Article

A pan-grass transcriptome reveals patterns of cellular divergence in crops

https://doi.org/10.1038/s41586-023-06053-0

Received: 8 June 2022

Accepted: 5 April 2023

Published online: 10 May 2023

Check for updates

Bruno Guillotin[1,2], Ramin Rahni[1], Michael Passalacqua[3], Mohammed Ateequr Mohammed[2], Xiaosa Xu[3], Sunil Kenchanmane Raju[1,4], Carlos Ortiz Ramírez[1,7], David Jackson[3], Simon C. Groen[5], Jesse Gillis[6] & Kenneth D. Birnbaum[1,2]✉

Different plant species within the grasses were parallel targets of domestication, giving rise to crops with distinct evolutionary histories and traits[1]. Key traits that distinguish these species are mediated by specialized cell types[2]. Here we compare the transcriptomes of root cells in three grass species—*Zea mays*, *Sorghum bicolor* and *Setaria viridis*. We show that single-cell and single-nucleus RNA sequencing provide complementary readouts of cell identity in dicots and monocots, warranting a combined analysis. Cell types were mapped across species to identify robust, orthologous marker genes. The comparative cellular analysis shows that the transcriptomes of some cell types diverged more rapidly than those of others—driven, in part, by recruitment of gene modules from other cell types. The data also show that a recent whole-genome duplication provides a rich source of new, highly localized gene expression domains that favour fast-evolving cell types. Together, the cell-by-cell comparative analysis shows how fine-scale cellular profiling can extract conserved modules from a pan transcriptome and provide insight on the evolution of cells that mediate key functions in crops.

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

Z. mays (maize) is a staple crop and *S. bicolor* (sorghum) is an important dryland crop and biofuel candidate that is closely related to maize, sharing a common ancestor about 12 million years ago[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[8–10]. The importance of the two crops, their evolutionary proximity, and their functional differences present an opportunity for comparative analysis of cellular evolution in plants[11,12]. In addition, 5 to 12 million years ago, subsequent to the shared ancestry with sorghum, maize underwent a whole-genome duplication (WGD), probably following a hybridization[7,13] (allopolypoidy). Comparing patterns of gene expression at the cell level in maize, sorghum and outgroup *S. viridis* (*Setaria*) provides an opportunity to examine cellular evolution and the role of gene duplications, including the paralogous genes generated by the WGD[7,14] (homeologues).

Cells provide depth and nuclei provide breadth

Single-cell analyses in plants have relied on the generation of protoplasts by enzymatic digestion of cell walls[15]. However, certain tissues and some species, including sorghum, are recalcitrant to digestion. There is also historic concern about the effects of protoplast generation on the cellular transcriptome, leading to a growing interest in nuclear profiling[16–18]. To assess the fidelity of nuclear profiling in detail across dicots and monocots, we first compared single-cell profiles (hereafter referred to as cell profiles) with single-nucleus profiles (hereafter referred to as nucleus profiles) in *Arabidopsis thaliana* (a dicot model with plentiful resources, 15,967 cells and 17,373 nuclei) and *Z. mays* (a monocot model, 4,235 cells[19] and 2,668 nuclei) (Supplementary Table 1).

The number of unique molecular indices (UMIs) was ten times (*Arabidopsis*) and six times (*Z. mays*) higher in cell profiles than in nucleus profiles (Extended Data Fig. 1a), similar to animal studies[20]. Accordingly, the average number of genes detected was 2.7 times (*Arabidopsis*) and 1.4 times (*Z. mays*) higher in cell profiles than in nucleus profiles (Extended Data Fig. 1b and Supplementary Table 1). However, despite the lower mRNA content, nucleus profiling detected 89% (*Arabidopsis*) and 88% (*Z. mays*) of the total genes present in cell profiles (Supplementary Table 1).

'Pseudo-bulked' transcriptomes derived from both cell and nucleus profiles displayed a high correlation to whole-root transcriptomes ($r \approx 0.7–0.8$; Extended Data Fig. 1c), confirming that both sampling methods generally reflected the expression patterns in intact tissue.

In both *Arabidopsis* and maize, cell and nucleus profiles enabled the generation of UMAP clusters corresponding to all the major cell identities[21] (Fig. 1a–c and Extended Data Figs. 2 and 3). However, in both species, the nuclear dataset generated fewer distinct clusters and frequently could not distinguish between closely related or subcellular identities (Extended Data Figs. 2 and 3). For example, in maize, stele cells contained a subcluster that we identified as

[1]Center for Genomics and Systems Biology, New York University, New York, NY, USA. [2]Center for Genomics and Systems Biology, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates. [3]Cold Spring Harbor Laboratory, New York, NY, USA. [4]Department of Plant Biology, Michigan State University, East Lansing, MI, USA. [5]Department of Nematology and Center for Plant Cell Biology, Institute for Integrative Genome Biology, University of California, Riverside, CA, USA. [6]Department of Physiology, University of Toronto, Toronto, Ontario, Canada. [7]Present address: UGA-LANGEBIO Cinvestav, Guanajuato, México. ✉e-mail: ken.birnbaum@nyu.edu

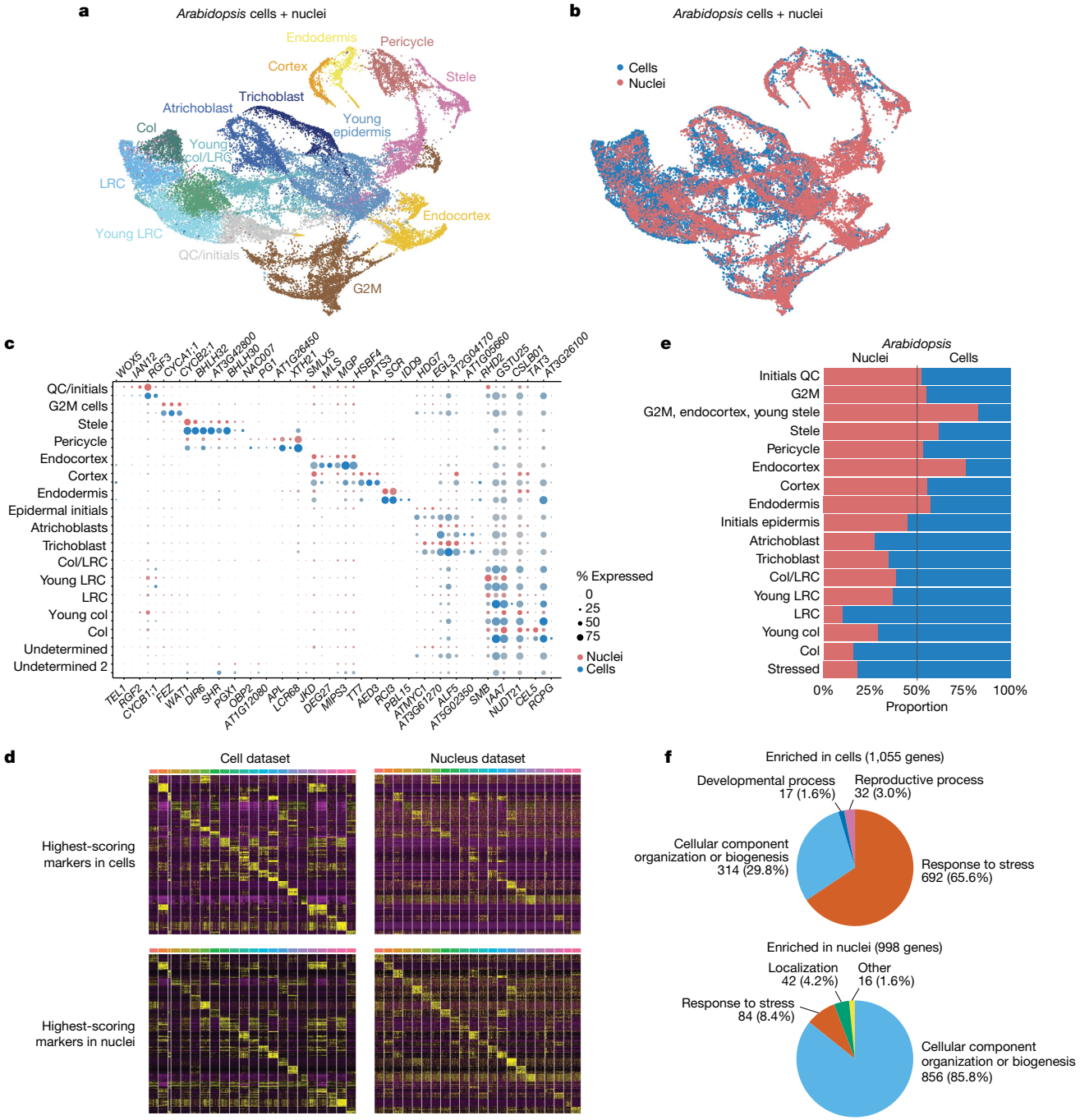


Fig. 1 | Cell and nucleus profiles identify the same markers but show different sensitivities and artifacts. a,**b**, Uniform manifold approximation and projection (UMAP) of combined *Arabidopsis* cell and nucleus transcriptomic profiles, with clusters coloured according to assigned cell identity (**a**) or cell profile versus nucleus profile origin (**b**). **c**, Dot plots of *Arabidopsis* marker genes in cell or nucleus profiles, showing all the cell types defined from clusters in this study. **d**, Heat maps of the 10 highest-scoring marker genes for each cell type found using Seurat (Methods). Top, the highest-scoring markers in the single-cell dataset (left) and their expression in the single-nucleus dataset (right). Bottom, the highest-scoring markers in the single-nucleus dataset (left) and their expression in the single-cell dataset (right). **e**, Proportion of cell profiles versus nucleus profiles present in each cell-type cluster. **f**, The difference in the prevalence of Gene Ontology (GO) terms among differentially expressed genes in each cluster between cell (top) and nucleus (bottom) profiles. Col, columella; LRC, lateral root cap; QC, quiescent centre.

xylem cells, whereas no such subcluster was apparent in the nuclear cluster analysis (Extended Data Fig. 3). Using a down-sampling approach on each dataset, a general rule of thumb emerged whereby twice as many nuclei were needed to discover the same number of clusters as cells or protoplasts (Extended Data Fig. 4a,b). Thus, the shallower depth of nucleus 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.

Combined and independent analysis of single-cell and single-nucleus transcriptomes generated clusters that reflected the same underlying biological patterns (Fig. 1a–c and Extended Data Fig. 4c,d). The highest-scoring markers extracted from nucleus profiles generally matched the highest-scoring ones extracted from cell profiles (Fig. 1c,d

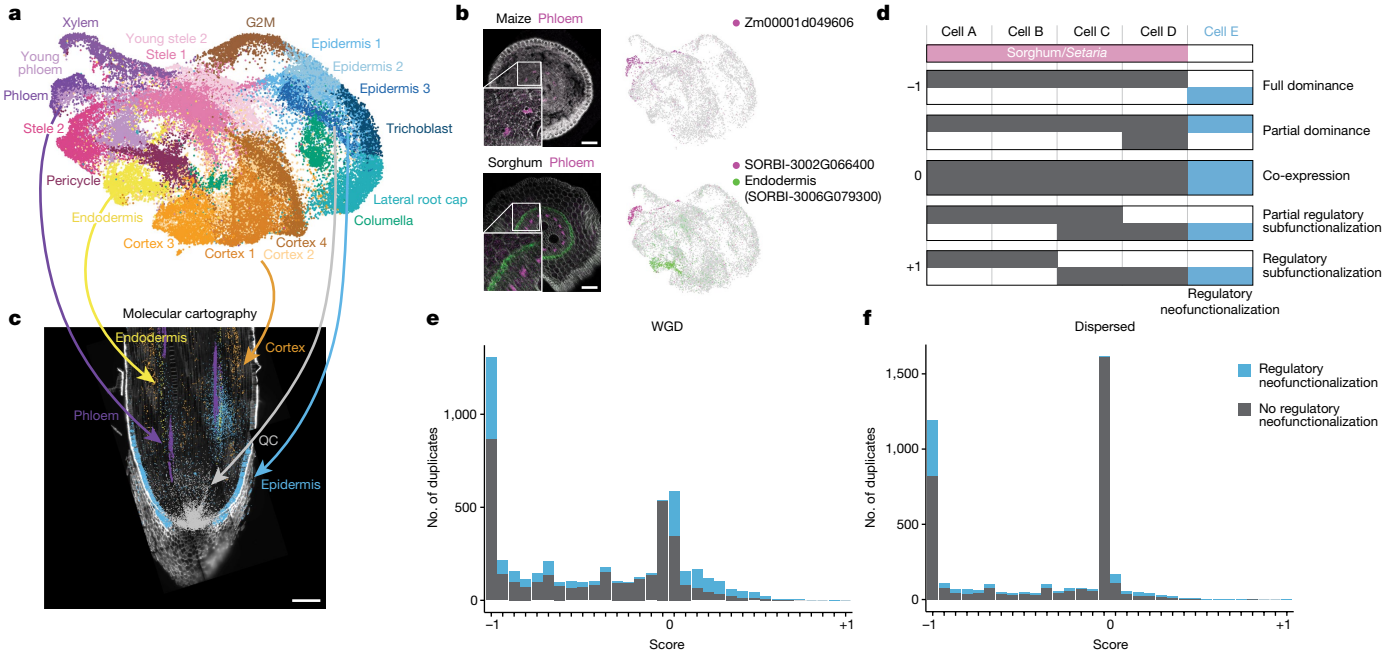


Fig. 2 | Mapping cell identities from maize to sorghum and gene duplicate analysis. a, UMAP of combined maize single-cell and single-nucleus transcriptomic profiles. Clusters are coloured and labelled according to cell identity. **b**, In situ hybridization in maize (top) and sorghum (bottom). The maize phloem marker is orthologous to the sorghum phloem marker. Bottom, green colouration corresponds to a sorghum endodermal marker that highlights the stele boundary. The minimum and 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 the images show the respective expression of each gene in the maize–sorghum co-clustered cell profiles, which were used initially to determine their expression pattern. **c**, Molecular cartography, showing markers used for the cell-cluster annotation of clusters in maize. **d**, Conceptual schematic of hypothetical expression patterns between duplicate gene pairs following a metric on the following scale: full dominance, −1; equal co-expression, 0; regulatory subfunctionalization, 1. Example intermediate states are also shown. Blue shows regulatory neofunctionalization. **e**,**f**, Distribution of duplicate gene expression patterns using the metric described in **d** for WGD homeologues (**e**) and dispersed duplicate pairs (**f**) having a similar median Ks (synonymous substitution rate). Number of genes: WGD homeologues, 10,104; dispersed duplicates, 7,552.

and Extended Data Fig. 4d). In addition, the assignment of cells to specific clusters was stable whether cell and nucleus profiles were clustered alone or together (Supplementary Table 2).

One advantage of nucleus profiles was their ability to capture cells from tissues that are recalcitrant to enzymatic digestion, improving the representation of cell identities (Fig. 1e and Extended Data Fig. 3d). For example, in maize, we detected a unique cluster in nucleus profiling that was not detected in cell profiling, which we confirmed as columella cells using previously published RNA-sequencing (RNA-seq) profiles of hand-sectioned root tissue[19].

In *Arabidopsis*, we found that 14% of total genes (3,218) were differentially expressed between cell profiles and nucleus profiles in a cluster-by-cluster analysis (Supplementary Table 3). Cell profiles showed a higher proportion of stress-related genes (Fig. 1f and Extended Data Fig. 5a,b). A similar analysis in maize, sorghum and *Setaria* also supported a lower stress response in nucleus profiles than in cell profiles (Supplementary Table 3). However, most of the differences between cell and nucleus profiles appeared to be related to compartmental RNA stability. For example, mRNAs enriched in nucleus versus cell profiles significantly overlapped with transcripts shown to have higher decay rates in the cytoplasm[22] ($P = 1.98 \times 10^{-11}$; Extended Data Fig. 5c). We conclude that combining cell and nucleus profiles has the advantage of uncovering cell type-specific protoplast responsive genes, while also providing depth in transcriptional readouts.

Conserved cell-type markers in cereals

Given the comprehensive coverage of a combined analysis, we generated cell and nucleus profiles to investigate cellular evolution in the maize–sorghum–*Setaria* clade. Thus, we generated profiles for sorghum (3,510 cells and 7,620 nuclei) and *Setaria* (10,613 cells and 12,192 nuclei) (Supplementary Table 1). We took advantage of previous comparative genomic sequence analyses in maize, sorghum and *Setaria* that mapped orthologues among the three species, including the homeologues created by WGD in maize[11,14] (hereafter referred to subgenome M1 and M2 to refer to maize's two parental genomes arising from hybridization). We used a set of single-copy orthologues in the three species to cluster all cellular and nuclear profiles together in a single step and then predicted cell identity using known cell type-specific marker genes for maize[19] (Fig. 2a, Supplementary Table 1 and Methods).

To validate the mapping, we: (1) performed an independent MetaNeighbor analysis, which uses neighbour voting to quantify the similarity of cell clusters across datasets using a given marker set of genes and their orthologues; (2) used an additional machine learning-based clustering method, scGen, to confirm the cluster membership[23] (Extended Data Fig. 6); (3) conducted whole-mount in situ hybridizations in maize and sorghum (Fig. 2b and Extended Data Figs. 7 and 8); and (4) performed spatial transcriptomics in maize (Fig. 2c and Extended Data Fig. 7), altogether confirming the maize-to-sorghum-to-*Setaria* mapping of cell identities. Thus, we could use the well-annotated maize cell-type map for rapid generation of a high-confidence cellular-resolution 'pan-transcriptome' of these key crop species, including hundreds of new cell type-specific marker genes (Supplementary Table 4).

One potential use of cell type-specific pan-transcriptome data is to search for highly localized and conserved gene expression modules. We used MINI-EX to identify cell type-specific networks across the three grass species[24]. The analysis revealed 15 transcription factors and putative targets (regulons) that were conserved in specific cell types across all three species (Extended Data Fig. 9a and Supplementary Table 5).

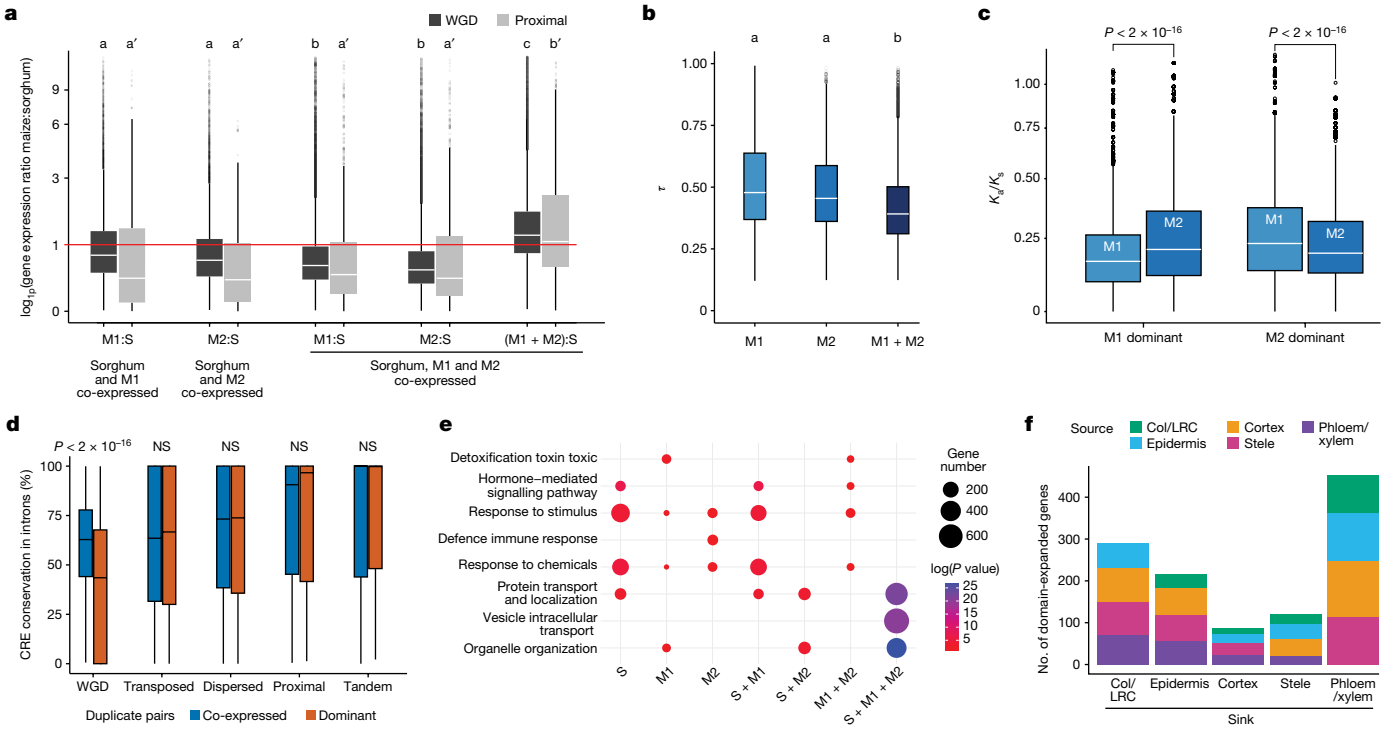


Fig. 3 | Detection of dosage compensation and cellular destination of regulatory neofunctionalized genes. a, Dosage compensation analysis with ratios of orthologous gene expression in maize and sorghum in the two duplication classes. The first two box plots represent cases in which a sorghum orthologue is expressed in the same cell type as a single maize homeologue (either M1 or M2). The third and fourth box plots represent cases in which both homeologues are expressed in the same cells. The fifth box plot shows the ratio when both of the co-expressed homeologues are added in the numerator over sorghum expression level in the denominator. Dosage compensation is inferred from a pattern in which lone expression of a homeologue is higher than expression of co-expressed homeologues. **b**, Tau (τ) value reflecting the degree of cell specificity in different expression categories within a cell, if M1 or M2 is dominant or if M1 and M2 are co-expressed. **c**, Ka (non-synonymous substitution rate)/Ks distribution of WGD homeologues: when either M1 or M2 is dominant in a cell type, the dominant homeologue displays stronger purifying selection than the non-dominant homeologue. **d**, The conservation rate of *Cis*-regulatory elements (CREs) between duplicate pairs in introns split into co-expressed and dominant categories. **e**, GO terms that are enriched within each expression category. S, M1 and M2, unique expression of the sorghum orthologue or one maize homeologue; S + M1 and S + M2, one maize homeologue is expressed in the same cell type as the sorghum orthologue. S + M1 + M2, both homeologues are expressed in the same cell type as the sorghum orthologue. **f**, Regulatory neofunctionalized genes categorized by their new expression domains. Colours within a bar show the ancestral cell-type domain (Methods). In **a**–**d**: $n = 10{,}104$ WGD, $n = 860$ proximal, $n = 3{,}154$ transposed, $n = 7{,}552$ dispersed, $n = 1{,}448$ tandem. **a**,**b**, One-way ANOVA followed by Tukey test for all pairwise comparisons. Groups that do not share a letter are significantly different from each other, unmarked letters refer to tests between WGD categories described in **a** and letters with prime symbol refer to tests between proximal duplicate categories ($P < 0.05$). **c**, Two-sided Wilcoxon test. **d**, Two-sided Wilcoxon signed-rank test, with Benjamini and Hochberg adjusted *P* value. In box plots, the centre line is the median, the top and bottom hinges correspond to the first and third quartiles, respectively, whiskers extend to quartiles ±1.5 × interquartile range and dots show potential outliers. NS, not significant.

In 5 out of the 15 cases, mutants in predicted transcription factors or direct *Arabidopsis* orthologues have been shown to exhibit cell type-specific phenotypes corresponding to the conserved regulon localization[25–29]. These results highlight the ability of comparative cell-type analyses to reveal conserved cellular mechanisms across species and connect specific genes to specific cellular functions.

The effect of maize WGD on cellular identity

The cellular map across species also provided the opportunity to examine how homologous cell types have diverged over the millions of years since the three species split. We first focused on the effects of gene duplication, comparing homeologues from the WGD to several other duplicate classes not identified as within WGD segments: gene pairs that arose from tandem, transposon-mediated, proximal (separated by up to ten genes) and dispersed (separated by more than ten genes) duplicate pairs[11] (Methods).

We used concordance between sorghum and *Setaria* to infer ancestral expression domains for each duplicate gene pair. We then developed a simple metric to represent the degree of overlap versus complementarity in cellular domains between duplicate pairs, ranging from consistently higher expression of one homeologue (dominance), to co-expression and then to regulatory subfunctionalization of homeologue pair expression[30,31] (Fig. 2d). We then identified duplicated genes that expanded their expression domain to new cell types compared with ancestral domains[32,33] (regulatory neofunctionalization; Fig. 2d and Methods). We note that we cannot determine whether differences in gene expression between duplicated genes occurred in the parent genomes or, more probably, after WGD[13,32,34]. Here we use the terms neo- and subfunctionalization to refer strictly to patterns in transcriptional domains at the cell-type level.

Overall, WGD homeologues made a more prevalent contribution to expression domain expansion (neofunctionalization) than other classes of duplicates. This was because they included a relatively low proportion of the co-expressed category, which showed no neofunctionalization (Fig. 2e,f and Extended Data Fig. 9b–d). Rather, WGD homeologues were enriched in dominance and subfunctionalized categories, both of which exhibited high levels of neofunctionalization in new cell types (Fig. 2e,f and Extended Data Fig. 9b–d). This trend did not appear to be driven by the age of the duplication as other duplicate classes had similar mean Ks values to WGD[35] (Methods, Extended Data Fig. 9b–h).

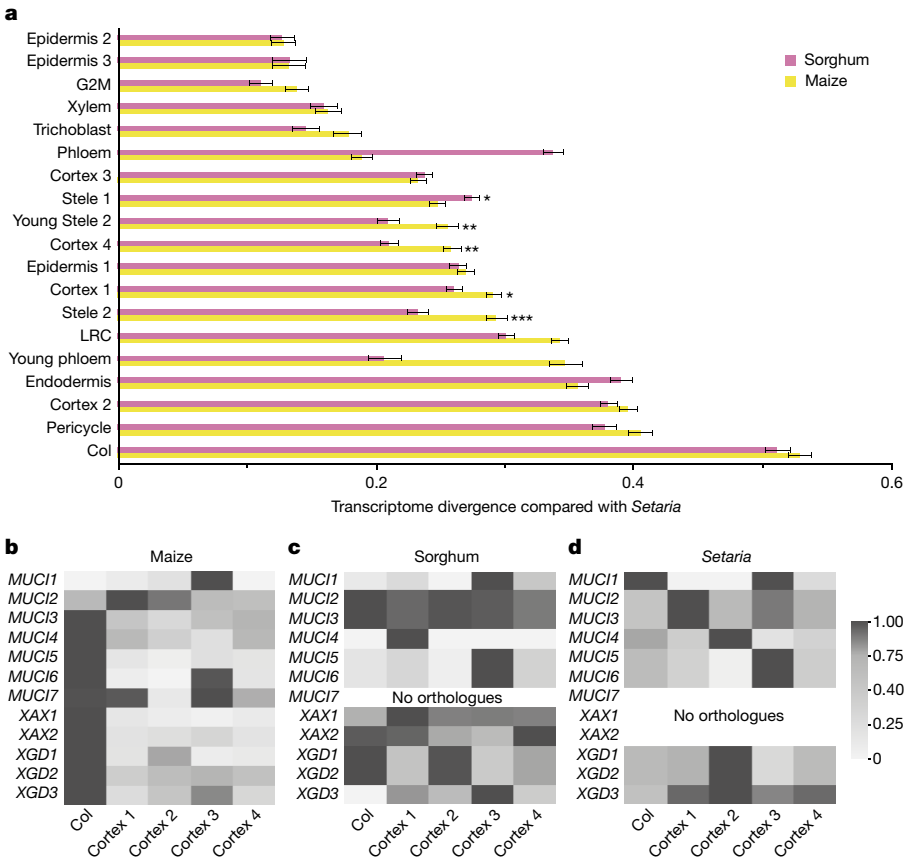


Fig. 4 | Differential divergence of cell types in maize compared with *Setaria*. **a**, MetaNeighbor analysis showing a quantification of transcriptome divergence among cell types in maize and sorghum compared with the outgroup *Setaria*. Two-sided Hanley McNeil test. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. Error bars indicate standard error. **b**,**c**, Heat maps showing expression of mucilage genes in maize (**b**), sorghum (**c**) and *Setaria* (**d**) columella cells and cortex layers.

In keeping with genome balance models, we observed that co-expressed WGD homeologues showed expression patterns indicative of dosage compensation[36,37], whereas this pattern was weaker or non-existent in other duplicate classes (Fig. 3a and Extended Data Fig. 10a–c).

In addition, 66% of all regulatory neofunctionalization cases in the WGD came from the dominance category, with a slightly higher proportion from the M1 subgenome[14,38] (Supplementary Table 6). Furthermore, dominant homeologues showed significantly higher cell-type specificity (τ) than co-expressed homeologues (Fig. 3b and Methods). Together, these trends meant that gene pairs that exhibited dominance patterns after WGD made the largest contribution to the transcriptional divergence of cell types.

As found in previous studies[34,39], dominant members of a homeologue pair showed greater purifying selection (Fig. 3c). In addition, we found that homeologues in the WGD class showed a marked decrease in the conservation of intronic *cis*-regulatory sites between the dominant and non-dominant homeologue compared with homeologues in the co-expressed class—a feature that is not observed in other duplicate classes or in promoters (Fig. 3d, Extended Data Fig. 10d and Supplementary Table 6). This could represent a possible loss of intron-mediated expression enhancement in the non-dominant homeologue. These two genomic features are consistent with previous findings that suggest that dominant homeologues may have retained ancestral gene functions[34,39], whereas non-dominant homeologues may adopt new functions or become pseudogenes.

However, pseudogenization appears to be a less likely possibility. When we analysed the same duplicate homeologue pairs in single-cell profiles of the maize inflorescence[40], we found that a subset (32%) of non-dominant homeologues in the root were instead dominant in cells of the inflorescence (Supplementary Table 6). Together, the relaxed purifying selection and the switch in dominance suggests that non-dominant homeologues may specialize in a subset of developmental contexts outside the root.

The dominance group showed an enrichment for GO term annotations related to immunity and response to stimulus or stress, even after removal of all potential protoplast-induced genes (Fig. 3e, Supplementary Table 7 and Methods). Thus, new cellular gene expression driven largely by WGD may contribute to tolerance to environmental stress, either constitutively or under our experimental conditions.

In addition, although subfunctionalization of cell-type domains between homeologue pairs was a minor outcome, this category of homeologues showed the highest rate of neofunctionalization (59%) compared with any other duplicate class (for example, in Fig. 2e,f and Extended Data Fig. 9b–d). The trend is consistent with models in which subfunctionalization is a transitory state that facilitates neofunctionalization[41]. Ultimately, 34% of all the neofunctionalized homeologues (that is, those with new cell-type expression after the WGD) came from the subfunctionalized category. Thus, although subfunctionalization via adopting complementary expression domains was relatively rare, it appeared to provide a high-probability route to cell-type domain expansion (neofunctionalization). This propensity for neofunctionalization made the subfunctionalized gene pair category a second major contributor to cellular divergence.

Finally, certain cell types appeared to be more likely domain-expansion destinations than others (Fig. 3f). The trends were similar for all duplicate classes, with the specialized vascular cells and root cap cells most frequently comprising the new expression domains.

Cortex was the least frequent sink for new domains, although it was one of the most frequent source domains (Fig. 3f and Extended data Fig. 10e–h). Overall, the data show how gene duplication, particularly WGD, frequently provides genetic material for the transcriptional divergence of specific cell types.

Root slime drives cellular divergence

To explore cellular divergence more broadly, we next examined the entire transcriptome of each cell cluster to determine which cell types were most substantially changed in maize and sorghum compared with the outgroup *Setaria*. For all comparative analyses, we combined cell and nucleus profile datasets, using MetaNeighbor to compare cell identities across species (Fig. 4a).

The analysis showed that in both maize and sorghum, the transcriptomes of columella, LRC, cortex subcluster 2, endodermis, pericycle and stele cell types are more divergent than other cell types when compared with *Setaria* (Fig. 4a). This shared divergence suggests that the function of these tissues diverged from *Setaria* before the maize–sorghum split. In addition, certain cell types—such as cortex subcluster 1 and 4, and several stele clusters—were significantly diverged between maize and sorghum, implying additional divergence after the maize–sorghum split. We note that the fast-evolving cell types were largely consistent with the sink tissues favoured for neofunctionalization by duplicate genes (compare Fig. 4a with Fig. 3f). Of note, 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[42] (Supplementary Table 8). We identified 443 genes displaying high expression divergence across species in columella cells. Many of these genes showed marked changes in cell-type localization between species, such as *DOWNY MILDEW RESISTANT 6* (*DMR6*), which is expressed in columella and epidermis in maize and in cortex and endodermis in sorghum (Extended Data Fig. 10i,j).

GO term analysis of the cortex-to-columella orthologues in maize showed enrichment in enzymes leading to the synthesis of mannose, raffinose and oligosaccharides (Supplementary Table 8). These sugars and carbohydrates are key components of mucilage (also known as slime), which can be secreted from many different cell types of the root and has multiple roles, such as the shaping of the root-associated microbiome and lubricating the root–soil interface[8,43–45].

We then examined all genes implicated in mucilage synthesis[8,9,46], finding the same general pattern of cortical expression in sorghum and *Setaria* and columella expression in maize (Fig. 4b–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 a putatively ancestral expression pattern in the cortex. The most parsimonious model is that the recruitment of the mucilage module occurred before the maize WGD, as both maize homeologues in the mucilage-annotated genes tended to share expression in the columella. However, the set of mucilage genes showed a significant overlap with genes previously defined as under selection during domestication[47] (Supplementary Table 8), suggesting that they have a role in agricultural traits.

Previous studies in animals have shown co-option of gene modules from one cell type to another as a mechanism of cellular diversification[48]. We tested how frequently gene expression modules, such as the mucilage group, switched cellular localization by focusing on regulons that have different cell type-specific expression patterns in maize compared with sorghum and/or *Setaria* (swapped regulons). Although annotated regulons comprise only a subset of all potential downstream targets of transcription factors, we identified more than 50 swapped modules across cell types. The swapped modules are prime candidates for genes that could mediate differences in cellular traits between maize and related species (Supplementary Table 5).

Overall, we identify two major trends in cellular divergence over a taxonomic span of 50 million years[49]. First, after WGD duplication, gene pairs in which one homeologue shows expression dominance have the strongest role in cell type-specific divergence. However, the rare class of subfunctionalized genes have the most likely evolutionary route to neofunctionalization. Second, homologous cell types appear to diverge in part by swapping gene expression modules[48], such as the mucilage genes found to be expressed in the maize columella. Finally, we illustrate how single-cell techniques can rapidly generate a pan-transcriptome to yield insights into plant cell-type evolution and provide new methods to explore the connection between genetic modules and cellular traits in important crop species.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-023-06053-0.

1. Woodhouse, M. R. & Hufford, M. B. Parallelism and convergence in post-domestication adaptation in cereal grasses. *Philos. Trans. R. Soc. B* **374**, 20180245 (2019).
2. Rich-Griffin, C. et al. Single-cell transcriptomics: a high-resolution avenue for plant functional genomics. *Trends Plant Sci.* **25**, 186–197 (2020).
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. 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**, 10956 (2020).
9. Ma, W. et al. The mucilage proteome of maize (*Zea mays* L.) primary roots. *J. Proteome Res.* **9**, 2968–2976 (2010).
10. Schittenhelm, S. & Schroetter, S. Comparison of drought tolerance of maize, sweet sorghum and sorghum–sudangrass hybrids. *J. Agron. Crop Sci.* **200**, 46–53 (2014).
11. Zhang, Y. et al. Differentially regulated orthologs in sorghum and the subgenomes of maize. *Plant Cell* **29**, 1938–1951 (2017).
12. Zheng, Z. et al. Shared genetic control of root system architecture between *Zea mays* and *Sorghum bicolor*. *Plant Physiol.* **182**, 977–991 (2020).
13. McKain, M. R. et al. Ancestry of the two subgenomes of maize. Preprint at *BioRxiv* https://doi.org/10.1101/352351 (2018).
14. 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. USA* **108**, 4069–4074 (2011).
15. Bawa, G., Liu, Z., Yu, X., Qin, A. & Sun, X. Single-cell RNA sequencing for plant research: insights and possible benefits. *Int. J. Mol. Sci.* **23**, 4497 (2022).
16. 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).
17. Long, Y. et al. FlsnRNA-seq: protoplasting-free full-length single-nucleus RNA profiling in plants. *Genome Biol.* **22**, 66 (2021).
18. 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).
19. Ortiz-Ramírez, C. et al. Ground tissue circuitry regulates organ complexity in maize and *Setaria*. *Science* **374**, 1247–1252 (2021).
20. Ding, J. et al. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat. Biotechnol.* **38**, 737–746 (2020).
21. Ray F. Evert. in *Esau's Plant Anatomy, Meristems, Cells, and Tissues of the Plant Body: their Structure, Function, and Development* 3rd edn 99 (Wiley, 2006).
22. 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. USA* **115**, E1485–E1494 (2018).
23. Lotfollahi, M., Wolf, F. A. & Theis, F. J. scGen predicts single-cell perturbation responses. *Nat. Methods* **16**, 715–721 (2019).
24. Ferrari, C., Manosalva Pérez, N. & Vandepoele, K. MINI-EX: integrative inference of single-cell gene regulatory networks in plants. *Mol. Plant* **15**, 1807–1824 (2022).
25. Donner, T. J., Sherr, I. & Scarpella, E. Regulation of preprocambial cell state acquisition by auxin signaling in *Arabidopsis* leaves. *Development* **136**, 3235–3246 (2009).

26. Wang, S. et al. RppM, encoding a typical CC-NBS-LRR protein, confers resistance to southern corn rust in maize. *Front. Plant Sci.* **13**, 951318 (2022).
27. Ingram, G. C., Magnard, J. L., Vergne, P., Dumas, C. & Rogowsky, P. M. *ZmOCL1*, an HDGL2 family homeobox gene, is expressed in the outer cell layer throughout maize development. *Plant Mol. Biol.* **40**, 343–354 (1999).
28. Li, Z., Tang, J., Srivastava, R., Bassham, D. C. & Howell, S. H. The transcription factor bZIP60 links the unfolded protein response to the heat stress response in maize. *Plant Cell* **32**, 3559–3575 (2020).
29. Guo, Z. et al. MRG1/2 histone methylation readers and HD2C histone deacetylase associate in repression of the florigen gene *FT* to set a proper flowering time in response to day-length changes. *New Phytol.* **227**, 1453–1466 (2020).
30. Grover, C. E. et al. Homoeolog expression bias and expression level dominance in allopolyploids. *New Phytol.* **196**, 966–971 (2012).
31. Lynch, M. & Force, A. The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154**, 459–473 (2000).
32. Chaudhary, B. et al. Reciprocal silencing, transcriptional bias and functional divergence of homeologs in polyploid cotton (*Gossypium*). *Genetics* **182**, 503–517 (2009).
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. Zhao, M., Zhang, B., Lisch, D. & Ma, J. Patterns and consequences of subgenome differentiation provide insights into the nature of paleopolyploidy in plants. *Plant Cell* **29**, 2974–2994 (2017).
35. Li, L. et al. Co-expression network analysis of duplicate genes in maize (*Zea mays* L.) reveals no subgenome bias. *BMC Genomics* **17**, 1–16 (2016).
36. Birchler, J. A. & Veitia, R. A. Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. *Proc. Natl Acad. Sci. USA* **109**, 14746–14753 (2012).
37. 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* **377**, 20210222 (2022).
38. Walsh, J. R., Woodhouse, M. R., Andorf, C. M. & Sen, T. Z. Tissue-specific gene expression and protein abundance patterns are associated with fractionation bias in maize. *BMC Plant Biol.* **20**, 4 (2020).
39. Renny-Byfield, S., Rodgers-Melnick, E. & Ross-Ibarra, J. Gene fractionation and function in the ancient subgenomes of maize. *Mol. Biol. Evol.* **34**, 1825–1832 (2017).
40. 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).
41. Rastogi, S. & Liberles, D. A. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evol. Biol.* **5**, 28 (2005).
42. 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).
43. Van Deynze, A. et al. Nitrogen fixation in a landrace of maize is supported by a mucilage-associated diazotrophic microbiota. *PLoS Biol.* **16**, e2006352 (2018).
44. 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).
45. 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).
46. 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**, 2481–2495 (2015).
47. Wang, B. et al. Genome-wide selection and genetic improvement during modern maize breeding. *Nat. Genet.* **52**, 565–571 (2020).
48. Arendt, D. The evolution of cell types in animals: emerging principles from molecular studies. *Nat. Rev. Genet.* **9**, 868–882 (2008).
49. Wang, X. et al. Genome alignment spanning major poaceae lineages reveals heterogeneous evolutionary rates and alters inferred dates for key evolutionary events. *Mol. Plant* **8**, 885–898 (2015).

Methods

Plant growth conditions

Seeds of *A. thaliana* Col-0, *Z. mays* B73, *S. bicolor* Btx623, *S. viridis* A10.1 and PI 669942 (US National Plant Germplasm System) were used in this study. *Arabidopsis* seeds were imbibed for 48 h at 4 °C before being surface-sterilized and placed on a nylon mesh (110 µm) within plates containing agar with 0.5× Murashige and Skoog salts (Sigma M5524), 0.5% sucrose, and 0.8% Agar (Sigma A1296). Plants 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 150 µm from the tip, and directly transferred to either the protoplast solution at room temperature or the nuclei lysis buffer at 4 °C.

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

Protoplast generation

Protoplasts were generated from primary and seminal roots as described previously[50]. For maize, sorghum and *Setaria*, roots were cut above the meristem as described above and placed in pretreatment solution containing L-cysteine for 40 min (3% sorbitol, 2.5 mM L-cysteine, 20 mM 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% corresponding to 400 mM, MES 20 mM, KCl 20 mM, $CaCl_2$ 40 mM, pH 5.8 with Tris, 100 µg ml^{-1} BSA; 2% cellulase 'Onozuka' RS, 1.2% cellulase 'Onozuka' R10, 0.4% macerozyme R-10 (all three Yakult Pharmaceutical); and 0.36% pectolyase Y-23 (MP Biomedicals)). Protoplasts 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%, 400 mM, MES 20 mM, KCl 20 mM, $CaCl_2$ 40 mM, pH 5.8 with Tris, BSA 100 µg ml^{-1}, 1.2% cellulase 'Onozuka' R10, 0.4% macerozyme R-10 (both Yakult Pharmaceutical)). Protoplasts were then filtered through a 20-µm cell strainer and transferred to microcentrifuge tubes for centrifugation.

Protoplasts were centrifuged for 3 min at 500*g* and the pellets were washed and resuspended in washing solution twice (mannitol 8%, MES 20 mM, KCl 20 mM, $CaCl_2$ 10 mM, pH 5.8 with Tris, and BSA 100 µg ml^{-1}) and used immediately for single-cell RNA-seq. An aliquot of protoplasts was stained with trypan blue (0.2% final) and checked on a haematocytometer under the microscope to determine cell viability and concentration before loading into the 10x Chromium.

Nuclei extraction

For all species, root tips were directly transferred to pre-chilled lysis buffer (0.3M sucrose, 15 mM Tris HCl at pH 8, 60 mM KCl, 15 mM NaCl, 2 mM EDTA, 0.5 mM Spermine, 0.5 mM Spermidine, 15 mM MES, 0.1% Triton, 5 mM DTT*, 1 mM PMSF*, 1% Plant Protease Inhibitors* 1 ml (Sigma P9599), BSA* 0.4%, RNase inhibitor* 0.2 µg $µl^{-1}$; asterisks indicate reagents added at the last minute). Roots were chopped on ice with scalpel blades for 5–10 min and transferred into a pre-chilled dounce homogenizer (Kimble, 885302). The pestle was moved up and down 10 times, samples were then kept on ice for 10 min before an additional 10 strokes with the pestle. Root extracts were filtered at 20 µm into a centrifuge tube and centrifuged for 10 min at 500*g* (maize, sorghum, and *Setaria*) or at 1,000*g* (*Arabidopsis*). Pellets were washed once with washing buffer (0.3 M sucrose, 15 mM Tris HCl at pH 8, 60 mM KCl, 15 mM NaCl, 0.5 mM Spermine, 0.5 mM Spermidine, 15 mM MES, 5 mM DTT*, 1 mM PMSF*, 1% Plant Protease Inhibitors* 1 ml (Sigma P9599), BSA* 0.4%, RNase inhibitor* 0.2 U $µl^{-1}$; asterisks indicate reagents added at the last minute). Finally, nuclei were resuspended into a final buffer (0.3 M sucrose, 15 mM Tris HCl at pH 8, 60 mM KCl, 15 mM NaCl, 0.5 mM Spermine, 0.5 mM Spermidine, 15 mM MES, 5 mM DTT*, 1% Plant Protease Inhibitors* 1 ml (Sigma P9599), BSA* 0.4%, RNase inhibitor* 0.2 U $µl^{-1}$; asterisks indicate reagents added at the last minute) and filtered using a 10-µm filter. An aliquot of nuclei was stained with DAPI for quality control and nuclei were counted under the microscope. Nuclei were used immediately for single-nucleus RNA-seq.

Single-cell RNA-seq

Sixteen-thousand cells or nuclei per replicate were loaded in a Single Cell B Chip (10x Genomics). 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 1×150 high-output chip (2 libraries per chip) or Novaseq 6000 chip SP V2.5 (4 libraries per chip). Raw single-cell RNA-seq data were analysed by Cell Ranger 5.0.1 (10x Genomics) to generate gene–cell matrices. Gene reads were aligned to the *Arabidopsis* TAIR10.38, Maize B73 v4, *S. bicolor* v3 and *S. viridis* v2 reference genomes.

UMAP

Replicates (see Supplementary Table 1) were integrated and cells mapped using the Seurat package v4.0[51] 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 library quality (see Supplementary Table 1). Clustering of cells or nuclei separately were done by log-normalized raw counts and the 2,000 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) was performed. The top 30 principal components were selected. Cells or nuclei were clustered using a *k*-nearest neighbour graph, which is based on the Euclidean distance in PCA space. The FindNeighbors and FindClusters functions with a resolution of 0.5. were applied. Next, non-linear dimensional reduction was performed using the UMAP algorithm with the top 30 principal components.

For the co-clustering of cells and nuclei, the datasets were treated similarly, all replicates were integrated at once using the seurat 'SCT' approach[52]. First, raw reads were normalized using the SCTransform function, then SelectIntegrationFeatures was used to identify anchors between the datasets, using 3,000 features.

For multiple species clustering, all orthologous genes names from ref. 11 were replaced by their corresponding maize ID in sorghum and *Setaria* raw features.tsv.gz files (gene conversion in Supplementary Table 1). Anchors are combined using PrepSCTIntegration and selected using FindIntegrationAnchors. For clustering of maize, sorghum and *Setaria* together, all species were considered equally using the FindIntegrationAnchors function. Finally, a PCA was performed using the first 100 principal components and a non-linear dimensional reduction was performed using the UMAP algorithm with the top 100 principal components.

Identification of WGD and non-WGD one-to-one gene duplicate pairs

We used prior studies to obtain a list of WGD homeologues in the maize1 and maize2 genomes[11,14]. To identify the other types of duplicated genes, DIAMOND v2.0.6 was used to perform blastp for the target genome (*Z. mays)* with itself, and the outgroup genome (*Amborella trichopoda*), retaining BLAST hits with e-value <10^{-5}. These BLAST hits were filtered to remove hits from different orthogroups as described[53]. Duplicate gene pairs were called using DupGen_finder.pl and DupGen_finder-unique.pl (https://github.com/qiao-xin/DupGen_finder) with the following parameters: -s 5 (requiring ≥ 5 genes to call a collinear block) -d 10 (≤ 10 intervening genes to call 'proximal' duplicates). Duplicate pairs are derived from five types of gene duplication, including whole-genome and four types of single-gene duplication: tandem, proximal, translocated and dispersed duplication. A custom R script was used to retain duplicate pairs with the lowest e-value to avoid over-counting pairs within gene families. Further, to filter out pericentric paralogues that are unlikely to be expressed, duplicate pairs where one of the paralogues was missing methylation information was removed, retaining only those pairs where both paralogues had methylation data (Supplementary Table 6). This procedure identified duplicates that were either not a part of the WGD (for example, in genome segments that were not retained) or duplicated after the WGD. It also filters out many ancient duplications whose one-to-one relationship becomes obscured over time.

GO term analysis

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

Cis-regulatory element prediction

Cis-regulatory element were predicted using the Meme suite FIMO algorythm v5.5.1 (https://meme-suite.org/meme/tools/fimo) on 500 bp in the promoters or introns. Maize transcription factor binding sites database used in FIMO was downloaded from http://plantregmap.gao-lab.org.

Gene expression analysis across species

Whole-root transcriptomes were obtained from ref. 19 for maize and ref. 54 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 taken from a previous report[53]. Low, mid and high Ks values were calculated from WGD Ks distribution using the 1/3 quartiles. τ was calculated as described[55], τ = sum of $i = 1$ to N $((1 - x_i)/(N - 1))$, where N is the total number of cell types and x_i is the expression profile component normalized by the maximal component value.

MetaNeighbor cell-type validation across species

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 used the MetaNeighbor package in Python (https://github.com/gillislab/pyMN)[56,57]. 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 labelled 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 (for example, technology) predict the identities of cell clusters in the other fraction of the data. This generates pairwise areas under the receiver operating characteristic (AUROCs) for each combination of clusters. To generate the heat maps, the PyMN.plotMetaNeighborUS was used with a brown–blue–green colour map. This plots the pairwise AUROCs generated previously.

For Fig. 4a, to generate *P* values for evaluating the significance of the differences between each pair of AUROCs generated by MetaNeighbor, we used the two-sided Hanley McNeil test, which produces a *Z*-score for the difference[58]. 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 used the scipy.stats.norm.sf function in Python to convert the *Z*-scores into *P* values for a two-sided test. For error bars on the AUROC in Fig. 4a, we calculated the standard deviation on the estimate of the AUROC, thus, a measure of the error in the mean standardized rank of the positives, so we term that measure of variability standard error.

Validation of integration using scGEN

To evaluate the integration of nuclei and cells across three plant species, we repeated the integration using the supervised integration method scGEN[23]. We used scGEN version 2.1 to train a model using the scgen.train function, and the scgen.model.batch_removal function to correct our data. Following correction, we used the ScanpyV1.9[59] calculate the nearest neighbours using scanpy.pp.neighbors, and generated a 2D projection using UMAP, via sc.tl.umap. We then used sc.tl.leiden clustering algorithm at a 0.6 resolution to identify clusters, which we evaluated for mixing and accuracy of integration.

Identification of single-cell regulatory networks using MINI-EX

We used MINI-EX, a pipeline specialized for inferring cell type-specific gene regulatory networks in plants[24] to identify the gene regulatory networks in our samples. As gene regulatory network inference is dependent upon datasets containing transcription factors and binding sites not available in sorghum and *Setaria*, we used maize transcription factors with one-to-one matches to sorghum and *Setaria* genes for those species. This converted list of transcription factors was used as the TF_list parameter in the miniex.config file. We ran the MINI-EX pipeline using the default parameters but modified it to run on 32 CPU cores.

Co-expression conservation between maize sub-genomes and sorghum

To generate co-expression conservation scores between the two maize sub-genomes and the sorghum genome (Supplementary Table 8), we used our existing aggregated co-expression networks[42]. In brief, these networks are built by taking all publicly available data and calculating average correlations between gene 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 the maize sub-genomes and the sorghum genome. 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[60]. We calculated 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 was then done again in the reverse direction, and the two scores were averaged (calculated as an AUROC). We then selected genes with the lowest co-expression scores (0.34 < FC.Score) and highest cell specificity ($\tau > 0.8$) in the root cap (Extended Data Fig. 10i and Supplementary Table 8).

Formulation of a dominance–co-expression-regulatory subfunctionalization metric

To calculate the Dominance versus regulatory subfunctionalization score, for each orthologue triplet (S, M1 or 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 and *Setaria* orthologue was expressed. We defined dominance if the average expression of one of the two duplicate is two time superior as the average expression of the other duplicate in the same cell type. Co-expression was defined when both duplicates were expressed in the same cell type and their respective expression was below a twofold range difference. Regulatory subfunctionalization was defined when both duplicates are dominant in different cell types. Regulatory neofunctionalization was defined when one or both duplicates are expressed in cell type in which the sorghum and *Setaria* orthologue were not expressed. In this dataset, a gene is defined as expressed if its expression is above the first quartile among genes detected in that cell type, this is necessary to normalize for cell-type quality (certain cell types display more UMI and more gene detected per cells than others). The procedure also helps remove the background of genes with very low expression that results from noise generated by combining cells and nuclei together. The score is given by the expression

Score = (number of cells in which M1 is dominant
× number of cells in which M2 is dominant)
− (number of cells of the dominant orthologue
− number of cells of the nondominant orthologue)

If the score is negative, the score is normalized by NormScore = Score/(no. of cells in which M1 and M2 are expressed). If the score is positive, the score is normalized by the NormScore = Score/(no. of cells in which M1 and M2 are expressed × 0.5)[2].

Cell-type marker identification

Marker genes for each species were identified using FindAllmarkers functions from Seurat (using 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 Fig. 2e and Extended Data Figs. 4c and 10a, statistical analysis was performed in R using a pairwise Wilcoxon test with p.adjust method BH, as the data are not normally distributed.

Correlation analysis in Extended Data Fig. 1c was performed using Pearson correlation function in R between whole-root data coming from single cell or single nuclei. In brief, averaged gene expression was calculated for each gene while combining every cell type using the AverageExpression function from Seurat.

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) were obtained from Molecular Instruments (Supplementary Table 9), with modifications to the hybridization protocol for plant tissue and for the half-mount technique[61].

For fixation, germination paper containing 7-day old maize or sorghum roots were unrolled and small volume of fixative FAA (4% formaldehyde, 5% glacial acetic acid, 50% ethanol in RNAse free water) was pipetted onto each root. Then longitudinal sectioning of root tips was performed using a 15° microscalpel. Roots were cut up to ~3 cm from the tip, then immediately fixed by transferring to FAA in 5-ml screw caps and put under vacuum several times until they no longer floated. Roots were then agitated at room temperature for at least 1 h in a tube revolver (all washes in the protocol were performed in a tube revolver or stated otherwise.)

Samples were dehydrated in a series of washes at room temperature: 70% ethanol for 15 min, 90% ethanol for 15 min, 100% ethanol 2 times for 15 min each, 100% methanol 2 times for 15 min each. Samples can then be stored at −20 °C for several weeks. Samples were washed 2 times for 15 min in 100% ethanol at room temperature before being permeabilized for 30 min in 50% Histo-Clear II:50% ethanol at room temperature. Then they were incubated 2 times for 30 min in a solution of 100% Histo-Clear II at room temperature. Each time, vacuum was applied for the first 10 min.

Samples were rehydrated through a series of washes: 50% Histo-Clear II/50% ethanol for 15 min, 100% ethanol for 15 min, 50% ethanol/50% DPBS-T (0.1% Tween-20, 1× DPBS) for 15 min (roots will float up then settle after a few minutes), 100% DPBS-T 2 times for 15 min (roots will float up again). Samples were 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 room temperature 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 min (roots should turn a little yellow after this step). Samples were washed 2 times for 15 min in DPBS-T at room temperature then incubated with Fixative II (4% formaldehyde in DPBS-T) under gentle vacuum for 10 min then in a tube revolver for 30 min at room temperature. They were then washed 2 times for 15 min each in DPBS-T at room temperature. Roots were aliquoted into 2 ml Eppendorf tubes and incubated in 500 µl of HCR probe hybridization buffer, vacuum was applied for 10 min then roots were incubated for 1 h at 37 °C in a thermomixer with agitation (1,000 rpm).

Samples can then be stored in probe hybridization buffer at −20 °C up to several weeks.

Probe buffers were made by adding 0.8 pmol of each probe set (for example, 2 µl of the 1 µM stock) to 500 µl of HCR probe hybridization buffer at 37 °C. Pre-hybridization solution was removed and replaced with probe solution. Samples were hybridized by incubating overnight (~20 h) at 37 °C in a thermomixer with agitation (1,000 rpm). The following day, excess probes were removed by washing 4 times for 15 min each with 1 ml of HCR Probe Wash Buffer at 37 °C in a thermomixer with agitation. Samples were washed 2 times for 5 min each with 1 ml of 5× SSC-T (25% 20× SSC, 0.1% Tween-20) at room temperature in a thermomixer with agitation. SSC-T was replaced with 500 µl of amplification buffer, gentle vacuum was applied in a fume hood for 10 min and then samples were pre-amplified by incubating in a tube rotator at room temperature for 50 min. While samples pre-amplify, 6 pmol of hairpin h1 and 6 pmol of hairpin h2 (that is, 5 µl of the 3 µM stocks) were prepared, each in its own separate tube. Hairpins were snap-cooled by heating at 95 °C for 90 s then kept in a dark drawer at room temperature for 30 min. Amplification solution was prepared by combining snap-cooled h1 and h2 hairpins in 250 µl of HCR amplification buffer at room temperature. Pre-amplification solution was removed and and replaced with amplification buffer containing hairpin solution overnight (~20 h) in the dark at room temperature in a thermomixer with agitation (1,000 rpm). Excess hairpins were removed by washing with 1 ml of 5× SSC-T at room temperature in a thermomixer with agitation, 2× for 5 min each, then 2× for 30 min each, 1× for 5 min. Samples were transferred onto a glass slide (in 5× SSC-T) and cut using a 30° microscalpel and arranged so that the cut face of the roots faced upwards. They were then covered with a coverslip and imaged on a confocal microscope.

Statistics and reproducibility

For HCR RNA-FISH experiments, all replicates are biological replicates.

Figure 2b: 1 experiment: transverse, 2 strong, 1 weak; longitudinal, 2 strong, 4 weak.

Extended Data Fig. 7a: 5 experiments: 7 outer cap, 11 transverse, 32 longitudinal—all consistent. Extended Data Fig. 7c: 1 experiment: transverse, 4 moderate signal; longitudinal, 5 moderate signal. Extended Data Fig. 7d: 1 experiment: transverse, 2 no signal, the rest moderate to strong; longitudinal, 1 too high, 5 moderate to strong. Extended Data Fig. 7e: 1 experiment: transverse, 4 weak, 11 none; longitudinal, 2 weak. Extended Data Fig. 7f: 1 experiment: transverse, 2 strong,

1 weak; longitudinal 2 strong, 4 weak. Extended Data Fig. 7g: 3 experiments: transverse, 2 weak, 1 very weak, 1 no signal; longitudinal, 2 weak, 8 no signal. Extended Data Fig. 7h: 2 experiments: transverse, 1 weak; longitudinal, 4 weak, 5 no signal. Extended Data Fig. 7i: 1 experiment: transverse, 4 moderate; longitudinal, 1 moderate, 1 no signal. Extended Data Fig. 7j: 1 experiment: outer cap, 2 weak; transverse, 2 weak, 3 no signal; longitudinal, 3 weak. Extended Data Fig. 7k: 1 experiment: transverse, 4 weak; longitudinal, 5 weak. Extended Data Fig. 7l: 1 experiment: outer cap, 1 weak; longitudinal, 3 weak, 1 imaged too low. Extended Data Fig. 7m: 1 experiment: transverse, 5 moderate; longitudinal, 3 moderate, 1 no signal. Extended Data Fig. 7n: 3 experiments: outer cap, 7 strong; transverse, 3 strong, 1 no signal; longitudinal, 25 strong.

Extended Data Fig. 8a: 4 experiments: 2 outer; 5 transverse; 20 longitudinal—all consistent. Extended Data Fig. 8c: 2 experiments; transverse, 3 strong, 2 moderate, 1 weak, 1 no signal; longitudinal, 6 strong, 2 moderate, 4 weak, 11 none. Extended Data Fig. 8d: 3 experiments: transverse, 7 strong, 2 no signal; longitudinal, 7 strong, 1 moderate, 5 imaged too low, 1 none. Extended Data Fig. 8e: 1 experiment: transverse, 3 weak; longitudinal, 3 weak. Extended Data Fig. 8f: 1 experiment: transverse, 3 no signal; longitudinal, 5 weak. Extended Data Fig. 8g: 1 experiment: outer, 1 moderate; longitudinal, 4 moderate. Extended Data Fig. 8h: 1 experiment: transverse, 1 very weak, 2 no signal; longitudinal: 2 weak, 2 no signal. Extended Data Fig. 8i: 2 experiments: transverse, 4 strong; longitudinal, 8 strong, 7 imaged too low.

Spatial transcriptomics

Tissue fixation and embedding was performed as described previously[62].

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 pretreatment and priming In preparation for hybridization. the coverslip was removed and the mounting reagent was washed twice in 1× PBS for 30 min 4 °C, followed by 1-min washes in 50% ehanol 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 labelled in a two-step procedure. Regions of interest were imaged as described below and fluorescent signals were removed after imaging via a decolourization procedure. Colour development, imaging and decolourization 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 was based on 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 by Resolve BioSciences on a Zeiss Cell-discoverer 7, using the 50× Plan Apochromat water immersion objective with an NA of 1.2 and the 0.5× magnification changer, resulting in a 25× 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 1,000 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 (0.5×). 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 neighbouring 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 two 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 neighbouring 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.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All reference genomes were downloaded from *Arabidopsis* TAIR10.38, at https://www.arabidopsis.org/, for Maize B73 v4, *S. bicolor* v3 and *S. viridis* v2 reference genomes at https://plants.ensembl.org/. All raw single-cell and single-nucleus RNA-seq data, expression matrices and analysed R-Seurat objects are available under Gene Expression Omnibus accession GSE225118. All data used to generate figures are available at https://figshare.com/articles/dataset/Data_for_Guillotin_et_al_/22331002, except for the following figures, for which the data can be found in the following deposited files under GEO accession GSE225118: Arabidopsis_Cells_Nuclei_Seurat_Obj.RData.gz (Fig. 1c and Extended Data Figs. 2c,d and 4a,b), Maize_Sorghum_Setaria_Cells_Nuclei_Seurat_Obj.RData.gz (Extended Data Figs. 3d and 5c,d). Data in Extended Data Figs. 2c,d and 3d, and those in Extended Data Fig. 5c,d are clustered separately. Data on single-cell RNA-seq quality control are provided in Supplementary Table 1. Analysis of single-cell versus single-nucleus RNA-seq data is provided in Supplementary Tables 2 and 3. Cell-specific marker genes for all species, including a shared pan library of marker genes, are provided in Supplementary Table 4. Data on regulon analysis are provided in Supplementary Table 5. Data on duplicate genes are provided in Supplementary Tables 6 and 7. Cellular divergence analysis is provided in Supplementary Table 8 and in situ probe information is provided in Supplementary Table 9.

50. 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).
51. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 e21 (2019).
52. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **20**, 296 (2019).
53. Raju, S. K. K., Ledford, S. M. & Niederhuth, C. E. DNA methylation signatures of duplicate gene evolution in angiosperms. *Plant Physiol.* kiad220 (2023).
54. Hernández-Coronado, M. et al. Plant glutamate receptors mediate a bet-hedging strategy between regeneration and defense. *Dev. Cell* **57**, 451–465.e6 (2022).
55. Yanai, I. et al. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650–659 (2005).
56. 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).
57. Fischer, S., Crow, M., Harris, B. D. & Gillis, J. Scaling up reproducible research for single-cell transcriptomics using MetaNeighbor. *Nat. Protoc.* **16**, 4031–4067 (2021).
58. Hanley, J. A. & McNeil, B. J. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* **148**, 839–843 (1983).
59. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
60. Crow, M., Suresh, H., Lee, J. & Gillis, J. Coexpression reveals conserved gene programs that co-vary with cell type across kingdoms. *Nucleic Acids Res.* **50**, 4302–4314 (2022).
61. Huang, T., Guillotin, B., Rahni, R., Birnbaum, K. & Wagner, D. A rapid and sensitive multiplex, whole mount RNA fluorescence in situ hybridization and immunohistochemistry protocol. Preprint at *bioRxiv* https://doi.org/10.1101/2023.03.09.531900 (2023).
62. 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).

Acknowledgements The authors thank M. Purugganan and G. 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., Human Frontiers of Science (LT000972/2018-L) to B.G. and 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).

Author contributions B.G. and K.D.B designed the research. B.G. generated all single-cell and single-nucleus RNA-seq data, with early profiles performed by C.O.R. 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. R.R., 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. S.K.R. performed the non-WGD duplication identification. M.P and J.G. performed the MetaNeighbor, MINI-EX, CoCoCoNet and validation analysis. B.G. analysed all the data. K.D.B., B.G. and R.R. wrote the manuscript.

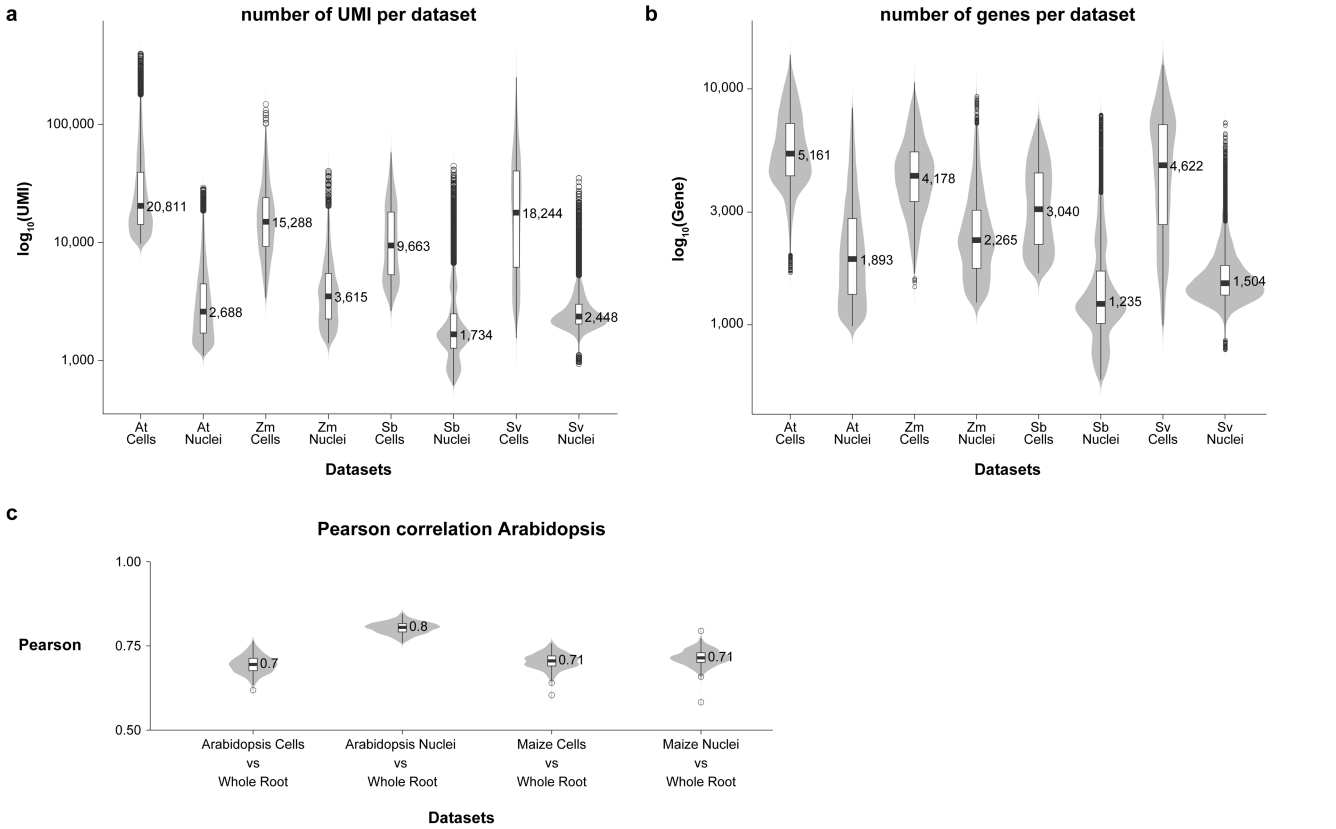
Competing interests The authors declare no competing interests.

Additional information

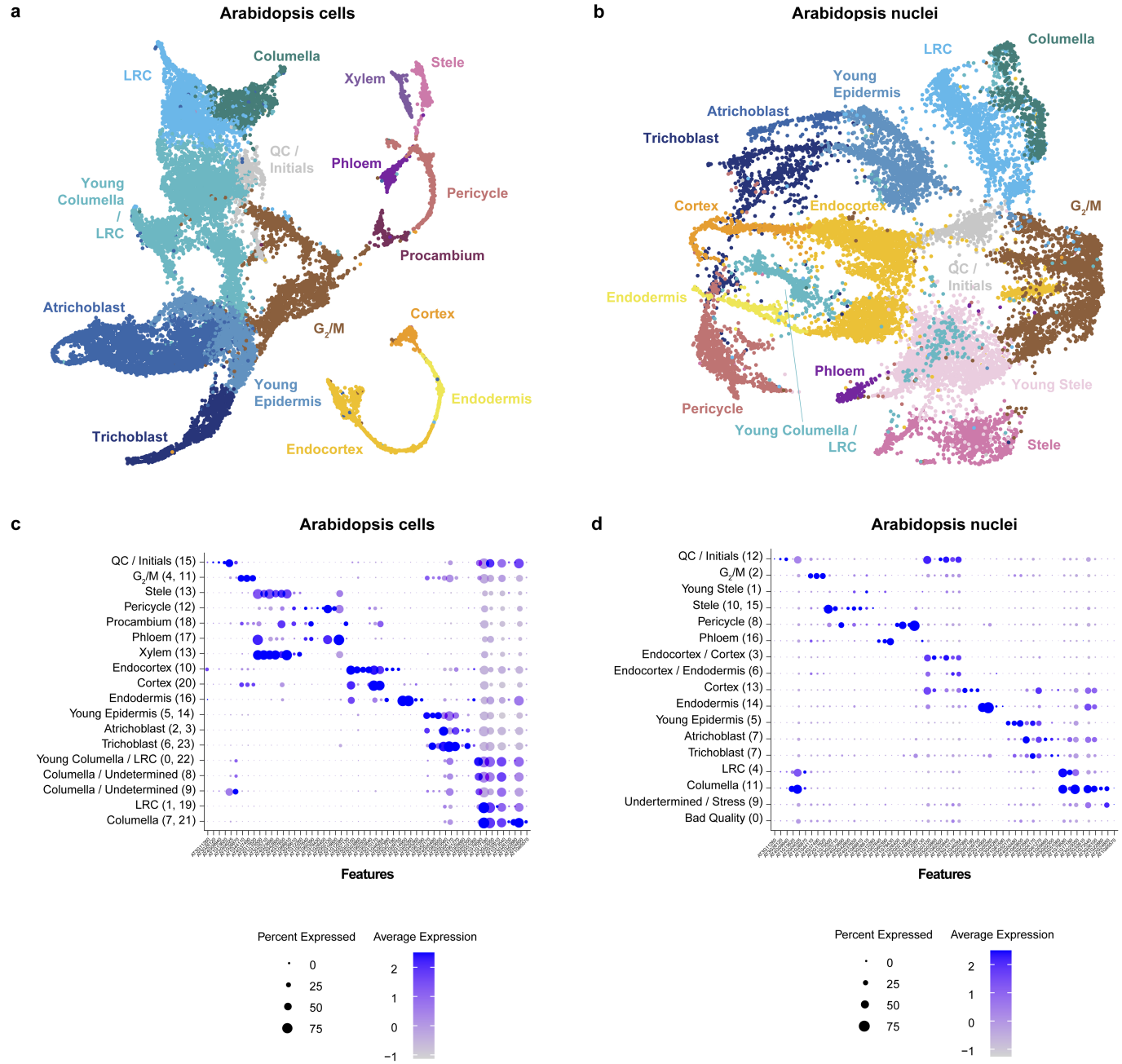
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-023-06053-0.

Correspondence and requests for materials should be addressed to Kenneth D. Birnbaum.

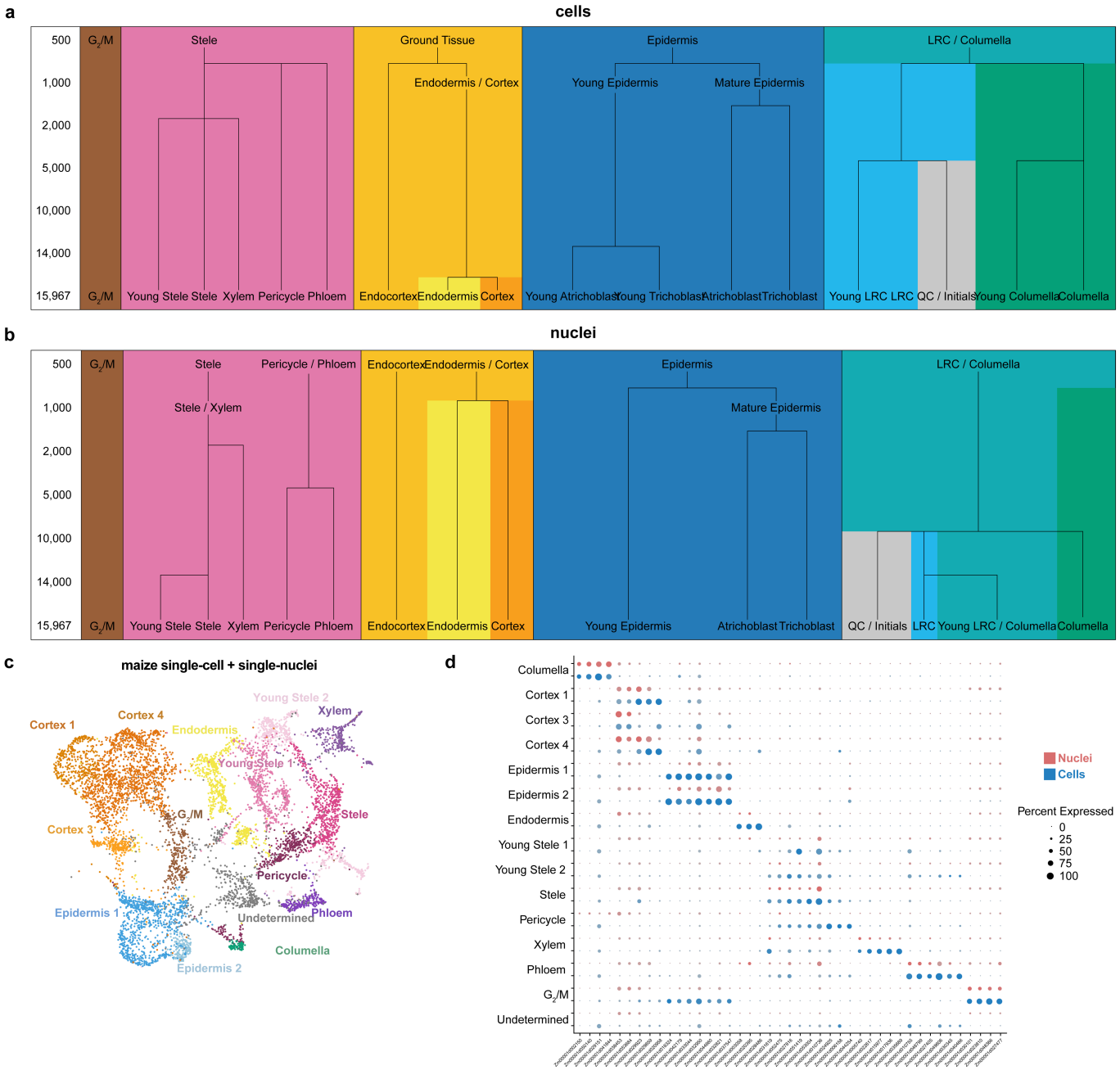
Peer review information *Nature* thanks Patrick Edger, James Schnable and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer review reports are available.



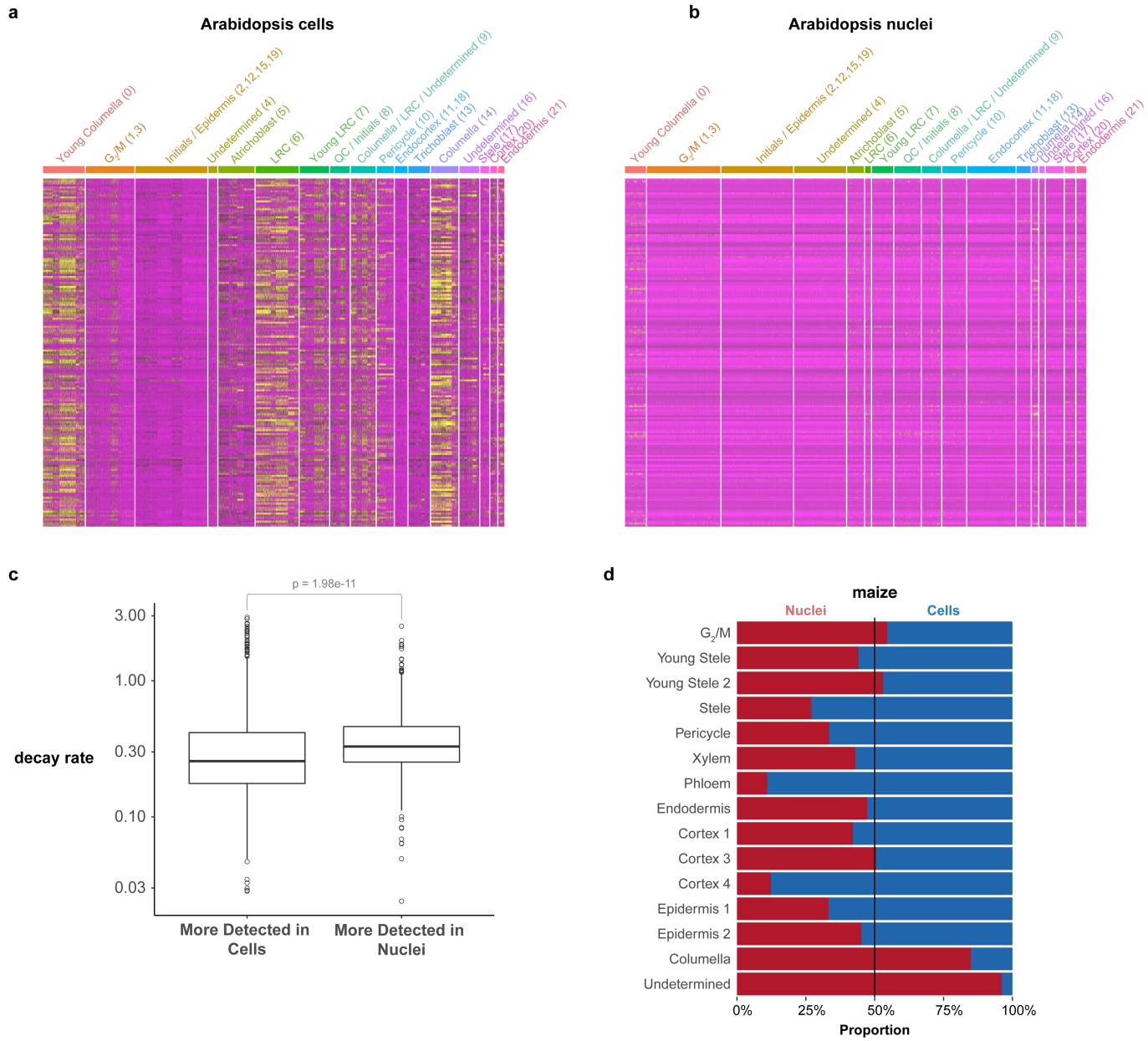
Extended Data Fig. 1 | Quality control and fidelity analysis of RNA-seq profiles using violin plots. a Distribution of the number of UMI detected among cells vs. nuclei. **b** Distribution of the number of genes detected among cells vs. nuclei. **c** Pearson correlation distributions of gene expression from single-cell or single-nucleus compared to whole-root RNAseq data in Arabidopsis and maize. The distributions are derived by randomly sampling 2,000 genes for correlation analysis between cells and nuclei. The random sampling was repeated 250 times to generate the distribution of correlation values. Violin plots display show the kernel probability density of the data at different values, boxplot inside display as the middle black line is the median, exact media is displayed on the graphs, the lower and upper hinges correspond to the first and third quartiles (Q1,Q3), extreme line shows Q3+1.5xIQR to Q1-1.5xIQR (interquartile range-IQR). Dots beyond the extreme lines shows potential outliers.



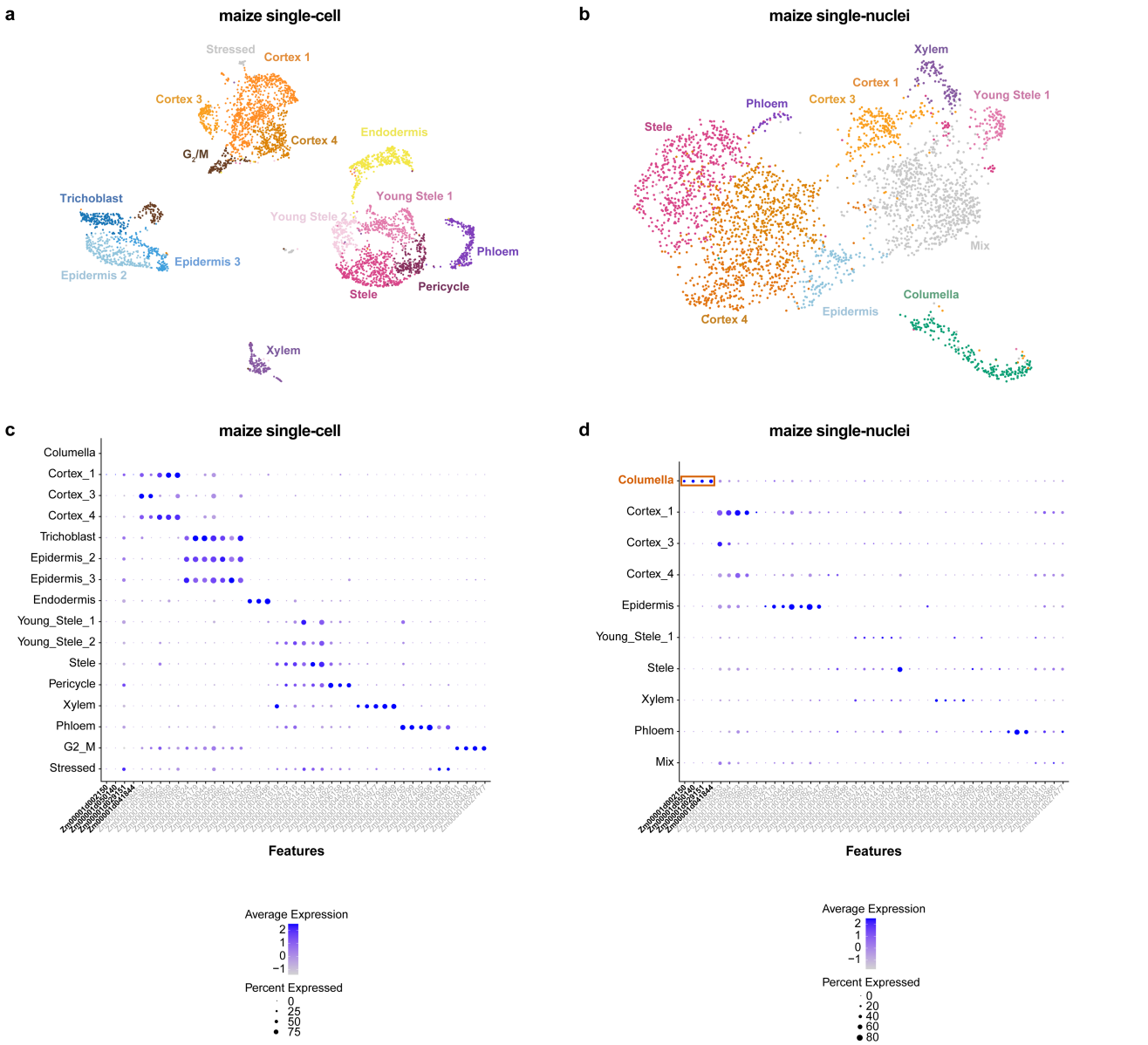
Extended Data Fig. 2 | Evaluation of agreement in nuclear and cell type profiles. a, b UMAP clustering in Arabidopsis single-cells (**a**) and single-nuclei (**b**) clustered independently, showing clusters with the same diagnosed cell identities. **c, d** Dot plots showing expression levels per cluster and expression in percent of cells of the same set of cell-type specific markers in cells (c) or nuclei (d). The markers are in the same order in both plots.



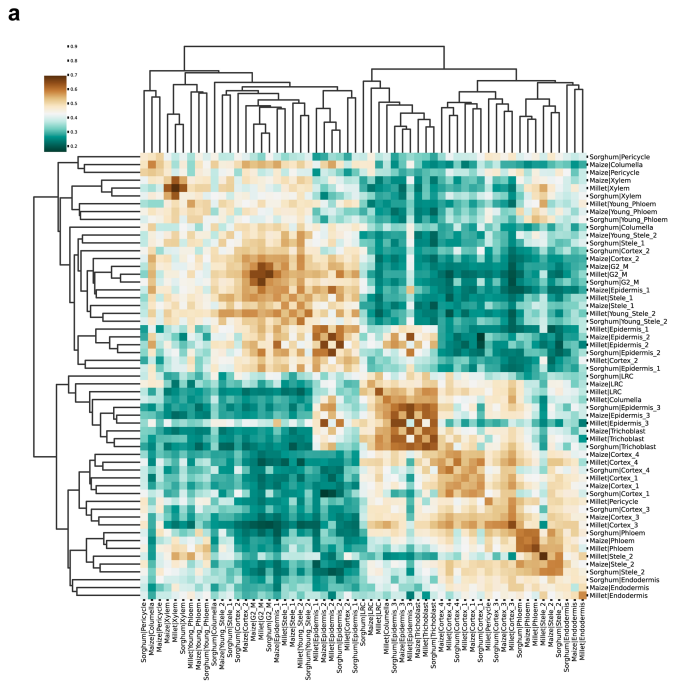
Extended Data Fig. 3 | Analysis of sensitivity of nuclear and cell profiles in distinguishing clusters and identifying markers. a Arabidopsis down sampling analysis shows the number of cells needed to resolve different clusters. A branch signifies that a new cluster with a known cell type identity was distinguished at a given sample size. **b** A similar analysis using the single nucleus RNA-seq dataset, showing that more nuclei are needed to resolve the same number of clusters compared to cells 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** UMAP of the combined maize single-cell and -nuclei datasets, clusters are colored by cell type identity. **d** Dotplot of maize marker genes in cells (blue) or in nuclei (red), showing overall concordance of marker gene expression in the two datasets.

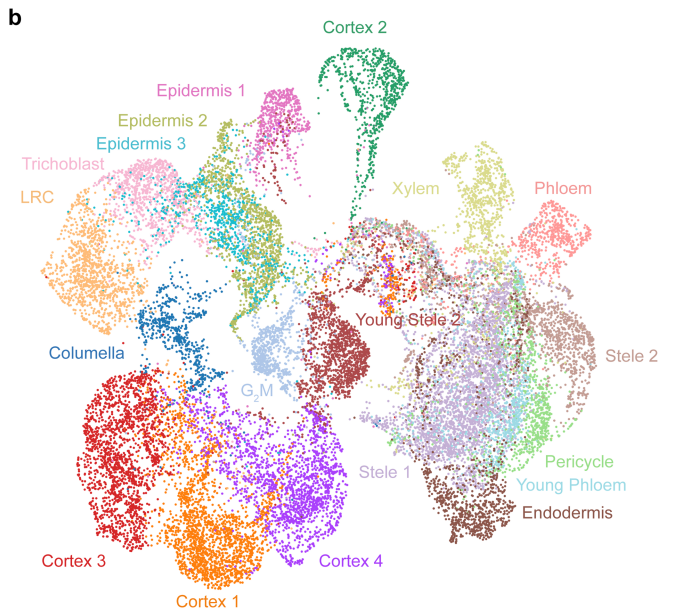


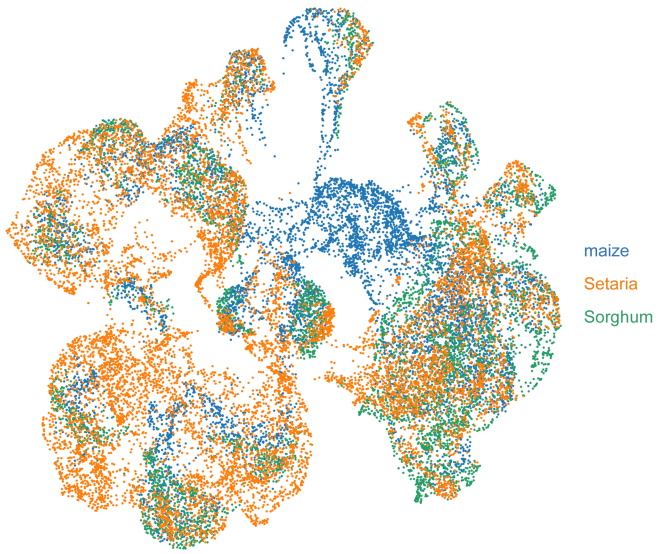
Extended Data Fig. 4 | Analysis of differentially regulated genes and cell capture efficiency in nuclear vs. cellular profiles. a, b Heatmaps of genes known to be induced by protoplast generation (Birnbaum et al., 2003) showing their expression in cells (**a**) vs. nuclei (**b**). The analysis shows that stress-induced genes also have higher expression in cells vs. nuclei, with a bias in specific cell types. **c** Distribution of expression levels of genes annotated for mRNA decay in cells or in nuclei, decay values from Sorenson et al., 2018. A significant increase in expression of mRNA decay-related genes was detected in nuclei, (n = 1965 genes, Wilcoxon rank sum test, two-sided, p-value = 1.98e-11), the boxplots display the middle line is the median, the lower and upper hinges correspond to the first and third quartiles (Q1,Q3), extreme line shows Q3+1.5xIQR to Q1-1.5xIQR (interquartile range-IQR). Dots beyond the extreme lines shows potential outliers. **d** Proportion of cells vs nuclei present in each cell type cluster.



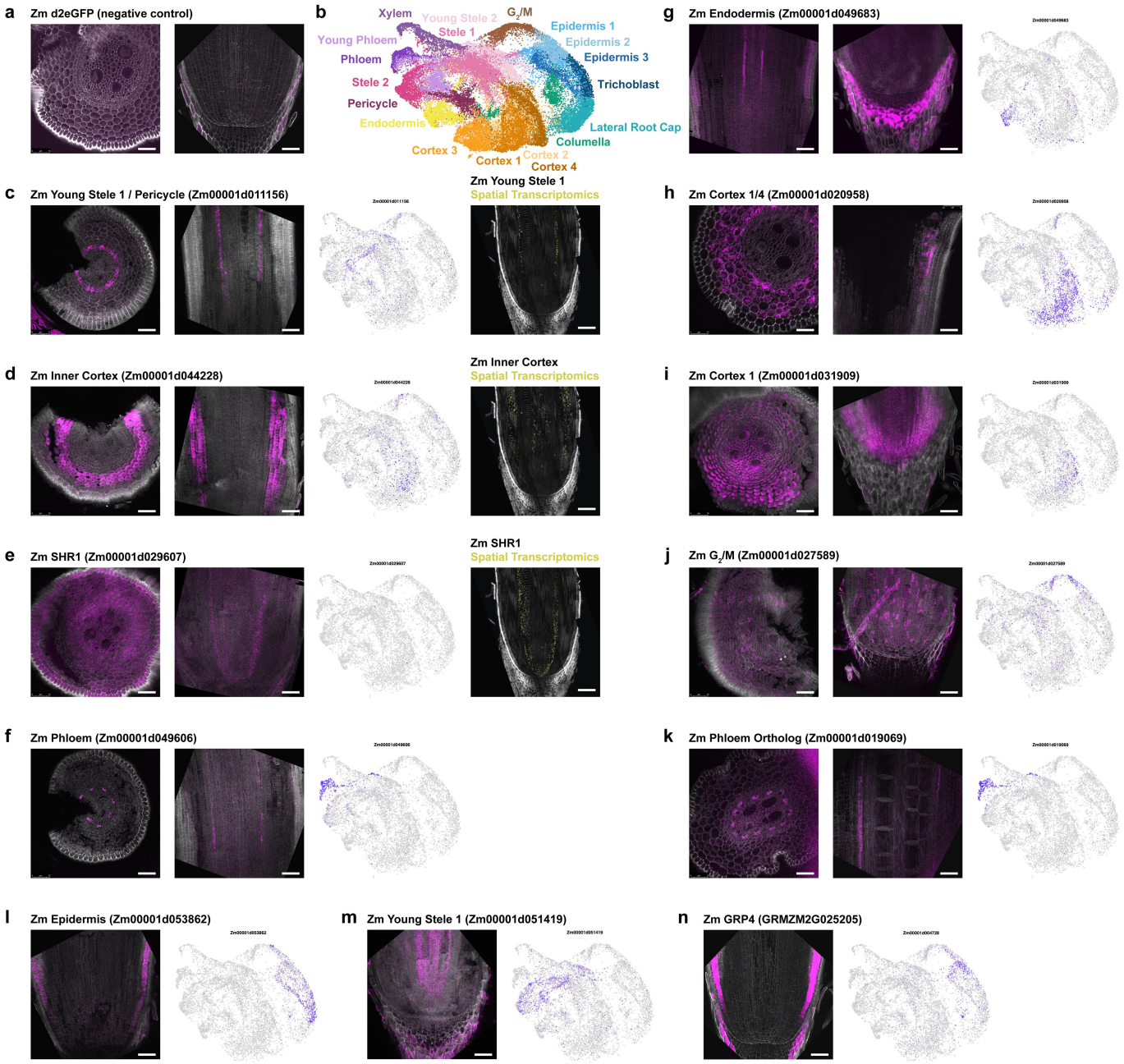
Extended Data Fig. 5 | Analysis of marker gene identification in maize single nucleus vs. cell profiles. a, b UMAPs of maize single-cell and single-nucleus RNA-seq data clustered independently. Only the single nucleus RNA-seq dataset displays a cluster annotated as columella, which is absent in the single-cell dataset. **c**, **d** Dotplot 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 single nucleus dataset.



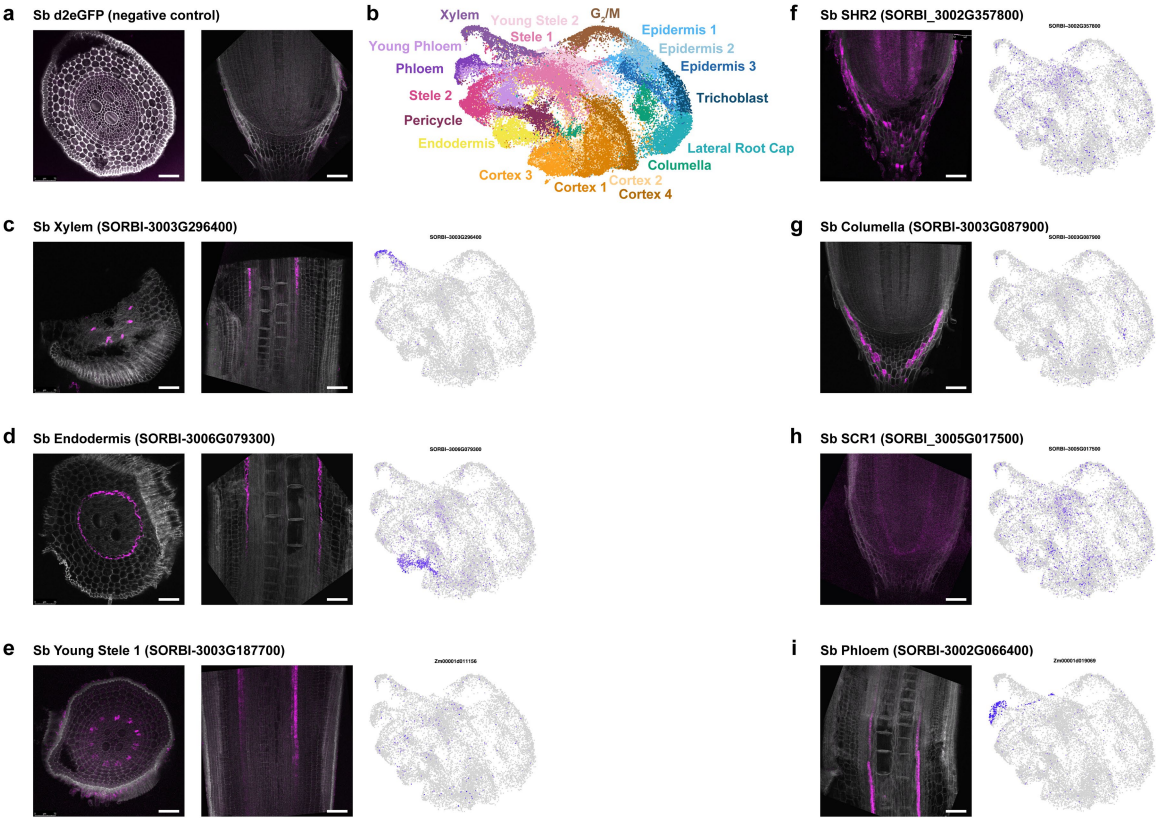




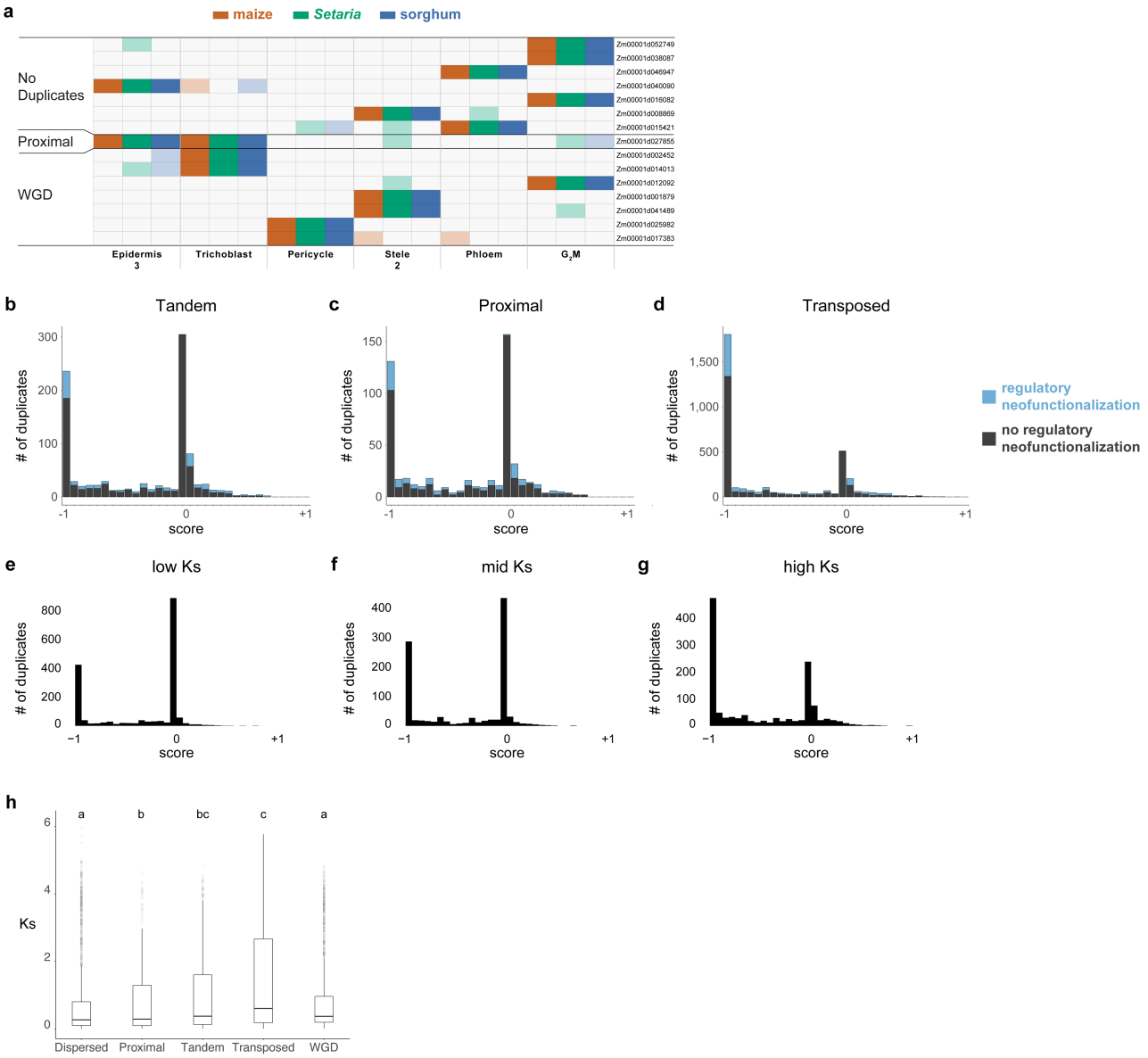
Extended Data Fig. 6 | Analysis of overall expression similarity among all cellular and nuclear clusters in the three monocot species studied. **a** AUROC test comparing every cell type in all species for both cell and nuclei datasets, showing that clusters discovered in either cell or nuclei group by like cell type and not by either species or source of material (cells or nuclei). **b**–**c** UMAPs generated by additional integration of the dataset using a Python supervised integration method scGen. This method uses a variational autoencoder to learn the underlying latent space for the cell types. **b** Different colors represent the clusters identified by the Seurat integration mapped onto the new scGen integration, showing Seurat classification was in relative agreement with the scGen classification. i.e., scGEN clusters have relatively homogenous coloration. **c** The same UMAP as in (**b**), this time showing the species distribution. Overall, each cluster has cells from each of the three species.



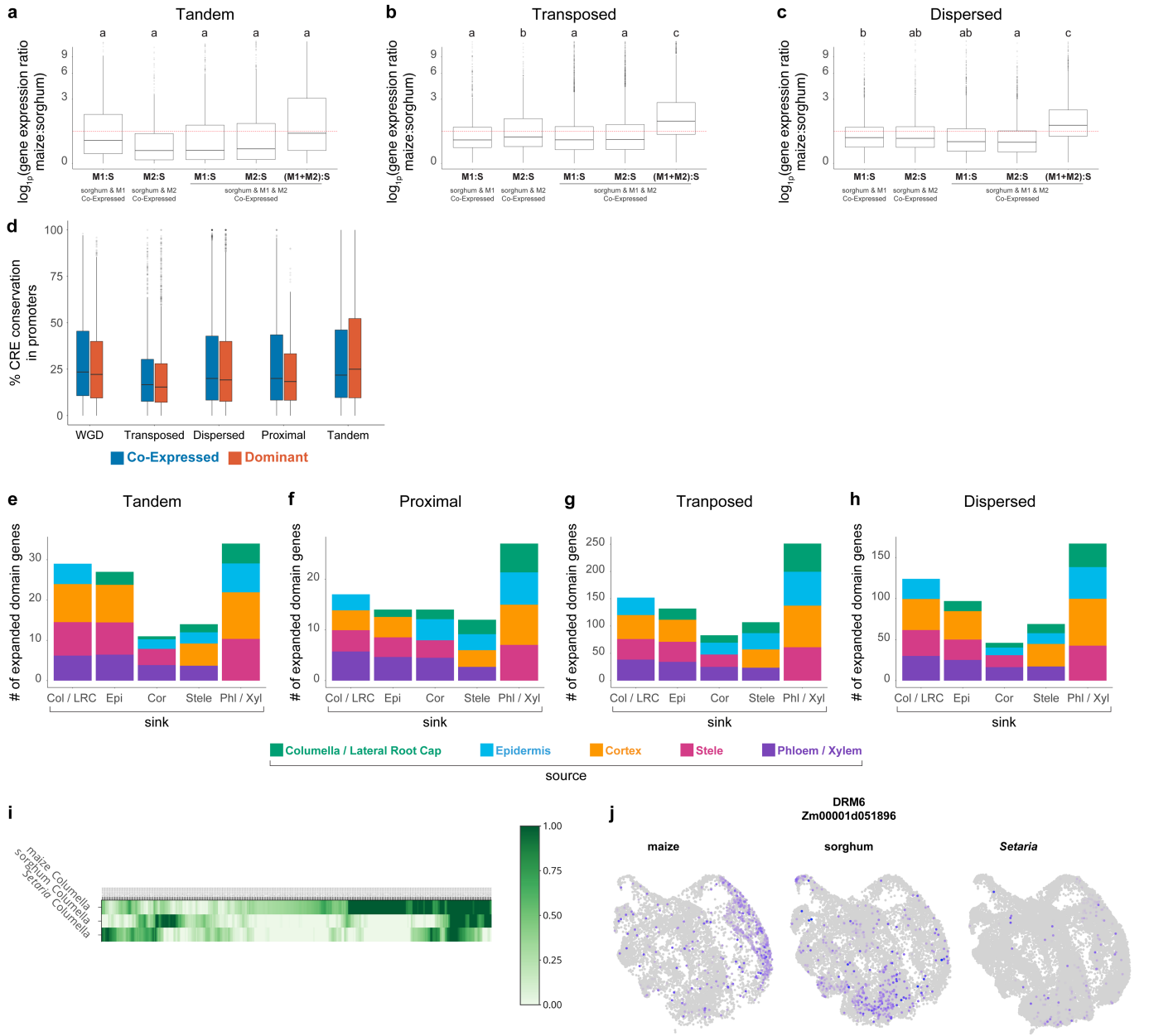
Extended Data Fig. 7 | *In-situ* hybridization corroborating evidence for marker localization in single cell/nuclei RNA-seq profiles in maize. **a**–**n** *in situ* hybridization using Hairpin Chain Reaction (HCR) probes labeling various transcripts. Cross sections are on the left and longitudinal sections are on the right. UMAPs showing each transcript's cluster localization are displayed next to each probe's fluorescent image. Additionally, spatial transcriptomics imaging data of the same probe is shown in the right column for (**c**–**e**). The minimum/maximum values for each fluorescence channel (grey: autofluorescence, magenta: HCR probes) have been adjusted to show the localization more clearly in the merged image.



Extended Data Fig. 8 | *In-situ* hybridization corroborates evidence for localization of marker gene expression from single-cell RNA-seq profiles in sorghum. a-i *In situ* hybridization using Hairpin Chain Reaction (HCR) probes labeling various transcripts. Cross sections are on the left and longitudinal sections on the right (a,c,d,e). Longitudinal sections are shown in (f,g,h,i). UMAPs showing each transcript's cluster localization are shown next to each probe's fluorescent image. The minimum/maximum values for each fluorescence channel (grey: autofluorescence, magenta: HCR probes) have been adjusted to show the localization more clearly in the merged image.



Extended Data Fig. 9 | Regulon conservation across species, and distribution of gene pair expression patterns. a Conserved regulons found using MINI-EX and their pattern of expression. The regulon is labeled by the transcription factor that putatively regulates it in each row. **b**–**d** Distribution of genes pairs on the dominance vs. regulatory subfunctionalization scale for transposed, tandem and proximal duplicate pairs. In blue, neofunctionalized duplicates are shown as a percentage of the bar. **e**–**g** Distribution on the dominance to regulatory subfunctionalization scale for dispersed gene duplicate pairs binned in thirds by their Ks value. The graphs suggest that duplicates tend to lose co-expressed patterns and gain dominance over time. **h** Boxplot of Ks values showing the distribution among all the duplicate classes used in the analysis. In h, statistical analysis was performed using a Kruskal-Wallis one-way ANOVA followed by the Tukey test for all pairwise comparisons. Not sharing a letter represents statistical significance at $p < 0.05$. In boxplots the middle line is the median, the lower and upper hinges correspond to the first and third quartiles (Q1,Q3), extreme line shows Q3+1.5xIQR to Q1-1.5xIQR (interquartile range-IQR). Dots beyond the extreme lines shows potential outliers. h. n = 10,104 WGD, n = 860 Proximal, n = 3,154 Transposed, n = 7,552 Dispersed, n = 1,448 Tandem.



Extended Data Fig. 10 | Overall analysis of expression conservation in duplicate classes and analysis of columella expression across species. **a**–**c** Dosage compensation analysis representing the expression ratios of maize over sorghum orthologous genes in tandem, transposed, and dispersed duplicate pairs. The first two boxplots represent cases in which a sorghum ortholog is expressed in the same homologous cell type as only a single maize duplicate (either M1 or M2). The third and fourth boxplots represent cases in which both homeologs are expressed in the same cell and a sorghum homolog is expressed in a homologous cell type. The last boxplot shows the ratio when both of the co-expressed homeologs are added together in the numerator, showing a mean ratio close to 1. The higher expression in the first two boxplots compared to the second two indicates dosage compensation. **d** Conservation rate of *cis*-regulatory elements between WGD homeolog pairs in promoters. The plot shows no major differences between co-expressed and dominant gene pairs, and no major differences among the different classes of duplication. **e**–**h** Distribution of maize genes displaying regulatory neofunctionalization of expression into new cell types. Colors signify the cell type of origin. **i** Expression heatmap of the 443 genes displaying high expression divergence across species in columella cells in maize, according to CoCoCoNet, with the orthologous gene expression in the other two species. **j** Example of the gene *DMR6* switching its expression between columella in maize to epidermis / cortex in sorghum. a-c, statistical analysis was performed using ANOVA followed by the Tukey test for all pairwise comparisons, Not sharing a letter represents statistical significance at $p < 0.05$. In boxplots the middle line is the median, the lower and upper hinges correspond to the first and third quartiles (Q1,Q3), extreme line shows Q3+1.5xIQR to Q1-1.5xIQR (interquartile range-IQR). Dots beyond the extreme lines shows potential outliers. a–h: n = 10,104 WGD, n = 860 Proximal, n = 3,154 Transposed, n = 7,552 Dispersed, n = 1,448 Tandem.

nature portfolio

Corresponding author(s): Kenneth Birnbaum

Last updated by author(s): Feb 27, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used
Data analysis	No additional custom code was develop in this study. Only basic function from R have been used to plot the data. Only the following softwares and packages were used: Cell Ranger 5.0.1 (10x Genomics, https://support.10xgenomics.com/single-cell-gene-expression/software/overview/welcome) was used to generate fastq files and align the reads from sequenced single cell or single nuclei Seurat package v4.0 (https://satijalab.org/seurat/) was used to integrate the expression matrices generated by Cell Ranger, according to tutorial present online. DIAMONDv2.0.6 (http://ab.inf.uni-tuebingen.de/software/diamond) and DupGen_finder.pl and DupGen_finder-unique.pl V1 (https://github.com/qiao- xin/DupGen_finder) were used to identify duplicated genes shinyGO V0.61 (http://bioinformatics.sdstate.edu/go/) was used to identify GO term enrichments FIMO algorythmv5.5.1 (https://meme- suite.org/meme/tools/fimo) was used to predict cis-regulatory elements in maize duplicates MetaNeighborv1 package in Python (https://github.com/gillislab/pyMN) was used to assess cross-species-cell transcriptional relationship scGENv2.1 (https://github.com/theislab/scgen) was used as an alternative method of single cell clustering Scanpyv1.9 (https://scanpy.readthedocs.io/en/stable/) was used calculate the nearest neighbors using scanpy.pp.neighbors, and generated a 2D projection using UMAP MINI-EXv1 (https://github.com/VIB-PSB/MINI-EX) was used to identify single cell regulatory networks ImageJ V2.9.0/1.53t

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All reference genomes were downloaded from Arabidopsis TAIR10.38, at https://www.arabidopsis.org/, for Maize B73 v4, Sorghum bicolor v3 and Setaria viridis v2 reference genomes at https://plants.ensembl.org/.
All raw scRNA-seq and snRNA-seq data, expression matrices and analyzed R-Seurat objects are available under GEO accession (GSE225118). All data on duplicate genes are provided in Supplementary Tables. All cell specific marker genes for all species, including a shared pan library of marker genes between species are provided in Supplementary Table 4.
All data used to generate figures is available at https://figshare.com/articles/dataset/Data_for_Guillotin_et_al_/22331002, except for the following figures, for which the data can be found under GEO accession GSE225118, in the following deposited files: Arabidopsis_Cells_Nuclei_Seurat_Obj.RData.gz (Fig. 1c; Extended Data Fig. 2c,d; Extended Data Fig. 4a,b), Maize_Sorghum_Setaria_Cells_Nuclei_Seurat_Obj.RData.gz (Extended Data Fig. 3d, Extended Data Fig. 5c,d). Extended Data Figs. 2c,d; 3d; and Fig. 5c,d are clustered separately.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	No human related research have been used in this manuscript
Population characteristics	No human related research have been used in this manuscript
Recruitment	No human related research have been used in this manuscript
Ethics oversight	No human related research have been used in this manuscript

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculation was performed. Each sampling was performed on a new batch of germinated seedlings, with a minimum of 30 root tips per sample. This amount was chosen to account for plant variability in root growth and minimum sample size requirement for statistical analysis. This amount was also calculated to avoid an excess of total amount of samples leading to technical difficulties in extracting cells or nuclei.
Data exclusions	No data was excluded
Replication	For each species, data were collected in the lab with at least 3 biological replicates for cells or nuclei (except for Sorghum, only one replicate was generated in single cells due to technical difficulties in digesting the cell wall of this species, as described in the manuscript, while 3 biological replicates were generated for Sorghum nuclei). Apart from the Sorghum cell-profiles for which protoplasts could not be generated (as above), all other replicates were successful. Each replicate is made by combining cells or nuclei from (30 to 200 root tips depending on the specie, as describe in Supplementary table 1). Each replicate allows the sequencing of 2000 to 7000 cells or nuclei. No technical repetitions were made in this study, ie: two single cell or single nuclei RNAseq made on the same day with the same batch of plant).
Randomization	No randomization was used in this study as plant were grown in only one control condition in the same environment, for all three biological replicates. Arabidopsis was grown in vitro and maize, sorghum and Setaria were grown in the same growth chamber in the same location at different time.
Blinding	Blinding was not possible in this study as there was no treatment effect to measure and as we compared species transcriptomes directly.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging