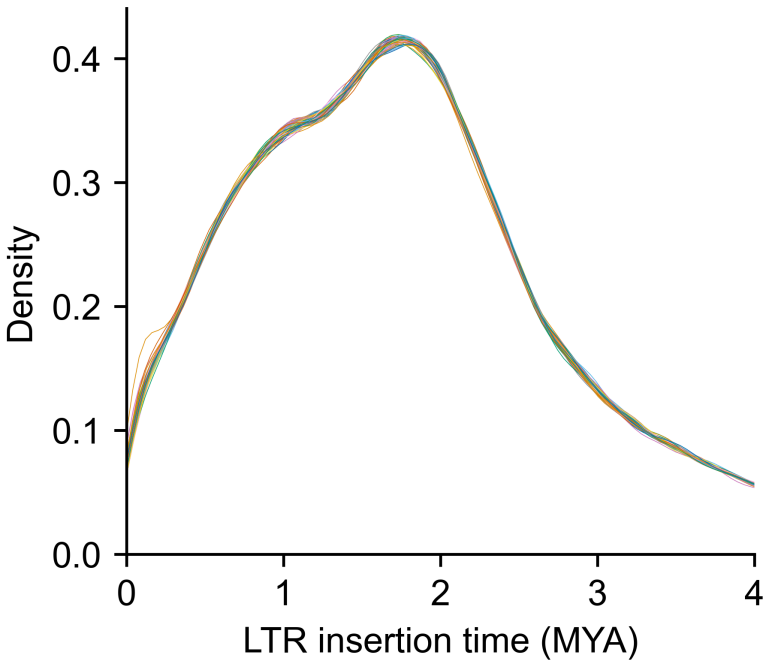
nature portfolio

https://doi.org/10.1038/s41586-024-08187-1

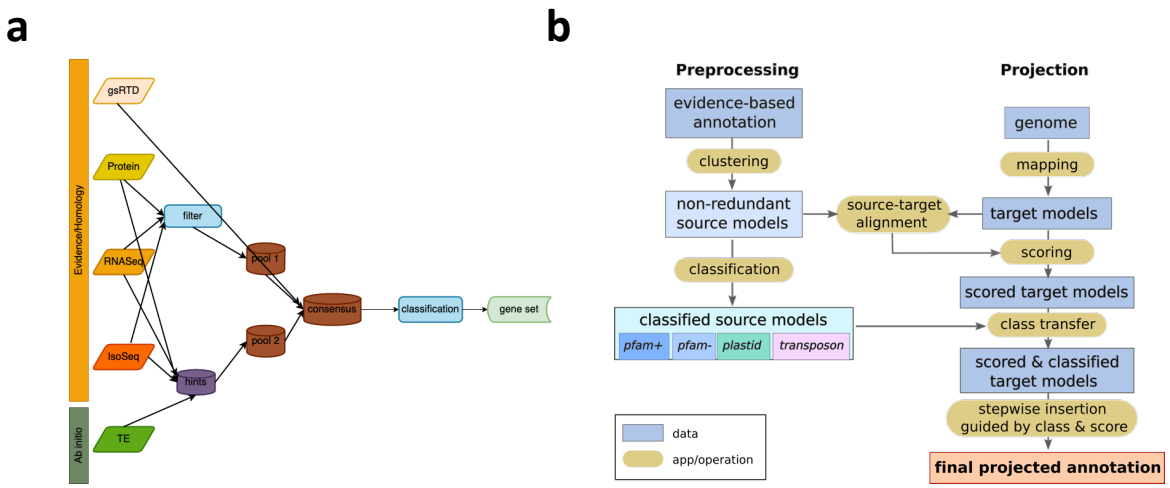
## Supplementary information

# Structural variation in the pangenome of wild and domesticated barley

In the format provided by the authors and unedited

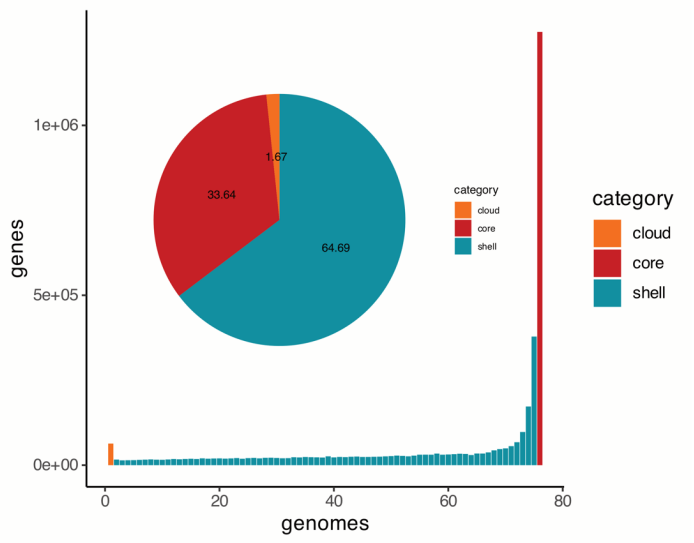



**Supplementary Figure 1: Insertion age distribution of de novo detected full-length LTR-retrotransposons for 76 barley genotypes.** Each coloured line stands for one genotype.

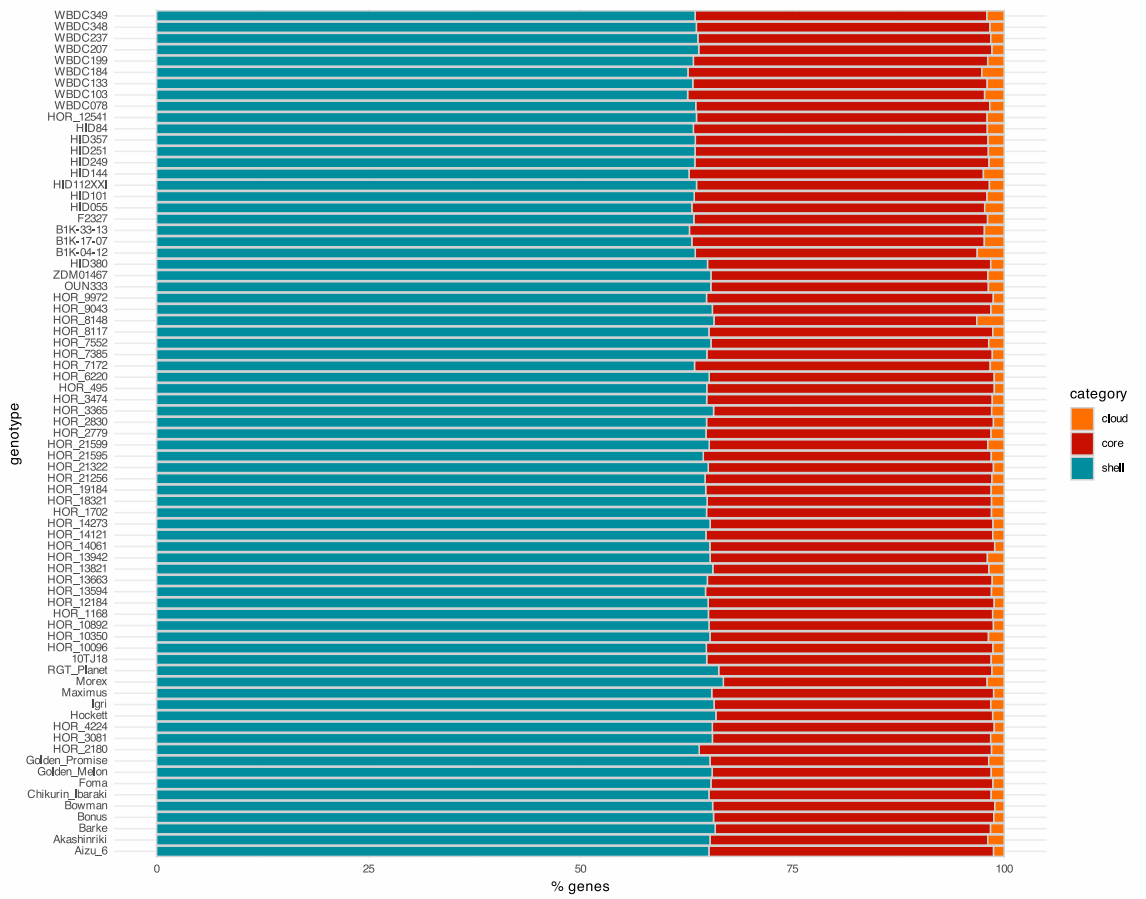



**Supplementary Figure 2. Workflow for annotating, projecting and clustering gene models. (a)** Workflow for the *de novo* gene predictions. **(b)** Workflow for gene projections.

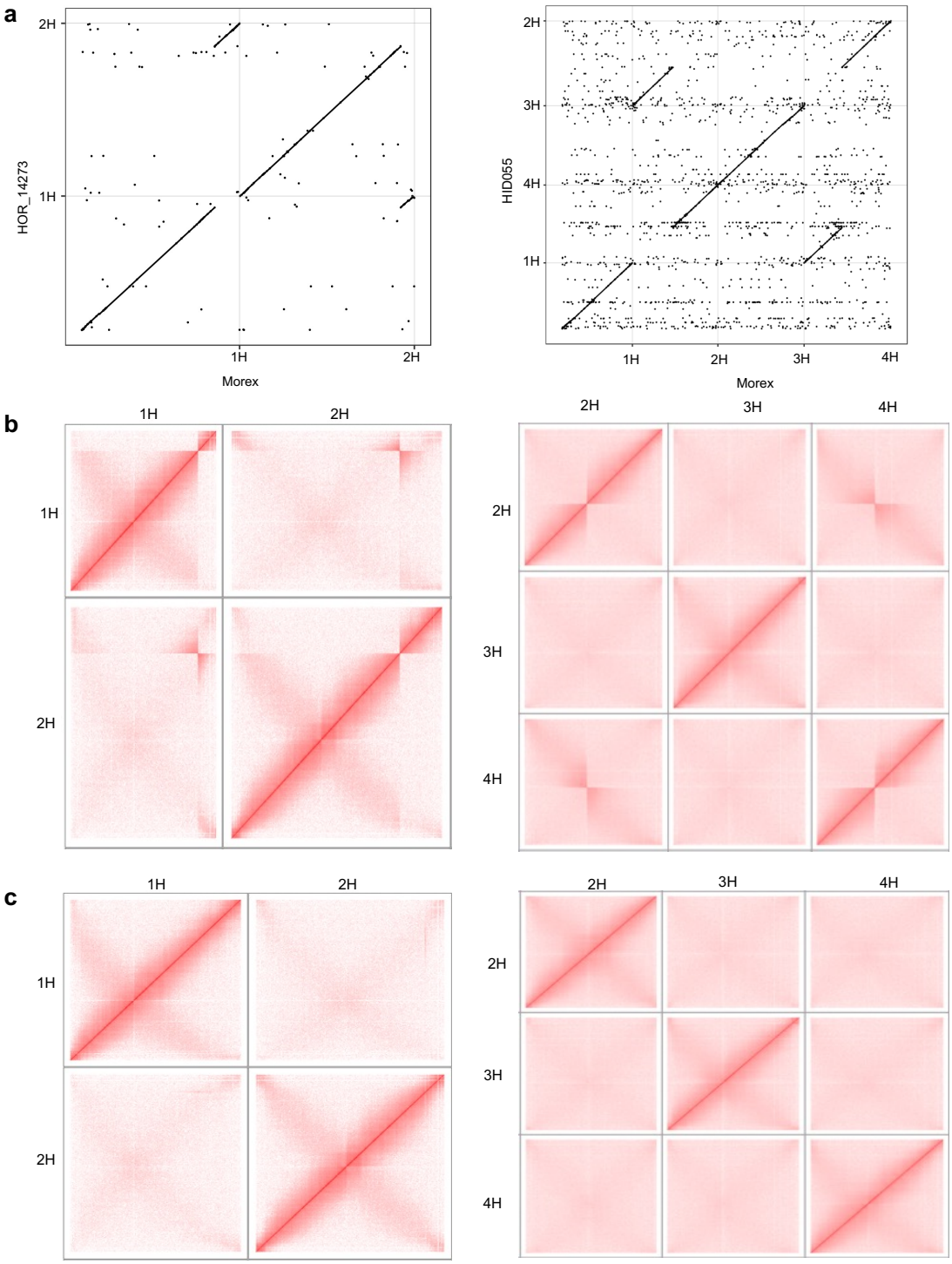
**a**



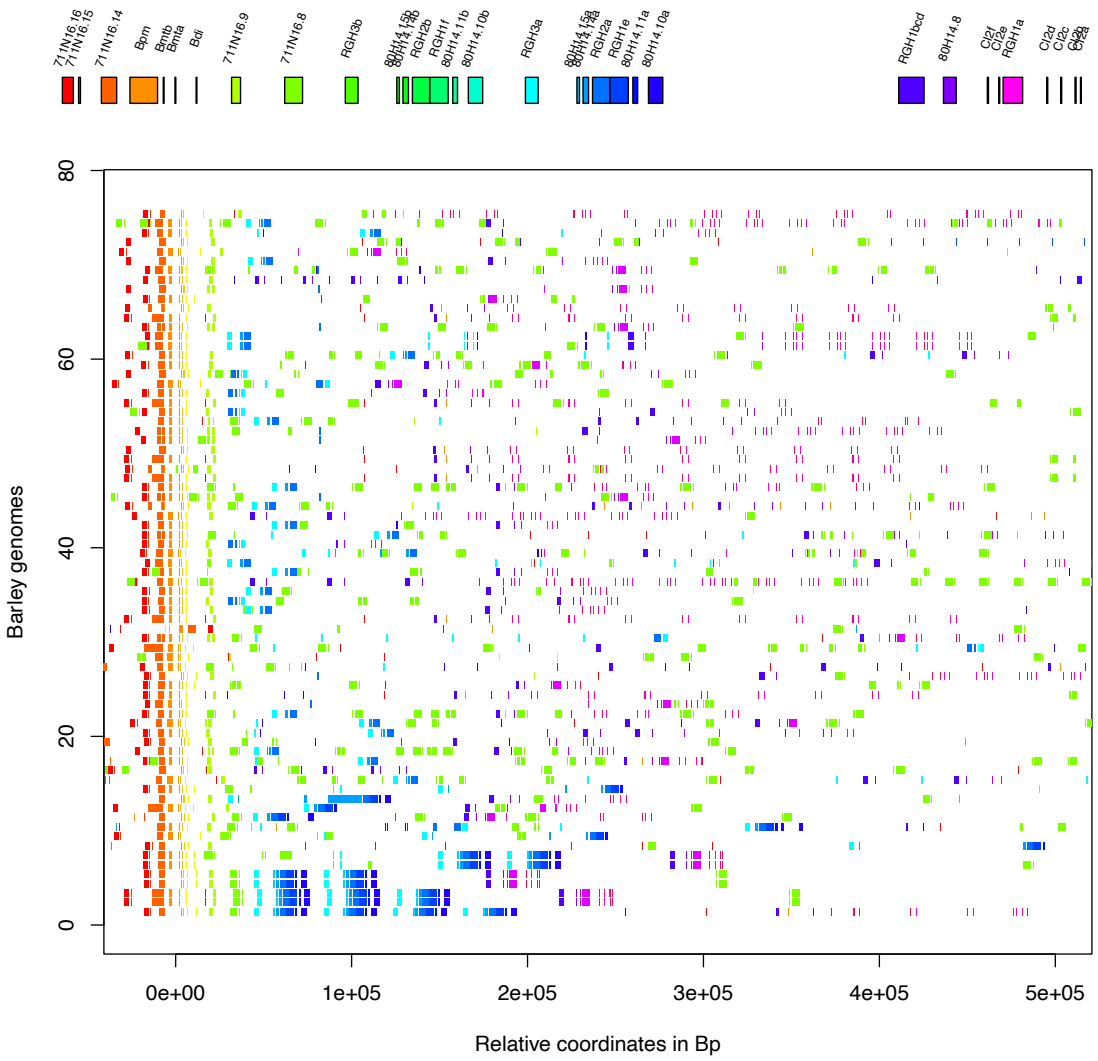

**b**



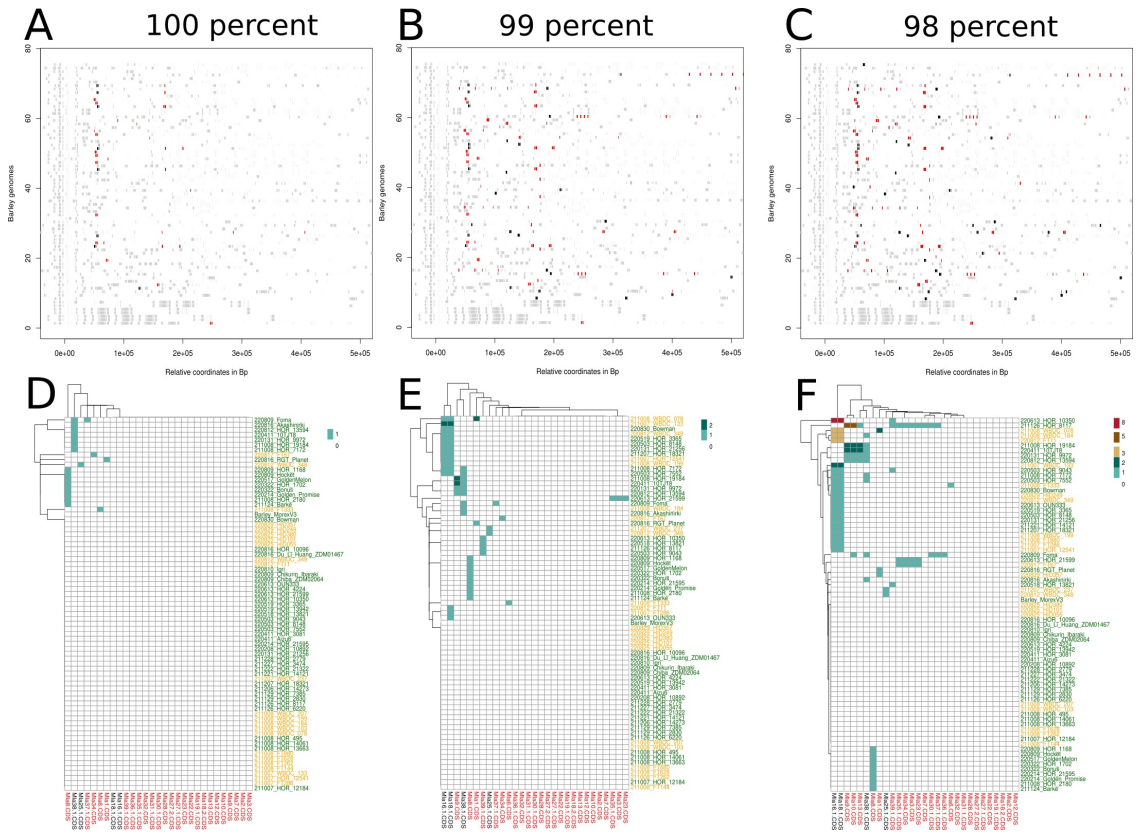

**Supplementary Figure 3: Orthologous framework. (a)** Histogram of the sizes of individual hierarchical orthologous groups (HOGs). "Size" means the number of pangenome accessions that have at least one member in a given HOG. The pie chart shows the ratio between conserved, shell and core genes. **(b)** Bar chart illustrating the proportion of genes contained in core, shell and cloud HOGs (by genotype).



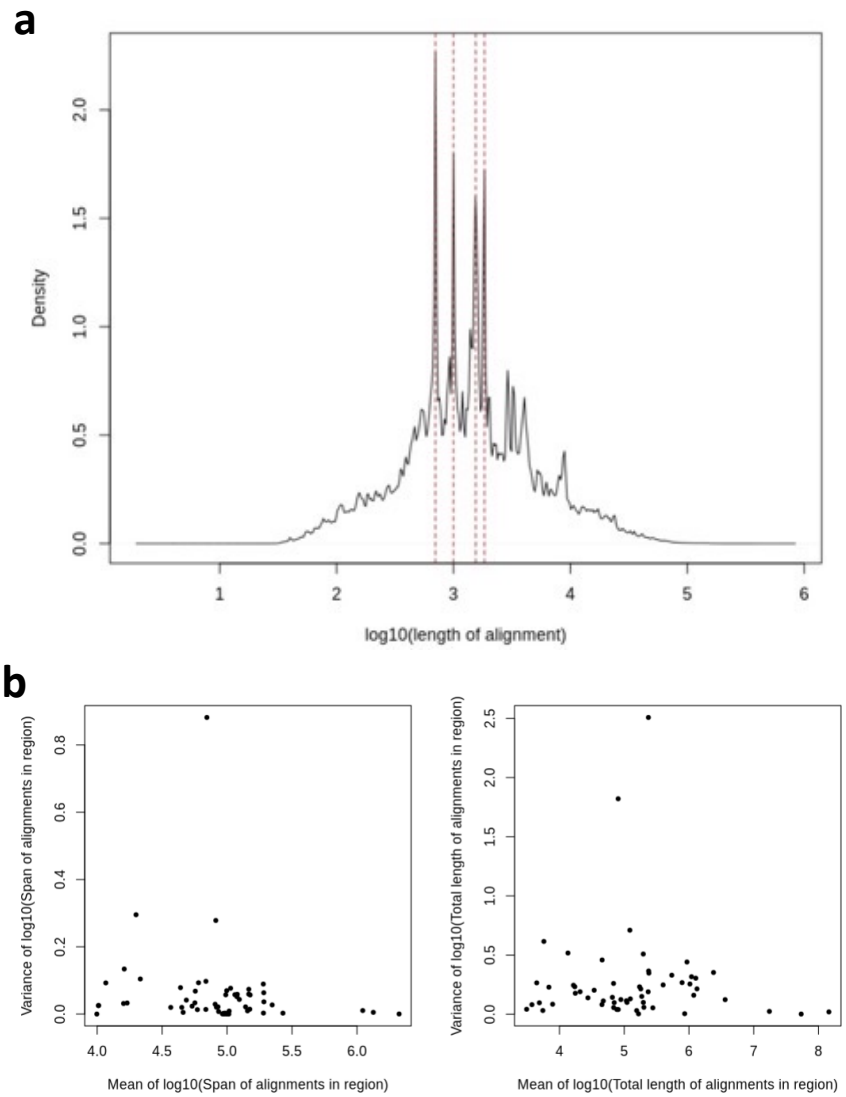

**Supplementary Figure 4: Translocations in HID055 and HOR 14273. (a)** Alignments of the final pseudomolecules to Morex. **(b)** Hi-C contact matrices using the chromosomal configuration of Morex. **(c)** Hi-C contact matrices using the chromosomal configuration of the respective native pseudomolecules. A clear interchromosomal signal is seen when the Morex reference is used, indicative of translocation relative to that genotype. If the native pseudomolecules are used no such off-diagonal signal is seen, supporting their structural integrity. Left column – HOR 14273; right column – HID055.



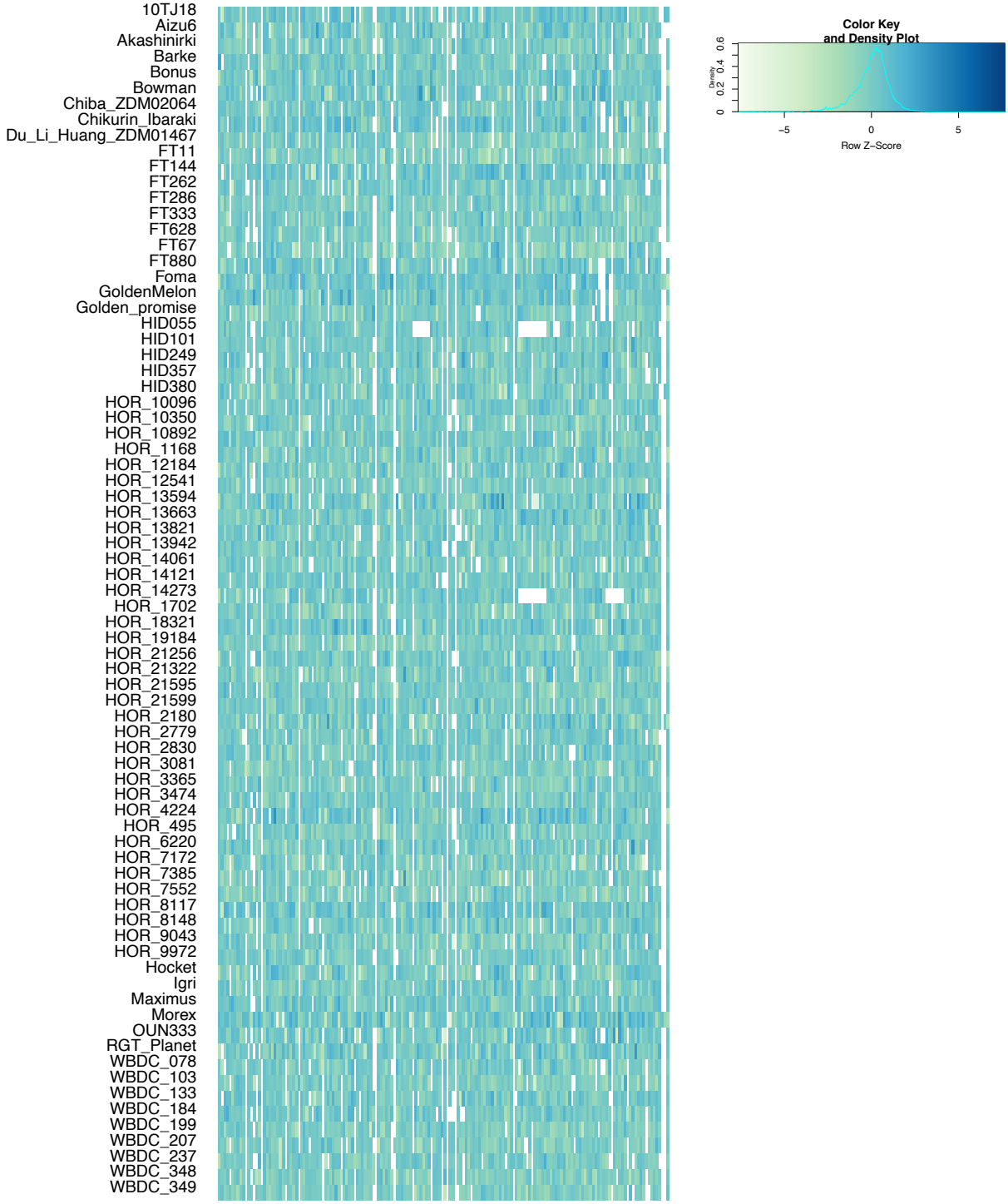

**Supplementary Figure 5: Structure of the *Mla* region across the 76 pangenome accessions**. The gene models present in the Morex genome are shown on top. The genomes are sorted based on the number of *RGH2a* copies with the most copies at the bottom. The positions on the *x*-axis are scaled based on the position of the *Bpm* gene.




**Supplementary Figure 6: Structure and copy number variation at *Mla* at different thresholds for alignment similarity.** Structural plot around the *Mla* locus for the 76 genomes. The gray rectangles represent all the homologous genes (based on BLAST alignments) in the regions in comparison with the Morex annotation. Each rectangle represents a blast result and the size on the rectangle is proportional to the length of the matching fragment. The red and black rectangles represent the position of the known *Mla* alleles from subfamily 1 and 2, respectively, as defined by Seeholzer et al.[25] Each gene coordinate is scaled based on the position of the *Bpm* gene in the respective genome. Three different thresholds, 100, 99, and 98%, for blasting the *Mla* alleles have been used for **(a)**, **(b)**, and **(c)**, respectively. **(d-f)**: Copy number variation of *Mla* alleles aross the 76 barley genomes. *Mla* alleles names in red and black represent the two subfamilies. The names of the accessions are coloured according to domestication states (green – domesticated; orange – wild). The colouring of the square indicates the number of copies according to the legend.

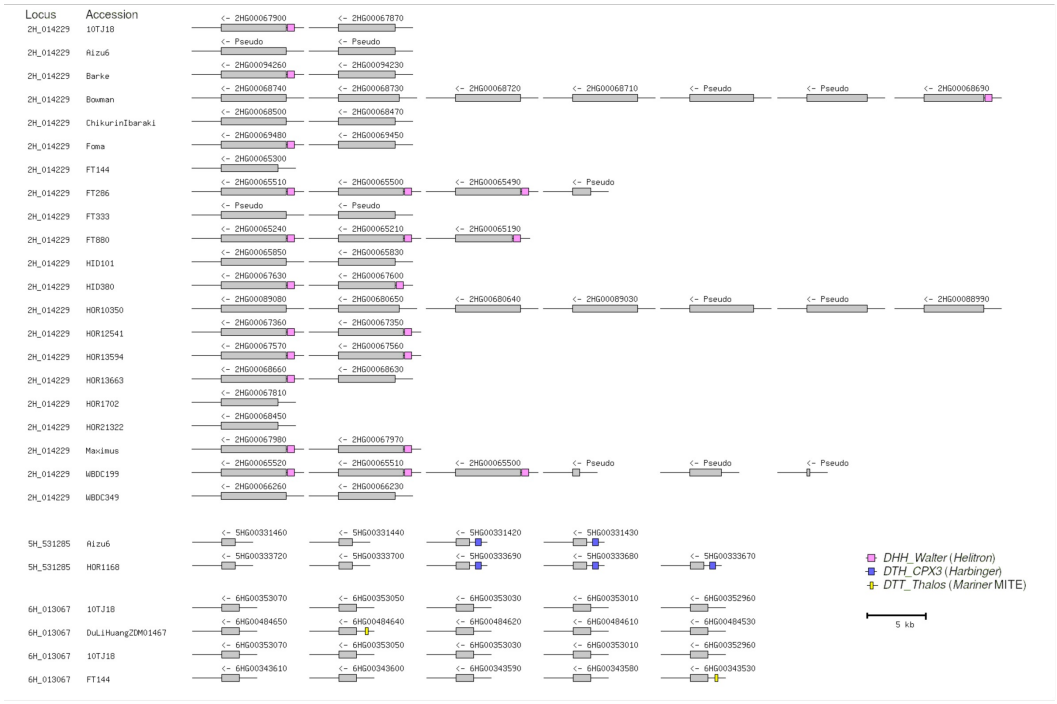



**Supplementary Figure 7. Structure and variability of complex loci. (a)** Length distribution of repeated units in structurally complex loci, ascertained by self-alignment. Peaks marked with dashed lines correspond to approximately 700, 1000, 1,550, and 1,840 bp. **(b)** Average lengths and variances of structurally complex loci across pangenome accessions. Only loci identified in at least 60 accessions were considered. Each point represents a locus.

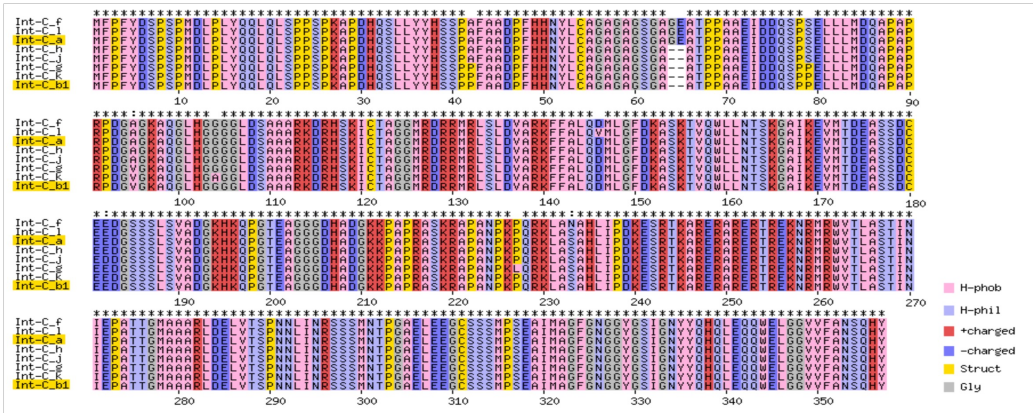



**Supplementary Figure 8: Read depth at complex loci.** Each cell in the heatmap shows the average per-bp read depth at one complex locus and one pangenome accession. The distribution of read depth is centered around the genomic median, i.e. the per-bp median coverage across the entire genome.

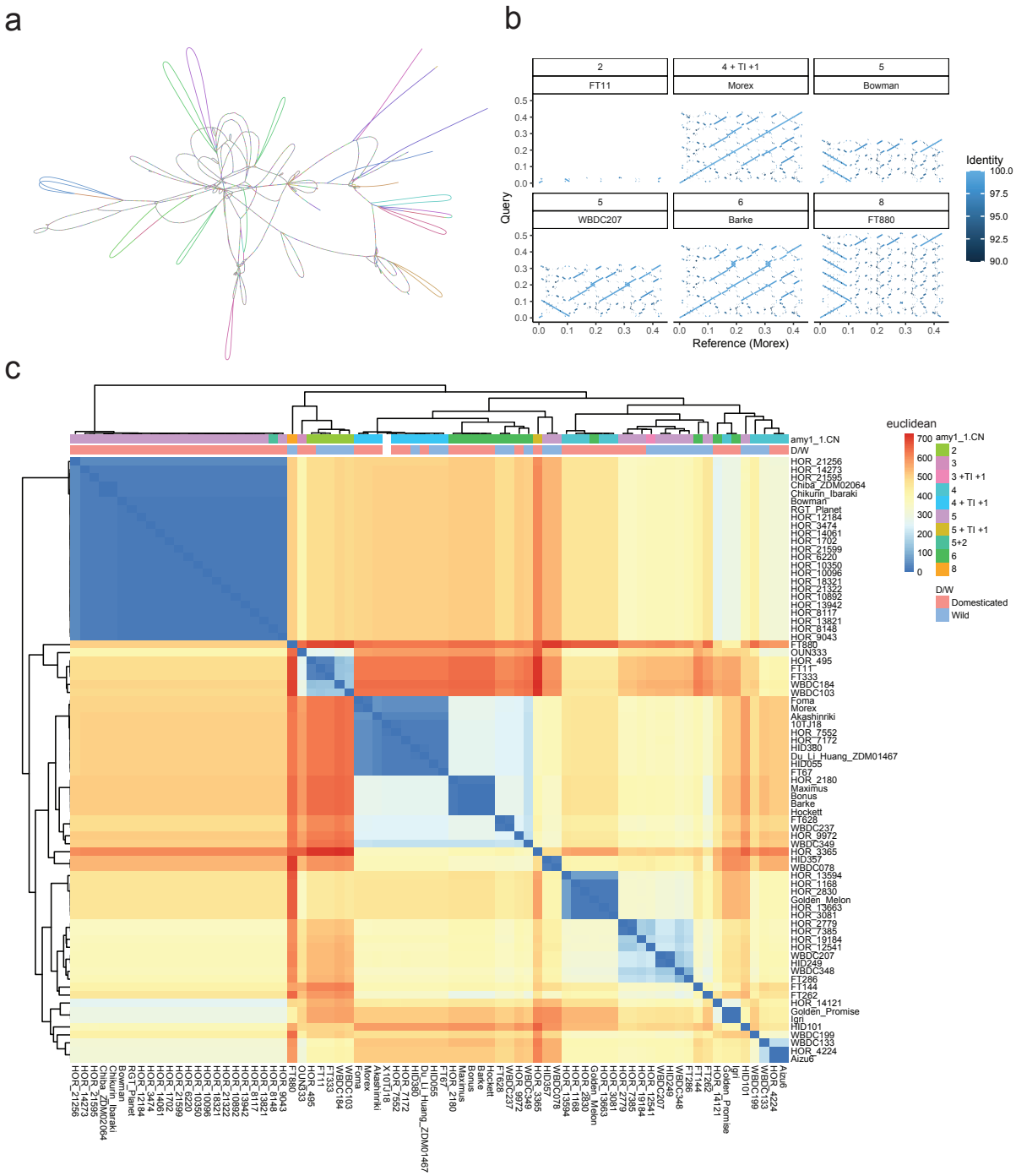
**a**



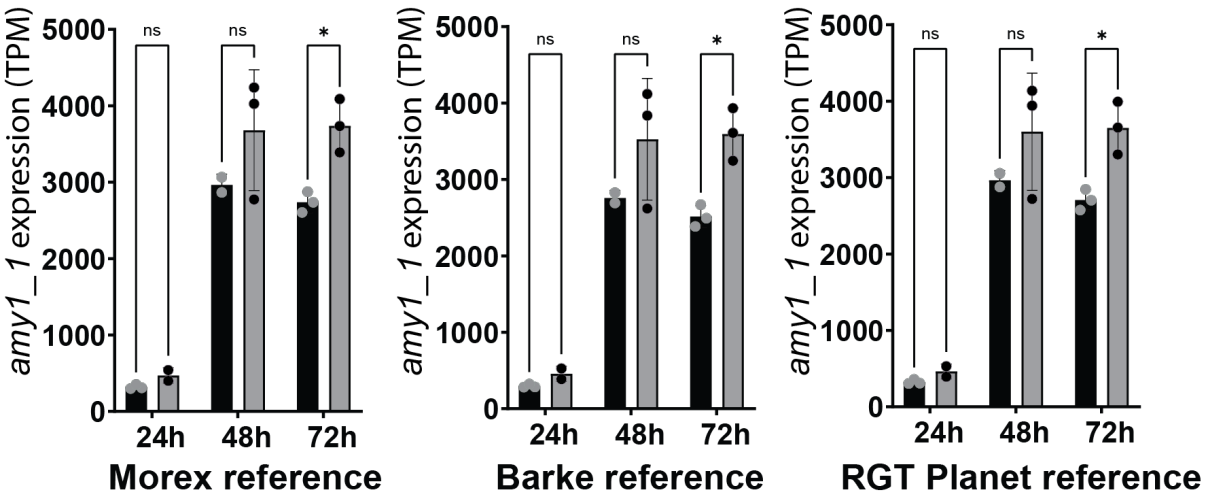

**b**



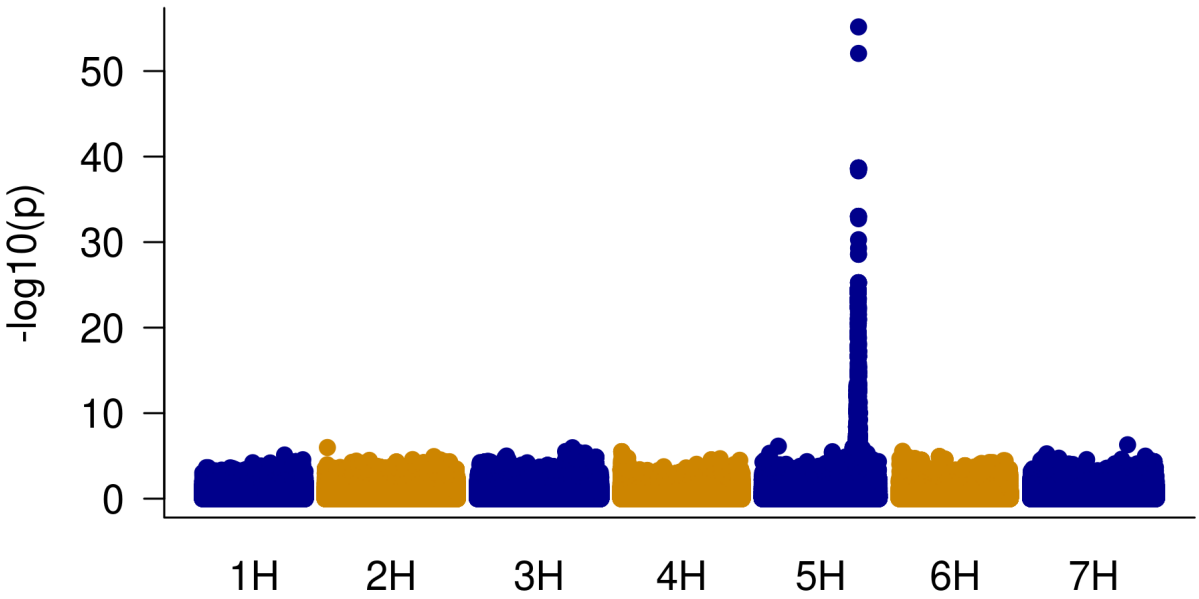

**Supplementary Figure 9: Assessment of sequence variants of individual genes in complex loci. (a)** Analysis of promoter sequences in tandem repeated genes in complex loci. We identified six instances where individual gene copies contain TE insertions in their promoters that are absent from other copies in a given gene cluster. Shown are three example loci. The genes are indicated by boxes and TEs by coloured boxes. Transcriptional orientation of genes are indicated by arrowheads next to the gene name For each locus, one representative of the identified haplotypes is shown (for example, for locus 5H_53185 there were two haplotypes which differ inthe number of gene copies that contain the TE insertion). When we searched for these TE presence/absence polymorphisms, we also identified 35 loci in which promoters of gene copies harbor insertions/deletions polymorphisms longer than 30 bp, which are not associated with TEs (not shown). **(b)** Predicted protein variants of *Int-c* (*HvTB1*) genes. Previously described alleles are highlighted in yellow. Color code: H-phob: Hydrophopic aa, H-phil: Hydrophilic aa, +charged: positively charged aa, - charged: negatively charged aa, Struct: structural aa, Cystein or Prolin, Gly: Gycin.




**Supplementary Figure 10: Local pangenome graph at *amy1_1*. (a)** A two-dimensional representation of the pangenome graph was drawn with Bandage. **(b)** Local alignments of the *amy1_1* locus between six representative haplotypes with different copy numbers and the reference cultivar Morex. **(c)** Clustering of *amy1_1* sequences according to structural features in the graph-based pangenome.

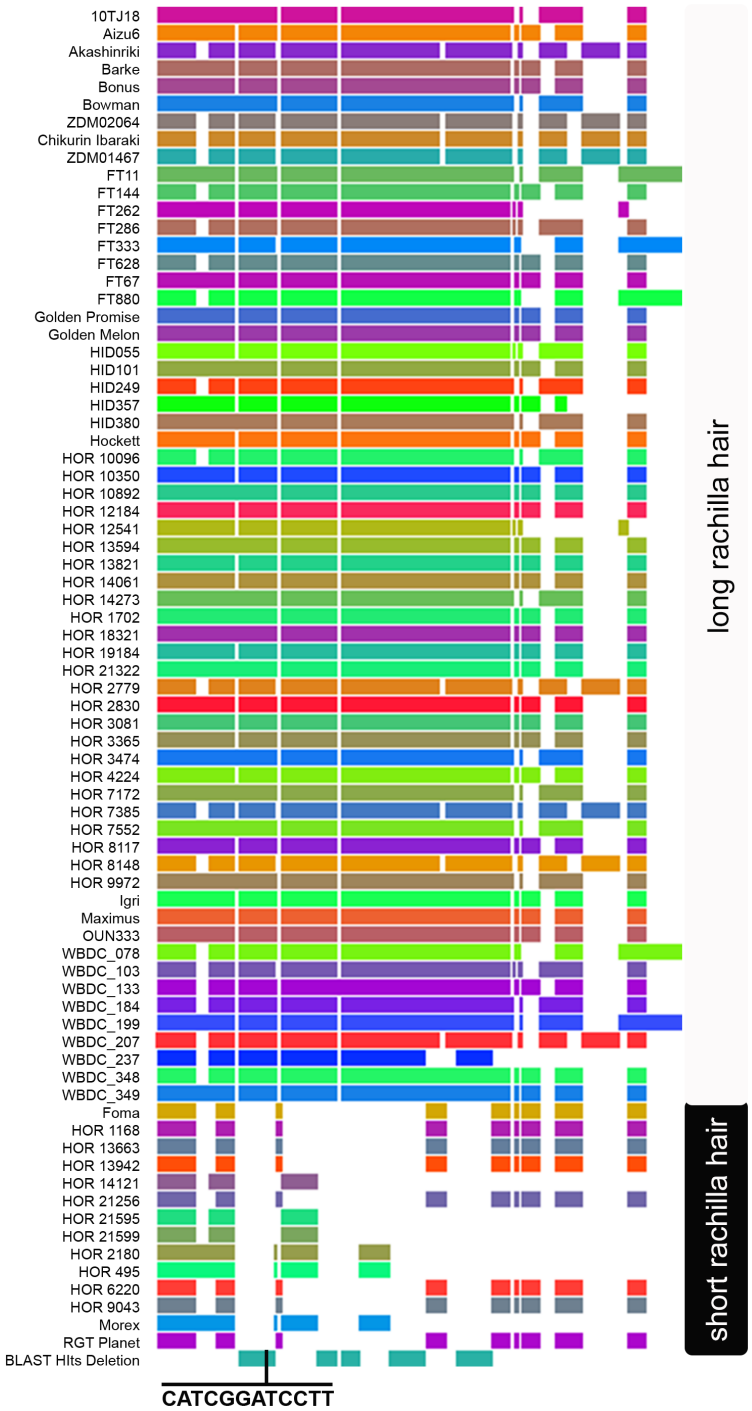



**Supplementary Figure 11:** *Amy1_1* gene expression of RGT Planet and *amy1_1*-Barke NIL during micro-malting. The panels show *amy1_1* expression in micro-malted grains of two European haplotypes (RGT Planet in black and its NIL with *amy1_1*-Barke haplotype in grey) after 24h, 48h and 72h. RNA-seq reads were mapped to three different reference. Due to multiple identical copies at this locus, expression level was presented as sum of TPM (transcripts per million reads mapped) from all a*my1_1* copies in each sample. To avoid bias from different *amy1_1* haplotypes in reference transcriptomes, quantification was done against the annotated Morex, Barke and RGT Planet reference genome sequences. Data was analysed by a mixed-effects model and the P value was adjusted using Tukey's multiple comparison test (*p ≤0.05, **p ≤0.01, ***p ≤ 0.001, ****p ≤ 0.0001, n=3 independent samples examined in one environment).




**Supplementary Figure 12:** Genome-wide association scan for rachilla hair length in the core1000 panel, **Supplementary Table 6**).


**Supplementary Figure 13: Targeted mutagenesis at *HvSRH1.*** **(a)** Structure of the *srh1* candidate gene with the only exon shown as a grey box and localization of target motifs 1a and 1b in a distance of 87 bp between the expected cleavage sites of Cas9. Target motif 1a is located at the non-coding strand and 1b at the coding strand. **(b)** Chromatogram of Sanger sequencing of PCR amplicons of target motif 1a in Golden Promise (GP) wild-type (wt) and the primary mutant brh72_E21 (**Supplementary Table 27**). Double peaks in the chromatogram indicate heterozygous insertions and/ or deletions. **(c)** Homozygous mutations in two M2 offspring of the plant E21. Dashes indicate deleted nucleotides, red letters indicate insertions, asterisks indicate unclear signal due to double peaks, protospacer adjacent NGG motif doubled underlined, target motif underlined, scissors indicate cleavage site of Cas9.




**Supplementary Figure 14: Local pangenome graph at *HvSRH1.*** ODGI plot of a minigraph-based pan-genome graph for the SV-affected region of the *Srh1* locus. Each coloured bar represents the respective haplotype graph of each of the 76 BPGv2 accessions (colours are assigned randomly). Accessions are arranged by rachilla hair phenotype (see **Supplementary Table 27**). The last (light blue) bar at the bottom represents the nodes identified by a BLAST run of the *srh1* enhancer region from cultivar Barke against all nodes in the graph. Multiple haplotypes span the region but none of the short-haired accessions' paths cover the nodes identified in Barke as associated with the long-haired *srh1* phenotype. The position of the conserved enhancer element is indicated.