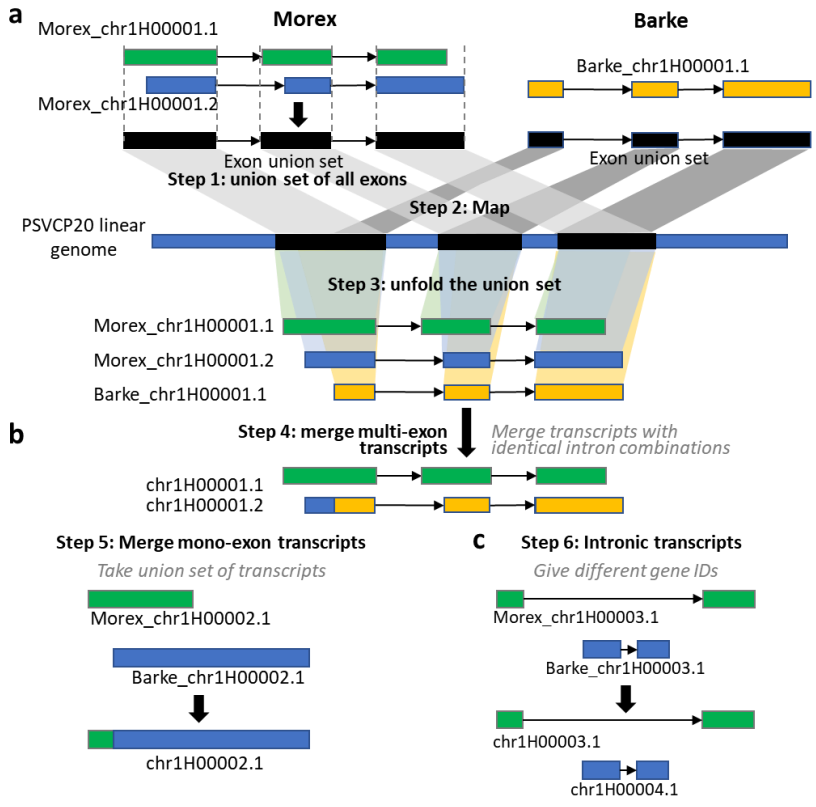


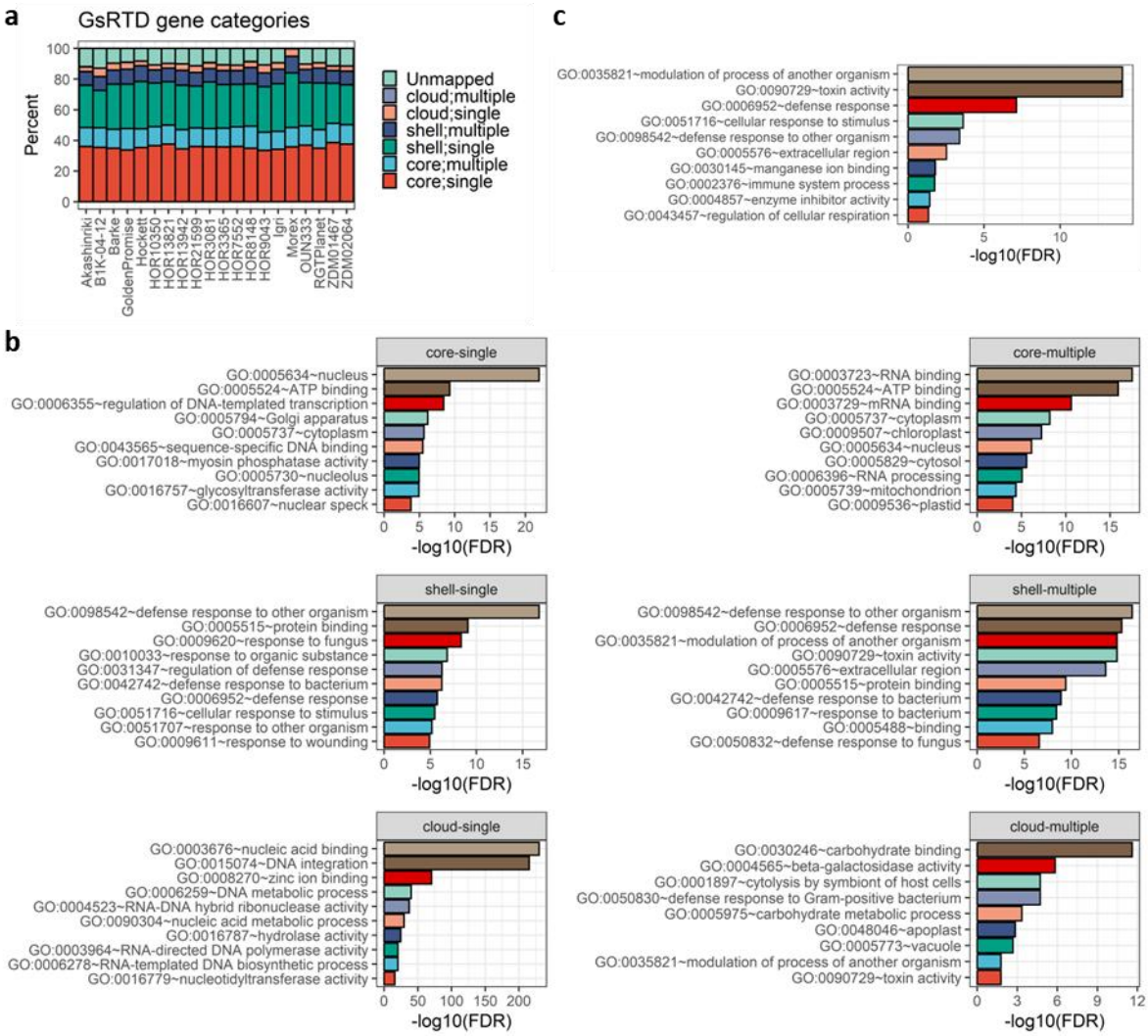
c

Look-up table

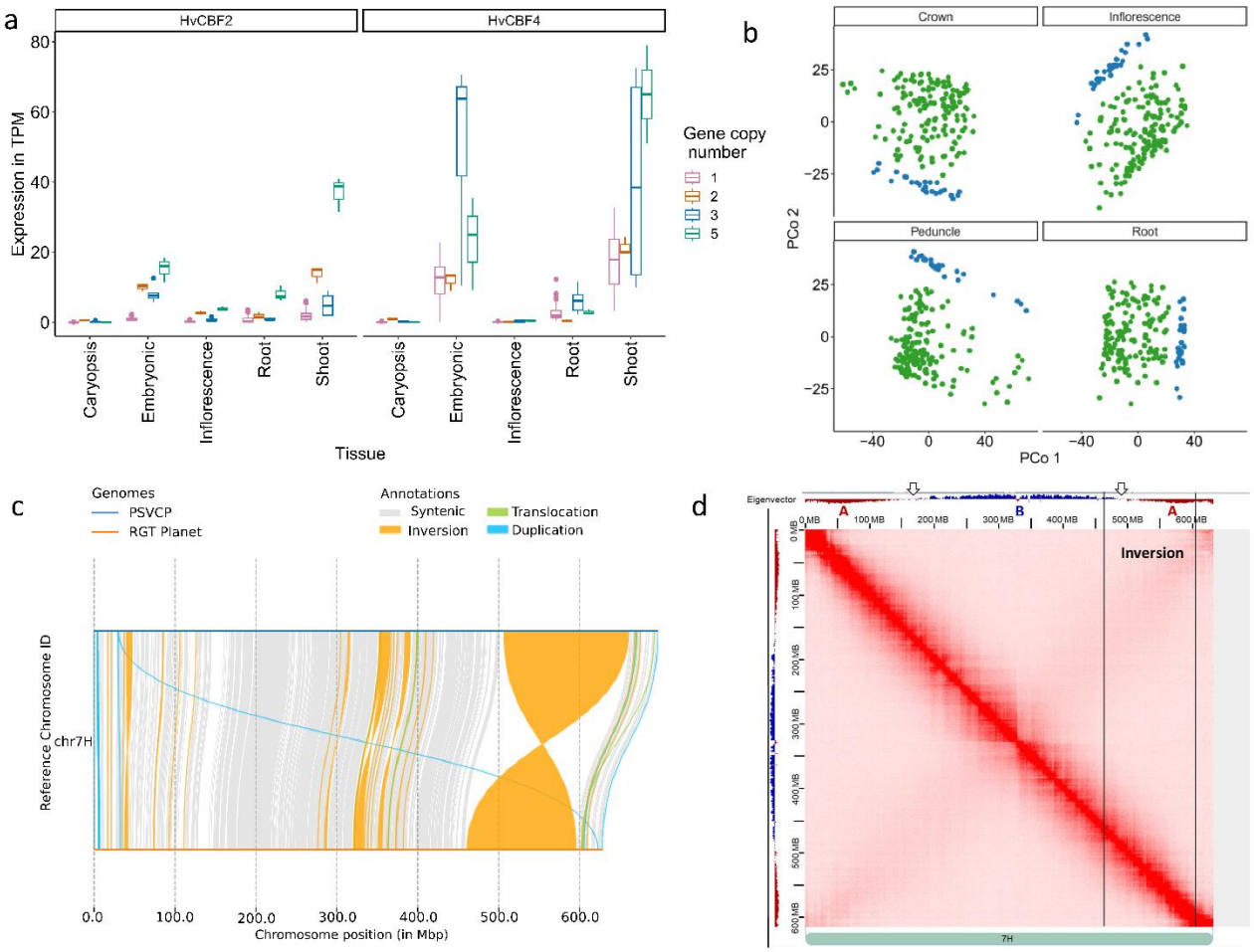
PanBaRT20_gene_id	PanBaRT20_transcript_id	GsRTD_gene_id	GsRTD_transcript_id
chr1H00001	chr1H00001.1	Morex_chr1H00001	Morex_chr1H00001.1
chr1H00001	chr1H00001.2	Morex_chr1H00001; Barke_chr1H00001	Morex_chr1H00001.2; Barke_chr1H00001.1
chr1H00002	chr1H00002.1	Morex_chr1H00002; Barke_chr1H00002	Morex_chr1H00002.1; Barke_chr1H00002.1
chr1H00003	chr1H00003.1	Morex_chr1H00002	Morex_chr1H00002.1
chr1H00004	chr1H00004.1	Barke_chr1H00003	Barke_chr1H00003.1

- **Core genes:** match 20 genotypes
- **Shell genes:** match 2-19 genotypes
- **Cloud genes:** match 1 genotype
- **Single copy:** only one gene in a genotype matched to a PanBaRT20 gene
- **Multiple copy:** multiple genes in the same genotype matched to a PanBaRT20 gene

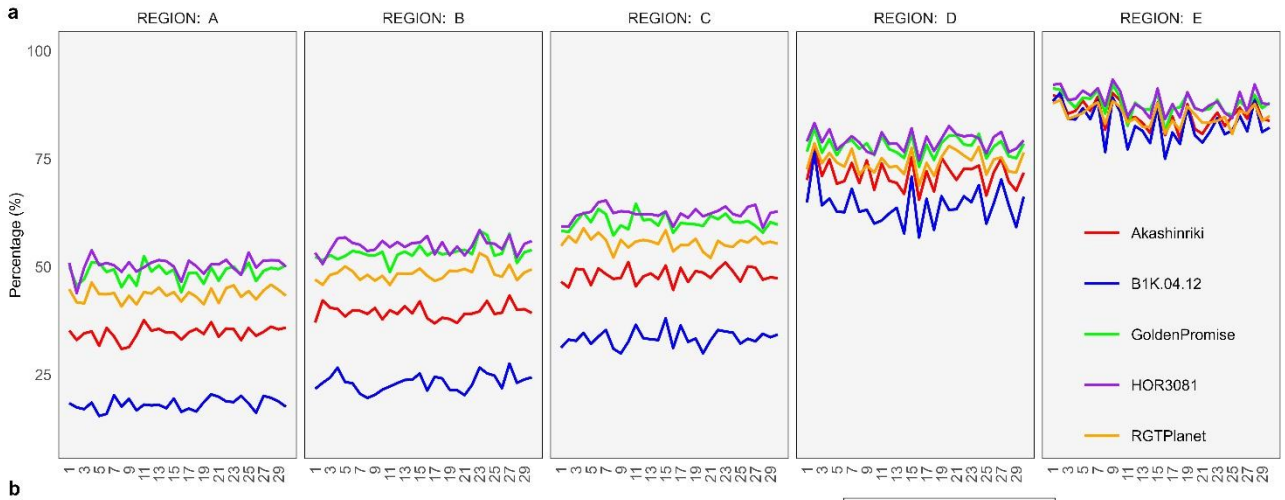
Extended Data Figure 1: Construction of PanBaRT20. The Morex and Barke GsRTDs were used as examples to illustrate the construction of PanBaRT20 from 20 GsRTDs. a, The transcripts in each GsRTD gene were collapsed into an exon union set (step 1). The union sets of all the GsRTD genes were mapped to the PSVCP pan-genome using Minimap2 (step 2). This ensured that all the transcripts from the same gene were mapped to the same genomic loci on the PSVCP pan-genome. (step 3). b, The overlapped transcripts were assigned the same gene ID. The multiple-exon transcripts that shared identical intron combinations were merged and the furthest start and end of these transcripts were taken as the transcript start site (TSS) and end site (TES) of the merged transcript (step 4). The overlapped mono-exon transcripts were merged into one transcript with the furthest starting and ending as the TSS and TES (step 5). If a set of overlapped transcripts were located entirely within the introns of other transcripts, they were assigned a separate gene ID (step 6). c, After assigning new gene and transcript IDs to the PanBaRT20 gene models, a look-up table was created to record the gene and transcript associations between PanBaRT20 and 20 GsRTDs.



Extended Data Figure 2: Characteristics of GsRTDs and PanBaRT20 a. The percentages of GsRTD genes in 20 genotypes that matched core-single copy, core-multiple copy, shell-single copy, shell-multiple copy, cloud-single copy and cloud-multiple copy genes in PanBaRT20. **b.** Significant gene ontology (GO) enriched terms of PanBaRT20 genes in core, shell and cloud gene categories. **c.** Significant GO enriched terms of 2,925 PanBaRT20 genes with zero TPM in any tissue in 1-19 genotypes.

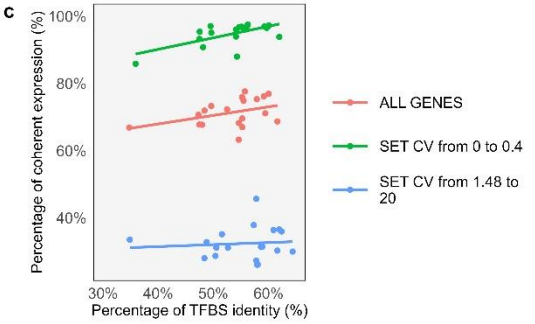


Extended Data Figure 3: Drivers of variation in transcript abundance. **a** Expression of HvCBF2 and HvCBF4 by tissue and Copy number. Significant correlation can be found for HvCBF2 in the Coleoptile (r^2=0.85), inflorescence (r^2=0.62), root (r^2=0.5) and shoot (r^2=0.71), and for HvCBF4 in the inflorescence (r^2=0.44) and shoot (r^2=0.49 **b** Principal component cluster analysis of the genes expressed in the inversion. The region contained 2508 expressed PanBaRT genes in at least one of the four tissues. The PCA splits the 201 genotypes into two clusters in every tissue corresponding to those genotypes carrying the inversion versus those carrying the wild type. **c** Alignment of chromosome 7H between cultivar RGT Planet and the linear pangenome generated through PSVCP highlighting the genomic 140Mb inversion. **d** Hi-C interaction matrix of chromosome 7H from the cultivar Morex. The division into A and B compartments emerges as principal component 1 in a PCA of Hi-C interaction frequencies. The A/B compartments boundary on the long arm of 7H is quite distinct and part of the inversion in RGT Planet.

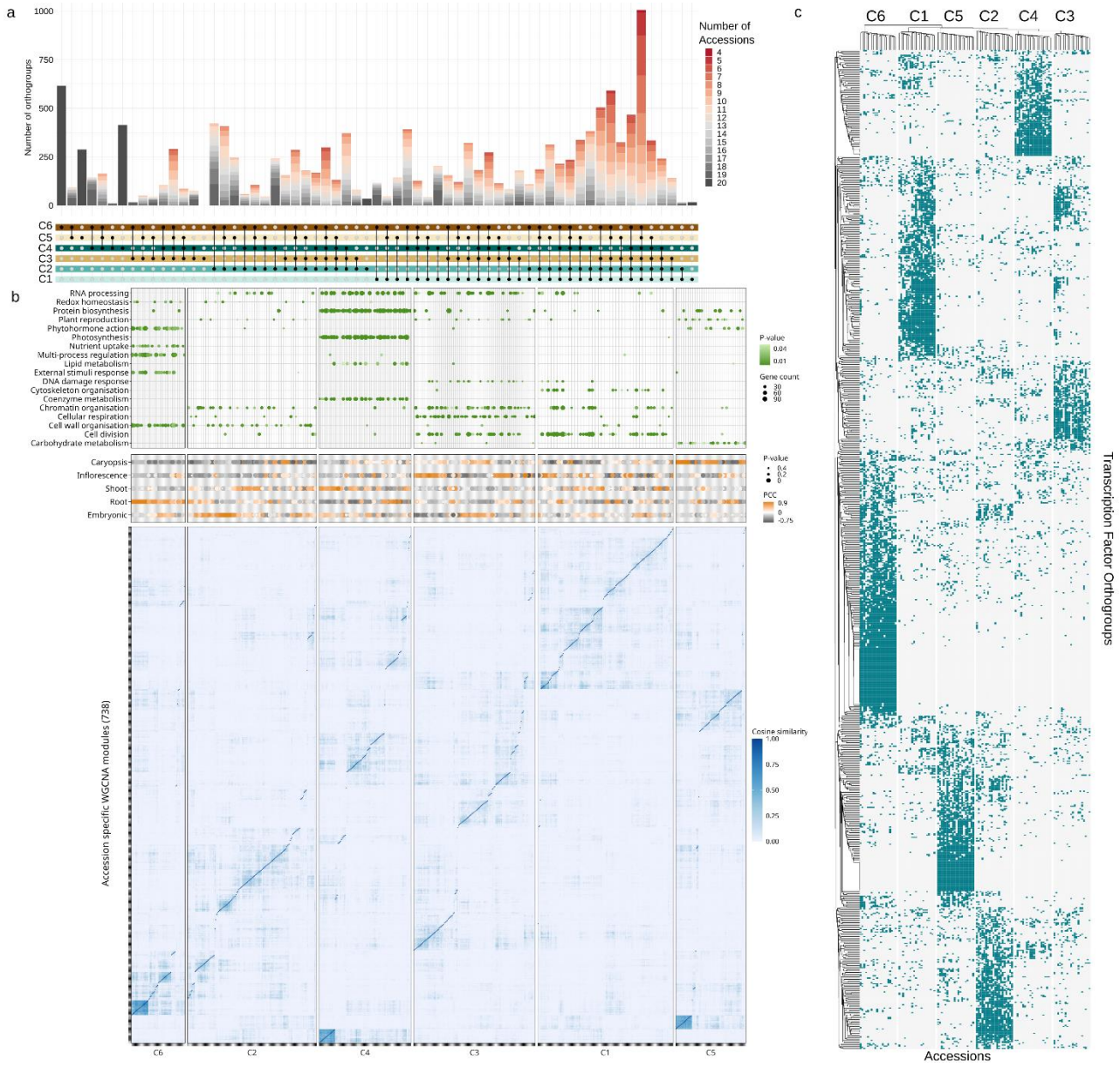


b

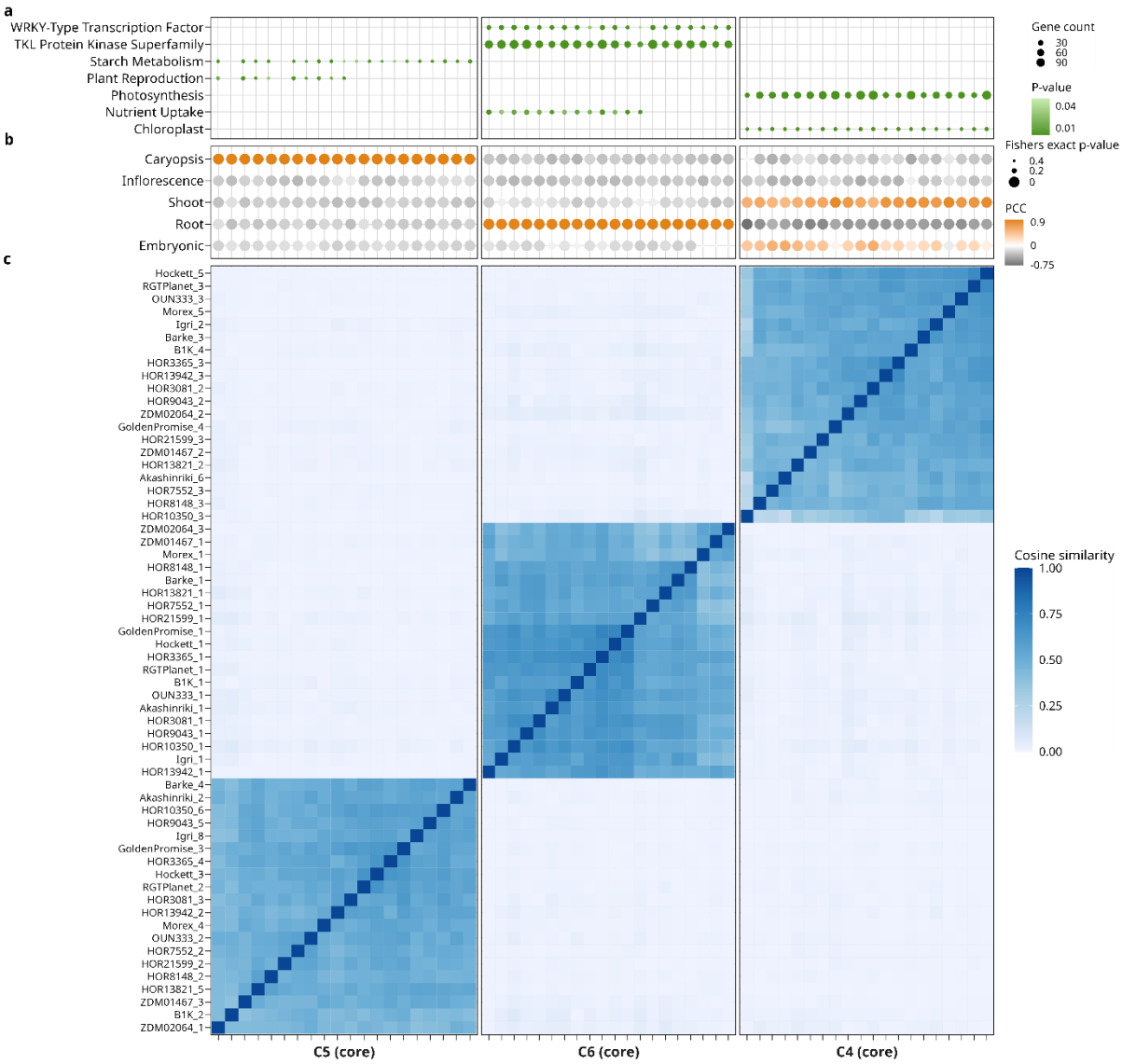
Code	TFBS	Code	TFBS
1	AP2-DREB - RCCGAC	16	GARE - TAACARA
2	AP2-ERF - CGCCGCC	17	HD-WOX13 - CAATCA
3	AP2-TOE2 - MTCGTA	18	HD-ZIP - AATNATT
4	B3-AuxRE - TGTCTC	19	LOB - TCCGGA
5	B3-AuxRE2 - TGTCGG	20	MRECHS - AACCTA
6	B3-LEC-AB3 - CATGCA	21	MYB-GARP-B-ARR - AGATWCG
7	B3-RAV - CACCTG	22	MYB-R2R3-MBSI - CNGTTR
8	bZIP-A-box - TACGTA	23	MYB-R2R3-MBSII - GKTWGTTR
9	bZIP-C-box - GACGTC	24	MYB-R2R3-MBSIIG - GKTWGGTR
10	bZIP-G-box - CACGTG	25	MYB-related - AGATAT
11	bZIP-GCN4 - TGASTCA	26	SPL - CGTAC
12	C2C2-ZF-DOF - AAAGY	27	STY1 - CCTAGG
13	C2C2-ZF-GATA - WGATAR	28	TCP - GGNCCC
14	C2C2-ZF-YABBY - WATNATW	29	trihelix - GRWAAW
15	C2H2-ZF - CAGCT	30	WRKY - TTGACY



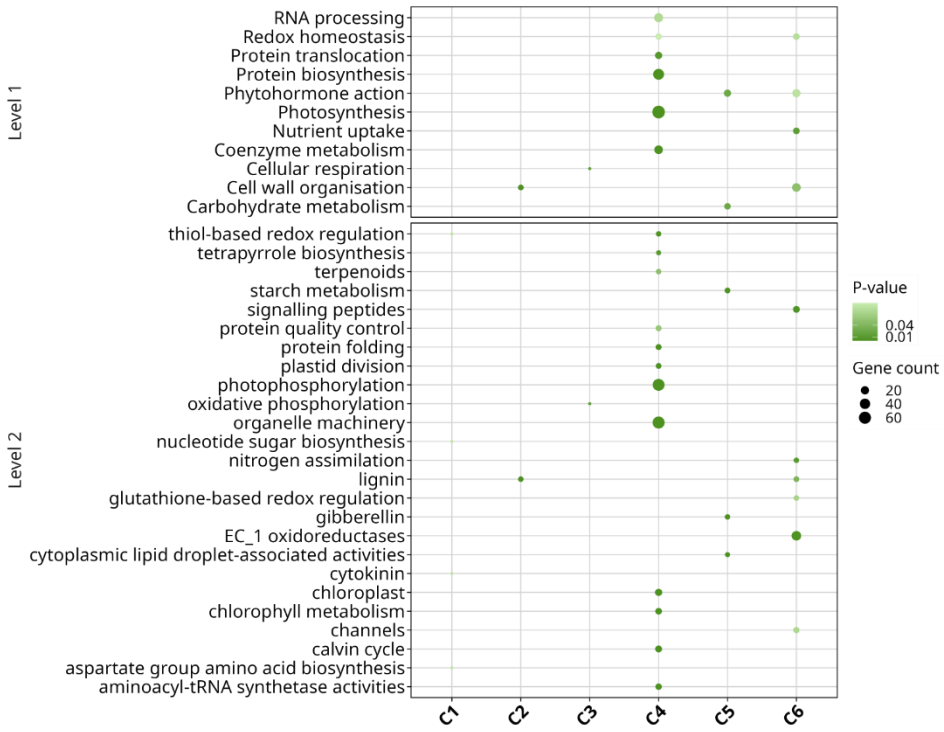
Extended Data Figure 4: Transcription factor binding site analysis **a.** Percent identities for transcription factor binding sites (TFBS) in 500bp segments of 2kb upstream (regions from A to D) and 500bp downstream (region E) of start codon computed for each genotype vs. Morex. Five representative genotypes among the 19 tested are plotted. The x-axis shows 30 TFBS consensus sequences[52,53] considered in the analysis. **b.** Table listing the TFBS indicated on x-axis with a code from 1 to 30 (see **Methods** for details). **c.** Percent TFBS identities against percent coherent expression in each pairwise comparison (individual genotypes vs Morex, each pair represented by a dot) for all genes (15,001 genes Pearson correlation value = 0.395), genes with low coefficient of variation (CV) (0 to 0.4 CV, 2,999 genes, corr. = 0.695) and genes with high CV (1.48 to 20 CV, 2,994 genes, corr. = 0.087).



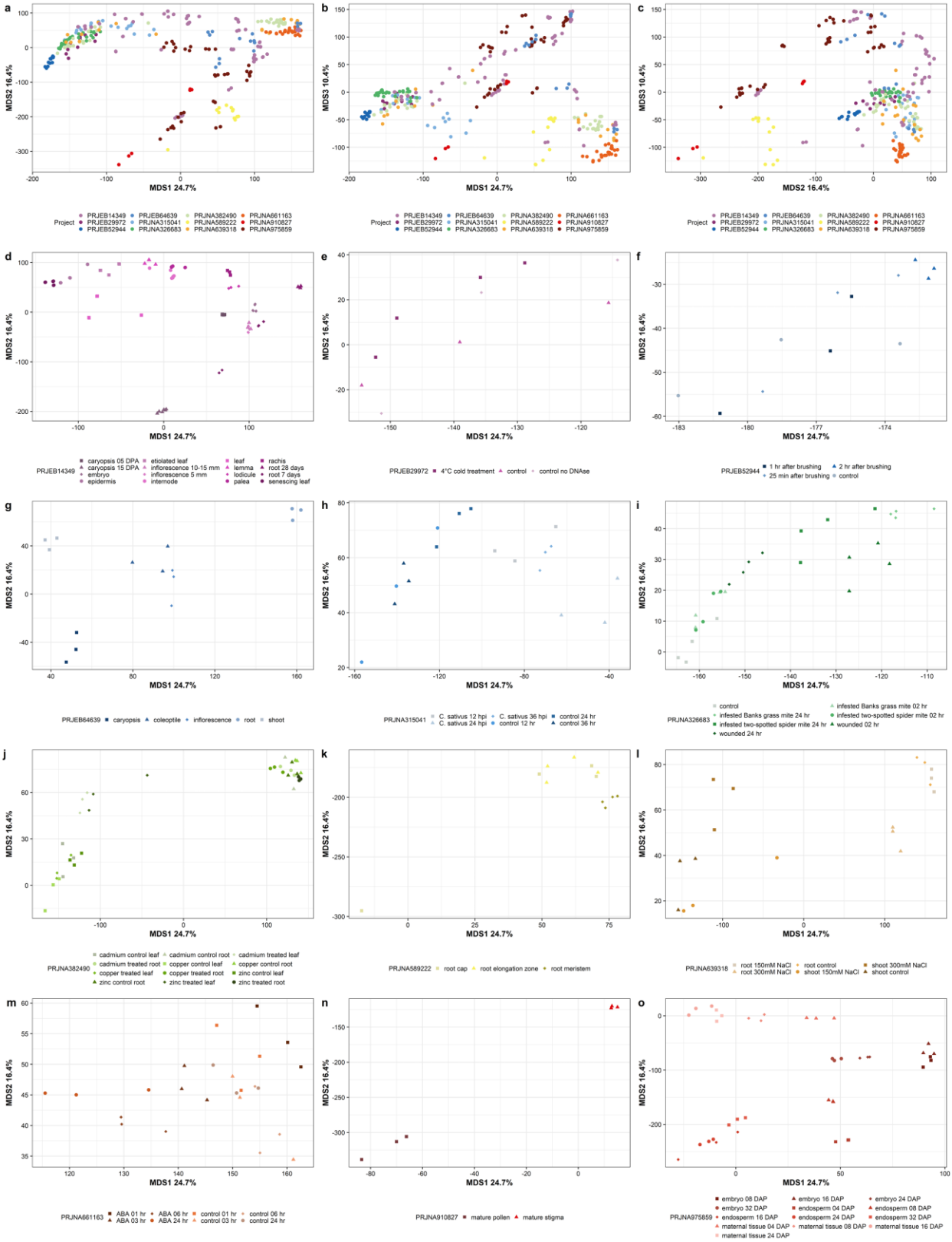
Extended Data Figure 5: Community Clustering a. Full distribution of 1:1 orthologous groups (exactly 20 genes, one from each genotype) across the six communities. Number of accessions refers to the largest cluster within a single community for the respective orthogroup. **b.** Heatmap displaying the Louvain community clustering of all 738 WGCNA modules (lower panel). 6 main communities were identified (X-axis C1-C6), which show distinct functional patterns (upper panel) and module-tissue correlations (middle panel). **c.** 861 transcription factors presence-absence clustering representation across the six communities.



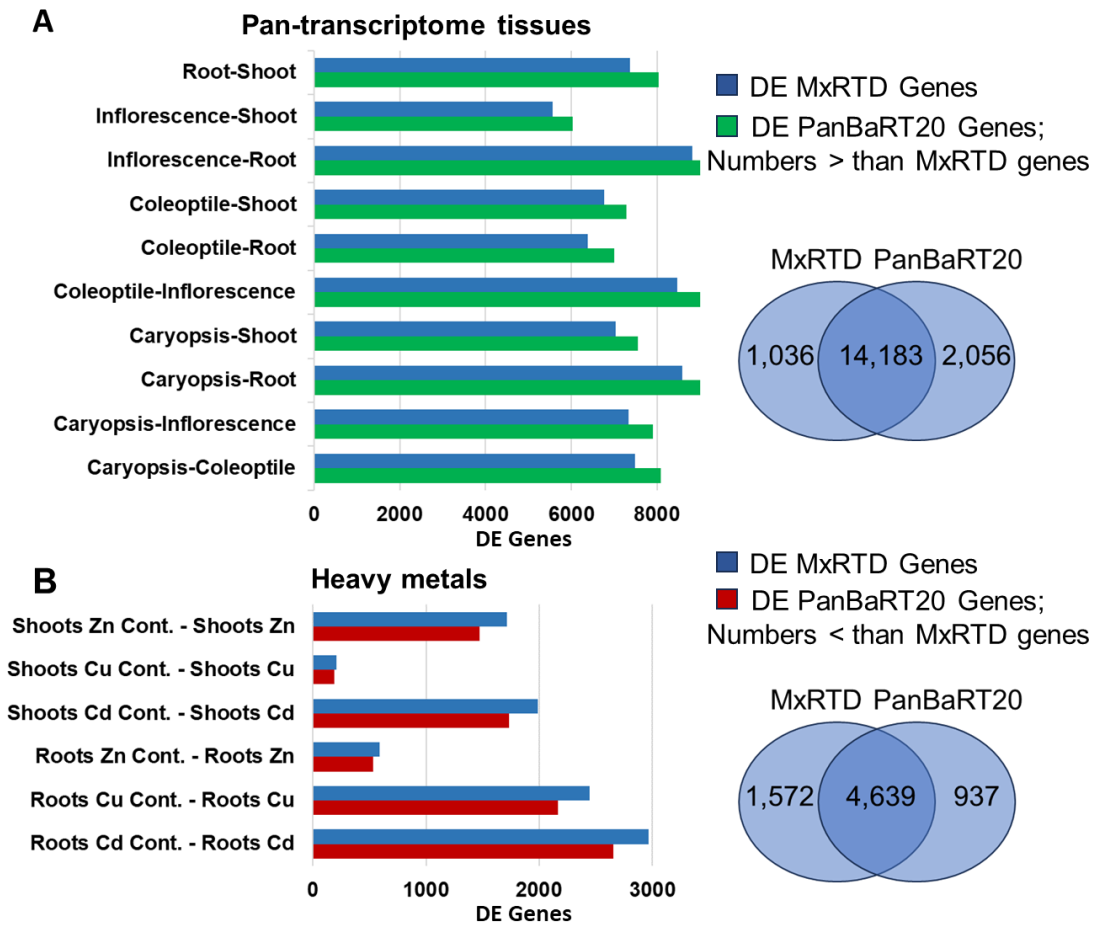
Extended Data Figure 6: Summary of functional enrichments (a) and module-tissue correlations (b) for the 3 most connected groups of modules with Cosine-similarity cutoff 0.43 (zoom-in of Extended Data Figure 4) displaying their Cosine similarity values and accession wise module names (c).



Extended Data Figure 7: Functional enrichment for orthologous groups exclusively present within one community for the two highest hierarchical Mercator categories (Level 1, Level 2) for 17 orthologous groups in C1, 34 in C2, 1 in C3, 414 in C4, 288 in C5 and 616 in C6.



Extended Data Figure 8: Multidimensional scaling plots on the gene expression of 12 different RNA-seq datasets used for generating the Morex Gene Atlas. a-c The first three dimensions showed the distribution of the samples from all projects. **d-o** The first two dimensions of the gene expression data from the individual projects.



Extended Data Figure 9: Differentially expressed (DE) genes in two experimental datasets. **a.** represents results from the pan-transcriptome sample dataset (PRJEB64639) while **b.** represents results from the heavy metal experimental dataset (PRJNA382490). Graphs show the numbers of DE genes in each named contrast group. Samples were pre-processed with cut off of 10 CPM in a least two of the different samples. DE genes were selected with a greater than 2-fold change and a significance value of p=<0.01. The blue box represents the total number of DE genes calculated using the Morex RTD (MxRTD), the green and red boxes represent the total number of DE genes calculated using the |Pan-BaRT20 RTD. Green represents Pan-BaRT20 which showed more DE genes compared to MxRTD and the red represents Pan-BaRT20 which showed less DE genes compared to MxRTD. The venn diagrams show the total number of unique genes across all the contrast groups tested using both Mx RTD and PanBaRT20 RTD. The overlap represents the number of DE genes common to both RTDs.