Single-nuclei RNA-sequencing of plants

Daniele Y. Sunaga-Franze[1,†], Jose M. Muino[2,†], Caroline Braeuning[1,†], Xiaocai Xu[3,†], Minglei Zong[3], Cezary Smaczniak[3], Wenhao Yan[3], Cornelius Fischer[1], Ramon Vidal[1], Magdalena Kliem[1], Kerstin Kaufmann[3,*], Sascha Sauer[1,*]

[1] Genomics Platforms, Max Delbrück Center for Molecular Medicine/Berlin Institute of Health, Berlin, Germany

[2] Systems Biology of Gene Regulation, Humboldt-Universität zu Berlin, Institute of Biology, Berlin, Germany

[3] Plant Cell and Molecular Biology, Humboldt-Universität zu Berlin, Institute of Biology, Berlin, Germany

† Joint Authors

*Correspondence to: kerstin.kaufmann@hu-berlin.de, sascha.sauer@mdc-berlin.de

ABSTRACT

Single-cell genomics offers a rich potential for research on plant development and environmental responses. Here, we introduce a generic procedure for plant nuclei isolation and nanowell-based library preparation for short-read sequencing. This plant-nuclei sequencing (PN-seq) method allows for analyzing the transcriptome in thousands of individual plant cells. We show the applicability of the experimental procedure to seedlings and developing flowers from *Arabidopsis thaliana*. The developmental trajectory of anther development is reconstructed, and stage-specific master regulators and their target genes are predicted. Novel marker genes for specific anther developmental stages are experimentally verified. The nuclei isolation procedure can be applied in different plant species, thus expanding the toolkit for plant single cell genomics experiments.

INTRODUCTION

The fundamental units of life, the cells, can vary tremendously within an organism. The analysis of specialized cells and their interactions is essential for a comprehensive understanding of the function of tissues and biological systems in general. Major biological roles such as growth, development and physiology ultimately gain plasticity from heterogeneity in cellular gene expression (1).

Without precise transcriptional maps of different cell populations, we cannot accurately describe all their functions and underlying molecular networks that drive their activities. Recent advances in single-cell (sc) and in particular single-nucleus (sn) RNA-sequencing have put comprehensive, high-resolution reference transcriptome maps of mammalian cells and tissues on the agenda of international consortia such as the Human Cell Atlas (2).

Similar efforts are made by the Plant Cell Atlas (3). Plant tissues and plant cells pose specific challenges compared to mammalian systems (4). Plant cells are immobilized in a rigid cell wall matrix, which is required to be removed for isolating single cells. Additional technical demands include size variability of plant cells, and the presence of plastids and vacuoles. Consequently, these characteristics require considerably different operational procedures compared with mammalian tissues.

Recently, plant single-cell RNA-sequencing studies using protoplast isolation (PI) have been published (5-9). This procedure allows to sensitively identify and classify plant cell types. However, it is known that enzymatic digestion of plant cell walls is an important stressor for the plant and thus can introduce artifacts at the transcriptome level, limiting the applicability of this approach. To overcome this limitation, PI-response genes can be identified through an independent bulk RNA-seq experiment and later eliminated from the scRNA-seq analysis (9), but this solution does not completely correct the bias introduced by PI. Here, we introduce a single-nucleus sequencing protocol as an alternative to the use of PI and illustrate it by studying the dynamics of *Arabidopsis* transcriptomes during flower development. Working with nuclei has the advantage to eliminate organelles and vacuoles, as well as secondary metabolites localized in the cytoplasm that can interact with RNA. SnRNA-seq experiments have specific challenges, such as lower RNA yield, that need to be overcome by optimized experimental procedures and data analysis strategies (10-13).

MATERIALS AND METHODS

Preparation of plant tissues

One gram of *Arabidopsis thaliana* (Col-0) seedlings or 10 inflorescences were collected and snap-frozen in liquid nitrogen. The same procedure was applied for the following samples: 10 unopened buds of *Petunia hybrida* (W115), 8 unopened buds of *Antirrhinum majus*, 20 fully developed flowers and 1.3 g leaves of *Solanum lycopersicum*.

Preparation of nuclei

Frozen tissue was carefully crushed to small pieces in liquid nitrogen using a mortar and a pestle and transferred to a gentleMACS M tube that was filled with 5 ml of Honda buffer (2.5 % Ficoll 400, 5 % Dextran T40, 0.4 M sucrose, 10 mM $MgCl_2$, 1 µM DTT, 0.5% Triton X-100, 1 tablet/50 ml cOmplete Protease Inhibitor Cocktail, 0.4 U/µl RiboLock, 25 mM Tris-HCl, pH 7.4). This buffer composition enables efficient lysis of cell membranes while keeping the nuclei membranes intact (14). The M tubes were put onto a gentleMACS Dissociator and a specific program (Supplementary Table 1) was run at 4 °C to disrupt the tissue and to release nuclei. The resulting suspension was filtered through a 70 µm strainer and centrifuged at 1000 *g* for 6 min at 4 °C. The pellet was resuspended carefully in 500 µl Honda buffer, filtered through a 35 µm strainer and stained with 3x staining buffer (12 µM DAPI, 0.4 U/µl Ambion RNase Inhibitor, 0.2 U/µl SUPERaseIn RNase Inhibitor in PBS). Nuclei were sorted by gating on the DAPI peaks using a BD FACS Aria III (200,000 – 400,000 events) into a small volume of landing buffer (4% BSA in PBS, 2 U/µl Ambion RNase Inhibitor, 1 U/µl SUPERaseIn RNase Inhibitor). Sorted nuclei were additionally stained with NucBlue from the Invitrogen Ready Probes Cell Viability Imaging Kit (Blue/Red), then counted and checked for integrity in Neubauer counting chambers. Quality of RNA derived from sorted nuclei was analyzed by

Agilent TapeStation using RNA ScreenTape or alternatively by Agilent's Bioanalyser 2100 system.

Preparation of single-nucleus libraries using SMARTer ICELL8 Single-Cell System

The NucBlue and DAPI co-stained single-nuclei suspension (60 cells/µl) was distributed to eight wells of a 384-well source plate (Cat. No. 640018, Takara) and then dispensed into a barcoded SMARTer ICELL8 3' DE Chip (Cat. No. 640143, Takara) by an ICELL8 MultiSample NanoDispenser (MSND, Takara). Chips were sealed and centrifuged at 500 g for 5 min at 4 °C. Nanowells were imaged using the ICELL8 Imaging Station (Takara). After imaging, the chip was placed in a pre-cooled freezing chamber, and stored at −80 °C for at least 2 h. The CellSelect software was used to support the identification of nanowells that contained a single nucleus. One chip yielded on average between 800 - 1200 nanowells with single nuclei. These nanowells were selected for subsequent targeted deposition of 50 nl/nanowell RT-PCR reaction mix from the SMARTer ICELL8 3' DE Reagent Kit (Cat. No. 640167, Takara) using the MSND. After RT and amplification in a Chip Cycler, barcoded cDNA products from nanowells were pooled by means of the SMARTer ICELL8 Collection Kit (Cat. No. 640048, Takara). cDNA was concentrated using the Zymo DNA Clean & Concentrator kit (Cat. No. D4013, Zymo Research) and purified with AMPure XP beads. Afterwards, cDNA was used to construct Nextera XT (Illumina) DNA libraries followed by AMPure XP bead purification. Qubit dsDNA HS Assay Kit, KAPA Library Quantification Kit for Illumina Platforms and Agilent High Sensitivity D1000 ScreenTape Assay were used for library quantification and quality assessment. Strand-specific RNA libraries for sequencing were prepared with TruSeq Cluster Kit v3 and sequenced on an Illumina HiSeq 4000 instrument (PE100 run).

Preparation of bulk RNA-seq libraries

Five 10-days-old *Arabidopsis thaliana* seedlings were collected into 1.5 ml screw-cap tubes with 5 glass beads, precooled in liquid nitrogen. Samples were homogenized by adding one half of the TRI-Reagent (Sigma-Aldrich, 1 ml per 100 mg) to each sample following sample disruption by using the Precellys 24 Lysis & Homogenization instrument for 30 sec and 4000 rpm. After homogenization, total RNA was extracted by adding the second half of the TRI-Reagent and the protocol was proceeded according to the manufacturer. To remove any co-precipitated DNA, a DNase-I digest was performed by using 1U DNase-I (NEB) in a total volume of 100 µl. Total RNA was cleaned-up by LiCl-precipitation using 10 µl 8 M LiCl and 3 vol 100% ethanol incubating at -20 °C overnight. Following a spin down at 4 °C, 17,900 xg for 30 min and 2 washing steps with 70% ethanol. The RNA pellet was dried on ice for 1 h and resuspended in 40 µl DEPC-water incubating at 56 °C for 5 min. Quality of total RNA was analyzed by Agilent TapeStation using RNA ScreenTape or alternatively by Agilent's Bioanalyser 2100 system. Concentration was measured by a Qubit RNA BR Assay Kit (Thermo Fisher Scientific). One µg of total-RNA was used for RNA library preparation with Illumina TruSeq® Stranded mRNA Library Prep, following the protocol according to the manufacturer. Quality and fragment peak size were checked by Agilent TapeStation using D1000 ScreenTape or alternatively by Agilent's Bioanalyser 2100 system. Concentration was measured by the Qubit dsDNA BR Assay Kit (Thermo Fisher Scientific). Three replicates, composed of 5 seedlings each, were used separately throughout the whole procedure. Strand-specific RNA libraries were prepared using TruSeq Stranded mRNA library preparation procedure and the three replicates were sequenced on an Illumina NextSeq 500 instrument (PE75 run).

Data pre-processing

Raw sequencing files (bcl) were demultiplexed and fastq files were generated using Illumina bcl2fastq software (v2.20.0). The command-line version of ICELL8 mappa analysis pipeline (demuxer and analyzer v0.92) was used for the data pre-processing and read mapping. Mappa_demuxer assigned the reads to the cell barcodes present in a predefined list of barcode sequences. Read trimming, genome alignment (*Arabidopsis thaliana* reference genome: TAIR10), counting and summarization were performed by mappa_analyzer with the default parameters. A report containing the experimental overview and read statistics for each PN-seq library was created using hanta software from the ICELL8 mappa analysis pipeline (Supplementary Data 1). The gene matrix generated by mappa_analyzer was used as input for the downstream analysis using R package Seurat v3 (15-16).

Quality control and data analysis

The analysis started by removing reads with barcodes representing the negative and positive controls included in all Takara Bio's NGS kits. For the seedling samples, Seurat was used to filter viable nuclei by i) removing genes detected in less than 3 nuclei, ii) nuclei with less than 200 genes, iii) nuclei with more than 5% of reads mapped to mitochondria and iv) nuclei with more than 5% mapped to chloroplasts. Seurat *SCTransform* normalization method was performed for each one of the seedling replicates separately. Data from 3 seedling replicates were integrated using *PrepSCTIntegration*, *FindIntegrationAnchors* and *IntegrateData* functions. After running the *RunPCA* (default parameters), we performed UMAP embedding using *runUMAP* with *dims*=1:20. Clustering analysis was performed using *FindNeighbors* (default parameters) and *FindClusters* function with *resolution*=0.5. Differentially expressed genes were found using *FindAllMarkers* function and “wilcox” test, *logfc.threshold = 0.25* and *min.pct*=0.25. The sub-clustering analysis of root was performed using the *subset* function and the seedling clusters containing root cells (clusters: 3, 4, 6, 7, 9, 11 and 12; Fig. 2b). *SCTransform* and *RunUMAP* with *dims*=1:15 and *resolution*=1.5 were re-run after sub-setting the data and subsequently *FindAllMarkers* to find the differentially expressed genes across the sub-clusters, with the “wilcox” test, *logfc.threshold = 0.25* and using the RNA assay (normalized counts).

For the flower PN-seq dataset (900 nuclei), only genes encoded in the nucleus were used (32,548 genes). Nuclei with i) less than 10,000 reads, ii) less than 500 genes containing 10 reads or iii) at least one gene covering more than 10% of the reads of a particular nucleus were filtered out. In addition, genes with less than 10 reads in at least 15 nuclei were also removed. The filtering step resulted in a dataset containing 856 nuclei and 14,690 genes. Seurat *SCTransform* normalization was applied to the filtered data using all genes as *variable.features*, and with parameters: *method*=”nb”, and *min_cells*=5. We used the *JackStraw* function in Seurat to estimate the optimal number of PCAs to be used in the analysis. After calculating the first 12 PCAs with *RunPCA,* we performed UMAP embedding using *runUMAP* with parameters *n.neighbors*=10, *min.dist*=.1, *metric*="correlation" and *umap.method*="umap-learn". Clustering was done with *FindNeighbors* (default parameters) and *FindClusters* function using the SLM algorithm, *resolution*=1.15 and *n.iter*=100. Markers genes were found with the function *FindAllMArkers*, using the “wilcox” test and *min.pct*=0.25.

Annotation

Annotation of the seedling and flower clusters was performed by visualizing the expression of the top 20 marker genes of each identified cluster on TraVaDB (Transcriptome Variation Analysis Database, http:/travadb.org (17)). TraVaDB is an open-access database containing the transcriptomes (RNA-seq) of a large type of Arabidopsis developmental stages and organs. The annotation of the root clusters was based on the top 200 marker genes (order by their p-value) reported by Denyer *et al.* (9). Their overlap with the marker genes from our root clusters was calculated. As the number of marker genes from our clusters is variable, we reported the proportion of marker genes from each of our clusters that overlaped the marker genes of each Denyer *et al.'s* clusters. The labels of Denyer *et al*.’s clusters were transferred to our clusters when the cluster showed the highest overlap (measured as proportion).

Reproducibility and correlation

To assess the reproducibility of our method, we compared the pooled number of reads overlapping each gene of each seedling replicate against one another in log2 space. The same was done to verify the similarity between unfixed and fixed seedling datasets.

The correlation between bulk and PN-seq datasets was investigated by comparing the average number of reads overlapping each gene in the PN-seq against bulk RNA-seq datasets. PN-seq (unfixed) and bulk RNA-seq of seedlings were generated in 3 biological replicates. Expression of bulk RNA-seq data was quantified with RSEM (18).

Network analysis

GENIE3 (19) was used to infer gene networks starting from the normalized expression data obtained from Seurat for each cluster independently, using the parameters *nTrees*=1000, and using as regulators the list of DNA binding proteins obtained from TAIR (www.arabidopsis.org). Genes expressed in less than 33% of the nuclei in a particular cluster were removed. Only the top 10,000 interactions were kept. Gene regulators with less than 10 predicted targets were also removed. Dynamics of the gene network through anther development were obtained by the following approach: first, all nuclei were ordered by their estimated developmental pseudotime using Monocle 3 (20) and cluster 0 (meristem/Early anther) as root cluster. Second, gene networks were estimated with GENIE, as described previously, using groups of non-overlapping sets of 50 nuclei that were previously ordered by its developmental pseudotime.

Generation and Confocal Imaging of Reporter Lines

To validate expression specificity of the marker genes from our single cell PN-seq approach, promoter::NLS-GFP (nuclear localization signal-green fluorescent protein) reporter lines were generated. The marker genes for validation were chosen from the pool of cluster-specific marker genes ($p<0.05$) that were not previously characterized in the literature (unknown marker genes). The genomic promoter region upstream of the ATG and until the closest neighboring gene was amplified by PCR and introduced into the entry vector pCR8:GW:TOPO by TA cloning (primers used for PCR are listed in Supplementary Table 3. Afterwards, the LR reactions were performed with the binary vector pGREEN:GW:NLS-GFP (21) to generate GFP transcriptional fusions to a nuclear localization signal (NLS) peptide. All reporter constructs were transformed into the Col-0 Arabidopsis background, and multiple independent lines per construct were analyzed under a Zeiss LSM800 laser-scanning confocal microscope. Different floral organs were dissected and screened for the GFP signal by confocal microscopy under 20× and 63× magnification objectives. Auto-fluorescence from chlorophyll was collected to give an outline of the flower organs. A 488-nm laser was used to excite GFP and chlorophyll and emissions were captured using PMTs set at 410–530 nm and 650–700nm. Z-stack screens were performed for the floral meristem and stigma tissues to give a 3D structure visualization.

RESULTS AND DISCUSSION

The problem of protoplast isolation for single-cell approaches

In order to evaluate the impact of PI on single cell RNA-seq experiments, we performed a re-analysis of root scRNA-seq data from Denyer *et al*. (9). In this work, an independent bulk RNA-seq experiment was performed to identify PI-responsive genes and subsequently eliminate them from the scRNA-seq analysis. We started our re-analysis observing the effect of the high PI-responsive genes in the clustering analysis. Several clusters were found containing cells with strong response to PI (Fig. 1a), with one cluster having up to 55% PI-responsive genes among the top 20 marker genes. This effect largely persisted when PI-responsive genes were excluded from the scRNA-seq analysis. After excluding PI-responsive genes from the clustering step, but still using them to identify markers, we observed that the clustering continued to be affected (Fig. 1b), with one cluster having up to 46% of its top 20 marker genes being PI-responsive genes.

Beyond previous reports showing the bias caused by the use of PI (9), these results highlight the need for alternative methods for plant single-cell genomics. In addition, protoplast isolation is not feasible in some plant tissues.

Nuclei isolation and scRNA-seq library preparation

Here we propose a single-nucleus sequencing strategy for transcriptome sequencing (PN-seq) in individual plant cells (Fig. 2a; full protocol in Materials and Methods). The key step of our plant-nuclei sequencing procedure consists of gentle but efficient isolation of plant nuclei. Snap-frozen *Arabidopsis* tissue was gently physically dissociated by pestle and transferred to Honda buffer for cell lysis (14). Cell walls and cell membranes were mechanically disrupted using a gentleMACS Dissociator, keeping the nuclei largely intact as observed by DAPI staining (Supplementary Fig. 1a). Released intact *Arabidopsis* nuclei were collected using Fluorescence-Activated Cell Sorting (FACS; Supplementary Fig. 1b). A clear separation between nuclei and debris was obtained (Supplementary Fig. 1c). To show the applicability of this method to different plant species/tissues, we successfully performed nuclei isolation in *Arabidopsis thaliana* (seedlings and flowers), *Petunia hybrida* (flowers), *Antirrhinum majus* (flowers), and *Solanum lycopersicum* (flowers and leaves) (Supplementary Fig. 1a,b). The RNA that was isolated from these nuclei was of high quality as observed by electrophoresis for *Arabidopsis* (Supplementary Fig. 1d).

The next step consists of generating high-quality cDNA libraries from the isolated nuclei. In principle, a number of library preparations and sequencing procedures can be combined (9,22). We opted for the Takara's ICELL8 system, a sensitive nanowell-based approach that includes a standardized lysis of nuclei by detergents and a freeze-thaw-cycle (23). One of the main advantages of this system is that it allows for manual selection of single-nucleus-containing wells, as well as visual inspection and selection of intact nuclei (i.e. nuclear rupture), thereby introducing additional quality control. Using SMARTer ICELL8 3' chemistry, we prepared DNA libraries for short paired-end sequencing using fresh, snap-frozen *Arabidopsis* seedlings.

PN-seq performance in Arabidopsis seedlings

To establish the method, we set up the protocol using pools of *Arabidopsis thaliana* seedlings (3 biological replicates), which feature diverse plant structures comprising the primary root, the hypocotyl and the cotyledons. This allowed us to characterize the performance of the procedures recovering the transcriptomes of a diverse set of tissues/organs. On average, we obtained 1,116 nuclei per replicate and 2,802 expressed genes per nucleus at ~220,000 sequenced reads per nucleus (Supplementary Data 1, Supplementary Fig. 2). A Pearson correlation coefficient of 0.9 was observed among the 3 biological replicates, indicating the high reproducibility of the method (Supplementary Fig. 3a,b). A good reproducibility was also observed between PN-seq and bulk RNA-seq (Pearson correlation of 0.74, Fig. 2d), even though the PN-seq data represent the nuclear transcriptome while the bulk RNA-seq data represent the nuclear and cytoplasmic transcriptome. This correlation is consistent with correlation coefficients found in previous publications (ranging from 0.7 to 0.85) (24), therefore indicating that the method was able to recover the main transcript abundances present in the bulk RNA-seq data.

The integration of the 3 seedling datasets by Seurat revealed 13 distinct clusters among the transcriptome of 2,871 nuclei (Fig. 2b). To annotate the cell types enriched in each cluster, we first obtained the top 20 marker genes of each cluster (Supplementary Table 2). The tissue-specificity of these markers were visualized in a heatmap (Supplementary Fig. 4) with the expression of these markers in a collection of tissue-specific transcriptome samples

(Transcriptome Variation Analysis Database; TraVaDB). This resulted in the annotation of 12 out of 13 clusters, which included all the expected main basic organ types of seedlings: Leaves/Cotyledons (n=643 nuclei), Hypocotyls (n=393 nuclei), Vasculature (Leaves/Roots) (n=342 nuclei), 2 clusters of Roots (n=267 nuclei), Shoot meristems (n=180 nuclei), 2 clusters of Root apices (n=192 nuclei), Roots/Hypocotyls (n=136 nuclei), 2 clusters of Leaves (n=152 nuclei) and Mature roots (n=27 nuclei). A cluster containing 539 nuclei was not annotated and labeled as Not Determined (Fig. 2b; Supplementary Table 2). A similar proportion of nuclei from each annotated cluster was observed across the 3 replicates, again indicating the good reproducibility of the method (Fig. 2c).

As the majority of the recent publications using scRNA-seq PI-based methods has been performed in roots, we also investigated the ability of our method to recover the main root cell types. For this purpose, we performed in-depth analysis of 964 nuclei that were identified as "root" in the seedling datasets. The 964 nuclei were re-clustered into 12 new sub-clusters. The marker genes of the predicted 12 sub-clusters were compared to the list of the top 200 markers from the recently published atlas of the *Arabidopsis* root (9). A good overlap among marker genes from both datasets was found (Supplementary Fig. 5b), with some clusters with an overlap of 40 genes. As result, the main root cell types could be recovered from our seedling dataset (Supplementary Fig. 5).

Similarity between snRNA-seq data generated from fixed and unfixed plant material

To allow for more technical flexibility in our method, i.e. the possibility to simplify the storage of plant samples and maintaining *in situ* expression states (25), we fixed seedlings using methanol directly after harvest and performed PN-seq as described before. We obtained a similar number of nuclei (850) and an average number of expressed genes (2,292) when using methanol fixation compared to no fixation (1,116 nuclei and 2,802 genes). A similar nuclei distribution was also observed between fixed and non-fixed samples (Supplementary Fig. 6a). Additionally, an expression correlation of 0.88 and p-value<2.2e-16 was observed among both group of samples (Supplementary Fig. 6b,c). These results indicate that fixation of the material does not introduce major differences in the number of nuclei and obtained cell-types.

PN-seq performance in *Arabidopsis* inflorescences

To evaluate the performance of PN-seq to study cell differentiation, we applied PN-seq to *Arabidopsis thaliana* inflorescences, which cover all stages of flower development prior to anthesis. After quality control filtering, we obtained transcriptomes of 856 nuclei with an average number of 2,967 expressed genes per nucleus (Supplementary Fig. 7a). The analysis identified 15 clusters corresponding to distinct organs and developmental stages (Fig. 3a; Supplementary Fig. 7b). To annotate these clusters with particular cell types, we first identified specific marker genes of each cluster (Supplementary Table 2), then plotted their expression profiles in the different floral organs and developmental stages obtained from TraVaDB (Fig. 3b). Last, we correlated the gene expression of each cluster with each TraVaDB sample and indicated these values in the UMAP plot (Supplementary Fig. 7c). A major proportion of clusters (37% of the nuclei population) were annotated as differentiating anthers at different developmental stages (clusters 3, 4, 6, 7, 10, 15). This can be explained by the fact that anthers comprise a large fraction of tissues (26-27) in developing flowers. Furthermore, anthers/pollen have very specific gene expression profiles (26-27) which may facilitate the computational identification of the clusters. Our data captured gene expression dynamics during anther/pollen development from undifferentiated stem cells (cluster 0; Fig. 3) to late anther stages close to organ maturity, prior to anthesis (cluster 3; Fig. 3). This led us to use Monocle 3 to estimate the pseudotime of each anther cell (Supplementary Fig. 8c). When we plotted the average pseudotime of the cells of each anther cluster against the developmental time of each cluster obtained with the TraVaDB annotation (Supplementary Fig. 8d), it showed a strong concordance with anther developmental stages, which indicates that we can use the estimated pseudotime of each cell as a proxy of its developmental stage, and therefore to study transcriptional dynamics of anther differentiation.

Gene regulatory trajectories of anther and pollen development

We used GENIE3 (19) to exemplify the capacity of the snRNA-seq data to infer the dynamics of gene regulatory networks (GRNs) during plant development. We reconstructed the GRNs for all clusters that were identified as “anthers” and estimated the strength of interactions between known transcription factors (TFs) versus all expressed genes. For example, the Figure 3d shows the GRN for cluster 15 representing an early anther stage. In our analysis, one of the main master TF (with most interactions) was *ABORTED MICROSPORES* (*AMS*), an already known regulator of anther development. We investigated more in-depth the regulatory dynamics of this TF using our data, the predicted targets of *AMS* and the related TF genes *bHLH089*, *bHLH091* and *bHLH010* (28-29) were expressed in a highly dynamic manner (Fig. 2c,d). AMS target genes at early stages were functionally enriched in chromatin remodeling (e.g. *BRAHMA*; *SET DOMAIN PROTEIN 16*) and pollen development (*DIHYDROFLAVONOL 4-REDUCTASE-LIKE1; ATP-BINDING CASSETTE G26*) (Fig. 3e). Late targets included metabolic enzymes as well as genes associated with RNA-regulatory processes. Newly identified marker genes covered the full anther developmental trajectory and are candidates for further mechanistic analyses.

Validation of cell type markers genes

To validate the clustering analysis and dynamic anther transcriptome trajectory, we assessed the expression patterns of genes using promoter::NLS-GFP reporter lines. We selected 10 previously uncharacterized genes predicted to be specific or preferentially expressed in one of the clusters (Fig. 4). Seven out of 10 selected genes showed a specific expression in line with predictions (Fig. 4, Supplementary Fig. 9). The genes *AT5G20030*, *AT5G08250*, *AT1G23520* and *AT2G16750* were expressed in anthers and showed stage-specific expression as predicted by our analysis shown in Fig. 3 (Fig. 4a-c, Supplementary Fig. 9a,c,d). Gene *AT5G08250* from cluster 7, the first cluster of anther lineage, showed very strong expression in young anthers from flower 16 to flower 18 (nomenclature according to TraVaDB; Fig. 4b, Supplementary Fig. 9a); *AT5G20030* from cluster 15, which is an 'early anther' cluster, showed a peak in expression in flower 12 (Fig. 4c, Supplementary Fig. 9b). *AT2G16750* from cluster 6, was expressed strongly in older anthers in flower 10 and flower 11 (Fig. 4d, Supplementary Fig. 9c). Finally, *AT1G23520* from cluster 3, the last cluster of anther lineage, was found to be expressed in old anthers in flower 6 to flower 8 (Fig. 4e, Supplementary Fig. 9d). Specific expression in the floral meristem was observed for genes *AT1G63100 and AT3G51740* from cluster 11 (Fig. 4f,g). Moreover, gene *AT4G11290* from cluster 14 showed highly specific expression in the stigma (Fig. 4h). On the other hand, *AT1G54500* was expressed in sepal primordia and developing sepals (Fig. 4i), indicating that it is not specific to meristems as predicted for cluster 5. *AT3G05570* and *AT2G38995* were found to be more broadly expressed (not shown).

Overall, although it is known that protoplast isolation (PI) procedure can affect the plant transcriptome, it has been the principle choice for plant single-cell sequencing and has been mostly applied to root samples so far (7-9, 30-31). Here, we introduced PN-seq that can be applied to analyze nucleic acids in bulk or in individual cells. The PN-seq methodology - based on efficient isolation of nuclei - is directly and easily applicable to a broad range of different plant tissues such as seedlings, flowers and leaves, and thus provides a versatile tool for plant single cell omics. In principle, various library preparation and sequencing methods can be combined with our generic nuclei isolation procedure.

Nanowell-based library preparation offered the possibility of visual quality control of individual nuclei, achieved high numbers of several thousand genes per cell and more than a thousand nuclei per run to sensitively detect plant cell (sub-) types. The number of nuclei can potentially be upscaled by using denser and/or larger nanowell-formats to further increase the number of nuclei for sequence analysis. The here applied nanowell-based approach resulting in deep cellular transcriptome data was of particular advantage to identify co-regulated genes and decipher gene networks underlying biological processes of interest. Along with the ever-growing range of nucleic acid sequencing technologies and plant genomics reference databases, single-nuclei genomics procedures are expected to become valuable tools to build maps of all plant cells of developing and adult tissues, and to measure cell-type-specific differences in environmental responses to gain novel mechanistic insights into plant growth and physiology (3).

DATA AVAILABILITY

The PN-seq and bulk RNA-seq data have been deposited in the ArrayExpress database at EMBL-EBI under accession number E-MTAB-9174.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

FUNDING

The work was supported by the European Commission [FP-7, grant agreement no. 262055 to S.S, ESGI and European Union’s Horizon 2020 research and innovation programme, grant agreement no. 824110 to S.S, EASI-Genomics] and DFG [grant no. KA 2720/5-1 to X.X, K.K].

CONFLICT OF INTEREST

The authors declare no competing interests.

ACKNOWLEDGEMENTS

We thank Solenne Bourdier and Dijun Chen for their support.

REFERENCES

1. Fischer,C., Metsger,M., Bauch,S., Vidal,R., Böttcher,M., Grote,P., Kliem,M. and Sauer,S. (2019) Signals trigger state-specific transcriptional programs to support diversity and homeostasis in immune cells. *Sci. Signal.*, 12, 1–17.

2. Regev,A., Teichmann,S.A., Lander,E.S., Amit,I., Benoist,C., Birney,E., Bodenmiller,B., Campbell,P., Carninci,P., Clatworthy,M., *et al.* (2017) The Human Cell Atlas. *Elife*, 6, e27041.

3. Rhee,S.Y., Birnbaum,K.D. and Ehrhardt,D.W. (2019) Towards Building a Plant Cell Atlas. *Trends Plant Sci.*, 24, 303–310.

4. Efroni,I. and Birnbaum,K.D. (2016) The potential of single-cell profiling in plants. *Genome Biol.*, 17, 65.

5. Efroni,I., Ip,P.-L., Nawy,T., Mello,A. and Birnbaum,K.D. (2015) Quantification of cell identity from single-cell gene expression profiles. *Genome Biol.*, 16, 9.

6. Efroni,I., Mello,A., Nawy,T., Ip,P.-L., Rahni,R., DelRose,N., Powers,A., Satija,R. and Birnbaum,K.D. (2016) Root Regeneration Triggers an Embryo-like Sequence Guided by Hormonal Interactions. *Cell*, 165, 1721–1733.

7. Zhang,T.-Q., Xu,Z.-G., Shang,G.-D. and Wang,J.-W. (2019) A Single-Cell RNA Sequencing Profiles the Developmental Landscape of Arabidopsis Root. *Mol. Plant*, 12, 648–660.

8. Jean-Baptiste,K., McFaline-Figueroa,J.L., Alexandre,C.M., Dorrity,M.W., Saunders,L., Bubb,K.L., Trapnell,C., Fields,S., Queitsch,C. and Cuperusa,J.T. (2019) Dynamics of gene expression in single root cells of arabidopsis thaliana. *Plant Cell*, 31, 993–1011.

9. Denyer,T., Ma,X., Klesen,S., Scacchi,E., Nieselt,K. and Timmermans,M.C.P. (2019) Spatiotemporal Developmental Trajectories in the Arabidopsis Root Revealed Using High-Throughput Single-Cell RNA Sequencing. *Dev. Cell*, 48, 840-852.e5.

10. Lake,B.B., Codeluppi,S., Yung,Y.C., Gao,D., Chun,J., Kharchenko,P. V., Linnarsson,S. and Zhang,K. (2017) A comparative strategy for single-nucleus and single-cell transcriptomes confirms accuracy in predicted cell-type expression from nuclear RNA. *Sci. Rep.*, 7, 1–8.

11. Macaulay,I.C., Ponting,C.P. and Voet,T. (2017) Single-Cell Multiomics: Multiple Measurements from Single Cells. *Trends Genet.*, 33, 155–168.

12. Selewa,A., Dohn,R., Eckart,H., Lozano,S., Xie,B., Gauchat,E., Elorbany,R., Rhodes,K., Burnett,J., Gilad,Y., *et al.* (2020) Systematic Comparison of High-throughput Single-Cell and Single-Nucleus Transcriptomes during Cardiomyocyte Differentiation. *Sci. Rep.*, 10, 1–

13. Ding,J., Adiconis,X., Simmons,S.K., Kowalczyk,M.S., Hession,C.C., Marjanovic,N.D., Hughes,T.K., Wadsworth,M.H., Burks,T., Nguyen,L.T., *et al.* (2020) Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat. Biotechnol.*, 38, 737–746.

14. Moreno-Romero,J., Santos-González,J., Hennig,L. and Köhler,C. (2017) Applying the INTACT method to purify endosperm nuclei and to generate parental-specific epigenome profiles. *Nat. Protoc.*, 12, 238–254.

15. Butler,A., Hoffman,P., Smibert,P., Papalexi,E. and Satija,R. (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*, 36, 411–420.

16. Stuart,T., Butler,A., Hoffman,P., Hafemeister,C., Papalexi,E., Mauck,W.M., Hao,Y., Stoeckius,M., Smibert,P. and Satija,R. (2019) Comprehensive Integration of Single-Cell Data. *Cell*, 177, 1888-1902.e21.

17. Klepikova,A. V., Kasianov,A.S., Gerasimov,E.S., Logacheva,M.D. and Penin,A.A. (2016) A high resolution map of the Arabidopsis thaliana developmental transcriptome based on RNA-seq profiling. *Plant J.*, 88, 1058–1070.

18. Li,B. and Dewey,C.N. (2014) RSEM: Accurate transcript quantification from RNA-seq data with or without a reference genome. *Bioinforma. Impact Accurate Quantif. Proteomic Genet. Anal. Res.*, 10.1201/b16589.

19. Huynh-Thu,V.A., Irrthum,A., Wehenkel,L. and Geurts,P. (2010) Inferring regulatory networks from expression data using tree-based methods. *PLoS One*, 5, 1–10.

20. 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.

21. Smaczniak,C., Muiño,J.M., Chen,D., Angenent,G.C. and Kaufmann,K. (2017) Differences in DNA binding specificity of floral homeotic protein complexes predict organ-specific target genes. *Plant Cell*, 29, 1822–1835.

22. 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.

23. Goldstein,L.D., Chen,Y.J.J., Dunne,J., Mir,A., Hubschle,H., Guillory,J., Yuan,W., Zhang,J., Stinson,J., Jaiswal,B., *et al.* (2017) Massively parallel nanowell-based single-cell gene expression profiling. *BMC Genomics*, 18, 1–10.

24. Beisang,D.J., Smith,K., Yang,L., Benyumov,A., Gilbertsen,A., Herrera,J., Lock,E., Racila,E., Forster,C., Sandri,B.J., *et al.* (2020) Single-cell RNA sequencing reveals that lung mesenchymal progenitor cells in IPF exhibit pathological features early in their differentiation trajectory. *Sci. Rep.*, 10, 1–12.

25. Alles,J., Karaiskos,N., Praktiknjo,S.D., Grosswendt,S., Wahle,P., Ruffault,P.L., Ayoub,S., Schreyer,L., Boltengagen,A., Birchmeier,C., *et al.* (2017) Cell fixation and preservation for droplet-based single-cell transcriptomics. *BMC Biol.*, 15, 1–14.

26. Smyth,D.R., Bowman,J.L. and Meyerowitz,E.M. (1990) Early flower development in Arabidopsis. *Plant Cell*, 2, 755–767.

27. Gómez,J.F., Talle,B. and Wilson,Z.A. (2015) Anther and pollen development: A conserved developmental pathway. *J. Integr. Plant Biol.*, 57, 876–891.

28. Xu,J., Yang,C., Yuan,Z., Zhang,D., Gondwe,M.Y., Ding,Z., Liang,W., Zhang,D. and Wilson,Z.A. (2010) The ABORTED MICROSPORES regulatory network is required for postmeiotic male reproductive development in Arabidopsis thaliana. *Plant Cell*, 22, 91–107.

29. Zhu,E., You,C., Wang,S., Cui,J., Niu,B., Wang,Y., Qi,J., Ma,H. and Chang,F. (2015) The DYT1-interacting proteins bHLH010, bHLH089 and bHLH091 are redundantly required for Arabidopsis anther development and transcriptome. *Plant J.*, 83, 976–990.

30. Shulse,C.N., Cole,B.J., Ciobanu,D., Malley,R.C.O., Brady,S.M., Dickel,D.E., Shulse,C.N., Cole,B.J., Ciobanu,D., Lin,J., *et al.* (2019) Profiling of Plant Cell Types Resource Profiling of Plant Cell Types. *Cell Reports*, 27, 2241-2247.e4.

31. Ryu,K.H., Huang,L., Kang,H.M. and Schiefelbein,J. (2019) Single-cell RNA sequencing resolves molecular relationships among individual plant cells. *Plant Physiol.*, 179, 1444–1456.

FIGURE LEGENDS

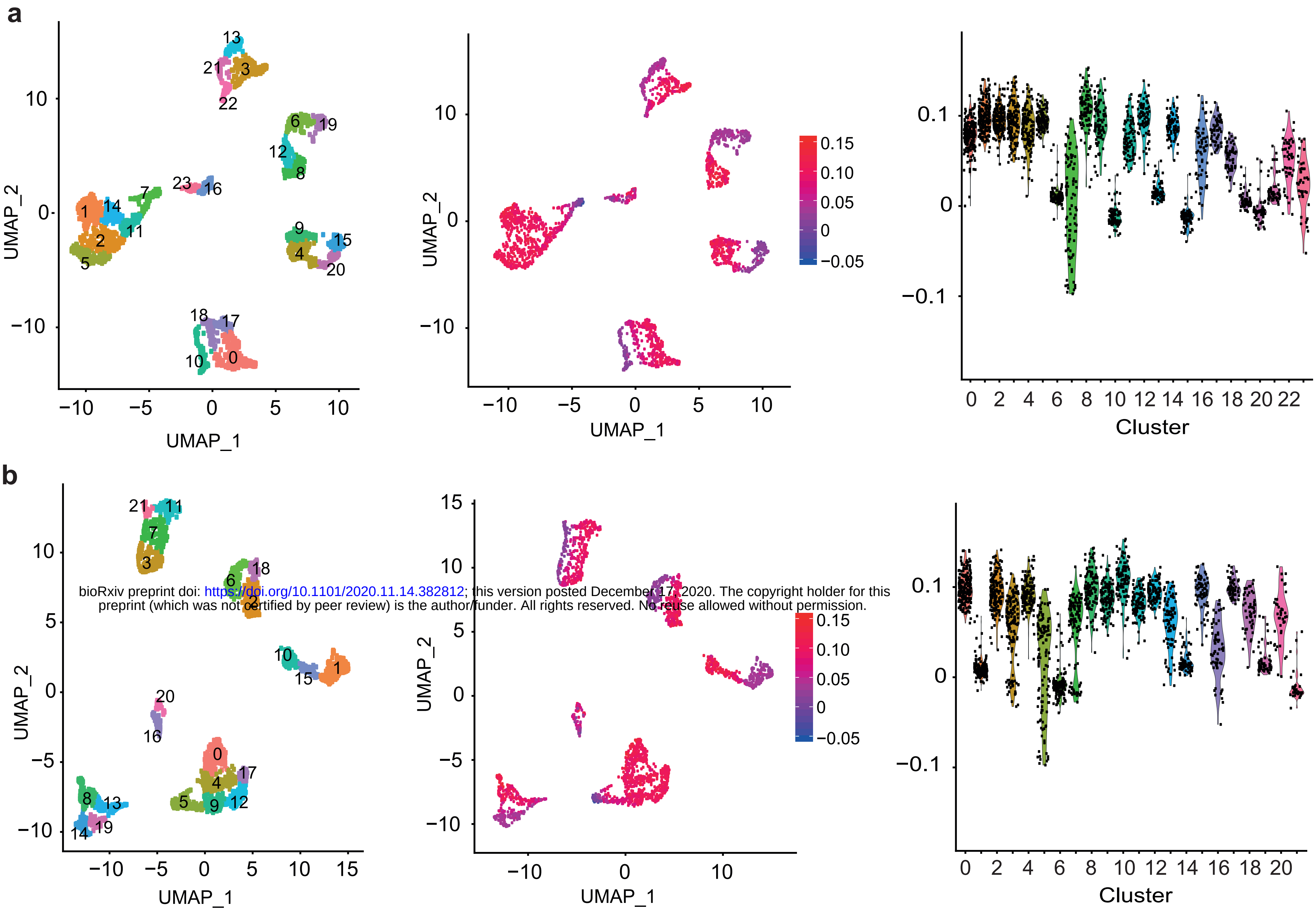
Figure 1: Effect of protoplast isolation (PI) in the Denyer *et al*. (9) root scRNA-seq experiment (data from ref. (9)). a) Re-analysis of the full scRNA-seq data. b) Re-analysis of the full scRNA-seq data after removing the top 6,000 PI-responsive genes. In a) and b), the left UMAP plots show the cell clusters of scRNA-seq data; the UMAP plots in the center show the difference between the correlation of each cell from scRNA-seq and bulk RNA-seq in PI and non-PI samples. Positive correlation numbers indicate cells with stronger similarity to the transcriptome of PI samples. The violin plots in the right show the difference in the correlation of cells between PI and non-PI per cluster.

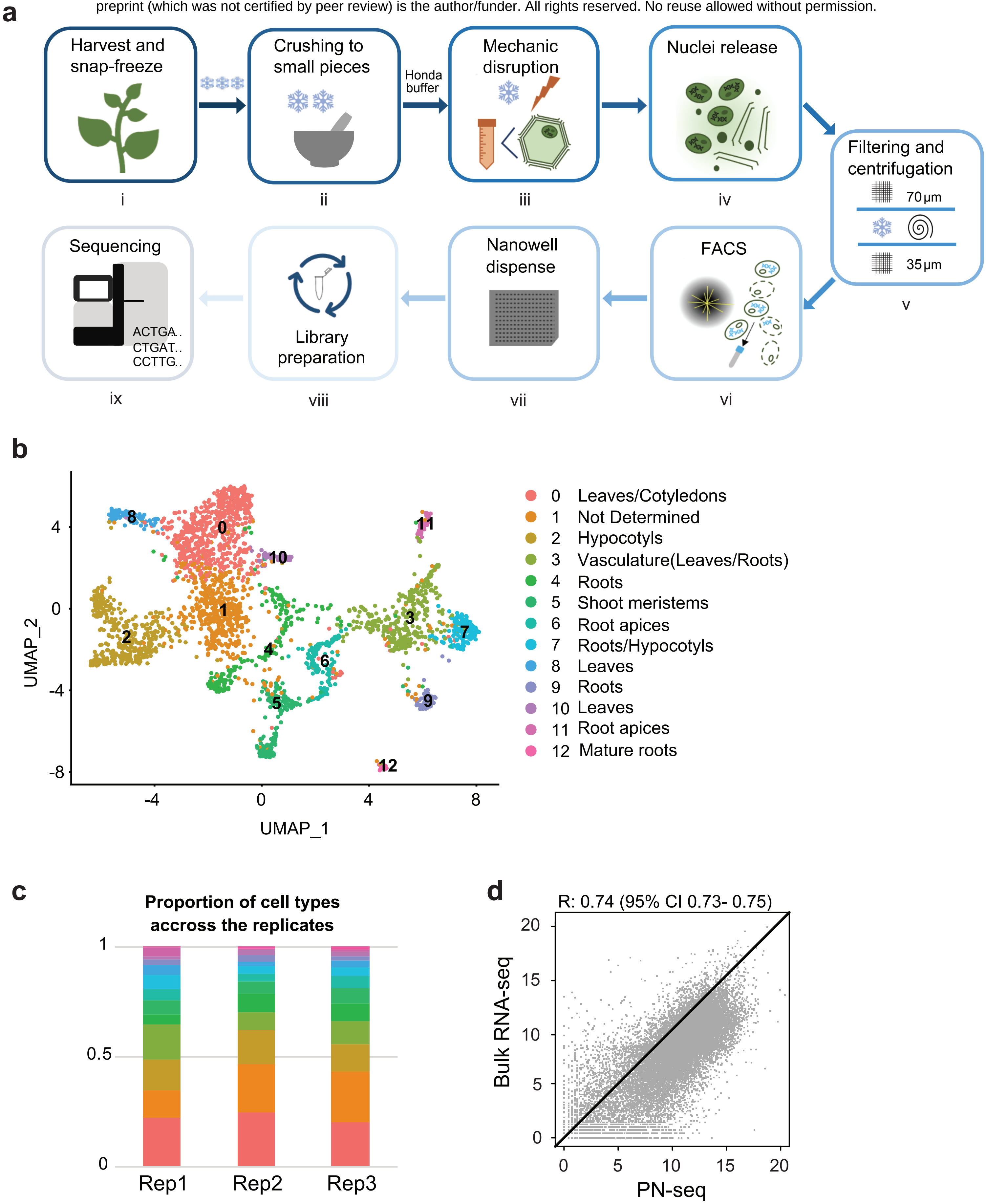
Figure 2: Single-nucleus RNA-sequencing. a) Schematic overview of PN-seq experimental strategy b) UMAP plot and clustering analysis of Arabidopsis seedlings samples (3 biological replicates, 13 clusters, 2,871 nuclei in total). c) Barplot showing that the three replicates have a similar proportion of nuclei across the identified clusters (the color code used to identify cluster cells is the same in b and c). d) Correlation (R= 0.74) of gene expression estimated from

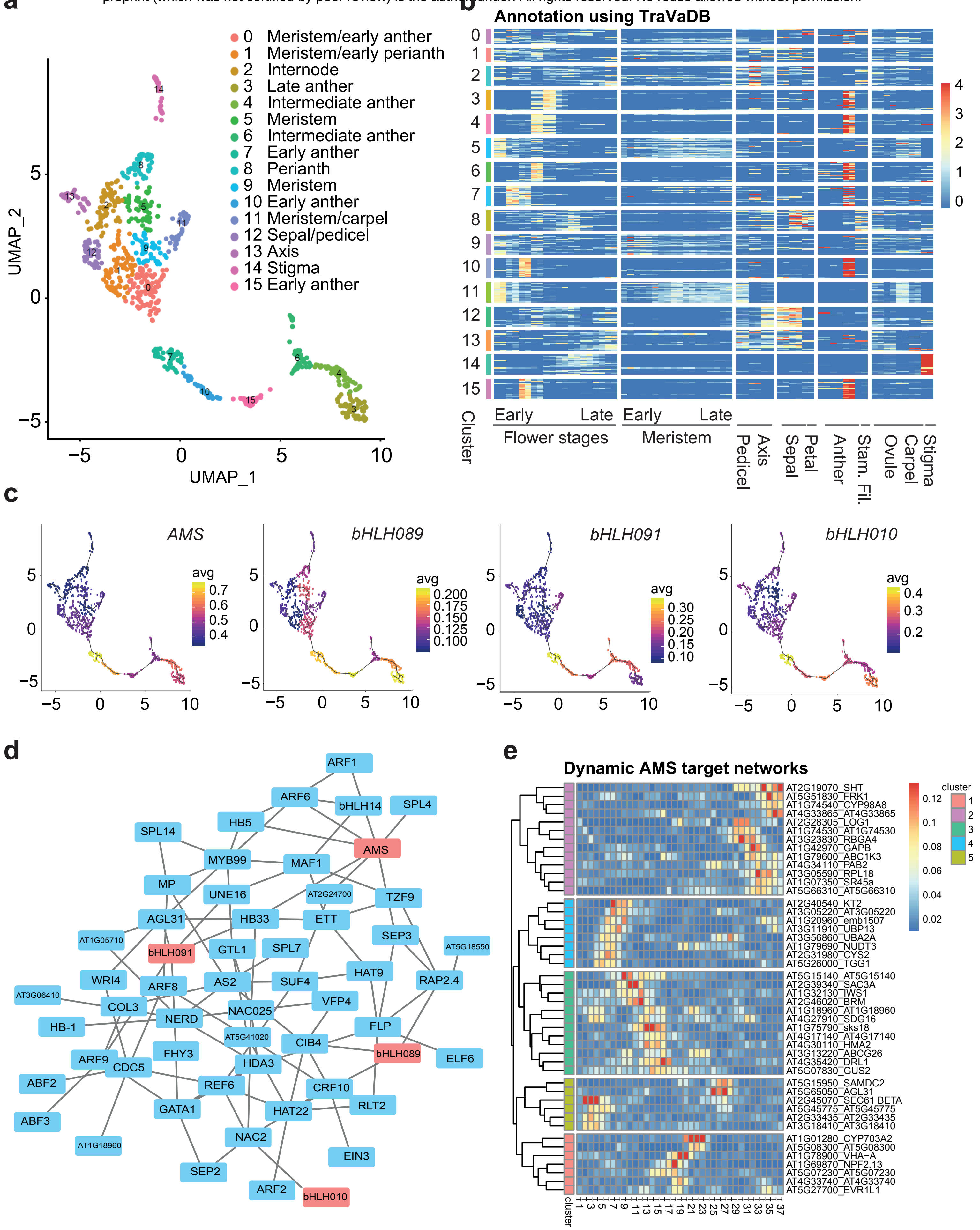
PN-seq (3 biological replicates) and bulk RNA-seq (3 biological replicates), indicating that PN-seq is able to recover similar transcriptomes than bulk RNA-seq.

Figure 3: Anther development at single-nuclei resolution. a) UMAP plot and clustering of the PN-seq data from Arabidopsis flowers before anthesis. b) Heatmap showing the expression of the top 20 significant marker genes for each cluster. c) Gene expression of known representative anther TF regulators *AMS*, *bHLH089*, *bHLH0901* and *bHLH010* plotted in the UMAP coordinates. d) Gene network estimated from cluster 15 (early anther) using GENIE3 (only TFs with more than 3 targets are shown). e) Heatmap showing the strength of the interaction between AMS and its target obtained by GENIE3 at different developmental stages. Namely, cells belonging to an anther cluster were ordered by their developmental stage predicted by Monocle3 pseudotime analysis (Supplementary Fig. 8c) and GRN networks were predicted independently for overlapping sets of 50 cells ordered by pseudotime; T1 is the first 50 cells (cluster 0, meristem/early anthers), and T37 is the latest stage (cluster 3, late anther).

Figure 4: Validation of cluster-specific marker genes with transcriptional reporter lines. a) Summary of expression-specificity validation of selected marker genes. Green dots indicate positive and grey dots indicate negative GFP signals at particular developmental stages or flower organs. Flower numbers (F4-F18) are after TraVaDB; flower developmental stages (S8-S12) are after Smyth *et al*. (26). b-i) Expression patterns of reporter lines: b) *AT5G08250*, anther from flower 16; c) *AT5G20030*, anther from flower 12; d) *AT2G16750*, anther from flower 11; e) *AT1G23520*, anther from flower 7; f) *AT1G63100*, meristem; g) *AT3G51740*, meristem; h) *AT4G11290*, stigma; i) *AT1G54500*, sepal. For the remaining confocal images see Supplementary Fig. 9. White arrowheads indicate exemplary GFP signals. Scale bars, 50 µm.


a
UMAP_2
UMAP_1
10
0
−10
−10
−5
0
5
10
13
21
3
22
6
19
12
8
23
16
7
1
14
11
2
5
9
15
4
20
18
17
10
0
0.15
0.10
0.05
0
−0.05
0.1
0
−0.1
0 2 4 6 8 10 12 14 16 18 20 22
Cluster
b
UMAP_2
UMAP_1
15
10
5
0
−5
−10
−10
−5
0
5
10
15
21
11
7
3
6
18
2
10
15
1
20
16
0
4
17
5
9
12
8
13
14
19
0.15
0.10
0.05
0
−0.05
0.1
0
−0.1
0 2 4 6 8 10 12 14 16 18 20
Cluster


a
Harvest and snap-freeze
i
Crushing to small pieces
ii
Honda buffer
Mechanic disruption
iii
Nuclei release
iv
Filtering and centrifugation
70 μm
35 μm
v
FACS
vi
Nanowell dispense
vii
Library preparation
viii
Sequencing
ACTGA..
CTGAT..
CCTTG..
ix
b
UMAP_2
UMAP_1
4
0
-4
-8
-4
0
4
8
0 Leaves/Cotyledons
1 Not Determined
2 Hypocotyls
3 Vasculature(Leaves/Roots)
4 Roots
5 Shoot meristems
6 Root apices
7 Roots/Hypocotyls
8 Leaves
9 Roots
10 Leaves
11 Root apices
12 Mature roots
c
Proportion of cell types accross the replicates
1
0.5
0
Rep1
Rep2
Rep3
d
R: 0.74 (95% CI 0.73- 0.75)
Bulk RNA-seq
PN-seq
0
5
10
15
20


a
0 Meristem/early anther
1 Meristem/early perianth
2 Internode
3 Late anther
4 Intermediate anther
5 Meristem
6 Intermediate anther
7 Early anther
8 Perianth
9 Meristem
10 Early anther
11 Meristem/carpel
12 Sepal/pedicel
13 Axis
14 Stigma
15 Early anther
UMAP_2
UMAP_1
5
0
−5
−5
0
5
10
b
Annotation using TraVaDB
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Cluster
Early
Late
Flower stages
Early
Late
Meristem
Pedicel
Axis
Sepal
Petal
Anther
Stam. Fil.
Ovule
Carpel
Stigma
4
3
2
1
0
c
AMS
avg
0.7
0.6
0.5
0.4
bHLH089
avg
0.200
0.175
0.150
0.125
0.100
bHLH091
avg
0.30
0.25
0.20
0.15
0.10
bHLH010
avg
0.4
0.3
0.2
5
0
−5
−5
0
5
10
d
ARF1
ARF6
bHLH14
SPL4
HB5
SPL14
AMS
MYB99
MAF1
MP
UNE16
AT2G24700
TZF9
AGL31
HB33
ETT
AT1G05710
SEP3
bHLH091
GTL1
SPL7
AT5G18550
WRI4
ARF8
AS2
SUF4
HAT9
RAP2.4
AT3G06410
COL3
VFP4
NAC025
NERD
HB-1
FLP
AT5G41020
CIB4
FHY3
bHLH089
ELF6
ARF9
CDC5
HDA3
ABF2
REF6
CRF10
GATA1
HAT22
RLT2
ABF3
NAC2
AT1G18960
EIN3
SEP2
ARF2
bHLH010
e
Dynamic AMS target networks
AT2G19070_SHT
AT5G51830_FRK1
AT1G74540_CYP98A8
AT4G33865_AT4G33865
AT2G28305_LOG1
AT1G74530_AT1G74530
AT3G23830_RBGA4
AT1G42970_GAPB
AT1G79600_ABC1K3
AT4G34110_PAB2
AT3G05590_RPL18
AT1G07350_SR45a
AT5G66310_AT5G66310
AT2G40540_KT2
AT3G05220_AT3G05220
AT1G20960_emb1507
AT3G11910_UBP13
AT3G56860_UBA2A
AT1G79690_NUDT3
AT2G31980_CYS2
AT5G26000_TGG1
AT5G15140_AT5G15140
AT2G39340_SAC3A
AT1G32130_IWS1
AT2G46020_BRM
AT1G18960_AT1G18960
AT4G27910_SDG16
AT1G75790_sks18
AT4G17140_AT4G17140
AT4G30110_HMA2
AT3G13220_ABCG26
AT4G35420_DRL1
AT5G07830_GUS2
AT5G15950_SAMDC2
AT5G65050_AGL31
AT2G45070_SEC61 BETA
AT5G45775_AT5G45775
AT2G33435_AT2G33435
AT3G18410_AT3G18410
AT1G01280_CYP703A2
AT5G08300_AT5G08300
AT1G78900_VHA-A
AT1G69870_NPF2.13
AT5G07230_AT5G07230
AT4G33740_AT4G33740
AT5G27700_EVR1L1
cluster
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
0.12
0.1
0.08
0.06
0.04
0.02
cluster
1
2
3
4
5

a

Genes	Prediction		Validation								
	Cluster	Organ/Stage	Flower Organs				Anther Development				
			Sepal	Stamen	Carpel (Stigma)	Meristem	S12 F4-5	S11 F6-8	S10 F9-11	S9 F12-14	S8 F15-18
AT1G63100	11	Meristem	●	●	●	🟢					
AT3G51740	11	Meristem	●	●	●	🟢					
AT1G54500	5	Meristem	🟢	●	●	●					
AT4G11290	14	Stigma	●	●	🟢	●					
AT5G08250	7	Anther/Early	●	🟢	●	●	●	●	●	●	🟢
AT5G20030	15	Anther/Early	●	🟢	●	●	●	●	●	🟢	●
AT2G16750	6	Anther/Intermediate	●	🟢	●	●	●	●	🟢	●	●
AT1G23520	3	Anther/Late	●	🟢	●	●	●	🟢	●	●	●