A Stroll Through the Developmental Landscape of Plants

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

[1]Universiteit Gent

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

April 5, 2023

Contents

1 Introduction

During its development, a stem cell becomes gradually more restricted in form and function, at last becoming one of a multitude of possible cell types that have been optimized during evolution to function as a part of a larger organism. It is this notion that C.H. Waddington tried to depict with his epigenetic landscape: that the sequence of developmental changes that a cell undergoes to achieve its final form proceeds along paths delineated by internal and external cues [**Waddington2011**] (figure 1A). Waddington referred to such robust developmental paths as canalizations, but nowadays they are more commonly referred to as developmental trajectories [**Waddington1957**]. This development is steered by molecular cues that depend on the cell's environment, among others. They determine the system's topology by creating additional possibilities via bifurcations, for example, or by taking away existing paths [**Ferrell2012**]. This metaphorical epigenetic landscape has formed the 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 that steer 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, it 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, where single-cell resolution gene expression *atlases* of the *Arabidopsis thaliana* root [**Denyer2019**, **Shahan2022**, **Wendrich2020**], shoot [**Zhang2021**], vascular tissue [**Otero2022**, **Kim2021**] and seed [**Picard2021**] have been 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 Waddington's developmental landscape [**Griffiths2018**, **Trapnell2015**]. 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**,

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

Street2018, **Bergen2020**, **Haghverdi2016**] (figure 1B). TI methods differ in their assumptions, their robustness, the underlying algorithm, or the topologies they can detect; the performance of most popular trajectory inference methods has been thoroughly reviewed in [**Saelens2019**]. TI algorithms assign 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, and one can identify temporal expression patterns or differential gene expression within and between lineages [**VandenBerge2020**], dynamic gene regulatory networks [**Nguyen2020**], or compare trajectories [**Alpert2018**, **Deconinck2021**] (figure 1C).

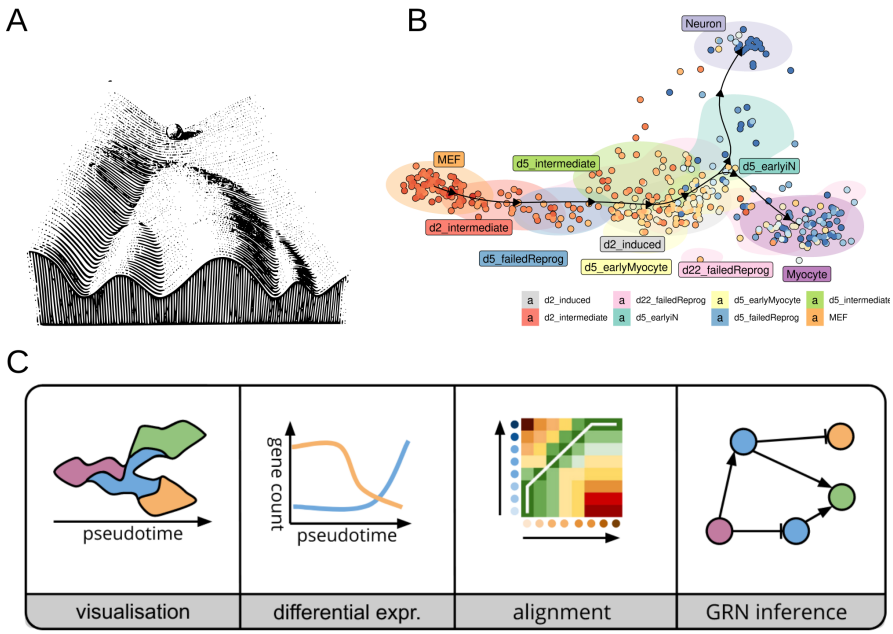

Figure 1: **An Overview of Trajectory Inference and Downstream Analyses** (A) Waddington's epigenetic landscape. [**Waddington1957**]. (B) Trajectory inference on the bifurcating developmental trajectory of mouse embryonic fibroblasts [**Dynversedyno**]. (C) Possible analyses downstream of trajectory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

Despite trajectory inference being developed by the single-cell genomics field, TI's conceptual framework can be applied to other applications where one wishes to study the temporal dynamics of an observable based on high-dimensional data. For example, the Maere lab (Center for Plant Systems Biology, VIB-UGent) aims to use TI methods on single plants to try and decipher developmental processes at the level of plant field trials. The plant root tip is a model system where TI methods have been extensively used in recent years [**Denyer2019**, **Ryu2019**, **Shulse2019**, **Wang2020**, **Zhang2019**, **Zhang2021**]. More specifically, the pathways underlying the differentiation of trichoblasts and atrichoblasts in the Arabidopsis epidermal layer have been characterized in quite some detail. Epidermal cell patterning is the result of a complex interplay between cell-cell communication involving cortical and epidermal cells, the exchange of mobile transcription factors, and lateral

inhibition [**Schiefelbein2014**, **Balcerowicz2015**, **SalazarHenao2016**, **Bruex2012**]. This research project illustrates the use cases of trajectory inference on a publicly available single-cell RNA-Seq dataset of the Arabidopsis root tip [**Denyer2019**]. We show how trajectory inference can be used to asses pseudo-temporal dynamics of gene expression of known regulators of the atrichoblast-trichoblast differentiation axis, and how the information gained can be used to elucidate how these processes are intertwined with one another and together regulate epidermal cell patterning.

2 Aim

Recent advances in sequencing technologies have established single-cell RNA-Seq as a mainstay in modern biological research. This is illustrated by the surge of published single-cell expression atlases, including multiple of the *Arabidopsis thaliana* root. The first aim of this project is to familiarize oneself with current community standards for processing single-cell RNA-Seq data (see [**Luecken2019**, **Amezquita2019**]), exploring different options along the way to find out their advantages and disadvantages. This includes, but is not limited to, exploratory data analysis and quality control, clustering methods such as Louvain community detection, dimensionality reduction (including tSNE and UMAP) and cell type annotation. The dataset of choice is a publicly available single-cell RNA-Seq dataset of the *Arabidopsis* root from [**Denyer**], who mainly focused on analyzing dyamic gene expression within differentiating trichoblasts, a type of epidermal cell. The analysis presented in the original publication only scratched the surface of epidermal cell patterning, which is the result of the coordination between multiple developmental processes such as cell-cell communication, lateral inhibition, and feedback loops. The second aim of this project is therefore to try and reproduce the findings of [**Denyer2019**] (to a certain extent) and to explore epidermal cell patterning using computational techniques that leverage the dynamic nature of the data. This reanalysis starts off with trajectory inference to identify developmental lineages using Slingshot [**Street2018**]. Once they have been identified, the reconstructed trajectories can be treated as a pseudo-time-series experiment, that allows us to assess biological patterns in a temporal manner to gain insight into epidermal cell differentiation. TradeSeq [**VandenBerge2020**] offers a statistical framework to compare pseudo-temporal patterns of gene expression both within a single trajectory, or between developmental trajectories. Within the confines of this project, TradeSeq can be used to analyze the dynamic expression of known regulators of epidermal cell patterning and compare how the expression patterns of these regulators may vary between developing cell types. Gene ontology enrichment can be an asset here to extract biological meaning from the (likely) numerous number of detected genes in some of the analyses [**Ashburner2000**]. If time permits, dynamical regulatory network inference could be used as a more rigorous means of detecting regulatory modules whose activity changes along the developmental axes. Further, computational approaches that try to determine cell-cell communication based on single-cell expression data [**Efremova2020**, **Xu2022**] can be used to add interaction between cell types as an additional piece in the biological puzzle.

3 Materials and Methods

3.1 Preprocessing, Clustering, Dimensionality reduction

The gene by cell expression matrix with raw scRNA-seq read counts 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**]. 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 one cell type, and this led to removing 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. 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 3000 most variable genes were identified as those having the highest residual variance in the regularized NB regression model. 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 (Seurat FindClusters, resolution = 1). 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 (cluster 9), atrichoblast (cluster 6) and cortex (cluster 14) as terminal clusters and cluster 10 as starting cluster. Simultaneous principal curves were then fitted based on this MST to infer smooth trajectories, and pseudotimes were assigned to each cell 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 analysis began with processing the scRNA-Seq dataset of the *Arabidopsis thaliana* root [**Denyer2019**]. Cells that are similar in terms of gene expression were clustered into 19 groups, and these clusters were annotated by mapping cluster-specific gene expression onto marker genes of isolated *Arabdopsis* root tissues (figure 2). 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 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), which is involved in pseudouridinylation, and ribosomal proteins including RPL16A (AT2G42740) ([**Bernstein2004**, **Bernstein2007**]). Further, mitotic regulators *CYCB1;1* and *AUR1* (AT4G32830) were highly expressed 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 as well as proliferating cells. The UMAP shows a disconnected star-like structure consisting of the central meristematic clusters extending out to different developmental lineages. These lineages include the cortex-endodermis differentiation axis, that of the root cap, consisting of columella and lateral root cap cells, as well as epidermal and vascular cell types (figure 2). 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 agreement on identity of these cells is rather low compared to e.g. the cortical and endodermal cells (figure 2). Interestingly, cells that are located terminally in the root-cap lineage showed high expression of genes involved in peroxisome biogenesis (PEROXIN family members) and glucosinolate metabolism (cytochrome P450 monooxygenases) and biogenesis of ER bodies (NAI1), which are organelles that are involved in stress responses and immunity [**Sarkar2020**, **Su2019**]. Other cluster-specific marker genes are in abiotic stress, such as GLUTAREDOXIN C1 (AT5G63030) and GLUTATHIONE PEROXIDASE 2 (AT2G31570), GUN2 (involved in stress-induced chloroplast dysfunction) [**Crawford2017**] and WRKY26 [**Li2011**]. Such stress-specific gene expression could play into the protective role of the root cap against biotic and abiotic stresses [**Kumar2020**], or, alternatively, this stress-response could be the result of programmed cell death during root cap cell sloughing [**Kumpf2015**]. Further, the resolution between the different vascular cell types is comparatively low. This lack of resolution is a known problem in plant single-cell transcriptomics and has been 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**]).

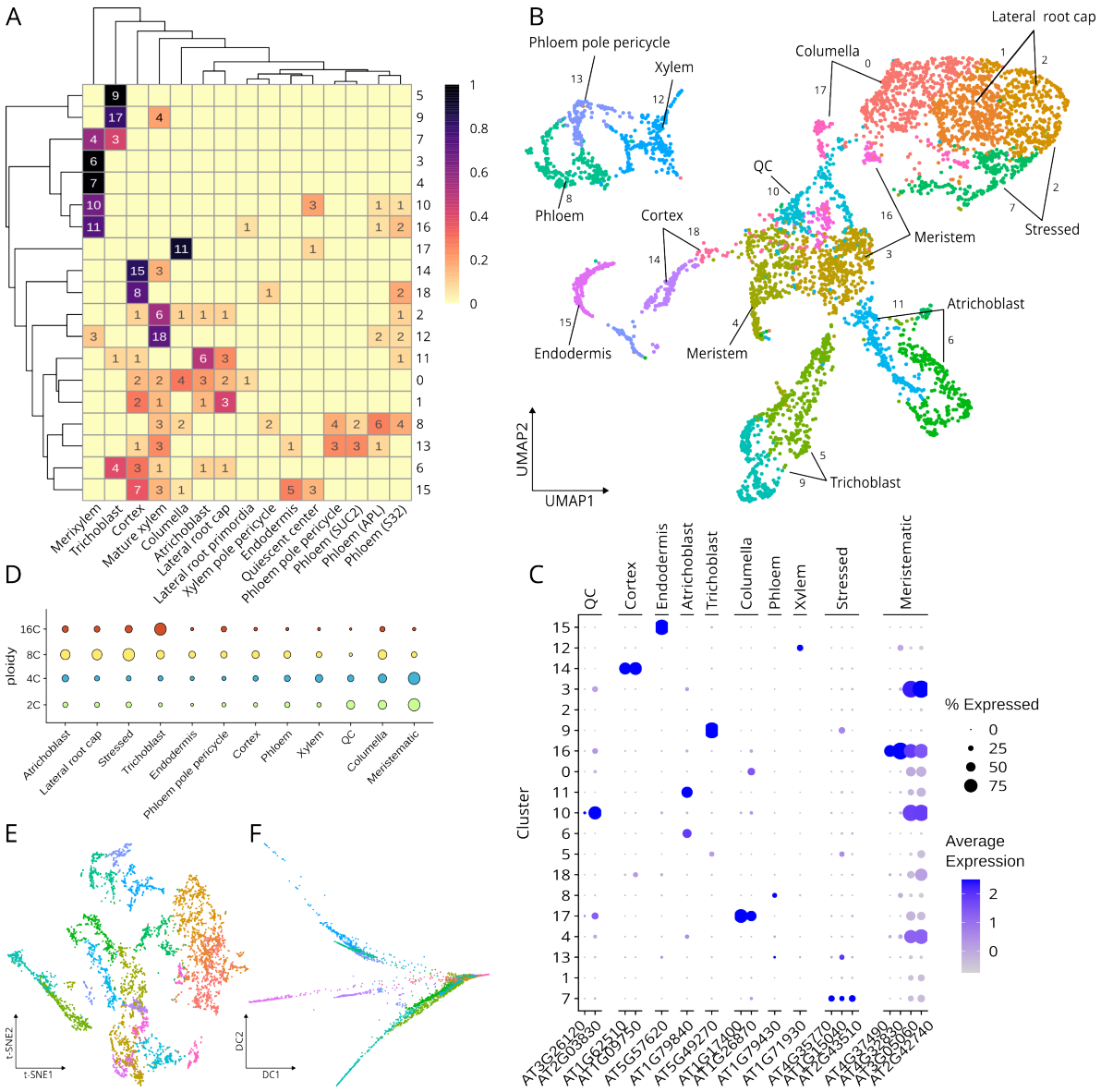

Figure 2: **Cell types of the *Arabidopsis* Root Identified by scRNA-Seq.**
(A) Heatmap showing the contribution of marker genes of different cell types to the cluster identity. The top 50 differentially expressed genes per cluster were mapped to marker genes of bulk RNA-Seq data of plant root tissue 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.