Gene duplication and cellular divergence in crops

Kenneth Birnbaum (ken.birnbaum@nyu.edu)
New York University https://orcid.org/0000-0002-8423-6859
Bruno Guillotin
NYU https://orcid.org/0000-0002-0117-2512
Ramin Rahni
New York University, Center for Genomics and Systems Biology United States
Michael Passalacqua
Cold Spring Harbor Laboratories https://orcid.org/0000-0002-6344-1175
Mohammed Mohammed
New York University Abu Dhabi, Center for Genomics and Systems Biology
Xiaosa Xu
Cold Spring Harbor Laboratories
David Jackson
Cold Spring Harbor Laboratory https://orcid.org/0000-0002-4269-7649
Simon Groen
University of California, Riverside
Jesse Gillis
University of Toronto

Biological Sciences - Article

Keywords:

Posted Date: June 17th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1739501/v1

Gene duplication and cellular divergence in crops

Bruno Guillotin[1,2], Ramin Rahni[1], Michael Passalacqua[3], Mohammed Ateequr Mohammed[2], Xiaosa Xu[3], David Jackson[3], Simon C. Groen[4], Jesse Gillis[5], and Kenneth D. Birnbaum[1,2,*]

[1]New York University, Center for Genomics and Systems Biology

[2]New York University Abu Dhabi, Center for Genomics and Systems Biology

[3]Cold Spring Harbor Laboratories

[4]University of California, Riverside

[5]University of Toronto, Physiology Department

*Corresponding author: ken.birnbaum@nyu.edu

Abstract:

Different plant species within the grasses were parallel targets of domestication, giving rise to crops with distinct evolutionary histories and traits. Key traits that distinguish these species are mediated by specialized cell types within organs. Here, we compare the transcriptomes of all cells within roots in three grasses—*Zea mays* (maize), *Sorghum bicolor* (sorghum), and outgroup *Setaria viridis* (*Setaria*). We first show that single-cell and single-nucleus RNA-seq provide complementary readouts of cell identity, warranting a combined analysis. Comparative cellular analysis shows that the transcriptomes of some cell types diverged more rapidly than others, in part by recruiting gene modules from other cell types. Furthermore, examining the whole genome duplication in maize, we detect extensive dosage compensation in surviving co-expressed homeologs, reinforcing genomic balance[1]. Homeolog pairs that underwent subfunctionalization[2], partitioning their expression among cell types, represented a minor pattern but showed the highest rate of acquiring a novel (non-ancestral) domain. These results fit a conjecture in which mechanisms that maintain stoichiometric balance at the molecular level aid in homeolog retention for extended periods to allow new functions to arise. An unexpected synergy between spatial sub- and neo-functionalization then contributes to changes in transcriptional cell identity.

Single-cell mRNA profiling has opened up new opportunities to study cellular evolution by comparing gene regulation in specialized cells across species[3,4]. In plants, high-resolution cellular profiling also has the potential to associate cell-level transcriptional regulation to key agricultural traits, many of which are mediated by specialized cells[5].

Zea mays (maize) is a staple crop and *Sorghum bicolor* (sorghum) is an important dryland crop and biofuel candidate that is closely related to maize, separated by about 12 million years[6,7]. However, the two species differ substantially in key traits such as drought and chilling tolerance, and release of root exudates that shape soil interactions[8–12]. The importance of the two crops, their evolutionary proximity, and their functional differences present a novel opportunity for comparative analysis of cellular evolution in plants[13,14]. In addition, since sharing a common ancestor with sorghum, maize underwent a whole genome duplication 5 to 12 million years ago [7], offering an opportunity to analyze changes in fine-scale gene regulation among paralogous

genes on duplicated chromosomes (homeologs) in the relatively early stages after a duplication[7,15].

Cells Provide Depth While Nuclei Give Breadth

Single-cell analyses in plants have relied on the generation of protoplasts by enzymatic digestion of cell walls[16–21]. However, certain tissues and even some species like sorghum are quite recalcitrant to digestion. There is also historic concern about the effects of protoplast generation on the cellular transcriptome, leading to growing interest in nuclear profiling[22–24]. To assess the fidelity of nuclear profiling in depth, we first compared single-cell vs single-nucleus vs whole-root profiles in both *Arabidopsis thaliana* (Arabidopsis/At, a dicot model with plentiful resources, 15,967 cells and 17,373 nuclei) and maize (Zm, a monocot model, 4,235 cells[25] and 2,668 nuclei) (Supplementary Table 1).

Measures of unique molecular indices (UMIs) per dataset showed an increase of 10x (At) and 6x (Zm) in cells than in nuclei (Extended Data Fig. 1a), similar to animal studies[26,27]. Accordingly, the average number of genes detected was 2.7x (At) and 1.4x (Zm) higher in cells than in nuclei (5,281 vs 1,895 in At, 4,198 vs 2,304 in maize) (Extended Data Fig.1b, Supplementary Table 1). However, despite the lower mRNA content, nuclear profiling detected 89% (At) and 88% (Zm) of total genes present in cells (Supplementary Table 1).

Both cell and nuclei "pseudo-bulked" transcriptomes displayed a high correlation to whole-root transcriptomes ($r \sim 0.7$-0.8, Extended Data Fig. 1c), confirming that both sampling methods generally reflected expression patterns of intact tissue.

In both Arabidopsis and maize, cells and nuclei generated UMAP clusters corresponding to all the major cell identities (Extended Data Fig. 2, 5). However, in both species, the nuclear dataset generated fewer distinct clusters, often failing to resolve between closely related or subcellular identities (Extended Data Fig. 2, 5). For example, in maize, stele cells contained a subcluster that we identified as xylem cells, where no such subcluster was apparent in the nuclear UMAP (Extended Data Fig. 5). Using a down-sampling approach on each dataset, a general rule-of-thumb emerged that twice as many nuclei as cells are needed to discover the same number of cell-type clusters as protoplasts (Extended Data Fig. 3a,b). Thus, the shallower depth of nuclear profiles provides less resolution for classification of cell identity—a drawback that down-sampling showed we could rectify, at least in part, by increasing the number of nuclei sampled compared to cells.

Either simultaneous or independent analysis of cells and nuclei generated clusters that reflected the same underlying biological patterns (Fig. 1a-c, Extended Data Fig. 3c,d). The highest-scoring markers extracted from nuclei generally matched the highest-scoring ones from cells (Fig. 1c,e, Extended Data Fig. 3d). In addition, the assignment of cells to specific clusters was stable when cells or nuclei were clustered either alone or together (Supplementary Table 2).

One advantage of nuclear profiles was their ability to capture a more representative sampling of cells within the tissue (Fig. 1d, Extended Data Fig. 4d). In one example in maize, we detected a unique cluster in single-nucleus profiling not present in single-cell/protoplast profiling, which we confirmed as columella cells using previously published hand-sectioned RNA-seq profiles (Extended Data Fig. 5,[25]).

In Arabidopsis, we found that 14% of total genes (3,218) were differentially expressed between cells and nuclei in a cluster-by-cluster analysis (Supplementary Table 3). While cells showed a

higher proportion of stress related genes (Fig. 1f, Extended Data Fig. 4a,b), most of the differences between cell and nuclear profiling appeared to be related to compartmental RNA stability, as mRNAs enriched in nuclei significantly overlapped with transcripts shown to have higher decay rates in the cytoplasm[28,29] (Extended Data Fig. 4c). This analysis showed that nuclear profiles did indeed show a lower stress response than protoplasts, but the difference was subtle and not a dominant component of the difference between cells and nuclei.

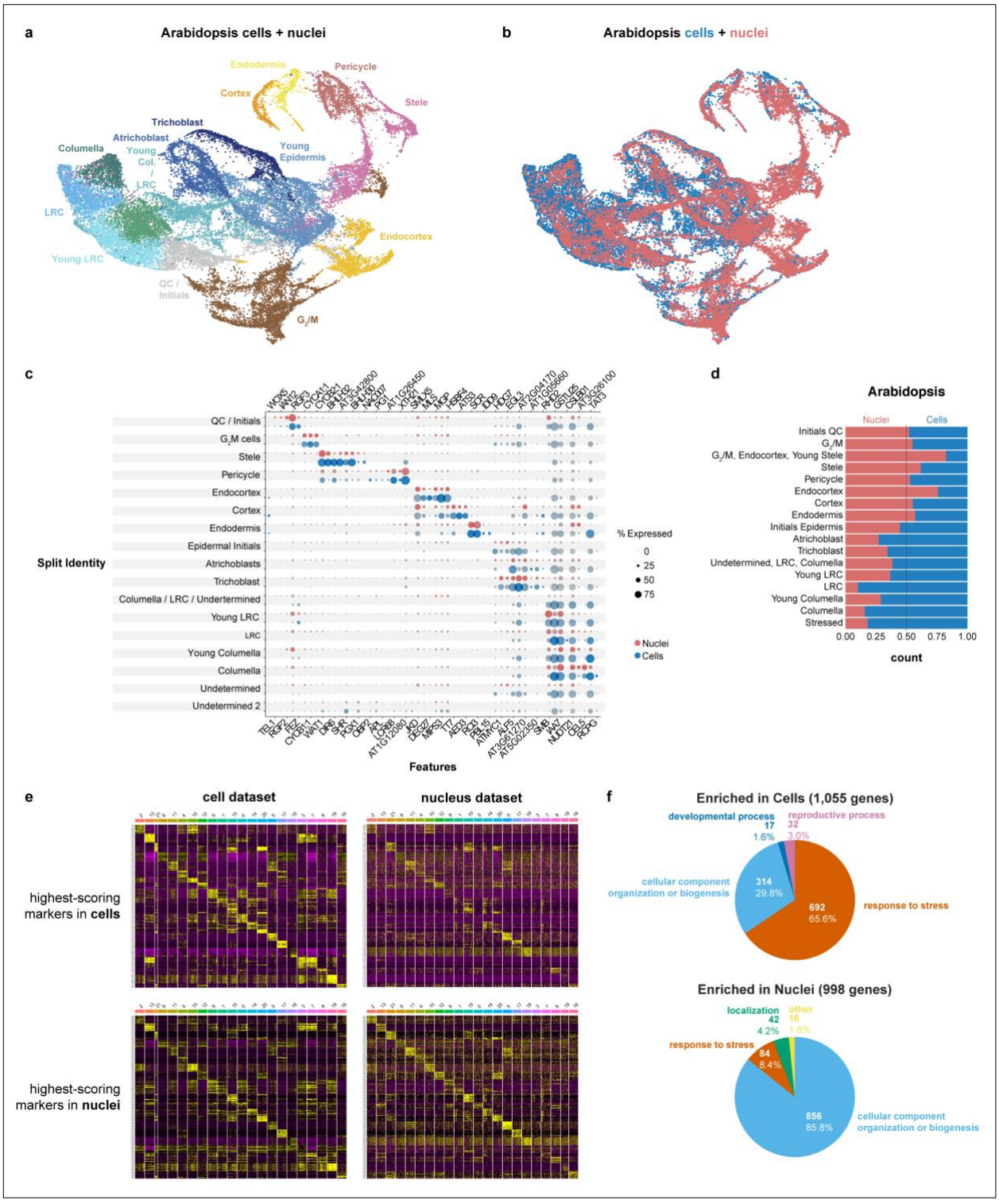


Fig. 1: Cell and nucleus profiles identify the same markers but show different sensitivities and artifacts. a, b UMAP of combined Arabidopsis cells and nuclei with clusters colored according to assigned cell identity (**a**) or cells vs. nuclei origin (**b**). **c** Dot plots of Arabidopsis marker genes in cells (blue) or in nuclei (red). **d** Proportion of cells vs nuclei present in each cell type cluster. **e** Heatmaps of the 10 highest-scoring marker genes for each cell type found using Seurat in the combined single-cell + single-nucleus dataset. This combined dataset was then divided into a cell dataset and a nucleus dataset. Within each of the four panels, rows correspond to genes, and columns correspond to cell-type clusters. Upper two heatmaps show highest-scoring markers within the cell dataset (left) vs expression of the same markers in the nucleus dataset (right). Lower row shows highest-scoring markers found in nucleus dataset (right) vs expression of the same markers in the cell dataset (left). **f** Pie charts showing the difference in the prevalence of GO terms among differentially expressed genes between cells (top) and nuclei (bottom), aggregated from cluster-by-cluster differences. GO term analysis is for up-regulated genes in either cells or nuclei.

A High-confidence cell-type Map in sorghum and *Setaria* Using a maize Reference.

Given the comprehensive coverage of a combined analysis, we pursued both whole cell and nuclear profiling to investigate cellular evolution in the maize-sorghum clade. We further generated profiles for sorghum (3,510 cells and 7,620 nuclei) and, as an outgroup, *Setaria viridis* (*Setaria*, 974 cells and 12,192 nuclei, Supplementary Table 1). We took advantage of comparative genomic sequence analyses in maize, sorghum, and *Setaria* that mapped orthologs among the three species, including the homeologs created by whole genome duplication in maize[13,15] (hereafter subgenome M1 and M2). We first used a set of single-copy orthologs in the three species to cluster the maize cells and nuclei and then mapped the sorghum and *Setaria* datasets onto the maize anchor[30] (Fig. 2a, Supplementary Table 1). To validate the mapping, we performed an independent MetaNeighbor analysis (Extended Data Fig. 6) as well as whole mount *in situ* hybridizations in maize and sorghum (Fig. 2b, Extended Data Fig. 7,8), and spatial transcriptomics in maize (Fig. 2c, Extended Data Fig. 7), confirming the maize-to-sorghum-to-*Setaria* mapping of cell identities. Thus, we could use the well-annotated maize cell type map to rapidly generate a high confidence single-cell “pan-transcriptome” of these key crop species, including hundreds of new cell-type specific marker genes (Supplementary Table 4).

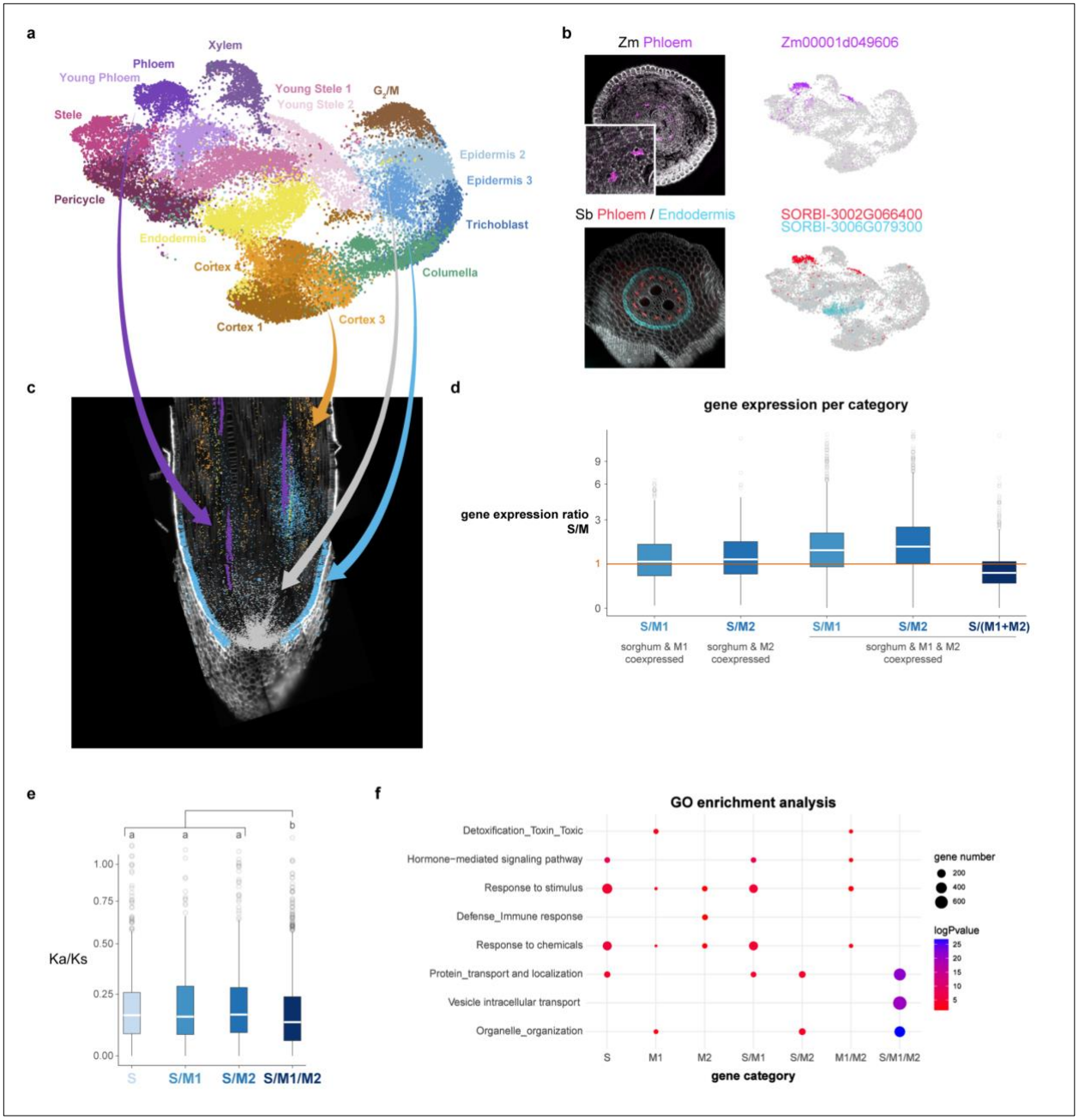


Fig. 2: Mapping cell identities from maize to sorghum reveals global dosage compensation between maize duplicates in cell types. a UMAP of combined maize single-cell and single-nucleus profiles. Clusters are colored and labeled according to cell identity. **b** *in situ* hybridization in maize (top) and sorghum (bottom). The maize phloem marker in magenta is orthologous to the sorghum phloem marker in red. Cyan in the lower panel corresponds to a sorghum endodermal marker. Autofluorescence highlighting anatomy is in grayscale. The minimum/maximum values for each channel in the fluorescence images have been adjusted to show the localization more clearly in the merged image. UMAPS next to images show the respective expression in single-cell / single-nucleus profiles, which were used initially to determine their expression pattern. **c** Spatial transcriptomics showing simultaneous localization of multiple markers that enable detailed mapping of single-cell profiles to specific tissues. **d** Expression ratios of sorghum over maize

orthologous genes, where the red line indicates equivalent expression levels. The first two boxplots represent cases where a sorghum ortholog is co-expressed with a single maize homeolog (either M1 or M2). The third and fourth boxplots represent cases in which both homeologs are expressed in the same cells. The last boxplot shows the ratio when both of the co-expressed homeologs are summed in the denominator. White bar is median, box limits represent 25th and 75th percentiles, lines are the 95th percentile. **e** Ka/Ks distribution of comparative patterns measured against sorghum. Cell type specific expression is averaged. Letters signify comparative patterns of cell type expression for sorghum and M1 and M2 subgenome expression: i.e. S = only the sorghum ortholog is present in the cell type cluster, SM1 = sorghum and maize 1 homeologs are co-expressed in the cell type cluster, etc. Statistical analysis was performed using ANOVA followed by the Tukey pairwise test. All "a" designates are statistically significant from "b" at $p < 0.05$. **f** GO terms enriched within each category of genes in each cell types.

Widespread Dosage Compensation between maize homeologs

The comparative cell-type maps provided an opportunity to quantify patterns in preserved gene duplicates – a potential source of innovation particularly following whole genome duplication. One hypothesis, known as the genomic balance model, holds that the viability of a duplicated gene is dependent on the capacity to adjust for dosage effects, such that the stoichiometry of molecular complexes is retained following a whole genome duplication[1].

However, we could observe changes in expression patterns in the 5 to 12 million years since whole genome duplication that were likely to disrupt the stoichiometric balance. Prior studies have demonstrated dosage compensation between gene duplicates in aneuploids and in newly generated polyploids[1,31]. However, it is not clear if this phenomenon has a role after ancient genome duplication events. The cellular resolution of our dataset offered an opportunity to test whether dosage compensation within cells could fine tune genomic balance.

To examine this issue, we adapted a ranked-based method to compare gene expression levels across genomes in each cell type[32]. This enabled us to use sorghum gene expression in each cell type as proxies for the "ancestral" state (expression level and cell-type domains). In addition, ancestral expression domains were supported by at least one maize homeolog. We also made the assumption that both homeologs had identical expression patterns at the time of whole genome duplication[33]. We categorized a homeolog as "dominant" if its average expression level per cell type is more than twice the expression of the other homeolog, and as "co-expressed" if both homeologs are detected and neither is dominant.

When one of the M1 or M2 homeologs was dominant in a cell type, this dominant homeolog was expressed at the same level as its sorghum ortholog (Fig. 2d). Interestingly, these dominant homeologs were enriched in GO terms for stress adaptation, immunity and response to stimulus (Fig. 2f, Supplementary Table 5). At the same time, the dominant homeologs display a higher cell-type specificity than co-expressed homeolog pairs[34] (τ), while the non-dominant ones show a higher nonsynonymous-to-synonymous substitution rate (Ka/Ks) suggesting a more relaxed purifying selection (Extended Data Fig. 9a-c).

Alternatively, when maize M1 and M2 homeolog pairs were co-expressed in the same cell types as their sorghum ortholog, each showed an expression level of about half that of the sorghum ortholog (Fig. 2d, Extended Data Fig. 9d,e). These co-expressed homeologs further displayed a marginally lower Ka/Ks ratio than dominant cases, suggesting that they are under more stringent purifying selection (Fig. 2e, Extended Data Fig. 9a). Finally, these homeologs showed GO term

enrichments for categories known to be favored for retention after whole genome duplication—cell homeostasis processes, translation, or ribosome biosynthesis[1,15] (Fig. 2f, Supplementary Table 5). The analysis therefore showed strikingly widespread dosage compensation in cell types where both homeologs are co-expressed likely fine tuning stoichiometric balance over long periods on a cell-type basis[35].

Neofunctionalization and subfunctionalization are linked

It has been suggested that mechanisms permitting stoichiometric balance could act as a bridge for the generation of new gene function, promoting the retention of gene duplicates to allow sufficient time for beneficial or complementary mutations to arise[1]. Two possible mechanisms for long-term duplicate gene retention are subfunctionalization and neofunctionalization[2], both of which we consider here at the transcriptional level. Thus, we define subfunctionalization here as a case in which duplicated genes partition the expression of the ancestral[2] (sorghum) domain (Fig.3a). Neofunctionalization is presumed to occur when one homolog retains ancestral and the second is free to diverge and adopt a new function[36]. Thus, at the transcriptional level, we define neofunctionalization as the dominance of one homolog in the ancestral domain and the expression of the second homolog in a cell type outside the ancestral domain (Fig.3c). We classified each homeolog pair across cell types on a scale from –1 (full dominance of one homeolog over the other), to 0 (full co-expression of both homeologs) to 1 (full subfunctionalization; Fig. 3a,b). Overall, 70% of the homeolog pairs showed dominance, 11% showed full co-expression, and 19% showed subfunctionalization (Fig. 3a,b, Supplementary Table 6).

To assess potential evolutionary forces driving each pattern, we assumed both homeologs matched the sorghum ortholog expression pattern at the time of duplication. We then randomly removed gene expression of either homeolog across cell types until their matrix of gene expression matched the overall presence/absence matrix of homeologs in the observed data.

The analysis revealed that dominance patterns were highly over represented in the observed data compared to the random model (Fig. 3b), with full dominance being the most abundant category. In instances of dominance, the non-dominant homeolog (in either subgenome) showed slightly relaxed purifying selection on average[33] (Extended Data Fig. 9b). There were many instances of domain expansion/neofunctionalization in the dominance categories (18%; Fig. 3d). However, counter to expectations, 80% of the time it was the dominant homeolog that expanded its domain (Supplementary Table 6). Dominance patterns were largely stable when we examined single cell profiles from the maize inflorescence[37], with 72% of the root dominance patterns falling into the same category in the inflorescence cells, swapping the dominant homeolog in a minority of cases, while 26% display a subfunctionalization pattern and 2% co-express (Supplementary Table 6). These observations suggest that the non-dominant homeolog could have a more essential role in at least one context in the plant or may possibly be in the early stages of pseudogenization.

While subfunctionalized homeolog patterns were highly underrepresented compared to the random model (Fig. 3b), homeolog pairs in this category showed a significantly higher proportion of neofunctionalization compared to dominance categories (33% vs. 18%, Fisher's exact test, p-value < 0.001; Fig. 3d, Supplementary Table 6). This was unexpected as it suggests that subfunctionalization, rather than dominance, was more likely to be accompanied by neofunctionalization. In addition, in contrast to the dominance categories above, there was no evidence of relaxed purifying selection in either the homeolog that extended or the one that retained the ancestral domain (Extended Data Fig. 10a). The result indicates that transcriptional neofunctionalization and subfunctionalization are somehow linked. We cannot determine which

precedes the other, although our model suggests that completely random forces acting on both homeologs could account for the observed subfunctionalization, which could then serve as a transition state to neofunctionalization[38].

Overall, we propose a model in which dosage compensation maintains stoichiometric balance for an extended period after whole genome duplication. This could be considered a form of quantitative subfunctionalization in which selection for full, ancestral dosage preserves the activity of both homeologs[2]. In any case, such a mechanism would prevent pseudogenization and permit the retention of gene duplicates. The prolonged survival of both homeologs would thus enable spatial subfunctionalization that then enhances the likelihood that homeologs expand their expression domains (spatial neofunctionalization)—presumably through changes in cis-regulatory regions[39]—leading to new transcriptional states in specialized cells.

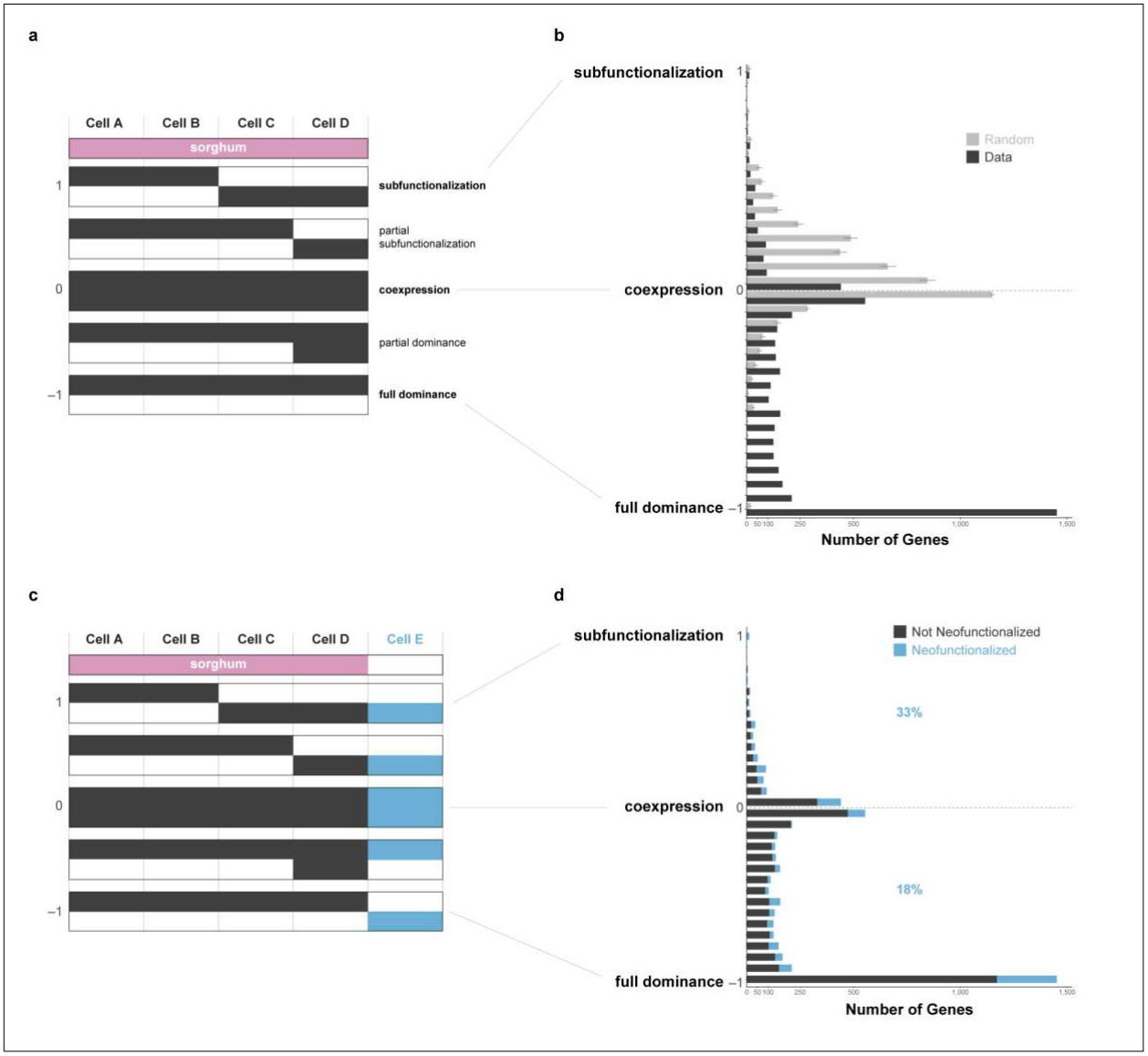


Fig. 3: Gene pairs that subfunctionalize undergo a high rate of neofunctionalization. a Conceptual schematic of patterns that characterize different categories of subfunctionalization vs.

dominance. **b** Observed distribution of genes pairs by their dominance (all categories <0) vs. subfunctionalization (all categories >0) score, where a score of 1 reflects complete dominance and -1 equal partitioning of the ancestral expression domain (subfunctionalization). Complete co-expression = 0. Dark bars represent the observed data while light gray bars represent the random model where expression domains of homeologs are randomly removed (see Methods), showing an over-representation of dominance patterns and an under-representation of subfunctionalization patterns. **c** Same schematic as in (a) but now showing novel expression domains (neofunctionalization) in blue. **d** The same distribution as shown in **(b)** with neofunctionalized events (domain expansion) mapped onto the distribution. Domain expansion is more frequent in subfunctionalized homeolog pairs, while dominance patterns show less frequent domain expansion with the vast majority being the dominant homeolog.

Root Cap "Slime" Drives a Case of Rapid Cell-Type Divergence

Using a three-taxa comparative approach, we next asked which cell types diverged most rapidly in maize and sorghum compared to the outgroup *Setaria*. To compare cell identity across species, we adapted MetaNeighbor, which uses neighbor voting to quantify the similarity of cell clusters across datasets using a given marker set of genes and their orthologs[40] (Fig. 4a).

The analysis showed that transcriptomes of columella, phloem, cortex subcluster 3, endodermis, pericycle, and stele cell types are the most divergent compared to *Setaria*, suggesting that the function of these tissues diverged from *Setaria* before the maize-sorghum split. In addition, certain cell types—such as cortex subcluster 1, phloem, and young stele subcluster 1—were statistically different between maize and sorghum, implying additional divergence after the maize-sorghum split. Interestingly, in maize, columella was among the most divergent cell types relative to *Setaria* (Fig. 4a).

To further investigate the potential functions involved in columella divergence, we used a measure of co-expression conservation to identify transcripts within clusters of interest that showed divergent patterns of expression across species in co-expression networks[35,41] (Supplementary Table 7). We identified 443 genes displaying high expression divergence across species in Columella. Many of these genes showed dramatic expression changes in spatial domain between species, such as downy mildew resistant 6 (DMR6), which is expressed in columella and epidermis in maize but in cortex and endodermis in sorghum (Extended Data Fig. 10b,c). Furthermore, GO term analysis showed enrichment in enzymes leading to the synthesis of mannose, raffinose, and oligosaccharides (Supplementary Table 7). These sugars and carbohydrates are key components of mucilage, also called slime, whose roles include shaping of the root-associated microbiome and lubricating the root-soil interface[8,10,42–44].

To further examine the potential change in mucilage-associated genes, we examined all genes implicated in mucilage component synthesis[10,11,45]. We found that this set of mucilage-annotated genes were mostly expressed in maize columella, while, in sorghum and *Setaria*, they were predominantly expressed in cortical layers (Fig. 4b,c,d). Overall, these results suggest that maize underwent a relatively rapid cellular divergence in columella, in part, by recruiting a mucilage gene expression module from ancestral expression pattern in the cortex. The most parsimonious model is that the recruitment of the mucilage module occurred before the maize whole genome duplication, as both maize homeologs, when conserved, tended to share expression in the columella.

Overall, the high-resolution comparative analysis provides evidence for three phenomena—dosage compensation, subfunctionalization, and neofunctionalization—as a driver of cell type

divergence following whole genome duplication. The results extend the genomic balance model[35] by showing its effects are reinforced after duplication at the transcriptional level. Furthermore, the data show that both single-cell and single-nucleus profiling produce high-fidelity maps of cell-type specific expression, with a combination of the two leveraging their complementary strengths. Cell type specific maps in a well-annotated species can then serve as an anchor for neighboring taxonomic groups, rapidly generating homologous maps in a related group of plant species. These maps provide powerful tools for analysis of cell type evolution that can identify genes associated with specific traits.

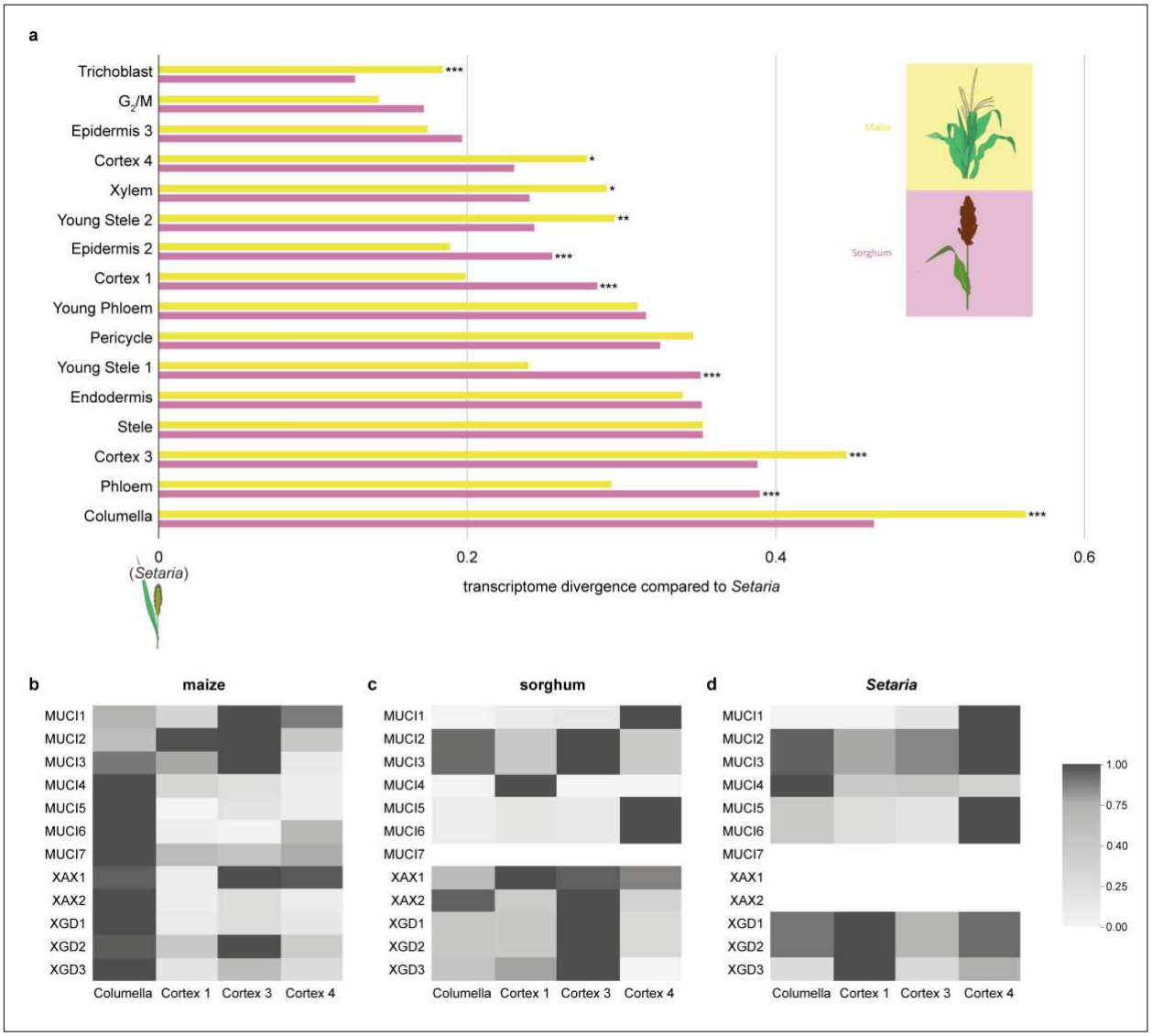


Fig. 4: Differential evolution between cell types reveals that columella is highly divergent compared to *Setaria*. a MetaNeighbor analysis showing transcriptome conservation scores between cell types in maize and sorghum compared to the outgroup *Setaria*. High conservation scores reflect similar transcriptomes across species. Statistical significance between maize and sorghum was performed using the Hanley McNeil test (see Methods; p-values: *<0.05,**<0.01,***<0.001). **b–d** Mucilage gene expression heatmaps for maize (b), sorghum (c), and *Setaria* (d) columella and cortical layers, showing predominant columella expression in maize

compared to cortical layers in sorghum and *Setaria*. Maize genes are identified in (b) and their corresponding sorghum and *Setaria* orthologs are listed on the same row in (c) and (d).

Acknowledgements

We thank Michael Purugganan and Gloria Coruzzi for helpful comments. This work was funded by National Science Foundation (IOS-1934388) to K.D.B., D.J, and J.G., the National Institutes of Health (R35GM136362) to K.D.B., and Human Frontiers of Science (LT000972/2018-L) to B.G., startup funds from the University of California Riverside to S.C.G. In addition, M.P. is funded by the William Randolph Hearst Scholarship from the School of Biological Sciences. J.G. is also supported by the National Institutes of Health (R01 LM012736 and R01 MH113005).

Contributions

B.G. and K.D.B designed the research. B.G. generated all single-cell and single-nucleus RNA-seq data. M.A.M. and B.G. designed the single-nucleus RNA-seq protocol. R.R. and B.G. performed the whole mount in-situ hybridization analysis. X.X. and D.J. performed the tissue preparation and histology for the spatial transcriptomics analysis. S.C.G. and B.G. conceived the analysis strategy and performed the tests for dosage compensation. M.P and, J.G. performed the MetaNeighbor and CoCoCoNet analysis. B.G. analyzed all the data. K.D.B., B.G., and R.R. wrote the manuscript.

The authors declare no competing interests.

Supplementary Information is available for this paper.

Material requests should be addressed to K.D.B.

References

1. Birchler, J. A. & Veitia, R. A. Gene balance hypothesis: Connecting issues of dosage sensitivity across biological disciplines. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 14746–14753 (2012).

2. Force, A. *et al.* Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531–1545 (1999).

3. Marioni, J. C. & Arendt, D. How Single-Cell Genomics Is Changing Evolutionary and Developmental Biology. *Annu. Rev. Cell Dev. Biol.* **33**, 537–553 (2017).

4. Shafer, M. E. R. Cross-Species Analysis of Single-Cell Transcriptomic Data. *Front. Cell Dev. Biol.* **7**, 175 (2019).

5. Kajala, K. *et al.* Innovation, conservation, and repurposing of gene function in root cell type development. *Cell* **184**, 3333-3348.e19 (2021).

6. Swigonova, Z. *et al.* On the tetraploid origin of the maize genome. *Comp. Funct. Genomics* **5**, 281–284 (2004).

7. Swigonova, Z. Close Split of Sorghum and Maize Genome Progenitors. *Genome Res.* **14**, 1916–1923 (2004).

8. Van Deynze, A. *et al.* Nitrogen fixation in a landrace of maize is supported by a mucilage-associated diazotrophic microbiota. *PLoS Biol.* **16**, 1–21 (2018).

9. Hasan, S. A., Rabei, S. H., Nada, R. M. & Abogadallah, G. M. Water use efficiency in the drought-stressed sorghum and maize in relation to expression of aquaporin genes. *Biol. Plant.* **61**, 127–137 (2017).

10. Kozlova, L. V., Nazipova, A. R., Gorshkov, O. V., Petrova, A. A. & Gorshkova, T. A. Elongating maize root: zone-specific combinations of polysaccharides from type I and type II primary cell walls. *Sci. Rep.* **10**, 1–20 (2020).

11. Ma, W. *et al.* The mucilage proteome of maize (Zea mays L.) primary roots. *J. Proteome Res.* **9**, 2968–2976 (2010).

12. Schittenhelm, S. & Schroetter, S. Comparison of Drought Tolerance of Maize, Sweet Sorghum and Sorghum-Sudangrass Hybrids. *J. Agron. Crop Sci.* **200**, 46–53 (2014).

13. Zhang, Y. *et al.* Differentially regulated orthologs in sorghum and the subgenomes of maize. *Plant Cell* **29**, 1938–1951 (2017).

14. Zheng, Z. *et al.* Shared Genetic Control of Root System Architecture between Zea mays and Sorghum bicolor1[OPEN]. *Plant Physiol.* **182**, 977–991 (2020).

15. Schnable, J. C., Springer, N. M. & Freeling, M. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 4069–4074 (2011).

16. Denyer, T. *et al.* Spatiotemporal Developmental Trajectories in the Arabidopsis Root Revealed Using High-Throughput Single-Cell RNA Sequencing. *Dev. Cell* **48**, 840-852.e5 (2019).

17. Efroni, I. *et al.* Root Regeneration Triggers an Embryo-like Sequence Guided by Hormonal Interactions. *Cell* **165**, 1721–1733 (2016).

18. Jean-Baptiste, K. *et al.* Dynamics of Gene Expression in Single Root Cells of Arabidopsis thaliana . *Plant Cell* **31**, 993–1011 (2019).

19. Ryu, K. H., Huang, L., Kang, H. M. & Schiefelbein, J. Single-Cell RNA Sequencing Resolves Molecular Relationships Among Individual Plant Cells. *Plant Physiol* **179**, 1444–1456 (2019).

20. Shulse, C. N. *et al.* High-Throughput Single-Cell Transcriptome Profiling of Plant Cell Types. *Cell Rep.* **27**, 2241-2247.e4 (2019).

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

22. Farmer, A., Thibivilliers, S., Ryu, K. H., Schiefelbein, J. & Libault, M. Single-nucleus RNA and ATAC sequencing reveals the impact of chromatin accessibility on gene expression in Arabidopsis roots at the single-cell level. *Mol. Plant* **14**, 372–383 (2021).

23. Long, Y. *et al.* FlsnRNA-seq: protoplasting-free full-length single-nucleus RNA profiling in plants. *Genome Biol.* **22**, 66 (2021).

24. Marand, A. P., Chen, Z., Gallavotti, A. & Schmitz, R. J. A cis-regulatory atlas in maize at single-cell resolution. *Cell* **184**, 3041-3055.e21 (2021).

25. Ortiz-Ramírez, C. *et al.* Ground tissue circuitry regulates organ complexity in maize and Setaria. *Science (80-.).* **374**, 1247–1252 (2021).

26. Allen, A. M. *et al.* A single-cell transcriptomic atlas of the adult Drosophila ventral nerve cord. *Elife* **9**, (2020).

27. Ding, J. *et al.* Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat. Biotechnol.* **38**, 737–746 (2020).

28. Narsai, R. *et al.* Genome-Wide Analysis of mRNA Decay Rates and Their Determinants in Arabidopsis thaliana. *Plant Cell Online* **19**, 3418–3436 (2007).

29. Sorenson, R. S., Deshotel, M. J., Johnson, K., Adler, F. R. & Sieburth, L. E. Arabidopsis mRNA decay landscape arises from specialized RNA decay substrates, decapping-mediated feedback, and redundancy. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E1485–E1494 (2018).

30. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *bioRxiv* 1–15 (2019). doi:10.1101/576827

31. Muyle, A., Marais, G. A. B., Bačovský, V., Hobza, R. & Lenormand, T. Dosage compensation evolution in plants: theories, controversies and mechanisms. *Philos. Trans. R. Soc. B Biol. Sci.* **377**, (2022).

32. Coate, J. E., Farmer, A. D., Schiefelbein, J. W. & Doyle, J. J. Expression Partitioning of Duplicate Genes at Single Cell Resolution in Arabidopsis Roots. *Front. Genet.* **11**, (2020).

33. Hughes, T. E., Langdale, J. A. & Kelly, S. The impact of widespread regulatory neofunctionalization on homeolog gene evolution following whole-genome duplication in maize. *Genome Res.* **24**, 1348–1355 (2014).

34. Yanai, I. *et al.* Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650–659 (2005).

35. Birchler, J. A. & Veitia, R. A. The Gene Balance Hypothesis: From Classical Genetics to Modern Genomics. *Plant Cell* **19**, 395–402 (2007).

36. Ohno, S. *Evolution by Gene Duplication*. *Teratology* **1**, (Springer Berlin Heidelberg, 1970).

37. Xu, X. *et al.* Single-cell RNA sequencing of developing maize ears facilitates functional analysis and trait candidate gene discovery. *Dev. Cell* **56**, 557-568.e6 (2021).

38. Rastogi, S. & Liberles, D. A. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evol. Biol.* **5**, 28 (2005).

39. Lu, Z. *et al.* The prevalence, evolution and chromatin signatures of plant regulatory elements. *Nat. Plants* **5**, 1250–1259 (2019).

40. Crow, M., Paul, A., Ballouz, S., Huang, Z. J. & Gillis, J. Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. *Nat. Commun.* **9**, 884 (2018).

41. Lee, J., Shah, M., Ballouz, S., Crow, M. & Gillis, J. CoCoCoNet: Conserved and comparative co-expression across a diverse set of species. *Nucleic Acids Res.* **48**, W566–W571 (2021).

42. Galloway, A. F., Knox, P. & Krause, K. Sticky mucilages and exudates of plants: putative microenvironmental design elements with biotechnological value. *New Phytol.* **225**, 1461–1469 (2020).

43. Roy, S. S., Mittra, B., Sharma, S., Das, T. K. & Babu, C. R. Detection of root mucilage using an anti-fucose antibody. *Ann. Bot.* **89**, 293–299 (2002).

44. Werker, E. & Kislev, M. Mucilage on the root surface and root Hairs of Sorghum: Heterogeneity in structure, manner of production and site of accumulation. *Ann. Bot.* **42**, 809–816 (1978).

45. Voiniciuc, C., Guenl, M., Schmidt, M. H.-W. & Usadel, B. Highly Branched Xylan Made by IRX14 and MUCI21 Links Mucilage to Arabidopsis Seeds. *Plant Physiol.* **169**, pp.01441.2015 (2015).

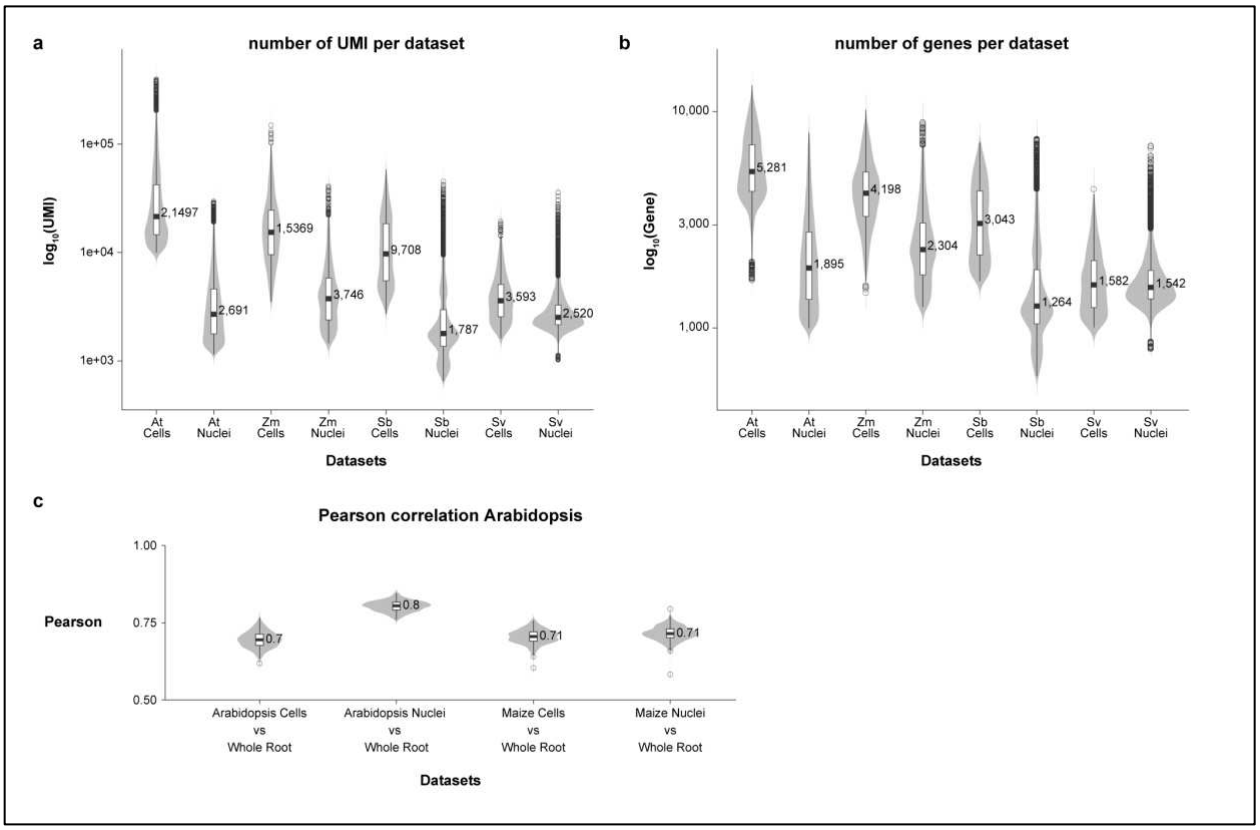
46. Efroni, I., Ip, P.-L., Nawy, T., Mello, A. & Birnbaum, K. D. Quantification of cell identity from single-cell gene expression profiles. *Genome Biol.* **16**, 9 (2015).

47. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902 e21 (2019).

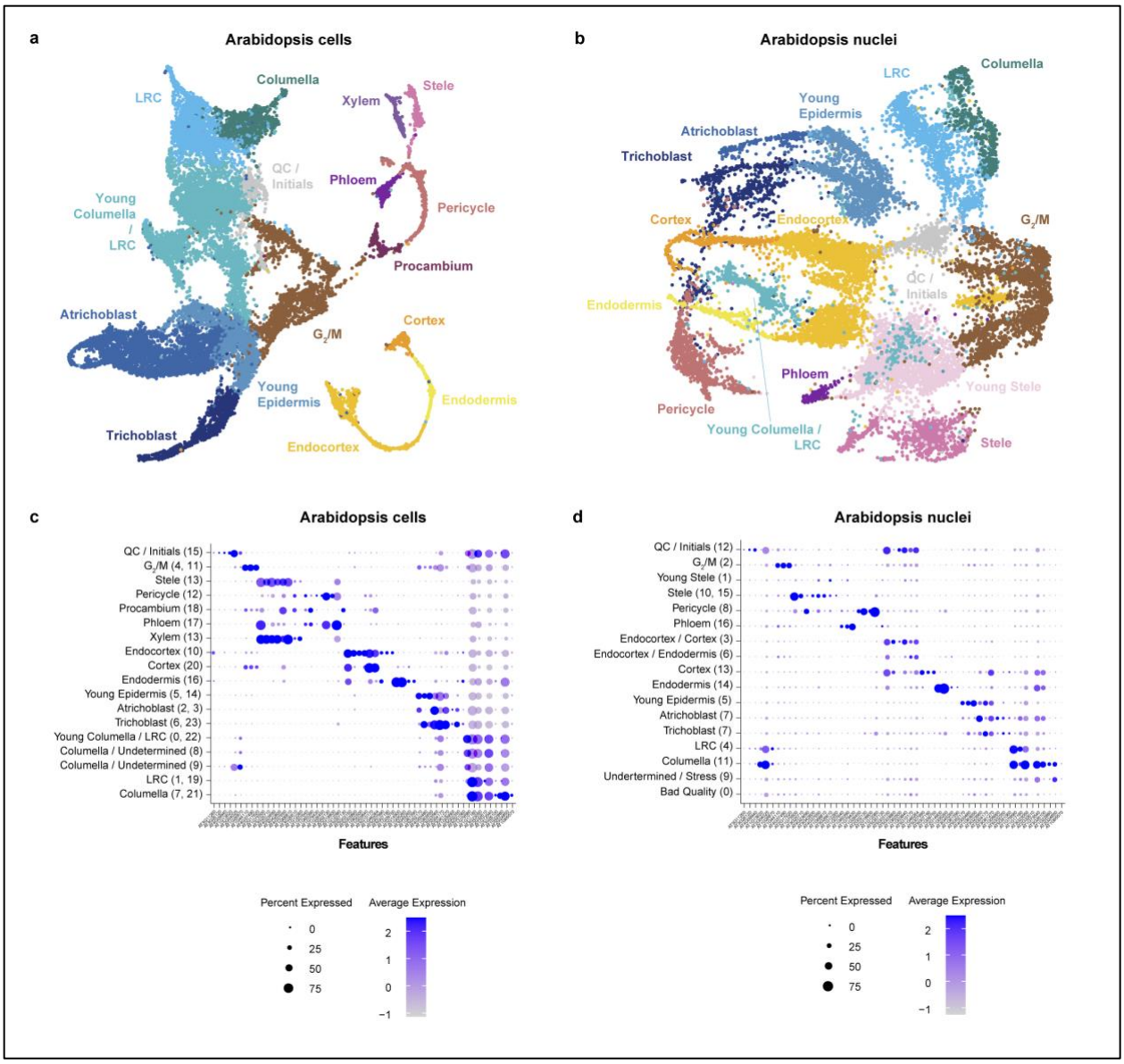
48. Hernández Coronado, M. *et al.* Repel or Repair: Plant Glutamate Receptor-Like Channels Mediate a Defense vs. Regeneration Tradeoff. *SSRN Electron. J.* (2021). doi:10.2139/ssrn.3818443

49. Jackson, D., Veit, B. & Hake, S. Expression of maize KNOTTED1 related homeobox genes in the shoot apical meristem predicts patterns of morphogenesis in the vegetative shoot. *Development* **120**, 405–413 (1994).

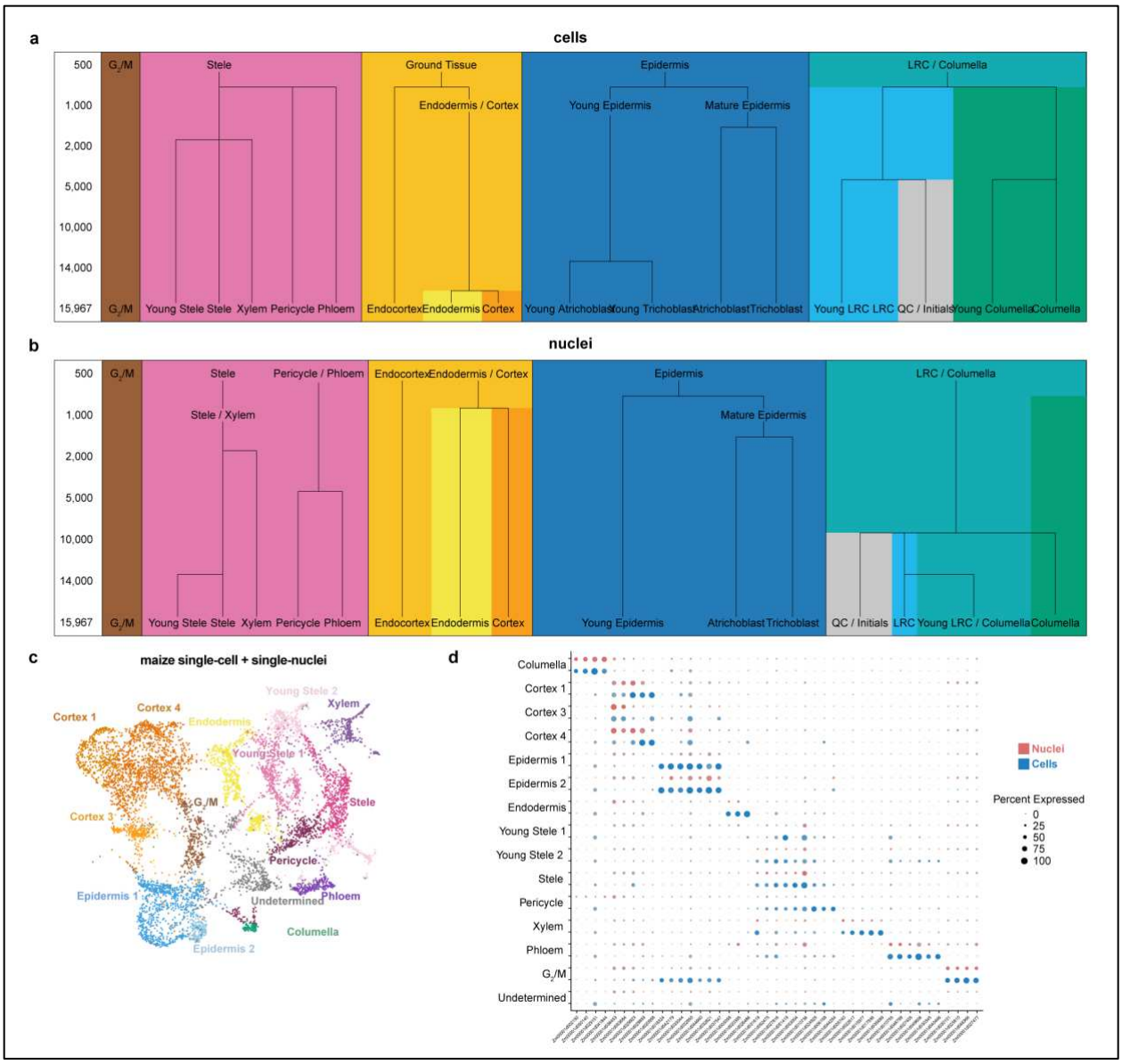
Extended Data



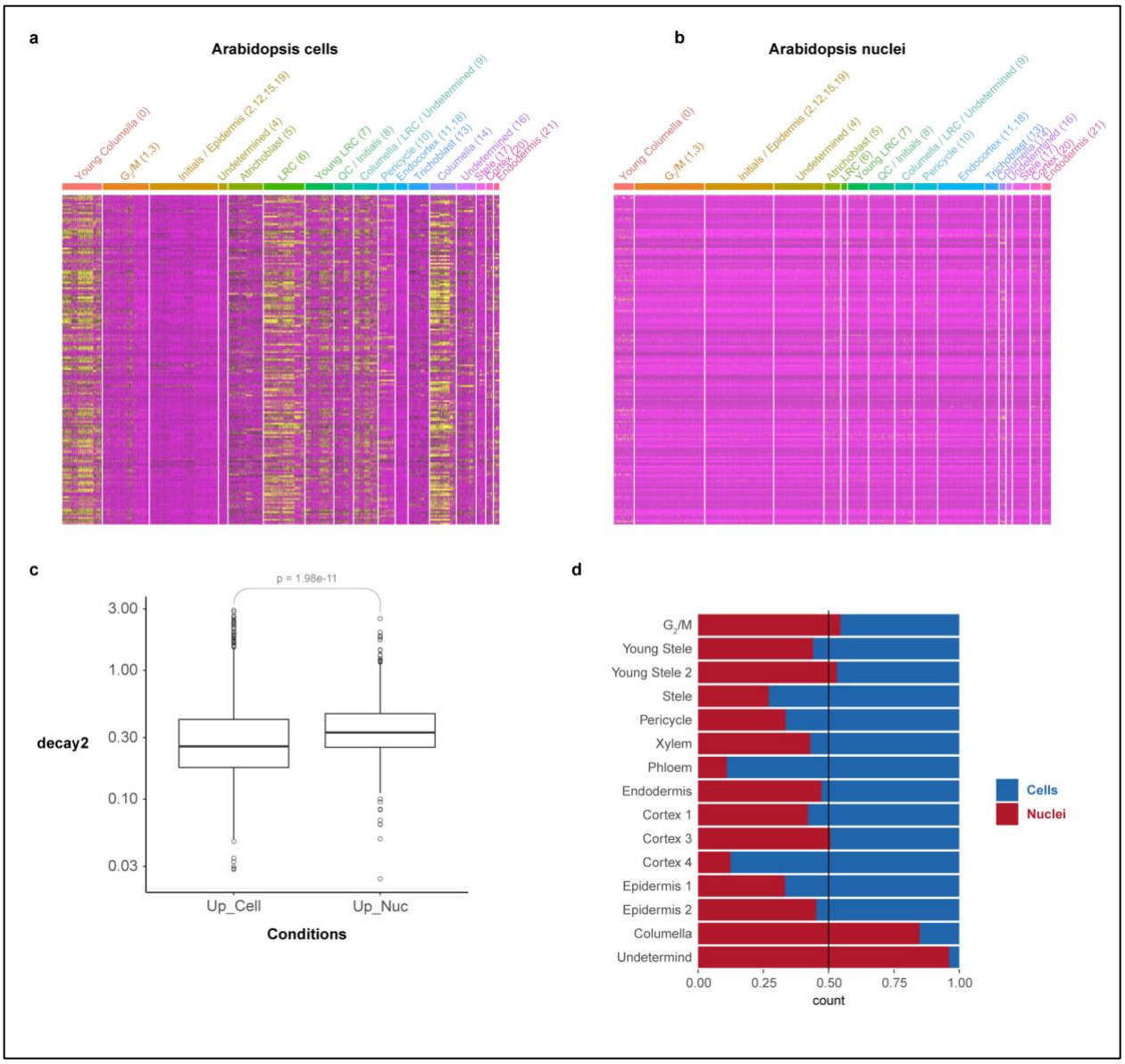
Extended Data Fig. 1: a Violin plot distribution of the number of UMI detected among cells vs. nuclei in Arabidopsis, maize, sorghum, and *Setaria*. **b** Violin plot distribution of the number of genes detected among cells vs. nuclei in the same species as in (a). Black bar is median, box represents 25th and 75th percentile, and vertical line is the 5th and 95th percentile. **c** Pearson correlation distributions of gene expression from single-cell or single-nuclei compared to whole-root RNAseq in Arabidopsis and maize, performed on replicate runs of single-cell or single-nucleus profiles.



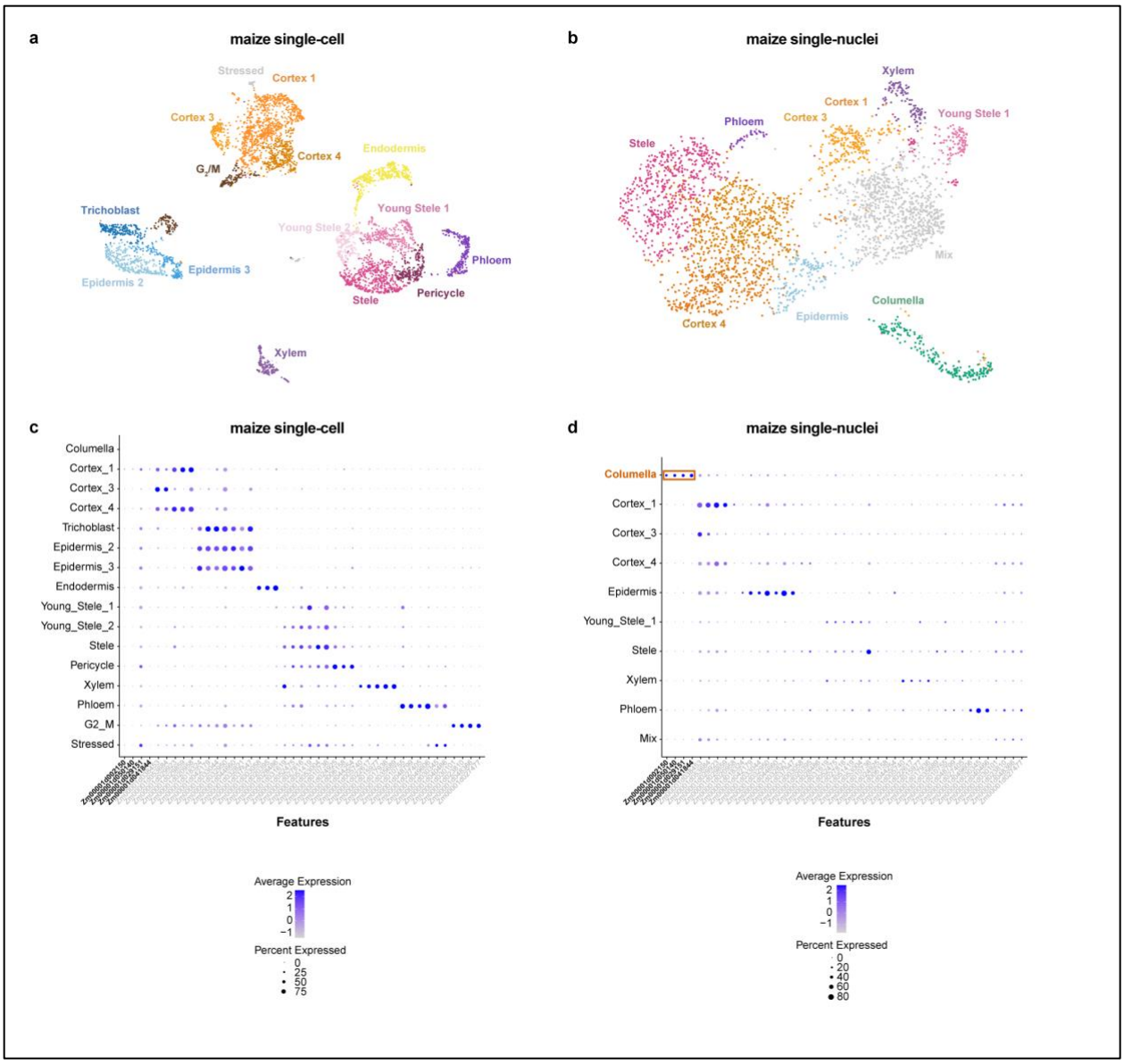
Extended Data Fig. 2: a, b UMAP clustering in Arabidopsis single-cell (**a**) and single-nucleus (**b**) datasets clustered independently, showing clusters with the same assigned cell identities. **c, d** Dot plots showing cluster-wise average expression levels and percent of cells detected for a set of Arabidopsis cell type markers in cells (**c**) vs. nuclei (**d**). The plot shows that markers for a given cell type identified in one profile type (cells or nuclei) show largely the same enriched expression in the second profile type.



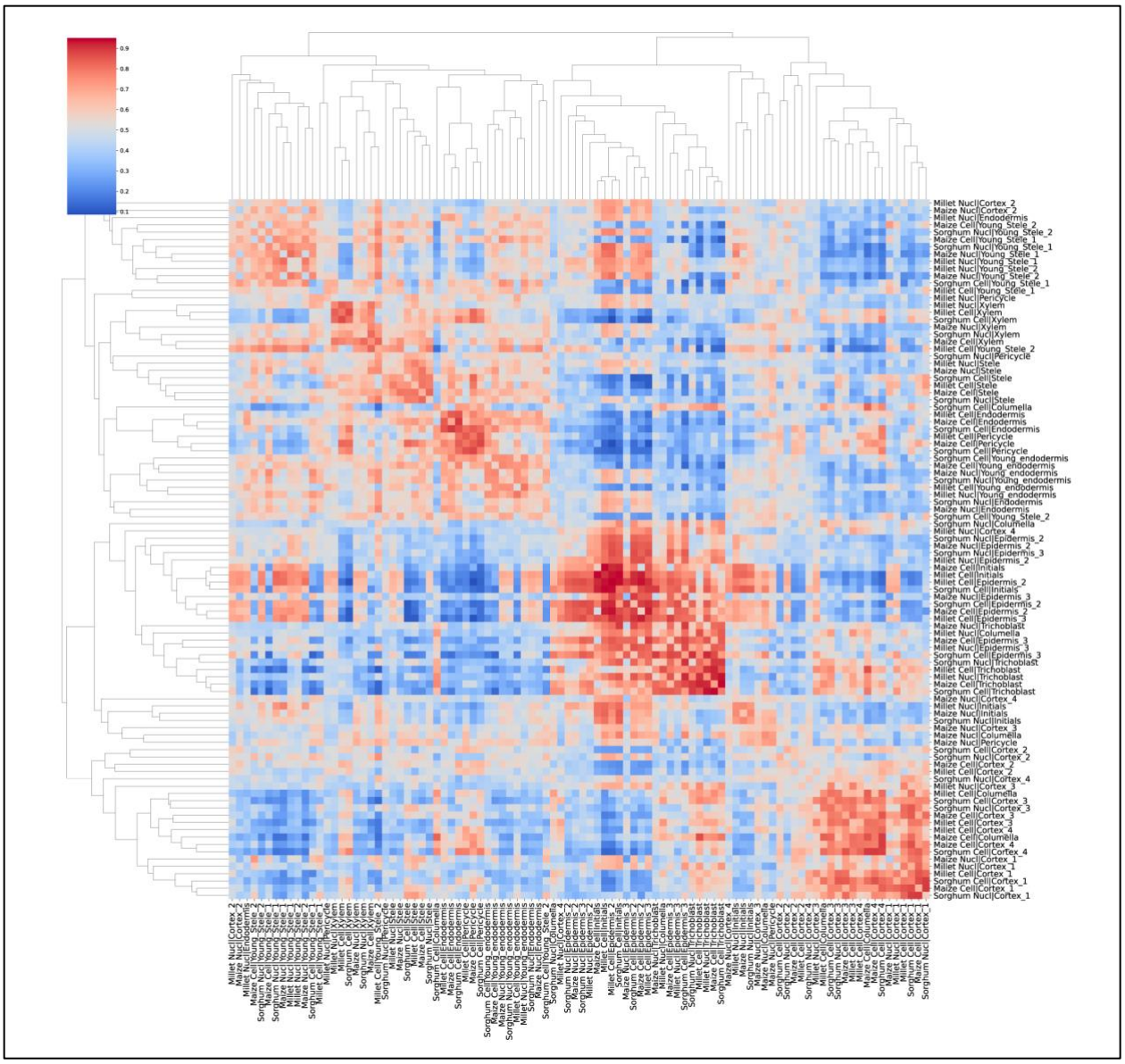
Extended Data Fig. 3: a Arabidopsis down-sampling analysis shows the number of cells needed to resolve clusters into different cell types within tissues. A branch signifies the number of cells needed for Seurat to distinguish a new cluster from a cloud of points when analyzing lower numbers of cells. **b** A similar analysis using the single-nucleus dataset, showing more nuclei are needed to resolve clusters compared to cell profiles in (a). Tracking the branches of graphs in (a) vs. (b) leads to a rule-of-thumb that two-fold more nuclei than cells are needed to identify clusters. **c** UMAPs of the combined maize single-cell and single-nucleus datasets with clusters colored by cell type. **d** Dot plot of maize marker genes in cells (blue) or in nuclei (red), showing concordance of marker expression in the two datasets.



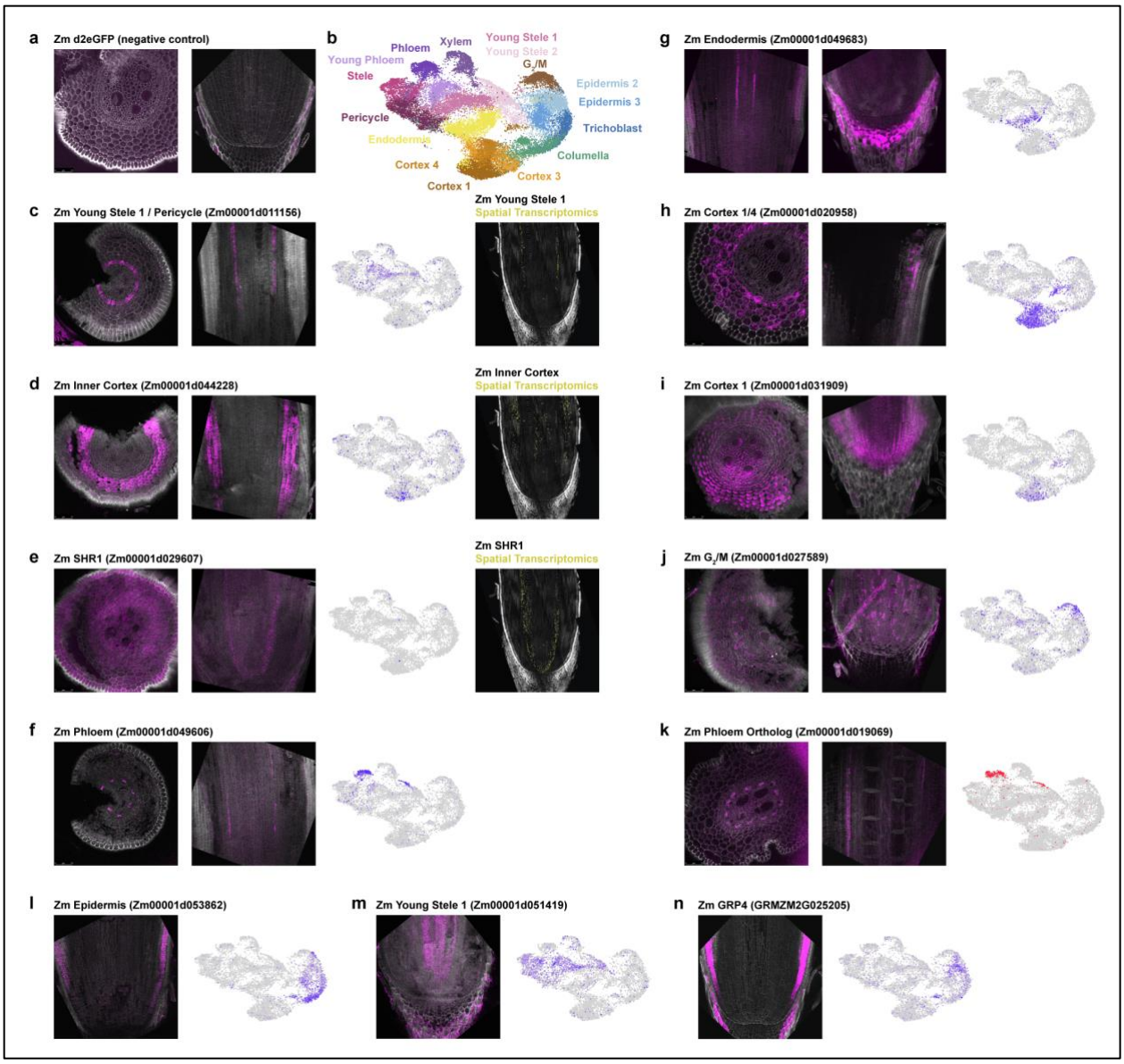
Extended Data Fig. 4: a, b Heatmaps of Arabidopsis genes known to be induced by protoplast generation (Birnbaum et al., 2003) showing their expression in cells (a) vs. nuclei (b). The analysis shows that stress-induced genes also have higher expression in cells vs. nuclei, with particular induction in specific cell types. **c** Distribution of expression levels of genes annotated for mRNA decay in cells or in nuclei. The same genes are analyzed in both cells and nuclei. A significant increase in mRNA decay gene expression level was detected in nuclei (Wilcoxon rank sum test, p-value = 1.98e-11). **d** Proportion of cells (blue) vs nuclei (red) present in each cell type cluster for maize.



Extended Data Fig. 5: a, b UMAPs of maize single-cell and single-nucleus datasets clustered independently. Only the single-nucleus dataset displays a cluster annotated as columella, which is absent in the single-cell dataset. **c, d** Dot plot of maize marker genes for each cell type cluster, showing expression in cells (**c**) and in nuclei (**d**) datasets independently. Markers for columella outlined in the red box are only present in the nuclei dataset.



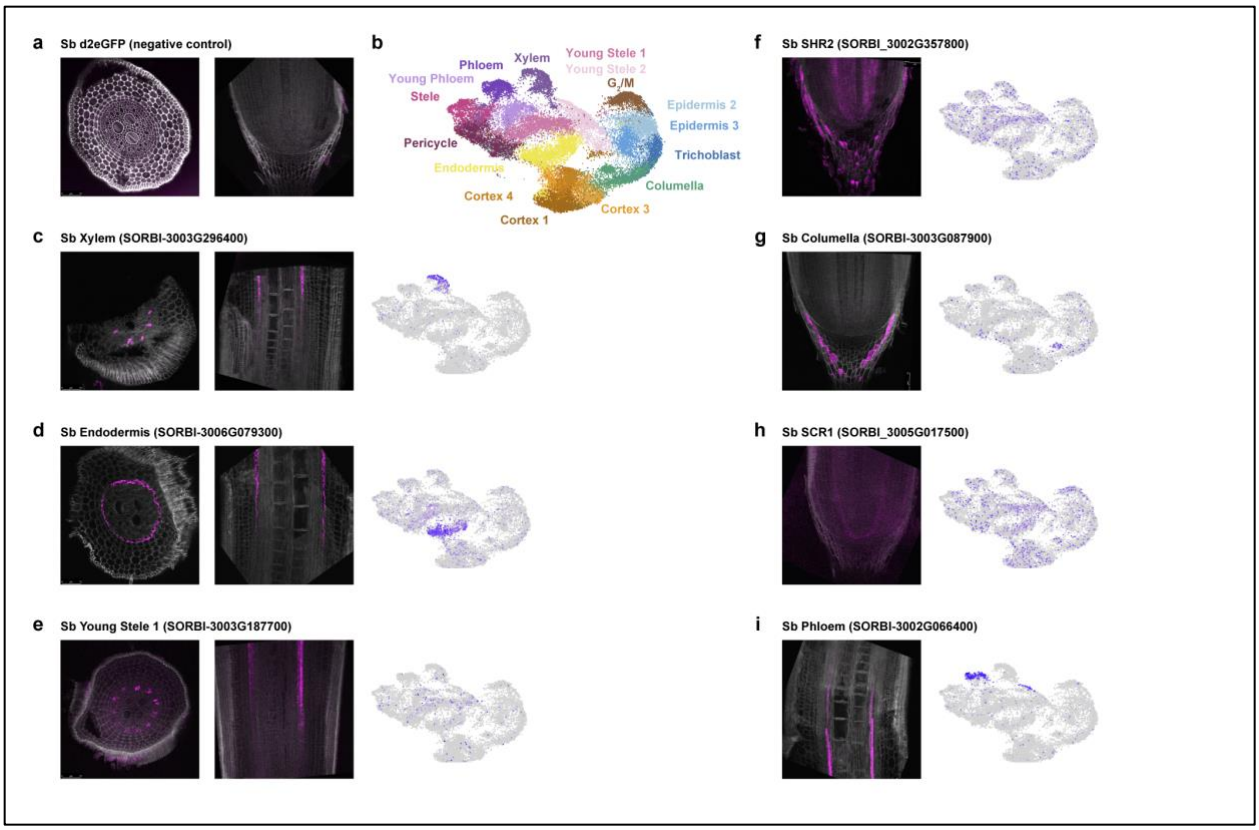
Extended Data Fig. 6: AUROC test as performed in MetaNeighbor comparing every cell type in maize, sorghum, and *Setaria* using single-cell and single-nucleus datasets separately, showing that cells and nuclei largely group by cell type and not by either species or profiling method (cells vs. nuclei).



506

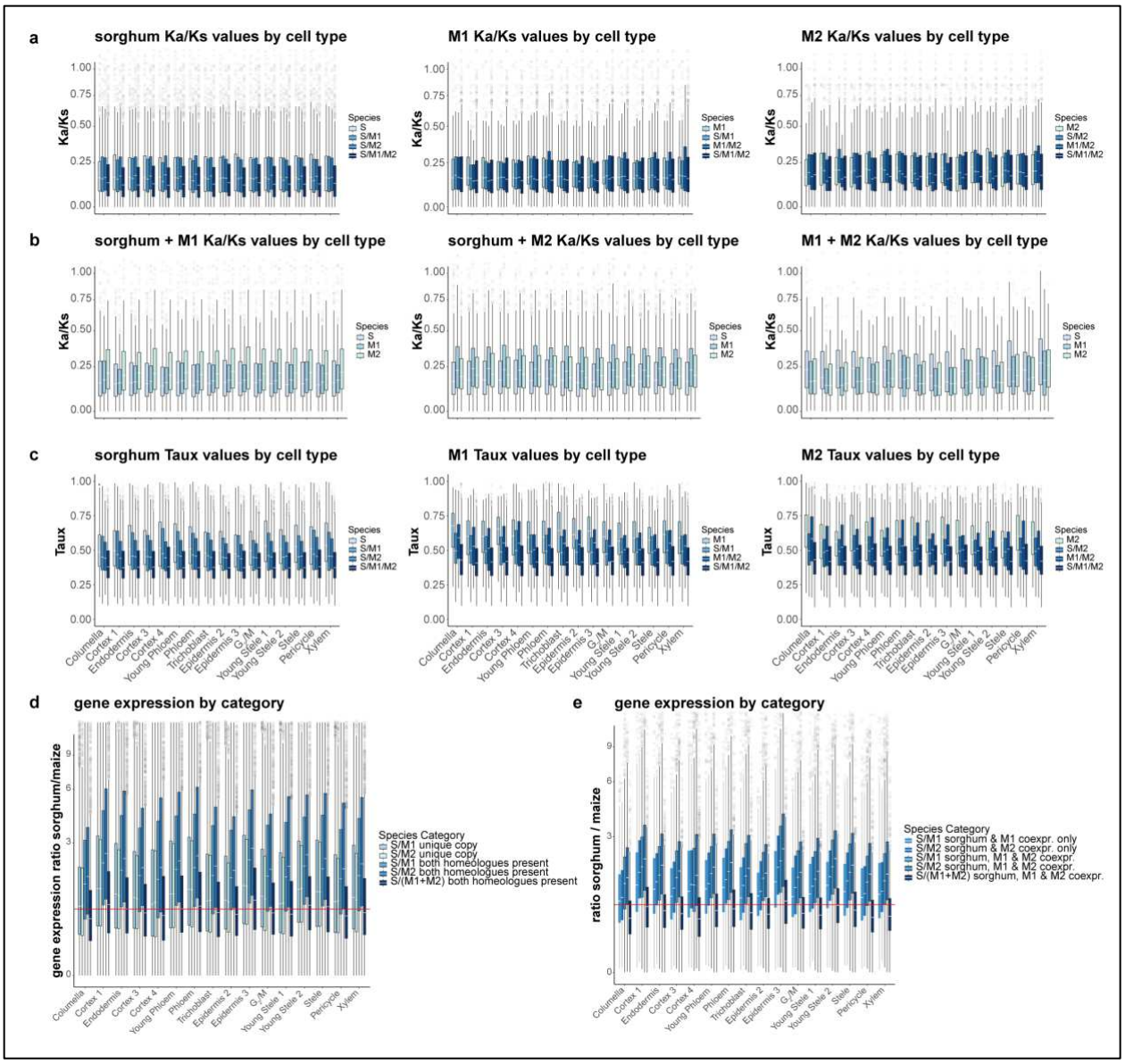
507 **Extended Data Fig. 7: a-n** *in situ* hybridization using Hairpin Chain Reaction (HCR) probes
508 labeling various transcripts in maize. UMAPs showing each transcript's cluster localization are
509 shown next to each probe's fluorescent image. Additionally, spatial transcriptomics imaging data
510 is shown for a three probes in maize (c-e), further validating the cluster annotations. The
511 minimum/maximum values for each fluorescence channel (grey: autofluorescence, magenta:
512 HCR probes) have been adjusted to show the localization more clearly in the merged image.

513



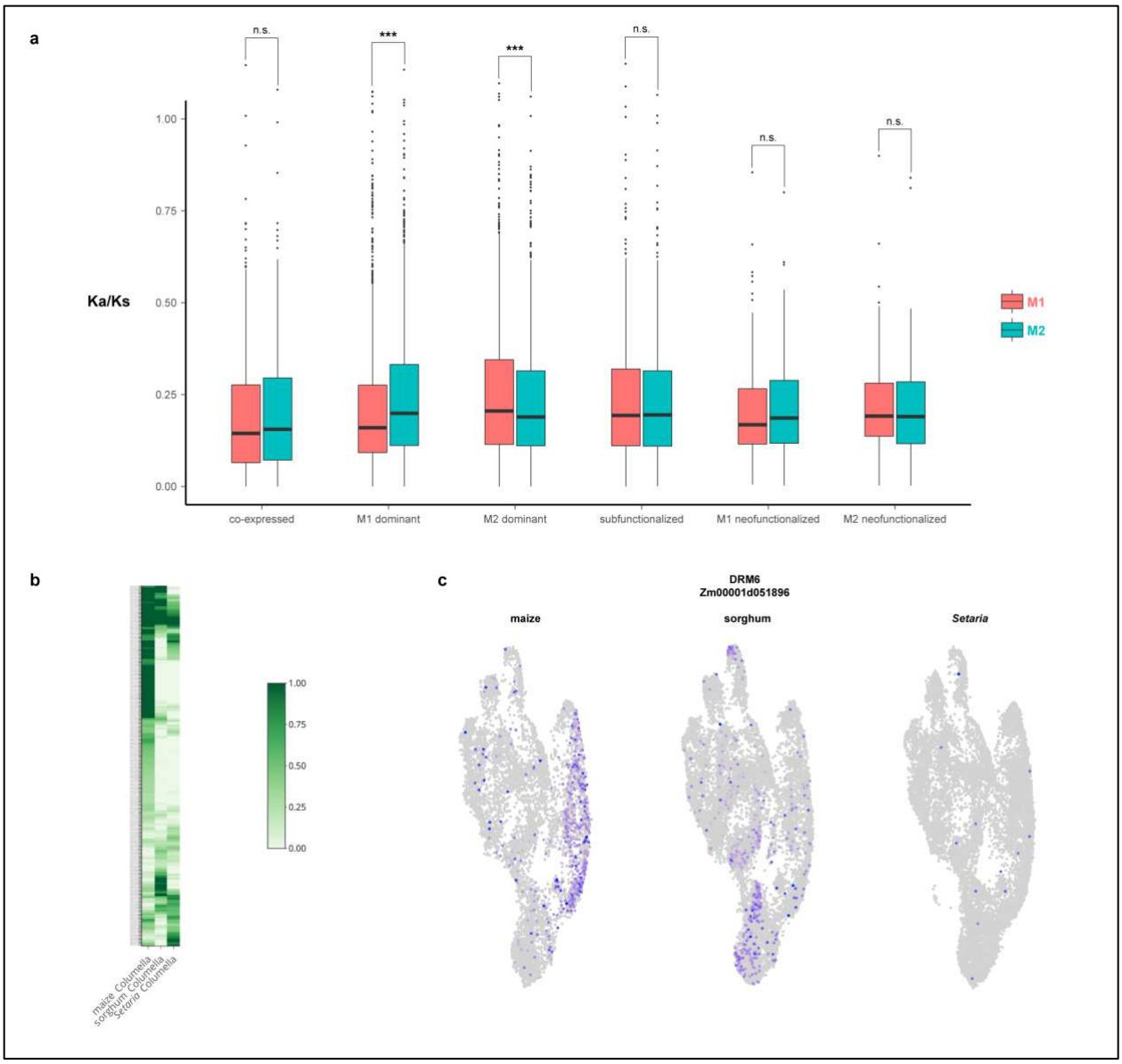
514

515 **Extended Data Fig. 8: a-i** *in situ* hybridization using Hairpin Chain Reaction (HCR) probes
516 labeling various transcripts in sorghum. UMAPs showing each transcript's cluster localization are
517 shown next to each probe's fluorescent image. The minimum/maximum values for each
518 fluorescence channel (grey: autofluorescence, magenta: HCR probes) have been adjusted to
519 show the localization more clearly in the merged image.
520



Extended Data Fig. 9: a Distribution of Ka/Ks values (relative to *Setaria*) of genes or homeologs among the different genomes or subgenomes across cell types (labeled on bottom graph in (c)), with expression pattern by cell type shown in subcategories. Letters represent co-expression categories of sorghum genes and maize homeologs in a given cell type (i.e. S = only sorghum gene expressed in cell type, SM1 = sorghum and M1 genes expressed in cell type, etc.). **b** Distribution of Ka/Ks values in cases of specific co-expression patterns (listed in title of graph) for each cell type (labeled in (c)). For example, in the first graph on the left, when sorghum and M1 are co-expressed in a given cell type, the M2 homeolog always has a slightly higher Ka/Ks value, showing they are under less stringent purifying selection. **c** Distribution of taux (cell specificity) measure of genes in different cell types according to comparative expression patterns. For example, in the first graph on the left, sorghum genes have a higher taux (i.e. more specificity) when they are not expressed in the same cell type as M1 or M2. **d** Analysis similar to Fig. 2d showing expression ratios between sorghum and M1 or M2 homeologs, here broken down by cell type and genomic status of M1 or M2 homeolog (retained or lost). The trends show the strongest dosage compensation at the cell-type level when both copies are retained and expressed. **e**

Analysis similar to Fig. 2d on only the retained homeologs showing ratios of expression between sorghum and M1 and/or M2 homeologs broken down by cell type depending on their co-expression with sorghum. The trends show dosage compensation by cell type only when the two homeologs are expressed in that cell type.



542

543 **Extended Data Fig. 10: a** Ka/Ks distribution of co-expressed vs dominant vs neofunctionalized
544 genes. Co-expressed genes display higher purifying selection. However, only the non-dominant
545 ortholog shows an increas in Ka/Ks distribution compared to the dominant ortholog, while other
546 categories don't display significant differences in purifying selection. Pairwise Wilcoxon Test ***
547 $p < 0.001$. **b** Heatmap represents the 443 most divergent genes across species in the CoCoCoNet
548 analysis, showing expression differences in the columella of the three grass species. **c** Example
549 of the gene DMR6 switching its expression between columella in maize vs epidermis / cortex in
550 sorghum.

551

552

Methods

Plant growth conditions

Seeds of *Arabidopsis thaliana* var Col0, *Zea maize* B73 and *Sorghum bicolor* Btx623, *Setaria viridis* (Accession PI 669942, U.S. National Plant Germplasm System)
Arabidopsis seeds were imbibed for 48h at 4°C before being surface sterilized and placed on a nylon mesh (110um) and agar plates-1/2 × Murashige and Skoog salts (Sigma M5524), 0.5% sucrose, 0.8% Agar (Sigma A1296). Plant were transferred vertically in growth chambers set to 23°C and a 16 h light/8 h dark cycle (400 µmol m^{-2} s^{-1}). Root tips were collected 7 days after transfer, cut with a feather scalpel at 150um from the tip and directly transferred to either the protoplast solution at room temperature or the nuclei lysis buffer at 4°C.

Maize, Sorghum seeds were sterilized using bleach (1.5% active chloride) and 0.001% tween 20 for 20mins then 4% chloramine T for 20mins. Setaria seed germination was induced by incubation in 4% liquid smoke (Colgin, Authentic Natural Hickory) at 29°C for 24h.Then setaria seeds were sterilized using bleach (1.5% active chloride) and 0.001% tween 20 for 20mins. All seeds were placed between two layers of brown paper (Anchor Paper&Cie., 38# regular), rolled and covered with aluminum foil to prevent roots from exposure to direct light. Rolls were placed in a bucket of tap water at 28/24°C at 16 h light/8 h dark cycle (250 µmol m^{-2} s^{-1}) for 7 days (15days for Setaria) before harvesting the root tips. Primary and seminal root tips were cut using a fine scalpel at 0.5 cm form the tip for Maize and Sorghum, 0.2cm from the tip for Setaria and transferred either to the pre-incubation solution for single cell or to the nuclei lysis buffer.

Protoplasting

Protoplasts were generated from primary and seminal roots as described previously [46]. For Maize, Sorghum and Setaria, roots were cut above the meristem as describe above and placed in pretreatment solution containing L-cysteine for 40 min (3% sorbitol, 2.5mM L-cysteine, 20mM MES, pH 5.8 with Tris) to improve enzyme efficiency and cell wall digestion. Cell walls were digested for 90 min in an enzyme solution optimized for monocot roots (Mannitol 8%, 400mM, MES 20mM, KCl 20mM, CaCl2 40mM, pH 5.8 with Tris, BSA 100ug/ml, 2% cellulase “Onozuka” RS, 1.2% cellulase “Onozuka” R10, 0.4% macerozyme R-10 (all three Yakult Pharmaceutical Industry CO.), 0.36% pectolyase Y-23 (MP Biomedicals)). Protoplast were then filtered through a 40 µm cell strainer and transferred to microcentrifuge tubes for centrifugation.
For Arabidopsis, roots were cut above the meristem as described above and placed in an enzyme solution optimized for Arabidopsis (Mannitol 8%, 400mM, MES 20mM, KCl 20mM, CaCl2 40mM, pH 5.8 with Tris, BSA 100ug/ml, 1.2% cellulase “Onozuka” R10, 0.4% macerozyme R-10 (all three Yakult Pharmaceutical Industry CO.)). Protoplast were then filtered through a 20 µm cell strainer and transferred to microcentrifuge tubes for centrifugation.

Protoplast were centrifuge for 3 min at 500 x g and the pellets were washed and resuspended in washing solution twice (Mannitol 8%, MES 20mM, KCl 20mM, CaCl2 10mM, pH 5.8 with Tris, BSA 100ug/ml) and used immediately for single-cell RNAseq.
An aliquot of protoplasts was stained with trypan blue (0.2% final) and check on hematocytometer under microscope to determine viability and cell concentration before loading into the 10x chromium.

Nuclei extraction

For all species, root tips are directly transferred in prechilled lysis buffer (0.3M sucrose, 15mM Tris HCl pH8, 60mM KCl, 15mM NaCl, 2mM EDTA, Spermine 0.5mM, Spermidine 0.5mM, 15mM MES, 0.1% Triton, 5mM DTT*, 1mM PMSF* , 1% Plant Protease Inhibitors* 1ml(Sigma P9599), BSA 0.4%*, RNase inhibitor 0.2u/ul*, (* add last minute)). Roots are chopped on ice with scalpel

blades for 5-10mins and transferred into a pre-chilled dounce homogenizer (Kimble, 885302). Pestle is moved for 10 back and forth, sample keep on ice for 10mins then additional 10 back and forth. Then root extracts are filtered at 20μm into a centrifuge tube and centrifuge for 10mins at 500g (Maize, Sorghum, Setaria) or 1000g (Arabidopsis). Pellet is washed once with washing buffer (0.3M sucrose, 15mM Tris HCl pH8 , 60mM KCl, 15mM NaCl, Spermine 0.5mM, Spermidine 0.5mM, 15mM MES, 5mM DTT*,1mM PMSF*, 1% Plant Protease Inhibitors* 1ml(Sigma P9599), BSA 0.4%*, RNase inhibitor 0.2u/ul*, (* add last minute)). Finally, nuclei are resuspended into final buffer (0.3M sucrose, 15mM Tris HCl pH8 , 60mM KCl, 15mM NaCl, Spermine 0.5mM, Spermidine 0.5mM, 15mM MES, 5mM DTT*, 1% Plant Protease Inhibitors* 1ml(Sigma P9599), BSA 0.4%*, RNase inhibitor 0.2u/ul*,(* add last minute)) and filtered using a 10μm filter. A nuclei aliquot is stained with DAPI for quality check and nuclei counting under microscope and used immediately for single-nuclei RNAseq.

Single cell RNA-seq

16,000 cells or nuclei were loaded in a Single Cell B Chip (10x Genomics) per replicate. Single-cell libraries were then prepared using the Chromium Single Cell 3´ library kit, following manufacturer instructions. Libraries were sequenced with an Illumina NextSeq 550 platform using a 1x150 high-output (2 libraries per chip) or Novaseq 6000 chip SP V2.5, 4 libraries per chip. Raw scRNAseq data was analyzed by Cell Ranger 5.0.1 (10x Genomics) to generate gene-cell matrices. Gene reads were aligned to Arabidopsis TAIR10.38, Maize B73 v4, Sorghum bicolor v3 and Setaria viridis v2 reference genome.

UMAP and ICI analysis

Replicates (see supplementary Table 1) were integrated and cells mapped using the Seurat package v3.0 [47] as follows: first, genes with counts in fewer than three cells were excluded from the analysis and their counts were removed. Second, low quality cells were removed using threshold variable depending on the libraries quality (see supplementary Table 1). Clustering of cells or nuclei separately were done by log-normalized raw counts and the 2000 most variable genes were identified for each replicate using the “vst” method in Seurat. Next, we used the *FindIntegrationAnchors* function to identify anchors between the three datasets, using 20 dimensions. A new profile with an integrated expression matrix containing cells from all replicates was produced with the *IntegrateData* function. For dimensionality reduction, the integrated expression matrix was scaled (linear transformed) using the *ScaleData* function, and Principal Component analysis (PCA) performed. The top 30 principal components were selected. Cells or nuclei were clustered using a K-nearest neighbor (KNN) graph, which is based on the Euclidean distance in PCA space. The *FindNeighbors* and *FindClusters* function with a resolution of 0.5. was applied. Next, non-linear dimensional reduction was performed using the UMAP algorithm with the top 30 PCs. Co clustering of cells and nuclei was performed using the SCT approach. First raw reads were normalized using the *SCTransform* function, then *SelectIntegrationFeatures* was used to identify anchors between the three datasets, using 3000 features. For multiple species clustering, all orthologous genes names were replace by their corresponding maize ID in sorghum and setaria raw features.tsv.gz files. Anchors are combined using PrepSCTIntegration and selected using *FindIntegrationAnchors*. For clustering of maize, sorghum and setaria together, maize was selected as a reference dataset in the *FindIntegrationAnchors* function, for single species integration all datasets are considered equally for the integration. Finally, a Principal Component analysis (PCA) is performed using the first 100 principal components and a non-linear dimensional reduction was performed using the UMAP algorithm with the top 100 PCs.

GO enrichments.

All GO enrichment were performed using shinyGO V0.61 (http://bioinformatics.sdstate.edu/go/) with an FDA of 0.05.

Gene expression analysis across species.

Whole-root transcriptomes were obtained from Ortiz-Ramírez *et al.*, 2021[25] for maize and Hernández Coronado et al., 2021[48] for arabidopsis.

Gene expression was normalized for each species using the *Normalizedata* function from Seurat. Then the average expression per cluster was calculated using *AverageExpression* from Seurat. Ka and Ks values were provided upon request by J.C. Schnable from the lab publication Zhang et al., 2017

Tau t was calculated as describe in (Yanai et al., 2.005) $\tau = \frac{\sum_{i=1}^{N}(1-x_i)}{N-1}$, where N is the total number of cell type and xi is the expression profile component normalized by the maximal component value.

Cell type prediction across species and technologies.

To determine how well the cell clusters characterized the shared identities of cells in their own clusters and the overlaps with the identities of all other cells, we utilized the MetaNeighbor package in Python (https://github.com/gillislab/pyMN) (Fischer et al. 2021; Crow et al. 2018). MetaNeighbor measures the replicability of cell-types by learning a model in one dataset (or subset) and testing for its ability to reconstruct cell-type clusters in the other dataset. First, we labeled all cells and nuclei by the technology used to sequence the transcriptome, by the cluster identity, and by the plant species to which they belonged. Then, we used the PyMN.variable_genes function from MetaNeighbor to subset the gene list to variable genes. This generates a list of genes that are variable across the technology and species. Next, we employed the PyMN.MetaNeighborUS function to measure how well the transcriptional profiles of cells from clusters in one division of the dataset (e.g., technology) predict the identities of cell clusters in the other fraction of the data. This generates pairwise AUROCs for each combination of clusters. To generate the heatmaps, the PyMN.plotMetaNeighborUS was used with a Brown Blue-green color map. This plots the pairwise AUROCs generated previously.

Co-expression Conservation between maize subgenomes and sorghum.

To generate co-expression conservation scores between the two maize sub genomes and sorghum, we use our existing aggregated co-expression networks (Lee et al. 2020). In brief, these networks are built by taking all publicly available data and calculating average correlations between genes pairs within experiments, standardizing within experiments, and then averaging to construct robust meta-analytic networks. We filtered these networks to a previously generated list of gene triplet pairs for maize sub genomes and sorghum. Next, for each gene, we compare the top co-expression partners across species to determine the degree of functional conservation, as described in more detail in previous work (Crow et al. 2022). We calculate this by taking the ranks of a gene's co-expression strength to all other genes in one species and using it to predict that gene's top 10 co-expressed partners in the second species. This is then done again in the reverse direction, and the two scores are averaged (calculated as an AUROC).

with the lowest co-expression scores (0.34 < FC.Score) and highest cell specificity (> 0.8) in the root cap (Supplementary Table 7; Extended Data Fig. 10b). Similar to trends in other cell types, these highly cell-type specific genes

Dominance vs. partition score:

To calculate the Dominance vs. partition score, for each ortholog triplet (S, M1, M2) we calculated the number of cells in which M1 or M2 was dominant or co-expressed together in the same cells where the sorghum ortholog was expressed.

*Score = (number of cells in which M1 is dominant * number of cells in which M2 is dominant) - (number number of cell of the dominant ortholog - number of cell of the non dominant ortholog)*

If the score is negative, the score is normalized by

$$NormScore = \frac{Score}{\#\ of\ cell \text{ in which M1 and M2 are expressed}}$$

If the score is positive, the score is normalized by dividing it by:

$$NormScore = \frac{Score}{(\#\ of\ cell \text{ in which M1 and M2 are expressed } * 0.5)^2}$$

Statistical analysis:

Each species marker genes were identified using *FindAllmarkers* functions from Seurat, log.FC= 0.25, pt.1 > 0.750 pt.2 < 0.250. Differential gene expression was done using the *Findmarkers* function from Seurat with default parameter function. For Fig2 e, Extended Data Fig. 4 c, 10 a, statistical analysis was performed on R using a pairwise Wilcoxon test with p.adjust method "BH" as data is not normally distributed.

Correlation analysis on Extended Data Fig 1 c was performed using Pearson correlation function on R between whole-root data coming from and single cell or single nuclei. Briefly averaged gene expression was calculated for each gene while combining every cell types using the *AverageExpression* function from Seurat.

For Fig 4 a, to generate p-values for evaluating the significance of the differences between each pair of AUROCs generated by MetaNeighbor, we utilized the Hanley McNeil test, which produces a Z-score for the difference (Hanley and McNeil 1983). As each MetaNeighbor AUROC is the averaged AUROC from two reciprocal tests between a pair of cell clusters, we chose the smaller of the two clusters as the number of true positives (NTP) to generate the most conservative p-value. The number of true negatives was the total number of cells, less the number of true positives. Following the calculation of Z-scores for each pairwise combination of AUROCs, we utilized the scipy.stats.norm.sf function in Python to convert the Z-scores into p-values for a two tailed test.

Hanley, J. A., and B. J. McNeil. 1983. “A Method of Comparing the Areas under Receiver Operating Characteristic Curves Derived from the Same Cases.” Radiology 148 (3): 839–43. https://doi.org/10.1148/radiology.148.3.6878708.

“Half mount” *in situ* hybridization:

Probes (Hairpin Chain Reaction (HCR) RNA-FISH) and reagents (including the Probe Hybridization Buffer, Probe Wash Buffer and Amplification buffer) are ordered from Molecular Instruments (https://www.molecularinstruments.com/shop)(**Supplementary Table 9**).

For fixation, germination paper containing 7 day old maize or sorghum roots are unrolled and small volume of fixative FAA (4% formaldehyde, 5% glacial acetic acid, 50% ethanol in RNAse

free water) is pipetted onto each root. Then longitudinal sectioning of root tips is performed using a 15° microscalpel. Roots are cut up to ~3cm from the tip, then immediately fixed by transferring to FAA in 5ml screw caps and put under vacuum several times until they no longer float. Roots are then agitated at RT for at least 1 hour in a tube revolver. (All washes in the protocol are performed in a tube revolver or stated otherwise.)
Samples are dehydrated in a series of washes at RT: 70% ethanol for 15 min, 90% ethanol for 15 min, 100% ethanol 2x for 15 min each, 100% methanol 2x for 15 min each. Samples can then be stored at -20°C for several weeks. Samples are washed 2x for 15 min in 100% ethanol at RT before being permeabilized for 30 min in 50% Histo-Clear II / 50% EtOH at RT. Then they are incubated 2x for 30 minutes in a solution of 100% Histo-Clear II at RT. Each time, vacuum is applied for the first 10 minutes.

Samples are rehydrated through a series of washes: 50% Histo-Clear II / 50% EtOH for 15 min, 100% EtOH for 15 min, 50% EtOH / 50% DPBS-T (0.1% Tween20, 1x DPBS) for 15 min (roots will float up then settle after a few minutes), 100% DPBS-T 2x for 15 min (roots will float up again). Samples are incubated with Proteinase K (0.1 M Tris-HCl (pH 8), 0.05 M EDTA (pH 8), Proteinase K 80 µg ml^{-1} final) at RT under vacuum for 5 min then digested with Proteinase K for 25 min in a 37°C water bath with manual agitation every 5-10 minutes (roots should turn a little yellow after this step). Samples are washed 2x for 15 min in DPBS-T at RT then incubated with Fixative II (4% formaldehyde in DPBS-T) under gentle vacuum for 10 min then in a tube revolver for 30 mins at RT. They are then washed 2x for 15 min each in DPBS-T at RT. Roots are aliquoted into 2 mL Eppendorf tubes and incubated in 500 µL of HCR Probe Hybridization Buffer, vacuum is applied for 10 mins then roots are incubated for 1 hour at 37°C in a thermomixer with agitation (1000 rpm).

Samples can then be stored in Probe Hybridization Buffer at -20°C up to several weeks.
Probe buffers are made by adding 0.8 pmol of each probe set (e.g. 2 µL of the 1 µM stock) to 500 µL of HCR Probe Hybridization Buffer at 37°C. Pre-hybridization solution is removed and replaced with probe solution. Samples are hybridized by incubating overnight (~20h) at 37°C in a thermomixer with agitation (1000 rpm). The following day, excess probes are removed by washing 4x for 15 min each with 1 mL of HCR Probe Wash Buffer at 37°C in a thermomixer with agitation. Samples are washed 2x for 5 min each with 1 mL of 5x SSC-T (25% 20x SSC, 0.1% Tween20) at RT in a thermomixer with agitation. SSC-T is replaced with 500 µL of amplification buffer, gentle vacuum is applied in a fume hood for 10 minutes and then samples are pre-amplified by incubating in a tube rotator at RT for 50 min. While samples pre-amplify, 6 pmol of hairpin h1 and 6 pmol of hairpin h2 (i.e. 5 µL of the 3 µM stocks) are prepared, each in its own separate tube. Hairpins are snap-cooled by heating at 95°C for 90 seconds then kept in a dark drawer at RT for 30 min. Amplification solution is prepared by combining snap-cooled h1 and h2 hairpins in 250 µL of HCR Amplification Buffer at RT. Pre-amplification solution is removed and and replaced with amplification buffer containing hairpin solution overnight (~20h) in the dark at RT in a thermomixer with agitation (1000 rpm). Excess hairpins are removed by washing with 1 mL of 5x SSC-T at RT in a thermomixer with agitation, 2x for 5 min each, then 2x for 30 min each, 1x for 5 min. Samples are transferred onto a glass slide (in 5x SSC-T) and cut using a 30° microscalpel and arranged so that the cut face of the roots is facing upwards. They are then covered with coverslip and imaged on confocal microscope.

Spatial transcriptomics:

Tissue fixation and embedding was performed as described in[49].

Sample slide preparation: Formaldehyde-fixed paraffin-embedded tissue sections (10 µm) were placed within capture areas on Resolve Bioscience slides and incubated on a hot plate for 10 min at 60 °C to attach the samples to the slides. Slides were treated to allow deparaffinization, permeabilization, acetylation, and refixation. After complete dehydration of the samples, a few drops of SlowFade-Gold Antifade reagent (Invitrogen) were added to the sections and covered with a thin glass coverslip to prevent damage during shipment to Resolve BioSciences (Germany).

Sample pre-treatment and priming: In preparation for hybridization, the coverslip is removed and the mounting reagent is washed twice in 1x PBS for 30 min 4 °C, followed by one min washes in 50% Ethanol and 70% Ethanol at room temperature. Samples were primed, after the aspiration of ethanol, by the addition of buffer BST1 for optimal hybridization of probes during the Molecular Cartography™ procedure, which uses a combination of probes and single-molecule fluorescence in-situ hybridization to identify 100 separate transcripts. Tissues were hybridized overnight at a constant temperature with all probes specific to the target genes. Samples were washed the next day to remove excess probes and fluorescently labeled in a two-step procedure. Regions of interest were imaged as described below and fluorescent signals were removed after imaging via a decolorization procedure. Color development, imaging, and decolorization were repeated over several cycles to develop a unique combinatorial code for every target gene that was derived from raw images as described below.

Probe design: The probes for 100 genes were designed based on full-length protein-coding transcript sequences (Supplementary Table 9). Probe design is based the manufacturer's proprietary algorithm, with probes available from the Resolve. After screening to generate probe candidates and discard ambiguous ones, the probes were mapped to the background transcriptome using *ThermonucleotideBLAST,* and probes with stable off-target hits were discarded.

Imaging: Samples were imaged on a Zeiss Celldiscoverer 7, using the 50x Plan Apochromat water immersion objective with an NA of 1.2 and the 0.5x magnification changer, resulting in a 25x final magnification. Standard CD7 LED excitation light source, filters, and dichroic mirrors were used together with customized emission filters optimized for detecting specific signals. Excitation time per image was fixed at 1000 ms for each channel, 20 ms for DAPI, and 1 ms for Calcofluor White. A z-stack was taken at each region with a distance per z-slice according to the Nyquist-Shannon sampling theorem. A custom CD7 CMOS camera (Zeiss Axiocam Mono 712, 3.45 µm pixel size) was used. The imaging for the cell-wall specific stain, Calcofluor White, was done at the end of all primary imaging. Before the preprocessing of the images, all images were corrected for background fluorescence. Based on the raw data image, the 20% darkest local pixel values and positions were determined and copied to a new empty image (background image) having the same size as the image to be corrected. The remaining 80% of pixels of the background image were generated based upon the surrounding existing pixel values using a distance-weighted average value. Finally, the background-corrected image (bc-image) was created by subtracting the background image values from the raw data image values.

Extraction of features: In the first step, a target value for the allowed number of maxima was calculated based on the area of the slice in µm² multiplied by an empirically optimized factor (0.5x). The resulting target value was used to adapt the threshold for the algorithm iteratively searching local 2D-maxima. The threshold leading to the closest number of maxima equal to or smaller than the target value was used for further steps and the respective maxima were stored in a reiterative process for every image slice independently. Maxima that did not have a neighboring maximum in an adjacent slice (termed as z-group) within a radius of one pixel were excluded. For the resulting list of maxima, the absolute brightness (Babs), the local background

(Bback), and the average brightness of the pixels surrounding the local maximum (Bperi) were measured and stored. The resulting maxima list was further filtered in an iterative loop by adjusting the allowed thresholds for (Babs-Bback) and (Bperi-Bback) to reach a feature target value based on the total volume of the 3D image. Only maxima still in a z-group with a size of at least 2 passed this stringent filter step. Each z-group was counted as one hit and the members of the z-groups with the highest absolute brightness were used as features to resemble 3D point clouds.

Determination of transformation matrices, pixel evaluation, and decoding: To align the raw data images from different imaging rounds, these images had to be corrected for the 6 degrees of freedom in 3D-space The extracted feature point clouds were used to find the transformation matrices to align the raw data images. Based on the transformation matrices, the corresponding images were processed by a rigid transformation using trilinear interpolation. The aligned images were used to create a profile for each pixel, which were then filtered for a variance from zero normalized by the total brightness of all pixels in the profile. Matched pixel profiles with the highest score were assigned as an ID to the pixel to further group the neighboring pixel with the same ID. The local 3D-maxima of the groups were determined as potential final transcript locations, which were additionally evaluated by the number of maxima in the raw data images where a maximum was expected. The finalized maxima were decoded by the fit to the corresponding code to be written to the results file and considered to resemble transcripts of the corresponding gene. The ratio of signals matching to codes used in the experiment and signals matching to codes not used in the experiment were used as estimation for specificity (false positives). Final image analysis was performed in ImageJ using the Polylux tool plugin from Resolve BioSciences to examine specific Molecular Cartography signals.

All R scripts related to models and statistical analyses are available upon request.

All raw RNA-seq data will be deposited in GEO upon publication.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTable1Summaryruns.xlsx
- SupplementaryTable2ArabidopsisclusteringrepartitionofNucleiorCellaloneandmergeddataset.csv
- SupplementaryTable3DifferentialyexpressedgenesbetweenAthNucleiandCells.xlsx
- SupplementaryTable4CellTypemarkersMaizeSorghumSetaria.xlsx
- SupplementaryTable5GOEnrichmentSSM1SM2M1M2SM1M2.csv
- SupplementaryTable6OrthologsDominance.xlsx
- SupplementaryTable7CococoNetAnalysisallfunctionalscorescalculated.csv
- SupplementaryTable8ProberefInsituhybridizationandmolecularcartography.xlsx