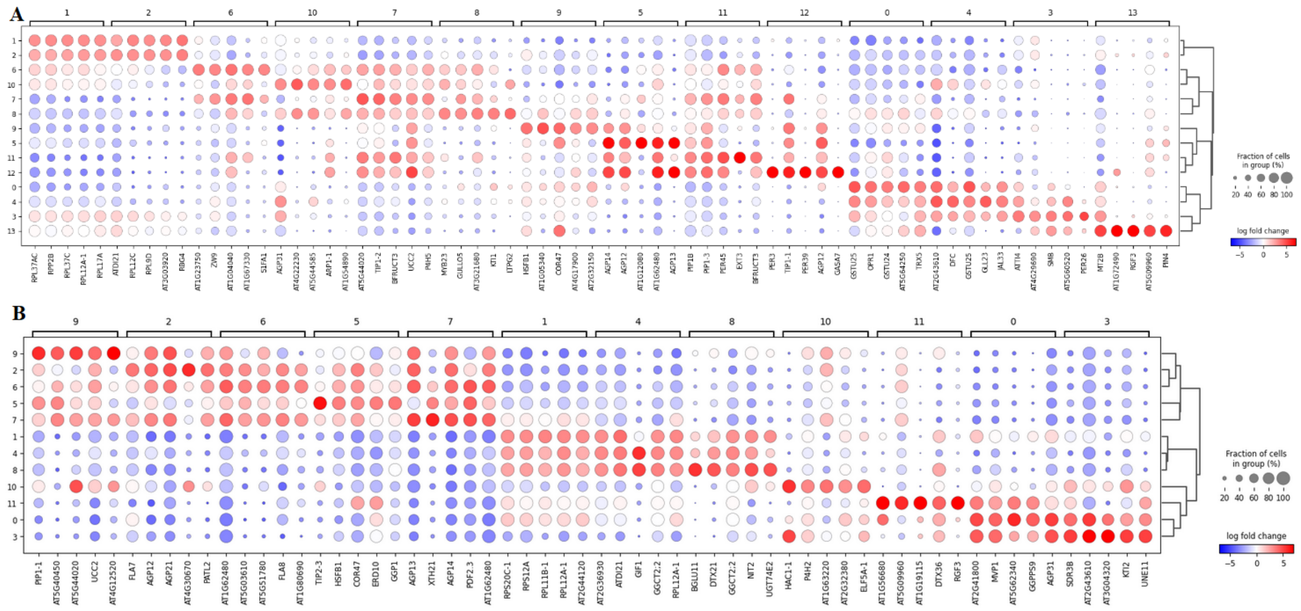

Fig. 5. Dotplot of log2 fold changes of the top 5 differentially expressed genes for each cluster. (A) Dotplot for Dataset 1. (B) Dotplot for Dataset 2.

marker genes. To investigate how similar the clusters with root cap cells are and to look at possible overclustering, a correlation matrix was made (Figure 4).

Between the clusters of Dataset 1, limited correlation was observed. Cluster 1 and 2 were highly correlated, which could mean that they contain the same cell type. Cluster 6 and 10 were also highly correlated. There was only limited correlation between cluster 10 and cluster 8, which indicates that cluster 10 contained a different cell type than cluster 8 which was classified as atrichoblast. High correlation could be seen between clusters 11 and 12. The endodermis is the innermost cortical cell layer. This explains the high similarity in expression profiles. Clusters 0, 3, 4 and 13 were also correlated to each other, with cluster 0 and 4 being the most correlated. This corresponds to the found marker genes, whereby cluster 0 and 4 had the most overlapping marker genes. Cluster 13 was classified as the QC, which provides precursor root cap cells. Therefore, this correlation with root cap cell types strengthens this classification.

Between the clusters of Dataset 2, a higher correlation is observed than in Dataset 1. Clusters 2, 5, 6, 7, and 9 are correlated. In all these clusters, marker genes for different cell types that are located in the stele were found. These cell types are all closely related but distinct enough to assign different cell types to them. Clusters 1, 4, and 8 are highly correlated to each other. Cluster 8 had marker genes for cortex cells. Cluster 1 and 4 had no marker genes. Because of this correlation matrix, it is very likely that cluster 1 and 4 also contain cortex cells. Cluster 0, 3 and 11 are correlated. Based on the presence of marker genes, cluster 0 and 3 contain root cap cells. However, the correlation matrix showed that they are correlated but distinct enough to form two cell types. The correlation matrix also supports the hypothesis that cluster 11 contains QC cells because it is moderately correlated to root cap cells.

3.3.3 Identification of clusters without marker genes

To assign a cell type to the clusters without marker genes, expression patterns of the top differentially expressed genes were studied using the BAR eFP browser. These patterns were compared with the dotplot of the expression of these genes in the different clusters (Figure 5).

For Dataset 1, no cell type was yet assigned to clusters 1, 2, 6, and 10. In cluster 1 and 2, predominantly ribosomal proteins are differentially upregulated. Expression patterns of the top 5 differentially expressed genes of cluster 1 and cluster 2 are very similar in both clusters. Using the BAR eFP browser, the expression profile of these genes were visualized. They all show expression in the meristematic zone. Because of their central location in the clustering, both clusters are thought to contain meristimatic cells that give rise to the different cell types. Cluster 6 and cluster 10 were difficult to assign a cell type to. It could be possible that cluster 6 contains trichoblast precursor cells, as the AT1G67330 is highly expressed in these cells and has moderate expression in the actual trichoblasts. This corresponds with the found expression profile in this analysis (Figure 5). The cell type of cluster 10 could not be determined.

For Dataset 2, no cell type was yet assigned to clusters 1, 4, and 10. Clusters 1, 4 and 8 are highly correlated. However, cluster 1 showed broader expression patterns than cluster 4 and 8. Cluster 1 contains a lot of ribosomal genes, which is typical for meristematic cells. Cluster 4 is a bit more specialized and could be classified as an intermediate between the meristem and the cortex, the cortex precursor cells. By looking at the expression of the top differential genes of cluster 10 with the the BAR eFP browser, it can be seen that all these genes are being expressed in the elongation zone. Based on the clustering, it can be concluded that the cells in cluster 10 are in a transition state between the meristem and the endodermis.

The known marker genes and the information provided by the BAR eFP browser was combined, resulting in the cluster annotation depicted in Figure 6.

3.4 Comparison between both datasets

Both analyzed datasets contain the same plant material, namely Arabidopsis roots of 6-day old seedlings. Therefore, these datasets should be comparable. In Table 1 and Table 2, the number of cells and the percentage assigned to each cluster are depicted. The total number of cells in Dataset 1 is much larger than in Dataset 2. Therefore, the absolute numbers of cells were converted to percentages to allow comparisons between both datasets.

To make the proportion of cells per cell type more comparable between both datasets, some clusters were aggregated. For the root cap, clusters

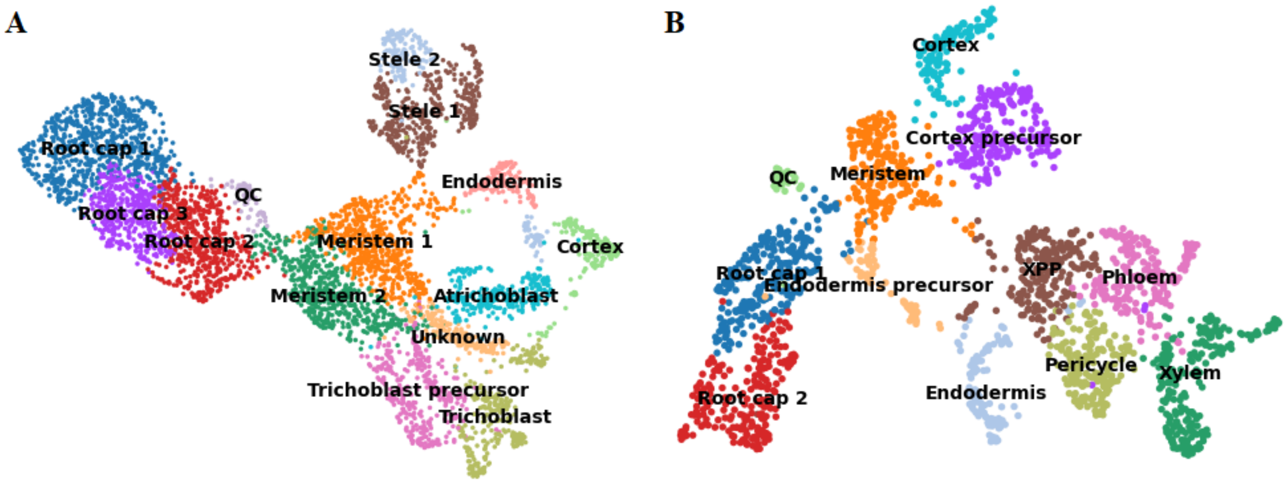

Fig. 6. Cluster annotation. The clustering was performed with the sc.tl.louvain function from the scanpy package with a resolution of 1.0. The clusters were annotated using known marker genes and expression profiles of top differentially expressed genes. (A) Annotated Louvain clustering of Dataset 1. (B) Annotated Louvain clustering of Dataset 2.

Cluster number	Cell type	Number of Cells	Percentage
0	Root cap 1	706	14.96
1	Meristem 1	592	12.54
2	Meristem 2	524	11.10
3	Root cap 2	523	11.08
4	Root cap 3	424	8.98
5	Stele 1	402	8.52
6	Trichoblast precursor	338	7.16
7	Trichoblast	261	5.53
8	Atrichoblast	255	5.40
9	Stele 2	173	3.67
10	Unknown	171	3.62
11	Cortex	157	3.33
12	Endodermis	128	2.71
13	QC	66	1.40
	Total	**4720**	**100**

Table 1. Number of cells and percentage of the total amount of cells per annotated cluster for Dataset 1.

Cluster number	Cell type	Number of Cells	Percentage
0	Root cap 1	246	12.47
1	Meristem	222	11.25
2	Xylem	218	11.05
3	Root cap 2	213	10.80
4	Cortex precursor	202	10.24
5	XPP	200	10.14
6	Phloem	181	9.17
7	Pericycle	166	8.41
8	Cortex	120	6.08
9	Endodermis	111	5.63
10	Endodermis precursor	62	3.14
11	QC	32	1.62
	Total	**1973**	**100**

Table 2. Number of cells and percentage of the total amount of cells per annotated cluster for Dataset 2.

Root cap 1, 2 and 3 of Dataset 1 were summed, which accounted for 35.02 % of all cells of this dataset. The sum of Root cap 1 and 2 of Dataset 2 resulted in a proportion of 23.26 % of all cells. For the stele, clusters Stele 1 and 2 of Dataset 1 were summed, which accounted for 12.18 % of all cells in this dataset. The sum of the clusters Xylem, XPP, Phloem and Pericycle of Dataset 2 resulted in a proportion of 38.77 % of all cells in this dataset. For the meristem, clusters Meristem 1 and 2 of Dataset 1 were summed, making a total proportion of 23.64 %.

4 Discussion

In this analysis, two single-cell datasets of 6-day old Arabidopsis roots were preprocessed, clustered and annotated.

After preprocessing, approximately the same amount of genes were present in both datasets. However, the number of cells differed tremendously. Dataset 1 contained more than twice as many cells as Dataset 2. The larger the amount of cells, the higher the chance to capture all cell types, including intermediate states. This allows researchers to discover new cell types and infer cell trajectories. The dimensions were reduced with PCA and the data was clustered with Louvain clustering. In Dataset 1, 14 clusters were identified. In Dataset 2, 12 clusters were found. Although both datasets contain the same plant material, a different number of clusters was found. This difference could be due to the differences in cell number, which makes that some intermediate cell types might be detected in Dataset 1 and not in Dataset 2.

The clusters were annotated by combining known marker genes, the correlation between clusters and the top differentially expressed genes for each cluster. Some cell types were found in both datasets, while others were specific for only one dataset. In both datasets, clusters containing root cap cells were found. In Dataset 1, 3 clusters were found whereby clusters 0 and 4 were much more similar to eachother than to cluster 3. In Dataset 2, 2 clusters with root cap cells were found. This indicates that there might be an overclustering in Dataset 1 and that there are in fact two distinct cell types present in the root cap. These cell types probably are collumella and the lateral root cap. However, it could not be determined which clusters contained each cell type. When taking all root cap clusters together for each dataset, 35.02% of all cells in Dataset 1 and 23.26% of Dataset 2 were marked as root cap. A larger proportion of cells leads often to more clusters, which explains why 3 root cap clusters were found in Dataset 1, compared to 2 in Dataset 2. Stele was also found in both datasets. In Dataset 1, different stele cell types could not be distinguished, while in Dataset 2, 4 different stele cell types were identified (XPP, Phloem, Pericycle and Xylem). The resolution of the stele cells was thus much higher in Dataset 2. When summing up all stele cells in both datasets, 12.18% of all cells in Dataset 1 and 38.77% of all cells in Dataset 2 were marked as stele. The larger proportion of stele cells in Dataset 2 resulted in a higher resolution and allowed the identification of specific stele cell types. The meristem was assigned to 2 clusters in Dataset 1 and 1 cluster in Dataset 2. All these clusters could not be identified by means of the known marker genes. The meristem is characterized by quick cell division. Therefore, mostly ribosomal proteins are found to be differentially expressed in these clusters, necessary to maintain fast growth and completely overshadowing cell type specific genes. The cortex and endodermis were in both datasets identified by known marker genes. Approximately, the same proportion of cells is present in these clusters, strengthening the annotation of these clusters. The smallest cluster in both datasets comprises the QC, which could also be identified by known marker genes. Morphologically seen, the QC is a small region in the root apex, which corresponds the small cluster sizes found in this analysis. Only in Dataset 1, clusters with trichoblasts and atrichoblasts could be identified by means of known marker genes. In Dataset 2, these cell types were not detected. Either these cells were incorporated in another cluster because they were not distinct enough or these cells were lost during data generation or data preprocessing steps. Lastly, some intermediate states were detected in both datasets. In Dataset 1, trichoblast precursor cells were found. In the UMAP representation, it can be seen that these cells form a smooth connection between the meristem and the true trichoblasts. In Dataset 2, the same pattern was observed for endodermis cells and cortex cells.

5 Conclusion and Future work

In this study, two single-cell datasets on Arabidopsis roots were preprocessed, clustered and annotated. The majority of the cell types found were present in both datasets. However, certain cell types like trichoblasts and atrichoblasts were only found in one dataset, creating an incomplete overview of cell type identities when only taking one dataset into account. This indicates the importance of combining multiple experiments to obtain a complete overview of all cell types and intermediate states present in Arabidopsis roots.

This research can form the basis of further analyses like cell trajectory inference or the identification of new marker genes, with as ultimate goal to better understand plant root development and use this information to improve crop productivity and yield.

References

Birnbaum, K. *et al.* (2003). A gene expression map of the <i>arabidopsis</i> root. *Science*, **302**, 1956–1960.

Denyer, T. *et al.* (2019). Spatiotemporal developmental trajectories in the arabidopsis root revealed using high-throughput single-cell rna sequencing. *Developmental Cell*, **48**, 840–852.e5.

Edgar, R. *et al.* (2002). Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Research*, **30**, 207–210.

Fitter, A. (2002). Characteristics and functions of root systems. *Plant Roots*, pages 15–32.

Kinsella, R. J. *et al.* (2011). Ensembl biomarts: a hub for data retrieval across taxonomic space. *Database*, **2011**, bar030–bar030.

Krämer, U. (2015). The natural history of model organisms: Planting molecular functions in an ecological context with *Arabidopsis thaliana*. *eLife*, **4**, e06100.

McInnes, L. *et al.* (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *ArXiv*.

Meyerowitz, E. M. (1987). *ARABIDOPSIS THALIANA*. *Annual Review of Genetics*, **21**, 93–111.

Schiefelbein, J. W. *et al.* (1997). Building a root: the control of patterning and morphogenesis during root development. *The Plant Cell*, **9**, 1089–1098.

Wendrich, J. R. *et al.* (2020). Vascular transcription factors guide plant epidermal responses to limiting phosphate conditions. *Science*, **370**.

Winter, D. *et al.* (2007). An electronic fluorescent pictograph browser for exploring and analyzing large-scale biological data sets. *PLoS ONE*, **2**, e718.

Wolf, F. A. *et al.* (2018). Scanpy: large-scale single-cell gene expression data analysis. *Genome Biology*, **19**, 15.