A Stroll Through the Developmental Landscape of Plants

Warre Dhondt[1], Michael Van de Voorde[2], and Prof. Dr. Ir. Steven Maere[3]

[1]Universiteit Gent

[2, 3]Evolutionary Systems Biology, VIB-Ugent Center for Plant Systems Biology

March 30, 2023

Contents

1 Introduction

Consider the following: Cells can be regarded as (meta)stable states of the biological state space, i.e. cells are fruitful combinations of biological parameters such as gene expression that influence the phenotype of said cell. Therefore, it would be likely that during its development a cell preferentially progresses through such stable states as well (most of the time). It is exactly this notion that C.H. Waddingtons epigenetic landscape tries to depict, the sequence of developmental changes that a cell undergoes to achieve its final, differentiated form [**Waddington2011**] (1A). Waddington referred to such robust developmental paths as canalizations, but nowadays they are more commonly referred to as developmental trajectories [**Waddington1957**]. This metaphorical epigenetic landscape has formed ideological basis for much of our understanding of cellular developmental biology, and hypotheses often boil down to inquiries into said trajectories: what are the biological principles underlying the developmental trajectory of a cell?

Attempts at understanding the developmental trajectory at the level of organs, tissues and more recently, single cells, have greatly benefited from advances in sequencing technologies and throughput [**Goodwin2016**]. Within the plant community, The first organ-wide mappings of gene expression patterns date from the early 2000s [**Birnbaum2003**, **Brady2007**]. Transcriptome-wide expression profiling at the level of single cells using RNA-Seq was introduced in 2009, but at its inception could only profile a handful of cells [**Tang2009**]. Key technological advancements have allowed for order-of-magnitude increases in throughput, with single-cell experiments currently routinely profiling thousands to hundreds of thousands of cells, making single-cell RNA-Seq a mainstay in modern biological research [**Svensson2018**]. These advancements have not escaped the field of plant developmental biology, with single-cell resolution gene expression *atlases* of the *Arabidopsis thaliana* root [**Denyer2019**, **Shahan2022**, **Wendrich2020**], shoot [**Zhang2021**], vascular tissue [**Otero2022**, **Kim2021**] and seed [**Picard2021**] being published in the last 4 years, to name a few.

Such single-cell expression[1] experiments offer a wealth of information on dynamic processes such as decision making during differentiation, and in some cases can be considered a direct readout of Waddingtons developmental landscape [**Griffiths2018**]. This has triggered a surge in computational methods, classified under the title *Trajectory Inference* (TI), that aim to infer the dynamic processes contained in the static snapshot that a single-cell experiment has to offer. Trajectory inference algorithms aim to reconstruct the developmental trajectory from a starting cell to a another cell by assuming that cells that lie on the path between the starting and terminal cells represent developmental intermediates, and then sometimes quite literally, connecting the dots [**Trapnell2014**, **Street2018**, **Bergen2020**, **Haghverdi2016**] (1B). TI methods differ in their assumptions, their robustness, the underlying algorithm or in the types of topology they can detect; the performance of most popular trajectory

[1] Other common single-cell 'omics modalities include chromatin availability, DNA methylation, genomics and even proteomics [**Vandereyken2023**]

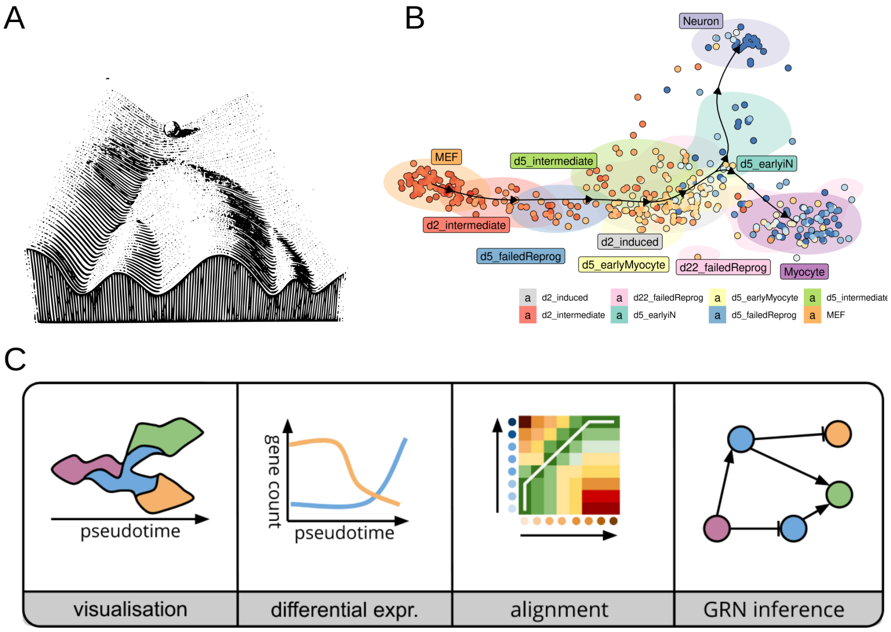

Figure 1: **An Overview of Trajectory Inference and Downstream Analyses** (A) Waddington's epigenetic landscape. [**Waddington1957**]. (B) Trajectory inference on the bifurcating developmental trajectory of mouse embryonic fibroblasts [**Dynversedyno**]. (C) Possible analysis downstream of trajectory inference include visualization, trajectory-based differential expression, alignment of trajectories and inference of gene regulatory networks.

inference methods has been thoroughly reviewed in [**Saelens2019**]. TI algorithms assigns cells along the trajectory a *pseudotime*, a unitless measure whose purpose is to give an indication of where along the trajectory the cell is located. Because of this pseudotime assignment, a lineage and its cells can be treated as a pseudo-time-series experiment one can identify temporal expression patterns and differential gene expression within and between lineages [**VandenBerge2020**], dynamic gene regulatory networks [**Nguyen2020**] or compare trajectories [**Alpert2018**] [**Deconinck2021**] (1C)

2 Aim

3 Materials and Methods

3.1 Preprocessing, Clustering, Dimensionality reduction

The processed single cell RNA-seq data for the wild-type Arabidopsis thaliana root atlas from [**Denyer2019**] were retrieved from the Gene Expression Omnibus database under accession GSE123818. All downstream processing was performed using the Seurat V4 framework [**Hao2021**]. Genes that are known to be differentially expressed upon protoplasting (Log2FC ¿ 2, q ¡ 0.05) were dismissed prior to further analysis [**Denyer2019**]. The counts were normalized using the variance stabilizing regularized negative binomial regression as implemented in SCTransform, using the percentage of reads mapping to mitochondrial and chloroplast genes as additional regressors [**Hafemeister2019**]. The dimensionality of the data was reduced by performing principal components analysis on the top 3000 most variable genes. Graph-based clustering was performed using the Louvain algorithm on the first 30 principal components. Two-dimensional embeddings were generated using the first 30 principal components as input for UMAP and t-SNE integrated in Seurat V4, diffusion pseudotime via the Destiny package ([**Angerer2015**]) and PaCMAP ([**pacmap**]).

3.2 Annotation of Clusters and Cell Types

For each cluster, markers were identified using Wilcoxon Rank Sum tests. The cluster markers were filtered for an average Log2FC ¿ 0.75 and to be expressed in $> 25\%$ of cells of that cluster. First, these markers were mapped to the cell-type specific markers from [**Shahan2022**] to infer the cell type associated with the clusters. Second, for every cluster, the top 30 differentially expressed markers ranked by logFC were mapped to tissue-specific cell type markers from [**Brady2007**]. These cluster annotations were reconciled with expression patterns of known developmental genes and manual inspection of the cluster-specific markers for the final cell-type annotation. Further, individual cells were annotated by their ploidy levels (2C, 4C, 8C, 16C) and developmental zone (meristem, elongation, maturation) by calculating the Pearson correlation coefficient with bulk RNA-seq reference expression profiles ([**Bhosale2018**, **Brady2007**]), using the annotation process kindly provided in the Github repository of [**Shahan2022**].

3.3 Trajectory inference

For trajectory analysis, Slingshot [**Street2018**] was used to build a minimum-spanning tree (MST) on the clusters, fixing trichoblast, atrichoblast and cortex as terminal clusters. Simultaneous principal curves were then fitted based on this MST to infer smooth trajectories, and cellular pseudotimes were determined based on these principal curves. Next, for every lineage, a negative binomial generalized additive model was fitted on the global top 3000 variable genes using TradeSeq [**VandenBerge2020**].

4 Results

4.1 Single-cell transcriptomics of the *Arabidopsis* root

The entire analysis used the dataset of a single-cell RNA-Seq (scRNA-Seq) experiment by [**Denyer2019**], who profiled 4727 *Arabidopsis thaliana* root cells. Plant cells have a rigid polysaccharide cell wall that can make it hard to isolate single cells. The principal method to make release single plant cells and make them amenable for scRNA-Seq is to release them from their cell wall by means of enzymatic digestion [**Seyfferth2021**]. To prevent systematic biases during the analysis, 6063 genes that showed significant differential expression in a bulk transcriptomic analysis of plant cell protoplasting [**Denyer2019**] were removed prior to further analysis. Initially, putative high quality cells were identified by filtering the raw library for cells with < 5 percent of genes mapping to mitochondrial and chloroplast genes, as well as ad hoc cutoffs for library size and number of counts. However, mapping the filtered cells to the final processed and annotated dataset via mutual nearest neighbors showed that cells that were removed during QC belonged predominantly to the trichoblast cell type, which would have led to the removal of approximately 30 percent of all trichoblasts (χ^2-test, $p = 2 \times 10^{-4}$). We therefore opted not to filter the cells based on the number of counts and/or features in order to prevent a systematic bias against one cell type during the analysis. The raw count data were normalized using a variance stabilizing regularized negative binomial regression [**Hafemeister2019**]. The resulting library contained information on the expression of 15313 features across 4727 cells. Cells that are similar in terms of gene expression were clustered into 19 groups via the Louvain algorithm on the shared nearest-neighbor graph. To assign cell types to these 19 clusters, the top differentially expressed genes per cluster were mapped to bulk RNA-Seq studies of isolated tissues of the Arabidopsis root ([**Wendrich2020**, **Brady2007**, **Seyfferth2021**]A). The annotated cells were embedded in two dimensions via UMAP for visualization (2B). Central clusters 3, 4, 10 and 16 all mapped to meristematic xylem tissue, but closer inspection of the genes used in this annotation showed that these were predominantly markers for cells with high mitotic/proliferative activity, without presence of markers relating to xylem development. For example, clusters 3, 4 and 16 showed high expression of genes related to DNA replication or nucleolar functionality, including those for histone family proteins H2A (AT1G51060, AT3G54560, AT4G27230) and H2B (AT1G07790), a subunit of the H/ACA complex (AT3G03920) that is involved in pseudouridinylation, and ribosomal proteins including RPL16A (AT2G42740) ([**Bernstein2004**, **Bernstein2007**]), as well as high expression of mitotic regulators *CYCB1;1* and *AUR1* (AT4G32830) in cluster 16, both of which are commonly used cell cycle markers ([**Schnittger2018**, **Weimer2016**]). Cluster 10 additionally showed expression of QC marker GLV6 ([**Fernandez2013**]) as well as GA3, which is involved in GA biosynthesis, a process that has been associated with QC identity ([**Nawy2005**]). Cluster 10 therefore likely consists of QC cells and its closely related initials. The UMAP shows a disconnected star-like structure consisting of a central mass of meristematic cells extending out to different developmental lineages, including trichoblast and atrichoblast lineages with their characteristic expression of *COBL9* (AT5G49270) and *GL2* (AT1G79840), respectively (2C). Additional identified lineages include the cortex-endodermis differentiation axis and that of the root cap, consisting of columella and lateral root cap cells. Many of the genes that are involved in the regulation of root cap development showed differential expression upon deprotoplasting and were therefore removed during the preprocessing of the data, which is reflected by the fact that the evidence for the identity of these cells is rather low compared to e.g. the cortex-endodermis lineage (**??**). Further, the resolution between the different vascular cell types is rather low. This lack of resolution is a known problem in plant single-cell transcriptomics and can be partly attributed to a high overlap in gene expression between vascular cells and its associated pericycle ([**Parizot2012**]), as well as issues with accessibility, causing to vascular cells to be underrepresented in single-cell atlases ([**Otero2022**]).

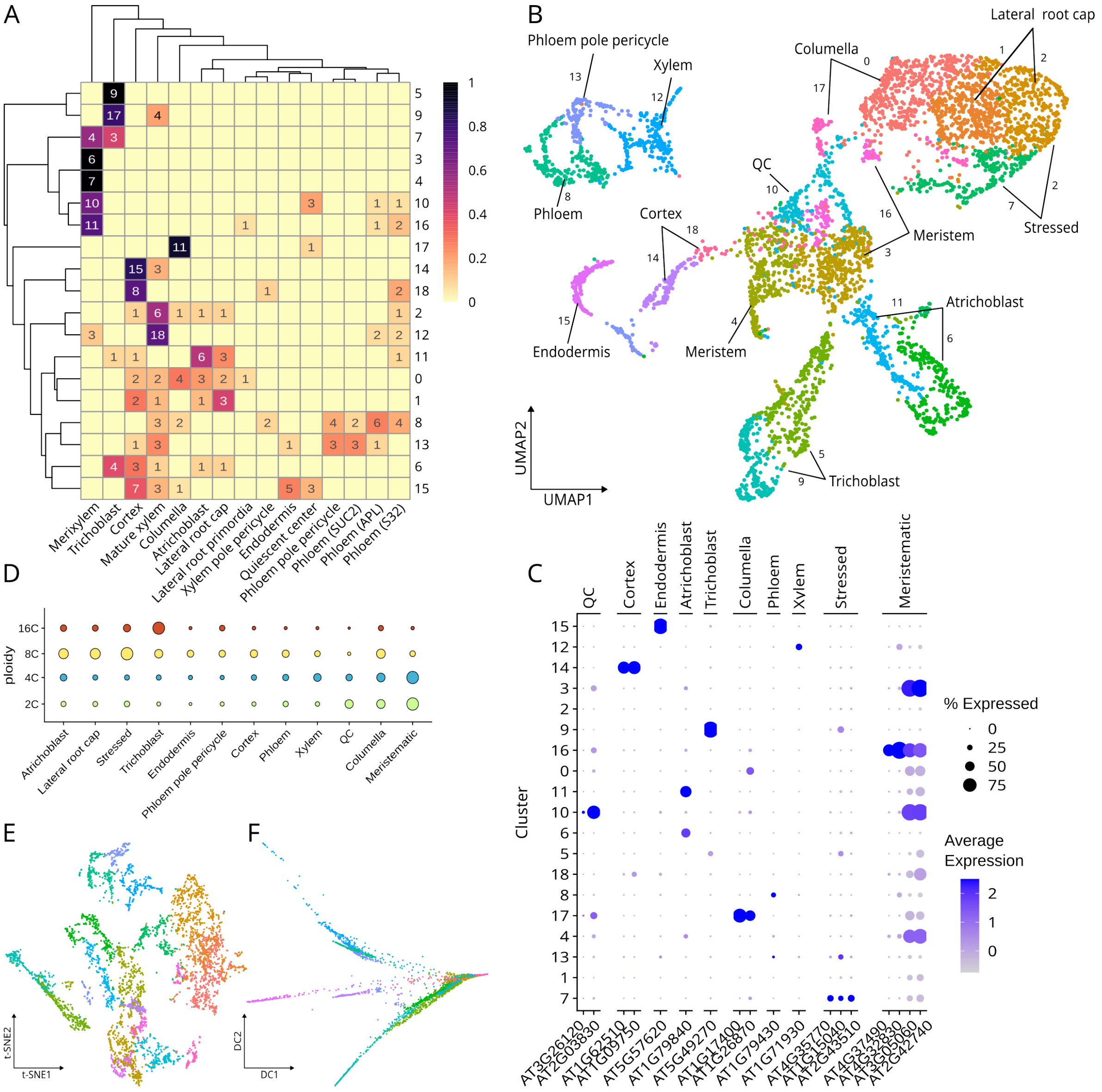

Figure 2: **Cell types of the *Arabidopsis* Root Identified by scRNA-Seq.**
(A) Heatmap showing the contribution of marker genes of different cell types to the cluster identity. The top 50 differentially expressed genes per cluster were mapped to marker genes of bulk RNA-Seq data of plant root tissue sections. The numbers in the matrix represent the number of DE genes of the top 50 of that cluster mapping to a certain tissue. The row colors are the normalized contribution of the cell types to the top DE genes of a given cluster. (B) UMAP of the 4727 *Arabidopsis* root cells. The (sub)clusters are colored and named according to their cell type annotation. Numbers besides the arrows indicate the cluster numbering as referred to in the text. (C) Dot plot showing expression of known cell type marker genes by the annotated cell types in the scRNA-Seq dataset. Dot sizes represent percentage expression in a cell type, color intensity reflects average normalized and scaled expression. (D) Dot plot showing the contribution of different ploidy levels to the cell types identified in the scRNA-Seq dataset. The ploidy level of each cell was identified by correlation with bulk RNA-seq reference expression profiles. (E) and (F): t-SNE and diffusion map, respectively, of the scRNA-Seq dataset. Cells are color-coded according to cell type annotation.

5 Discussion