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] Max Delbrück Center for Molecular Medicine/Berlin Institute of Health, Berlin, Germany
[2] Humboldt-Universität zu Berlin, Institute of Biology, Systems Biology of Gene Regulation, Berlin, Germany
[3] Humboldt-Universität zu Berlin, Institute of Biology, Plant Cell and Molecular Biology, Berlin, Germany

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

Abstract

Single-cell genomics has slowly advanced in plant research. 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 several thousands of genes in thousands of individual plant cells. In this way, we expand the toolset for single-cell genomics in the field of plant biology to generate plant transcriptome atlases in development and environmental response.

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 have been 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 [5,6,7,8,9]. 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. PI-response genes can be identified through an independent bulk RNA-seq experiment to eliminate at least the most strongly affected genes from a scRNA-seq [9], as is shown in the Supplementary Fig. 1. In summary, there is an urgent need for alternative efficient single-cell genomics methods tailored for plant research.

Here, we introduce a single-nucleus sequencing protocol for plants (Fig. 1a; full protocol in Supplementary Materials and Methods). 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 [10,11,12,13]. As yet, there is no report of snRNA-seq methodology in plants.

Results and Discussion

Preparation of plant tissue and nuclei

Here we propose a single-nucleus sequencing strategy to detect nucleic acids derived from individual plant cells. The key step of our plant-nuclei sequencing (PN-seq) procedure consists of gentle but efficient isolation of plant nuclei. Plant tissue was frozen, gently physically dissociated and transferred to sucrose-rich, protease and RNAse inhibitor containing Honda buffer to support cell lysis. Cell walls and cell membranes were mechanically disrupted using a gentleMACS Dissociator, keeping the nuclei largely intact as detected by microscopy (Supplementary Fig. 2a, Materials and Methods). Finally, released intact plant nuclei were collected using Fluorescence Activated Cell Sorting (FACS), as demonstrated for a variety of samples including complex seedlings of the model plant *Arabidopsis thaliana*, as well as flowers of *Arabidopsis thaliana*, *Petunia hybrida* and *Antirrhinum majus* (snapdragon) and flowers and leaves from *Solanum lycopersicum* (tomato) (Fig. 1b and Supplementary Fig. 2b and c). The RNA that was isolated from nuclei was of higher quality than RNA that was conventionally purified from plant tissue, as observed by chromatrography analysis (Supplementary Fig. 2d).

Plant-nuclei (PN) sequencing of *Arabidopsis thaliana* seedlings

The next step consists of generating high quality cDNA libraries from the isolated nuclei. In principle, a number of library preparation and sequencing procedures can be combined. Of note, nuclei from *Arabidopsis thaliana* (~2 µm diameter) are much smaller than typical human or mouse nuclei (~10 µm diameter), and may contain only a fraction of the average mammalian nuclei RNA amount. We thus opted for a sensitive nanowell-based approach that includes lysis of nuclei by detergents and a freeze-thaw-cycle. In microdroplet-based single-cell RNA-sequencing methods such as the popular commercial 10x Genomics Chromium procedures, only a relatively mild lysis by detergents can be applied, since reverse transcription (RT) reactions take place in the same environment. Moreover, nanowells allow for selection of single-nucleus-containing wells and exclusion of no-nucleus- and multiple-nuclei-wells, thereby introducing additional quality control (Supplementary Materials and Methods). Using SMARTer ICELL8 3' chemistry, we prepared DNA libraries for short paired-end sequencing.

Raw sequencing data were preprocessed with ICELL8 mappa analysis pipeline and the R package Seurat v3 was used for downstream analysis [14]. Global properties of single-nuclei RNA-sequencing libraries including the number of sequenced, barcoded and mapped reads were summarized using hanta software from ICELL8 mappa analysis pipeline (Supplementary Data 1).

To validate our method, we set up our protocol using pools of *Arabidopsis thaliana* seedlings (3 biological replicates), which feature diverse plant structures comprising the radicle (embryonic root), the hypocotyl (embryonic shoot), and the cotyledons (seed leaves). On average, we obtained 1,116 nuclei per replicate and 2,802 genes per nucleus at ~220,000 reads per nucleus from these complex samples (Supplementary Data 1, Supplementary Fig. 3). Previous studies using PI of much less complex roots and droplet-based scRNA-seq obtained on average of 2,300 cells and the median number of 4,300 genes per cell [9]. As expected, the lower number of expressed genes per nuclei found by PN-seq can be explained by the fact that plant nuclei were small (~2 µm diameter) and thus contained less RNA. The ICELL8 system provides power for unbiased isolation of up to 1,800 single cells on a single chip, which can be upscaled by denser and/or bigger nanowell formats. As shown below, we achieved a reasonable number to perform further biology analysis, indicating the potential of PN-seq as a broadly applicable method.

Next, we analyzed the reproducibility of PN-seq. Correlation between the replicates was assessed using MA plots and Pearson's correlation, which ranged between 0.90 and 0.91 (Supplementary Fig. 4a-b). In silico pooling of our PN-seq data and subsequent correlation with gene expression data derived from conventional bulk RNA-sequencing resulted in Pearson correlation of 0.74 (Supplementary Fig. 4c, Supplementary Materials and Methods). This correlation is consistent with correlation coefficients found in previous publications (ranging from 0.7 to 0.85) [15] and therefore we found good agreement between both experiments. As shown in Supplementary Fig. 5, gene expression differences between both experiments were similar across all the chromosomes. No bias versus expression of specific sets of genes was observed.

Main organs and cell types of seedlings

The three seedling replicates were subsequently assessed using Seurat integration analysis, which initially revealed 13 distinct nuclei clusters (n=2,871) (Supplementary Fig. 6a, Supplementary Materials and Methods). Similar distribution of nuclei across clusters was observed, indicating that highly similar nuclei populations were recovered in each biological

replicate (Fig. 1c-d). The Wilcoxon rank sum test was applied to identify the significant marker genes of the clusters (Supplementary Table 1). Annotation of the initial 13 clusters was done based on the TraVaDB (Transcriptome Variation Analysis Database - http://travadb.org [17], Supplementary Materials and Methods), a plant gene expression resource, which resulted in 10 cluster labels that could be roughly classified into expected main basic organ types of seedlings: Leaves/Cotyledons (n=643 nuclei), Shoot meristems (n=180 nuclei), Hypocotyls (n=393 nuclei), Root apices (n=192 nuclei), Vasculature (Leaves/Roots) (n=342 nuclei), Roots (n=267 nuclei), Leaves (n=152 nuclei), Mature roots (n=27 nuclei), Roots/Hypocotyls (n=136 nuclei) and non-determined nuclei (n=539 nuclei) (Fig. 1e). The heatmap showing the expression of marker genes recovered from TraVaDB is displayed in Supplementary Fig. 6b. The entire annotation process is illustrated in Supplementary Fig. 6c. The clusters contain similar numbers of nuclei from each replicate, corroborating again the reproducibility of the method (Fig. 1f).

Next, we performed in depth analysis of root nuclei using 964 nuclei derived from root tissue of the seedlings (Root apices = 192 nuclei, Vasculature (Leaves/Roots) = 342 nuclei, Roots = 267 nuclei), Mature roots = 27 nuclei, Roots/Hypocotyls = 136 nuclei). The 964 nuclei were reorganized into 15 sub-clusters. The marker genes of the predicted 15 sub-clusters were compared to the list of markers from recently published atlas of the *Arabidopsis* root [5,9] (Supplementary Table 1, Supplementary Materials and Methods). PN-seq faithfully recovered major root cell types from our complex seedling dataset: mature, cortex/endodermis, stele, trichoblast, atrichoblast, endodermis and xylem (Supplementary Fig. 7). When looking at the expression of those genes in the RNA-seq based TraVa dataset, we notably observed the clusters 0, 1, 4, 10, 11, 12 and 14, in particular cluster 10 (trichoblast), enriched with multiple marker genes known from flowers, indicating i) the higher complexity of cell and organ type of inflorescence data compared to roots, ii) the still poor spatiotemporal resolution currently available, as TravA dataset is not cell-type specific and iii) the basic principle of the biology in which a gene may play role in multiple biological processes (Supplementary Table 1).

High similarity between fixed and unfixed material

In order to allow for more technical flexibility, i.e. the possibility to simplify the harvest and storage of plant samples, we fixed seedlings using methanol prior to our workflow (Supplementary Materials and Methods). High similarity across the samples from fixed and unfixed procedures was observed (Supplementary Fig. 8a-c) and the output was similar to the unfixed procedure: 850 nuclei and 2,292 genes (mean) per nucleus (Supplementary Fig. 3), implying that fixation of the material does not introduce major differences in the results. The

option of PN-seq to process frozen or methanol-fixed materials offers an additional advantage over protoplast-based procedures, since the latter requires immediate processing of fresh, unfrozen plant material.

Developmental flower stages covered by *Arabidopsis thaliana* inflorescences

Next, to study cell differentiation in plants, we applied PN-seq to *Arabidopsis thaliana* inflorescences covering all developmental stages 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. 9a), and with 14,690 genes expressed in at least five nuclei. The analysis identified 15 clusters corresponding to distinct organs and developmental stages (Fig. 2a; Supplementary Fig. 9b). In order to annotate these clusters with particular cell types, we first identified specific marker genes of each cluster (Supplementary Table 1), then plotted their expression profiles in the different floral organs and developmental stages obtained from TraVaDB [17] (Fig. 2b). Last, we correlated the gene expression of each cluster with each TraVaDB sample and indicated these values in the UMAP plot (Supplementary Fig. 9c). 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 six anthers per developing flower comprise a large fraction of floral tissues [18,19]. Furthermore, anthers/pollen express unique genes [18,19] which facilitates the bioinformatic identification of the clusters. Our data captured the developmental gene expression profiles during anther/pollen development from undifferentiated stem cells (cluster 0; Fig. 2) to late anther stages close to organ maturity, prior to anthesis (cluster 3; Fig. 2). Pseudotime analysis using Monocle 3 showed a strong concordance with anther developmental stages (Supplementary Fig. 10).

Regulatory link during anther and pollen development

Next, we analyzed how the gene regulatory network dynamically changes during anther and pollen development. Gene regulatory networks (GRNs) were inferred from transcriptome data using GENIE3 [20] to estimate the strength of interaction between known transcription factors (TFs) versus all the expressed genes for each cluster independently (Supplementary Fig. 11). One of the most connected TFs representing anthers was *ABORTED MICROSPORES* (*AMS*). *AMS* and the related TF genes *bHLH089*, *bHLH090* and *bHLH010* [21,22] were expressed in a highly dynamic manner (Fig. 2c, 2d). 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. 2e). Late targets included metabolic enzymes as well as genes associated with RNA-regulatory processes. Newly identified marker genes covered the full anther developmental trajectory (Supplementary Fig. 12) and are candidates for further mechanistic analyses.

To further validate our clustering analysis, we assessed the expression patterns of genes using promoter: GFP -NLS reporter lines. We selecteded genes with significant specificity and unknown function (Supplementary Fig. 13). In general, all genes showed expression in line with predictions. Specific expression in the floral meristem was observed for genes (AT1G63100, AT3G51740) from clusters 11, while genes AT4G11290 selected from cluster 14 showed highly specific stigma tissue expression. AT2G38995 was expressed in the sepals and petals, as expected for a marker from cluster 8, and it also showed slight expression in anthers. Genes AT5G20030, AT1G23520 and AT2G16750 were expressed in anthers and showed stage specificities that correlated with our analysis. Gene AT5G20030 from cluster 15, which is early anther cluster, showed expression in young anthers from flower 9 to flower 14. Finally, genes AT1G23520 and AT2G16750 were found in clusters of older anthers and indeed showed expression in old flowers (from 6-8 and 4-5), respectively.

In conclusion, PN-seq allowed for efficiently building transcriptome maps of plant samples and for studying at the level of individual cells dynamic GRNs during development, and revealed cell-type and stage-specific TF target pathways in an unprecedented manner.

Conclusion

Althought it is known that protoplast isolation (PI) procedure can significantly affect the plant transcriptome, it has been the basic choice for plant single-cell sequencing and had been mostly applied to roots samples [7,8,9,23,24]. Here, we introduced PN-seq that can be applied to analyze nucleic acids in bulk or in individual cells. Our new PN-seq methodology - based on efficient isolation of nuclei - is directly and easily applicable to a broad range of different plant tissues such as complex seedlings, flowers and leaves, and thus provides a versatile tool for multiple plant studies. In principle, various library preparation and sequencing methods can be combined with our 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 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 arsenal of nucleic acids 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].

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*. A step-by-step protocol for the preparation of plant tissues, nuclei and single-nucleus libraries as well as the steps for the data pre-processing analysis can be found at Protocol Exchange.

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 $MgCl_2$, 1µM DTT, 0.5% Triton X-100, 1 tablet/50mL cOmplete Protease Inhibitor Cocktail, 0.4 U/µl RiboLock, 25 mM Tris-HCl, pH 7.4). The M tubes were put onto a gentleMACS Dissociator and a specific program (Supplementary Table 3) 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 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 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 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 2nd half of 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 pa incubating at -20 °C overnight. Following a spin down at 4 °C, 13,000 rpm for 30 min and 2 washing steps with 70% Ethanol pa. The RNA pellet was dried on ice for 1 h and resuspended in 40 µl DEPC-H2O 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 by five seedlings each, were used separately through 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

The overall data analysis workflow is shown in Supplementary Fig. 14. 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. Mappa_demuxer assigned the reads to the cell barcodes present in the well-list file. 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 Seurat v3.

Quality control and data analysis

The downstream analysis started by removing the negative and positive controls included in all Takara Bio's NGS kits. For the seedling samples, R package Seurat v3 was used to filter viable nuclei, removing genes detected in less than 3 nuclei, nuclei with less than 200 genes, nuclei with more than 5% of mitochondria and nuclei with more than 5% of chloroplasts. Seurat *SCTransform* normalization method was performed for each one of the seedling replicates separately. Data from three 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. 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; Supplementary Fig. 6a). *SCTransform* and *RunPCA* 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 and using the RNA assay (normalized counts). The annotation of the clusters was based on the top 20 markers of each cluster.

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 filtered out. The filtering step resulted in a dataset containing 856 nuclei and 14,690 genes. Seurat v3 *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 (Supplementary Fig. 8b). 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. The annotation of the clusters was based on the top 20 markers of each cluster.

Annotation

Annotation of the seedling and flower clusters was based on TraVaDB (Transcriptome Variation Analysis Database, http:/travadb.org). TraVaDB is an open-access database based on RNA-Seq data, which includes 79 samples, each with at least two biological replicates, corresponding to different developmental stages and parts of roots, leaves, flowers, seeds, siliques and stems. The top 20 differentially expressed genes of each cluster was used as input for the analysis with TraVaDB [17]. The complete TraVaDB was downloaded. The heatmaps showing the tissue types in which genes were found expressed were created using a R script.

For the annotation of the root clusters, we developed a function in R (available at https://github.com/ramonvidal/punyplatypus). This function processes the output of Seurat *FindAllMarkers* and predicts cell-type(s) for each cluster based on the match between the genes from the input list and the marker genes from a reference list containing one or more single-cell experiments. An adjusted p-value (by Bonferroni correction for multiple hypothesis testing) and a PPV (positive predictive value) are calculated for each predicted cell-type. The smaller

the p-value the bigger the evidence the genes is a cell-type specific marker. The PPV describes the performance of the prediction. It represents the proportion of positive results that are truly positive. High PPV can be interpreted as indicating the accuracy of the prediction. Only differentially expressed genes with adjusted p-value <= 0.05 were used as input for the punyplatypus. The reference list was created using marker genes from recently published single-cell RNA expression data of the *Arabidopsis* root [5,9]. They are listed in Supplementary Table 1.

Reproducibility and correlation

To assess technical reproducibility of our protocol, we used MA plots to evaluate the variability in each pair of seedling replicates. We compared three replicates against one another, resulting in three comparisons. Pearson's correlation coefficient was calculated across seedling replicates. The consistency between bulk and PN-seq experiments was investigated throught the comparison between the log2 mean expression of genes detected in both experiments. Expression of bulk RNA-seq data was quantified with RSEM[25]. They are listed in Supplementary Table 3.

Network analysis

GENIE3 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 [26] and cluster 0 (meristem/Early anther) as root cluster. Next, gene networks were estimated with GENIE, as described before, 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 locatisation 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 Supplemental Table 4. Afterwards, the LR reactions were performed with the binary vector pGREEN:GW:NLS-GFP (Smaczniak et al. 2017) to generate GFP transcriptional fusions to a nuclear-localization signal 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.

Data availability

All relevant data have been deposited in EBI ArrayExpress, accession number E-MTAB-9174.

Code availability

R function developed for the annotation of the root clusters is available on Github: https://github.com/ramonvidal/punyplatypus).

Acknowledgements

The study was supported by the European Commission (FP-7, grant agreement no. 262055, ESGI, and Horizon 2020, grant agreement no. 824110, EASI-Genomics) (S.S.), and by DFG grant KA 2720/5-1 (X.X, K.K). We thank Solenne Bourdier and Dijun Chen for support.

References

1. Fischer, C. *et al.* Signals trigger state-specific transcriptional programs to support diversity and homeostasis in immune cells. *Sci. Signal.* **12**, 1–17 (2019).
2. Regev, A. *et al.* The Human Cell Atlas. *Elife* **6**, e27041 (2017).
3. Rhee, S. Y., Birnbaum, K. D. & Ehrhardt, D. W. Towards Building a Plant Cell Atlas. *Trends Plant Sci.* **24**, 303–310 (2019).
4. Efroni, I. & Birnbaum, K. D. The potential of single-cell profiling in plants. *Genome Biol.* **17**, 65 (2016).
5. Efroni, I., Ip, P.-L., Nawy, T., Mello, A. & Birnbaum, K. D. Quantification of cell identity from single-cell gene expression profiles. *Genome Biol.* **16**, 9 (2015).
6. Efroni, I. *et al.* Root Regeneration Triggers an Embryo-like Sequence Guided by Hormonal Interactions. *Cell* **165**, 1721–1733 (2016).
7. Zhang, T.-Q., Xu, Z.-G., Shang, G.-D. & Wang, J.-W. A Single-Cell RNA Sequencing

Profiles the Developmental Landscape of Arabidopsis Root. *Mol. Plant* **12**, 648–660 (2019).

8. Jean-Baptiste, K. *et al.* Dynamics of gene expression in single root cells of arabidopsis thaliana. *Plant Cell* **31**, 993–1011 (2019).
9. Denyer, T. *et al.* Spatiotemporal Developmental Trajectories in the Arabidopsis Root Revealed Using High-Throughput Resource Spatiotemporal Developmental Trajectories in the Arabidopsis Root Revealed Using. *Dev. Cell* **48**, 840-852.e5 (2019).
10. Lake, B. B. *et al.* 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 (2017).
11. Macaulay, I. C., Ponting, C. P. & Voet, T. Single-Cell Multiomics: Multiple Measurements from Single Cells. *Trends Genet.* **33**, 155–168 (2017).
12. Selewa, A. *et al.* Systematic Comparison of High-throughput Single-Cell and Single-Nucleus Transcriptomes during Cardiomyocyte Differentiation. *Sci. Rep.* **10**, 1–13 (2020).
13. Ding, J. *et al.* Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat. Biotechnol.* **38**, 737–746 (2020).
14. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
15. Beisang, D. J. *et al.* 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 (2020).
16. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11**, 740–742 (2014).
17. Klepikova, A. V., Kasianov, A. S., Gerasimov, E. S., Logacheva, M. D. & Penin, A. A. A high resolution map of the Arabidopsis thaliana developmental transcriptome based on RNA-seq profiling. *Plant J.* **88**, 1058–1070 (2016).
18. Smyth, D. R., Bowman, J. L. & Meyerowitz, E. M. Early flower development in Arabidopsis. *Plant Cell* **2**, 755–767 (1990).
19. Gómez, J. F., Talle, B. & Wilson, Z. A. Anther and pollen development: A conserved developmental pathway. *J. Integr. Plant Biol.* **57**, 876–891 (2015).
20. Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. & Geurts, P. Inferring regulatory networks from expression data using tree-based methods. *PLoS One* **5**, 1–10 (2010).
21. Xu, J. *et al.* The ABORTED MICROSPORES regulatory network is required for postmeiotic male reproductive development in Arabidopsis thaliana. *Plant Cell* **22**, 91–107 (2010).
22. Zhu, E. *et al.* The DYT1-interacting proteins bHLH010, bHLH089 and bHLH091 are redundantly required for Arabidopsis anther development and transcriptome. *Plant J.* **83**, 976–990 (2015).
23. Shulse, C. N. *et al.* Profiling of Plant Cell Types Resource Profiling of Plant Cell Types. *CellReports* **27**, 2241-2247.e4 (2019).
24. Ryu, K. H., Huang, L., Kang, M. & Schiefelbein, J. Single-Cell RNA Sequencing Resolves Molecular Relationships Among Individual Plant Cells 1 [OPEN]. **179**, 1444–1456 (2019).
25. Li, B. & Dewey, C. N. RSEM: Accurate transcript quantification from RNA-seq data with or without a reference genome. *Bioinforma. Impact Accurate Quantif. Proteomic Genet. Anal. Res.* 41–74 (2014) doi:10.1201/b16589.
26. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).

Acknowledgements

The study was supported by the European Commission (FP-7, grant agreement no. 262055, ESGI, and Horizon 2020, grant agreement no. 824110, EASI-Genomics) (S.S.), and by DFG grant KA 2720/5-1 (X.X, K.K). We thank Solenne Bourdier and Dijun Chen for support.

Author information

Conception and design: S.S. and K.K.; Protocol optimization to obtain plant nuclei: W.Y.,X.X.,C.S., C.B. C.F., and S.S.; Experimental work: X.X. and C.B.; FACS analysis: C.B.; Library preparation for single-cell and bulk RNA-sequencing: C.B. and C.S.; Nuclei Isolation and selection by ICELL8 Imaging System: X.X.,C.B. and C.S.; Validation with transgenic lines and confocal imaging: X.X. and M.Z.; Provision of tools and materials: K.K., M.K. and S.S. Data analysis: D.Y.S.F. and J.M.M.; R function development: R.V.; Data interpretation: D.Y.S.F., J.M.M., C.S., K.K. and S.S; Drafting of the manuscript: S.S. and D.Y.S.F.; Manuscript writing, review and edition: all authors. Funding and supervision: K.K. and S.S.

Ethics declarations

The authors declare no competing interests.

Figures

Figure 1: Single-nucleus RNA-sequencing. a) Schematic overview of the here-applied experimental strategy consisting of i) harvest and snap-freeze of plant material, ii) crushing to small pieces in liquid nitrogen, iii) mechanic disruption in Honda buffer (with a gentleMACS Dissociator), iv) nuclei release, v) filtering and centrifugation, vi) FACS, vii) nanowell-based single-nucleus preparation, viii) library preparation and finally ix) sequencing. After nuclei isolation (FACS) alternative experimental approaches are conceivable to produce sequence data. b) FACS histogram plots of DAPI fluorescent nuclei from different plants and different tissue types: *Arabidopsis thaliana* seedlings and flowers, *Petunia hybrid* flowers, *Antirrhinum majus* flowers (snapdragon), *Solanum lycopersicum* (tomato) flowers and leaves are shown after conventional gating for rough debris exclusion and doublet discrimination. The grey filled sections represent the gate that was set for sorting. The different tissue types produce different amount of nucleus-like sized, low DAPI-fluorescent debris that can only be separated from intact nuclei by gating the high DAPI-fluorescent peaks (Suppl. Fig. 2). c-d) UMAP plots showing the reproducibility among three independent biological replicates of *Arabidopsis thaliana* seedlings. e) UMAP plot of 2,871 nuclei showing the single-nuclei cluster by identity.

f) Barplot showing the proportion of nuclei corresponding to the identified cluster across the biological replicates.

Figure 2: Anther development at single-nuclei resolution. a) UMAP plot of the flower PN-seq data. b) Heatmap showing the expression of the top 20 significant markers genes for each cluster. c) Gene expression of known anther regulators *AMS*, *bHLH089*, *bHLH090* and *bHLH010* plotted in the UMAP coordinates. d) Gene network estimated from cluster 15 (early anther) using GENIE3 (Supplementary Materials and Methods); only top 5,000 interactions were used, and 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 across overlapping sets of 50 cells from anther clusters ordered by pseudotime; T1 is the first 50 cells (cluster 0, meristem/early anthers), and T37 is the latest stage (cluster 3, late anther).

Supplementary Information

Supplementary Figures

Suppl. Fig. 1: Effect of protoplast isolation (PI) on root scRNA-seq. a) Expression correlation of each cell from scRNA-seq with bulk RNA-seq sample using PI and with bulk RNA-seq sample without using PI (y and x-axis, respectively; Denyer *et al.* data [9]). Cells have different response to protoplasting, with some groups of cells being more sensitive to protoplasting (higher correlation with PI data than with data from intact tissues) than others. b) Re-analysis of scRNA-seq full data and c) re-analysis of scRNA-seq data removing the top 6,000 PI-responsive genes (Denyer *et al.* data [9]) from the clustering step. In b) and c), 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 with bulk RNA-seq PI and non-PI samples. Thus, positive correlation numbers indicates 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. When no PI-responsive genes were removed (b) we observed several clusters containing cells with strong response to PI, with the most extreme cluster with up to 55% of the top 20 marker genes being PI-responsive genes. This effect largely persisted when PI-responsive genes were excluded from the primary scRNA-seq analysis. After the exclusion of PI-responsive genes from the clustering step, but still using them to identify markers, we observed the most extreme cluster with up to 46% of the top 20 marker genes being PI-responsive genes (c). These results highlight a need for alternative methods beyond PI for plant single-cell genomics.

Suppl. Fig. 2: Generic single-nuclei isolation procedure. a) Microscopy analysis. Sections from disposable Neubauer counting chambers with DAPI stained nuclei after FACS from *Arabidopsis* thaliana seedlings and flowers, *Petunia hybrid* flowers, *Antirrhinum majus* flowers (snapdragon), *Solanum lycopersicum* (tomato) flowers and leaves. The brightfield images are overlain with the blue fluorescence images. Images of *Arabidopsis thaliana* samples were captured with a DMi8 microscope by Leica and the others by a BZ-X700 Series microscope by Keyence. The images show that FACS yields clean, debris-free nuclei suspensions irrespective of the initial amount of debris. b) Contour plots of flow cytometry experiments. c) Gating strategy used for flow cytometry, exemplified for *Arabidopsis thaliana* flower (inflorescence) samples. d) Quality control of RNA nuclei samples through chromatrography-based analysis of RNA derived from pooled nuclei of *Arabidopsis thaliana* seedlings and RNA derived from conventional purification (RIN = RNA Integrity Number). B1 corresponds to bulk RNA from tissue and C1 corresponds to RNA from sorted nuclei.

Suppl. Fig. 3: Summary of PN-seq seedling datasets. Violin plots showing the total number of detected genes (nFeature), reads counts (nCount), proportion of mitochondria (percent.mt) and chloroplast (percent.ch) contamination per nucleus for each replicate.

Suppl. Fig. 4: Correlation among PN-seq replicates and between PN-seq and bulk RNA-seq libraries. a) MA-plot showing the differences between samples: Replicate 1 versus Replicate 2, Replicate 1 versus Replicate 3 and Replicate 2 versus Replicate 3, plotted against the average gene count value (A). The red line shows the average differences. b) Scatterplot of gene expression obtained by different PN-seq replicates. c) Scatterplot of gene expression obtained by pooling the three PN-seq replicates against the bulk RNA-seq. Read counts from both datasets, PN-seq and bulk RNA-seq, were log2 transformed.

Suppl. Fig. 5: Comparison of bulk-RNA and PN-seq data indicates no expression bias of specific gene groups. Expression of top 25% genes with highest variance between bulk (red boxes) and PN-seq (green boxes) experiments. *A. thaliana* chromosomes are shown on the x-axis and the log2 mean expression of the genes in 3 replicates are shown on the y-axis. The thick line in the middle of the box represents 50% of observations, with the lower end of the box at 25% and the upper end of the box at 75%.

603 Suppl. Fig. 6: Annotation of seedling cell-clusters using TraVaDB dataset. a) UMAP plot of
604 2,871 nuclei organized in 13 clusters before the annotation. b) Heatmaps of the 13 clusters
605 showing the expression level of the top 20 differentially expressed genes from each cluster in
606 the TraVaDB dataset. c) Illustration of the cluster annotation process.
607
608 Suppl. Fig. 7: Single-nuclei transcriptome analysis of a subset of root cells derived from
609 seedling data. a) UMAP of 15 clusters (n=964 nuclei). Eight clusters were faithfully annotated
610 using punyplatypus function: cluster 0 - mature (p-value=7.785874e-04, PPV=0.55), cluster 1
611 - endodermis (p-value=1.550274e-13, PPV=0.86), cluster 1 - cortex (p-value=1.689028e-10,
612 PPV=0.71), cluster 4 - stele (p-value=8.827076e-12, PPV=0.90), cluster 9 - mature (p-
613 value=5.837800e-10, PPV=1.0, cluster 10 - trichoblast (p-value=1.655356e-92, PPV=0.90),
614 cluster 11 - trichoblast (p-value=3.641153e-11, PPV=0.47), cluster 12 - endodermis (p-
615 value=2.486261e-05, PPV=0.54) and cluster 14 - xylem (p-value=8.777430e-37, PPV=0.94).
616 Punyplatypus calculates a p-value and PPV per cell type, which indicate the performance of
617 annotation. The smaller the p-value the bigger the evidence that the genes are cell-type specific
618 markers. High PPV can be interpreted as indicating the proportion of genes in a cluster found
619 annotated with the same cell type in the reference list. b) Violin plots showing expression of
620 marker genes per annotated cluster. c) Heatmap corroborating the annotation by punyplatypus.
621 It shows the top 100 markers of Denyer *et al.*[9] among the top 1000 expressed genes per nuclei
622 of PN-seq. Almost all cell types could be recovered using a subset of root nuclei from the
623 complex seedling data.
624 Suppl. Fig. 8: Correlation between unfixed and fixed seedling samples. a) UMAP plot showing
625 similar cell distribution of unfixed and fixed seedling samples. b) UMAP plot showing the
626 overlapping between cells from unfixed and fixed seedling samples. c) Pearson's correlation
627 coefficient across unfixed and fixed seedling samples. The thick line in the middle of the box
628 represents 50% of observations (log2 of read counts), with the lower end of the box at 25% and
629 the upper end of the box at 75%. Bars extend to the lowest and highest values that are not
630 outliers. The correlation found over all samples was 0.90.
631
632 Suppl. Fig. 9: Single-nuclei transcriptome analysis of *Arabidopsis thaliana* flower
633 development. a) Number of genes per nuclei (nFeature) and number of reads per gene (nCount).
634 b) JACKSTRAW plot to identify the optimal number of PCAs for the analysis of the
635 inflorescence dataset. c) Annotation of clusters based on correlation: the average gene
636 expression of each cluster was Spearman correlated against each one of the TraVaDB

transcriptome dataset considered. The two labels plotted on top of each cluster indicate the two TraVa samples with highest correlation.

Suppl. Fig. 10: Temporal trajectories in the floral PN-seq dataset. a) Annotation of clusters based on correlation with the "stages" samples in the TraVa dataset. b and c) Pseudotime analysis using Monocle 3. d) Flower developmental stages recovered from the TraVa dataset.

Suppl. Fig. 11: Main master regulators in flower cells. GENIE3 was used to estimate the gene network of each cluster. The heatmap shows the number of predicted target genes for each TF among the top 10,000 strongest interaction in the network. Only the top 4 TFs with most predicted target genes per cluster are shown.

Suppl. Fig. 12: Novel marker genes covering the developmental trajectory of anther/pollen development.

Suppl. Fig. 13: a) Summary of functional validation. Clearly visually validated genes are indicated by green dots, whereas grey dots indicate negative results. b) Validation for cluster-specific genes with transcriptional reporter lines. Expression patterns of reporter lines for the following genes reveal the predicted floral organ specificities: (a) AT1G63100, floral meristem; (b) AT3G51740, floral meristem; (c) AT4G11290, stigma; (d) AT2G38995, sepal; (e) AT2G38995, petal; (f) AT2G38995, anther. Expression patterns of reporter lines for the following genes reveal the predicted stage specificities for anther development: (g) AT5G20030, flower 11-13; (h) AT1G23520, flower 6-8; (i) AT2G16750, flower 4-5. White arrowheads indicate GFP signals in nuclei;. Scale bars, 50 µm.

Suppl. Fig. 14: PN-seq data analysis pipeline. Blue boxes represent the main analysis steps, the software used in each step is shown in italic and the steps using *mappa* analysis pipeline are shown in the dashed gray box. The in/output files from each step of analysis are represented with file icons.

Supplementary Tables:

Suppl. Table 1: List of marker genes in *Arabidopsis thaliana* seedlings, roots and flowers.

Suppl. Table 2: Program steps for gentle tissue disruption and nuclei release on the gentleMACS Dissociator and the instrument specific commands. The instrument specific tubes

have got a stator and rotor. Latter can be moved at certain a certain speed (rpm = rounds per minute) for a certain time (time in s) in several rounds (loops).

Suppl. Table 3: Sequencing data metrics and read counts of PN-seq and bulk RNA-seq libraries.

Supplementary Data:

Suppl. Data 1: Summary and read statistics of the PN-seq data (adapted from the html reports generated by ICELL8 hanta software)