www.sciencemag.org/content/364/6435/52/suppl/DC1

Supplementary Material for

Defining the developmental program leading to meiosis in maize

Brad Nelms* and Virginia Walbot*

*Corresponding author. Email: bnelms.research@gmail.com (B.N.); walbot@stanford.edu (V.W.)

Published 5 April 2019, *Science* **364**, 52 (2017)
DOI: 10.1126/science.aav6428

This PDF file includes:

Materials and Methods
Figs S1 to S13
Tables S6 and S7
References

Other Supplementary Material for this manuscript includes the following:
(available at www.sciencemag.org/content/364/6435/52/suppl/DC1)

Tables S1 to S5 as separate Excel files

Materials and Methods

Experimental methods:

Plant material

Maize (*Zea mays*) inbred line W23 *bz2* was grown in a greenhouse at Stanford, CA with a 14-h day / 10-h night light cycle. For single cell RNA-seq experiments, plants were harvested between 3:30 pm – 4:30 pm (9.5 to 10.5 h into the day cycle) and processed immediately to reduce variability caused by circadian changes in gene expression.

Tissue dissociation

All tissue handling procedures were miniaturized so that cells could be isolated from a single floret. Anthers were dissected from the middle third of the central tassel spike, and anther length was measured using a stage micrometer (Fisher Scientific). Anthers from a single upper floret were cut with a #11 scalpel every 0.5 mm and placed on a microscope slide in a 20 µL drop of anther protoplasting buffer (10 mM MES pH 5.7, 400 mM Trehalose, 2 mM CaCl2, 10 mM KCl, 0.1% BSA) with protoplasting enzymes: 1.25% w/v Cellulase-RS, 0.5% Macerozyme R10, 0.5% Hemicellulose, and 0.5% Pectolyase Y23 (Sigma-Aldrich catalog numbers C0615, P2401, H2125, and P5936, respectively). A second microscope slide was placed on top to minimize evaporation, using a thin piece of tape on both sides and two #1.5 coverslips to provide sufficient space between the slides to avoid crushing the anthers. Slides were incubated at 31 ºC for 50 minutes, and then coverslip spacers were removed and shear force was applied by moving the top microscope slide back and forth 5-6 times to release anther protoplasts (space was still maintained between the slides by the thin pieces of tape). The use of trehalose as an osmoprotectant instead of mannitol was important for miniaturization: mannitol was prone to crystallization at the edge of small liquid droplets, while trehalose was not. Furthermore, trehalose is compatible with all downstream enzymatic steps.

Cell isolation

Pre-meiotic and early meiotic protoplasts were identified by their large size, picked up with a blunted 29 gauge insulin needle (BD Biosciences), and washed twice in protoplasting buffer without $CaCl_2$. Single washed protoplasts were then picked up with a blunted 33 gauge syringe needle (Sigma-Aldrich CAD4112) and placed on the cap of an 8-tube PCR strip (Axygen Low Profile 8-Strip PCR Tubes; Fisher Scientific 14-223-505). The presence of an isolated single cell without attached debris was confirmed microscopically (10X magnification, Nikon Diaphot) during both cell pick-up and release. Each isolated cell was photographed, providing both an image with which to measure cell size and a time stamp for when each cell was isolated. Based on the diameter of the droplets visible in these images, we estimate that cells were isolated in a typical buffer volume of 10-50 pL. After isolating each set of 8 cells, the PCR caps were attached to PCR tubes and flash frozen in liquid nitrogen. Frozen cells were stored at -80 ºC.

Illumina library preparation

Sequencing libraries were prepared using a protocol based on CEL-seq2 *(21),* with three modifications: (i) the lysis step was modified to enable split-cell technical replicates, (ii) comparable but different enzymes (e.g. different companies) were used, either for convenience or because the enzymes were found to be more efficient with maize RNA, and (iii) the unique molecular identifying (UMI) length was increased from 6 to 10 nucleotides. To avoid correlations between batch effects and biological information, sample locations and pooling schemes were randomized at several points during library synthesis (Table S1).

Briefly, frozen cells were thawed and immediately lysed by adding 2.1 µL lysis buffer (0.1% Triton X-100, 1.5 mM dNTP mix, 1:315,000 Ambion ERCC spike in controls) and incubating at 65 ºC for 2 minutes. Then 1 µL (half) of the cell lysate was transferred to a fresh tube and 1 µL 1.5 µM barcoded oligo(dT) primer was added to each split-cell sample. Samples were incubated at 65 ºC for 3 minutes then placed on ice. Reverse transcription was initiated by adding 1 µL of Life Technologies SuperScript IV reverse transcription mix (4:1:1:1 ratio of 5x Superscript IV buffer, RnaseOUT, Superscript IV, 100 mM DTT) and incubated at 42 ºC for 2 minutes, 50 ºC for 10 minutes, 55 ºC for 5 minutes, and 70 ºC for 10 minutes. Excess primers were digested with exonuclease 1 (New England Biolabs).

Next, second strand synthesis was performed using the NEBNext mRNA Second Strand Synthesis Module (New England Biolabs). Sets of 16 individually barcoded split-cell samples were pooled and DNA was purified using a 1.2 ratio of AMPure XP beads (Beckman Coulter). Samples were then amplified with the MEGAscript T7 *in vitro* transcription kit (Thermo Fisher Scientific) at 16 ºC overnight. Amplified RNA was purified with RNAClean XP beads (Beckman Coulter), and a second round of reverse transcription was performed with SuperScript II reverse transcriptase (Life Technologies) using a tagged random hexamer primer (5'-GCCTTGGCACCCGAGAATTCCANNNNNN). Libraries were then amplified with 10-15 rounds of PCR using the NEBNext Ultra II Q5 Master Mix (New England Biolabs) and Illumina primers, and libraries were purified and size selected with AMPure XP beads. Library quantity and size distribution was assessed with an Agilent Biolanalyzer High Sensitivity DNA kit.

Libraries were sequenced on an Illumina HiSeq 2500 instrument at the University of Delaware Sequencing and Genotyping Center with paired-end 76 bp reads. Primer sequences can be found in Table S6. All primers were synthesized by phosphoramidite chemistry at the Stanford Protein and Nucleic Acid Facility (Stanford, CA).

Quantitative PCR (qPCR)

Two of three anthers from a floret were placed on a cleaned microscope slide in 10 µL water. Anthers were then cut in half with a #11 scalpel and meiotic cells were gently squeezed out. Remaining anther material was removed, and meiotic cells with excess water were placed in a PCR tube (Axygen) and flash frozen in liquid nitrogen. Extruded meiocytes were lysed with 0.1% Triton X-100 (Thermo Fisher Scientific) and reverse transcription was performed without further purification using SuperScript IV reverse transcriptase (Life Technologies) and oligo$(dT)_{12-18}$ primer (Thermo Fisher). Quantitative PCR was performed using TaqMan primers synthesized by Integrated DNA Technologies

(Table S7) on a BioRad CFX instrument. Three to 4 replicates were performed per condition, then cycle threshold (Ct) values for replicate samples were averaged and the mean Ct of the reference gene (*pco080082*) was subtracted.

Cytological staging and image acquisition

One of three anthers from a floret was fixed in 4% paraformaldehyde at room temperature for 2 h. Anthers were washed and then fixed meiotic cells were extruded in Prolong Diamond antifade media (Life Technologies) with 10 µg/mL Hoechst 33342 (Sigma-Aldrich). Finally, a #1.5 coverslip (Zeiss) was placed on top. Confocal image stacks were taken with a Leica SP8 microscope using a 63X 1.4 n.a. oil immersion objective, a 405 nm excitation laser, an AOTF emmission filter set to capture light between 495 nm and 562 nm, and a Nyquist resolution voxel size of 58.2 nm by 58.2 nm by 299 nm. All images were processed uniformly in ImageJ *(22)* Fiji *(23)* by applying rolling ball background subtraction with a 50 pixel radius and then linearly scaling the maximum and minimum intensity values.

Cytological stage was scored using established criteria *(11)*: (i) premeiotic interphase / early leptotene – centrally located nucleolus, round knobs, chromatin threads not visible or visible as localized patches; (ii) leptotene stage 1 – centrally located nucleolus, round knobs, chromatin threads condensed throughout the nuclear volume; (iii) leptotene stage 2 – same as leptotene stage 1 except the nucleolus was located at the nuclear periphery; (iv) prezygotene – same as leptotene stage 2 except the knobs were clearly elongated. Cytological scoring was performed blinded to sample identity (i.e. without any knowledge of the anther length or qPCR staging).

Computational methods:

Library mapping

One of each paired-end sequencing read contained transcript sequence; these reads were mapped to the B73 reference genome (AGPv4) using Hisat2 *(24)*. The second read pair contained cell barcodes and a 10 bp unique molecular identifier *(25)* (UMI). For transcript counting, read-pairs that mapped to the same gene with one or fewer bp differences in their UMIs were counted as originating from a single first-strand cDNA molecule. Allowing for a 1 base error in the UMIs reduced over-counting artifacts resulting from sequence read errors.

During library mapping, two errors were detected in the cell barcode reverse transcription primers. First, primer ‘CelSeq_dT-1s’ had an extra nucleotide added to the cell barcode (Table S6), which was apparently the result of a systematic primer synthesis error; the cell barcode was updated to account for this synthesis error and no data were lost. Second, very few reads mapped to the primer ‘CelSeq_dT-4s’, which could be attributed to a second primer synthesis error or a mistake during primer dilution; samples with primer dT-4s were lost from the dataset, and so a subset of cells do not have technical replicates (16 cells out of the 144 that passed QC criteria). Cells without technical replicates were not enriched for any particular anther length. For analyses

comparing technical replication (e.g. Fig. 2A, Fig S4), the 128 cells with replicates were sufficient to assess reproducibility.

Quality control

Before applying quality control (QC) criteria, data from split-cell technical replicates were pooled together. Pooling split-cell replicates for QC was important because technical replicates were later used to assess analysis quality, and so blinding QC to replicate information avoided selecting for cells where the technical replicates were in agreement. For QC, cells were excluded if under 71.8% of mapped reads mapped to the 10 maize chromosomes; the 71.8% cutoff was chosen because it is 1.5 median absolute deviations from the population median (Fig. S3A). This QC criterion excluded cells with a high proportion of reads mapping to the ERCC spike-in controls and/or the organellar (mitochondria/chloroplast) genomes. Four additional cells were excluded because their transcript abundance was poorly correlated with every other cell in the dataset (Pearson's correlation < 0.6; Fig. S3B). These quality control criteria did not substantially enrich for cells at any given anther length.

Certain features of our data suggest possible causes of low quality cells. The majority of cells that failed QC (57/61 cells; 93%) were removed because a low fraction of reads mapped to the 10 maize chromosomes. Second, if we compare technical replicates for the cells that failed the QC criteria, we found that in all cases both replicates would have individually failed QC; thus, we conclude that the causes of low quality cells occurred prior to splitting the cell lysate for technical replication. Two possible explanations for these observations are (i) some cells lysed during the brief time between cell isolation and adding lysis buffer, leading to RNA degradation by cellular or external RNases, and (ii) some cells fell off the PCR cap prior to adding lysis buffer. While careful and rapid sample handling can help minimize both of these events, some cell drop-out during sample handling is likely unavoidable because of the miniaturization of single-cell RNA-seq.

Sequencing data normalization

Genes with a low transcript abundance (<100 UMIs across all samples) were excluded from analysis; 12,902 genes passed this abundance criterion. In addition, 375 genes that appeared to be expressed at specific cell cycle phases (see below) were excluded. The resulting data (12,527 genes by 272 split-cell samples) were normalized by dividing by the total number of transcripts (UMIs) in each cell and multiplying by one million (transcripts per million normalization). Then the dataset was log transformed after adding a pseudocount of 11, corresponding to 0.5 UMIs per cell on average.

Calculation of pseudotime and pseudotime velocity

To calculate pseudotime, the top 2000 most variable genes were selected, the data were transformed by principal component analysis (PCA), and a principal curve was fit to the top 10 PCs using the R package princurve *(9)*.

Pseudotime velocity was calculated repeatedly after drawing a subset of samples from the complete dataset (bootstrapping), which served to both smooth out noise in the calculation and provided a way to assess uncertainty in the velocity estimate. For each of 1000 bootstrap rounds, 24 samples were randomly drawn from each of 7 anther size bins

(Fig. S8A). Binning the cells by anther size enforced even sampling across developmental time. Pseudotime was then calculated using the principal curves method. Each step of the pseudotime calculation was repeated for every bootstrap sample, including selecting the top 2000 most variable genes, PCA, and the principal curve fit. Then samples were ordered by increasing pseudotime, and pseudotime velocity was calculated as the slope of a linear least squares fit to pseudotime rank vs pseudotime with a rolling window size of 10 samples (see Fig. 2B). Ninety-five percent confidence intervals were calculated as the 0.025 and 0.975 quantiles of the bootstrap values. Estimates of pseudotime velocity were robust to the binning scheme for bootstrapping (Fig. S8B) and the window size selected for the linear slope calculations (Fig. S8C).

To link pseudotime velocity to estimated developmental time (e.g. bottom legend in Fig. 2C), the mean pseudotime rank of each sample was first calculated from the bootstrap estimates. Each sample was then attributed a developmental time so as to preserve (i) the rank order estimated from pseudotime and (ii) the total actual time (168 hours) represented by the single-cell time-course. Estimates of developmental time recovered the expected relationship with the age of each anther, as a linear fit of estimated developmental time (in hours) vs estimated anther age (in days, based on anther length) had a slope of 24.1 ± 0.6 (mean ± standard error; R 'lm' function).

Comparison between pseudotime velocity and clustering

Eight single-cell RNA-seq clustering methods were applied to identify cell stages using the log-transformed dataset after selecting the 2000 most-variable genes. SC3 *(26)*, RaceID3 *(27)*, SNN-cliq *(28)*, and SEURAT *(29)* were applied using the published R or python packages with standard parameters. SINCERA *(30)* was applied using Z-transformed data and the 'graph' clustering method. PCAreduce *(31)* was applied using the largest probability method (method = 'M'). Hierarchical clustering was performed using base R functions and the 'ward.D' clustering method. PCA + k means was performed using the R function 'kmeans' applied to the first two principal components of the PCA-transformed data.

SEURAT, SINCERA, and SNN-cliq provided a single clustering solution. The other 5 methods required the selection of a predefined number of clusters. These methods were applied with a range of cluster numbers from 2 to 5, and the largest number of clusters that correctly grouped at least 90% of the split-cell technical replicates was selected. The performance of all methods as a function of cluster number (k) can be seen in Fig. S7B (assessed based on the percentage of split-cell replicates grouped together). A consensus clustering solution was also considered by plotting a heatmap of how often every method grouped the same cells together (Fig. S7A), similar to the approach taken by SC3 *(26)*.

Pseudotime velocity was used to identify cell stages by selecting each peak in pseudotime velocity as a potential stage boundary. A range of stage numbers was considered (k = 2 to 5) by ranking each stage boundary based on the height of the corresponding pseudotime velocity peak; for example, the 2 stage solution (k = 2) for pseudotime velocity was provided by splitting the samples at the position of the largest velocity peak, the 3 cluster solution (k = 3) was selected by splitting the samples at the largest and second largest velocity peaks, etc. Pseudotime velocity performed better than all clustering methods at grouping split-cell technical replicates for every value of k (Fig. S7B).

Estimating the duration of the Pr1 and Pr2 transitions

In the single-cell RNA-seq data, we did not capture any cells that appeared to be intermediates during the Pr1 or Pr2 transitions; for instance, in Fig 2D both Pr1 and Pr2 split cells into distinct groups without any cells showing intermediate expression of Pr1 or Pr2 markers. To estimate an upper bound for the duration of Pr1 and Pr2, we asked how long these expression transitions could be without any intermediate cells being sampled for scRNA-seq. Given the synchrony of cells during meiosis, we considered each floret to be sampled independently but every cell from the same floret to be from the same expression stage; this will produce a more conservative upper bound than if we had considered cells rather than florets to be sampled independently.

Cells were collected from 11 florets with anthers between 1.15 and 1.6 mm in length; an anther range that represents ~72 hours of germinal development *(4)* and includes ~1/3 of anthers before Pr1, 1/3 between Pr1 and Pr2, and 1/3 after Pr2. Given we observed 0 florets out of 11 within the Pr1 or Pr2 transitions, an upper bound for the true fraction of florets within either Pr1 or Pr2 is 0.24 (95% confidence interval, binomial distribution). Thus, an upper bound for the combined time spent within the Pr1 and Pr2 transitions is 17.3 hours (0.24 x 72 total hours), or less than 10 hours per transition on average.

Estimating the fraction of genes that change in expression during germinal differentiation

Transcript counts for split-cell technical replicates were pooled (to provide a single value for each gene and each cell) and then the data were TPM-normalized and log-transformed as described in the 'calculation of pseudotime' section (above). Next, the mean $\log_2$ expression level of each gene was calculated for 'early archesporial cells' (cells with pseudotime ≤ 10, a total of the 20 cells), meiotic cells before the Pr1 transition (orange in Fig. 2C), and meiotic cells after the Pr2 transition (dark red in Fig. 2C). The $\log_2$ fold change in expression during Pr1 and Pr2 ($FC_{Pr1+Pr2}$) was calculated as the difference between the mean $\log_2$ expression in meiotic cells before Pr1 compared to meiotic cells after Pr2, and the $\log_2$ fold change in expression from early archesporial cells to pollen mother cells ($FC_{AR>PMC}$) was calculated similarly.

To quantify the fraction of genes that change >2 fold during the Pr1 and Pr2, the $\log_2$ fold-change values for each gene, $\overline{\log_2(FC_g)}$, were modeled as a response variable with homoscedastic, normally distributed measurement errors:

$$\overline{\log_2(FC_g)} = \log_2(FC_g) + N\{0, \sigma\}$$

Under this noise model, the population estimate for the fraction of genes with a $\log_2$(FC) exceeding any given threshold becomes a deconvolution problem. Deconvolved estimates of the fraction of genes with a $\log_2$(FC) greater than 1 or less than -1 were calculated using the 'cdf.decon' function of the R package 'spsurvey' *(32)*. This calculation was repeated for 1000 bootstrap samples using the R package 'boot' *(33)*, and bias-adjusted estimates ±95% confidence intervals were reported.

Identifying differentially expressed genes

For each gene, a cubic smoothing spline with 5 degrees of freedom was fit to the curve of pseudotime vs expression level. An example of these smooth splines can be seen in Fig. 2D (black lines on the scatter plots to the right). Then the proportion of variance explained by the spline was calculated as:

$$1-\frac{\sigma^2_{res,g}}{\sigma^2_{tot,g}}$$

where $\sigma^2_{res,g}$ is the variance of the residual expression levels for gene *g* after subtracting the smooth spline, and $\sigma^2_{tot,g}$ is the variance in the original data for gene *g*. The variance explained by the spline was then used as a test statistic for each gene, and a *p*-value was calculated from this statistic using a permutation test. *p*-values were adjusted for multiple hypothesis testing with Holm's method.

Gene clustering, pathway enrichment analysis, and gene family identification

All differentially expressed genes were grouped by kmeans clustering. A range of cluster numbers (from 2 to 15) was considered and k = 6 was selected because it was the largest k that returned clearly distinguishable clusters. GO term enrichment was calculated for each cluster with the AgriGO *(14)* web tool, using the *Zea mays* landing page and default parameters. Meiotic 'recombination and synapsis genes' (Table S4) were genes from Table 1 of Ref. 2 acting in pre-meiosis through zygotene (labeled as "A", "B", or "C" in the "Link to Figures 1 and 2" column). Maize orthologs of genes with established meiotic function in other plant species were included if there was a one-to-one relationship between orthologs in each species.

Identifying cell cycle-regulated genes and cell cycle phases

To find cell cycle regulated genes, we first took advantage of our split-cell technical replicates and identified transcripts with significant cell-to-cell variability after controlling for developmental progress. A permutation-test ANOVA was performed to determine which transcripts had significantly greater variation between cells isolated from the same floret than can be explained by technical variability. Of the 182 genes with significant cell-to-cell variability, the majority (74%) formed clusters readily attributable to phase in the cell cycle. These initial cell cycle clusters were expanded to include other genes with a correlated expression pattern (Pearson's correlation ≥ 0.4) in cells isolated from ≤ 1 mm anthers (the transit amplifying stage). A total of 375 cell cycle-regulated genes was identified (Table S2).

By grouping the cell cycle-regulated genes using hierarchical clustering (Fig. S6), each gene was assigned to a specific cell cycle phase. For instance, one cluster of genes contained 35 histone subunits (43% of the 81 genes in the cluster). Histone transcripts are upregulated during S-phase in most eukaryotes, and so this cluster was assigned to S-phase. The inferred order of expression of each cell cycle gene cluster based on principal component analysis (Fig. 5A) or hierarchical clustering (Fig. S6) reproduced the known order of the associated cell cycle phases during mitosis. This provides the first genome-wide, experimentally determined list of cell cycle-regulated genes in plants without the use of chemical inhibitors for synchronization. Each cell was assigned to a cell cycle

phase by hierarchical clustering – i.e. clustering all cells based on the expression of the 375 cell cycle-regulated genes (Fig. S6).

A gene cluster could be associated with every phase of the mitotic cell cycle except G1. Putative G1 genes were defined as those genes that were both (i) negatively correlated with the cumulative expression of all 375 cell cycle genes (Pearson's correlation ≤ -0.3) in cells isolated from ≤ 1 mm anthers and (ii) not strongly correlated with genes from any individual cell cycle phase (Pearson's correlation ≤ 0.1 for the most similar cell cycle cluster). There was far less confidence in the G1 gene assignments compared to the 375 genes associated with every other cell cycle phase. These 'G1 genes' were included in Fig. 5A and 5B for visualization purposes and were reported in Table S2, but were not considered as 'cell cycle-regulated' genes for other analyses.

For Fig. 5D and Fig. 5E, cyclins were defined as genes that encode proteins with a cyclin N-terminal domain (Interpro ID IPR006671).

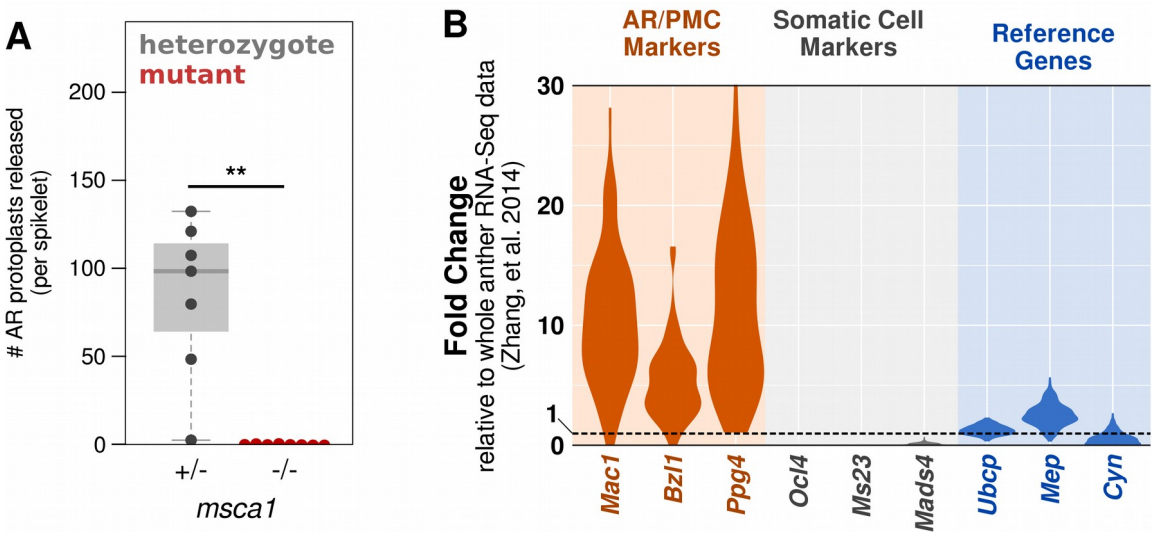

Fig. S1. Validation of archesporial/pollen mother cell isolation protocol.

Germinal cells were identified by their large size, as they are 2-4 times larger in diameter than all other cells of the anther *(4)*. (**A**) Counts of large protoplasts (20-25 µm in diameter) released from *msca1* mutant plants, which lack archesporial cells, or heterozygous siblings. A single floret containing three anthers between 0.4 and 0.5 mm was dissected from each plant and protoplasted, then the number of resulting large (putative archesporial) cells was counted. The *msca1* mutant plants lacked any large protoplasts. **, $p < 0.05$; Mann-Whitney U test. (**B**) Expression of germinal and somatic marker genes from archesporial and pollen mother cells isolated for single-cell RNA-seq. For each cell and each gene, the fold change of expression was calculated by dividing the expression of each marker gene (transcripts per million (TPM)) by the expression of that gene in published *(34)* whole anther RNA-seq data (also normalized by TPM). Violin plots show the distribution of calculated fold changes for each gene across all 144 cells in the single-cell dataset. AR, archesporial cell; PMC, pollen mother cell.

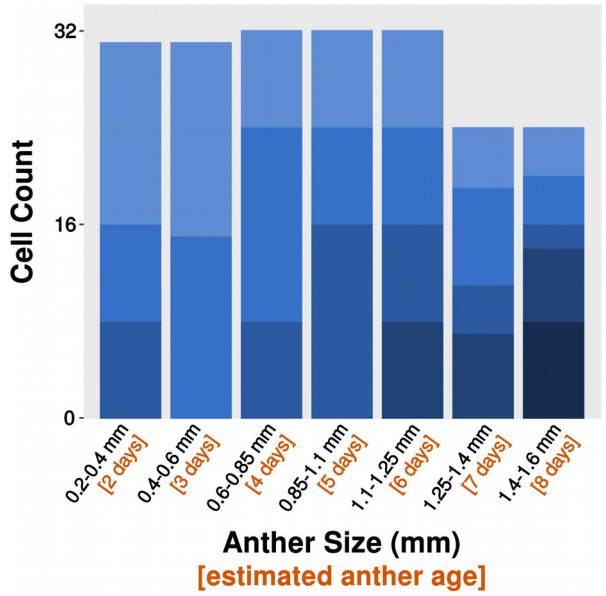

Fig. S2. Histogram of the number of cells collected at each anther size.

Anther sizes were binned by approximate anther age. Each shade of blue denotes cells from a different plant collected at a particular anther size class; all cells collected from a plant were from a single floret.

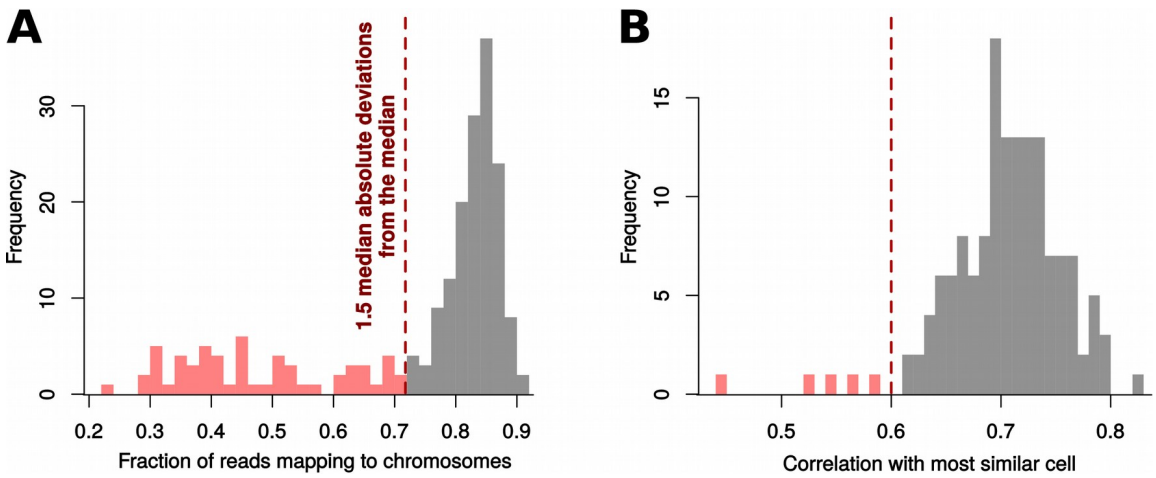

Fig. S3. Quality control criteria.

Before applying quality control (QC) criteria, data from split-cell technical replicates were pooled together. This way the cell selection was blind to technical reproducibility; therefore, downstream analyses comparing technical replicates were not biased by the QC procedure. (**A**) Histogram of the fraction of mapped reads that mapped to the 10 maize chromosomes (rather than mapping to ERCC spike-in controls or organellar chromosomes). Fifty-seven cells were excluded (red histogram bars) because they had fewer than 71.8% of reads mapping to the chromosomes (71.8% is 1.5 median absolute deviations below the median). (**B**) Histogram of the Pearson's correlation coefficient between the expression profile of every cell and the most similar other cell. Four cells were excluded because their expression profile had a low correlation with all other cells in the dataset.

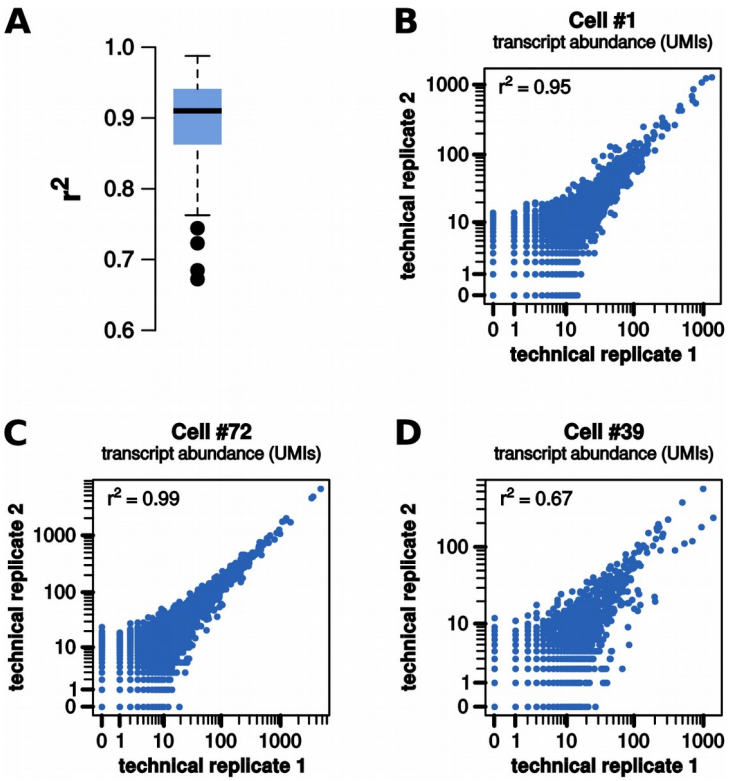

Fig. S4. Reproducibility of technical replicates.

(**A**) Boxplot showing the r^2 between technical replicates for all cells that passed the quality control criteria. (**B-D**) Scatter plot of transcript abundance for split-cell replicates. Cells were numbered by the order in which they were isolated. (**B**) A representative cell, arbitrarily chosen as the first cell isolated. (**C**) The cell with the highest r^2 between split-cell replicates. (**D**) The cell with the lowest r^2 between split-cell replicates.

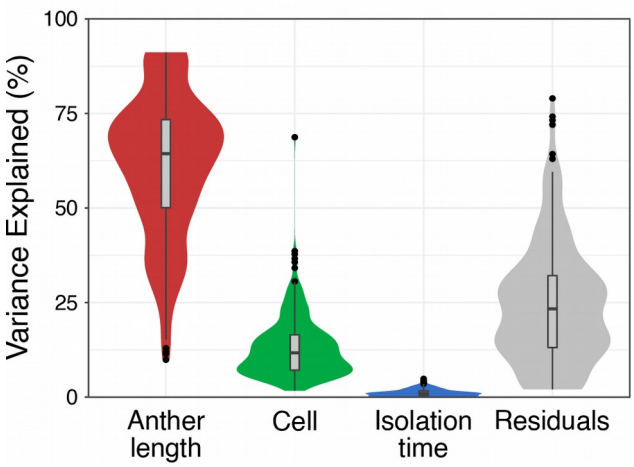

Fig. S5. Variance explained by tissue stage, cell, and isolation time.

A partitioning of variance was performed on the sample-by-sample correlation matrix using the R package ‘variancePartition’ *(35)* to ask how well the correlation with all other cells could be explained by tissue stage (anther length), cell, or isolation time (which ranged from 1 to 2 h). The violin plots show the distribution of this partitioning for all samples (a sample is a split-cell replicate). The majority (63.2%) of the correlation matrix is explained by tissue stage. Isolation time does not explain a significant portion of the sample-by-sample correlation matrix, suggesting the effects of protoplasting are small relative to pre-existing biological variability.

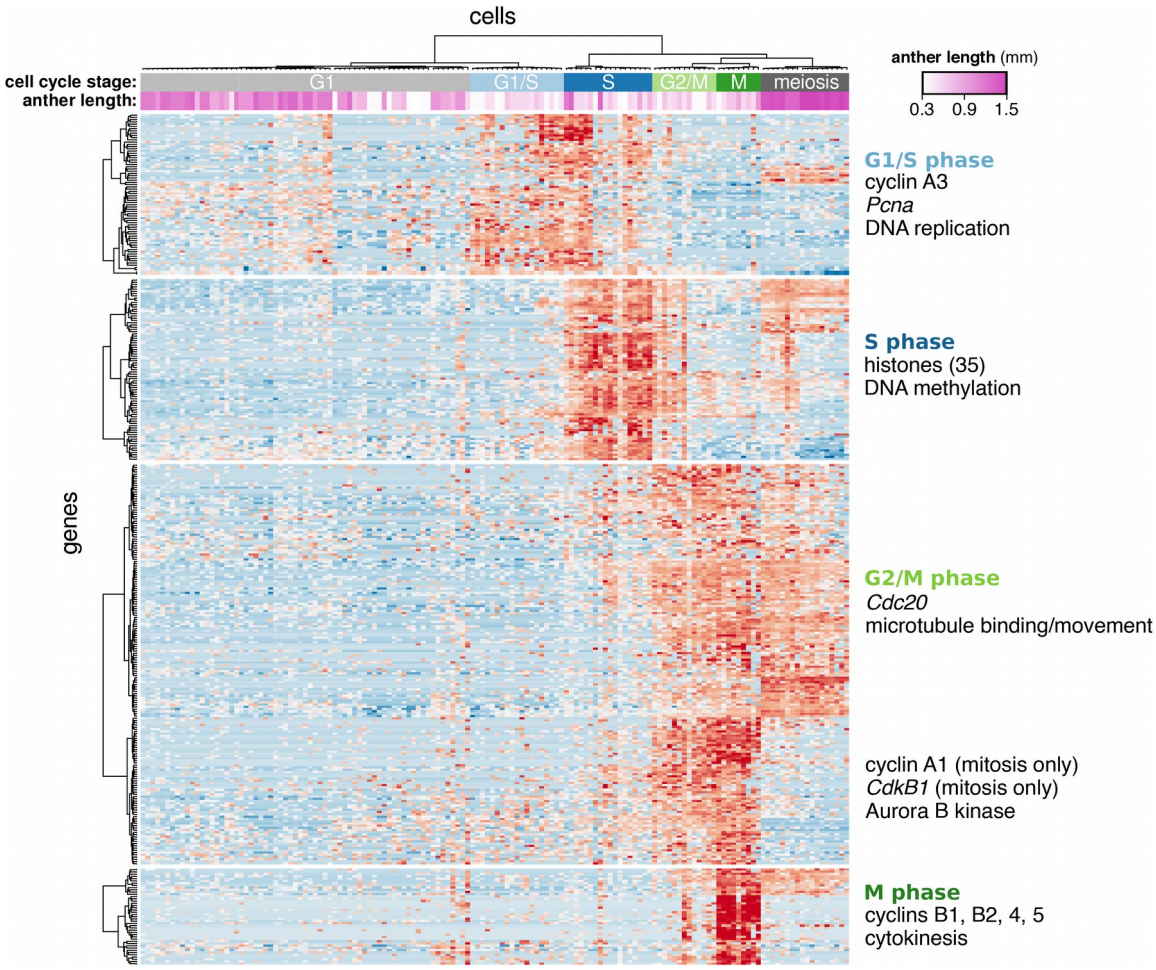

Fig. S6. Cell-cycle regulated gene clusters.

In archesporial stage anthers (≤ 1 mm), there were 182 genes that varied significantly in cells isolated from the same anther (ANOVA). The majority of these (135 genes; 74%) could be associated with cell cycle-specific expression based on the presence of known cell cycle-regulated genes such as cyclins and histones. Shown is a heatmap of these genes plus 240 others with a Pearson's correlation ≥0.4 with at least one of the initial 135 genes (135 + 240 = 375 total cell cycle genes). Clusters were assigned to cell cycle phases based on the presence of genes or enriched Gene Ontology terms known to be expressed during that phase (listed on the right). The inferred order of these expression clusters by principal component analysis (Fig. 5A) or the heatmap matches the order of the assigned cell cycle phases. Colors denote expression levels scaled from the minimum log TPM to maximum log TPM for each gene. Meiotic cells (cells after the Pr1 transition) express a subset of genes associated with multiple phases of the mitotic cell cycle. TPM, transcripts per million.

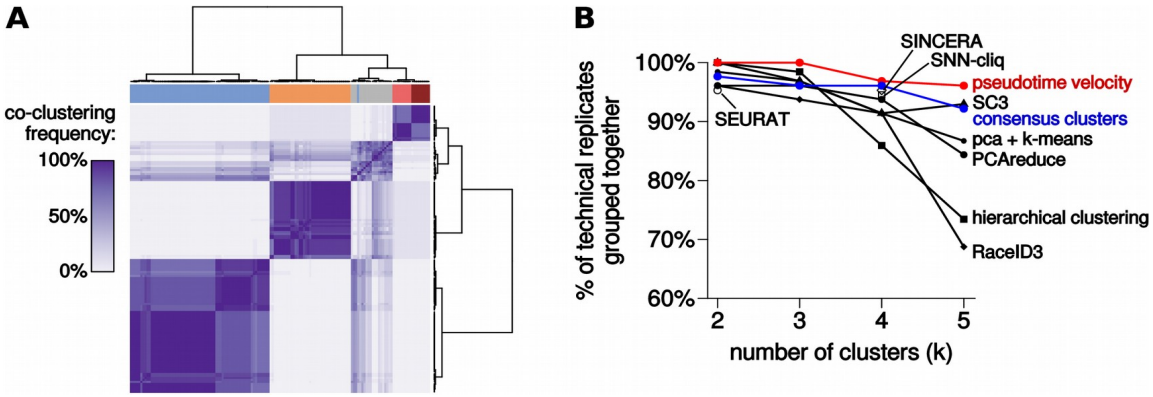

Fig. S7. Comparison of stages determined by pseudotime velocity with single-cell clustering methods.

Eight single-cell clustering methods were tested using default parameters, as described in the Supplementary Materials. (**A**) Heatmap displaying how often every cell was grouped with every other cell by 8 single-cell clustering methods. A co-clustering frequency of 100% (purple) indicates that both technical replicates for two cells were grouped together by all 8 clustering methods. The top bar indicates the stages defined for each cell by pseudotime velocity (colored identically to Fig. 2). Pseudotime velocity stages agree closely with the consensus clusters. (**B**) Evaluation of cell stages defined by pseudotime velocity and the 8 clustering methods. Every method was applied to find 2-5 clusters (except for 3 methods that returned a single clustering solution: SEURAT, SINCERA, and SNN-cliq). Then the quality of the resulting clusters was assessed based on the percentage of split-cell technical replicates that were correctly placed within the same group. Pseudotime velocity grouped the split-cell replicates more accurately than all clustering methods.

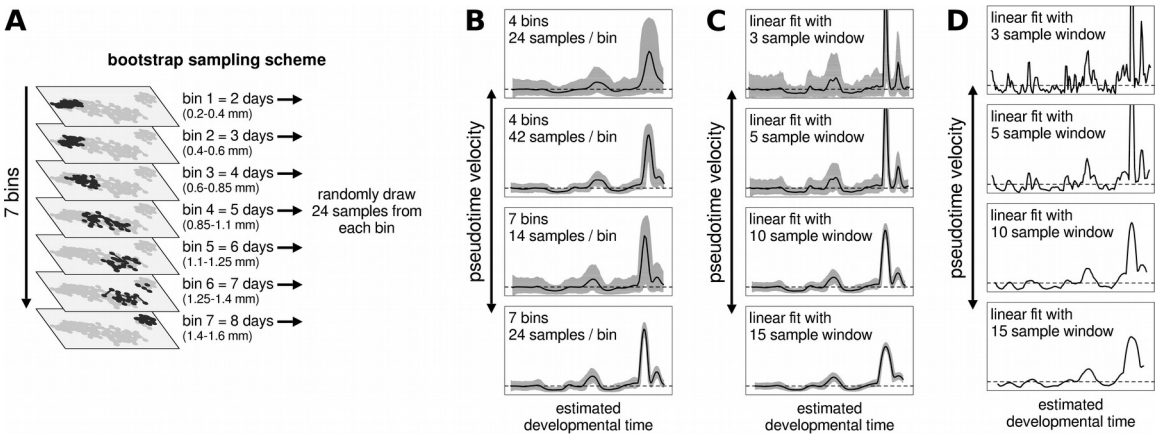

Fig. S8. Evaluation of bootstrap sampling and parameter choice for pseudotime velocity calculations.

(**A**) Illustration of bootstrap sampling scheme with 7 anther bins and 24 samples drawn from each bin. This binning scheme was used for all calculations except for panel B of this figure. (**B**) Sensitivity of pseudotime velocity to the number of sampling bins and number of samples drawn per bin. Samples were divided into 4 or 7 bins by anther length, a fixed number of samples were drawn randomly from each bin, and pseudotime velocity was calculated from each bootstrap sample. The shape of the pseudotime velocity curve is robust to the binning scheme used. Gray outlines, 95% confidence interval. (**C**) Sensitivity of pseudotime velocity to the linear fit window size. Pseudotime velocity was calculated as the linear slope in pseudotime with a rolling window of 3, 5, 10, or 15 samples (see Fig. 2A). Larger window sizes reduced variability but smoothed out finer details in the velocity curve. (**D**) Equivalent to panel C except with no bootstrapping. Bootstrapping dampened noise and allowed for estimates of uncertainty in the pseudotime velocity calculations.

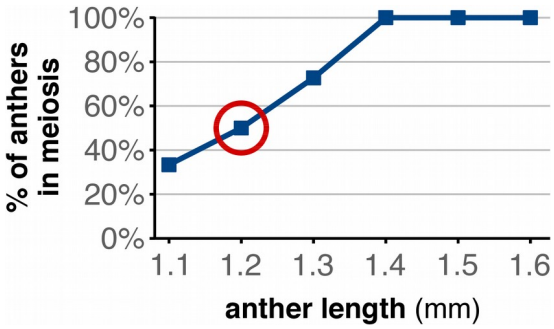

Fig. S9. Staging meiotic start by anther length.

The % of anthers in meiosis by anther length. Half of 1.2 mm anthers were in meiosis (red circle), indicating that this is the average anther length at the start of meiosis.

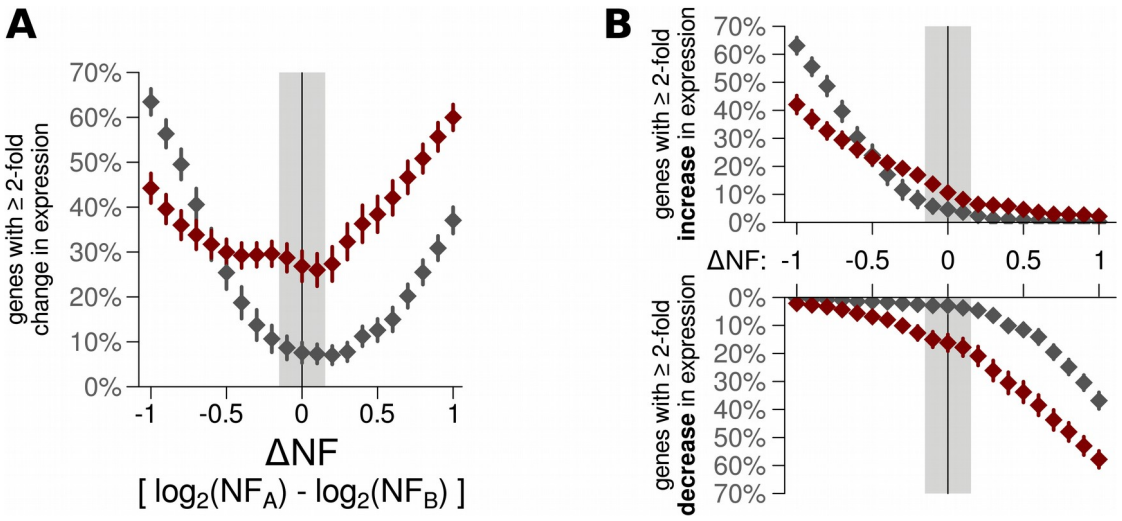

Fig. S10. Effect of normalization on the estimated percentage of genes that change ≥2 fold during prophase 1 transitions.

The fraction of genes with a ≥2 fold absolute change in expression (**A**) or increase / decrease in expression (**B**) were estimated with different normalization factors. ΔNF is the difference in $\log_2$ normalization factors for population A and population B, centered so that transcripts per million (TPM) normalization results in ΔNF of 0. Several different methods for data normalization were tested, including TPM normalization, mean fold change normalization, upper quartile normalization, trimmed mean normalization, and no normalization. Every normalization method tested produces a ΔNF that falls within the light gray boxes. Red diamonds are the estimates for the combined change in expression during the Pr1 and Pr2 transitions, and dark gray diamonds are estimates for the change in expression between the early archesporial cells and pollen mother cells prior to Pr1. Estimates for the fraction of genes up- or down-regulated were somewhat sensitive to normalization, but the fraction of genes with ≥2 fold absolute change in expression were not.

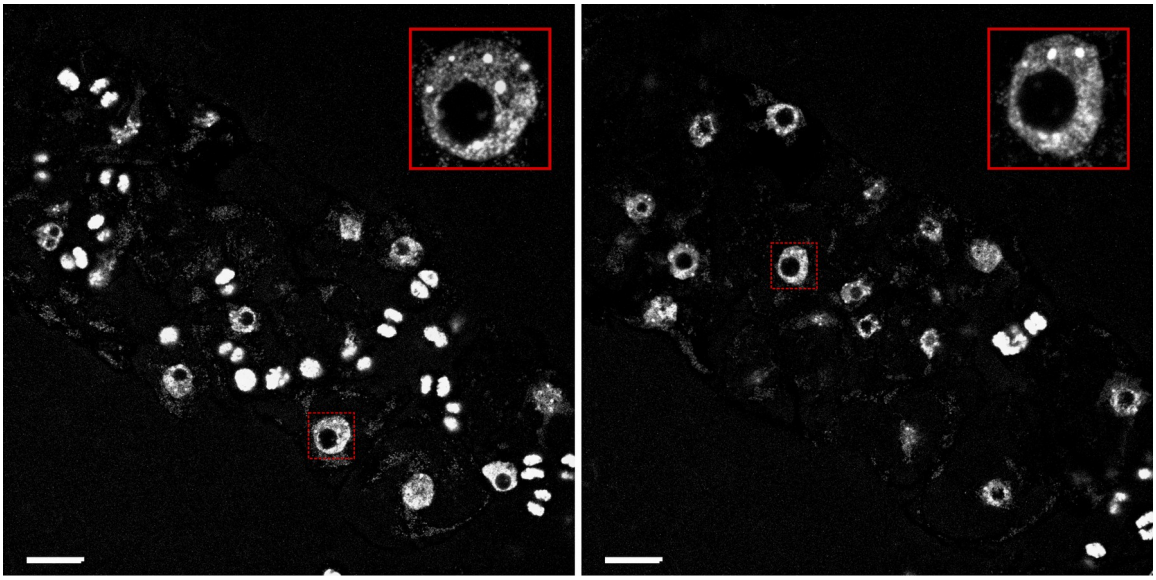

Fig. S11. Cytology of *am1-489* in 2.5 mm anthers.

Two focal planes of an extruded germinal column from a 2.5 mm *am1-489* anther. The majority of cells are in interphase with diffuse chromatin, a central nucleolus, and rounded knobs. Cells were never observed with meiotic-like cytology. Insets on the top right of each image are higher resolution views of a representative cell (marked with a dotted red outline in the main image). Scale bar, 40 μm.

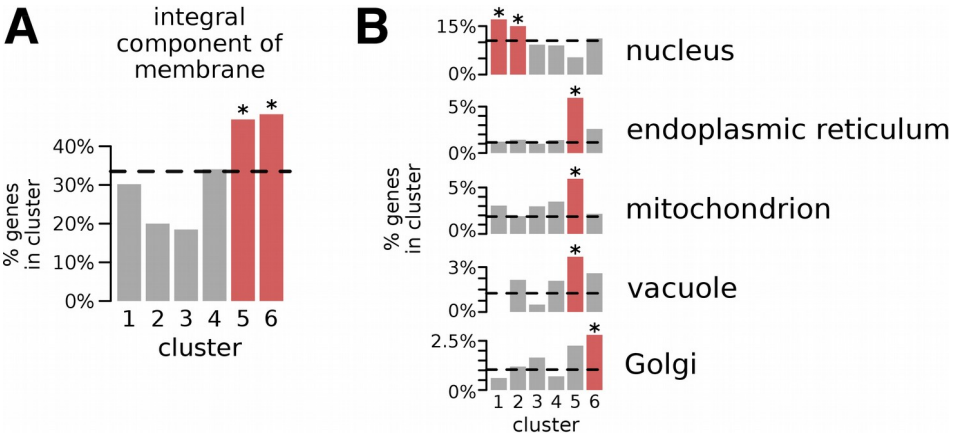

Fig. S12. Dynamic regulation of genes linked to sub-cellular organelles during early germinal development.

Enrichment for (**A**) genes that encode for integral membrane proteins or (**B**) genes linked to membrane-bound organelles in each cluster of differentially expressed genes. Cluster numbers refer to the gene clusters in Fig. 4A and Table S3. Red bars with an asterisk denote significantly enriched GO terms ($p < 0.05$; Fisher's exact test).

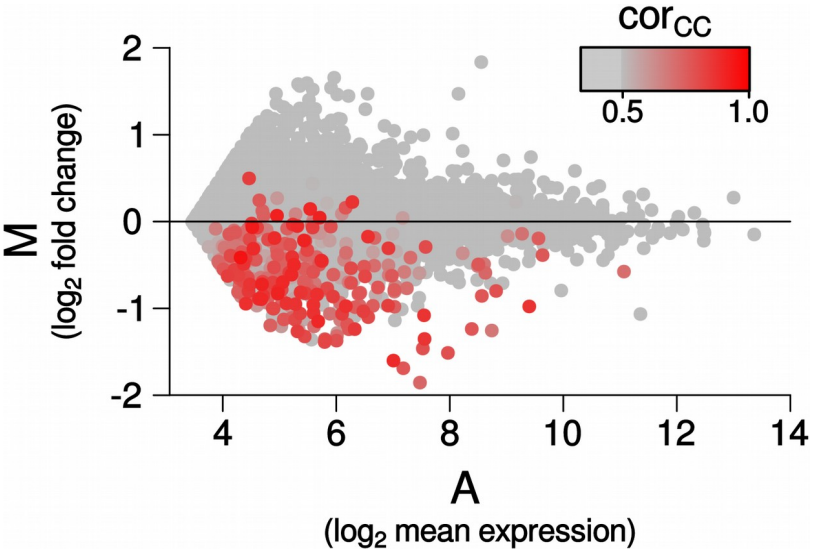

Fig. S13. MA plot of simulated bulk data depicting gene expression changes between cells from 0.5 mm and 1.0 mm anthers.

Data for cells from 0.4 – 0.6 mm anthers or from 0.9 – 1.1 mm anthers were pooled together to simulate a bulk dataset; cell cycle-regulated genes were not removed. The color indicates the correlation of each gene's expression level with the most similar cell cycle cluster (excluding the G1-phase cluster). Cell cycle-regulated genes were strongly down-regulated between 0.5 mm and 1.0 mm anthers because of the decreasing proportion of cells in different phases of the cell cycle (e.g. Fig. 5C). From bulk data, it would not be possible to distinguish genes down-regulated during archesporial development from genes expressed at a specific (non-G1) cell cycle phase.

Table S1. Sample information for isolated maize germinal cells. (separate file)

Table S2. Cell cycle-regulated genes. (separate file)

Table S3. Differentially expressed genes during early germinal development. (separate file)

Table includes gene cluster assignments and columns specifying which genes are transcription factors or 'meiotic candidate genes'. Meiotic candidate genes were defined as any genes in clusters 5 or 6 that were also identified in *(5)* as spatially enriched in meiotic cells.

Table S4. Established meiotic genes. (separate file)

Genes required for early meiotic progression in plants, based off of *(2)*.

Table S5. Enriched Gene Ontology terms. (separate file)

Gene Ontology term enrichment from AgriGO *(14)*. Excel sheets are results for each cluster in Fig. 4A. FDR, false discovery rate.

Table S6. Primer sequences for scRNA-seq library contruction.

Cel-Seq barcodes are underlined.

Name	Sequence
dT_1s	GCCGGTAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATC$(N)_{10}$<u>AGACCTC</u>$(T)_{24}$V
dT_2s	GCCGGTAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATC$(N)_{10}$<u>AGCTAG</u>$(T)_{24}$V
dT_4s	GCCGGTAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATC$(N)_{10}$<u>AGCTTC</u>$(T)_{24}$V
dT_5s	GCCGGTAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATC$(N)_{10}$<u>CATGAG</u>$(T)_{24}$V
dT_6s	GCCGGTAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATC$(N)_{10}$<u>CATGCA</u>$(T)_{24}$V
dT_9s	GCCGGTAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATC$(N)_{10}$<u>CAGATC</u>$(T)_{24}$V
dT_10s	GCCGGTAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATC$(N)_{10}$<u>TCACAG</u>$(T)_{24}$V
dT_19s	GCCGGTAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATC$(N)_{10}$<u>TCGATC</u>$(T)_{24}$V
dT_23s	GCCGGTAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATC$(N)_{10}$<u>GTCTAG</u>$(T)_{24}$V
dT_25s	GCCGGTAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATC$(N)_{10}$<u>GTTGCA</u>$(T)_{24}$V
dT_26s	GCCGGTAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATC$(N)_{10}$<u>GTGACA</u>$(T)_{24}$V
dT_31s	GCCGGTAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATC$(N)_{10}$<u>ACTCGA</u>$(T)_{24}$V
dT_37s	GCCGGTAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATC$(N)_{10}$<u>CTAGGA</u>$(T)_{24}$V
dT_46s	GCCGGTAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATC$(N)_{10}$<u>TGCAGA</u>$(T)_{24}$V
dT_101X	GCCGGTAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATC$(N)_{10}$<u>TCACGC</u>$(T)_{24}$V
dT_102X	GCCGGTAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATC$(N)_{10}$<u>GAGTCG</u>$(T)_{24}$V

Table S7. Primer sequences for quantitative PCR.

6-FAM, Fluorescein; ZEN, Zen fluorescent quencher; IBFQ, Iowa black fluorescent quencher; HEX, Hexachlorofluorescein. Gene IDs are Zm00001d039101 (*C3h3*), Zm00001d006555 (*Rpl38e*), Zm00001d039352 (*Rmf*), Zm00001d050069 (*Trps8*), Zm00001d006351 (*Ago18a*), Zm00001d048163 (*pco080082*).

Gene	Fwd	Rev	TaqMan Probe
C3h3	TTGAAGATGTCAGTCCACCAC	GTTCTGAAGACTGACGATGGTATAA	6-FAM/TTCCACCCT/ZEN/CTCCAGCTCTACTTCA/IBFQ
Rpl38e	CTGACTGTCCAGGAGATTTAAGG	GGAGCCCACCTACAAACTAAA	6-FAM/TCAGGGACG/ZEN/TTTGCTCCAGTGTT/IBFQ
Rmf	TCTGAGCTGTTCCTGGAGAA	ACCTCCTCGTCCTGAAAGT	6-FAM/TCTCGTACT/ZEN/CCAAGGTCTCGTCCC/IBFQ
Trps8	CACAAGGTGGAGCAGGTATT	TCAGTTACAGCCACGTAAAGG	6-FAM/AGTCACCCG/ZEN/ACCTGAAGGTAGACT/IBFQ
Ago18a	CGCTGCAGTCTCTCACATATAA	GCTAACTTGTGGGCGTAGTAT	6-FAM/AAGCTGCAC/ZEN/TCGCTCTGTGTCAAT/IBFQ
pco080082	CCCGCAAGCGAAATGATAAAG	CAAATGGGCACGGTAGTGTA	HEX/TGCGTTGCC/ZEN/TGGTATAGCAAAGCTA/IBFQ

References and Notes

1. T. Kelliher, R. L. Egger, H. Zhang, V. Walbot, Unresolved issues in pre-meiotic anther development. *Front. Plant Sci.* **5**, 347 (2014). doi:10.3389/fpls.2014.00347 Medline

2. R. Mercier, C. Mézard, E. Jenczewski, N. Macaisne, M. Grelon, The molecular biology of meiosis in plants. *Annu. Rev. Plant Biol.* **66**, 297–327 (2015). doi:10.1146/annurev-arplant-050213-035923 Medline

3. W. Crismani, C. Girard, R. Mercier, Tinkering with meiosis. *J. Exp. Bot.* **64**, 55–65 (2013). doi:10.1093/jxb/ers314 Medline

4. T. Kelliher, V. Walbot, Emergence and patterning of the five cell types of the *Zea mays* anther locule. *Dev. Biol.* **350**, 32–49 (2011). doi:10.1016/j.ydbio.2010.11.005 Medline

5. S. Dukowic-Schulze, A. Sundararajan, J. Mudge, T. Ramaraj, A. D. Farmer, M. Wang, Q. Sun, J. Pillardy, S. Kianian, E. F. Retzel, W. P. Pawlowski, C. Chen, The transcriptome landscape of early maize meiosis. *BMC Plant Biol.* **14**, 118 (2014). doi:10.1186/1471-2229-14-118 Medline

6. H. Zhang, R. L. Egger, T. Kelliher, D. Morrow, J. Fernandes, G.-L. Nan, V. Walbot, Transcriptomes and proteomes define gene expression progression in pre-meiotic maize anthers. *G3 (Bethesda)* **4**, 993–1010 (2014). doi:10.1534/g3.113.009738 Medline

7. T. L. Yuan, W. J. Huang, J. He, D. Zhang, W. H. Tang, Stage-specific gene profiling of germinal cells helps delineate the mitosis/meiosis transition. *Plant Physiol.* **176**, 1610–1626 (2018). doi:10.1104/pp.17.01483 Medline

8. S. Dukowic-Schulze, A. Harris, J. Li, A. Sundararajan, J. Mudge, E. F. Retzel, W. P. Pawlowski, C. Chen, Comparative transcriptomics of early meiosis in *Arabidopsis* and maize. *J. Genet. Genomics* **41**, 139–152 (2014). doi:10.1016/j.jgg.2013.11.007 Medline

9. T. Hastie, W. Stuetzle, Principal curves. *J. Am. Stat. Assoc.* **84**, 502–516 (1989). doi:10.1080/01621459.1989.10478797

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

11. R. K. Dawe, J. W. Sedat, D. A. Agard, W. Z. Cande, Meiotic chromosome pairing in maize is associated with a novel chromatin organization. *Cell* **76**, 901–912 (1994). doi:10.1016/0092-8674(94)90364-6 Medline

12. C.-J. R. Wang, G.-L. Nan, T. Kelliher, L. Timofejeva, V. Vernoud, I. N. Golubovskaya, L. Harper, R. Egger, V. Walbot, W. Z. Cande, Maize *multiple archesporial cells 1* (*mac1*), an ortholog of rice TDL1A, modulates cell proliferation and identity in early anther development. *Development* **139**, 2594–2603 (2012). doi:10.1242/dev.077891 Medline

13. W. P. Pawlowski, C.-J. R. Wang, I. N. Golubovskaya, J. M. Szymaniak, L. Shi, O. Hamant, T. Zhu, L. Harper, W. F. Sheridan, W. Z. Cande, Maize AMEIOTIC1 is essential for multiple early meiotic processes and likely required for the initiation of meiosis. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 3603–3608 (2009). doi:10.1073/pnas.0810115106 Medline

14. T. Tian, Y. Liu, H. Yan, Q. You, X. Yi, Z. Du, W. Xu, Z. Su, agriGO v2.0: A GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Res.* **45** (W1), W122–W129 (2017). doi:10.1093/nar/gkx382 Medline

15. S. Lee, H. Warmke, Organelle size and number in fertile and T-cytoplasmic male-sterile corn. *Am. J. Bot.* **66**, 141–148 (1979). doi:10.1002/j.1537-2197.1979.tb06206.x

16. H. G. Dickinson, J. Heslop-Harrison, Ribosomes, membranes and organelles during meiosis in angiosperms. *Philos. Trans. R. Soc. London B Biol. Sci.* **277**, 327–342 (1977). doi:10.1098/rstb.1977.0021 Medline

17. Y. Azumi, D. Liu, D. Zhao, W. Li, G. Wang, Y. Hu, H. Ma, Homolog interaction during meiotic prophase I in *Arabidopsis* requires the SOLO DANCERS gene encoding a novel cyclin-like protein. *EMBO J.* **21**, 3081–3095 (2002). doi:10.1093/emboj/cdf285 Medline

18. C. D. Green, Q. Ma, G. L. Manske, A. N. Shami, X. Zheng, S. Marini, L. Moritz, C. Sultan, S. J. Gurczynski, B. B. Moore, M. D. Tallquist, J. Z. Li, S. S. Hammoud, A comprehensive roadmap of murine spermatogenesis defined by single-cell RNA-seq. *Dev. Cell* **46**, 651–667.e10 (2018). doi:10.1016/j.devcel.2018.07.025 Medline

19. J. D. Richter, P. Lasko, Translational control in oocyte development. *Cold Spring Harb. Perspect. Biol.* **3**, a002758 (2011). doi:10.1101/cshperspect.a002758 Medline

20. A. P. Caryl, G. H. Jones, F. C. Franklin, Dissecting plant meiosis using *Arabidopsis thaliana* mutants. *J. Exp. Bot.* **54**, 25–38 (2003). doi:10.1093/jxb/erg041 Medline

21. T. Hashimshony, N. Senderovich, G. Avital, A. Klochendler, Y. de Leeuw, L. Anavy, D. Gennert, S. Li, K. J. Livak, O. Rozenblatt-Rosen, Y. Dor, A. Regev, I. Yanai, CEL-Seq2: Sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17**, 77 (2016). doi:10.1186/s13059-016-0938-8 Medline

22. C. A. Schneider, W. S. Rasband, K. W. Eliceiri, NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012). doi:10.1038/nmeth.2089 Medline

23. J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J.-Y. Tinevez, D. J. White, V. Hartenstein, K. Eliceiri, P. Tomancak, A. Cardona, Fiji: An open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012). doi:10.1038/nmeth.2019 Medline

24. M. Pertea, D. Kim, G. M. Pertea, J. T. Leek, S. L. Salzberg, Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* **11**, 1650–1667 (2016). doi:10.1038/nprot.2016.095 Medline

25. T. Kivioja, A. Vähärautio, K. Karlsson, M. Bonke, M. Enge, S. Linnarsson, J. Taipale, Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* **9**, 72–74 (2011). doi:10.1038/nmeth.1778 Medline

26. V. Y. Kiselev, K. Kirschner, M. T. Schaub, T. Andrews, A. Yiu, T. Chandra, K. N. Natarajan, W. Reik, M. Barahona, A. R. Green, M. Hemberg, SC3: Consensus clustering of single-cell RNA-seq data. *Nat. Methods* **14**, 483–486 (2017). doi:10.1038/nmeth.4236 Medline

27. D. Grün, A. Lyubimova, L. Kester, K. Wiebrands, O. Basak, N. Sasaki, H. Clevers, A. van Oudenaarden, Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* **525**, 251–255 (2015). doi:10.1038/nature14966 Medline

28. C. Xu, Z. Su, Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* **31**, 1974–1980 (2015). doi:10.1093/bioinformatics/btv088 Medline

29. R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, A. Regev, Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015). doi:10.1038/nbt.3192 Medline

30. M. Guo, H. Wang, S. S. Potter, J. A. Whitsett, Y. Xu, SINCERA: A pipeline for single-cell RNA-seq profiling analysis. *PLOS Comput. Biol.* **11**, e1004575 (2015). doi:10.1371/journal.pcbi.1004575 Medline

31. J. Žurauskienė, C. Yau, pcaReduce: Hierarchical clustering of single cell transcriptional profiles. *BMC Bioinformatics* **17**, 140 (2016). doi:10.1186/s12859-016-0984-y Medline

32. A. Olsen, T. Kincaid, spsurvey: Spatial survey design and analysis. R package version 3.4 (2017).

33. B. Ripley, A. Canty, boot: Bootstrap R (S-Plus) functions. R package version 1.3-20 (2017).

34. J. Zhai, H. Zhang, S. Arikit, K. Huang, G.-L. Nan, V. Walbot, B. C. Meyers, Spatiotemporally dynamic, cell-type-dependent premeiotic and meiotic phasiRNAs in maize anthers. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 3146–3151 (2015). doi:10.1073/pnas.1418918112 Medline

35. G. E. Hoffman, E. E. Schadt, variancePartition: Interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics* **17**, 483 (2016). doi:10.1186/s12859-016-1323-z Medline