Developmental and conditional dynamics of gene expression in single root cells of *A. thaliana*

Ken Jean-Baptiste[1], José L. McFaline-Figueroa[1], Cristina M. Alexandre[1], Michael W. Dorrity[1,2], Lauren Saunders[1], Kerry L. Bubb[1], Cole Trapnell[1], Stanley Fields[1,3], Christine Queitsch[1]*, and Josh T. Cuperus[1]*

*Corresponding authors: Cuperusj@uw.edu, Queitsch@uw.edu

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) is: Josh Cuperus (Cuperusj@uw.edu).

Short title: Single-cell RNA-seq in Arabidopsis roots

1 Department of Genome Sciences, University of Washington, Seattle, WA 98195.
2 Department of Biology, University of Washington, Seattle, WA 98195.
3 Department of Medicine, University of Washington, Seattle, WA 98195.

Single-cell RNA-seq can yield high-resolution cell-type-specific expression signatures that reveal new cell types and the developmental trajectories of cell lineages. Here, we apply this approach to *A. thaliana* root cells to capture gene expression in 3,121 root cells. We analyze these data with Monocle 3, which orders single cell transcriptomes in an unsupervised manner and uses machine learning to reconstruct single-cell developmental trajectories along pseudotime. We identify hundreds of genes with cell-type-specific expression, with pseudotime analysis of several cell lineages revealing both known and novel genes that are expressed along a developmental trajectory. We identify transcription factor motifs that are enriched in early and late cells, together with the corresponding candidate transcription factors that likely drive the observed expression patterns. Finally, by applying heat stress to whole seedlings, we address the longstanding question of possible heterogeneity among cell types in the response to an abiotic stress. Although the response of canonical heat shock genes dominates expression across cell types, subtle but significant differences in other genes can be detected among cell types. Taken together, our results demonstrate that single-cell transcriptomics holds promise for studying plant development and plant physiology with unprecedented resolution.

Introduction

Many features of plant organs such as roots are traceable to specialized cell lineages and their progenitors (Irish, 1991; Petricka et al., 2012). The developmental trajectories of these lineages have been based on tissue-specific and cell-type-specific expression data derived from tissue dissection and reporter gene-enabled cell sorting (Birnbaum et al., 2003; Brady et al., 2007; Li et al., 2016). However, tissue dissection is labor-intensive and imprecise, and the cell sorting requires prior knowledge of cell-type-specific promoters and genetic manipulation to generate reporter lines. Few such lines are available for plants other than the reference plant *Arabidopsis thaliana* (Rogers and Benfey, 2015). Advances in single-cell transcriptomics can replace these labor-intensive approaches. Single-cell RNA-seq has been applied to heterogeneous samples of human, worm, and virus origin, among others, yielding an unprecedented depth of cell-type-specific information (Cao et al., 2017; Irish, 1991; Packer and Trapnell, 2018; Russell et al., 2018; Trapnell, 2015; Trapnell et al., 2014). Here, we take advantage of expression data from root-specific reporter lines in *A. thaliana* (Birnbaum et al., 2003; Brady et al., 2007; Cartwright et al., 2009; Li et al., 2016) to explore the potential of single-cell RNA-seq to capture the expression of known cell-type-specific genes and to identify new ones. We focus on roots of mature seedlings and probe the developmental trajectories of several cell lineages.

Results

Single cell RNA-seq of whole *A. thaliana* roots reveals distinct populations of cortex, endodermis, hair, non-hair, and stele cells

We used whole *A. thaliana* roots from seven-day-old seedlings to generate protoplasts for transcriptome analysis using the 10x Genomics platform (**Supplemental Figure 1A**). We captured 3,121 root cells to obtain a median of 6,152 unique molecular identifiers (UMIs) per cell. These UMIs corresponded to the expression of a median of 2,445 genes per cell and a total of 22,419 genes, close to the gene content of *A. thaliana*. Quality measures for sequencing and read mapping were high. Of the approximately 79,483,000 reads, 73.5% mapped to the TAIR10

A. thaliana genome assembly, with 67% of these annotated transcripts. These values are well within the range reported for droplet-based single-cell RNA-seq in animals and humans.

For data analysis, we applied Monocle 3, which orders transcriptome profiles of single cells in an unsupervised manner without *a priori* knowledge of marker genes (Qiu et al., 2017a; Qiu et al., 2017b; Trapnell et al., 2014). We used the 1500 genes in the data set (**Supplemental Table 1**) that showed the highest variation in expression (empirically calculated dispersion at least 2.7 times higher than the dispersion curve fit for the expression of all genes (**Supplemental Figure 1B**)). For unsupervised clustering, we used 25 principal components (PC). These 25 accounted for 72.5% of the variance explained by the first 100 PCs, with the first PC explaining 11% and the 25th PC explaining 0.9% (**Supplemental Figure 1C**). Cells were projected on to two dimensions using the uniform manifold approximation and projection (UMAP) method (McInnes and Healy, 2018) and clustered, resulting in 11 clusters (**Figure 1A**) (Blondel et al., 2008). Most clusters showed similar levels of total nuclear mRNA, although clusters 9 and 11 were exceptions with higher levels (**Supplemental figure 1D**).

To assign these clusters to cell types, we performed two complementary analyses relying on expression data from tissue-specific and cell-type-specific reporter lines (Brady et al., 2007; Cartwright et al., 2009; Li et al., 2016). First, we compared the expression data to the gene expression values in each single cell, using Spearman's rank correlations to assign each cell to a cell type based on the highest correlation (**Figure 1B, Supplemental Table 2**) (Brady et al., 2007; Cartwright et al., 2009). Second, we examined the expression of 530 cell-type-specific marker genes (Brady et al., 2007) by defining seven marker gene clusters with k-means clustering and calculating their average expression for each cell. We then compared each cell's UMAP cluster assignment (**Figure 1A**) with its marker-gene-based assignment. In general, the UMAP clusters showed high and cluster-specific expression of marker genes. For example, cells in cluster 10 showed high and specific mean expression of cortex marker genes (**Figure 1C, Supplemental Figure 2, Supplemental Table 3**). Both methods allowed us to assign the Louvain components to five major groups: root hair cells, non-hair cells (split between an early and late cluster), cortex cells, endodermis cells and stele cells (containing both xylem and phloem cells) (**Figure 1A**). Although some cells showed the highest correlation to the cell type

Columella in Spearman's rank tests, these cells co-clustered with non-hair cells (**Figure 1B**). This finding is consistent with bulk data comparisons. Specifically, the *PET11* (Columella-sorted) bulk expression data are most similar to bulk expression data sorted for *GL2* and *WER,* both of which mark non-hair cells (Li et al., 2016). Therefore, these cells were grouped with other non-hair cells in Louvain component 8, identifying this cluster as early non-hair cells.

To assess the extent to which combined single-cell expression data resembled bulk expression data (Li et al., 2016), we compared the single-cell root data to bulk root data (**Figure 1D, E**). We observed strong correlations between these data sets, with a Pearson's correlation coefficient (R^2) of 0.52 (Spearman's ρ=0.71). We also compared the combined single-cell expression data to three bulk expression data sets representing the major developmental zones in the *A. thaliana* root: the meristematic zone, the elongation zone, and the maturation zone (**Figure 1E**). We observed the highest correlation of single-cell and bulk expression in the elongation zone (R^2=0.70, ρ=0.83) and a lower correlation in the maturation zone (R2=0.58, ρ=0.70), consistent with the developmental stage of the harvested roots (**Supplemental Figure 1A**). As expected, single-cell and bulk expression were poorly correlated in the meristematic zone (R^2=0.11, ρ=0.43), as meristematic tissue accounts for only a small proportion of mature roots. Furthermore, we compared tissue-specific expression (Li et al., 2016) to expression both in the annotated cell clusters and in cells expressing appropriate marker genes. In general, we found strong correlations among these data sets, suggesting that the clusters are annotated correctly (**Supplemental Table 4**).

We also compared the relative representation of root cell types between our data set and estimates based on microscopy studies (**Figure 1F**) (Brady et al., 2007; Cartwright et al., 2009). Independent of annotation method, we observed the expected numbers of cortex (222 Spearman's/ 233 UMAP), endodermis (306/ 304), non-hair cells (1201/ 1061) and Columella cells (111/ no UMAP cluster). Hair cells (565/ 898) were overrepresented whereas stele cells (508/ 490) were underrepresented, possibly reflecting a bias in the protoplast preparations of whole root tissue.

Some marker genes are not expressed exclusively in a single cell type, making it desirable to identify additional genes with cell-type specific expression. We confirmed the high and cluster-specific expression of well-known marker genes (**Figure 2A, Supplemental Figure 3**) (Li et al., 2016) such as the root-hair-specific *COBL9*, the endodermis-specific *SCR* and the three stele-specific genes *MYB46* (xylem-specific), *APL* (phloem-specific), and *SUC2* (phloem-specific). The nonspecific expression of the quiescent center cell marker genes *WOX5* and *AGL42* is likely due to the failure to capture sufficient numbers of these rare cells. The nonspecific expression of *WOL* and the more heterogeneous pattern of both *WER* and *GL2* expression have been previously observed (Brady et al., 2007; Winter et al., 2007).

To find novel marker genes, we identified genes with significantly different expression within and among clusters by applying the Moran's I test implemented in Monocle 3. We thereby identified 164 novel genes with cluster-specific expression, including at least one in each cluster (**Figure 2A, Supplemental Figure 3, Supplemental Table 5)**. Using cell-type annotations rather than Louvain clusters, we identified an additional 125 novel genes with cell-type-specific expression, some of which have been implicated in the development of a cell lineage in targeted molecular genetics studies. For example, the stele-specific AT1G8810 (*ABS5, T5L1*) gene (cluster 7, **Figure 2A**) encodes a bHLH protein that promotes vascular cell division and differentiation as part of a heterodimer with another bHLH protein, LHW (Katayama et al., 2015; Ohashi-Ito et al., 2014). Another stele-specific gene, AT4G36160 (*ANAC076*, *VND2*) (cluster 7), encodes a ClassIIB NAC-domain transcription factor that contributes to xylem vessel element differentiation by promoting secondary cell wall formation and programmed cell death (Tan et al., 2018). In tissue-specific bulk data (Brady et al., 2007; Winter et al., 2007), both genes show xylem-specific expression consistent with their biological functions; *T5L1* expression is high only in the meristematic and elongation zones, while *VND2* expression starts in the elongation zone and persists throughout the maturation zone. Other genes, not previously implicated in root development, show tissue-specific bulk expression patterns consistent with the single-cell data. For example, AT1G54940 (*GUX4*, *PGSIP4*), which encodes a xylan glucuronosyltransferase (Lee et al., 2012; Mortimer et al., 2010), is specifically expressed in hair cells (cluster 9) and is most highly expressed in cells destined to become hair cells in the elongation zone and in differentiated hair cells in the maturation zone.

Some transcription factors show high correlation with specific cell types

We asked whether we could determine which transcription factors govern the cluster-specific expression patterns. To do so, we tested for transcription factor motif enrichments in the proximal regulatory regions of genes with cluster-specific expression, examining 500 bp upstream of the transcription start site (Alexandre et al., 2018; Sullivan et al., 2014) and a comprehensive collection of *A. thaliana* transcription factor motifs (O'Malley et al., 2016). This analysis revealed transcription factor motif enrichments among clusters and annotated major tissues and cell types (**Figure 2B**).

Using these cluster-specific transcription factor motif enrichments, we sought to deduce the identity of the corresponding transcription factors, a challenging task given that transcription factors in *A. thaliana* often belong to large gene families without factor-specific motif information (Riechmann et al., 2000). For each highly enriched motif, we determined which transcription factor family members were expressed in the corresponding cluster or tissue, or in neighboring cell layers (some factors move between cells (Petricka et al., 2012)) (**Supplemental Table 5**). We focused first on the small *BZR/BEH* gene family whose motif was specifically enriched in cortex cells (cluster 10). Of the six genes *BEH1-4*, *BES1*, and *BZR1*, recessive mutants of *BEH3*, *BEH4*, *BES1* and *BEH4* show reduced hypocotyl growth in the dark, with double mutant analysis indicating extensive but incomplete redundancy in function (Lachowiec et al., 2016). Neither *beh1* and *beh2* single mutants nor the respective double mutant show hypocotyl defects or other phenotypic defects (Lachowiec et al., 2018). Nevertheless, *BEH2* was the most highly expressed *BZR/BEH* family member across clusters and annotated tissue and cell types (**Supplemental Figure 4A, B**). *BEH4*, the most ancient member of this gene family and the one with the strongest known phenotypic impact, showed the strongest cortex-specific expression (**Supplemental Figure 4A, B**). In a test for Pearson's correlation in expression of *BZR/BEH* family members across all cells (**Supplemental Figure 4C**), *BES1* and *BZR* expression was highly correlated, consistent with these genes being the most recent duplicates in the family (Lachowiec et al., 2013; Lan and Pritchard, 2016). *BES1* expression was also highly correlated with *BEH2* expression, suggesting a role for *BEH2* in root development.

In contrast to the *BEH/BZR* gene family, we found stronger cluster specificity for some TCP transcription factors. The TCP motif was strongly enriched in cortex (cluster 10), endodermis (cluster 1) and stele (cluster 7). Of the 24 TCP transcription factors, we detected expression for eight. Of these, *TCP14* (AT3G47620) and *TCP15* (AT1G69690) were expressed specifically in stele (clusters 7 and 4) (**Figure 2B, Supplemental Figure 4D, E, Supplemental Table 5**). *TCP14* and *TCP15* are class I TCP factors thought to promote development. Acting together, *TCP14* and *TCP15* promote cell division in young internodes (Kieffer et al., 2011), seed germination (Resentini et al., 2015), cytokinin and auxin responses during gynoecium development (Lucero et al., 2015), and repression of endoreduplication (Peng et al., 2015). Both genes are expressed in stele in bulk tissue data (Brady et al., 2007; Winter et al., 2007), with *TCP14* expression also observed in the vasculature by *in situ* hybridization (Tatematsu et al., 2008). *TCP14* can affect gene expression in a non-cell-autonomous manner.

To further investigate the co-occurrence of cluster-specific transcription factor motif enrichments with transcription factor expression, we examined the novel genes with cluster-specific expression. Eight of these encode transcription factors with corresponding highly enriched cluster-specific binding motifs. For one of these, *BRN2* (AT4G10350), cluster-specific expression coincided with enrichment for its motif NAC (cluster 8, non-hair-cells, **Figure 2B**). *BRN2* encodes a ClassIIB NAC transcription factor implicated in root cap maturation together with *BRN1* and *SMB*. Class IIB NAC transcription factors are thought to contribute to terminal cell differentiation accompanied by strong cell wall modifications (Bennett et al., 2010). In our data, *BRN2* was most highly expressed in cluster 8 (non-hair cells) and less so in cluster 6 (**Supplemental Table 5**).

Pseudotime trajectories coincide with the development stages of cortex, endodermis, and hair cells

We next sought to visualize the continuous program of gene expression changes that occurs as each cell type in root differentiates. Because whole roots contain a mix of cells at varying developmental stages, we reasoned that our experiment should have captured a representative snapshot of their differentiation. Monocle not only clusters cells by type but also placed them in "pseudotime" order along a trajectory that describes their maturity. To make these trajectories,

Monocle 3 learns an explicit principal graph from the single-cell-expression data through reversed graph embedding, an advanced machine learning method (Qiu et al., 2017a; Qiu et al., 2017b; Trapnell et al., 2014). To dissect the developmental dynamics of individual clusters, we first focused on the well-defined root-hair cells, in which combined single-cell expression values highly correlated with those from bulk protoplasts sorted for expression of the *COBL9* root-hair marker gene (**Supplemental Table 4**). To annotate the unsupervised trajectory that Monocle 3 created for hair cells, we used the Spearman's rank test to compare expression in all cells to bulk expression data representing 13 different developmental stages in root tissues (**Supplemental Figure 5**) (Brady et al., 2007; Cartwright et al., 2009). Each cell was assigned the developmental stage most correlated with its expression values (**Figure 3A)**. The hair cells with the earliest developmental stage assignment were designated as the root of the trajectory. Next, pseudotime was calculated for all other hair cells based on their distance from the root of the trajectory (**Figure 3B**). We compared this calculated pseudotime with the most highly correlated developmental assignment from bulk data, finding close agreement (**Figure 3B**). Examples of genes that are expressed early and late in pseudotime in the UMAP hair cluster are shown in **Figure 3C**.

Hair cells undergo endoreduplication as they mature, to contain up to 16N genomic copies in the developmental stages assayed (Bhosale et al., 2018). Endoreduplication is thought to increase transcription rates (Bourdon et al., 2012). However, as hair cell differentiation progresses, general transcription might decrease as hair-cell-specific genes become more expressed. To distinguish among these scenarios, we explored whether transcription rates differed in hair cells across pseudotime. The total amount of mRNA captured was drastically reduced across hair cell development (**Figure 3D**). This result may simply be due to technical bias; for example, larger, endoreduplicated cells may be more difficult to assess with this droplet-based method and thus the observed reduction in captured transcripts should affect all genes more or less equally. Alternatively, this observation may reflect hair cell differentiation, whereby transcription of hair-cell-specific genes should remain unaffected or increase over pseudotime. Indeed, transcription of hair-cell-specific genes appeared to increase over pseudotime, consistent with these cells undergoing differentiation towards terminally differentiated hair cells (**Figure 3E, Supplemental Figure 6A**).

To further explore this transcriptional dynamic, we used RNA velocity measurements. These assess errors in priming during 3' end reverse transcription to determine splicing rates per gene and cell, which correlate with transcriptional rates. In our data, only ~4% of reads were informative for annotating splicing rates. Using data for 1000 genes, including all hair-cell-specific genes, mean RNA velocity increased across pseudotime (**Supplemental Figure 6B,** p = 2.2 e-16 linear model, rho = 0.68). This increase in velocity was associated with the predicted changes in endoreduplication (Bhosale et al., 2018), especially between the 4N and 8N stages (**Supplemental Figure 6C,** Tukey's multiple comparison p-value = 0.0477).

We also observed developmental signals in other cell types, including cortex and endodermis (**Figure 4A-D, Supplemental Figure 7**). Combined single-cell expression values for cortex cells highly correlated with those from bulk protoplasts sorted for expression of the COR cortex marker gene (**Figure 4B**, R^2= 0.74, rho=0.86). As Monocle 3 did not identify a trajectory for cortex cells in the context of all cells, we isolated the cortex cells and re-performed UMAP dimensionality reduction, clustering, and graph embedding (Supplemental Table 1). Each cortex cell was assigned a developmental stage based on its Spearman's rank correlation with bulk expression data (Brady et al., 2007; Cartwright et al., 2009). Cortex cells with the earliest developmental signal were designated as the root of the cortex trajectory, and pseudotime was assigned to the remaining cortex cells based on their distance from the root (**Figure 4A-D, Supplemental Figure 4**). As pseudotime increased for cortex cells, so did their age, indicating good agreement of the trajectory with developmental bulk RNA-seq data.

We asked whether we could assign the transcription factors that drive gene expression in early and late hair, cortex, and endodermis cells. As before, we first analyzed transcription factor motif enrichments and then explored the expression of the corresponding transcription factor gene families. Indeed, for most developmentally enriched transcription factor motifs, we could pinpoint candidate transcription factors that are expressed either early or late. For example, the *AP2/EREBP* (APETALA2/ethylene responsive element binding protein) transcription factor family is one of the largest in *A. thaliana* (Riechmann et al., 2000), with nearly 80 covered in our data set; of these, only four showed strong expression in late hair cells (AT2G25820,

At5G65130, AT1G36060, AT1G44830) (**Figure 3F, G, Supplemental Figure 8**). One of these, AT1G36060 (Translucent Green), regulates expression of aquaporin genes (Zhu et al., 2014). Overexpression of this gene confers greater drought tolerance (Zhu et al., 2014), consistent with its expression in older hair cells. Similar examples of developmental stage-specific motif enrichments with corresponding transcription factor expression were also found for cortex and endodermis (**Figure 4E, F**, **Supplemental Figure 7, Supplemental Figure 8**).

Heat-shocked root cells show subtle expression differences among cell types

A major question in studying plant responses to abiotic stress, such as heat or drought, is the extent to which such responses are non-uniform across cell types. Canonically, the heat stress response is characterized by rapid and massive up-regulation of a few loci, mostly encoding heat shock proteins, with dramatic down-regulation of most other loci, in part because of altered mRNA splicing and transport (Saavedra et al., 1996; Yost and Lindquist, 1986, 1988). In plants, a set of about 100 genes, mostly encoding heat shock proteins, shows extreme chromatin accessibility at both promoter and gene body upon heat stress, consistent with their high expression (Sullivan et al., 2014). In mammals and insects, not all cells are competent to exhibit the hallmarks of the heat shock response (Dura, 1981; Morange et al., 1984); specifically, cells in early embryonic development fail to induce heat shock protein expression upon stress. Thus, we explored whether all cells within developing roots were capable of exhibiting a typical heat shock response. To do so, we applied a standard heat stress (45 min, 38ºC) to eight-day-old seedlings, harvested their roots along with roots from age- and time-matched control seedlings, and generated protoplasts for single-cell RNA-seq of both samples. Approximately 1000 cells from each sample were assessed for gene expression.

Due to global gene expression changes upon heat shock, we could not simply assign clusters and tissue types as before for control and heat-shocked cells. The genes previously used for assigning clusters and tissue types were down-regulated in heat-stressed cells, leading to different cluster calls (**Supplemental Figure 9**). To enable comparisons, we used a mutual nearest neighbor approach to embed cells conditioned on treatment in UMAP space (Haghverdi et al., 2018). This procedure yielded corresponding clusters in control and heat-shocked cells, albeit with varying cell numbers for some and loss of others (*e.g.* endodermis, see **Supplemental Figure 9**). To

compare control and heat-shocked cells across assigned clusters and tissue types, we first focused on the genes involved in the acute heat shock response that show extreme chromatin accessibility along the gene body and high expression upon heat stress. Hierarchical clustering of control and heat shock treated single-cell transcriptomes by expression of genes near dynamically accessible sites showed numerous genes whose expression was greatly increased or decreased between conditions regardless of cluster and tissue type (**Supplemental Figure 9C**). To more clearly display the dynamics of genes expression across tissue types, we grouped genes by their expression patterns across all cells and used each group to devise a signature by which to score individual cells within tissue types. This analysis confirmed that multiple tissue types display a similar induction in gene expression modules upon heat shock treatment (see **Figure 5,** gene cluster 4 and **Supplemental Figure 10**, cluster 2**)**.

Next, we broadened our analysis to genes known to have differential expression (1783 genes) and to genes residing near regulatory regions that change in accessibility upon heat stress (1730 genes) as observed in bulk analyses (Alexandre et al., 2018; Sullivan et al., 2014). Although these gene sets overlap (942 genes), they contain complementary information as changes in accessibility do not necessarily translate into altered expression, and vice versa (Alexandre et al., 2018). Differential gene analysis of our single cell RNA-seq of control and heatshock treated transcriptomes identified 752 of bulk and 564 of dynamic chromatin linked genes as varying significantly by treatment. These analyses were dominated by the canonical heat shock response, including almost all measured heat shock genes residing in cluster 4, which had consistent high expression upon heat shock (**Figure 5B**). However, the expression signature analysis revealed subtle but significant differences among some tissue types (**Figure 5B, C,** *e.g.* clusters 3 and 8). These results demonstrate both the promise and the challenges inherent in comparing single-cell data across different conditions and treatments.

Discussion

Despite its potential, single-cell RNA-seq has not heretofore been broadly applied to plants. Here, we use *A. thaliana* roots to establish both the experimental and analytic procedures for this methodology. Using Monocle 3, we assigned with high confidence over 3000 cells to expected cell and tissue types. In particular, cortex, endodermis and hair cells were identifiable. In

contrast, non-hair cells and Columella cells were challenging to distinguish, reflecting the high similarity in their expression profiles, as observed in bulk expression data (Brady et al., 2007; Cartwright et al., 2009). Some tissues like stele were difficult to dissect into individual cell types, likely due to the difficulty of digesting the cell walls of the tightly packed vascular bundle to obtain single cells (Brady et al., 2007; Cartwright et al., 2009). We identified hundreds of novel genes with cell-type-specific and tissue-type-specific expression, which may allow the generation of new marker lines for detailed genetic analyses. These genes, together with cluster-specific enriched transcription factor motifs and corresponding transcription factors, are candidates for driving differentiation and cell-type identity.

Similarly, the findings on developmental trajectories highlight the potential of single cell transcriptomics to advance a high resolution view of plant development. Plants have a continuous body plan with new cells continuously arising while older cells persist, allowing us to detect developmental trajectories without spatial information. We describe specific enrichments of transcription factor motifs for early and late hair, cortex and endodermis cells, together with their corresponding candidate transcription factors. Future analyses, which will require greater number of cells than assayed here, may include combinatorial expression of multiple transcription factor family members.

The well-established developmental trajectory for hair cells allowed us to explore the relationships of endoreduplication, transcriptional rates, and differentiation. Our findings suggest that transcriptional rates, measured as mRNA velocity, increase with increasing ploidy. However, this transcriptional increase appears to be limited to genes specifically expressed in hair cells, as overall levels of RNA decreased over pseudotime. Our observations are consistent with hair cells becoming more specialized and moving towards a terminally differentiated state over time.

Our exploration of the heat shock response with single-cell RNA-seq was driven by prior data that not all cells and tissues are equally competent to respond to stress. The identification of plant cell types that most strongly respond to abiotic stresses such as heat, drought, and nutrient starvation may enable cell-type-specific genetic manipulation to generate stress-tolerant crops

without pleiotropically affecting plant fitness and yield. Although all cells in these experiments showed gene expression changes typical of the canonical heat shock genes, we detected subtle but highly significant expression differences among cells and tissue types for other genes. Thus, single-cell transcriptomics across stress conditions holds potential for future crop breeding and genetic engineering. However, such analyses require much larger numbers of cells than currently accessible by droplet-based methods. Moreover, such analyses should focus on treatments that are less overwhelmed by a strong canonical signal to increase resolution in detecting cell-type-specific differences.

In this study, we relied on the extensive and detailed expression data for bulk *A. thaliana* cell and tissue types to establish the validity of our approaches. The overwhelming correspondence of our findings with these prior data and with other data derived from traditional molecular genetics provides confidence that less well-characterized *A. thaliana* tissues and other plants, including crops, will be amenable to these approaches to the analysis of development and environmental response.

Materials and Methods:

Plant Material and Growth Conditions. *Arabidopsis thaliana* Col-0 seedlings were grown vertically at 22°C, on 1xMS + 1% sucrose plates covered with one layer of filter paper. Seven or eight days-old seedlings (LD, 16h light/8h dark, ~100 µmol m2 s) were collected around ZT3, and the roots/shoots excised with a sharp razor blade. For the heat shock, seedling plates were transferred from 22°C to 38°C for 45 min (Conviron TC-26, light ~100 µmol m2 s), and the roots harvested immediately after.

Protoplast Isolation. Protoplast isolation was done according to Bargmann and Birnbaum (2010), with slight modifications. Briefly, 1 g of whole-roots was incubated in 10 ml of protoplasting solution for 1.5 h at 75 rpm. After passing through a 40 µm strainer, protoplasts were centrifuged at 500 g for 5 min and washed once in protoplasting solution without enzymes. Final suspension volume was adjusted to a density of 500 – 1,000 cells/µl. Protoplasts were placed on ice until further processing.

Single-cell RNA-seq protocol

Single-cell RNA-seq was performed on fresh Arabidopsis root protoplast using the10X scRNA-seq platform, the Chromium Single Cell Gene Expression Solution (10X Genomics).

Data Analysis

Estimating gene expression in individual cells

Single-cell RNA-seq reads were sequenced and then mapped to the TAIR10 Arabidopsis genome using Cellranger (version 2.1.0) (https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger). Cellranger produces as output matrices with a row for each gene and a column for each cell, in which entries denote UMIs observed. The ARAPORT gene annotation was used. For the heat shock analysis, reads from a control sample and reads from heat-shocked sample were aggregated using “cellranger aggr” to normalize libraries to equivalent number of mean reads per cell across libraries.

Running Monocle 3: Dimensionality Reduction, and Cell Clustering The output of the cellranger pipeline was parsed into R (version 3.5.0) using the cellranger R kit (version 2.0.0) and converted into a CellDataSet (cds) for further analysis using Monocle 3 alpha (version 2.99.1) (http://cole-trapnell-lab.github.io/monocle-release/monocle3/). All Monocle 3 analysis was performed on a High Performance Computing cluster using 128GB of RAM spread across 8 cores. The lower detection limit for the cds was set at 0.5, and the expression family used set to negbinomial.size().

We visualized cell clusters and trajectories using the standard Monocle workflow. Monocle internally handles all normalization needed for dimensionality reduction, visualization, and differential expression via "size factors" that control for variability in library construction efficiency across cells. After estimating the library size factors for each cell (via estimateSizeFactors), and estimating the dispersion in expression for each gene (via estimateDispersions) in the dataset, the top 1500 genes in terms of dispersion, *i.e*. 1500 genes with the most expression variability in our dataset, were selected to order the cells into clusters. The expression values of these 1500 genes for each cell were log-transformed and projected onto the first 25 principal components via Monocle's data pre-processing function (preprocessCDS). Then, these lower-dimensional coordinates were used to initialize a nonlinear manifold learning algorithm implemented in Monocle 3 called Uniform Manifold Approximation and Projection, or UMAP (via reduceDimension) (McInnes and Healy, 2018). This allows us to visualize the data unto two or three dimensions. Specifically, we projected onto 2 components using the cosine distance metric, setting the parameters "n_neighbors" to 50, and "min_dist" to 0.1.

The Louvain method was used to detect cell clusters in our two dimensional representation of the dataset (partitionCells); this resulted in 11 cell clusters, or Louvain components. Cells were then clustered into "super" groups using a method derived from "approximate graph abstraction" (Wolf et al., 2018) and for each super group, a cell trajectory was drawn atop the projection using Monocle's reversed graph embedding algorithm, which is derived from "SimplePPT" (learnGraph) (Mao et al., 2015). This yielded 6 cell trajectories.

Identifying Cell Types

In order to categorize the cells into cell types and to apply developmental information, a deconvolved root expression map was downloaded from AREX LITE: The Arabidopsis Gene Expression Database (http://www.arexdb.org/data/decondatamatrix.zip). Using this data matrix, the Spearman's rank correlation was calculated between each cell in our dataset and each cell type and longitudinal annotation in the data matrix (3121 x 128 Spearman's rank correlations total). Specifically, we looked at the correlation of 1229 highly variable genes in our dataset. These 1229 genes represents the overlap between our 1500 highly variable genes and genes in the root expression map data matrix. Cells in our dataset were assigned a cell type and a developmental label based on the annotation with which each cell had the highest correlation. (*i.e.* if a cell correlated highest with endodermis cells in longitudinal zone 11, then it would be called as endodermis_11).

In addition to using the Spearman's rank correlation to assign cells their cell type, a set of known marker genes derived from GFP marker lines of the Arabidopsis root were used to identify cell types based on the high gene expression of these marker genes. These genes were obtained from (Brady et al., 2007; Cartwright et al., 2009). Specifically Supplemental Table 2 (Cartwright et al., 2009) was used. For the analysis comparing bulk RNA and pseudo bulk scRNA-seq data, the bulk data was obtained from Li et al. 2016 (Li et al., 2016); specifically, we used Table S5 from this study. Isoforms of each gene were averaged in order to be comparable to the pseudo bulk data.

Running Monocle 3: Identifying High Specificity Genes

In order to identify differentially expressed genes between cell clusters the moran's I test was performed on our UMAP (principalGraphTest), with the projection being broken up into 25 x 25 spatial units. Then marker genes were identified for each cluster, and each annotated grouping of clusters using a moran's I threshold of 0.1 and a qval threshold of 0.05. In order for a gene to be considered highly specific, it must have had a specificity rating of above 0.7.

Transcription factor motif analysis

Highly specific genes were identified for each cell cluster, and their promoters were analyzed for presence of transcription factor motifs. Promoters were defined as 500 base pairs upstream of the

start site of each gene. Instances of each motif were identified using (Grant et al., 2011) at a p-value cutoff of 1e-5 for each match. The input position weight matrices for each motif were enumerated in a previous study of binding preferences for nearly all Arabidopsis transcription factors (O'Malley et al., 2016). Motif frequencies in genes specific to each cell cluster were compared to a background set of motif frequencies across all promoters in the Arabidopsis genome to determine a log2 enrichment score. TF family genes were pulled from the gene family page of TAIR10 (https://www.arabidopsis.org/browse/genefamily/index.jsp).

Running Monocle 3: Assigning Pseudotime

Since cluster 10, which predominately contained cortex cells did not have a trajectory, those cells were reclustered using the UMAP settings previously described except the number of neighboring points used to approximate local manifold structure was decreased to 25.

Pseudotime analysis requires the selection of a cell as an "origin" for the pseudotime trajectory. Origin assignment was based on the Spearman's rank assignments for each cell. The following cells were used as origins for their respective cell type trajectories: cortex_2, hair_2, endodermis_2, nonHair_3.

Calculating significance with the Permutation Test

The permutation test was used to calculate the significance of the observed trends that the total mRNA of hair marker genes and hair specific genes increased as pseudotime increased in hair cells. To do this, 10000 random samplings of 441 genes (the number of hair marker genes), and 201 genes (the number of hair specific genes) were taken respectively. Next, the median total mRNA was calculated across pseudotime for each random sampling and the slope of this data was calculated using a generalized linear model. The observed slope of the marker genes and the hair specific genes was compared to the distribution of slopes generated by 10000 random samplings. No random sampling of genes had a slope that was higher than the observed slopes generated by the hair marker genes or the hair specific genes. The significance, or the p-value, of the trend seen in the hair marker genes and the hair specific genes can then be calculated simply as the proportion of sampled permutations that have a slope that is equal to or greater than slope generated by our genes of interest. This gives us a p-value of 1/10001 or roughly 1×10^{-4}.

Calculating RNA velocity

We used the Velocyto R and Python packages (version 0.6 and 0.17, respectively) to estimate RNA velocity for root hair cells (La Manno et al., 2018). Matrices of spliced and unspliced RNA counts were generated from Cellranger outputs using velocyto.py CLI and "run10x" defaults. We followed the velocyto.py and velocyto.R manuals (http://velocyto.org/) and used spliced (emat) and unspliced (nmat) matrices to estimate RNA velocity. With predefined cell type annotations, we performed gene filtering with the parameter "min.max.cluster.average" set to 0.2 and 0.05 for “emat” and “nmat”, respectively. RNA velocity using the selected 996 genes was estimated with the defaults to the function gene.relative.velocity.estimates() except parameters "kCells" and "fit.quantile" which were set to 5 and 0.05, respectively. Velocity measurements for each cell were calculated as the difference between "$projected" and "$current" (with $deltaT = 1) results from the estimated velocity output.

Analysis of heat shocked data

Differentially expressed genes as a function of heatshock treatment where identified using the differentialGeneTest function in Monocle specifying a full model of Treatment*UMAP cluster and a residual model of UMAP cluster. Hierarchical clustering of these genes across control and heatshock treated cells was performed using the pheatmap function in the pheatmap R package (version 1.0.10) specifying ward.D2 as the clustering method. Genes with similar dynamics across treatment and cell types were recovered using the cutree function from the stats package in R specifying k = 8 for both DHS linked genes and bulk differentially expressed genes. To generate signatures from these 8 groups of clustered genes we log normalized expression values using a pseudocount of 1 and for each cell calculated the mean normalized expression value across genes that belong to one of the 8 gene cluster.

Data Availability

All sequencing data can be found on GEO at:

Acknowledgments

We thank members of the Trapnell Lab for input and discussion. This work was supported by NSF MCB-1516701 to CQ and NSF RESEARCH-PGR 1748843 to CQ and SF. KJB was supported by the University of Washington NIH Big Data for Genomics and Neuroscience Training Grant (T32LM012419).

References

Alexandre, C.M., Urton, J.R., Jean-Baptiste, K., Huddleston, J., Dorrity, M.W., Cuperus, J.T., Sullivan, A.M., Bemm, F., Jolic, D., Arsovski, A.A.*, et al.* (2018). Complex Relationships between Chromatin Accessibility, Sequence Divergence, and Gene Expression in Arabidopsis thaliana. Mol Biol Evol *35*, 837-854.

Bennett, T., van den Toorn, A., Sanchez-Perez, G.F., Campilho, A., Willemsen, V., Snel, B., and Scheres, B. (2010). SOMBRERO, BEARSKIN1, and BEARSKIN2 regulate root cap maturation in Arabidopsis. Plant Cell *22*, 640-654.

Bhosale, R., Boudolf, V., Cuevas, F., Lu, R., Eekhout, T., Hu, Z., van Isterdael, G., Lambert, G., Xu, F., Nowack, M.K.*, et al.* (2018). A spatiotemporal DNA endoploidy map of the Arabidopsis root reveals roles for the endocycle in root development and stress adaptation. Plant Cell.

Birnbaum, K., Shasha, D.E., Wang, J.Y., Jung, J.W., Lambert, G.M., Galbraith, D.W., and Benfey, P.N. (2003). A gene expression map of the Arabidopsis root. Science *302*, 1956-1960.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment *2008*.

Bourdon, M., Pirrello, J., Cheniclet, C., Coriton, O., Bourge, M., Brown, S., Moise, A., Peypelut, M., Rouyere, V., Renaudin, J.P.*, et al.* (2012). Evidence for karyoplasmic homeostasis during endoreduplication and a ploidy-dependent increase in gene transcription during tomato fruit growth. Development *139*, 3817-3826.

Brady, S.M., Orlando, D.A., Lee, J.Y., Wang, J.Y., Koch, J., Dinneny, J.R., Mace, D., Ohler, U., and Benfey, P.N. (2007). A high-resolution root spatiotemporal map reveals dominant expression patterns. Science *318*, 801-806.

Cao, J., Packer, J.S., Ramani, V., Cusanovich, D.A., Huynh, C., Daza, R., Qiu, X., Lee, C., Furlan, S.N., Steemers, F.J.*, et al.* (2017). Comprehensive single-cell transcriptional profiling of a multicellular organism. Science *357*, 661-667.

Cartwright, D.A., Brady, S.M., Orlando, D.A., Sturmfels, B., and Benfey, P.N. (2009). Reconstructing spatiotemporal gene expression data from partial observations. Bioinformatics *25*, 2581-2587.

Dura, J.M. (1981). Stage dependent synthesis of heat shock induced proteins in early embryos of Drosophila melanogaster. Mol Gen Genet *184*, 381-385.

Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics *27*, 1017-1018.

Haghverdi, L., Lun, A.T.L., Morgan, M.D., and Marioni, J.C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol *36*, 421-427.

Irish, V.F. (1991). Cell lineage in plant development. Curr Opin Cell Biol *3*, 983-987.

Katayama, H., Iwamoto, K., Kariya, Y., Asakawa, T., Kan, T., Fukuda, H., and Ohashi-Ito, K. (2015). A Negative Feedback Loop Controlling bHLH Complexes Is Involved in Vascular Cell Division and Differentiation in the Root Apical Meristem. Curr Biol *25*, 3144-3150.

Kieffer, M., Master, V., Waites, R., and Davies, B. (2011). TCP14 and TCP15 affect internode length and leaf shape in Arabidopsis. Plant J *68*, 147-158.

Lachowiec, J., Lemus, T., Thomas, J.H., Murphy, P.J., Nemhauser, J.L., and Queitsch, C. (2013). The protein chaperone HSP90 can facilitate the divergence of gene duplicates. Genetics *193*, 1269-1277.

Lachowiec, J., Mason, G.A., Schultz, K., and Queitsch, C. (2016). Redundancy, feedback, and robustness in the Arabidopsis thaliana BZR/BEH gene family. bioRxiv.

Lachowiec, J., Mason, G.A., Schultz, K., and Queitsch, C. (2018). Redundancy, feedback, and robustness in the *Arabidopsis thaliana BZR/BEH* gene family. bioRxiv 053447; doi: https://doiorg/101101/053447.

Lan, X., and Pritchard, J.K. (2016). Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. Science *352*, 1009-1013.

Lee, C., Teng, Q., Zhong, R., and Ye, Z.H. (2012). Arabidopsis GUX proteins are glucuronyltransferases responsible for the addition of glucuronic acid side chains onto xylan. Plant Cell Physiol *53*, 1204-1216.

Li, S., Yamada, M., Han, X., Ohler, U., and Benfey, P.N. (2016). High-Resolution Expression Map of the Arabidopsis Root Reveals Alternative Splicing and lincRNA Regulation. Dev Cell *39*, 508-522.

Lucero, L.E., Uberti-Manassero, N.G., Arce, A.L., Colombatti, F., Alemano, S.G., and Gonzalez, D.H. (2015). TCP15 modulates cytokinin and auxin responses during gynoecium development in Arabidopsis. Plant J *84*, 267-282.

Mao, Q., Yang, L., Wang, L., Goodison, S., and Sun, Y. (2015). SimplePPT: A Simple Principal Tree Algorithm Proceedings of the 2015 SIAM International Conference on Data Mining, 792–800.

McInnes, L., and Healy, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv e-prints 180203426.

Morange, M., Diu, A., Bensaude, O., and Babinet, C. (1984). Altered expression of heat shock proteins in embryonal carcinoma and mouse early embryonic cells. Mol Cell Biol *4*, 730-735.

Mortimer, J.C., Miles, G.P., Brown, D.M., Zhang, Z., Segura, M.P., Weimar, T., Yu, X., Seffen, K.A., Stephens, E., Turner, S.R.*, et al.* (2010). Absence of branches from xylan in Arabidopsis gux mutants reveals potential for simplification of lignocellulosic biomass. Proc Natl Acad Sci U S A *107*, 17409-17414.

O'Malley, R.C., Huang, S.C., Song, L., Lewsey, M.G., Bartlett, A., Nery, J.R., Galli, M., Gallavotti, A., and Ecker, J.R. (2016). Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. Cell *165*, 1280-1292.

Ohashi-Ito, K., Saegusa, M., Iwamoto, K., Oda, Y., Katayama, H., Kojima, M., Sakakibara, H., and Fukuda, H. (2014). A bHLH complex activates vascular cell division via cytokinin action in root apical meristem. Curr Biol *24*, 2053-2058.

Packer, J., and Trapnell, C. (2018). Single-Cell Multi-omics: An Engine for New Quantitative Models of Gene Regulation. Trends Genet *34*, 653-665.

Peng, Y., Chen, L., Lu, Y., Wu, Y., Dumenil, J., Zhu, Z., Bevan, M.W., and Li, Y. (2015). The ubiquitin receptors DA1, DAR1, and DAR2 redundantly regulate endoreduplication by modulating the stability of TCP14/15 in Arabidopsis. Plant Cell *27*, 649-662.

Petricka, J.J., Winter, C.M., and Benfey, P.N. (2012). Control of Arabidopsis root development. Annu Rev Plant Biol *63*, 563-590.

Qiu, X., Hill, A., Packer, J., Lin, D., Ma, Y.A., and Trapnell, C. (2017a). Single-cell mRNA quantification and differential analysis with Census. Nat Methods *14*, 309-315.

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H.A., and Trapnell, C. (2017b). Reversed graph embedding resolves complex single-cell trajectories. Nat Methods *14*, 979-982.

Resentini, F., Felipo-Benavent, A., Colombo, L., Blazquez, M.A., Alabadi, D., and Masiero, S. (2015). TCP14 and TCP15 mediate the promotion of seed germination by gibberellins in Arabidopsis thaliana. Mol Plant *8*, 482-485.

Riechmann, J.L., Heard, J., Martin, G., Reuber, L., Jiang, C., Keddie, J., Adam, L., Pineda, O., Ratcliffe, O.J., Samaha, R.R.*, et al.* (2000). Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. Science *290*, 2105-2110.

Rogers, E.D., and Benfey, P.N. (2015). Regulation of plant root system architecture: implications for crop advancement. Curr Opin Biotechnol *32*, 93-98.

Russell, A.B., Trapnell, C., and Bloom, J.D. (2018). Extreme heterogeneity of influenza virus infection in single cells. Elife *7*.

Saavedra, C., Tung, K.S., Amberg, D.C., Hopper, A.K., and Cole, C.N. (1996). Regulation of mRNA export in response to stress in Saccharomyces cerevisiae. Genes Dev *10*, 1608-1620.

Sullivan, A.M., Arsovski, A.A., Lempe, J., Bubb, K.L., Weirauch, M.T., Sabo, P.J., Sandstrom, R., Thurman, R.E., Neph, S., Reynolds, A.P.*, et al.* (2014). Mapping and dynamics of regulatory DNA and transcription factor networks in A. thaliana. Cell Rep *8*, 2015-2030.

Tan, T.T., Endo, H., Sano, R., Kurata, T., Yamaguchi, M., Ohtani, M., and Demura, T. (2018). Transcription Factors VND1-VND3 Contribute to Cotyledon Xylem Vessel Formation. Plant Physiol *176*, 773-789.

Tatematsu, K., Nakabayashi, K., Kamiya, Y., and Nambara, E. (2008). Transcription factor AtTCP14 regulates embryonic growth potential during seed germination in Arabidopsis thaliana. Plant J *53*, 42-52.

Trapnell, C. (2015). Defining cell types and states with single-cell genomics. Genome Res *25*, 1491-1498.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., and Rinn, J.L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol *32*, 381-386.

Winter, D., Vinegar, B., Nahal, H., Ammar, R., Wilson, G.V., and Provart, N.J. (2007). An "Electronic Fluorescent Pictograph" browser for exploring and analyzing large-scale biological data sets. PLoS One *2*, e718.

Wolf, F.A., Hamey, F., Plass, M., Solana, J., Dahlin, J.S., Gottgens, B., Rajewsky, N., Simon, L., and Theis, F.J. (2018). Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. bioRxiv, doi: https://doiorg/101101/208819

Yost, H.J., and Lindquist, S. (1986). RNA splicing is interrupted by heat shock and is rescued by heat shock protein synthesis. Cell *45*, 185-193.

Yost, H.J., and Lindquist, S. (1988). Translation of unspliced transcripts after heat shock. Science *242*, 1544-1548.

Zhu, D., Wu, Z., Cao, G., Li, J., Wei, J., Tsuge, T., Gu, H., Aoyama, T., and Qu, L.J. (2014). TRANSLUCENT GREEN, an ERF family transcription factor, controls water balance in Arabidopsis by activating the expression of aquaporin genes. Mol Plant *7*, 601-615.

Figure Legends:

Figure 1. **Annotation of cell and tissue types for single-cell RNA-seq of whole *A. thaliana* roots.** (**A**) Root cells were clustered and projected onto two-dimensional space with UMAP (McInnes and Healy, 2018). Filled circles represent individual cells; colors represent their respective Louvain component. Monocle 3 trajectories (black lines) are shown for clusters in which a trajectory could be identified. (**B**) Filled circles represent individual cells; colors indicate cell and tissue type based on highest Spearman's rank correlation with tissue-specific bulk expression data (Brady et al., 2007; Cartwright et al., 2009). (**C**) Known marker genes (Brady et al., 2007; Cartwright et al., 2009) were used to cluster single-cell gene expression profiles based on similarity. The expression of 530 known marker genes was grouped into 7 clusters, using k-means clustering. Mean expression for each cluster (rows) is presented for each cell (columns). Cells were ordered by their respective Louvain component indicated above by color (see **A,** starting at component 1 at left). Number of genes in each cluster is denoted at right. (**D**) Single-cell RNA-seq pseudo-bulked expression data are compared to bulk expression data of whole roots (Li et al., 2016). (**E**) Single-cell pseudo-bulk expression data are compared to bulk-expression data of the three developmental regions of the *A. thaliana* root (Li et al., 2016). (**F**) Proportions of cells as annotated by either UMAP (in **A**) or Spearman's rank correlation (in **B**) are compared to proportions determined by microscopy (Brady et al., 2007; Cartwright et al., 2009).

Figure 2. **Novel cluster-specific and tissue-specific genes and enriched transcription factor motifs.** (**A**) Proportion of cells (circle size) and mean expression (circle color) of genes with cluster-specific and tissue-specific expression are shown, beginning with known marker genes labeled with their common name (right) and their systematic name (left). For novel genes, the top significant cluster-specific genes are shown, followed by the top significant tissue-specific genes; both were identified by principal graph tests (Moran's I) as implemented in Monocle 3. Note the correspondence between Louvain components and cell and tissue types. For all novel cluster-specific and tissue-specific genes, see **Supplemental Table 3**. (**B**) Enrichments of known transcription factor motifs (O'Malley et al., 2016) 500 bp upstream of genes with cluster-specific expression compared to genome background. Motifs are specific to transcription factor gene

families rather than individual genes. The plot is clustered based on similarity in enrichments with Louvain components and cell and tissue types (filled circles) indicated.

Figure 3. Developmental trajectory of hair cells. (**A**) UMAP-clustered hair cells were assigned a developmental time point based on highest Spearman's rank correlation with bulk expression data of staged tissue (13 developmental stages) (Brady et al., 2007; Cartwright et al., 2009). Cell type and developmental time point are indicated in shades of blue (and pink). (**B**) Cells were ordered in pseudotime; columns represent cells, rows represent expression of the 1500 genes. Rows were grouped based on similarity in gene expression, resulting in 6 clusters (indicated left), with genes in clusters 2 and 5 expressed early in pseudotime and genes in cluster 1 expressed late. Hair cells with the earliest developmental signal (Brady et al., 2007; Cartwright et al., 2009) were designated as the root of the trajectory. The graph above represents the average best-correlation of developmental stage (Brady et al., 2007; Cartwright et al., 2009) in a scrolling window of 20 cells with pseudotime, showing the expected increase in developmental age with increasing pseudotime. (**C**) Examples of an early and a late expressed hair-cell-specific gene. Gene expression in each cell is superimposed onto the UMAP cluster and trajectory, with lighter colors indicating higher gene expression. **(D)** Total median RNA captured in cells across pseudotime decreases across pseudotime. Number of genes included is indicated. **(E)** Comparison of total RNA for hair-cell-specific genes (in red) to a comparable random set of genes (in blue). Number of genes is indicated (Permutation test p-value $\approx 1 \times 10^{-4}$). (**F**) Different transcription factor motifs reside in the 500 bp upstream regions of genes expressed early (clusters 2, 5) compared to genes expressed late (cluster 1). Transcription factor motifs specific to early hair cells are denoted with blue bars, those for late hair cells with green bars; bar length indicates motif frequency. Thresholds on either side (grey box, dotted lines) refer to 1.5 standard deviation above mean motif frequency. (**G**) Expression of individual members of transcription factors families highlighted in **D** across pseudotime identifies candidate factors driving early or late gene expression.

Figure 4. Developmental trajectory of cortex cells. (**A**) Cortex cells were re-clustered to create a trajectory, in which each cell was assigned a developmental time point and identity (shades of yellow, brown, pink) based on the highest correlation of a cell's gene expression with prior data

(Brady et al., 2007; Cartwright et al., 2009). (**B**). Comparison of pseudo-bulk expression data from cells annotated as cortex cells with bulk expression data from protoplasts sorted for expression of the cortex marker gene *COR* (Li et al., 2016). (**C**) Cells were ordered in pseudotime; columns indicate cells and rows the expression of the 1500 genes. Rows were grouped based on similarity in gene expression, resulting in 6 clusters (indicated left), with genes in clusters 2 and 3 expressed early in pseudotime and genes in cluster 1 expressed late. Cortex cells with the earliest developmental signal (Brady et al., 2007; Cartwright et al., 2009) were designated as the root of the trajectory. The graph above represents the average best-correlation of developmental stage (Brady et al., 2007; Cartwright et al., 2009) in a scrolling window of 20 cells with pseudotime, showing the expected increase in developmental age with increasing pseudotime. (**D**) Examples of an early and a late expressed novel cortex-cell-specific gene. Gene expression in each cell is superimposed onto the UMAP cluster and trajectory, with lighter colors indicating higher gene expression. (**E**) Different transcription factor motifs reside in the 500 bp upstream regions of genes expressed early (clusters 2, 3) compared to genes expressed late (cluster 1). Transcription factor motifs specific to early cortex cells are denoted with blue bars, those for late cortex cells with green bars, bar length indicates motif frequency. Thresholds on either side (grey box, dotted lines) refer to 1.5 standard deviation above mean motif frequency. (**F**) Expression of individual members of transcription factors families highlighted in **D** across pseudotime identifies candidate factors driving early or late gene expression.

Figure 5. Single cell RNA-seq highlights canonical and novel aspects of the heat shock response. (**A**) A nearest neighbor approach used on control and heat-shocked cells yields largely comparable UMAP clusters. (**B**) Single-cell expression values for genes differentially regulated in response to a similar heat shock in bulk tissue (Alexandre et al., 2018) were clustered based on similarity in gene expression in control versus heat-shocked cells. Clusters are indicated at right. Cluster 4 contains the vast majority of acute heat-shock induced genes, for which chromatin accessibility extends along the gene body upon heat shock (Sullivan et al., 2014). See **Supplemental Figure 9** for a similar analysis considering genes residing near chromatin accessible sites that are dynamic in response to heat shock (Sullivan et al., 2014). (**C**) Pairwise comparisons of cell and tissue types reveal subtle but highly significant differences in the heat shock response for some clusters. Significance was determined with a generalized linear model.

Shown are data for clusters 3 and 8. There is no cell or tissue-type specific response for cluster 4, as all cells show the canonical heat stress response typical of genes in this cluster.

Supplemental Figure Legends

Supplemental Figure 1. **General tissue and data features**. (**A**) Image of seven-day-old roots used for protoplast generation. (**B**) Dispersion of gene expression across cells is compared to average gene expression. The 1500 genes marked in black were used for further analysis. (**C**) Principal components are plotted against explained variance in gene expression; 25 components were selected. (**D**) Total nuclear mRNA per Louvain component is shown. Louvain components are colored as in **Figure 1A**.

Supplemental Figure 2. Marker gene expression. Known marker gene expression (Brady et al., 2007; Cartwright et al., 2009) was clustered based on expected respective cell types and tissues; mean gene expression is shown for each cell. Cells are clustered by Louvain components across the x-axis.

Supplemental Figure 3. Example gene expression. Expression of several example genes is shown superimposed onto cells within the UMAP derived map; shades of green to yellow indicate expression levels, with lighter color indicating higher expression; cells without detectable expression are in grey. Examples include known and novel genes with tissue-specific expression such as *COBL9*, *SCR*, *GL2*, *WER*, *SPL15* and *AT2G45180*.

Supplemental Figure 4. Transcription factor family expression patterns. (**A**) Louvain and (**B**) tissue-specific expression for individual members of the *BEH/BZR* transcription factor family. Circle size indicates proportions of cells expressing a given gene; circle color indicates expression level, with lighter colors corresponding to higher expression. (**C**) *BEH/BZR* gene family members co-expressed in higher numbers of cells are indicated with lighter colors. Specific patterns are identified for several genes, *e.g. BES1* (AT1G75080) and *BEH2* (AT4G36780). (**D**) Louvain and (**E**) tissue-specific expression for the TCP transcription factor family. (**F**) TCP gene family members co-expressed in higher numbers of cells are indicated with lighter colors. Specific patterns are identified for several TCP members, especially for *TCP15*

(AT1G69690) and *TCP14* (AT3G47620), two genes specifically expressed in stele. (**G**) Louvain and (**H**) tissue-specific expression for the NAC transcription factor family. (**I**) NAC gene family members co-expressed in higher numbers of cells are indicated with lighter colors. Specific patterns are identified for several NAC members, especially between *ATAF1* (AT1G01720) and *ATAF2* (AT5G08790).

Supplemental Figure 5. Developmental and tissue Spearman's rank correlations. Spearman's rank correlation coefficients for each cell's expression were determined for each tissue and developmental stage (Brady et al., 2007; Cartwright et al., 2009). Development stages are ordered left to right, youngest to oldest. Not all cell and tissue types are represented in every developmental stage; *e.g.* Columella is represented only in stages 1 and 2, lateral root primordia are represented only in stage 12, and phloem is represented from stages 3 to 13. All others are represented from stages 2 to 13.

Supplemental Figure 6. Changes in transcription across hair cell development. (A) Comparison of total RNA captured for hair-cell-specific genes (in red) versus a comparable set of random genes (in blue). Median total RNA is plotted; standard deviation is shaded in respective colors. Expression of hair-cell-specific genes differed significantly (Permutation test p-value $\approx 1 \times 10^{-4}$). **(B)** Using published predictions of ploidy in hair cells (Bhosale et al., 2018), and our best rank assignment of developmental time point, we assigned each cell a ploidy and compared RNA velocity across these differing predicted ploidy (Tukeys multiple comparsion p-value = 0.0477). **(C)** RNA velocity, which assesses the ratio of spliced to unspliced transcripts, was measured across pseudotime and shows increases over pseudotime.

Supplemental Figure 7. Developmental trajectory of endodermal cells. (**A**) UMAP-clustered endodermal cells were assigned a developmental time point based on highest Spearman's rank correlation with bulk expression data of staged tissue (13 developmental stages) (Brady et al., 2007; Cartwright et al., 2009). Cell type and developmental time point are indicated in shades of magenta. (**B**) Cells were ordered in pseudotime; columns represent cells, rows represent expression of the 1500 genes. Rows were grouped based on similarity in gene expression, resulting in 6 clusters (indicated left), with genes in clusters 1 and 2 expressed early in

pseudotime and genes in clusters 3 and 4 expressed late. Endodermal cells with the earliest developmental signal (Brady et al., 2007; Cartwright et al., 2009) were designated as the root of the trajectory. The graph above represents the average best-correlation of developmental stage (Brady et al., 2007; Cartwright et al., 2009) in a scrolling window of 20 cells with pseudotime, showing the expected increase in developmental age with increasing pseudotime. (**C**) Examples of an early and a late expressed endodermis-specific gene; gene expression in each cell is superimposed onto the UMAP cluster, with lighter colors indicating higher gene expression. (**D**) Different transcription factor motifs reside in the 500 bp upstream regions of genes expressed early (clusters 1, 2) compared to genes expressed late (clusters 3, 4). Transcription factor motifs specific to early hair cells are denoted with blue bars, those for late hair cells with green bars; bar length indicates motif frequency. Thresholds on either side (grey box, dotted lines) refer to 1.5 standard deviation above mean motif frequency.(**E**) Expression of individual members of transcription factors families highlighted in **D** across pseudotime identifies candidate factors driving early or late gene expression.

Supplemental Figure 8. **Developmental expression of individual transcription factors.** Expression of individual members of transcription factor families with enriched early (indicated with blue bar) or late motifs (indicated in green) for hair, cortex, and endodermis cells. Lighter colors indicate higher expression; x-axis denotes pseudotime progressing from early to late as left to right.

Supplemental Figure 9. **Heat shock clustering and expression profiling** (**A**) Conventional Monocle clustering and UMAP projection shows that cells are primarily clustered by treatment (control or heat shock) rather than cell or tissue type. (**B**) A mutual nearest neighbor approach takes treatment into account to generate comparable clusters; however, some clusters change in size, become ambiguous, or are lost. **(C)** Heatmap of expression of known gene-body accessible upon heat shock genes for each cell, hierarchically clustered based on similarity, resulting in little Louvain cluster grouping of cells suggesting these genes do not have a cell type specific signature. (**D**) Amongst the ambiguous clusters, non-hair and Columella cells are the least reliable as they no longer co-cluster as early non-hair cells; rather, their second best correlation is not to different developmental stage of the same tissue but to a different tissue.

Supplemental Figure 10. **Conditional expression in genes with dynamic chromatin accessibility during heat shock** (**A**) Single-cell expression values for genes previously identified as residing near chromatin accessible sites that are dynamic in response to heat shock (Sullivan et al., 2014) were clustered based on similarity in gene expression in control versus heat-shocked cells. Clusters are indicated at right. Cluster 2 contains the vast majority of acute heat-stress induced genes, for which chromatin accessibility extends along the gene body upon heat shock (Sullivan et al., 2014). **(B)** Pairwise comparisons of cell and tissue types reveal subtle but highly significant differences in the heat shock response for clusters from genes associated with dynamic chromatin accessibility in (A). (**C**) Pairwise comparisons of cell and tissue types reveal subtle but highly significant differences in the heat shock response for some additional clusters from differentially expressed genes from Figure 5. Significance was determined with a generalized linear model (Supplemental Table 6).

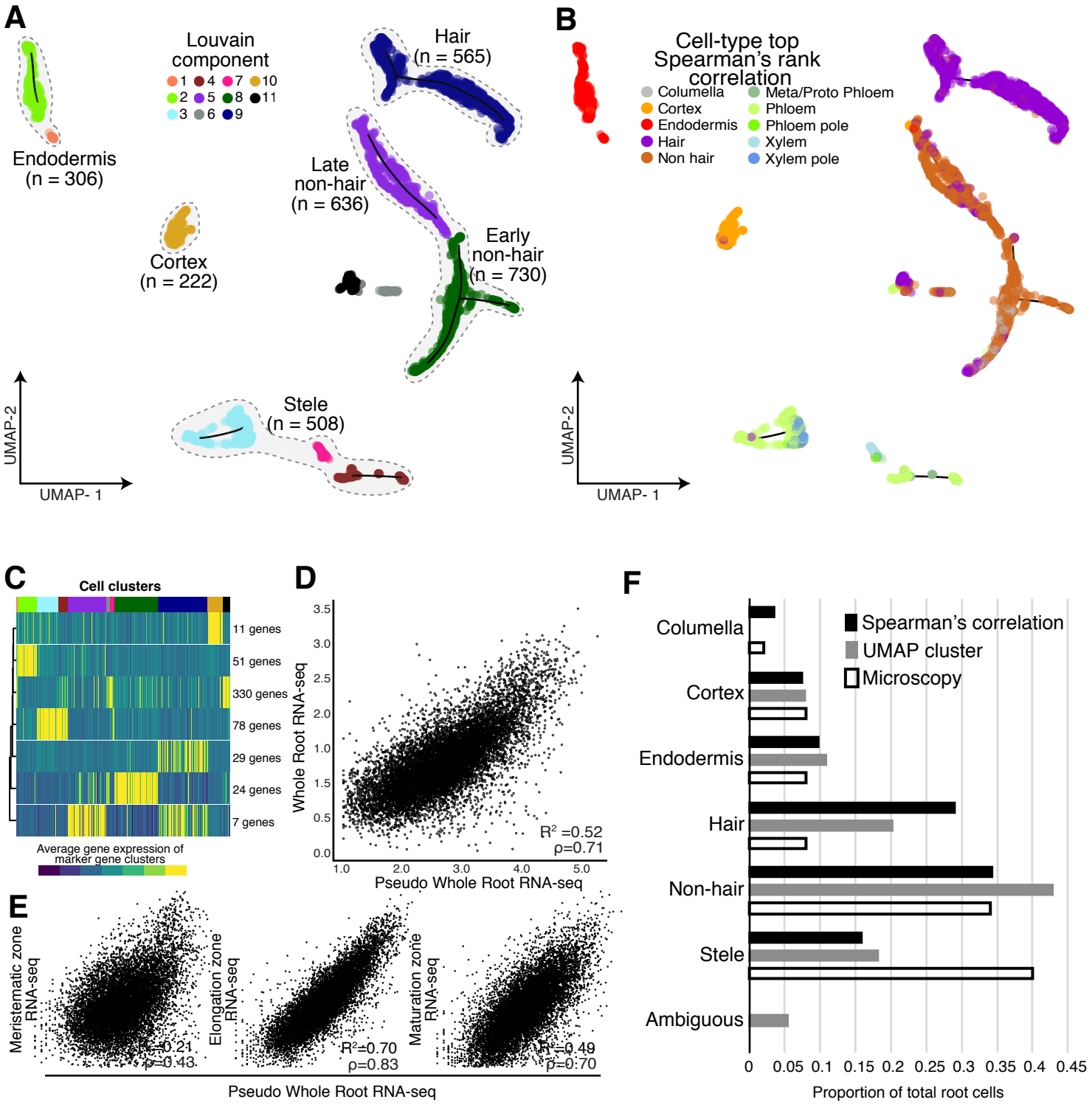


Figure 1. Annotation of cell and tissue types for single-cell RNA-seq of whole A. thaliana roots. (A) Root cells were clustered and projected onto two-dimensional space with UMAP (McInnes and Healy, 2018). Filled circles represent individual cells; colors represent their respective Louvain component. Monocle 3 trajectories (black lines) are shown for clusters in which a trajectory could be identified (B) Filled circles represent individual cells; colors indicate cell and tissue type based on highest Spearman's rank correlation with previously published tissue-specific bulk expression data (Brady et al., 2007; Cartwright et al., 2009). (C) Using known marker genes (Brady et al., 2007; Cartwright et al., 2009), single-cell gene expression profiles were clustered based on similarity. The expression of 530 known marker genes was grouped into 7 clusters, using k-means clustering. Mean expression for each cluster (rows) is presented for each cell (columns), cells were ordered by their respective Louvain component indicated above by color (see A, starting at component 1 at left). Number of genes in each cluster is denoted at right. (D) Single-cell RNA-seq pseudo-bulked expression data is compared to bulk expression data of whole roots (Li et al., 2016). (E) Single-cell pseudo-bulk expression data is compared to bulk-expression data of the three developmental regions of the A. thaliana root (Li et al., 2016). (F) Proportions of cells as annotated by either UMAP (in A) or Spearman's rank correlation (in B) are compared to proportions determined by microscopy in a previous study (Brady et al., 2007; Cartwright et al., 2009).

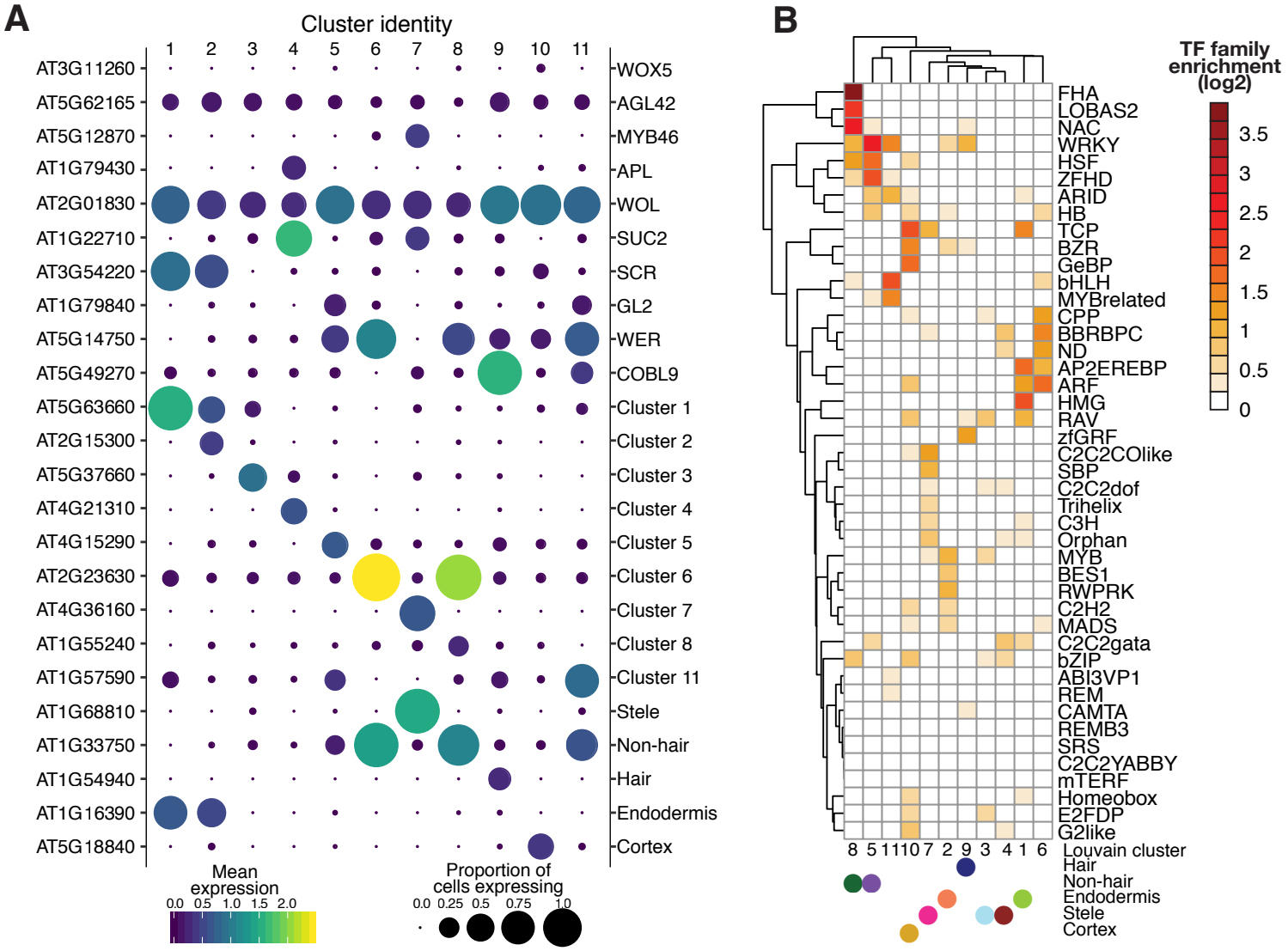


Figure 2. Novel cluster- and tissue-specific genes and enriched transcription factor motifs. (A) Proportion of cells (circle size) and mean expression (circle color) of genes with cluster- and tissue-specific expression are shown, beginning with known marker genes labeled with their common name (right) and their systematic name (left). For novel genes, the top significant cluster-specific genes are shown, followed by the top significant tissue-specific genes; both were identified by principal graph tests (Moran's I) as implemented in Monocle 3. Note the correspondence between Louvain components and cell and tissue types. For all novel cluster-and tissue-specific genes see Supplemental Table 3. (B) Enrichments of known transcription factor motifs (O'Malley et al., 2016) 500bp upstream of genes with cluster-specific expression compared to genome background. Motifs are specific to transcription factor gene families rather than individual genes. Plot is clustered based on similarity in enrichments with Louvain components and cell and tissue types (filled circles) indicated.

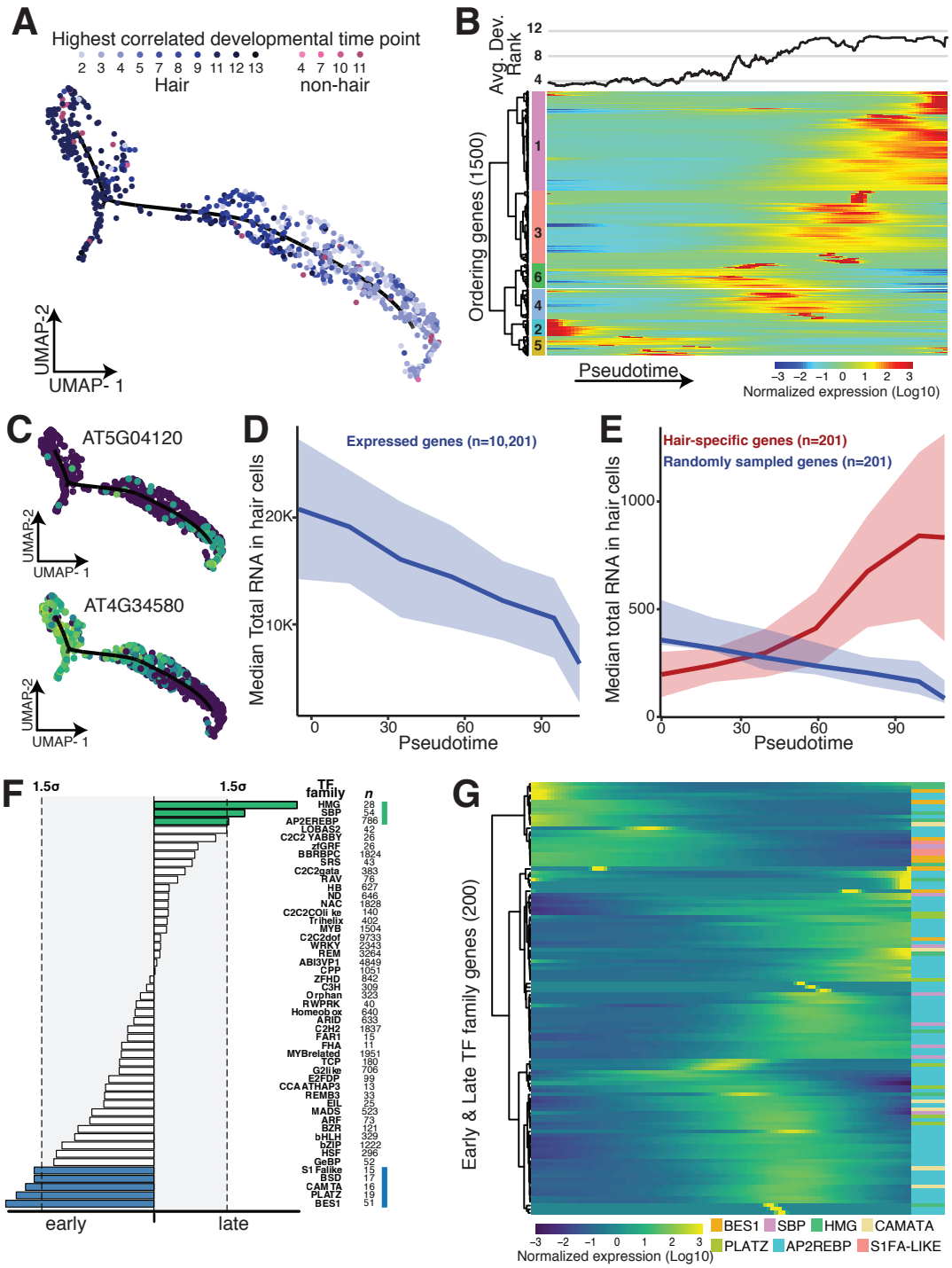


Figure 3. Developmental trajectory of hair cells. (A) UMAP-clustered hair cells were assigned a developmental time point based on highest Spearman's rank correlation with bulk expression data of staged tissue (13 developmental stages) (Brady et al., 2007; Cartwright et al., 2009). Cell type and developmental time point are indicated in shades of blue (and pink). (B) Cells were ordered in pseudotime; columns represent cells, rows represent expression of the 1500 genes. Rows were grouped based on similarity in gene expression, resulting in 6 clusters (indicated left), with genes in clusters 2 and 5 expressed early in pseudotime and genes in cluster 1 expressed late. Hair cells with the earliest developmental signal (Brady et al., 2007; Cartwright et al., 2009) were designated as the root of the trajectory. The graph above represents the average best-correlation of developmental stage (Brady et al., 2007; Cartwright et al., 2009) in a scrolling window of 20 cells with pseudotime, showing the expected increase in developmental age with increasing pseudotime. (C) Examples of an early and a late expressed hair-cell-specific gene. Gene expression in each cell is superimposed onto the UMAP cluster and trajectory, with lighter colors indicating higher gene expression. (D) Total median RNA captured in cells across pseudotime decreases across pseudotime. Number of genes included is indicated. (E) Comparison of total RNA for hair-cell-specific genes (in red) to a comparable random set of genes (in blue). Number of genes is indicated (Permutation test p-value ≈ 1 x 10-4). (F) Different transcription factor motifs reside in the 500 bp upstream regions of genes expressed early (clusters 2, 5) compared to genes expressed late (cluster 1). Transcription factor motifs specific to early hair cells are denoted with blue bars, those for late hair cells with green bars; bar length indicates motif frequency. Thresholds on either side (grey box, dotted lines) refer to 1.5 standard deviation above mean motif frequency. (G) Expression of individual members of transcription factors families highlighted in D across pseudotime identifies candidate factors driving early or late gene expression.

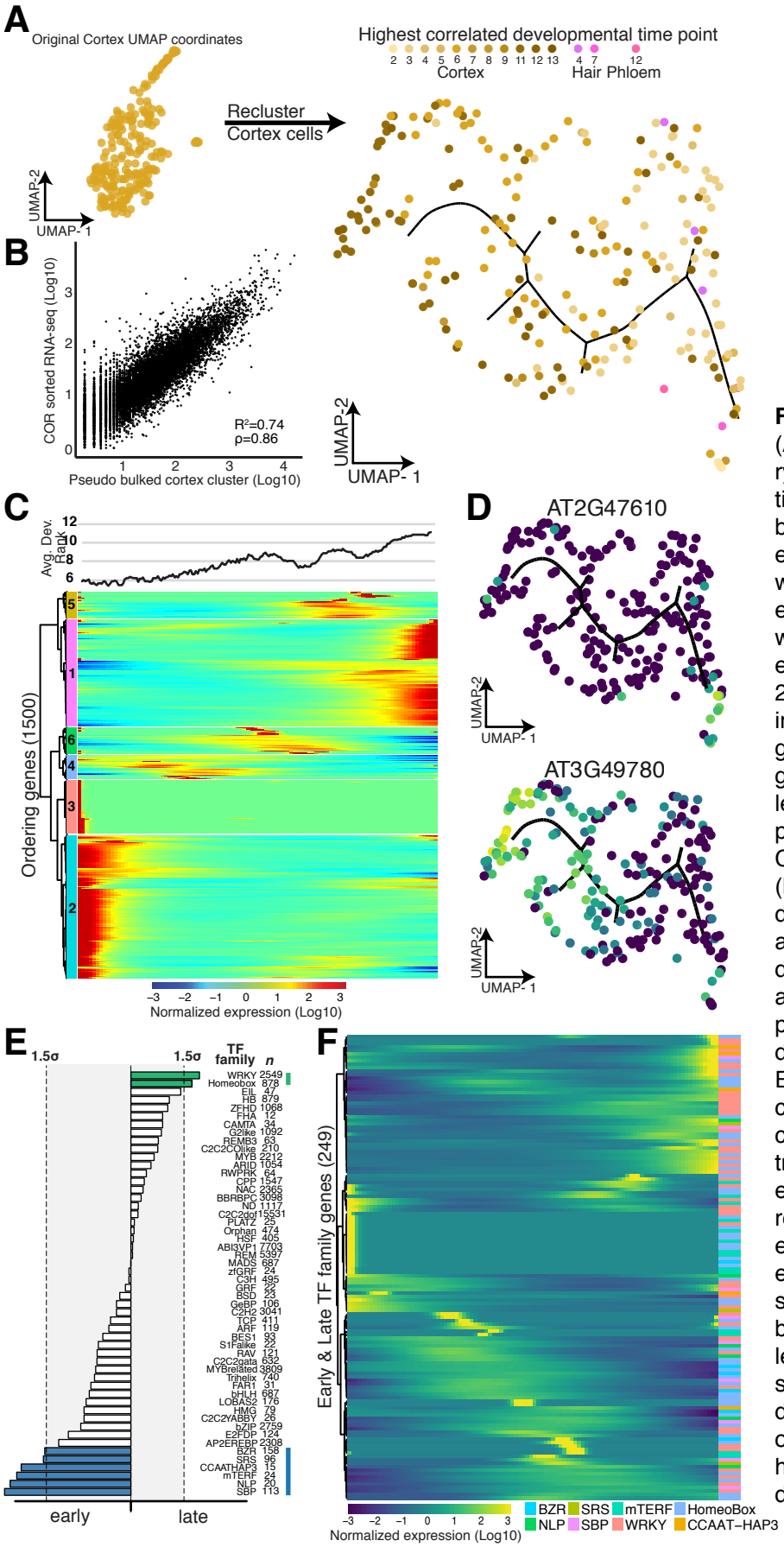
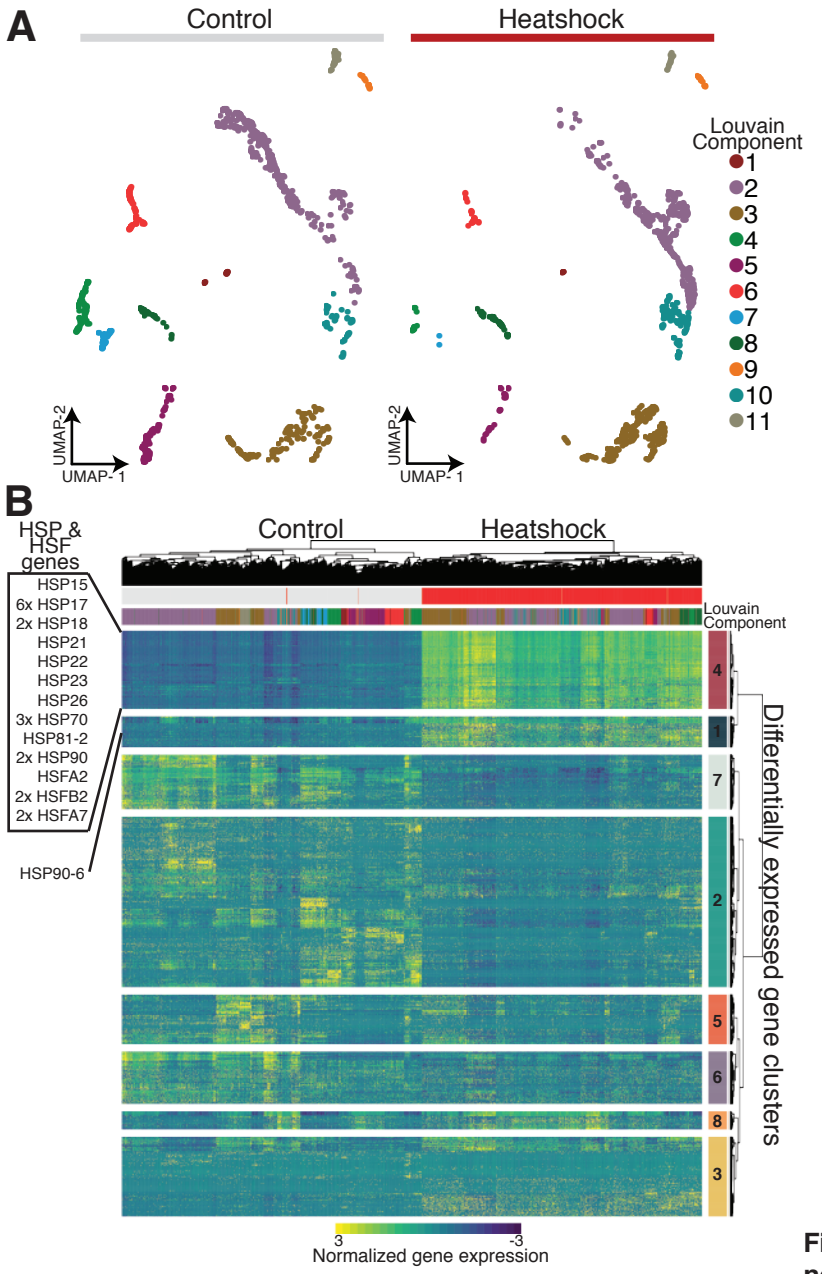


Figure 4. Developmental trajectory of cortex cells. (A) Cortex cells were re-clustered to create a trajectory, in which each cell was assigned a developmental time point and identity (shades of yellow, brown, pink) based on the highest correlation of a cell's gene expression with prior data (Brady et al., 2007; Cartwright et al., 2009). (B). Comparison of pseudo-bulk expression data from cells annotated as cortex cells with bulk expression data from protoplasts sorted for expression of the cortex marker gene COR (Li et al., 2016). (C) Cells were ordered in pseudotime; columns indicate cells and rows the expression of the 1500 genes. Rows were grouped based on similarity in gene expression, resulting in 6 clusters (indicated left), with genes in clusters 2 and 3 expressed early in pseudotime and genes in cluster 1 expressed late. Cortex cells with the earliest developmental signal (Brady et al., 2007; Cartwright et al., 2009) were designated as the root of the trajectory. The graph above represents the average best-correlation of developmental stage (Brady et al., 2007; Cartwright et al., 2009) in a scrolling window of 20 cells with pseudotime, showing the expected increase in developmental age with increasing pseudotime. (D) Examples of an early and a late expressed novel cortex-cell-specific gene. Gene expression in each cell is superimposed onto the UMAP cluster and trajectory, with lighter colors indicating higher gene expression. (E) Different transcription factor motifs reside in the 500 bp upstream regions of genes expressed early (clusters 2, 3) compared to genes expressed late (cluster 1). Transcription factor motifs specific to early cortex cells are denoted with blue bars, those for late cortex cells with green bars, bar length indicates motif frequency. Thresholds on either side (grey box, dotted lines) refer to 1.5 standard deviation above mean motif frequency. (F) Expression of individual members of transcription factors families highlighted in D across pseudotime identifies candidate factors driving early or late gene expression.



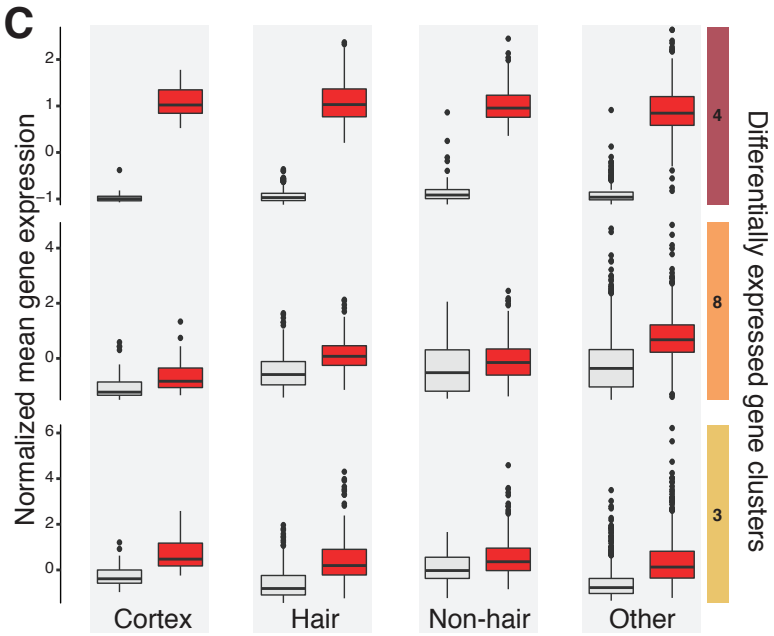


Figure 5. Single cell RNA-seq highlights canonical and novel aspects of the heat shock response. (A) Using a nearest neighbor approach, control and heat-stressed cells yielded largely comparable UMAP clusters. (B) Single-cell expression values for genes previously identified as differentially regulated in response to a similar heat stress in bulk tissue (Alexandre et al., 2018) were clustered based on similarity in gene expression in control versus heat- shocked cells. Clusters are indicated at right. Cluster 4 contains the vast majority of acute heat-stress induced genes, for which chromatin accessibility extends along the gene body upon heat shock (Sullivan et al., 2014). See Supplemental Figure 9 for a similar analysis considering genes residing near chromatin accessible sites that are dynamic in response to heat shock (Sullivan et al., 2014). (C) Pair-wise comparisons of cell and tissue types reveal subtle but highly significant differences in the heat shock response for some clusters. Significance was determined with a generalized linear model. Shown are data for clusters 3 and 8. There is no cell or tissue-type specific response for cluster 4, all cells show the canonical heat stress response typical of genes in this cluster.