A Stroll Through the Developmental Landscape of Plants: Trajectory Inference on Single-Cell Transcriptomics Data

Warre Dhondt

Supervisor: Michael Van de Voorde[1,2]

[1]Department of Plant Biotechnology and Bioinformatics, Ghent University

[2]VIB Center for Plant Systems Biology

April 27, 2023

Contents

Acronyms

BH	Benjamini-Hochberg
DR	Dimensionality Reduction
LR	Ligand-receptor
PCA	Principal Components Analysis
scRNA-Seq	single-cell RNA-sequencing; single-cell transcriptomics
t-SNE	t-distributed stochastic neighbor embedding
TI	Trajectory Inference
UMAP	Uniform Manifold Approximation and Projection

1 Introduction

During its development, a stem cell becomes gradually more restricted in form and function, at long last becoming one of a multitude of possible cell types that have been shaped 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: the sequence of developmental changes that a cell undergoes to achieve its final form proceeds along paths delineated by internal and external cues (figure 1a); (Waddington, 2011). Waddington referred to such robust developmental paths as canalizations, but nowadays they are more commonly referred to as developmental trajectories (Waddington, 1957). 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 paths via bifurcations, for example, or by taking away existing paths (figure 1b); (Ferrell, 2012). This metaphorical epigenetic landscape has formed the epistemological 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 (Goodwin et al., 2016). Within the plant community, the first organ-wide mappings of gene expression patterns date back to the early 2000s (Birnbaum et al., 2003; Brady et al., 2007). 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 (Tang et al., 2009). 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 (scRNA-Seq) a mainstay in modern biological research (Svensson et al., 2018). These advancements did not go unnoticed in plant developmental biology, where single-cell resolution gene expression *atlases* of the *Arabidopsis thaliana* root (Denyer et al., 2019; Shahan et al., 2022; Wendrich et al., 2020), shoot (Zhang et al., 2021), vascular tissue (Otero et al., 2022; Kim et al., 2021) and seed (Picard et al., 2021) 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 (Griffiths et al., 2018; Trapnell, 2015). 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 rep-

[1] Other common single-cell 'omics modalities include chromatin availability, DNA methylation, genomics and even proteomics (Vandereyken et al., 2023)

resent developmental intermediates, and then, sometimes quite literally, connecting the dots (figure 1c); (Trapnell et al., 2014; Street et al., 2018; Bergen et al., 2020; Haghverdi et al., 2016). Trajectory inference methods differ in their assumptions, their robustness, the underlying algorithm, or the topologies they can detect; the performance of most popular TI methods has been thoroughly reviewed and benchmarked in Saelens et al. (2019). Trajectory inference 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 (pseudo-)temporal expression patterns or differential gene expression within and between lineages (Van den Berge et al., 2020), dynamic gene regulatory networks (Nguyen et al., 2020), or compare trajectories (figure 1d); (Alpert et al., 2018; Deconinck et al., 2021). 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 (VIB Center for Plant Systems Biology) aims to use TI methods on single plants to try and decipher developmental processes at the level of plant field trials.

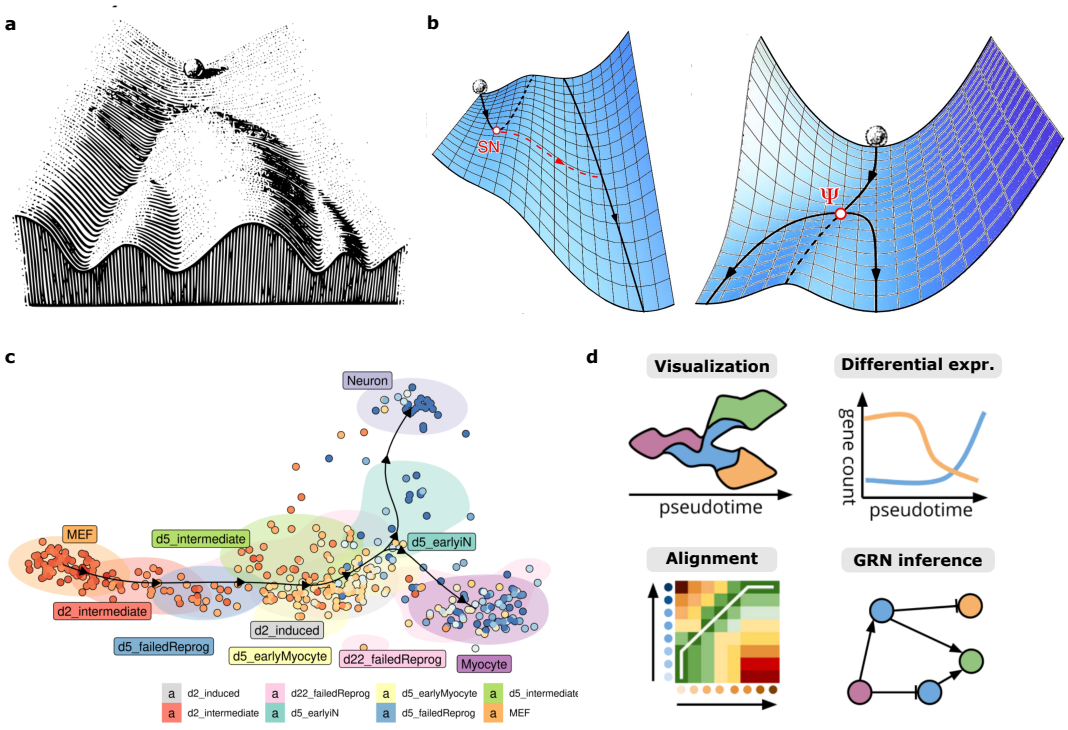

Figure 1 | An overview of trajectory inference and possible downstream analyses. (a) Waddington's epigenetic landscape (Waddington, 1957). (b) The topology of the developmental landscape can change because of developmental cues. Cell-fate commitment by an inductive stimulus occurs via a saddle node (SN) bifurcation (left). Lateral inhibition creates a pitchfork bifurcation (Ψ, right); (Ferrell, 2012). (c) Trajectory inference on the bifurcating developmental trajectory of mouse embryonic fibroblasts (Saelens et al., 2019). (d) Possible analyses downstream of trajectory inference include visualization, trajectory-based differential expression, alignment of trajectories and inference of gene regulatory networks (Deconinck et al., 2021).

The plant root tip is one model system where TI methods have been extensively used in recent years (Denyer et al., 2019; Ryu et al., 2019; Shulse et al., 2019; Wang et al., 2020; Zhang et al., 2019; Zhang et al., 2021). Stem cells from the root apical meristem are committed to a cell type according to their position along the radial axis, giving the root a concentric organization of it's tissues (figure 2a). Meristematic cells from the root apical meristem divide continuously to provide new cells to developing root tissue. As the root elongates, the developing cells stray away further from the stem cell niche and start to lose their proliferative status, while they increase in length and gain their cellular identity in the elongation and maturation zones (figure 2b); (Petricka et al., 2012). The defined radial pattern of cell specification and the longitudinal differentiation axis have been studied thoroughly and make the root tip a good model system for computationally inferring developmental lineages using TI (Seyfferth et al., 2021). The pathways underlying the differentiation of hair cells (H-cell or trichoblast) and non-hair cells (N-cell or atrichoblast) in the *Arabidopsis* epidermis 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 (Schiefelbein et al., 2014; Balcerowicz et al., 2015; Salazar-Henao et al., 2016; Bruex et al., 2012). This research project illustrates the use cases of trajectory inference on a scRNA-Seq dataset of the *Arabidopsis* root tip from Denyer et al. (2019). 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 from TI can be applied to elucidate the regulation of epidermal cell patterning.

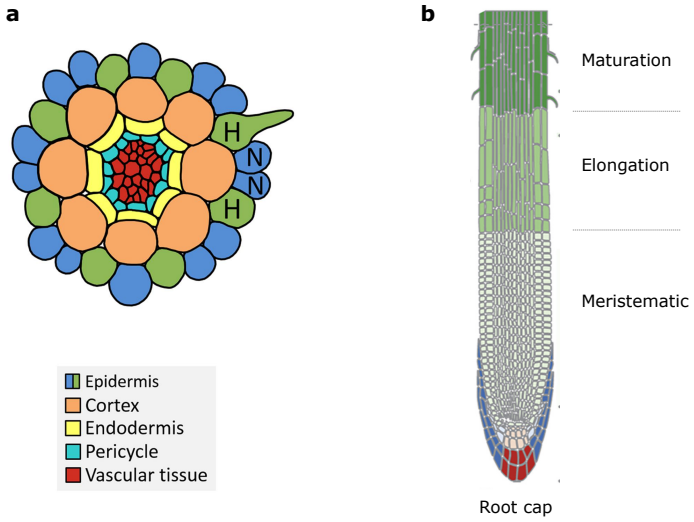

Figure 2 | Cellular organization of the *Arabidopsis* root. (a) Schematic representation of the cellular organization on a transverse section through the *Arabidopsis* root depicting the position of root hair (H) and non-root hair (N) cells (Balcerowicz et al., 2015). (b) Schematic representation of the developmental zones along the longitudinal axis of the root (Shahan et al., 2022).

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 in 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 scRNA-Seq data (see Luecken and Theis, 2019; Amezquita et al., 2019), 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 t-distributed stochastic neighbor embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP)) and cell type annotation. The dataset of choice is a publicly available scRNA-Seq dataset of the *Arabidopsis* root from Denyer et al. (2019), who mainly focused on analyzing dynamic gene expression within differentiating trichoblasts, a type of epidermal cell. They showed that trichoblast development is associated with progressive waves of gene expression that are functionally enriched towards the developmental stage, and reconstructed a gene regulatory network that captured regulatory events across pseudotime. 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 Denyer et al. (2019) 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 (Street et al., 2018). Once they have been identified, the reconstructed trajectories can be treated as a pseudo-time-series experiment, which allows us to assess biological patterns in a temporal manner to gain insight into epidermal cell differentiation. TradeSeq offers a statistical framework to compare pseudo-temporal patterns of gene expression both within a single trajectory, or between developmental trajectories (Van den Berge et al., 2020). 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) large number of differentially expressed genes in some of the analyses (Ashburner et al., 2000). 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 (E. Y. Su et al., 2022). Further, computational approaches that try to determine cell-cell communication based on single-cell expression data can be used to add interaction between cell types as an additional piece in the biological puzzle (Efremova et al., 2020; Xu et al., 2022).

3 Overview of Techniques

Exploratory data analysis and preprocessing in Seurat

- Quality control of scRNA-Seq libraries
- Normalization methods (log-normalization, variance-stabilizing normalizations)
- Louvain clustering
- Dimensionality reduction (PCA, t-SNE, UMAP, Diffusion maps, PaCMAP)
- Cluster and cell type annotation (cell-type markers, correlation-based)

Trajectory inference and differential expression

- Reconstructing trajectories using Slingshot
- Pseudotime differential expression analysis using tradeSeq

Additional analyses

- GO functional enrichment analysis using GOseq
- Inference of ligand-receptor interactions using PlantPhoneDB

General tools

- Various visualization methods (ComplexHeatmap package, ggplot)
- Coding in R
- Using a computer cluster
- Version control (Git)

4 Methodological Principle: Neighborhood-based Methods for Dimensionality Reduction

Dimensionality reduction is one of the essential tools in the bioinformatician's toolbox for analyzing data of high-throughput experiments. The high number of dimensions of this data[2] requires techniques that can pick out the dimensions that are most informative, which alleviates some computational burdens, as well as specialized visualization techniques. Reducing the dimensionality of scRNA-Seq data generally consists of two stages: feature selection and feature extraction. In *feature selection* we select a subset of features $p < d$ out of the d observed features, such that we map a cell $x \in \mathbb{R}^d$ to $y \in \mathbb{R}^{p<<d}$ while retaining as much information as possible that is present in the ambient space $\mathbb{R}^d$ (Ghojogh et al., 2019). These informative features can, for example, be defined as those that on average show higher variability than is expected based on their mean expression (Hao et al., 2021). The top n most variable features are then selected to perform the dimensionality reduction/manifold learning on, already reducing the dimensionality up to tenfold. Other approaches for feature selection involve identification of features that provide redundant information via dependency measures such as correlation or mutual information (Ghojogh et al., 2019). In *feature extraction*, we transform the selected features into a lower dimensional vector of latent features, and it is this procedure of feature extraction that is known as bona fide dimensionality reduction. Such dimensionality reduction techniques can be divided in to linear and non-linear methods. The next section will use principal components analysis (PCA) to illustrate linear dimensionality reduction methods and their uses, before we move on to non-linear methods such as neighborhood graph-based techniques, which are often more useful for scRNA-Seq data (Xiang et al., 2021).

4.1 Linear dimensionality reduction

Principal components analysis is a linear transformation that transforms the data to a new coordinate system such that the variance along every orthogonal principal component (PC) is maximized. (Hastie et al., 2009). More formally, each PC is a vector containing a certain weight[3] for every feature that reflects that feature's contribution to the variation along that principal component. PCA generally speaking does not resolve scRNA-Seq data well in two-dimensional embeddings, as it is a linear technique and therefore best captures variation along a single direction, whereas the underlying manifold is not necessarily linear (Xiang et al., 2021). The non-linear manifold learning methods are, however, often preceded by PCA to further reduce the dimensionality down from *variable feature space* to some reasonably low dimensional space consisting of e.g. the top 50 principal components. The idea here is that our feature $\times$ observation matrix can be approximated by some lower rank matrix such that minimal information is lost and that the structure of the data is preserved (Eckart and Young, 1936). This step aids in further removing redundant dimensions and/or noise present in the data and eases some computational burdens associated with the methods described next.

[2] E.g. in this project the dataset contained information on approx. 4700 cells in 15 000 dimensions (genes)
[3] Also called loading

4.2 Neighborhood Graphs for Non-Linear Dimensionality Reduction

Non-linear dimensionality reduction methods try to create lower dimensional embeddings that recapitulate the structure of the data in ambient high-dimensional space. Most methods of this sort share the same approach: they generate a weighted, high-dimensional graph based on the distance between a point and its neighbors, then do the same for the lower dimensional representation and try to minimize the error by optimizing some objective function that describes how well they approximate the high-dimensional structure (McInnes et al., 2018). The next sections describe two of the most commonly used non-linear methods for dimensionality reduction, how they optimize the projection and what the effect of this is on the structure of the resulting embedding.

4.2.1 t-Distributed Stochastic Neighbor Embedding (t-SNE)

Stochastic Neighbor Embedding (SNE) and its successor, t-distributed SNE (t-SNE), optimize the embedding of high-dimensional data points in lower dimensional space in a way that maintains the neighborhoods of the cells (Van der Maaten and Hinton, 2008; Hinton and Roweis, 2002). Put more simply, SNE makes sure that points that are close in the original data remain close after projection. This local neighborhood is described by the probability of picking j as the neighbor of i with a dissimilarity (distance) d_{ij} according to a Gaussian centered on i: $p_{j|i} = exp(-d_{ij}^2)/\sum_{i \neq k} exp(-d_{ik}^2)$ (figure3). In symmetric SNE, these conditional probabilities are converted to symmetric probablities as $p_{ij} = \frac{p_{j|i}+p_{i|j}}{2n}$. The following step defines the same neighborhood but for the lower dimensional embedding of the datapoints, and at this point SNE and t-SNE part ways. SNE defines the neighborhoods of the embedding just like before by computing a Gaussian neighborhood probability distribution for every datapoint (figure 3); (Hinton and Roweis, 2002). The objective function is then to minimize the difference between the neighborhoods in ambient space and the neighborhoods in lower dimensional space over all observations, which SNE does by minimizing the Kullback-Leibler divergence via gradient descent. However, this results in crowding, i.e. placing cells close together whereas they were dissimilar in ambient space. This phenomenon is rooted in the fact that in lower number of dimensions there is inherently less space to fit those neighbors, i.e. neighbors that are equidistant in a higher number of dimensions will be pushed in a lower number of dimensions (Cook et al., 2007). Van der Maaten and Hinton (2008) provided a solution for this by using a t-distribution for the lower dimensional probability function, effectively a heavy tailed Gaussian, which makes the objective function less stringent and in doing so creates the space that is lost by reducing the number of dimensions. Heavy-Tailed Symmetric Stochastic Neighbor Embedding (HSSNE; Yang et al., 2009 generalized this idea for using other heavy tailed probability distribution functions. UNI-SNE provided a solution for the crowding problem by using a background distribution in the lower dimension to add a small repulsion during gradient descent optimization (Cook et al., 2007). One thing we have not discussed is the size of the local neighborhood that is considered. Because the data may have varying densities at different points in ambient space, the neighborhood is evaluated on a case-to-case basis by adjusting the normalization factor of the Gaussian (σ) such that the probability distribution achieves a specified

perplexity (Van der Maaten and Hinton, 2008). This perplexity is an information theoretic concept (exponentiated Shannon entropy) that reflects the uncertainty or information content of a probability distribution, which in practice comes down to the effective size (number of neighbors) contained in the local neighborhood (Kobak and Berens, 2019). By changing this perplexity one can steer (t-)SNE to focus more on local structure (small neighborhoods) versus global structure (large neighborhoods) (Kobak and Berens, 2019).

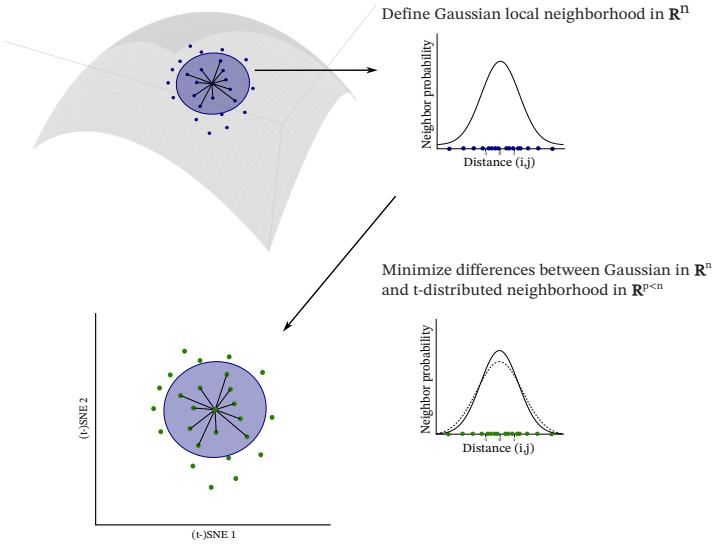

Figure 3 | Dimensionality reduction using t-distributed stochastic neighbor embedding (t-SNE). For each point in ambient space, its local neighborhood is defined according to a Gaussian based on pairwise distances. The same neighborhood is defined for data points in the embedding but using a t-distribution, and the difference between these distribution is minimized using gradient descent. For details see text.

4.2.2 Uniform Manifold Approximation and Projection (UMAP)

The general idea of UMAP is similar to that of (t-)SNE: it learns a (graph-based) representation of the data in ambient space, and then optimizes the embedding in a lower number of dimensions (figure 4). The novelty in UMAP lies in how it learns the structure of the data in ambient space, borrowing its theoretical foundations from topological data analysis. McInnes et al. (2018) identified the following problem in manifold learning: the assumption that data points are uniformly sampled from the underlying manifold is not always met. What they proposed, was to construct a local metric for each datapoint such that the neighbors of that datapoint *are* uniformly distributed, ensuring that the assumption of a uniform distribution is met. UMAP does this by normalizing the distance from a point to its neighbors with respect to the distance to some k-th neighbor. Note how this is similar to how each cell's neighborhood is defined in t-SNE. McInnes went on to use category to prove that the local metric space (one for each point) can be mapped to a graph-like topological representation of the data (effectively a weighted graph) without losing any information. Once we have this topological representation of the data in ambient space, such a representation is defined on the lower dimensional embedding as well. The error between the two is then minimized by reorganizing the lower dimensional

graph using a force directed graph layout algorithm, where the forces are derived from the gradients that optimize the cross-entropy between the two graphs. Again, one can draw parallels between the UMAP optimization procedure and that of t-SNE. The former assigns an attractive force to edges based on the weight of the high-dimensional graph, whereas the latter does so based on the Gaussian neighborhood function in ambient space. UMAP has two parameters that affect how the original data is approximated in the embedding procedure, i.e. the number of neighbors and minimum separation. The number of neighbors hyperparameter determines how large the local metric space is in the **u**niform **m**anifold **a**pproximation part of **UMA**P, and affects how many neighbors are directly considered in the vicinity of each datapoint. The effect of this hyperparameter is similar to that of the k-neighbors parameter in t-SNE: considering a higher number of neighbors shifts the embedding to a more global representation (figure 4b). The minimum separation hyperparameter determines how closely the data points can be packed together in the embedding, which is a trade-off between faithfully representing the data and interpretability.

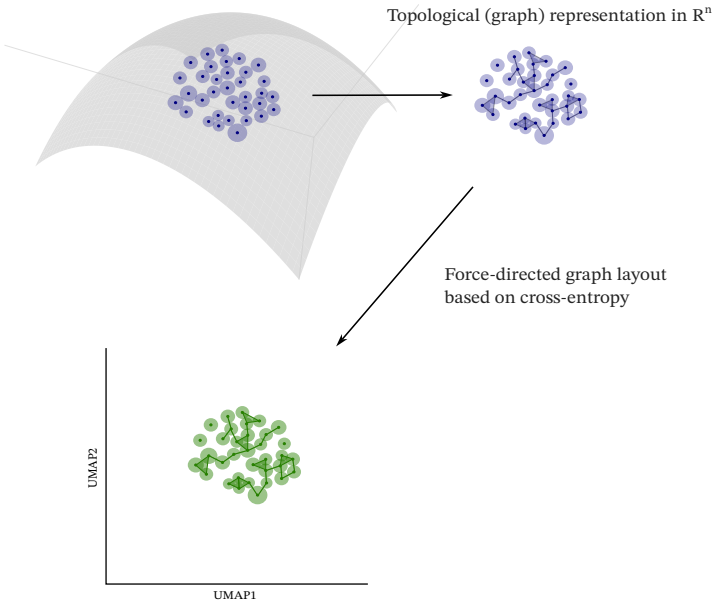

Figure 4 | Dimensionality reduction using Uniform Manifold Approximation and Projection (UMAP). UMAP generates a topological (graph-like) representation of the data in ambient space, and optimizes the embedding such that the topological representation of the embedding resembles the original using a force-directed graph layout algorithm. For details see text.

4.2.3 Evualuating dimensionality reduction tools

The principal use case of dimensionality reduction tools is to provide visualizations of high-dimensional biological data that are amenable for human interpretation. The increasing use of high throughput assays that generate such high-dimensional data has created a pressing demand for tools that generate interpretable yet accurate visualizations, and the supply of different methods has increased accordingly. (t-)SNE was one of the first methods that provided an effective solution and remains one of the most used tools to date. UMAP was introduced as an alternative to t-SNE, claiming that is was better at preserving global structure and computationally more efficient, and has dominated the field from

2018 onward (McInnes et al., 2018). However, benchmarking has showed that t-SNE's runtime nears that of UMAP (figure 5a); (Becht et al., 2018; Xiang et al., 2021). The computational performance of UMAP is the best all around, but UMAP is surpassed by FFT-accelerated Interpolation-based t-SNE (FIt-SNE) for larger input sizes. The other aspect of evaluating dimensionality reduction methods involves quantifying how well the tools preserves the structure that is present in the ambient data. This usually involves a evaluating a combination of metrics such as the Pearson correlation between all pairwise distances (global structure) and conservation of k-nearest neighbors (local structure) between the datapoints in ambient space and the embedding (Chari and Pachter, 2021). Using this approach, Becht et al. (2018) proposed that UMAP is better at preserving global structure than t-SNE, but this was countered by Kobak and Linderman (2021), who showed that both methods perform similarly given the optimal initialization (figure 5c). The problem with defining such evaluation metrics is that they are inherently biased towards methods that use these properties in their loss function, which prevents reliable and large-scale benchmarking of dimensionality reduction tools. Further, tuning hyperparameters of DR algorithms allows the researcher to change the aesthetics of the visualization at will, which reinforces the qualitative nature of such low-dimensional embeddings (figure 5b); (Chari and Pachter, 2021). To counteract the misuse of these hyperparameters to achieve visualizations that match a biased idea of what the data should look like, one could generate unbiased visualization by optimize the 'structure preservation' metrics as a function of the hyperparameters.

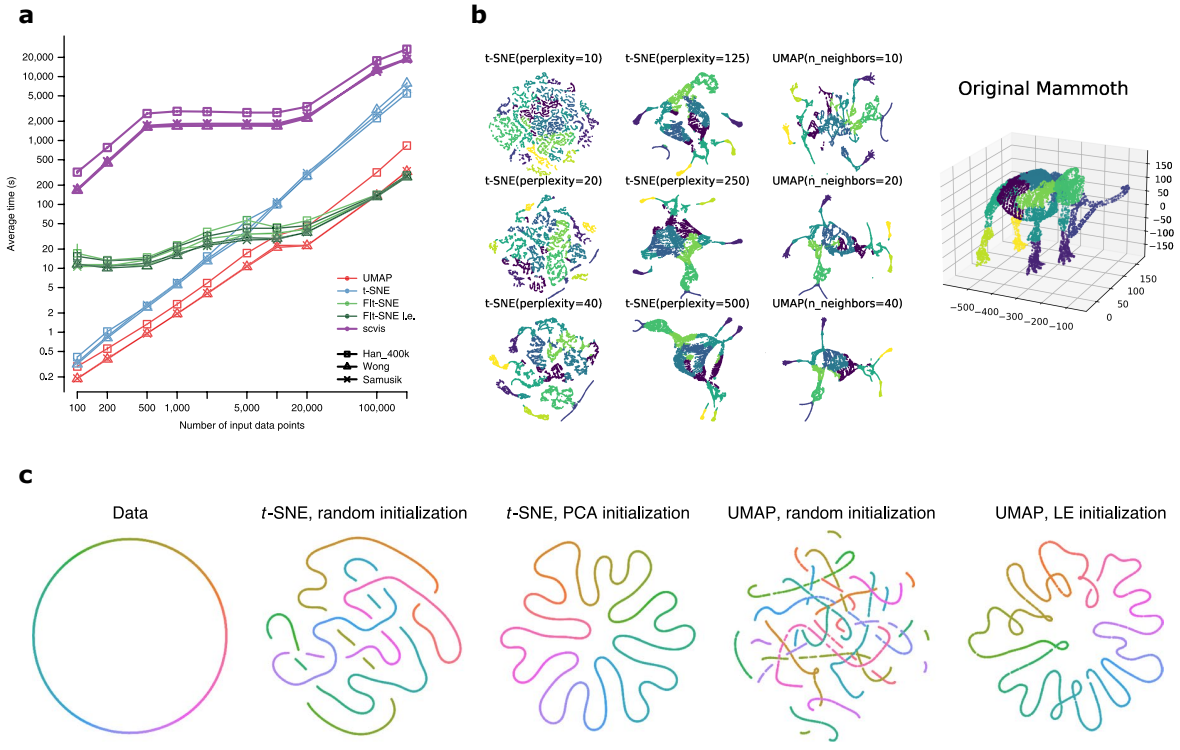

Figure 5 | t-SNE and UMAP tend to perform similarly depending on the task at hand. (a) Run times of five dimensionality reduction methods for subsets of varying sizes from independent datasets (Becht et al., 2018). (b) 2D embeddings of a 3D woolly mammoth dataset generated using t-SNE and UMAP with varying hyperparameter settings (Wang et al., 2021). (c) t-SNE and UMAP with random and non-random initialization (Kobak and Linderman, 2021).

5 Materials and Methods

5.1 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 Denyer et al. (2019) were retrieved from the Gene Expression Omnibus database under accession GSE123818. All downstream processing was performed using the Seurat V4 framework (Hao et al., 2021). Genes that are known to be differentially expressed upon protoplasting ($\log_2$ fold change >2, q<0.05) were dismissed prior to further analysis as in Denyer et al. (2019). Initially, putative high quality cells were identified by filtering the raw library for cells with less than 5% of genes mapping to mitochondrial and chloroplast genes, as well as 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 filtered were predominantly trichoblasts ($p = 2 \times 10^{-4}, \chi^2$-test), and this would lead to removing approx. 30% of all trichoblasts. The χ^2-test was performed on the contingency table describing the number of cells for cell type in the filtered dataset and the cells that were filtered. We therefore decided not to filter any 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 % reads mapping to mitochondrial and chloroplast genes as additional regressors (Hafemeister and Satija, 2019). The 3000 most variable genes were identified as those having the highest residual variance in this regression model. The dimensionality of the data was reduced by performing principal components analysis on the top 3000 most variable genes (Luecken and Theis, 2019). 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 the Seurat function *RunUMAP* using standard settings (n.neighbors = 30; min.dist = 0.3; 500 epochs).

5.2 Annotation of Clusters and Cell Types

For each cluster, markers were identified using Wilcoxon Rank Sum tests to find genes that are differentially expressed in that cluster compared to cells not in that cluster. The cluster markers were filtered for an average $\log_2$ fold change >0.75 and to be expressed in >25% of cells of that cluster. Only markers with a positive log fold-change for a given cluster compared to the remaining cells were considered. First, these markers were mapped to the cell-type specific markers from Shahan et al. (2022) to infer the cell type associated with the clusters. Second, for every cluster, the top 30 differentially expressed markers ranked by fold change were mapped to tissue-specific cell type markers from Brady et al. (2007). 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 (Bhosale et al., 2018; Brady et al., 2007) as in the annotation process kindly provided in the Github repository of Shahan et al. (2022).

5.3 Trajectory inference and Trajectory Differential Expression

For trajectory analysis, Slingshot (Street et al., 2018) was used to build a minimum-spanning tree on the a subset of clusters 3, 4, 5, 6, 9, 10, 11, 14 an 18, fixing trichoblast (cluster 9), atrichoblast (cluster 6) and cortex (cluster 14) as terminal clusters and cluster 10 (QC/meristematic cells) as starting cluster. Simultaneous principal curves were then fitted based on this minimum spanning 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 (Van den Berge et al., 2020). The generalized additive models were fitted with six knots, which was determined to be optimal using the akaike information criterion in the `evaluateK` function from tradeSeq. Genes that were differentially expressed between differentiated trichoblasts and atrichoblasts were detected using the tradeSeq `diffEndTest`.

5.4 Pseudotime Expression Heatmaps

First, genes that are differentially expressed across pseudotime were identified using the tradeSeq `associationTest`, retaining only those genes whose Wald test statistic had a $p<0.01$. For those genes, the mean smoother was predicted along 30 pseudotime bins in each lineage with `predictSmooth`. Per lineage, the genes were grouped into 7 clusters via hierarchical clustering using Ward's linkage. For each cluster, over-represented Gene Ontology categories were identified using GOseq (Young et al., 2010). For all genes in the dataset, GO annotations ('biological process') were retrieved from *plants.ensembl.org* using BioMart. To account for the gene selection bias based on its length, a probability weighting function was estimated from the raw count matrix. Over-represented categories were identified as those having GoSeq `'Hypergeometric'` (Fisher's exact test) or `'Wallenius'` $p<0.05$ after Benjamni-Hochberg (BH) correction. The heatmaps were produced using the ComplexHeatmap package in R (Gu et al., 2016). For the heatmap annotations, each bin's developmental zone and ploidy level (see above for details on annotation) was determined by majority rule within that bin.

5.5 Ligand-Receptor Interactions

For trichoblast, atichoblast and cortex cells, ligand-receptor (LR) interactions were inferred from co-expression data using PlantPhoneDB (Xu et al., 2022). For all cell type combinations, LR interactions were scored using the mean ligand and receptor expression levels as in Efremova et al. (2020), and significant scores were identified using permutation tests (100 iterations). We filtered the significant LR pairs (BH-corrected $p<0.01$) on the requirement that cells cannot interact with less mature cells from the same developmental lineage. For the remaining pairs, their score across 30 pseudotime bins was determined as the mean of the tradeSeq smoother of the ligand and receptor in the respective cell types from which the interaction was inferred.

6 Results

6.1 Single-cell transcriptomics of the *Arabidopsis* root

We began the analysis with processing the scRNA-Seq dataset of the *Arabidopsis thaliana* root from Denyer et al. (2019). We clustered cells that are similar in terms of gene expression into 19 groups, and these clusters were annotated by mapping cluster-specific gene expression onto marker genes of isolated *Arabdopsis* root tissues from Brady et al. (2007) (figure 6a-c). The UMAP shows a disconnected, star-like structure consisting of the central meristematic clusters extending out to different developmental lineages. Central clusters 3, 4, 10, and 16 all mapped to meristematic xylem tissue, but closer inspection of the identified marker genes showed that these were predominantly markers for cells with high mitotic/proliferative activity, without markers related to xylem development. For example, clusters 3, 4, and 16 highly express genes related to DNA replication or nucleolar functionality, including genes coding for histone family proteins and ribosomal proteins (Bernstein and Baserga, 2004; Bernstein et al., 2007). Further, the mitotic regulators cyclin *CYCB1;1* and Aurora kinase *AUR1* are highly expressed in cluster 16, both of which are commonly used as markers for actively dividing cells (Schnittger and Veylder, 2018; Weimer et al., 2016). Cluster 10 additionally expresses QC marker *GOLVEN 6/RGF8* (Fernandez et al., 2013), and the gene for ent-kaurene oxidase 1 (*GA3, AT5G25900*), which is involved in gibberellin biosynthesis, a process that has been associated with QC identity (Nawy et al., 2005). Cluster 10 therefore likely consists of QC cells as well as proliferating cells. The central mass of meristematic cells radiates outwards towards 5 distinct lineages (figure 6a). These lineages were identified as the vascular cell types (clusters 8, 12, 13), cortex and endodermis (14, 18, and 15, respectively), trichoblast and atrichoblast (5, 9 and 6, 11, resp.), and the root cap, comprising cells of the central columella (0, 17) and lateral root cap (1, 2). Many genes that regulate root cap development were differentially expressed upon protoplasting and were therefore not included in the analysis, and this is reflected in the comparatively low agreement on the identity of root cap cell types compared to others. Interestingly, cells that are located terminally in the root-cap lineage highly express genes involved in peroxisome biogenesis (PEROXIN family members), glucosinolate metabolism, and biogenesis of ER bodies, which are organelles that are involved in stress responses and immunity (Sarkar et al., 2020; T. Su et al., 2019). Other cluster-specific marker genes are in involved responses against abiotic stress, such as those encoding defensin-like proteins (*TI1, AT2G43510*), GENOMES UNCOUPLED 2[4] (*GUN2, AT2G26670*), SENESCENCE 1 (*SEN1, AT4G35770*) and WRKY26 (*AT5G07100*; Phukan et al., 2016). Such stress-specific gene expression could play into the root cap's protective role against biotic and abiotic stresses (Kumar and Iyer-Pascuzzi, 2020), or, alternatively, this stress-response could be the result of programmed cell death during root cap cell sloughing (Kumpf and Nowack, 2015). This hypothesis could not be tested, as the genes involved in root cap cell sloughing are differentially expressed upon protoplasting and were therefore not included in the analysis.

[4] GUN2 is involved in stress-induced chloroplast dysfunction (Crawford et al., 2017)

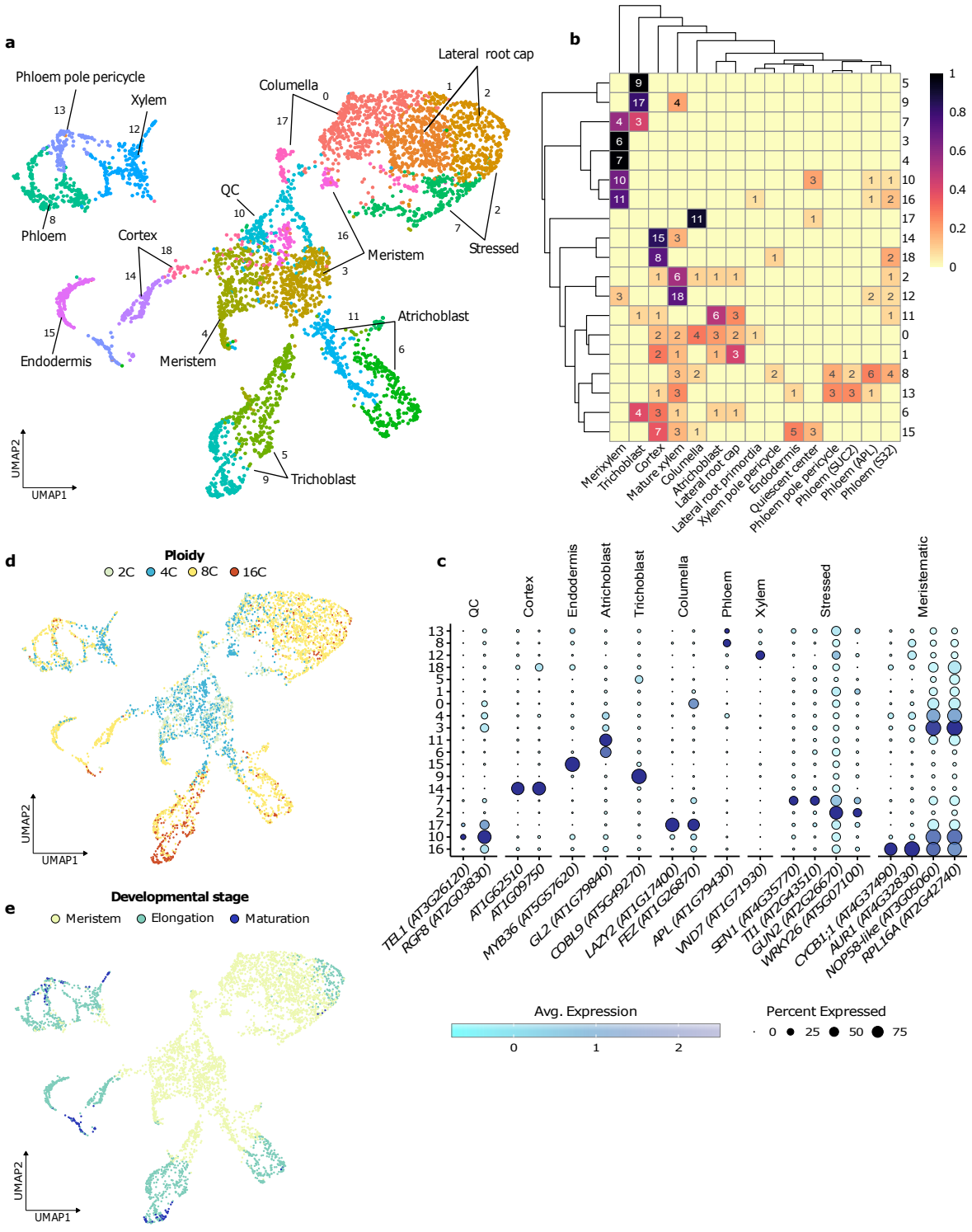

Figure 6 | Cell types of the *Arabidopsis* Root Identified by scRNA-Seq. (a) UMAP of the 4727 *Arabidopsis* root cells. The (sub)clusters are colored and named according to their cell type annotation. Numbers besides the arrows indicate the cluster numbering as referred to in the text. (b) Heatmap showing 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 30 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. (c) Dot plot showing expression of known cell type marker genes per cell cluster identified in the scRNA-Seq dataset. Dot sizes represent % expression in a cluster, color reflects average normalized and scaled expression. (d-e) UMAP with cells annotated according to their inferred ploidy level (d) and developmental zone (e). The ploidy level and developmental zone of each cell was identified by correlation with bulk RNA-seq reference expression profiles.

Further, we note that the resolution between the different vascular cell types is comparatively low (figure 6b). Lastly, the trichoblast and atrichoblast epidermal cells were identified by the characteristic expression of *GLABRA 2* (*GL2, AT1G79840*), a repressor of the trichoblast fate in atrichoblast cells, and the root hair tip growth regulator *COBRA-LIKE 9* (*COBL9, AT5G49270*), respectively (Jones et al., 2005; Masucci et al., 1996).

As the root elongates, developing cells stray away further from the stem cell niche and start to lose their proliferative status, while they increase in length and gain their cellular identity in the elongation and maturation zones (figure 2); (Petricka et al., 2012). During this differentiation, plant cells trade the standard, mitotic cell cycle for endoreduplication and the concomitant increase in cell endoploidy. To infer the developmental status of the cells in our dataset, cellular gene expression profiles were correlated with bulk RNA-Seq reference expression profiles for different root cell ploidy levels and developmental zones (figure 6d-e); (Shahan et al., 2022). The meristematic cells were annotated as having regular ploidy (2C) or having undergone a single endocycle (4C). Given the mutually exclusive nature of the mitotic cells cycle and endoreduplication, and that these cells are annotated as being meristematic, all but the outermost 4C cells are more likely mitotically dividing G2 cells. Such cells cannot be distinguished from 4C cells by flow cytometry and therefore could have ended up in the 4C reference expression profile, which would lead to this observation (Bhosale et al., 2018). The transition from the central meristematic cells to the different developing lineages is accompanied by an increase in ploidy to 8C. The highest ploidy level occurs predominantly in trichoblast cells nearing the maturation stage, with other 16C cells scattered across other mature cell types in small numbers. This additional endoreduplication in root hair cells is a known phenomenon and has been proposed to be involved in tip outgrowth (Bhosale et al., 2018). The transition between the developmental stages is delayed compared to the increase in ploidy levels. This is in agreement with Hayashi et al. (2013), and implies that an increase in cellular ploidy precedes the transition between developmental zones of the root. This developmental stage annotation offers little resolution for root cap cells, as the root cap is almost entirely contained in the meristematic zone of the root (figure2b). More sophisticated methods to segment the root cap into proximal and distal zones exist, but these lie outside the scope of this project (see Shahan et al., 2022). Importantly, the pattern of the endoploidy and developmental stage annotations confirm that the developmental lineages stretch from the central meristematic cells outward to the differentiated cell types, and that the different lineages encompass the entire development from immature cells in the root apical meristem to mature cell types.

6.2 Developmental Trajectories of Root Epidermal Cells

Next, we set out to determine patterns of gene expression that are involved in epidermal cell differentiation. The developmental lineages for trichoblast, atrichoblast, and cortex cells were reconstructed using Slingshot (Street et al., 2018), starting from cluster 10 (QC and meristematic cells), and ending in the differentiated cell types. This approach is supported by the patterns observed in the ploidy and developmental stage annotations, and ensures a pseudotime ordering of the cells that represents a

biologically meaningful signal (figure 7a). The epidermal and cortex trajectories diverge early on during the meristematic stage, which is in agreement with the cell types developing from distinct types of initial cells (the cortex-endodermis and epidermal initials, respectively). The trichoblast and atrichoblasts trajectories have a larger part of their trajectory in common and emerge from a shared body of meristematic cells (cluster 3).

To determine the temporal patterns of gene expression within these lineages, the most variable genes that were significantly differentially expressed across pseudotime were grouped based on their expression across 30 pseudotime bins, and the resulting clusters were evaluated for functional enrichment of GO biological processes (figure 7b). The early meristematic stage was functionally enriched for genes involved in transcriptional activity and genome architecture, as well as auxin-activated and regulating processes. By the time ploidy levels increase towards the later meristematic stage, these is a surge gene expression for ribosome function and translation. The developmental transition from the meristematic to the elongation stage is accompanied by an increase in gene expression for cell wall biogenesis and remodelling, as in Bhosale et al. (2018) and Denyer et al. (2019). Elongation-stage trichoblasts were also functionally enriched for synthesis of thalianol, a triterpentene phytohormone. Late-stage gene expression in either cell type is largely targeted towards transmembrane transport and stress responses. At this point, trichoblasts additionally express genes for cell wall remodelling and tip growth, and known regulators of root hair development, including *ROP4*, *ROOT HAIR SPECIFIC 11* and *13*, and *ROOT HAIR DEFECTIVE 6-LIKE 2* and *4* (figure 7c). The increasing expression of the latter, *RSL4*, harmonizes with the decreasing expression of *GLABRA 2*. GLABRA 2 is master regulator in atrichoblast specification and inhibits a set of bHLH transcription factors that activate a cascade of root hair cell-specifying genes in developing trichoblasts (Lin et al., 2015). These dynamic expression patterns of *RSL4* and *GL2* are in line with the hypothesized involvement of RSL4 in the inhibition of GL2-mediated atrichoblast specification in trichoblast cells (figure 7c, bottom left); (Qiu et al., 2021). To account for possible false positive enrichments due to selection bias because of a gene's length and expression level, we also tested for GO enrichment using GOseq, modeling the selection bias using the mean read count for each gene(Young et al., 2010). Many of the functionally enriched categories that seem relevant to the biological context, e.g. auxin-related processes, root hair elongation, are no longer detected as enriched after correcting for this selection bias, with only terms relating to translation, transcription and membrane/water transport remaining (figure 7, terms annotated '**').

6.3 The role of ligand-receptor interactions in trichoblast development

Trichoblasts and atrichoblasts patterning in the root epidermis occurs in a position-dependent manner (figure 2); (Balcerowicz et al., 2015). Intercellular signals produced by cells of the neighboring cortex specify epidermal cells to become either hair cells (H-cells) of non-hair cells (N-cells). To infer inter-lineage crosstalk between the developing cortex and epidermis, putative ligand-receptor (LR) interactions were inferred based on their co-expression between the different cell type clusters (figure

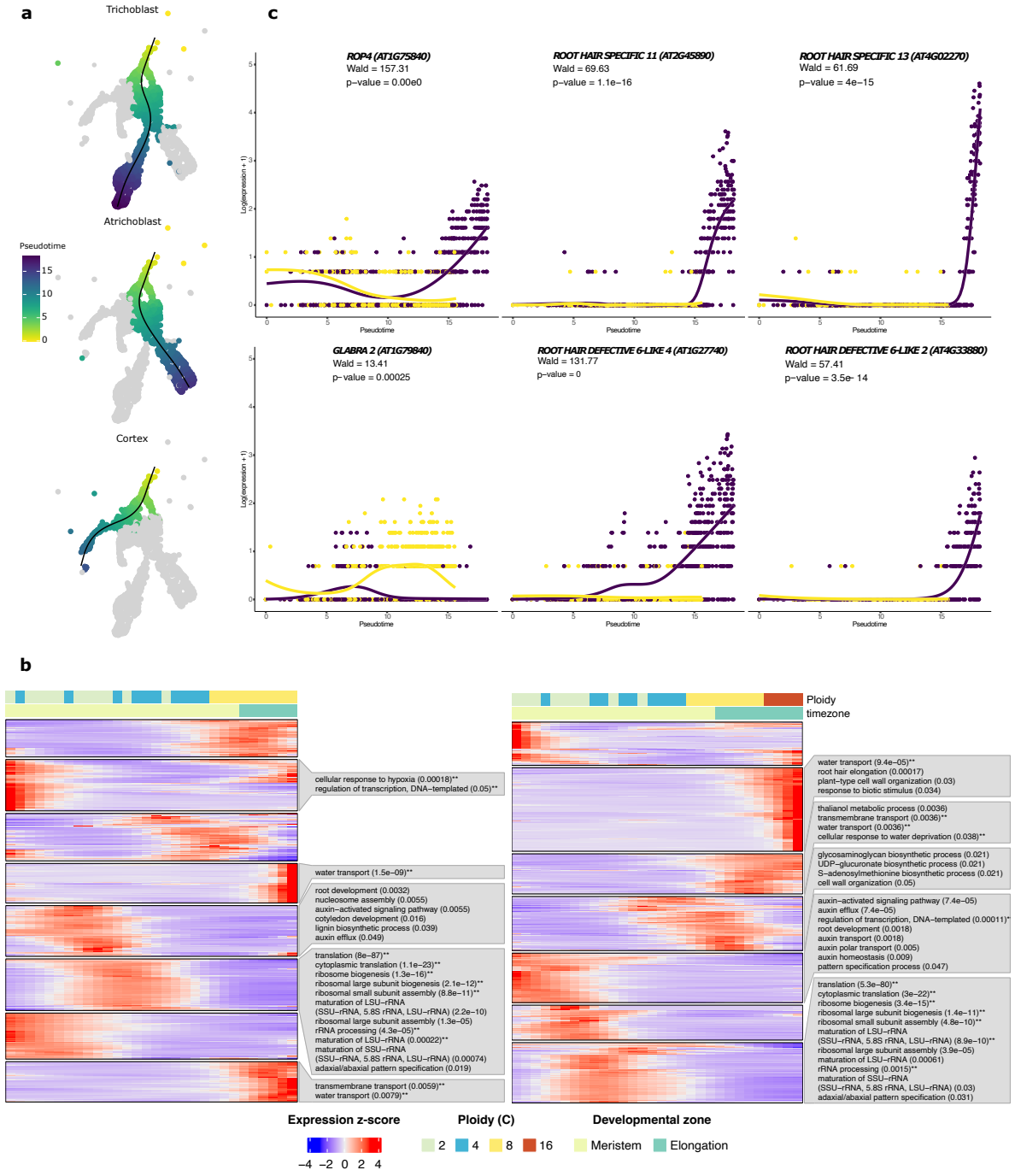

Figure 7 | Differential expression of root hair growth regulators across pseudotime. (a) UMAP describing the developmental lineages from a central cluster (QC/meristematic cells) to differentiated trichoblast, atrichoblast and cortex cell types. The lineages were identified via Slingshot; the black line indicates the principal curve passing through the cells belonging to that lineage. (b) For the atrichoblast (left) and trichoblast (right) lineages, the 3000 most variable genes were clustered according to their expression across 30 pseudotime bins. Each cluster was tested for functional enrichment in GO biological processes using the hypergeometric test and the Wallenius approximation taking into account selection bias due to gene length and expression level. BH-corrected p for the hypergeometric test is shown between brackets; genes that were also significantly enriched after bias correction are annotated with '**'. The developmental zone and ploidy annotations on top correspond were determined by majority-rule in each pseudotime bin. (c) Expression of trichoblast growth regulators across pseudotime within the trichoblast and atrichoblast developmental lineages. For each gene, the line indicates the tradeSeq smoother for the gene's generalized additive model, which describes the expression of that gene across pseudotime within a given lineage (Van den Berge et al., 2020).

6a). Computational tools that infer cell-cell communication from co-expression data typically consider all pairwise combination of cell types. Because a cell's developmental state depends on it's position along the longitudinal axis of the root, we imposed an additional requirement that cells cannot interact with less mature cells from the same developmental lineage. Next, the mean expression scores of significant interactions were determined across 30 pseudotime bins to asses the dynamics of these LR interactions, and the LR pairs were clustered into 6 groups according to their expression across pseudotime (figure 8). The majority of significant LR pairs are most highly expressed in the later developmental stages (figure 8a). We tested the clusters of LR pairs with similar expression patterns for functional enrichment, but none of the clusters was enriched for any GO biological processes (all q>0.3; data not shown). For each cluster of LR-pairs with similar temporal dynamics, the top scoring ligand-receptor pairs are dominated by interactions involving highly abundant cell wall proteins, such as those of the arabinogalactan protein family (AGP), and these top LR pairs are detected across multiple (putative) interacting cell types. The aspecificity of the proposed interactions is also noticeable when looking at the expression of the involved genes across pseudotime, and suggests the need for other methods to score and rank the methods for these applications (figure 8b).

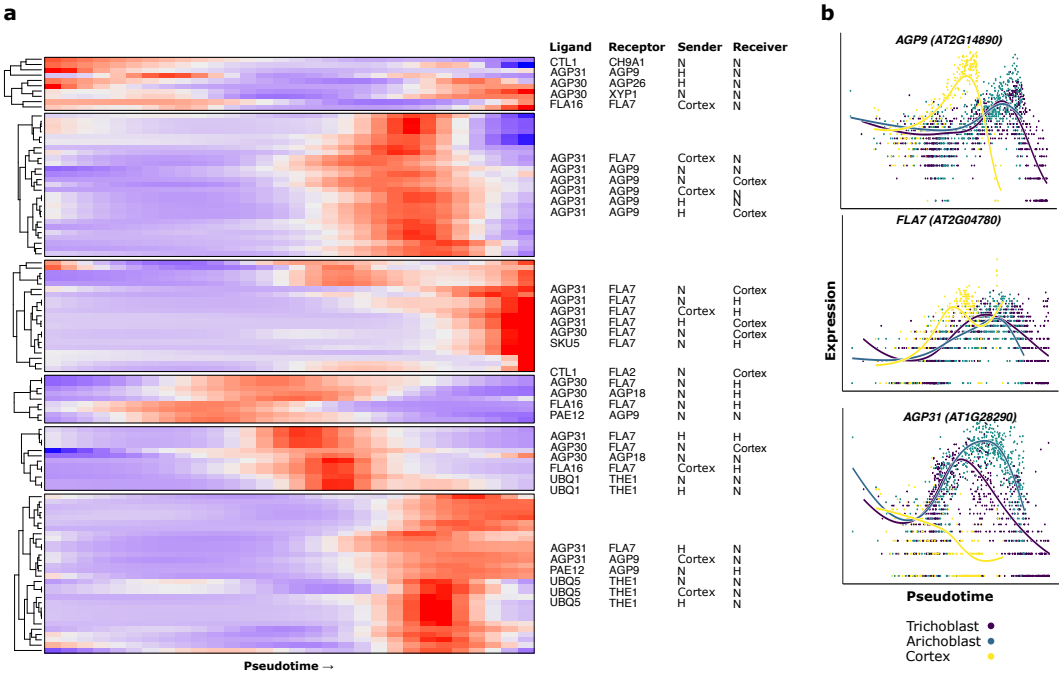

Figure 8 | Crosstalk between cells of the developing root inferred from single-cell expression data. (a) Heatmap showing the mean expression score of putatively interacting ligand-receptor pairs between developing trichoblast (H), atrichoblast (N) and cortex cells across pseudotime. The top (q< 0.01) putative interactions were clustered according to their mean expression in the partaking cell types across 30 pseudotime bins. For each cluster, the six highest ranking ligand-receptor pairs are shown, along with the cell types involved, ranked according to decreasing mean expression score. (b) Expression of a top scoring ligands (*AGP31*) and receptors (*AGP9*, *FLA7*) across pseudotime within developing trichoblast, atrichoblast and cortex cells. For each gene, the line indicates the tradeSeq smoother for the gene's generalized additive model, which describes the expression of that gene across pseudotime within a given lineage (Van den Berge et al., 2020).

7 Discussion

In this project we reanalyzed a single-cell RNA-Seq dataset of the *Arabidopsis* root from Denyer et al. (2019), with the goal of illustrating how trajectory inference can be used to study epidermal cell differentiation. The reconstructed root atlas appears to capture the developmental lineages of the root cap, cortex/endodermis and epidermal cells fairly well. The relations involving the vascular cell types are harder to interpret based on their 2D UMAP representation, and are further hampered by the poor resolution between the different vascular cell types. 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 their associated pericycle (Parizot et al., 2012), as well as issues with accessibility, causing the vascular cells to be underrepresented in single-cell atlases (Otero et al., 2022). The fact that approx. 6000/27000 genes were removed from the dataset because they were differentially expressed during protoplasting presented another nuisance, because this included not only many cell type marker genes (e.g. the ubiquitous QC marker *WOX5*), but also many known regulators of developmental processes, which precluded performing regulatory network inference. Removing genes from the transcriptome that are known to be affected by protoplasting is considered good practice (Denyer and Timmermans, 2022), but is not always performed. For example, Wendrich et al. (2020) state that these genes contributed less than 8% to the top differentially expressed genes, but did not remove them from their dataset.

In previous work, trichoblast development was studied by looking at dynamic enrichment of transcription factor motifs (Jean-Baptiste et al., 2019) or functional enrichment and gene regulatory networks (Denyer et al., 2019). This project expanded on the latter approach by comparing the temporal dynamics of functional specialization between trichoblasts and their atrichoblast counterpart, and layering information about the cell's ploidy and developmental zone on top of this. We reconstructed developmental trajectories for trichoblast and atrichoblast differentiation, and identified temporally regulated gene modules whose functional enrichment caters towards the requirements of the cell's developmental stage. For example, genes that were expressed the highest in the early meristematic stage were involved in auxin-activated and regulating processes, which should be no surprise, given auxin's cardinal role in pattern specification in the root stem cell niche (Pardal and Heidstra, 2021). Other waves of expression appeared to cater towards the needs of the cell when transitioning between developmental zones. Examples include the expression of genes for ribosome function at the transition between ploidy levels, which could be required to maintain steady-state protein concentrations for endoreduplication-driven cell growth, and increased expression of genes for cell wall biogenesis in elongation-stage cells. In agreement with established work, trichoblasts only start to functionally diverge from atrichoblasts starting from the elongation zone, whereas transcriptomic and morphological differences are detected much earlier in the meristematic zone (Balcerowicz et al., 2015). We note that many of the functional enrichments that were detected using the hypergeometric test, and that appear to be biologically relevant (e.g. 'root hair elongation' in late trichoblasts), were no longer

detected after correcting for the selection bias due to gene length and expression level (Young et al., 2010). Importantly, this is something that Denyer et al. (2019) did not take into account. This means that these findings could either be false positives, or marginally significant true positives that benefited from the added power from the selection bias, or this could mean that using gene counts is not a good proxy for gene selection bias (i.e. it is too stringent). These results should therefore be judged in combination with, for example, the pseudo-temporal expression patterns of the contributing genes, as we did for regulators of root-hair growth.

We also attempted to profile inter-lineage crosstalk between the developing cortex and epidermal cell types. To do so, we inferred putative ligand-receptor pairs based on their co-expression between interacting cell types, imposing an additional requirement that cells cannot interact with earlier cells of the same lineage. The pseudo-temporal expression dynamics of these putative LR-pairs then revealed that these significant pairs were predominantly expressed during later stages of development, and that the top-ranking pairs were dominated by pairs involving genes for abundant cell wall proteins (*AGP31*, *AGP9*, *FLA7*), and that they were expressed across all lineages. Multiple arabinogalactan cell wall proteins have been implicated in root cell elongation and root (hair) growth, but most reports are based on phenotypic evidence and lack mechanistic support (Hromadová et al., 2021). Another interesting putative receptor is *THESEUS1*, which is involved in brassinosteroid-mediated cell elongation and is most expressed in mid-to-late development (Gonneau et al., 2018; Guo et al., 2009). Although these examples show that we managed to detect ligand-receptor interactions whose pseudo-temporal expression matches their proposed function, our approach was not optimal. The issue lies in how the significance of an interacting pair is evaluated; this is done by determining the probability of a score under the null hypothesis that co-expression is not enriched between cell types. The LR pairs deemed to be significant in this way are not necessarily informative if you wish to find LR pairs that are significant across pseudotime, because the two questions have different null hypotheses. That is, an LR pair that is significantly co-expressed in two cell types compared to *all* possible combinations is not guaranteed to be informative within some smaller pseudotime window. The immobility of plant cells poses the additional challenge that putative interactions have to be restricted to cells that are in proximity of eachother, which could in theory be accomplished by aligning the lineages by some landmark such as the developmental zones. To be able to perform such an analysis in a valid way would therefore require a more sophisticated approach that assesses the significance of ligand-receptor co-expression over, for example, pseudotime windows that comprise cells of the different lineages that can interact.

All things considered, this project used epidermal cell differentiation in the *Arabidopsis* root as an example to illustrate what single-cell transcriptomics can offer plant biologist studying developmental processes. Comparing our results to those of the original analysis of the dataset showed that the power and validity of the results is dependent on correct statistical practice and can be aided integrating different levels of information. We also illustrate that while computational inference of cell-cell communication between developing lineages is not an easy task, it may be a useful source of information to study developmental processes.

8 Addendum

All supporting information for the analysis that was performed in this project has been made available at github.com/WaDhondt/CompPlantDev.

References

Alpert, A., Moore, L. S., Dubovik, T., & Shen-Orr, S. S. (2018). Alignment of single-cell trajectories to compare cellular expression dynamics. *Nature Methods*, *15*(4), 267–270. https://doi.org/10.1038/nmeth.4628

Amezquita, R. A., Lun, A. T. L., Becht, E., Carey, V. J., Carpp, L. N., Geistlinger, L., Marini, F., Rue-Albrecht, K., Risso, D., Soneson, C., Waldron, L., Pagès, H., Smith, M. L., Huber, W., Morgan, M., Gottardo, R., & Hicks, S. C. (2019). Orchestrating single-cell analysis with bioconductor. *Nature Methods*, *17*(2), 137–145. https://doi.org/10.1038/s41592-019-0654-x

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29. https://doi.org/10.1038/75556

Balcerowicz, D., Schoenaers, S., & Vissenberg, K. (2015). Cell fate determination and the switch from diffuse growth to planar polarity in arabidopsis root epidermal cells. *Frontiers in Plant Science*, *6*. https://doi.org/10.3389/fpls.2015.01163

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., Ginhoux, F., & Newell, E. W. (2018). Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, *37*(1), 38–44. https://doi.org/10.1038/nbt.4314

Bergen, V., Lange, M., Peidli, S., Wolf, F. A., & Theis, F. J. (2020). Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology*, *38*(12), 1408–1414. https://doi.org/10.1038/s41587-020-0591-3

Bernstein, K. A., & Baserga, S. J. (2004). The small subunit processome is required for cell cycle progression at g1. *Molecular Biology of the Cell*, *15*(11), 5038–5046. https://doi.org/10.1091/mbc.e04-06-0515

Bernstein, K. A., Bleichert, F., Bean, J. M., Cross, F. R., & Baserga, S. J. (2007). Ribosome biogenesis is sensed at the start cell cycle checkpoint (K. Weis, Ed.). *Molecular Biology of the Cell*, *18*(3), 953–964. https://doi.org/10.1091/mbc.e06-06-0512

Bhosale, R., Boudolf, V., Cuevas, F., Lu, R., Eekhout, T., Hu, Z., Isterdael, G. V., Lambert, G. M., Xu, F., Nowack, M. K., Smith, R. S., Vercauteren, I., Rycke, R. D., Storme, V., Beeckman, T., Larkin, J. C., Kremer, A., Höfte, H., Galbraith, D. W., … Veylder, L. D. (2018). A spatiotemporal DNA endoploidy map of the arabidopsis root reveals roles for the endocycle in root development

and stress adaptation. *The Plant Cell*, *30*(10), 2330–2351. https://doi.org/10.1105/tpc.17.00983

Birnbaum, K., Shasha, D. E., Wang, J. Y., Jung, J. W., Lambert, G. M., Galbraith, D. W., & Benfey, P. N. (2003). A gene expression map of the iarabidopsis/i root. *Science*, *302*(5652), 1956–1960. https://doi.org/10.1126/science.1090022

Brady, S. M., Orlando, D. A., Lee, J.-Y., Wang, J. Y., Koch, J., Dinneny, J. R., Mace, D., Ohler, U., & Benfey, P. N. (2007). A high-resolution root spatiotemporal map reveals dominant expression patterns. *Science*, *318*(5851), 801–806. https://doi.org/10.1126/science.1146265

Bruex, A., Kainkaryam, R. M., Wieckowski, Y., Kang, Y. H., Bernhardt, C., Xia, Y., Zheng, X., Wang, J. Y., Lee, M. M., Benfey, P., Woolf, P. J., & Schiefelbein, J. (2012). A gene regulatory network for root epidermis cell differentiation in arabidopsis (D. C. Bergmann, Ed.). *PLoS Genetics*, *8*(1), e1002446. https://doi.org/10.1371/journal.pgen.1002446

Chari, T., & Pachter, L. (2021). The specious art of single-cell genomics. https://doi.org/10.1101/2021.08.25.457696

Cook, J., Sutskever, I., Mnih, A., & Hinton, G. (2007). Visualizing similarity data with a mixture of maps. In M. Meila & X. Shen (Eds.), *Proceedings of the eleventh international conference on artificial intelligence and statistics* (pp. 67–74). PMLR. https://proceedings.mlr.press/v2/cook07a.html

Crawford, T., Lehotai, N., & Strand, Å. (2017). The role of retrograde signals during plant stress responses. *Journal of Experimental Botany*, *69*(11), 2783–2795. https://doi.org/10.1093/jxb/erx481

Deconinck, L., Cannoodt, R., Saelens, W., Deplancke, B., & Saeys, Y. (2021). Recent advances in trajectory inference from single-cell omics data. *Current Opinion in Systems Biology*, *27*, 100344. https://doi.org/10.1016/j.coisb.2021.05.005

Denyer, T., Ma, X., Klesen, S., Scacchi, E., Nieselt, K., & Timmermans, M. C. (2019). Spatiotemporal developmental trajectories in the arabidopsis root revealed using high-throughput single-cell RNA sequencing. *Developmental Cell*, *48*(6), 840–852.e5. https://doi.org/10.1016/j.devcel.2019.02.022

Denyer, T., & Timmermans, M. C. (2022). Crafting a blueprint for single-cell RNA sequencing. *Trends in Plant Science*, *27*(1), 92–103. https://doi.org/10.1016/j.tplants.2021.08.016

Eckart, C., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, *1*(3), 211–218. https://doi.org/10.1007/bf02288367

Efremova, M., Vento-Tormo, M., Teichmann, S. A., & Vento-Tormo, R. (2020). CellPhoneDB: Inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nature Protocols*, *15*(4), 1484–1506. https://doi.org/10.1038/s41596-020-0292-x

Fernandez, A., Hilson, P., & Beeckman, T. (2013). GOLVEN peptides as important regulatory signalling molecules of plant development. *Journal of Experimental Botany*, *64*(17), 5263–5268. https://doi.org/10.1093/jxb/ert248

Ferrell, J. E. (2012). Bistability, bifurcations, and waddington's epigenetic landscape. *Current Biology*, *22*(11), R458–R466. https://doi.org/10.1016/j.cub.2012.03.045

Ghojogh, B., Samad, M. N., Mashhadi, S. A., Kapoor, T., Ali, W., Karray, F., & Crowley, M. (2019). Feature selection and feature extraction in pattern analysis: A literature review. https://doi.org/10.48550/ARXIV.1905.02845

Gonneau, M., Desprez, T., Martin, M., Doblas, V. G., Bacete, L., Miart, F., Sormani, R., Hématy, K., Renou, J., Landrein, B., Murphy, E., Cotte, B. V. D., Vernhettes, S., Smet, I. D., & Höfte, H. (2018). Receptor kinase THESEUS1 is a rapid alkalinization factor 34 receptor in arabidopsis. *Current Biology*, *28*(15), 2452–2458.e4. https://doi.org/10.1016/j.cub.2018.05.075

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(6), 333–351. https://doi.org/10.1038/nrg.2016.49

Griffiths, J. A., Scialdone, A., & Marioni, J. C. (2018). Using single-cell genomics to understand developmental processes and cell fate decisions. *Molecular Systems Biology*, *14*(4). https://doi.org/10.15252/msb.20178046

Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, *32*(18), 2847–2849. https://doi.org/10.1093/bioinformatics/btw313

Guo, H., Li, L., Ye, H., Yu, X., Algreen, A., & Yin, Y. (2009). Three related receptor-like kinases are required for optimal cell elongation in iarabidopsis thaliana/i. *Proceedings of the National Academy of Sciences*, *106*(18), 7648–7653. https://doi.org/10.1073/pnas.0812346106

Hafemeister, C., & Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology*, *20*(1). https://doi.org/10.1186/s13059-019-1874-1

Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F., & Theis, F. J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods*, *13*(10), 845–848. https://doi.org/10.1038/nmeth.3971

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. M., Yeung, B., … Satija, R. (2021). Integrated analysis of multimodal single-cell data. *Cell*, *184*(13), 3573–3587.e29. https://doi.org/10.1016/j.cell.2021.04.048

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference and prediction* (2nd ed.). Springer.

Hayashi, K., Hasegawa, J., & Matsunaga, S. (2013). The boundary of the meristematic and elongation zones in roots: Endoreduplication precedes rapid cell expansion. *Scientific Reports*, *3*(1). https://doi.org/10.1038/srep02723

Hinton, G. E., & Roweis, S. (2002). Stochastic neighbor embedding. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems*. MIT Press. https://proceedings.neurips.cc/paper/2002/file/6150ccc6069bea6b5716254057a194ef-Paper.pdf

Hromadová, D., Soukup, A., & Tylová, E. (2021). Arabinogalactan proteins in plant roots – an update on possible functions. *Frontiers in Plant Science*, *12*. https://doi.org/10.3389/fpls.2021.674010

Jean-Baptiste, K., McFaline-Figueroa, J. L., Alexandre, C. M., Dorrity, M. W., Saunders, L., Bubb, K. L., Trapnell, C., Fields, S., Queitsch, C., & Cuperus, J. T. (2019). Dynamics of gene expression in single root cells of iarabidopsis thaliana/i. *The Plant Cell*, *31*(5), 993–1011. https://doi.org/10.1105/tpc.18.00785

Jones, M. A., Raymond, M. J., & Smirnoff, N. (2005). Analysis of the root-hair morphogenesis transcriptome reveals the molecular identity of six genes with roles in root-hair development in arabidopsis. *The Plant Journal*, *45*(1), 83–100. https://doi.org/10.1111/j.1365-313x.2005.02609.x

Kim, J.-Y., Symeonidi, E., Pang, T. Y., Denyer, T., Weidauer, D., Bezrutczyk, M., Miras, M., Zöllner, N., Hartwig, T., Wudick, M. M., Lercher, M., Chen, L.-Q., Timmermans, M. C. P., & Frommer, W. B. (2021). Distinct identities of leaf phloem cells revealed by single cell transcriptomics. *The Plant Cell*, *33*(3), 511–530. https://doi.org/10.1093/plcell/koaa060

Kobak, D., & Berens, P. (2019). The art of using t-SNE for single-cell transcriptomics. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-13056-x

Kobak, D., & Linderman, G. C. (2021). Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nature Biotechnology*, *39*(2), 156–157. https://doi.org/10.1038/s41587-020-00809-z

Kumar, N., & Iyer-Pascuzzi, A. S. (2020). Shedding the last layer: Mechanisms of root cap cell release. *Plants*, *9*(3), 308. https://doi.org/10.3390/plants9030308

Kumpf, R. P., & Nowack, M. K. (2015). The root cap: A short story of life and death. *Journal of Experimental Botany*, *66*(19), 5651–5662. https://doi.org/10.1093/jxb/erv295

Lin, Q., Ohashi, Y., Kato, M., Tsuge, T., Gu, H., Qu, L.-J., & Aoyama, T. (2015). GLABRA2 directly suppresses basic helix-loop-helix transcription factor genes with diverse functions in root hair development. *The Plant Cell*, tpc.15.00607. https://doi.org/10.1105/tpc.15.00607

Luecken, M. D., & Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: A tutorial. *Molecular Systems Biology*, *15*(6). https://doi.org/10.15252/msb.20188746

Masucci, J. D., Rerie, W. G., Foreman, D. R., Zhang, M., Galway, M. E., Marks, M. D., & Schiefelbein, J. W. (1996). The homeobox gene iGLABRA 2/i is required for position-dependent cell differentiation in the root epidermis of iarabidopsis thaliana/i. *Development*, *122*(4), 1253–1260. https://doi.org/10.1242/dev.122.4.1253

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. https://doi.org/10.48550/ARXIV.1802.03426

Nawy, T., Lee, J.-Y., Colinas, J., Wang, J. Y., Thongrod, S. C., Malamy, J. E., Birnbaum, K., & Benfey, P. N. (2005). Transcriptional profile of the arabidopsis root quiescent center. *The Plant Cell*, *17*(7), 1908–1925. https://doi.org/10.1105/tpc.105.031724

Nguyen, H., Tran, D., Tran, B., Pehlivan, B., & Nguyen, T. (2020). A comprehensive survey of regulatory network inference methods using single cell RNA sequencing data. *Briefings in Bioinformatics*, *22*(3). https://doi.org/10.1093/bib/bbaa190

Otero, S., Gildea, I., Roszak, P., Lu, Y., Vittori, V. D., Bourdon, M., Kalmbach, L., Blob, B., Heo, J.-o., Peruzzo, F., Laux, T., Fernie, A. R., Tavares, H., & Helariutta, Y. (2022). A root phloem pole cell atlas reveals common transcriptional states in protophloem-adjacent cells. *Nature Plants*, *8*(8), 954–970. https://doi.org/10.1038/s41477-022-01178-y

Pardal, R., & Heidstra, R. (2021). Root stem cell niche networks: It's complexed! insights from arabidopsis (K. Vissenberg, Ed.). *Journal of Experimental Botany*, *72*(19), 6727–6738. https://doi.org/10.1093/jxb/erab272

Parizot, B., Roberts, I., Raes, J., Beeckman, T., & Smet, I. D. (2012). Iin silico/ianalyses of pericycle cell populations reinforce their relation with associated vasculature iniarabidopsis/i. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1595), 1479–1488. https://doi.org/10.1098/rstb.2011.0227

Petricka, J. J., Winter, C. M., & Benfey, P. N. (2012). Control of iarabidopsis/i root development. *Annual Review of Plant Biology*, *63*(1), 563–590. https://doi.org/10.1146/annurev-arplant-042811-105501

Phukan, U. J., Jeena, G. S., & Shukla, R. K. (2016). WRKY transcription factors: Molecular regulation and stress responses in plants. *Frontiers in Plant Science*, *7*. https://doi.org/10.3389/fpls.2016.00760

Picard, C. L., Povilus, R. A., Williams, B. P., & Gehring, M. (2021). Transcriptional and imprinting complexity in arabidopsis seeds at single-nucleus resolution. *Nature Plants*, *7*(6), 730–738. https://doi.org/10.1038/s41477-021-00922-0

Qiu, Y., Tao, R., Feng, Y., Xiao, Z., Zhang, D., Peng, Y., Wen, X., Wang, Y., & Guo, H. (2021). EIN3 and RSL4 interfere with an MYB–bHLH–WD40 complex to mediate ethylene-induced ectopic root hair formation in iarabidopsis/i. *Proceedings of the National Academy of Sciences*, *118*(51). https://doi.org/10.1073/pnas.2110004118

Ryu, K. H., Huang, L., Kang, H. M., & Schiefelbein, J. (2019). Single-cell RNA sequencing resolves molecular relationships among individual plant cells. *Plant Physiology*, *179*(4), 1444–1456. https://doi.org/10.1104/pp.18.01482

Saelens, W., Cannoodt, R., Todorov, H., & Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. *Nature Biotechnology*, *37*(5), 547–554. https://doi.org/10.1038/s41587-019-0071-9

Salazar-Henao, J. E., Vélez-Bermúdez, I. C., & Schmidt, W. (2016). The regulation and plasticity of root hair patterning and morphogenesis. *Development*, *143*(11), 1848–1858. https://doi.org/10.1242/dev.132845

Sarkar, S., Stefanik, N., Kunieda, T., Hara-Nishimura, I., & Yamada, K. (2020). The arabidopsis transcription factor NAI1 activates the iNAI2/i promoter by binding to the g-box motifs. *Plant Signaling &amp Behavior*, *16*(2), 1846928. https://doi.org/10.1080/15592324.2020.1846928

Schiefelbein, J., Huang, L., & Zheng, X. (2014). Regulation of epidermal cell fate in arabidopsis roots: The importance of multiple feedback loops. *Frontiers in Plant Science*, *5*. https://doi.org/10.3389/fpls.2014.00047

Schnittger, A., & Veylder, L. D. (2018). The dual face of cyclin b1. *Trends in Plant Science*, *23*(6), 475–478. https://doi.org/10.1016/j.tplants.2018.03.015

Seyfferth, C., Renema, J., Wendrich, J. R., Eekhout, T., Seurinck, R., Vandamme, N., Blob, B., Saeys, Y., Helariutta, Y., Birnbaum, K. D., & Rybel, B. D. (2021). Advances and opportunities in single-cell transcriptomics for plant research. *Annual Review of Plant Biology*, *72*(1), 847–866. https://doi.org/10.1146/annurev-arplant-081720-010120

Shahan, R., Hsu, C.-W., Nolan, T. M., Cole, B. J., Taylor, I. W., Greenstreet, L., Zhang, S., Afanassiev, A., Vlot, A. H. C., Schiebinger, G., Benfey, P. N., & Ohler, U. (2022). A single-cell arabidopsis root atlas reveals developmental trajectories in wild-type and cell identity mutants. *Developmental Cell*, *57*(4), 543–560.e9. https://doi.org/10.1016/j.devcel.2022.01.008

Shulse, C. N., Cole, B. J., Ciobanu, D., Lin, J., Yoshinaga, Y., Gouran, M., Turco, G. M., Zhu, Y., O'Malley, R. C., Brady, S. M., & Dickel, D. E. (2019). High-throughput single-cell transcriptome profiling of plant cell types. *Cell Reports*, *27*(7), 2241–2247.e4. https://doi.org/10.1016/j.celrep.2019.04.054

Street, K., Risso, D., Fletcher, R. B., Das, D., Ngai, J., Yosef, N., Purdom, E., & Dudoit, S. (2018). Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics*, *19*(1). https://doi.org/10.1186/s12864-018-4772-0

Su, E. Y., Spangler, A., Bian, Q., Kasamoto, J. Y., & Cahan, P. (2022). Reconstruction of dynamic regulatory networks reveals signaling-induced topology changes associated with germ layer specification. *Stem Cell Reports*, *17*(2), 427–442. https://doi.org/10.1016/j.stemcr.2021.12.018

Su, T., Li, W., Wang, P., & Ma, C. (2019). Dynamics of peroxisome homeostasis and its role in stress response and signaling in plants. *Frontiers in Plant Science*, *10*. https://doi.org/10.3389/fpls.2019.00705

Svensson, V., Vento-Tormo, R., & Teichmann, S. A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols*, *13*(4), 599–604. https://doi.org/10.1038/nprot.2017.149

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K., & Surani, M. A. (2009). mRNA-seq whole-transcriptome analysis of a single cell. *Nature Methods*, *6*(5), 377–382. https://doi.org/10.1038/nmeth.1315

Trapnell, C. (2015). Defining cell types and states with single-cell genomics. *Genome Research*, *25*(10), 1491–1498. https://doi.org/10.1101/gr.190595.115

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., & Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, *32*(4), 381–386. https://doi.org/10.1038/nbt.2859

Van den Berge, K., de Bézieux, H. R., Street, K., Saelens, W., Cannoodt, R., Saeys, Y., Dudoit, S., & Clement, L. (2020). Trajectory-based differential expression analysis for single-cell sequencing data. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-14766-3

Van der Maaten, L., & Hinton, G. (2008). Visualizing high-dimensional data using t-sne [Pagination: 27]. *Journal of Machine Learning Research*, *9*(nov), 2579–2605.

Vandereyken, K., Sifrim, A., Thienpont, B., & Voet, T. (2023). Methods and applications for single-cell and spatial multi-omics. *Nature Reviews Genetics*. https://doi.org/10.1038/s41576-023-00580-2

Waddington, C. H. (2011). The epigenotype. *International Journal of Epidemiology*, *41*(1), 10–13. https://doi.org/10.1093/ije/dyr184

Waddington, C. (1957). *The strategy of the genes*. Routledge. https://doi.org/10.4324/9781315765471

Wang, Y., Huang, H., Rudin, C., & Shaposhnik, Y. (2021). Understanding how dimension reduction tools work: An empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization.

Wang, Y., Huan, Q., Chu, X., Li, K., & Qian, W. (2020). Single-cell transcriptome analyses recapitulate the cellular and developmental responses to abiotic stresses in rice. https://doi.org/10.1101/2020.01.30.926329

Weimer, A. K., Demidov, D., Lermontova, I., Beeckman, T., & Damme, D. V. (2016). Aurora kinases throughout plant development. *Trends in Plant Science*, *21*(1), 69–79. https://doi.org/10.1016/j.tplants.2015.10.001

Wendrich, J. R., Yang, B., Vandamme, N., Verstaen, K., Smet, W., de Velde, C. V., Minne, M., Wybouw, B., Mor, E., Arents, H. E., Nolf, J., Duyse, J. V., Isterdael, G. V., Maere, S., Saeys, Y., & Rybel, B. D. (2020). Vascular transcription factors guide plant epidermal responses to limiting phosphate conditions. *Science*, *370*(6518). https://doi.org/10.1126/science.aay4970

Xiang, R., Wang, W., Yang, L., Wang, S., Xu, C., & Chen, X. (2021). A comparison for dimensionality reduction methods of single-cell RNA-seq data. *Frontiers in Genetics*, *12*. https://doi.org/10.3389/fgene.2021.646936

Xu, C., Ma, D., Ding, Q., Zhou, Y., & Zheng, H.-L. (2022). PlantPhoneDB: A manually curated pan-plant database of ligand-receptor pairs infers cell–cell communication. *Plant Biotechnology Journal*, *20*(11), 2123–2134. https://doi.org/10.1111/pbi.13893

Yang, Z., King, I., Xu, Z., & Oja, E. (2009). Heavy-tailed symmetric stochastic neighbor embedding. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, & A. Culotta (Eds.), *Advances in neural*

information processing systems. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2009/file/2291d2ec3b3048d1a6f86c2c4591b7e0-Paper.pdf

Young, M. D., Wakefield, M. J., Smyth, G. K., & Oshlack, A. (2010). Gene ontology analysis for RNA-seq: Accounting for selection bias. *Genome Biology*, *11*(2), R14. https://doi.org/10.1186/gb-2010-11-2-r14

Zhang, T.-Q., Chen, Y., & Wang, J.-W. (2021). A single-cell analysis of the arabidopsis vegetative shoot apex. *Developmental Cell*, *56*(7), 1056–1074.e8. https://doi.org/10.1016/j.devcel.2021.02.021

Zhang, T.-Q., Xu, Z.-G., Shang, G.-D., & Wang, J.-W. (2019). A single-cell RNA sequencing profiles the developmental landscape of arabidopsis root. *Molecular Plant*, *12*(5), 648–660. https://doi.org/10.1016/j.molp.2019.04.004