Supplementary Material

Transcriptional and imprinting complexity in Arabidopsis seeds at single-nucleus resolution

Colette L. Picard, Rebecca A. Povilus, Ben P. Williams, Mary Gehring

Other Supplementary Material for this manuscript includes the following:

Supplementary Text

Defining micropylar and peripheral endosperm nuclei cluster identity

After fertilization, Arabidopsis endosperm undergoes three rounds of rapid synchronous nuclear replication without cytokinesis, forming a syncytium of nuclei surrounded by cytoplasm (*1,2*). Then, three morphologically distinct domains are formed: the micropylar, peripheral, and chalazal endosperm regions (Fig. 1e, *3*). Nuclei divide synchronously within a domain but asynchronously with other domains. At one pole of the seed, the dense micropylar endosperm surrounds the embryo, while at the other pole the chalazal endosperm lies adjacent to the termination of maternal vascular tissue. The peripheral endosperm lies between the poles and consists of a large central vacuole with nuclei arranged along the outer periphery, connected by cytoplasmic bridges (*2*). Several days post-fertilization, endosperm cellularization proceeds in a wave from the micropylar to chalazal pole. The chalazal endosperm does not fully cellularize.

We first evaluated the expression of previously defined marker genes (Supplementary Data 2) for micropylar, peripheral, and chalazal endosperm tissue obtained from published microarray analysis of laser capture microdissection (LCM) of wild-type Ws-0 seeds at a similar stage of development (*4*, *5*). Markers for each endosperm domain were enriched in several clusters, while some clusters lacked enrichment for markers of any domain (Fig. 1c). To further refine cluster identity, we identified differentially expressed genes by performing all possible comparisons among nuclei clusters (Fig. S4, Supplementary Data 2). GO term analysis indicated that several probable peripheral and micropylar endosperm clusters had increased expression of genes characteristic of S- and M-phase (Fig. 1d, Extended Data Fig. 5, Supplementary Data 2), suggesting that some of the variability between endosperm nuclei could be attributed to differences in cell cycle stage. We therefore performed a trajectory analysis to map each nucleus onto a linear path representing progression through the cell cycle (Extended Data Fig. 7, Supplementary Data 3) (*6,7*). To determine the extent to which cell cycle

differences were driving endosperm nuclei clustering, we repeated the clustering while omitting 1,065 genes whose expression varied significantly along the cell cycle trajectory (Supplementary Data 3). Similar endosperm clusters were recovered, indicating that cell cycle stage is not sufficient to explain the observed clustering (Fig. S7).

Multiple endosperm nuclei clusters showed elevated expression of genes related to M-phase, such as mitotic chromosome condensation and cytokinesis (Fig. 1d, Extended Data Fig. 5). To examine where these nuclei were distributed in the endosperm, we performed RNA *in situ* hybridization for AT4G11080/*3XHMG-BOX1*, which was most frequently expressed in VxC E2 and CxV E4, E8 and E10 nuclei (Extended Data Fig. 6). In each of these clusters, the majority of nuclei were identified as being in M-phase (Extended Data Fig. 7). Hybridization signal was detected in a spotty pattern in the embryo, in the micropylar endosperm, and along the cellularization boundary of the peripheral endosperm (Extended Data Fig. 6), consistent with dividing nuclei. Among the clusters expressing AT4G11080, VxC E2 and CxV E4 had the strongest expression of other M-phase related genes (Fig. S4, Supplementary Data 2), had similar overall expression patterns (Fig. S4, S5), and were enriched in genes associated with peripheral endosperm (Fig. 1c). Based on these distinct lines of evidence, we concluded that both VxC E2 and CxV E4 clusters represent M-phase peripheral endosperm nuclei (Fig. 1e). CxV E8, consisting of only 6 nuclei, appears to be a distinct class of possibly mid M-phase peripheral endosperm nuclei (Fig. 1e).

In contrast, CxV E10 and VxC E4 showed some upregulation of M-phase related genes (Fig. 1d) and AT4G11080/*3XHMG-BOX1*, but were transcriptionally distinct from the peripheral M-phase clusters (Fig. S4) and had some expression of LCM micropylar endosperm markers (Fig. 1c). Yet only one cluster, VxC E3, was strongly enriched for the expression of LCM micropylar endosperm markers (Fig. 1c, Fig. S5). However, VxC E3 shared some gene expression features with VxC E4 (Fig. 1b, Fig. S4) and CxV E9, E10, and E11, including

expression of AT1G09380/*UMAMIT25*, which encodes an amino acid exporter (*8*) (Extended Data Fig. 6). We found that *UMAMIT25* transcript was detected in the micropylar endosperm in both crosses and in 25% of Col x Cvi F_1 seeds it was also detected on the leading edge of the cellularization boundary in peripheral endosperm (Extended Data Fig. 6). RNA was also detected in the seed coat, as expected based on snRNA-seq data from the seed coat clusters (Extended Data Fig. 6). This suggests that both VxC E3 and E4 represent distinct populations of micropylar nuclei. VxC E4 and CxV E10 both express M-phase related genes (Fig. 1d) and have similar expression patterns (Fig. S4), suggesting that together CxV E10 and VxC E4 correspond to actively dividing micropylar nuclei. CxV E11 has upregulation of S-phase related genes (Fig. 1d) and may correspond to micropylar nuclei in S-phase. Unlike these clusters, cell cycle trajectory analysis indicated that VxC E3 contains nuclei only in G2 or G0, suggesting these are fully differentiated, non-dividing micropylar nuclei (Fig. 1e, Extended Data Fig. 7). In Col x Cvi seeds, *RGE1*/*ZOU*/AT1G49770, which is expressed in the embryo surrounding region of the micropylar endosperm and is required for embryo cuticle formation (*9*), is highly expressed in CxV E9, but not E10 or E11 (Supplementary Data 2). This suggests that CxV E9 represents the micropylar nuclei in the embryo surrounding region in CxV (Fig. 1e).

Several clusters corresponded to peripheral endosperm, which comprises the bulk of the nuclei in endosperm (Fig. 1d,e). While some were most strongly associated with a particular cell cycle phase, a distinctive subset were characterized by upregulation of the photosynthetic machinery and were classified as ‘energy-generating’ peripheral endosperm. These three clusters were distinguished from each other by GO terms specific to certain clusters, such as translation and fatty acid biosynthesis for energy-generating peripheral I (Fig. 1d, Extended Data Fig. 5, Supplementary Data 2). This suggests that there are distinct nuclei populations within peripheral endosperm whose function is primarily to carry out photosynthesis and biosynthesis of different nutrients. In contrast to these clusters, VxC E5 was enriched for down-

regulation of photosynthesis genes and fructose 1,6 bisphosphate metabolism, but had few upregulated genes overall and appeared relatively metabolically inactive. This cluster was therefore classified as 'inactive' peripheral endosperm (Fig. 1e).

Defining seed coat nuclei cluster identity

The developing seed coat consists of five distinct cell layers and the chalazal seed coat region (Extended Data Fig. 3, Fig. S3). The innermost cell layer ii1, or endothelium, is characterized by elevated production of proanthocyanidins (*10*). Genes related to proanthocyanidin production were strongly upregulated in VxC E7 and E8 (Extended Data Fig. 3). Two other clusters, CxV S4 and VxC S4, had high expression of *BAN*, a gene known to be expressed in the endothelium, but not *TT12*, another endothelium marker. By contrast, CxV S6 and VxC S7, S8 had high expression of both genes (Fig. S3). Based on these and other gene expression data, we characterized CxV S4 and VxC S4 as 'potential endothelium' and VxC S7, S8 as 'proanthocyanidin-synthesizing endothelium'. The 'potential endothelium' clusters were also characterized by enriched expression of genes involved in dNTP synthesis and glutamine metabolism (Extended Data Fig. 3). Clusters CxV S2 and VxC S2 likely correspond to the epidermal layer (oi2). Both clusters expressed the epidermal cell fate transcription factor *GLABRA2*, and were enriched for cutin biosynthesis and transport (Extended Data Fig. 3). The clusters with the most nuclei, CxV S1 and VxC S1, had elevated expression of genes related to the synthesis of flavonol, which is produced in the outer integument layers (Extended Data Fig. 3, *11*). The chalazal seed coat clusters, CxV S3 and VxC S6, were enriched for sucrose biosynthesis, mannose response, and glucose response, all consistent with a nutrient transfer function for this region. The chalazal seed coat also had elevated expression of multiple transcription factors, including the MADs domain transcription factors *AGL15*, *AGL16*, *AGL18*,

AGL2/SEP1, *AGL69/MAF4*, and the interacting factors *AGL9/SEP3* and *AGL11/STK*, which influence mechanical properties of the seed coat (*12*).

Supplementary Methods

Confocal microscopy

Controlled floral pollinations were performed for each cross depicted in Fig. 1a. Siliques were collected 4 DAP, dissected to open up the carpel wall, and fixed in FAA overnight at 4°C. Samples were dehydrated with an ethanol series to 100% ethanol, and cleared by gradual infiltration with immersion oil (Immersol 518F, Zeiss). Seeds were removed from cleared siliques and mounted on slides. Whole seeds were imaged on a Zeiss LSM700 Confocal Microscope, with settings optimized to detect auto-fluorescence (Excitation: 405nm at 0.1% power, 488nm 11% power, 555nm at 28% power. Detection: from 415 to 735 nm).

Probes for in situ hybridization

Name	Gene ID	Length (bp)	Forward Primer (5'-3')	Reverse Primer(5'-3')
CYCD4;2	AT5G10440	437	ACCAAAGCAAATTCTACAATAGAG	GTCCAAATTGAAGTTCCTCACAG
GA20OX5	AT1G44090	530	TAGGCAACCATCGTCAAGAGATC	CCTGTGGTAACAACTCCTGTAG
MEE56	AT4G13380	472	AGGCAGAAATCTTGACGATGAATG	GCCATTTCTTTCTCATCATCACTAC
UMAMIT25	AT1G09380	585	AGGCATCAGCACAAGCCAAAG	ATACGAACTAGAGACCAATAACCAG
HMG1,2	AT4G11080	634	GACGGAGCTGAAGAAACTGC	CGCCATCTGATCATAAGGAGCC
NEP-interacting	AT2G44240	622	GGTTTTGCGGTAGCTCTGATG	GATAAACCTGCCAACCAGCTTA

Metaplots of coverage over genes and introns

To generate plots in Fig. S2, bigWig coverage files for each library were generated using the deepTools function bamCoverage (*13*, v.3.2.0) with options -bs 1 --normalizeUsing CPM. Coverage over genes and introns was calculated using deepTools computeMatrix in scale-regions mode with options -a 500 -b 500 -m 2000 -bs 50 for genes and -a 200 -b 200 -m 200 -bs 10 for introns. Average profiles for each library were then obtained using deepTools plotProfile

with option --perGroup and extracting matrix file with --matrixFile option. Average and s.d. of coverage across all libraries was calculated for each bin in the profile and used to generate plots in Fig. S2.

Plots of average total or allelic expression of genes or gene groups over clusters

Dot plots of average total expression of either individual genes (e.g. Fig. 2a, Extended Data Fig. 6) or groups of genes (e.g. Fig. 1c, Extended Data Fig. 2) were made using a custom script (single_cell_RNAseq_plots.R) available in the Github repository, in 'dot' mode. Briefly, the raw count matrix (rows: n genes x m columns: nuclei) was first converted into $\log_2$(CPM) values, and for each gene, average $\log_2$(CPM>0) was calculated across all nuclei in each cluster to obtain a matrix of n genes x c clusters. Only nuclei with nonzero CPM values (CPM > 0) were used to calculate the average. The fraction of nuclei in each cluster with at least one detected read (CPM > 0) was also calculated. Plots were generated by indicating average $\log_2$(CPM>0) using dot color, and fraction of nuclei with at least one detected read in gene as dot size. For plots aggregating groups of genes (e.g. Fig. 1c), average $\log_2$(CPM>0) and fraction of zeros were also averaged across all genes in each group to obtain a single value per group.

Plots of average total or allelic expression of individual genes shown in Fig. 3c,e, Extended Data Fig. 3D, Extended Data Fig. 9, Fig. S3, S13, and S16 were made using the same custom script (single_cell_RNAseq_plots.R) in 'lin' mode. For total expression, average $\log_2$(CPM>0) and fraction of zeros were obtained as above. For allelic expression, $\log_2$(CPM>0) and fraction of zeros were obtained for maternal and paternal counts separately. Paternal counts from each cluster were also fit to the appropriate distribution (NB or ZINB, see 'Modeling maternal and paternal counts' in Supplementary Material), and two values were randomly drawn from this distribution and summed for each nucleus in the cluster to simulate a doubling of

paternal dosage (dotted blue lines). Plots in Fig. S8 were made using the same script in 'cmp' mode.

Lists of marker genes from seed compartment RNA-seq data obtained by laser capture microdissection (LCM) of seed tissue *(4)* were obtained from Schon et al. 2017 (*5*). All genes that were considered markers for different tissues at different timepoints (n = 181) were discarded. Markers for globular and heart stages (the stages used in this study) were then kept for analysis (n = 1,945, list of markers provided in Supplementary Data 2). Note that some tissue marker genes are confounded with other features. For example, embryo 'markers' are confounded with cell cycle phase, such that an apparent enrichment for embryo identity in some differentially expressed gene groups is because of the expression of M-phase genes (Fig. S4, Supplementary Data 2).

Identifying differentially expressed genes

To identify genes that were differentially expressed within our dataset, DEsingle was run on all possible pairs of clusters within the 14 CxV endosperm clusters, within the 11 VxC endosperm clusters, between the CxV and VxC endosperm clusters, within the 6 CxV seed coat clusters, within the 8 CxV seed coat clusters, between the CxV and VxC seed coat clusters, and between each endosperm cluster and each seed coat cluster (741 comparisons). Finally, to control for batch effects that might confound the DE analysis, DEsingle was also used to look for DE genes between all batches (batches = 96-well plates, prepared on different days) within the CxV and VxC seed coat and endosperm (8 batches represented in each of CxV endo, CxV seed coat, VxC endo and VxC seed coat, 28 * 4 = 112 comparisons).

For each non-batch comparison, all genes that passed stringent significance cutoffs (pval < 0.0001 and abs[$\log_2$(fold change)] > 2) were extracted. These were combined across all comparisons to obtain a list of 4,540 genes that were significantly differentially expressed

across at least one experimental comparison (cluster, genotype or tissue). This was repeated for all batch comparisons, resulting in 97 genes called as differentially expressed across at least one pair of batches/plates, suggesting that batch effects in our data are minimal. 40 of the batch differentially expressed genes were also differentially expressed in the main analysis and were censored, yielding a final set of 4,500 differentially expressed genes (Supplementary Data 2).

Cell cycle analysis

A list of 22 genes associated with various phases of the cell cycle, including G0, was manually curated from the literature (*6*, *7*). The count matrix (n=1,437 nuclei) was first filtered to remove lowly expressed genes, and count values were converted to CPM. Of the 1,437 nuclei, 1,309 (91%) had expression of at least one of the 22 cell cycle marker genes and were retained for further analysis. t-SNE was first performed over CPM values for the 22 cell cycle genes, using the Rtsne package (*14*) with initial_dims = 2 and perplexity = 100. Projected points were then clustered using k-means clustering with k = 6, corresponding approximately to the G0, G1, G1 to S, S, G2 and M phases of the cell cycle (Supplementary Data 3).

Dijkstra's algorithm was used to trace a trajectory that was required to pass through the medoid of each k-means cluster. The trajectory was recalculated 200 times, sampling only 50% of the nuclei each time. Each of the 200 trajectories were then combined (*15*) and averaged to build a single smoothed, consensus trajectory through the tSNE plot. Each point in the plot (representing a single nucleus) was then projected onto the consensus trajectory to estimate the positioning of each nucleus along the cell cycle. To identify other genes that vary significantly according to the cell cycle in our dataset, library depth-normalized counts were obtained using the procedure used by edgeR (*16*). A hurdle model, which uses a truncated distribution to model the additional zeros present in single-cell RNA-seq data and has been used previously in similar contexts (*17*) was used to model the normalized counts. The distribution used for the non-zero

component of the hurdle model was the negative binomial distribution, which is used frequently to model count data from RNA-seq (*16,18*). We used this count model to test each gene using a regression approach (hurdle() function in R pscl package (*19*) to determine whether the cell cycle trajectory explains a significant amount of the variation in each gene's expression. P-values were adjusted using the Bonferroni method, and genes with adjusted *p*-values < 0.01 were considered significantly cell-cycle dependent. This analysis identified a total of 1,065 cell-cycle-dependent genes (Supplementary Data 3).

Identifying imprinted genes from snRNA-seq data

We developed a method for assessing imprinting that accounts for maternal and paternal dosage in endosperm (single_cell_ASE_analysis.R, in Github repository).

1. Modeling maternal and paternal counts and testing for parental bias

Paternal counts were fit to the negative binomial (NB) and zero-inflated negative binomial (ZINB) distributions, and the Akaike Information Criterion was used determine which had the best fit (Fig. S8). The NB distribution is commonly used to model counts in bulk RNA-seq (*16,18*), while the ZINB extends the NB with an additional parameter to model dropout events and has been used for scRNA-seq analysis (*20*,*21*). To account for maternal genome dosage, maternal count data were instead modeled as the sum of two independent, identically distributed ZINB or NB random variables. The NB has parameters μ, σ, representing the mean and variance respectively, and the ZINB has an additional ν parameter modeling zero inflation (*22*):

Distribution	Equation	Mean	Variance
$X \sim NB(\mu, \sigma)$	$P_X(x \mid \mu, \sigma) = \frac{\Gamma\left(x + \frac{1}{\sigma}\right)}{\Gamma\left(\frac{1}{\sigma}\right)\Gamma(x+1)}\left(\frac{\sigma\mu}{1+\sigma\mu}\right)^x\left(\frac{1}{1+\sigma\mu}\right)^{\frac{1}{\sigma}}$	$E[X] = \mu$	$Var(X) = \mu + \sigma\mu^2$
$X \sim ZINB(\nu, \mu, \sigma)$	$P(x \mid \nu, \mu, \sigma) = \begin{cases} \nu + (1-\nu) * p_{X'}(0 \mid \mu, \sigma) & if\ x = 0 \\ (1-\nu) * p_{X'}(x \mid \mu, \sigma) & if\ x = 1,2,3 \ldots \end{cases}$ where $X' \sim NB(\mu, \sigma)$	$E[X] = (1-\nu)\mu$	$Var(X) = \mu(1-\nu) * (1 + (\nu + \sigma)\mu)$

Parameter estimates were obtained by Maximum Likelihood (*23*). Maternal and paternal parameter estimates were well-correlated with each other (Fig. S9), suggesting that the overall transcriptional kinetics of the maternal and paternal alleles are similar for most genes in our dataset. Both imaging studies and single-cell studies have found that the two alleles of most genes function independently and share similar kinetics (*24*).

Our analysis revealed minor but consistent skew in favor of higher maternal allele expression at most genes, and maternal (1-ν) and μ estimates were consistently slightly higher than paternal estimates (Fig. S9). This effect was more pronounced in Col x Cvi than in Cvi x Col F_1 nuclei, suggesting that technical factors, such as mapping bias in favor of the sequenced strain Col, played a role, though we cannot rule out biological factors (also see 'identifying imprinted genes below'). In light of this, we tested an adjusted null hypothesis for deviation from the expected 2m:1p ratio that accounted for the average skew in the data.

Under the null hypothesis that the gene was not imprinted and that the maternal and paternal ZINB/NB parameters were equal after adjusting for average maternal skew, p-values were obtained using a likelihood ratio test.

2. Evaluating our approach using simulated count data

To test the power and accuracy of our approach, we simulated maternal and paternal count data resembling the CxV and VxC datasets, but with varying levels of overall expression, parental bias, nuclei-to-nuclei variability, and number of nuclei (Extended Data Fig. 8). Counts were drawn from ZINB distributions with known parameters. Our approach was able to accurately identify genes with parental bias under a wide range of conditions, with accuracy increasing as expression levels or parental bias increased. Simulations included a degree of maternal skew similar to the one in the real data, and showed that failing to adjust for this skew when testing for parental bias led to high false positive maternal bias calls, while an adjusted hypothesis had very few false positives (Extended Data Fig. 8).

Our method also identified maternal and paternal bias in simulations of highly expressed or highly biased genes from as few as 5 or 10 simulated nuclei, though with a high false negative rate (Extended Data Fig. 8). Our approach therefore likely has a high false negative rate for lowly expressed genes or genes expressed only in rare subpopulations. For example, *SUVH7* and *MEA* are both well-known imprinted genes (*25, 26*), but were not among the PEGs and MEGs identified in this study. However, *SUVH7* expression is limited to the chalazal cyst, a relatively rare population in our dataset. Although our method found that *SUVH7* is clearly paternally biased in VxC, which has more cyst nuclei, we lacked power to identify bias in CxV, which has only 6 cyst nuclei (Supplementary Data 4). Similarly, *MEA* displays maternal bias in the data (Supplementary Data 4), but is only strongly expressed in CxV E14, a cluster with only 11 nuclei. Therefore, our analysis likely undercounts the true number of imprinted genes. However, false positive rates were low across all conditions tested, suggesting that our method has high specificity but tends to be conservative (Extended Data Fig. 8).

3. Identifying imprinted genes

Using the method outlined above, *p*-values for significant parental bias were generated for each gene in that had at least 10 allelic counts, and corrected using the Benjamini-Hochberg method (*27*). The 4 DAP CxV and VxC datasets were analyzed separately. Genes with adjusted *p*-value < 0.05 and a $\log_2$(m/p) value more than 0.5 away from the median $\log_2$(m/p) value across all genes were considered significantly biased (Fig. S10, Supplementary Data 4). Genes were divided into weak, medium, or strong bias based on $\log_2$(m/p) values, and results from both CxV and VxC datasets were combined into a final 'overall status': MEGs/PEGs (maternal/paternal bias in both crosses), Col/Cvi bias (bias in favor of the same strain in both crosses), bias in only one cross, or no bias (Fig. S10, Fig. S11, Supplementary Data 4). Unbiased genes with fewer than 20 allelic counts overall likely lacked statistical power and were classified as having 'too few reads'.

Of the MEGs and PEGs previously identified from whole-endosperm RNA-seq (*28*) with sufficient data for our analysis, most were also considered imprinted based on our dataset (Fig. S12). However, a number of the MEGs and PEGs identified in this study were not previously identified (Fig. S12). Some of these lacked data in the previous study, either because they were not detected or the annotation did not yet exist. However, a number of the MEGs identified in this study were censored from the analysis in the previous study because laser-capture microdissection datasets suggest that these genes are more highly expressed in the seed coat than in the endosperm (*4*). An ongoing concern with bulk endosperm datasets is that contamination from the seed coat, a maternal tissue, can lead to false positives among MEGs; this type of contamination has been shown to occur in some manually dissected endosperm datasets (*5*). However, with our data, it is possible to carefully separate nuclei derived from seed coat from nuclei derived from endosperm based on a number of expression features (Fig. 1), so our CxV and VxC endosperm datasets should not have seed coat contamination. Thus, many MEGs identified here are indeed more highly expressed in seed coat than in endosperm (Figs.

S12, S13, *3*). After also noting a widespread skew towards maternal expression in the data (Figs. S9, S11), we tested the possibility that this phenomenon was caused by mRNAs from the seed coat being carried over in the sorting buffer during FANS by adding an additional wash step prior to sorting (see 'Seed Nuclei FANS'). Surprisingly, endosperm nuclei that had been treated with the extra wash step showed an even more pronounced maternal skew compared to unwashed nuclei (Supplementary Data 1), and washed nuclei had even stronger maternal expression of most MEGs compared to unwashed nuclei (Fig. S14). It remains unclear whether this phenomenon is biological, or a consequence of technical factors related to snRNA-seq data. However, overall these results suggest that our approach is able to identify imprinted genes that could not previously be evaluated by bulk RNA-seq.

Supplementary References

1. Mansfield, S. G., Briarty, L. G. Development of the free-nuclear endosperm in Arabidopsis thaliana. *Arabidopsis Information Service* **27**, (1990).
2. Brown, R. C., Lemmon, B. E., Nguyen, H., Olsen, O.-A. Development of endosperm in Arabidopsis thaliana. *Sex. Plant Reprod.* **12**, 32-42 (1999).
3. Brown, R. C., Lemmon, B. E., Nguyen, H. Events during the first four rounds of mitosis establish three developmental domains in the syncytial endosperm of Arabidopsis thaliana. *Protoplasma* **222**, 167-174 (2003).
4. Belmonte, M. F., *et al.* Comprehensive developmental profiles of gene activity in regions and subregions of the Arabidopsis seed. *Proc. Natl. Acad. Sci. USA* **110**, E435–E444 (2013).

5. Schon, M. A. & Nodine, M. D. Widespread contamination of Arabidopsis embryo and endosperm transcriptome data sets. *Plant Cell* **29**, 608–617 (2017).
6. Menges, M., de Jager, S. M., Gruissem, W., Murray, J. A. H. Global analysis of the core cell cycle regulators of Arabidopsis identifies novel genes, reveals multiple and highly specific profiles of expression and provides a coherent model for plant cell cycle control. *Plant J.* **41**, 546-566 (2005).
7. Menges, M., Hennig, L., Gruissem, W., Murray, J. A. H. Cell Cycle-regulated Gene Expression in *Arabidopsis*. *J. Biol. Chem.* **277**, 41987-42002 (2002).
8. Besnard, J., *et al.* Arabidopsis UMAMIT24 and 25 are amino acid exporters involved in seed loading. *J. Exp. Bot.* **69**, 5221-5232 (2018).
9. Moussu, S., *et al.* ZHOUPI and KERBEROS mediate embryo/endosperm separation by promoting the formation of an extracuticular sheath at the embryo surface. *Plant Cell* **29**, 1642-1656 (2017).
10. Debeaujon, I., *et al.* Proanthocyanidin-accumulating cells in Arabidopsis testa: regulation of differentiation and role in seed development. *Plant Cell* **15**, 2514-2531 (2003).
11. Pourcel, L., *et al. TRANSPARENT TESTA10* encodes a laccase-like enzyme involved in oxidative polymerization of flavonoids in *Arabidopsis* seed coat. *Plant Cell* **17**, 2966-2980 (2005).
12. Ezquer, I., *et al.* The developmental regulator SEEDSTICK controls structural and mechanical properties of the Arabidopsis seed coat. *Plant Cell* **28**, 2478-2492 (2016).
13. Ramírez, F., *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160-W165 (2016).

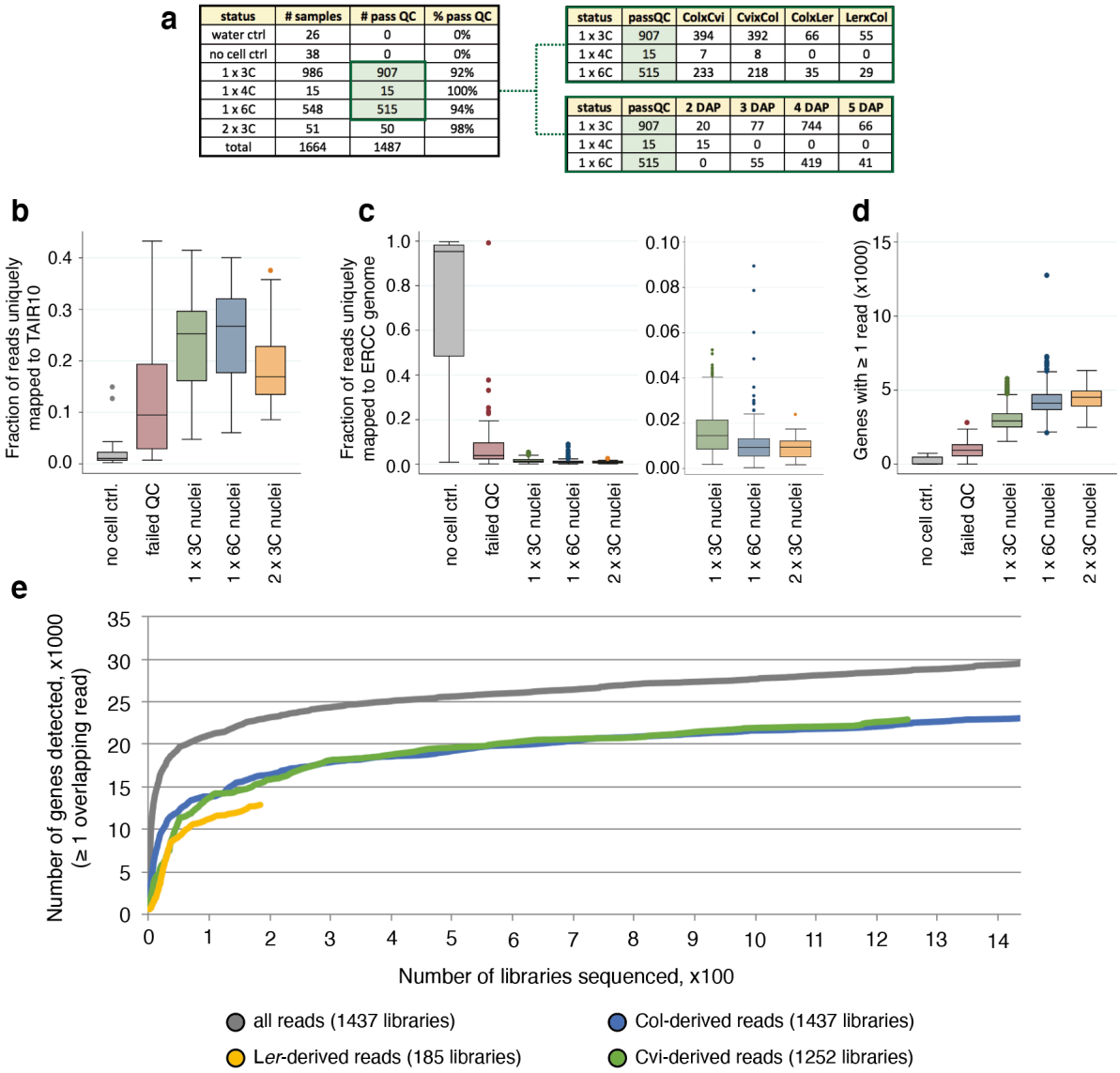
14. Krijthe, J. H. Rtsne: T-Distributed Stochastic Neighbor Embedding using a Barnes-Hut Implementation, https://github.com/jkrijthe/Rtsne. (2015).
15. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal Complex Systems* **1695**, (2006).
16. Robinson, M. D., McCarthy, D. J., Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
17. McDavid, A., *et al.* Modeling bi-modality improves characterization of cell cycle on gene expression in single cells. *PLoS Comput. Biol.* **10**, e1003696 (2014).
18. Love, M. I., Huber, W., Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
19. Zeileis, A., Kleiber, C., Jackman, S. Regression Models for Count Data in R. *J. Stat. Softw.* **27**, 1-25 (2008).
20. Miao, Z., Deng, K., Wang, X., Zhang, X. DEsingle for detecting three types of differential expression in single-cell RNA-seq data. *Bioinformatics* **34**, 3223-3224 (2018).
21. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., Vert, J.-P. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun*. **9**, 284 (2018).
22. Rigby, R. A. & Stasinopoulos, D. M. *A flexible regression approach using GAMLSS in R* (London Metropolitan University, London, 2010).
23. Henningsen, A. & Toomet, O. maxLik: A package for maximum likelihood estimation in R. *Comput. Stat.* **26**, 443-458 (2011).
24. Jiang, Y., Zhang, N. R., Li, M. SCALE: modeling allele-specific gene expression by single-cell RNA sequencing. *Genome Biol.* **18**, 74 (2017).

25. Gehring, M., Missirian, V., Henikoff, S. Genomic analysis of parent-of-origin allelic expression in *Arabidopsis thaliana* seeds. *PLoS ONE* **6**, e23687 (2011).
26. Kinoshita, T., Yadegari, R., Harada, J. J., Goldberg, R. B., Fischer, R. L. Imprinting of the MEDEA Polycomb gene in the Arabidopsis endosperm. *Plant Cell* **11**, 1945-1952 (1999).
27. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289-300 (1995).
28. Pignatta, D., *et al.* Natural epigenetic polymorphisms lead to intraspecific variation in Arabidopsis gene imprinting. *Elife* **3**, e03198 (2014).
29. Picard, C. L. & Gehring, M. Proximal methylation features associated with nonrandom changes in gene body methylation. *Genome Biol.* **18**, 73 (2017).
30. Windsor, J. B., Symonds, V. V., Mendenhall, J., Lloyd, A. M. Arabidopsis seed coat development: Morphological differentiation of the outer integument. *Plant J.* **22,** 483–493 (2000).
31. Nakaune, S., *et al.* A vacuolar processing enzyme, deltaVPE, is involved in seed coat formation at the early stage of seed development. *Plant Cell* **17**, 876-887 (2005).
32. Mizzotti, C., *et al.* SEEDSTICK is a master regulator of development and metabolism in the Arabidopsis seed coat. *PLoS Genet*. **10**, e1004856 (2014).
33. Stadler, R., Lauterbach, C., Sauer, N. Cell-to-cell movement of green fluorescent protein reveals post-phloem transport in the outer integument and identifies symplastic domains in Arabidopsis seeds and embryos. *Plant Physiol*. **139**, 701-712 (2005).
34. Debeaujon, I., Peeters, A. J., Léon-Kloosterziel, K. M., Koornneef, M. The TRANSPARENT TESTA12 gene of Arabidopsis encodes a multidrug secondary

transporter-like protein required for flavonoid sequestration in vacuoles of the seed coat endothelium. *Plant Cell* **13**, 853-871 (2001).

35. Coen, O., *et al.* A TRANSPARENT TESTA transcriptional module regulates endothelium polarity. *Front Plant Sci*. **10**,1801 (2020).
36. Le, B. H., *et al*. Global analysis of gene activity during Arabidopsis seed development and identification of seed-specific transcription factors. *Proc Natl Acad Sci U S A*. **107**, 8063-8070 (2010).
37. Gao, M. J., *et al.* SCARECROW-LIKE15 interacts with HISTONE DEACETYLASE19 and is essential for repressing the seed maturation programme. *Nat Commun*. **6**, 7243 (2015).
38. Molina, I., Ohlrogge, J. B., Pollard, M. Deposition and localization of lipid polyester in developing seeds of Brassica napus and Arabidopsis thaliana. *Plant J.* **53**, 437-449 (2008).

Supplementary Figures



status	# samples	# pass QC	% pass QC
water ctrl	26	0	0%
no cell ctrl	38	0	0%
1 x 3C	986	907	92%
1 x 4C	15	15	100%
1 x 6C	548	515	94%
2 x 3C	51	50	98%
total	1664	1487	

status	passQC	ColxCvi	CvixCol	ColxLer	LerxCol
1 x 3C	907	394	392	66	55
1 x 4C	15	7	8	0	0
1 x 6C	515	233	218	35	29

status	passQC	2 DAP	3 DAP	4 DAP	5 DAP
1 x 3C	907	20	77	744	66
1 x 4C	15	15	0	0	0
1 x 6C	515	0	55	419	41

Fig. S1. Summary of library quality and comparison to controls. (a) Left: summary of libraries obtained: water ctrl = Nextera XT library built with water (no DNA); no cell ctrl = no nucleus sorted into well, but carried through the entire prep; 1 x 3C = single nucleus from 3C FANS peak, 1 x 4C = single nucleus from 4C FANS peak (2 DAP only, due to no visible 3C peak); 1 x 6C = single nucleus from 6C FANS peak, 2 x 3C = two nuclei sorted from 3C peak. The 1,437 non-control libraries highlighted in green passed basic QC cutoffs (see Methods). Right: 1,437 high-quality libraries by ploidy and genotype (top) or stage (bottom). (b-d) Distribution in control, failed, and high-quality libraries, of (b) the fraction of reads that mapped to TAIR10, (c) the fraction of reads that mapped to the ERCC spike-in genome, and (d) the number of genes detected. 1x3C, 1x6C and 2x3C are libraries that passed QC filters. Number of nuclei in each category shown in (a). Median, interquartile range and upper-/lower-adjacent values (1.5*IQR) indicated by center line, box, and whiskers of each boxplot, respectively. In (c), panel on right is zoomed-in view of libraries that passed QC. (e) Total number of genes detected as the number of single nuclei sequenced that passed QC increases, defined as at least 1 read mapping to gene. For allelic expression, gene detection was defined as at least 1 allele-specific read (read overlapping a SNP) mapping to gene.

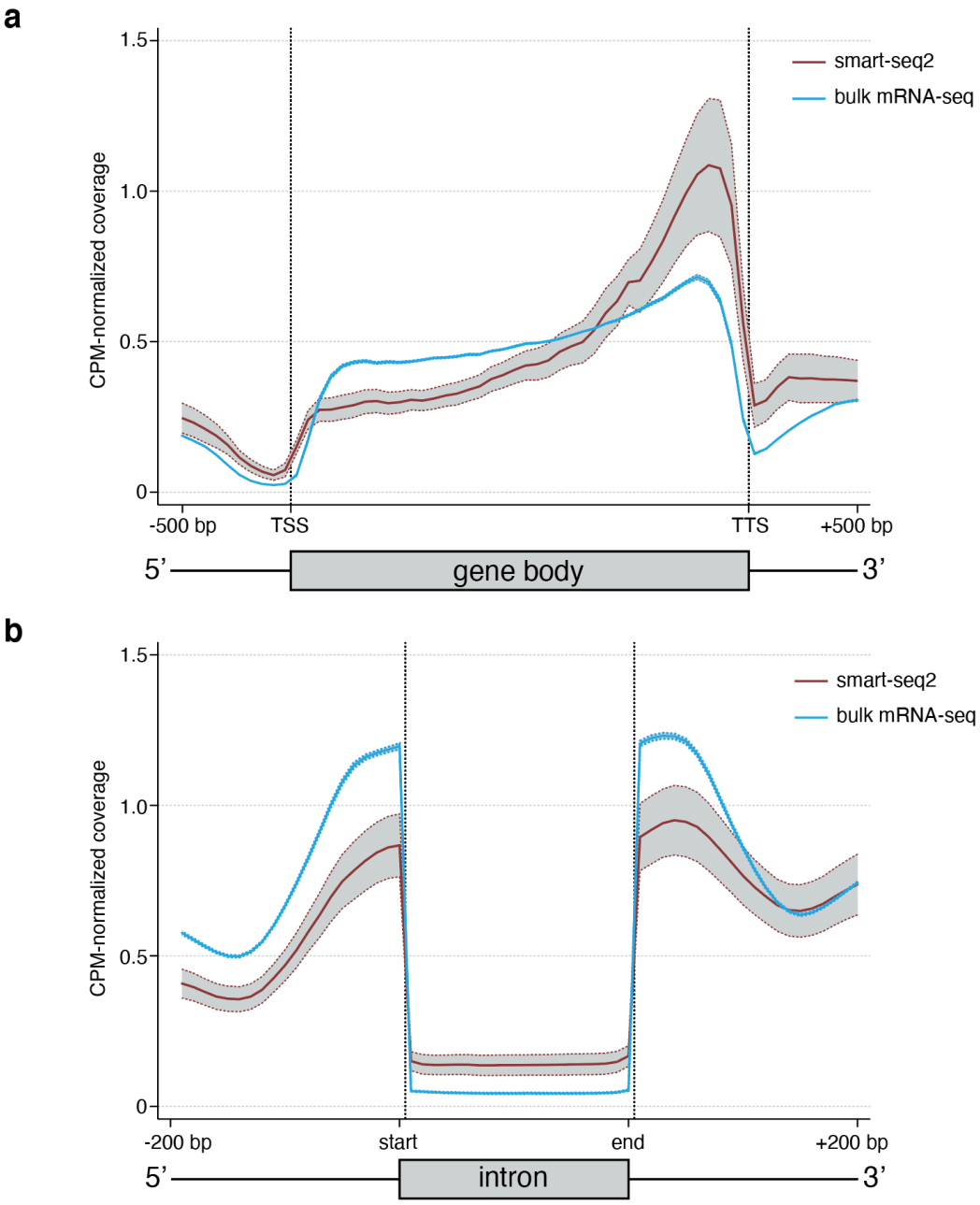


Fig. S2. Comparison of bias in gene and intron coverage in Smart-seq2 libraries vs. published bulk mRNA-seq data. Single-nuclei Smart-seq2 libraries generated in this study were compared to 6 published bulk whole-cell RNA-seq libraries (3 replicates each of Col and Cvi parent lines, *29*). Central solid line indicates overall average, red for Smart-seq2 libraries (average of all 1437 Smart-seq2 libraries that passed QC) and blue for bulk mRNA-seq (average of 6 libraries, *29*). Dotted lines and shaded area indicate +/- 1 SD.

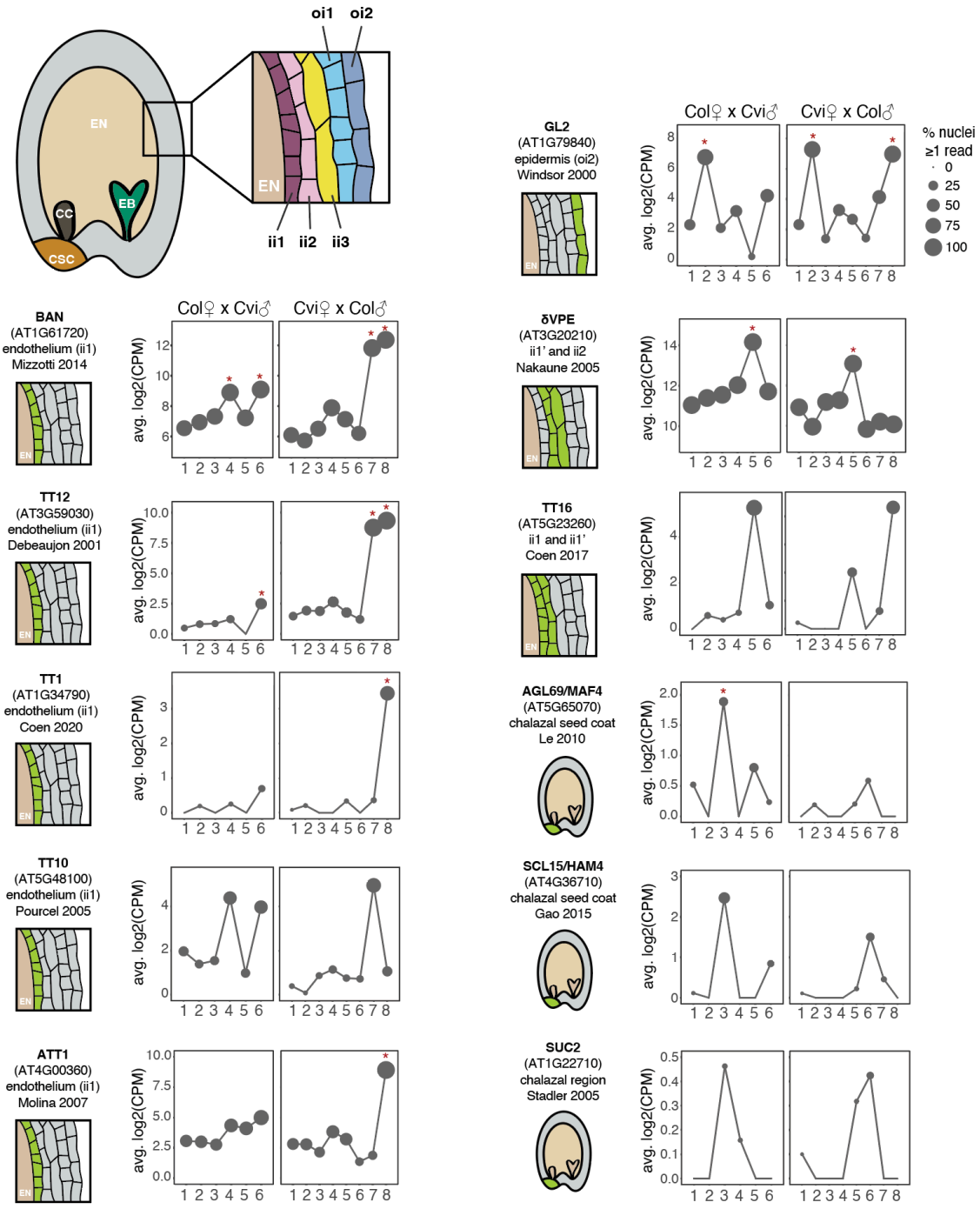


Fig. S3. Expression of genes with known cell layer or region-specific seed coat expression patterns in seed coat nuclei clusters. Genes were curated from the literature and only considered if expression pattern was determined either by GUS staining or in situ hybridization (*11*, *30-38*). Schematic at left of each plot shows the localization of the gene according to the indicated study. Red star: gene was significantly upregulated in that cluster ($p < 0.05$, permutation test). *TT1*, *TT10*, *TT16*, *SCL15* and *SUC2* were too lowly expressed to evaluate statistical significance. (Top Left) Schematic of seed showing location of the chalazal seed coat region (CSC), as well as embryo (EB), endosperm (EN), and chalazal cyst (CC). Inset: schematic of the five cell layers in seed coat, from ii1 (the endothelium, innermost cell layer) to oi2 (epidermis, outermost cell layer).

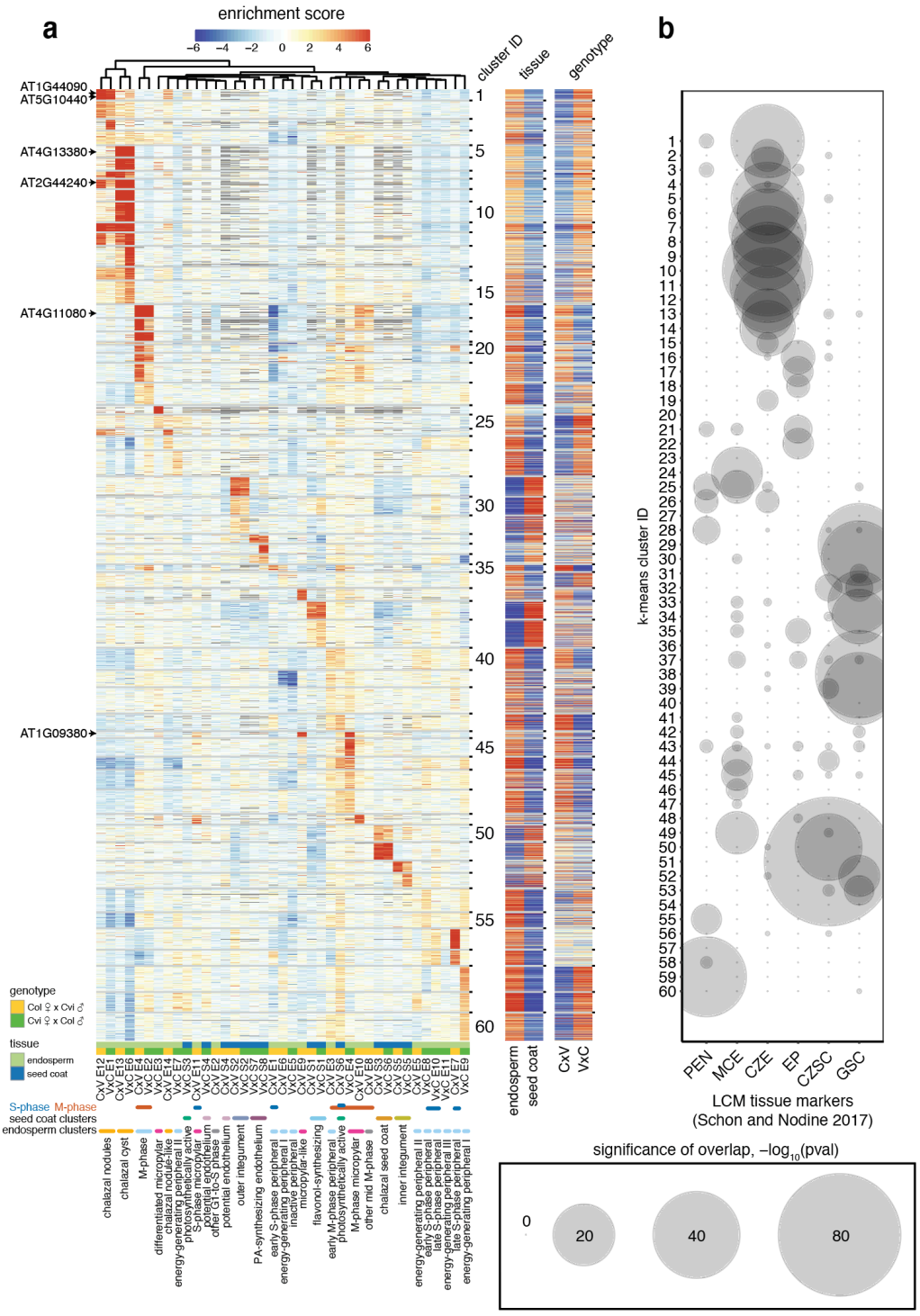


Fig. S4. Overall expression patterns of 4,500 genes significantly differentially expressed among the endosperm and seed coat clusters. (a) Heatmap of gene expression 'enrichment scores' for all 4,500 genes clustered by k-means clustering (k =60). Genes used for in situ hybridization (Fig. 2, Extended Data Fig. 6) indicated at left. Nuclei clusters are labeled according to cluster identity (Fig. 1e); clusters enriched for genes related to M-phase or S-phase (Fig. 1d) also indicated. Columns on right indicate relative enrichment score in seed coat vs. endosperm and Col x Cvi vs. Cvi x Col. (b) Significance of overlap between genes in each k-means cluster from (a), and marker genes for different seed compartments (*4*, curated by *5*). P-values were computed using the hypergeometric test.

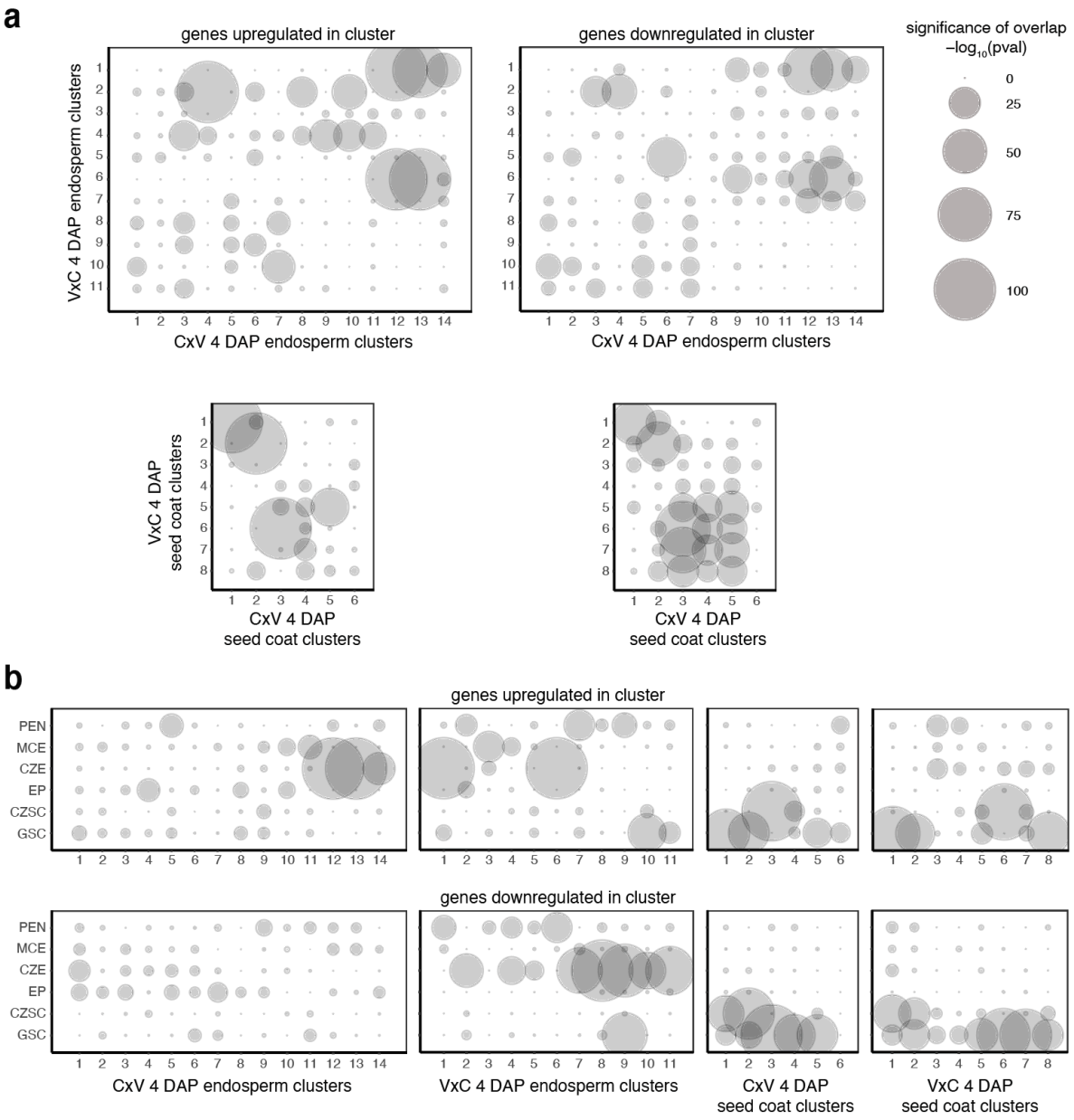


Fig. S5. Overlap between significantly up- and down-regulated genes across all 4 DAP nuclei clusters and with published seed domain markers. (a) Significance of overlap of significantly up- and down-regulated genes between Col x Cvi and Cvi x Col 4 DAP endosperm or seed coat clusters. p-value obtained using hypergeometric test. Values above 100 (-log(pval)) were truncated to 100 for plotting. (b) Significance of overlap between genes significantly up- or down-regulated in the nuclei clusters, and marker genes for six seed tissues (*4, 5*). Legend same as (a). PEN = peripheral endosperm, MCE = micropylar endosperm, CZE = chalazal endosperm, EP = embryo proper, CZSC = chalazal seed coat, GSC = general seed coat

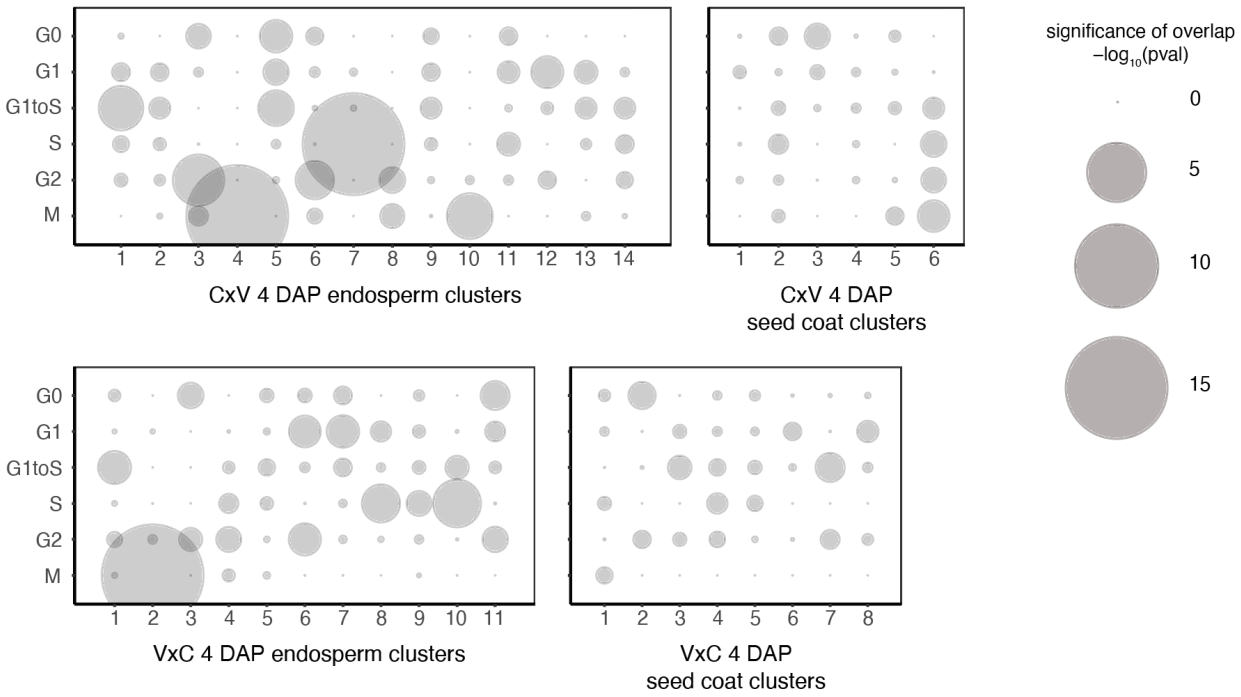


Fig. S6. Significance of overlap between nuclei in each of the endosperm and seed coat clusters and nuclei in each stage of the cell cycle. Cell cycle phase obtained from clustering analysis shown in Extended Data Fig. 7. p-value obtained from hypergeometric test.

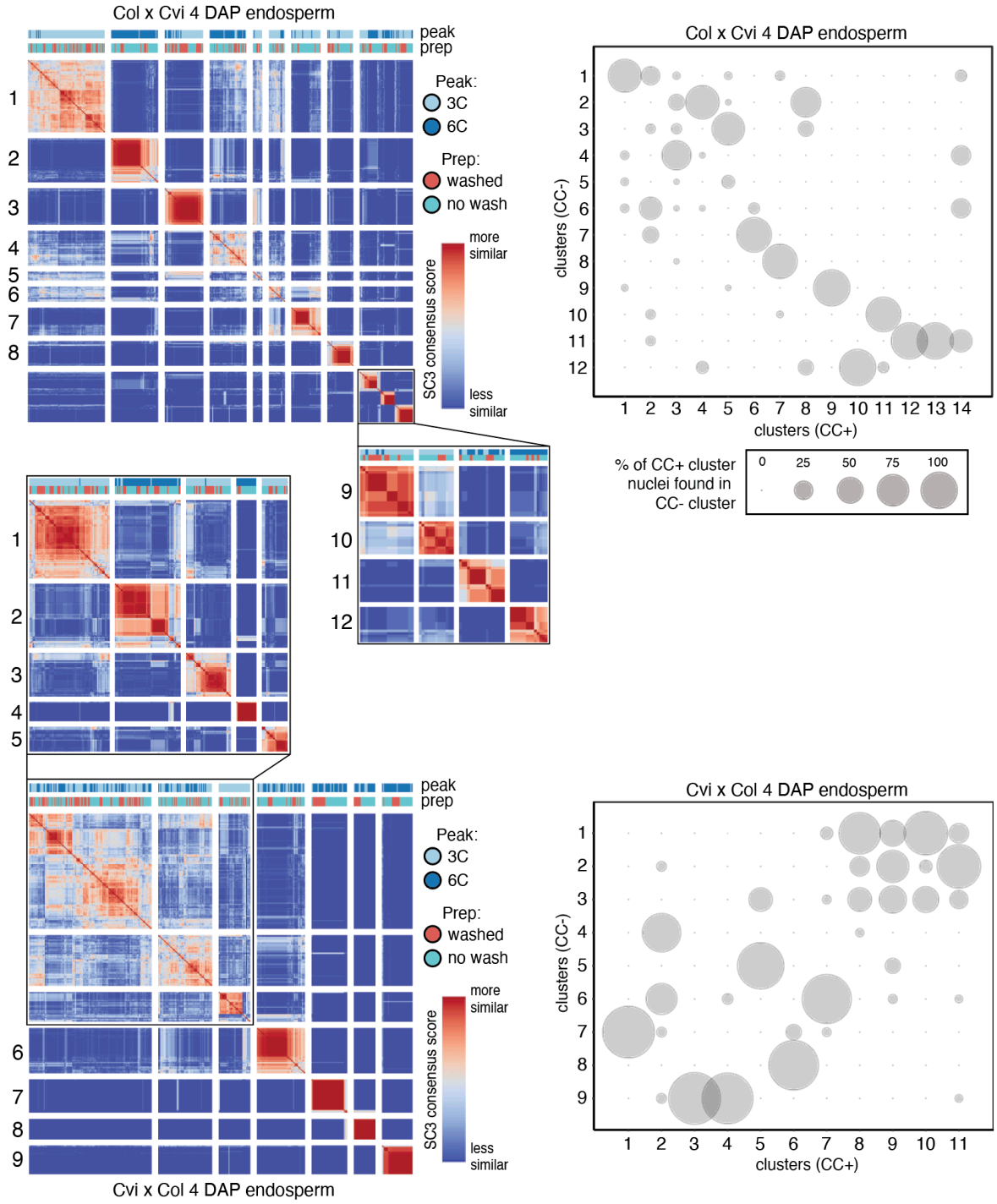


Fig. S7. Omitting cell cycle correlated genes does not substantially alter clusters. SC3 clustering was repeated while omitting the 1065 genes whose expression patterns correlated significantly with cell cycle trajectory (Supplementary Data 3). The nuclei in each of the new clusters obtained without cell-cycle correlated genes (CC-) were compared to those in the original clusters obtained from the full dataset (CC+), and the percent of CC+ nuclei that were present in the indicated CC- cluster are shown in plots at right. Most nuclei from CC+ clusters fell largely (>75%) into a single CC- cluster, suggesting the clustering was mostly unaffected by omitting the cell-cycle correlated genes. Clusters with strong expression of cell-cycle related genes (e.g. CxV E7; see Fig. 1d) were also not strongly affected.

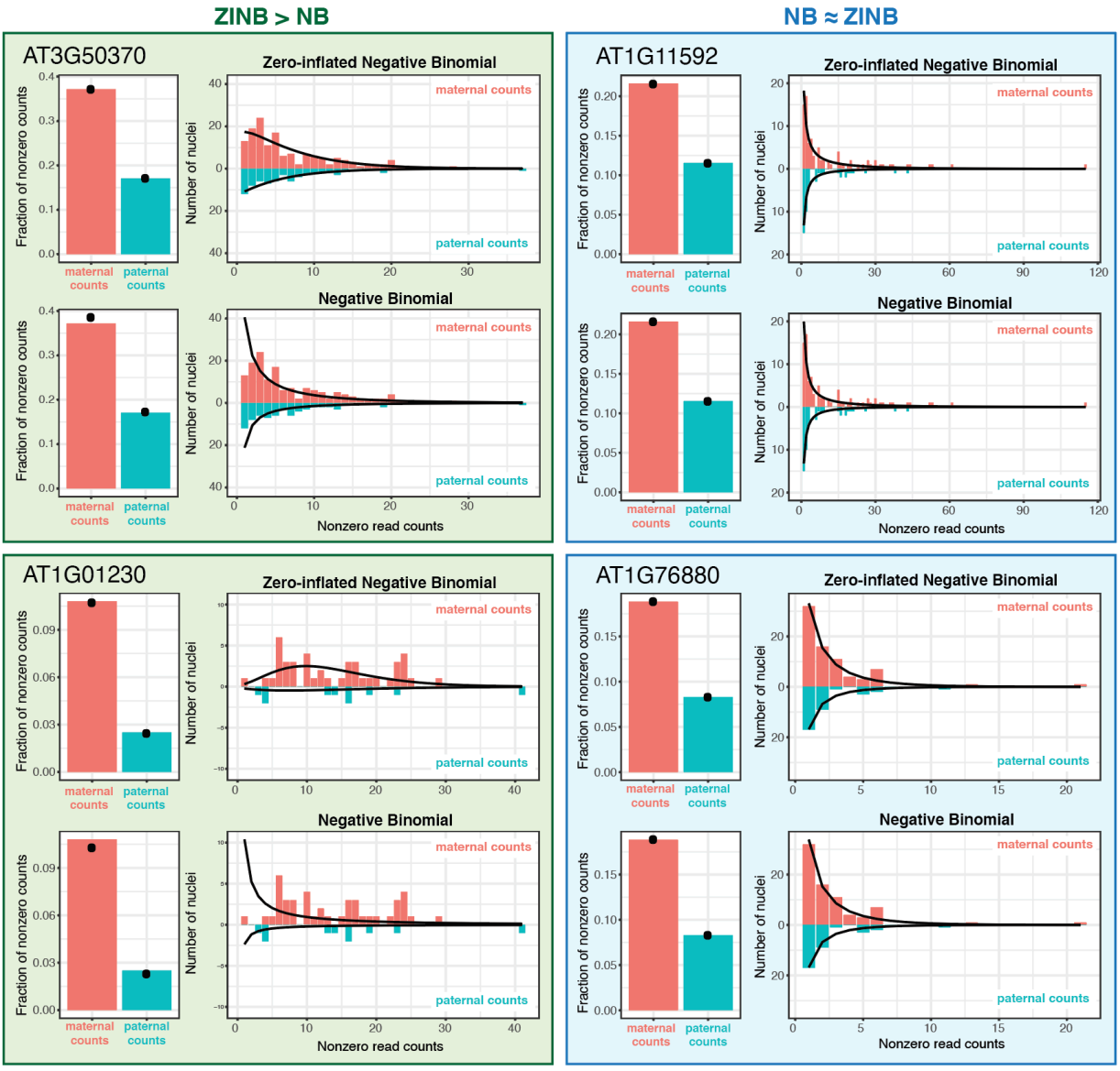


Fig. S8. ZINB distribution accurately models allelic count data. Examples of maternal and paternal count data from four genes fit to the zero-inflated negative binomial (ZINB) and negative binomial (NB) distributions. The fraction of nuclei with nonzero count values is given in bar chart to the left of each plot, with a histogram of the nonzero values on the right. Black dots on the bar chart and black lines in histogram indicate predicted values from ZINB (top) or NB (bottom) fit to the count data. On the left are genes where the ZINB distribution is better than the NB, on the right are genes where they are equivalent.

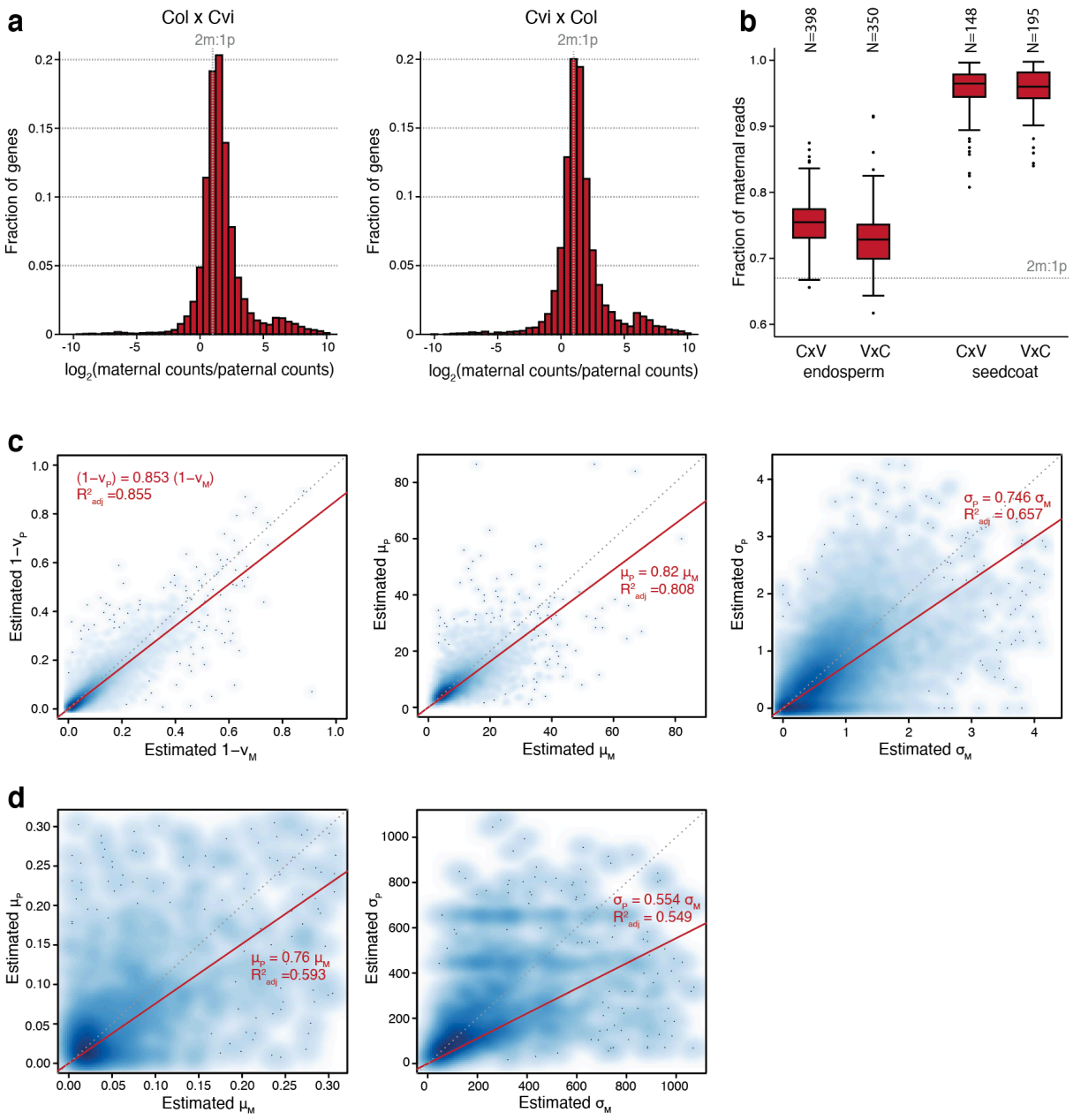


Fig. S9. Exploration of maternal:paternal ratio and parameter estimates for fitted ZINB and NB models from the single-nuclei allelic count data. (a) Histogram of the $\log_2$ ratio of maternal to paternal counts for all genes in the dataset, with the expected 2:1 ratio indicated as a grey vertical dotted line. (b) Fraction of maternal reads in CxV (Col x Cvi) and VxC (Cvi x Col) endosperm and seed coat. Maternal bias is more pronounced in CxV libraries than VxC, suggesting that the maternal bias is partially due to mapping bias in favor of the sequenced strain (Col). Number of nuclei in each category indicated at top of plot. Median, interquartile range and upper-/lower-adjacent values (1.5*IQR) indicated by center line, box, and whiskers of each boxplot, respectively. (c) Scatterplot of parameter estimates for the three ZINB parameters v, μ, and σ. Parameter estimates for maternal counts are plotted against estimates for paternal counts. Results shown from Col x Cvi. Each dot represents a single gene. (d) Scatterplot of parameter estimates for the two NB parameters μ and σ, for maternal and paternal count distributions.

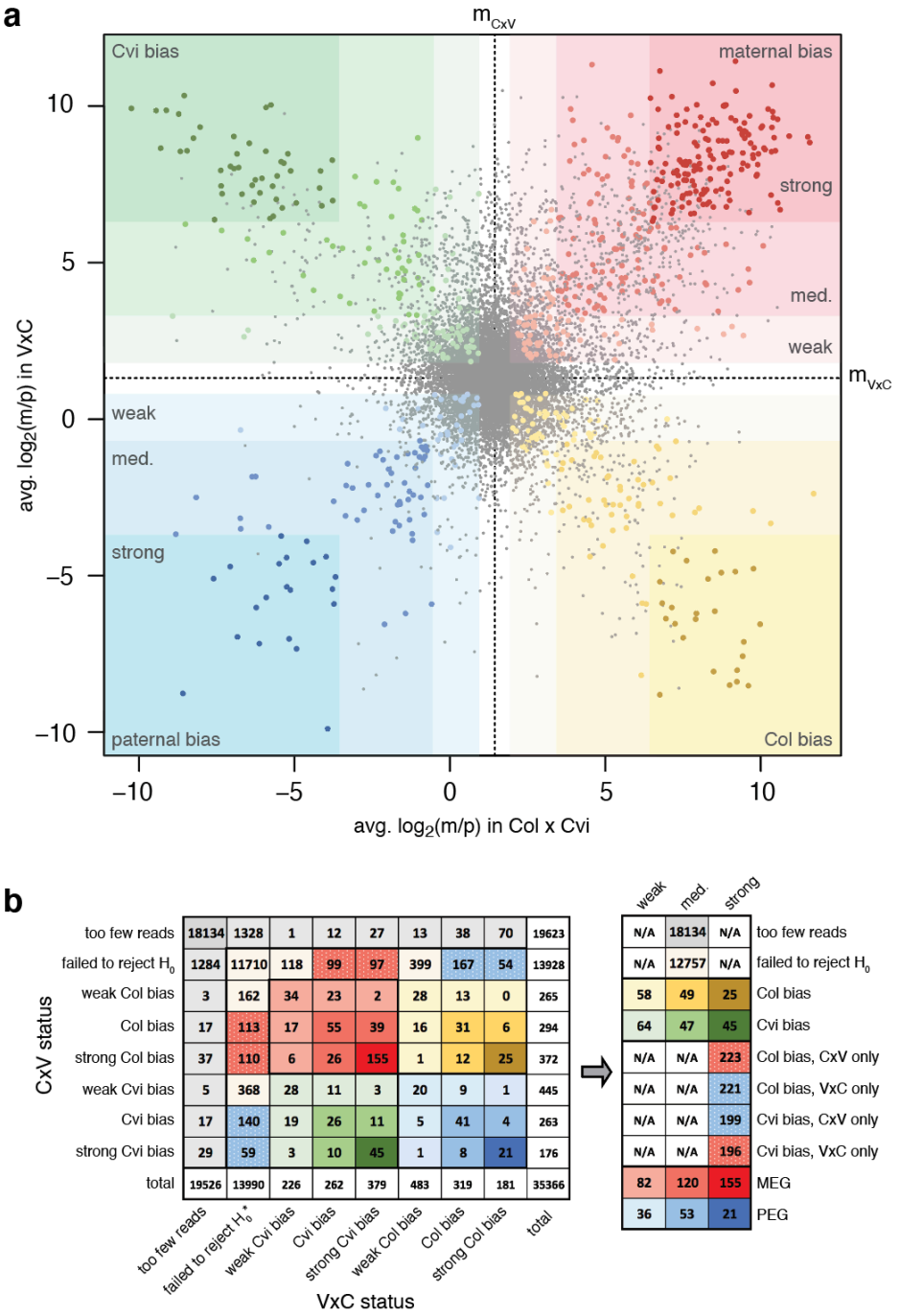


Fig. S10. Genes displaying allelic bias in reciprocal crosses of Col and Cvi. (a) Scatterplot comparing weighted average $\log_2$(m/p) across all nuclei with allelic reads, in Col x Cvi compared to Cvi x Col. Quadrants of the plot corresponding to maternal, paternal, Col, and Cvi bias indicated. Genes that were significantly biased highlighted in corresponding color, with larger point size. Average $\log_2$(m/p) across all genes in Col x Cvi (mCxV) and Cvi x Col (mVxC) indicated by dotted black lines. Quadrants are divided into three subregions: weak bias (m ± 0.5), medium bias (m ± 2) and strong bias (m ± 5). (b) Table corresponding to (a), indicating number of genes in each category. Final classifications based on both Col x Cvi and Cvi x Col results, along with combined gene counts for each category, are shown in table at right.

a

	Gene Status	# genes	% mat. CxV	% mat. VxC	log_2(m/p) CxV	log_2(m/p) VxC
MEGs	strong bias	155	99.62	99.61	8.45	8.36
	medium bias	120	97.05	96.98	5.55	5.75
	weak bias	82	89.43	88.53	3.46	3.19
PEGs	strong bias	21	3.17	2.60	-5.39	-5.79
	medium bias	53	21.89	18.58	-2.39	-2.43
	weak bias	36	47.79	43.61	-0.25	-0.49
Col bias	strong bias	25	99.59	1.76	8.29	-6.40
	medium bias	49	97.35	16.40	5.80	-2.70
	weak bias	58	88.79	46.16	3.18	-0.25
Cvi bias	strong bias	45	1.72	99.54	-6.55	8.11
	medium bias	47	18.97	96.83	-2.64	5.38
	weak bias	64	43.76	87.82	-0.60	3.10
other	CxV Col bias	223	97.40	82.48	6.23	3.29
	CxV Cvi bias	199	21.73	78.58	-2.55	2.61
	VxC Col bias	221	77.35	19.93	2.44	-2.58
	VxC Cvi bias	196	79.38	97.17	2.61	6.03
no bias		12,757	71.60	69.88	1.56	1.43

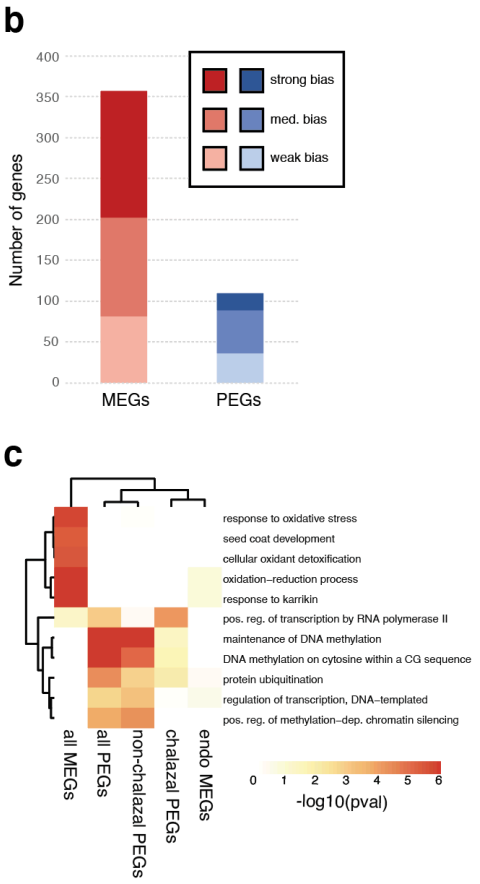


Fig. S11. Summary of allelic bias under all conditions evaluated. (a) Table of average percent maternal in CxV (Col x Cvi) and VxC (Cvi x Col) 4 DAP endosperm data, as well as average log_2(m/p) in each cross. (b) Number of MEGs and PEGs displaying strong, medium and weak bias (> 5, > 2, and > 0.5 s.d. away from median log_2(m/p) in dataset, respectively). (c) GO-term analysis of all MEGs, all PEGs, PEGs preferentially expressed in chalazal endosperm (chalazal PEGs, Fig. 3a), PEGs not preferentially expressed in chalazal endosperm (non-chalazal PEGs), and MEGs with higher expression in endosperm than in seed coat (endo MEGs). Heatmap colors according to -log_{10}(p-value) of significance of enrichment of indicated GO-term among genes of indicated category. Color scale is truncated at -log_{10}(pval) = 6.

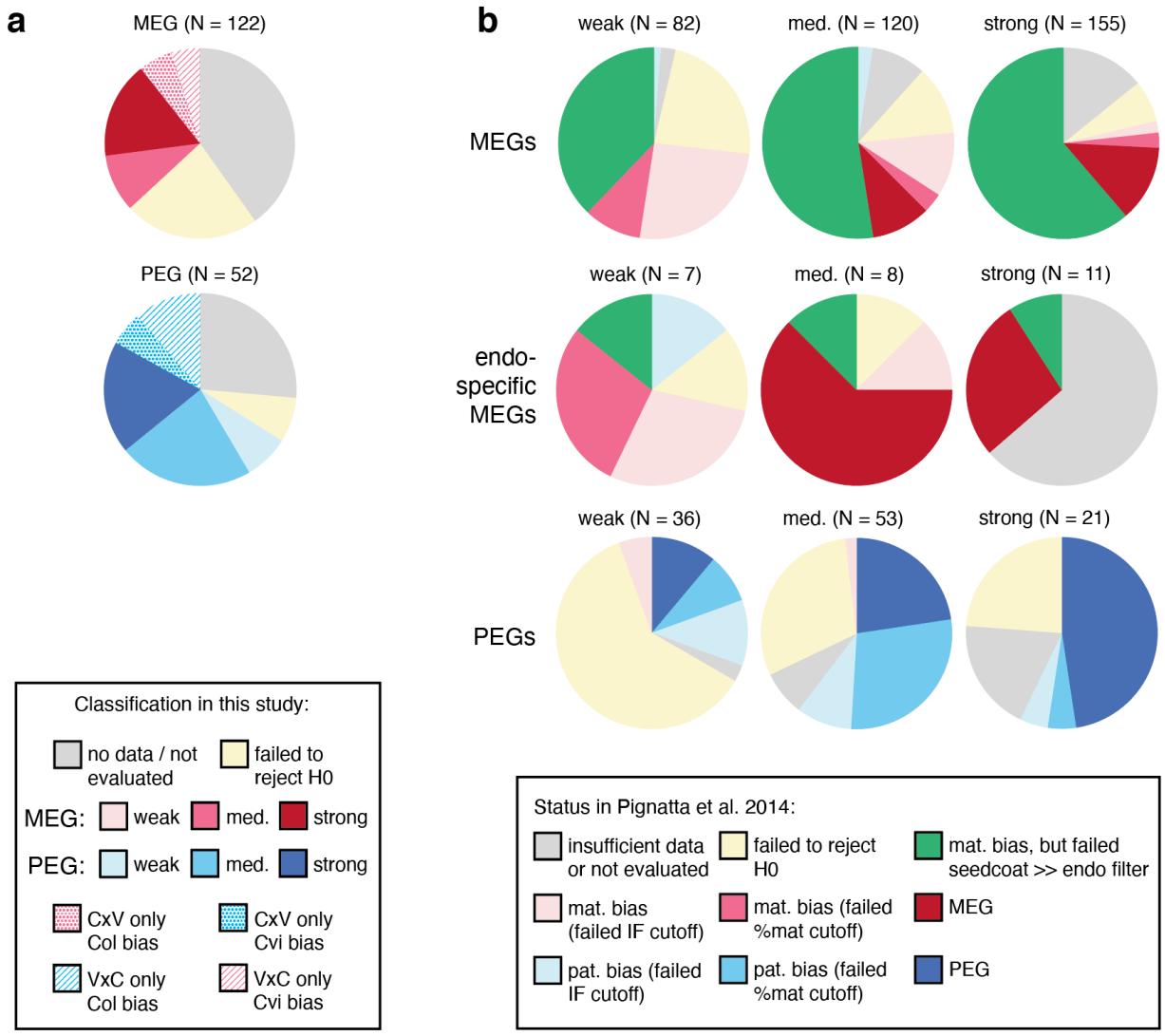


Fig. S12. Comparison of imprinting status determined by snRNA-seq and conventional whole-endosperm mRNA-seq. (a) Classification results from this study for the 122 MEGs and 52 PEGs identified between Col and Cvi 7 DAP endosperm in (*28*). Categories as shown in Fig. S10. For genes in the 'CxV only, Col bias' and similar categories, the null hypothesis could only be rejected for the indicated cross, but the other cross direction may also exhibit bias (but lack sufficient data to reject). (b) Status in (*28*) of all MEGs and PEGs identified in this study. (*28*) involved successive thresholds for assessing a gene's imprinting status: p-value < 0.01 -> IF (imprinting factor) > 2 -> % maternal either >85% (mat. bias) or <50% (pat. bias) -> imprinted. Genes that failed to meet all the imprinting criteria in (*28*) were therefore re-classified according to which cutoff they had failed. Additionally, maternally biased genes ($p < 0.01$) were subjected to an additional criterion that eliminated genes with much higher expression in seed coat vs. endosperm, to minimize potential biases that might be caused by seed coat contamination (*28*).

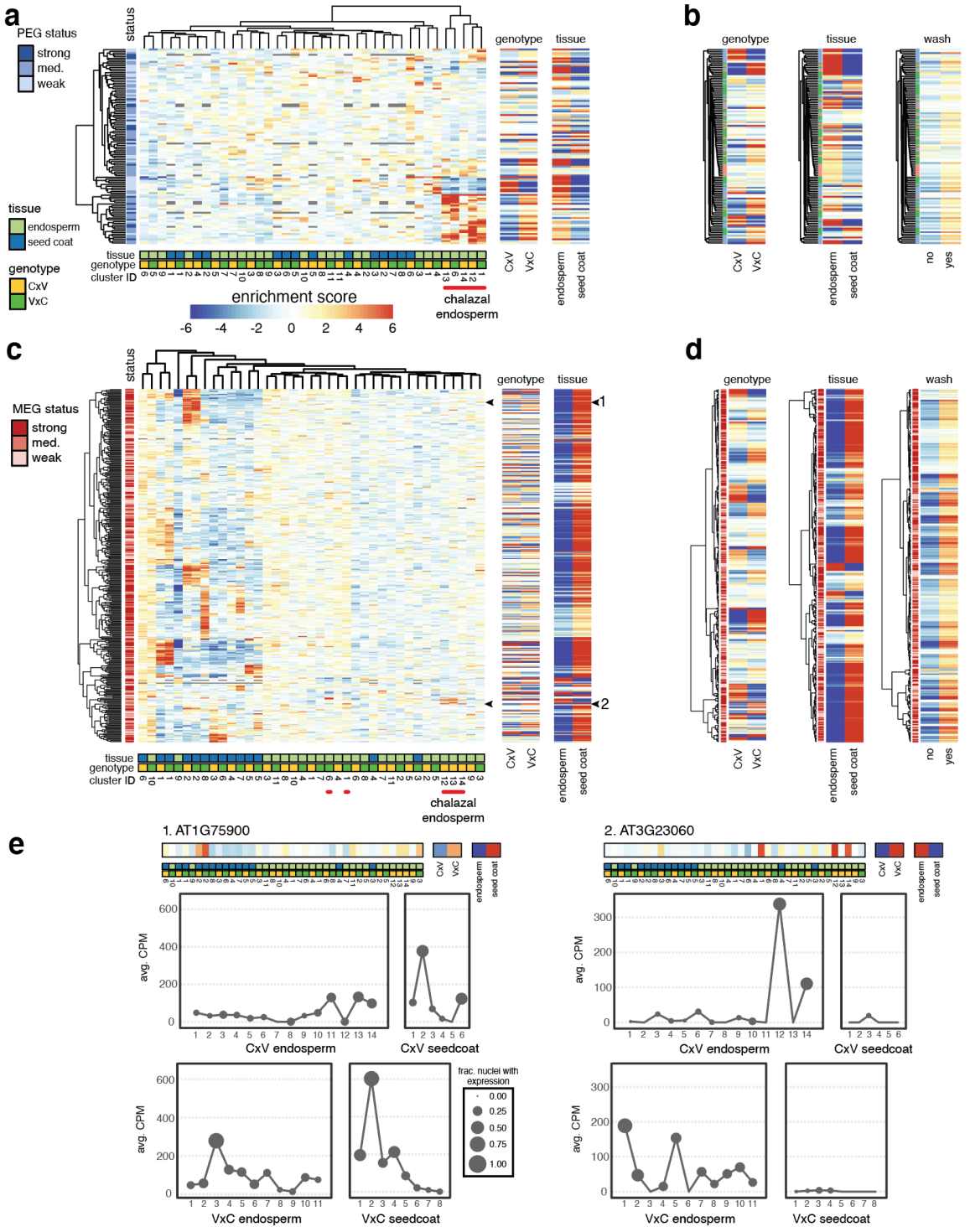


Fig. S13. Summary of MEG and PEG expression patterns across endosperm and seed coat clusters. (a,c) Hierarchical clustering of gene expression enrichment scores (ES) for all (a) PEGs and (c) MEGs, controlling for tissue and genotype. Separate heatmaps of ES calculated by genotype and by tissue on right. (b,d) Hierarchical clustering for (b) PEGs and (d) MEGs performed separately over ES from genotype, tissue and wash (yes = isolated nuclei were washed 2x prior to sorting, no = no extra wash steps). (e) Example expression profiles across nuclei clusters for two MEGs, AT1G75900 (strong MEG) and AT3G23060 (medium MEG). Locations of these genes in (c) indicated by black arrows.

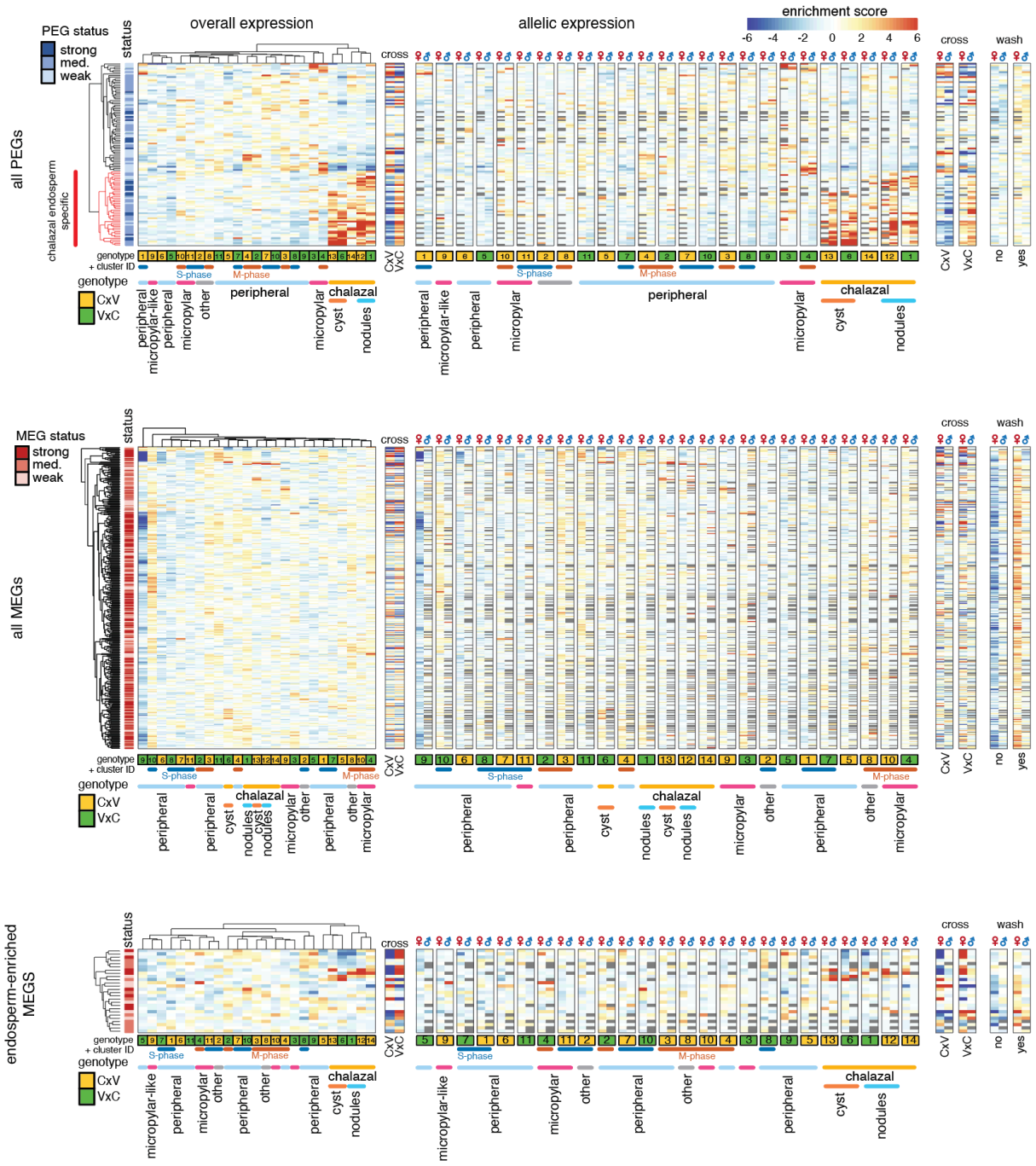


Fig. S14. Total and allelic expression patterns across the endosperm clusters, for all imprinted genes. Top panel shows all PEGs, middle all MEGs, and bottom panel shows a subset of MEGs whose expression is enriched in endosperm. Expression enrichment scores (ES) for total expression shown in heatmap on the left. On right, ES for maternal (♀) and paternal (♂) allelic expression enrichment scores are shown separately for each cluster, with row and column order otherwise identical to the leftmost heatmap. Scores were also calculated across factors cross (genotype of seed, CxV = Col x Cvi or VxC = Cvi x Col), and wash (nuclei were either washed 2x in sorting buffer prior to sorting, or not washed).

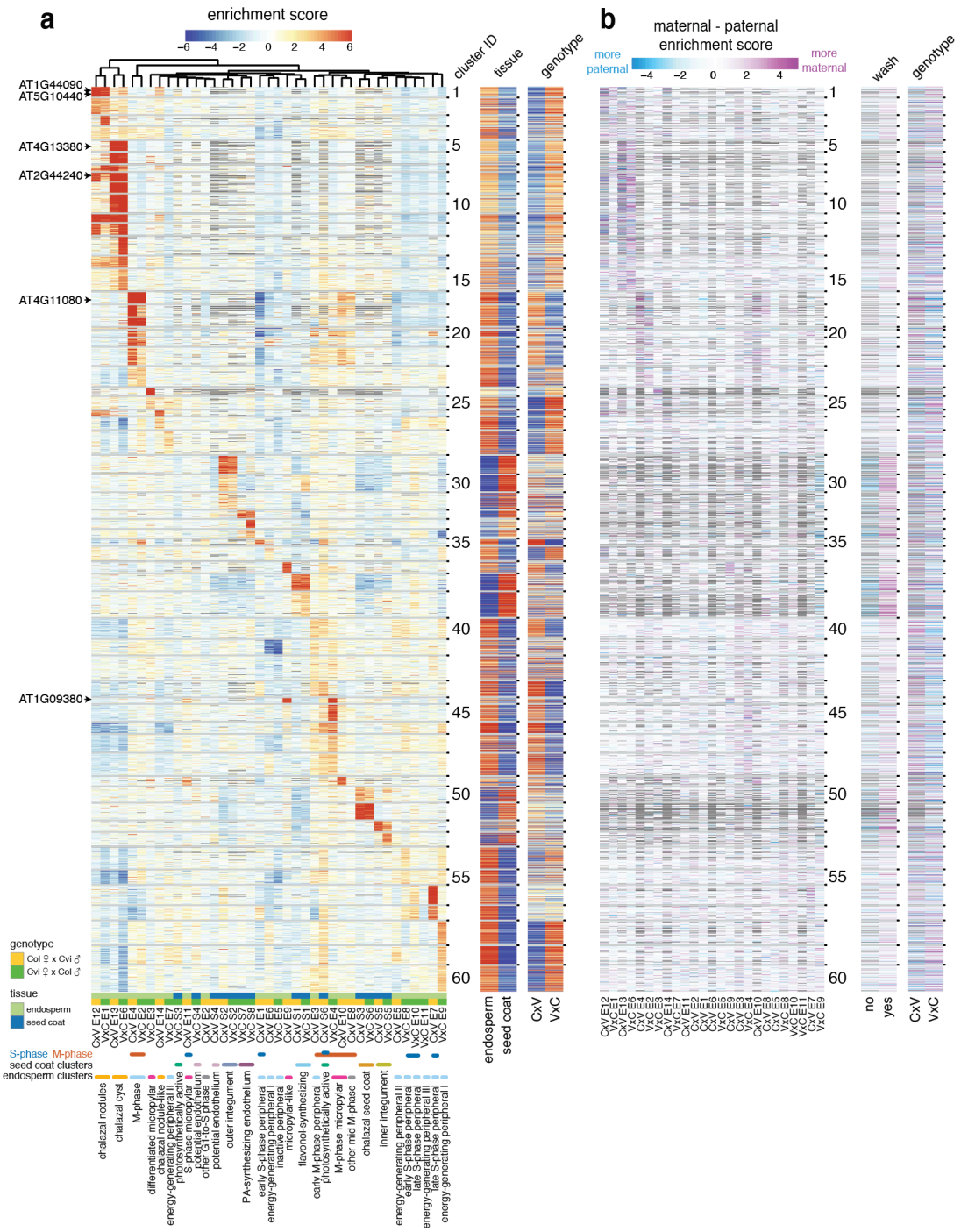


Fig. S15. Total and allelic expression patterns of 4,500 genes significantly differentially expressed among the endosperm and seed coat clusters. (a) Heatmap of gene expression enrichment scores (ES) (same as Fig. S4a). (b) Heatmap of difference between the maternal (ESMat) and paternal (ESPat) expression enrichment scores in the Col, Cvi 4 DAP endosperm dataset. Pink indicates relatively higher enrichment of maternal expression compared to paternal expression, while blue indicates the opposite. Far right: heatmaps comparing the ESMat - ESPat values as a function of sample prep (nuclei were either washed 2x in sorting buffer prior to sorting, or not washed) and genotype are also shown. Order of rows same in all plots.

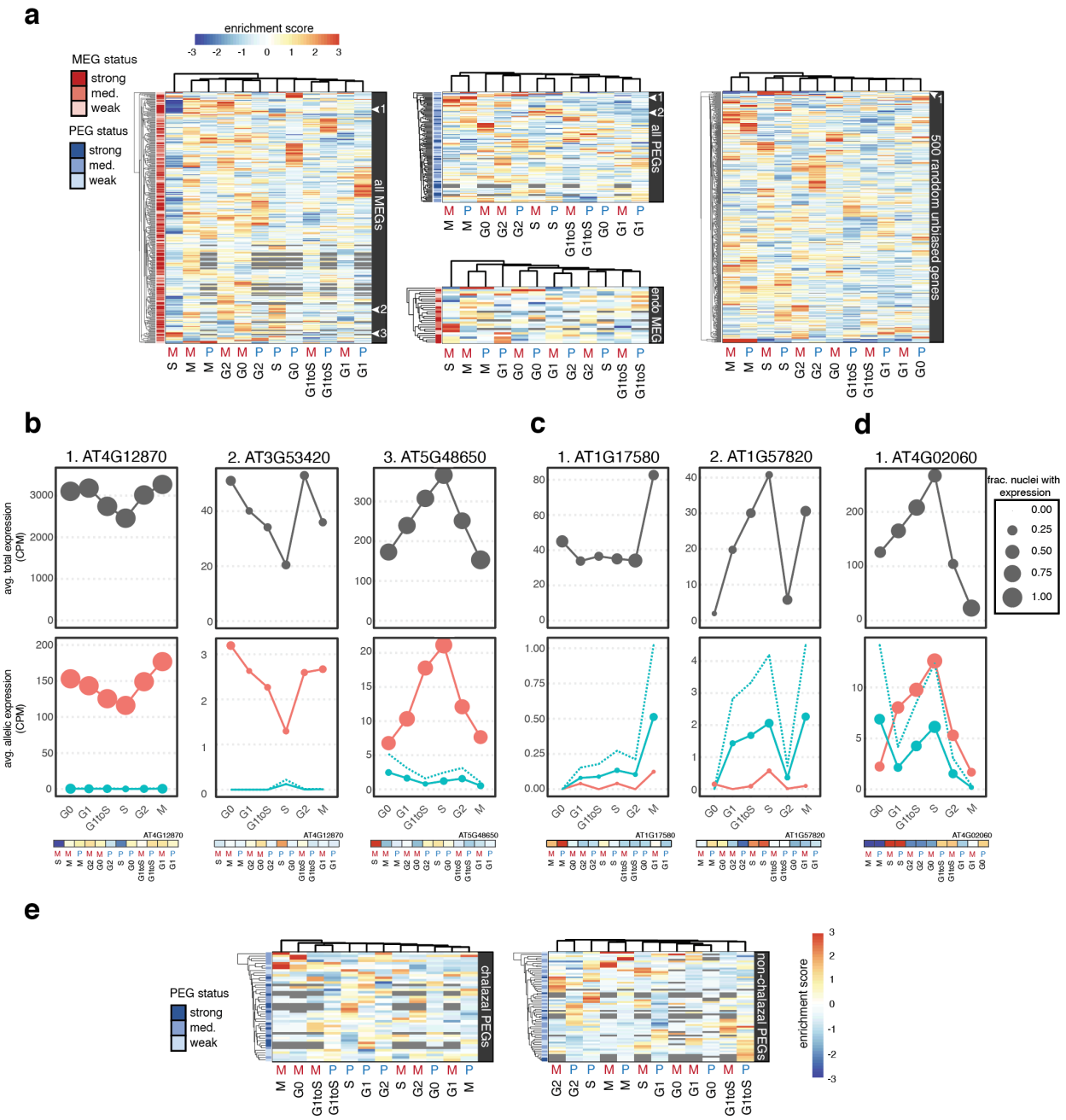


Fig. S16. Effect of cell cycle on expression of imprinted genes. (a) Enrichment scores for maternal (M) and paternal (P) allelic expression over different phases of the cell cycle, for all MEGs, all PEGs, MEGs specifically expressed in endosperm ('endo MEG'), and a random subset of 500 genes that show no overall allelic bias. Controlled for genotype and prep (extra washes). (b-d) Example total (top, grey) and allelic (bottom, red = maternal, blue = paternal, dotted blue line = simulated 2x paternal) expression patterns across cell cycle phases for (b) three example MEGs, (c) two example PEGs and (d) an example non-imprinted gene. Genes in b-d are marked with white arrows in appropriate heatmap in (a). Values from heatmaps in (a) shown at bottom of each plot. (e) Same as (a), except enrichment scores were computed by permuting cell cycle labels over nuclei in chalazal clusters only (CxV E12, 13 and 14, VxC E1, 6, N = 76 chalazal nuclei with cell cycle information). Controlled for genotype. Legend and scale same as (a). Chalazal-enriched and non-chalazal-enriched PEGs (Fig. 3a) shown.