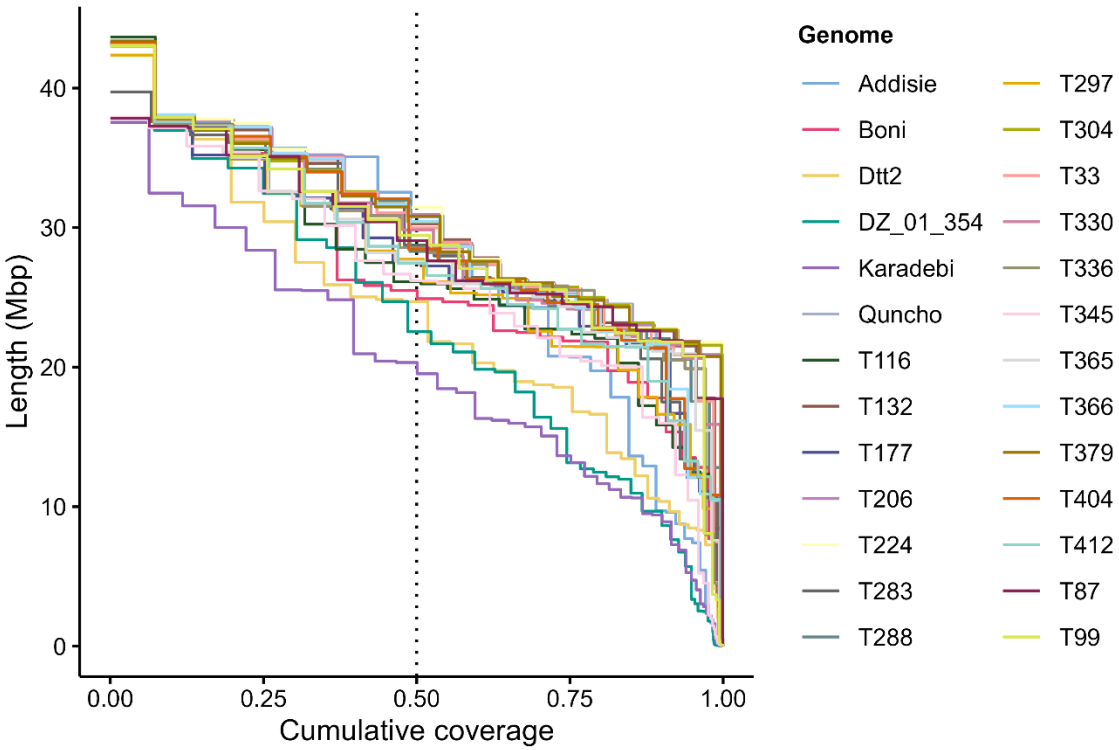
Supplementary information

The teff (*Eragrostis tef*) pangenome reveals haplotypic diversity and targets for molecular improvement

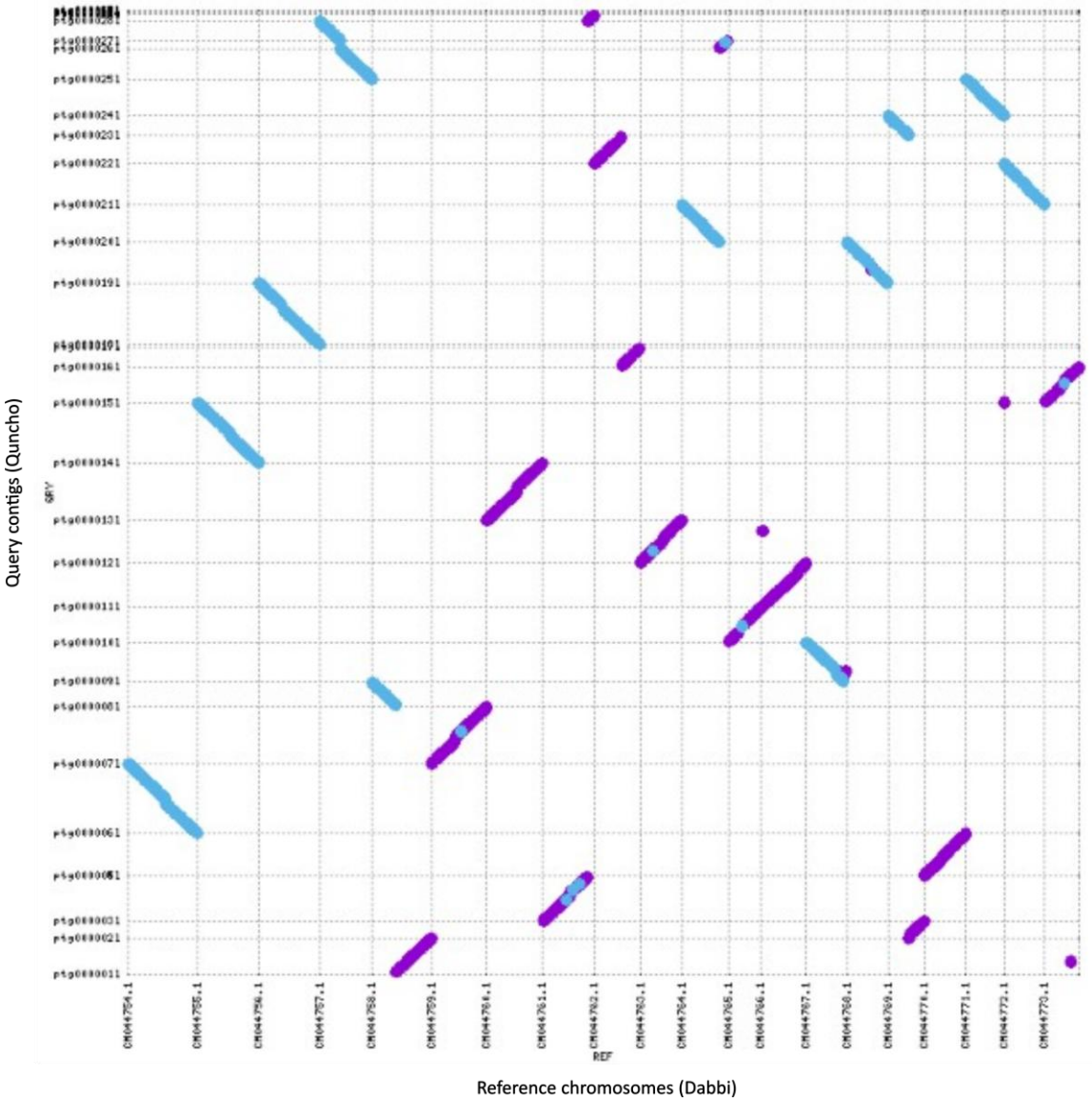
Correspondence to:
m.dellacqua@santannapisa.it
o.shorinola@bham.ac.uk

Contains Supplementary Figures 1-20 and Supplementary Tables 1-16.

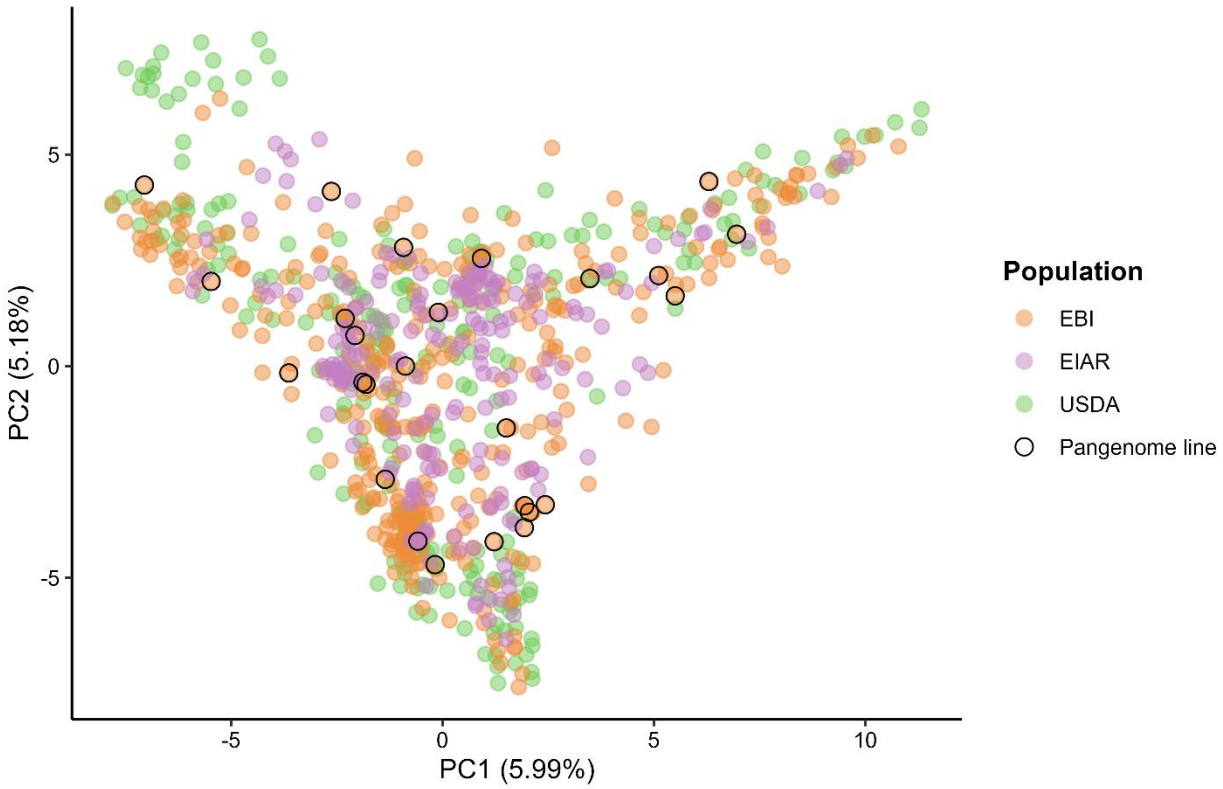
Supplementary Figures 1-20



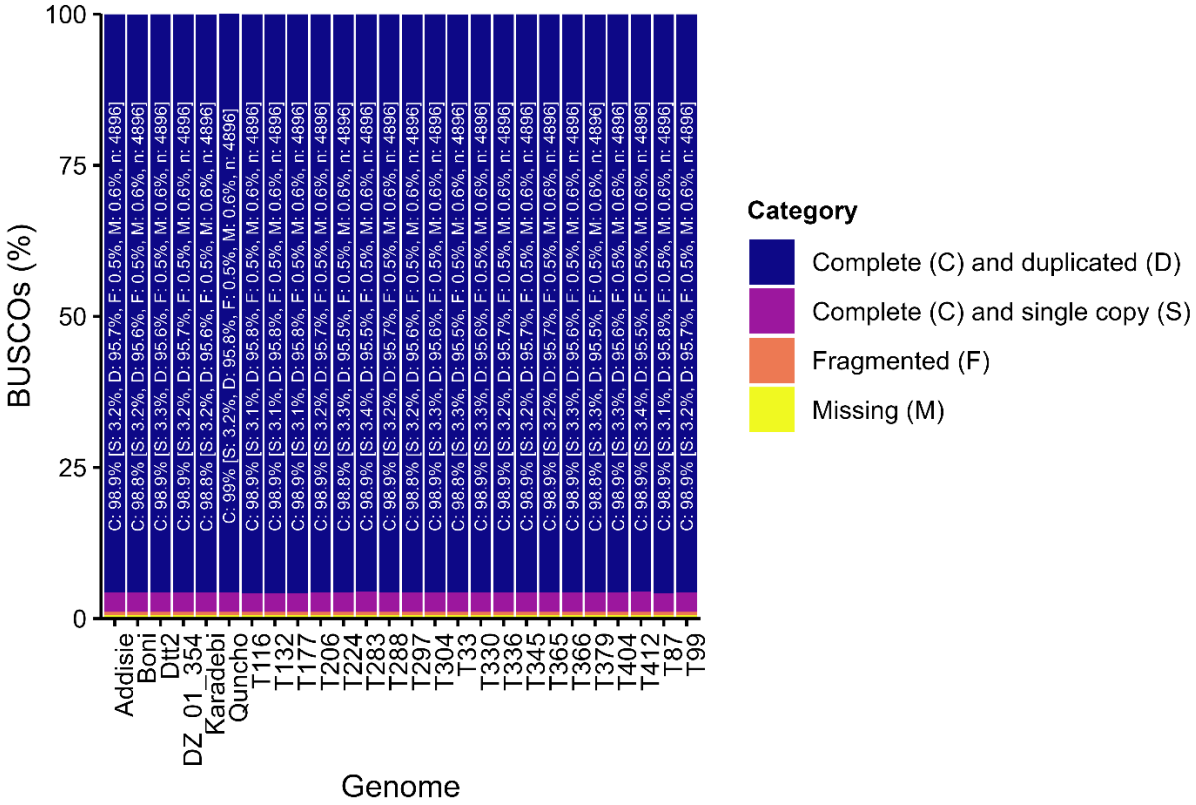
Supplementary Fig. 1| Assembly contiguity. Contiguity of assemblies in the teff pangenome, with colors according to legend. The vertical dotted line marks the N50 value. The height of the curve at this intersection represents the N50 contigs length (Mbp) for 26 assembled teff pangenome accessions.



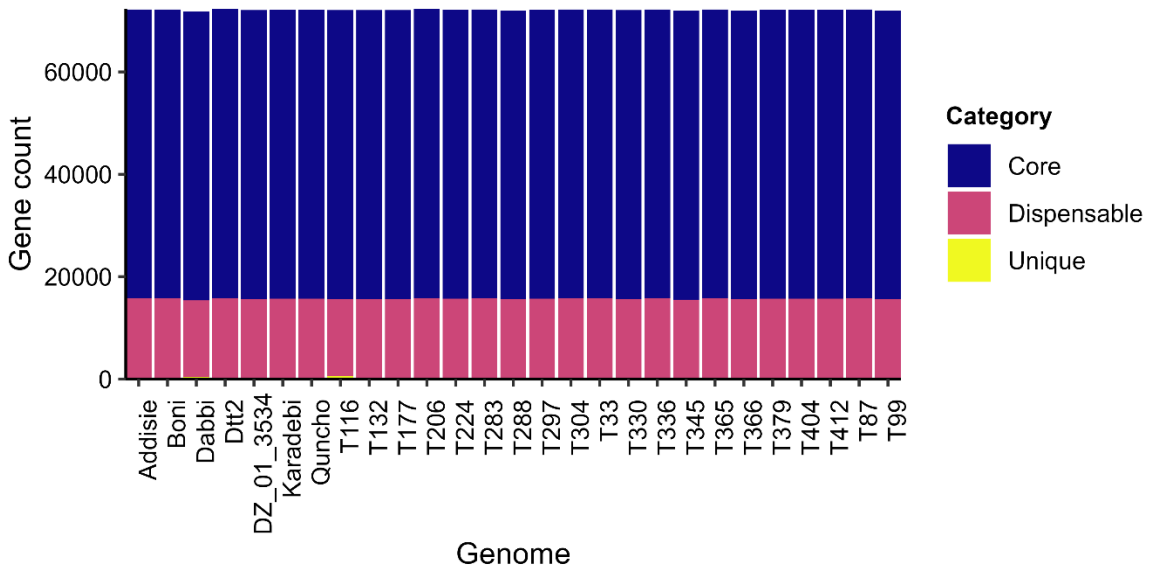
Supplementary Fig. 2 | Assembly contiguity. Dot-plot displaying comparison between Quncho *de novo* assembly contigs (y-axis) and the previously available Dabbi reference chromosomes (x-axis). Diagonal lines represent the alignments between contigs, blue and purple according to direction.



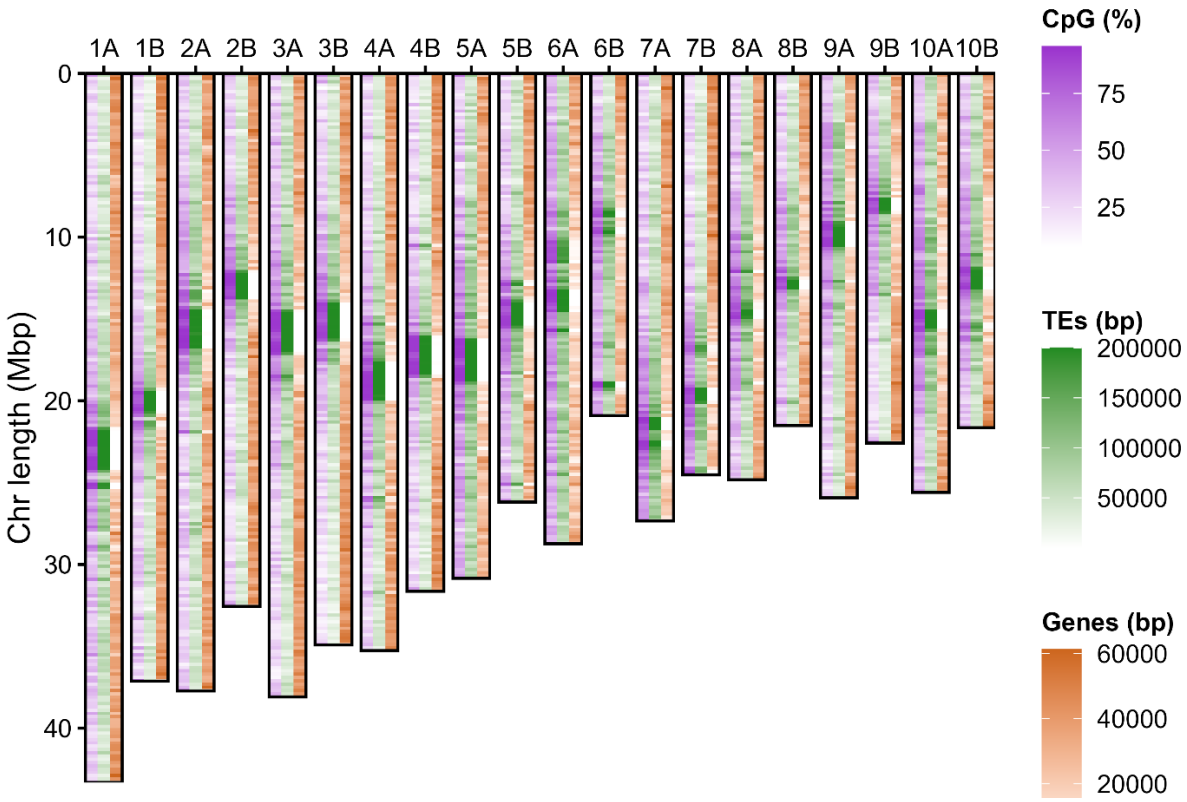
Supplementary Fig. 3 | Global teff diversity. Principal component analysis (PCA) showing the distribution of teff accessions across three major sources of ex-situ teff germplasm: Ethiopian Biodiversity Institute (EBI), Ethiopian Institute of Agricultural Research (EIAR), United States Department of Agriculture (USDA). Black open circles indicate the accessions included in the pangenome



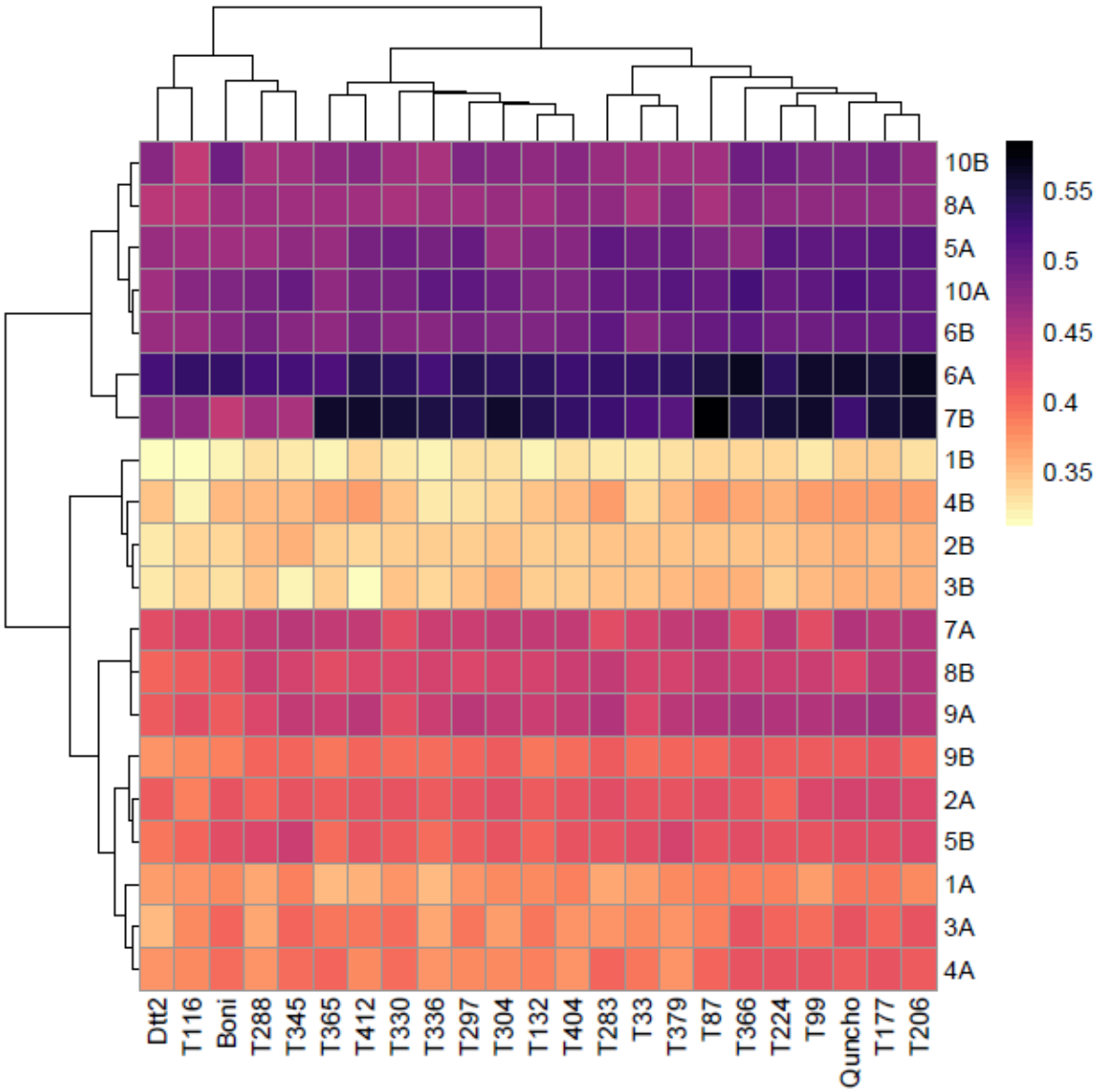
Supplementary Fig. 4 | Assembly completeness. Stacked bar displaying Benchmarking Universal Single-Copy Ortholog (BUSCO) score for 26 assembled teff pangenome accessions. Assessment using the *poales*_odb10 lineage dataset. The Bars color represents the percentage of complete and duplicated (dark blue), complete and single copy (purple), fragmented (orange) and missing (yellow).



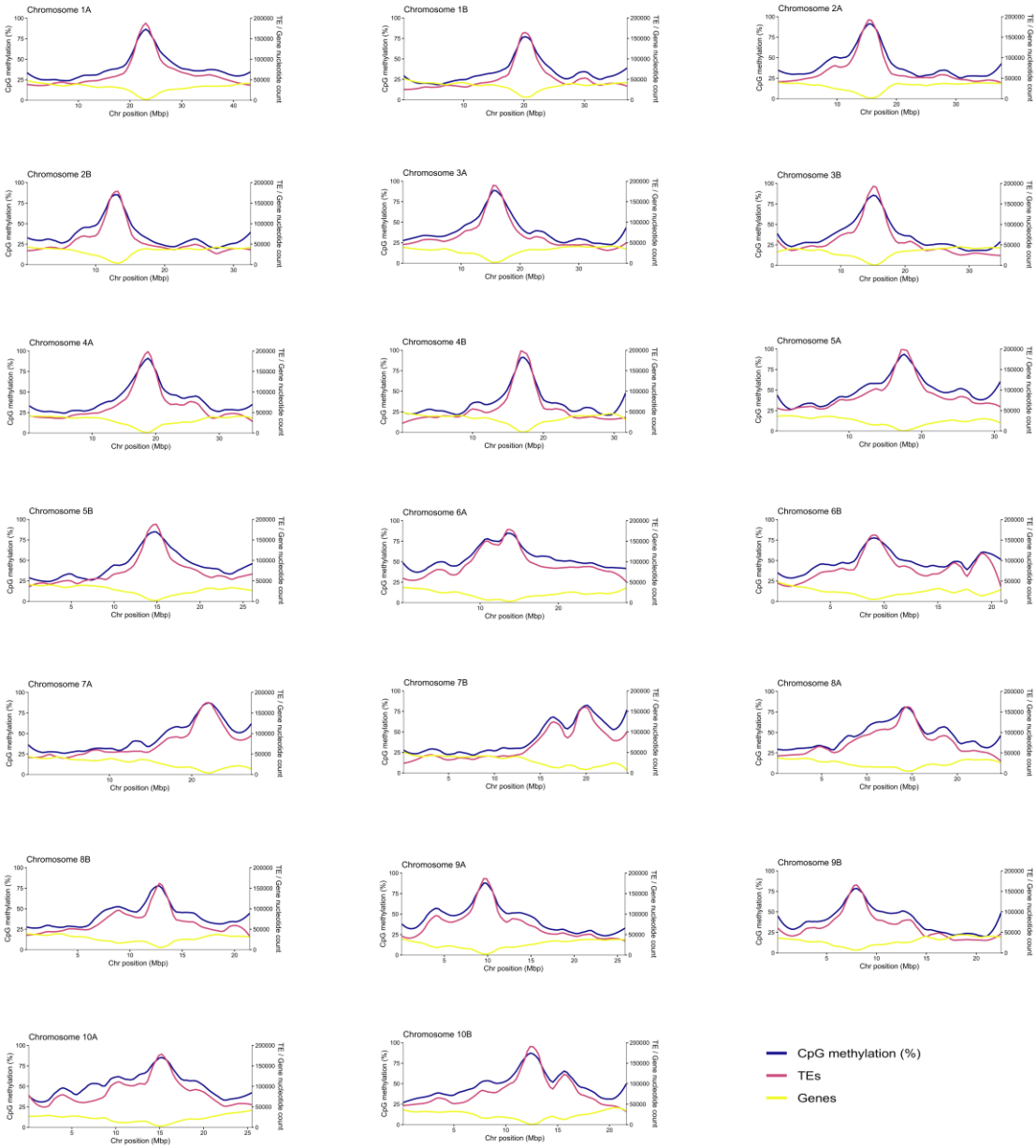
Supplementary Fig. 5 | Pangenome gene content. Distribution of core, dispensable and unique genes across all pangenome accessions, with colors according to legend.



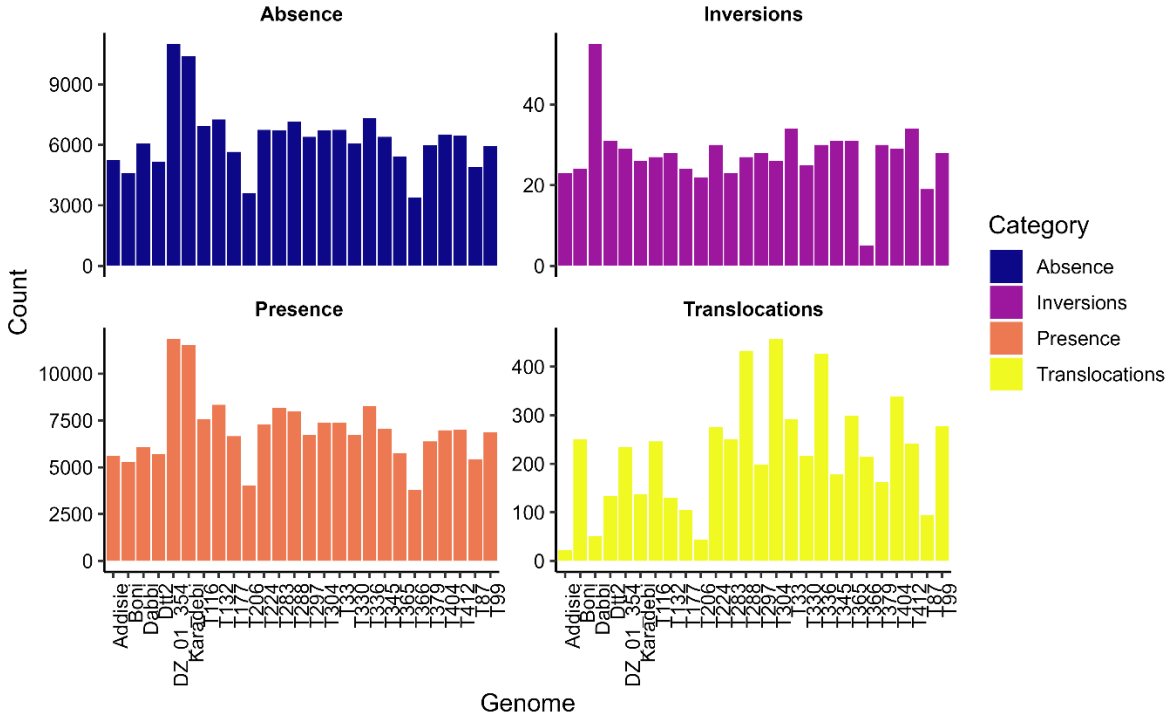
Supplementary Fig. 6 | Methylation density in Quncho genome. Genes (bp), TEs (bp) and CpG methylation (%) density across Quncho chromosomes (200 kbp window size).



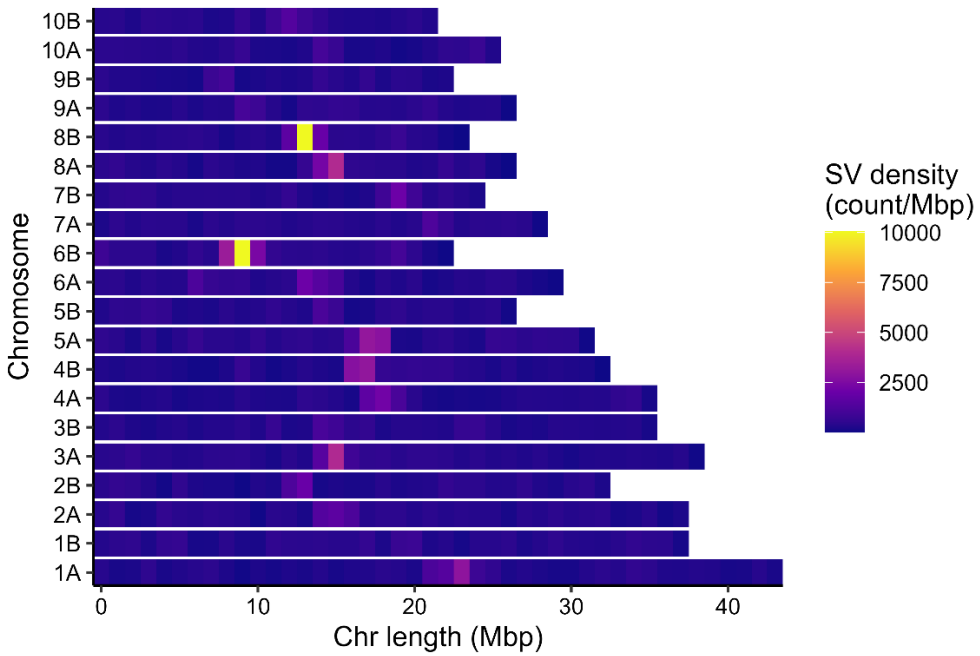
Supplementary Fig. 7 | Methylation landscape across teff pangenome accessions. Heatmap of mean chromosomal methylation densities across genomes. Hierarchical clustering using Euclidean distance.



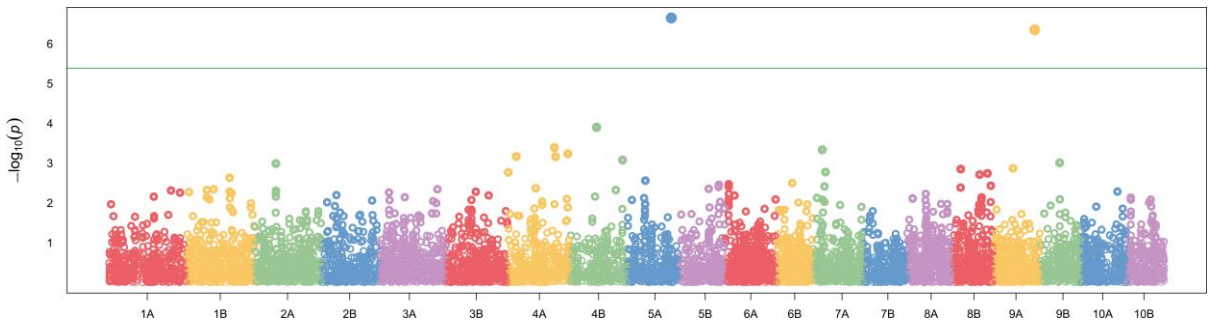
Supplementary Fig. 8 | Genome wide distribution of methylation, gene content, and transposable elements in Quncho. Line plots showing the relationship between CpG methylation (%, blue), TE density (pink) and gene density (yellow) across all 20 Quncho chromosomes. The left y-axis indicates methylation percentage, while the right y-axis indicates nucleotide count per 200 kb windows and x-axis chromosome position. Lines are smoothed using LOESS (span = 0.2).



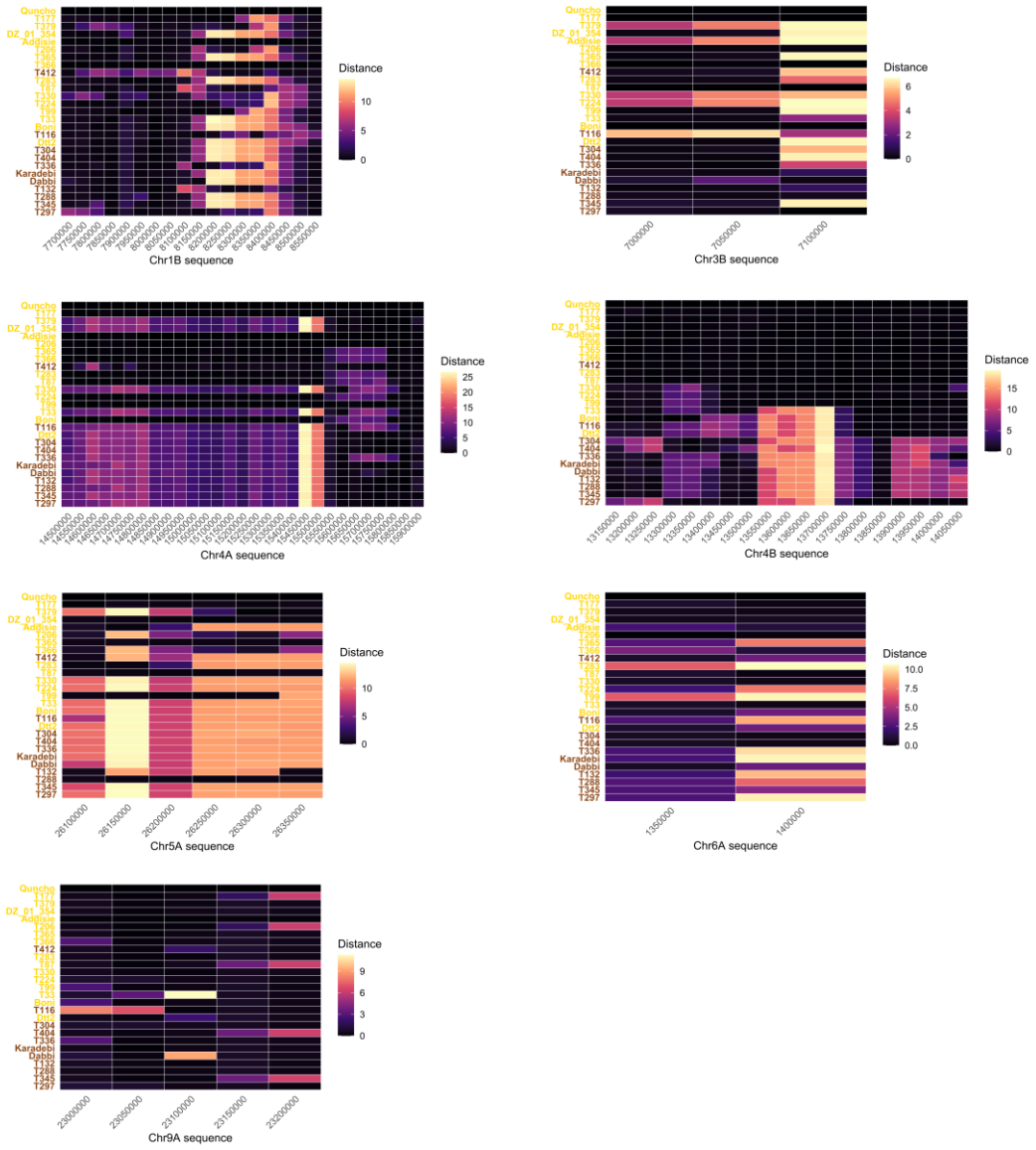
Supplementary Fig. 9 | Structural variation in the teff pangenome accessions. Bar plots showing the total count of SVs (size greater 50 bp) identified in each of the 26 pangenome accessions plus Dabbi relative to the Quncho genome, categorized by variants type (presence, absence, translocations and inversions). Absence SVs included copy loss (CPL), deletions (DEL), copy-loss duplications (DUP/INVDP), and reference-specific sequence highly divergent regions (HDR), unaligned regions (NOTAL), and tandem repeat (TDM). Presence SVs included insertions (INS), copy gains (CPG), copy-gain duplications (DUP/INVDP), and query-specific sequences in HDR, NOTAL, and TDM regions. Inversions (INV) and translocations and inverted translocations (TRANS/INVTR) were grouped as translocations.



Supplementary Fig. 10 | Structural variation density across the teff chromosomes. SV density (count per 1Mbp windows) across the 20 chromosomes of the teff pangenome accessions.

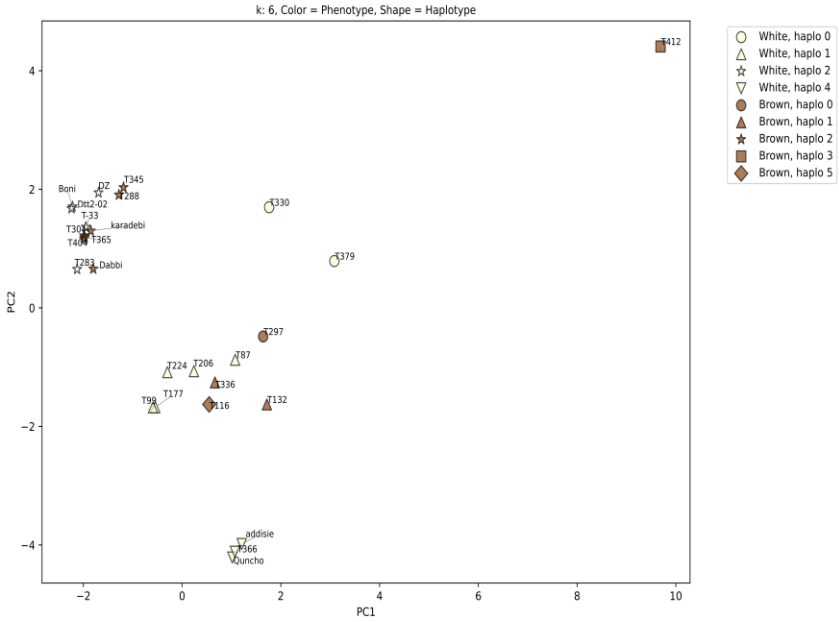


Supplementary Fig. 11 | Manhattan plot showing genome wide association study (GWAS) results for seed color on teff accessions from the EBI collection. Each dot represents a single nucleotide polymorphism (SNP), plotted by genomic position (*x*-axis) and -log10 (p-value) (*y*-axis). The horizontal green line marks the genome-wide significance threshold corresponding to a false discovery rate of 0.05. Chromosomes are shown in alternating colors and peaks highlight genomic regions significantly associated with the trait. We used the BLINK model.

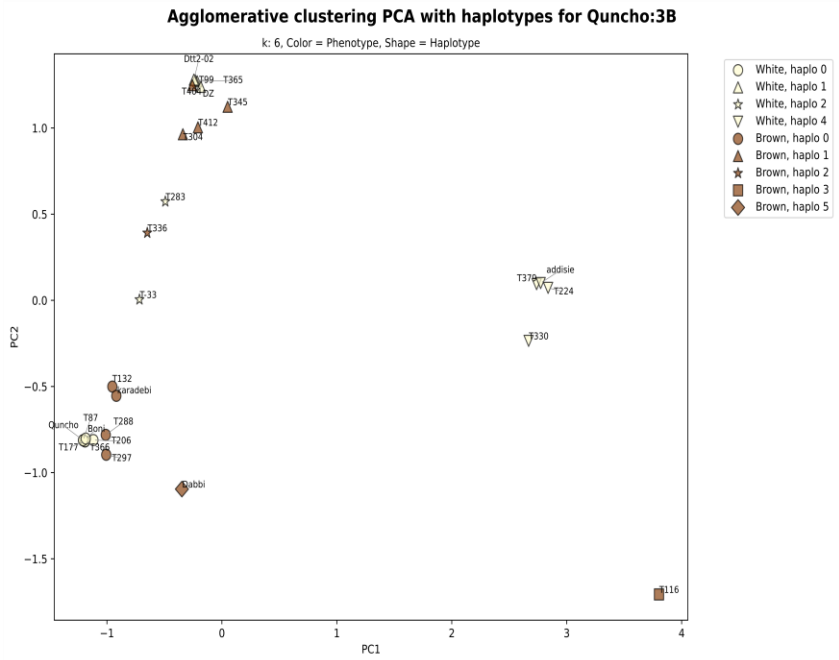


Supplementary Fig. 12 | Haplotype blocks underlying seed color variation on chromosome 1B, 3B, 4A, 4B, 5A, 6A and 9A. Heatmaps of *k*-mer-based variation profiles generated in 50 kbp windows relative to the reference accession Quncho across genomic intervals associated with phenotypic variation in seed color. The x-axis represents genomic coordinates (bp) and the y-axis represents pangenome accessions, with yellow and brown names distinguishing white and brown accessions. Cell colors show the level of distance to Quncho as reported in the legend

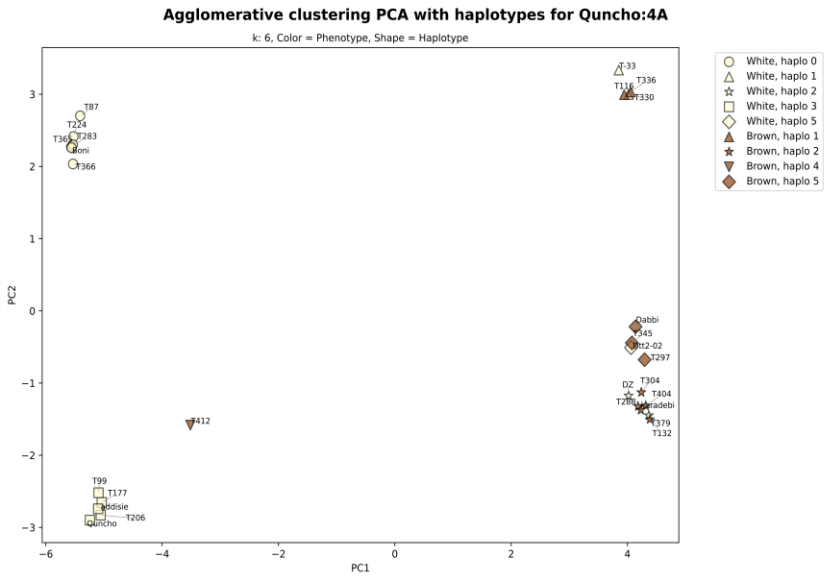
Agglomerative clustering PCA with haplotypes for Quncho:1B



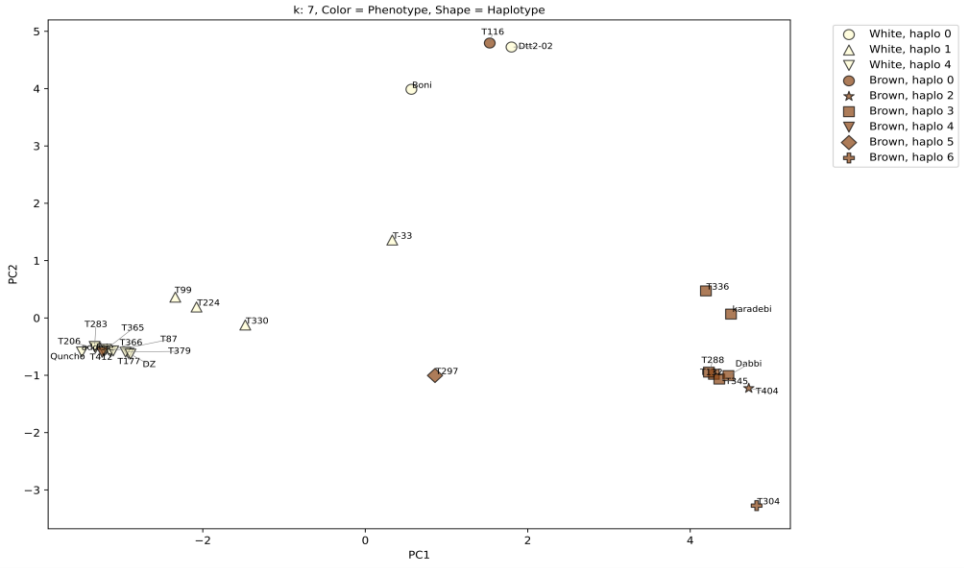
Agglomerative clustering PCA with haplotypes for Quncho:3B



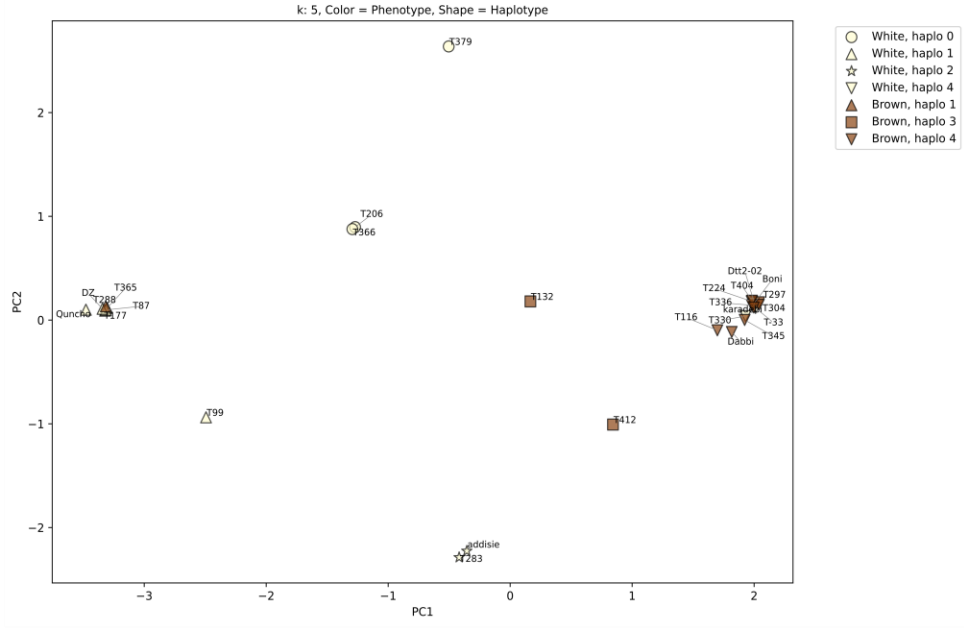
Agglomerative clustering PCA with haplotypes for Quncho:4A



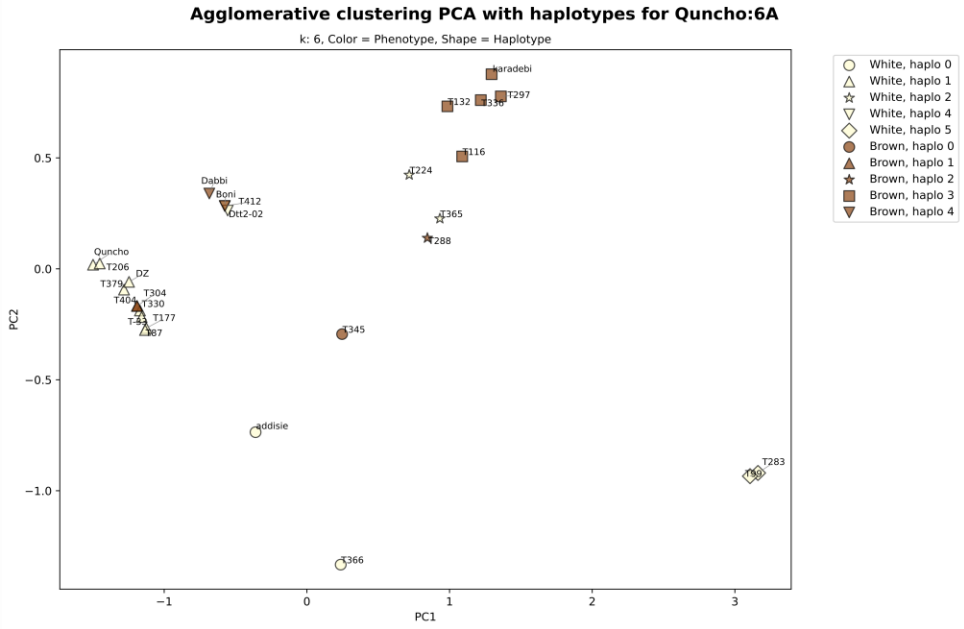
Agglomerative clustering PCA with haplotypes for Quncho:4B

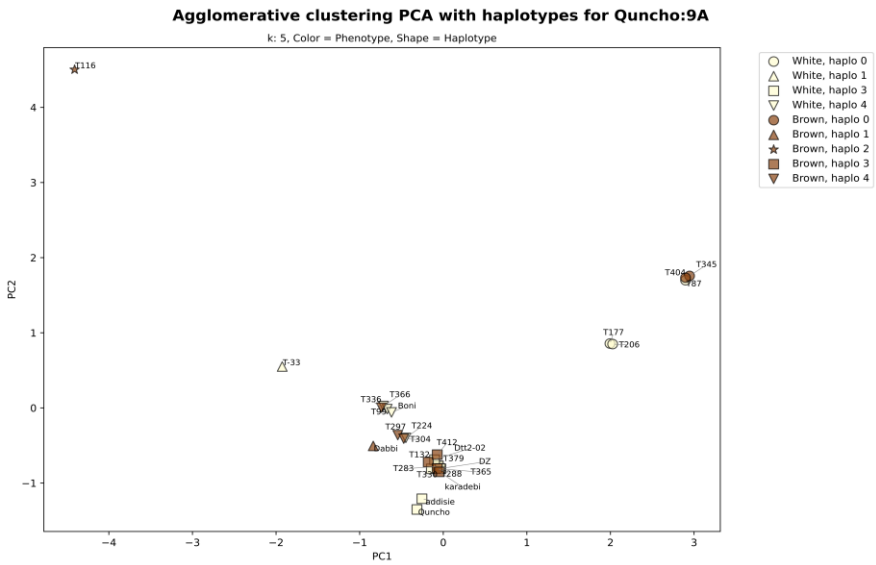


Agglomerative clustering PCA with haplotypes for Quncho:5A



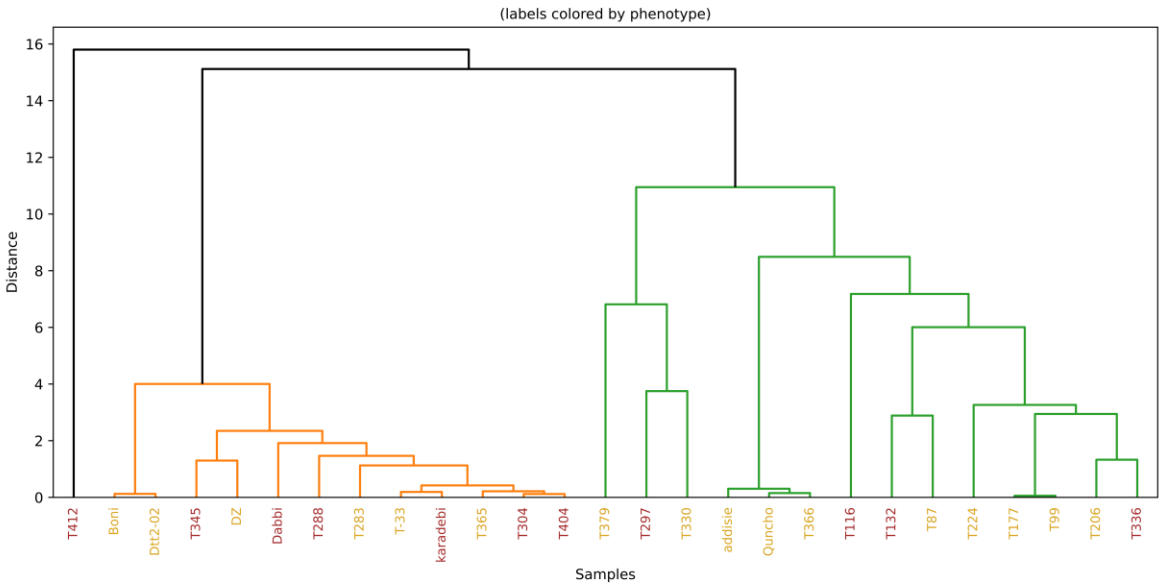
Agglomerative clustering PCA with haplotypes for Quncho:6A



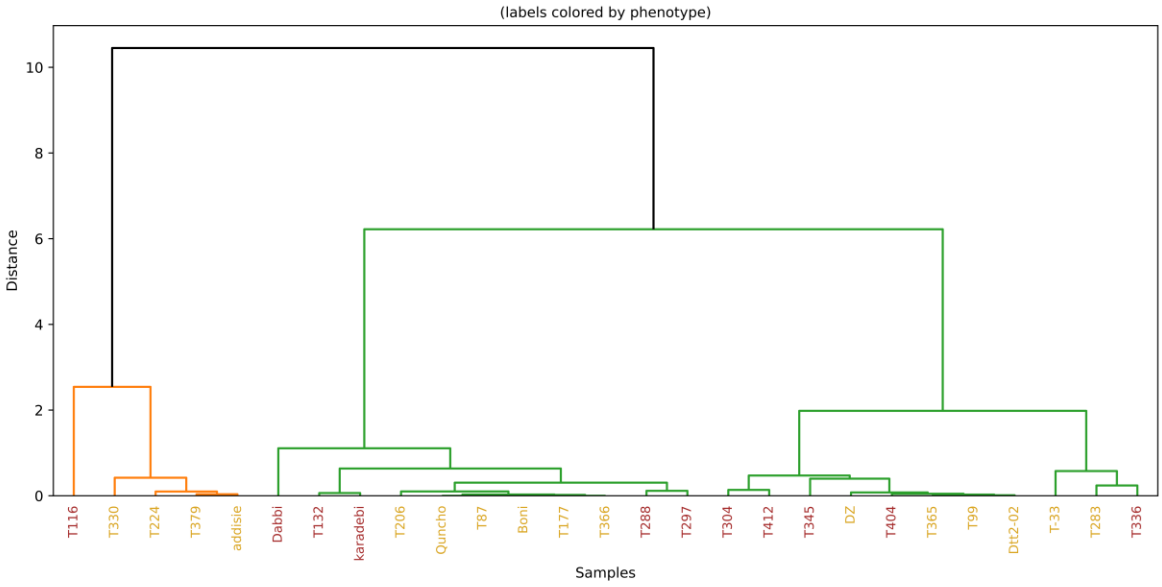


Supplementary Fig. 13 | Principal component analysis (PCA) of genomic interval associated with seed color variation on chromosome 1B, 3B, 4A, 4B, 5A, 6A and 9A. PCAs were performed on standardized k-mer count matrices, retaining principal components (PCs) explaining 90 % of the variance. PC1 and PC2 for each interval are shown. Colors indicate the seed color phenotype, while shapes denote specific haplotype groups.

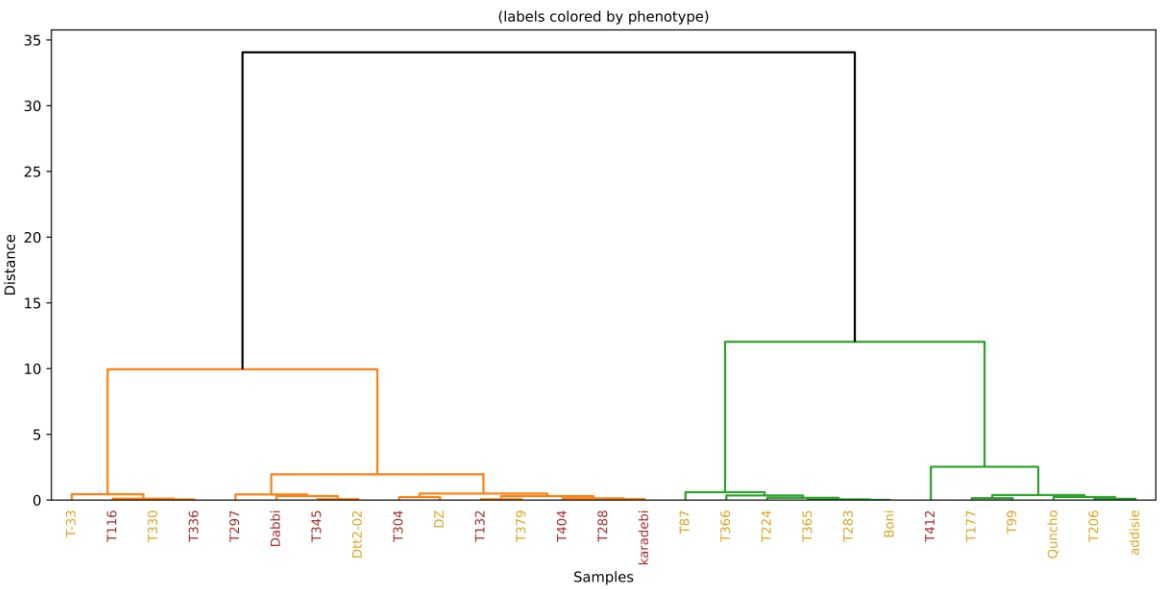
Hierarchical Clustering Dendrogram for Quncho:1B



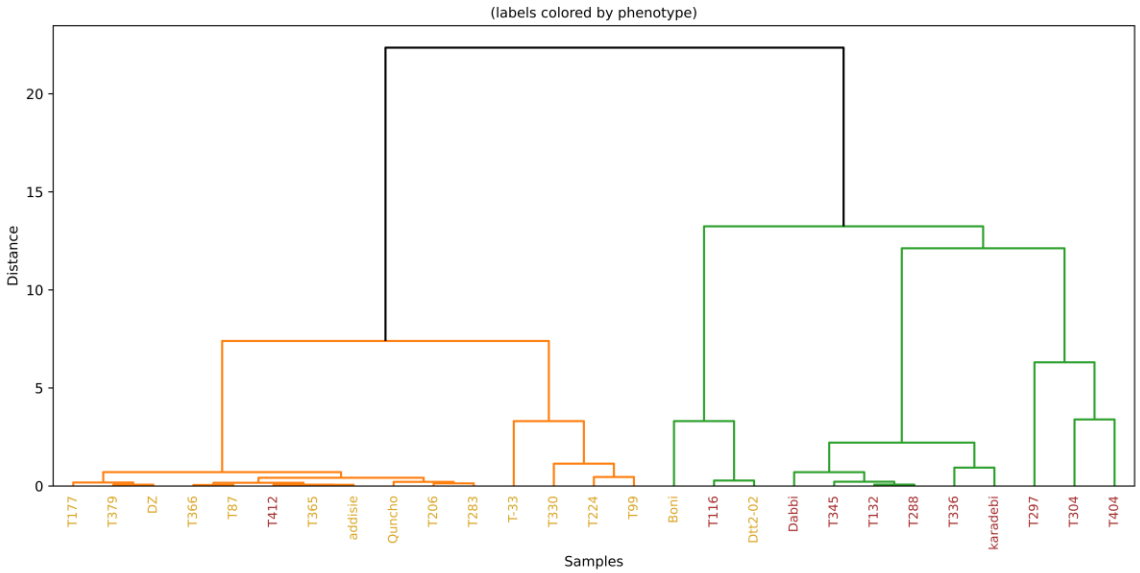
Hierarchical Clustering Dendrogram for Quncho:3B



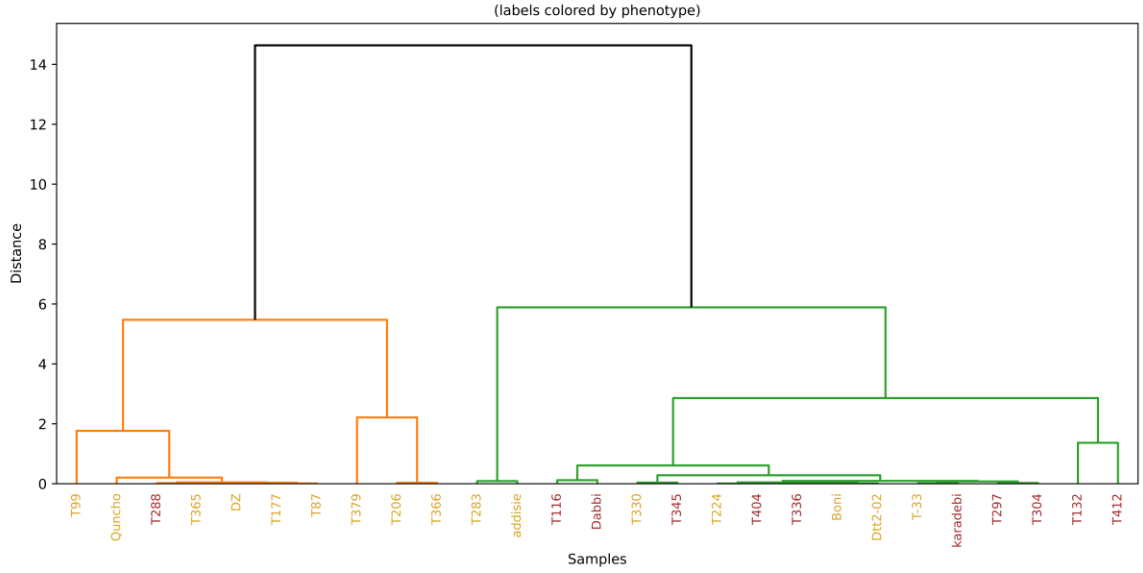
Hierarchical Clustering Dendrogram for Quncho:4A



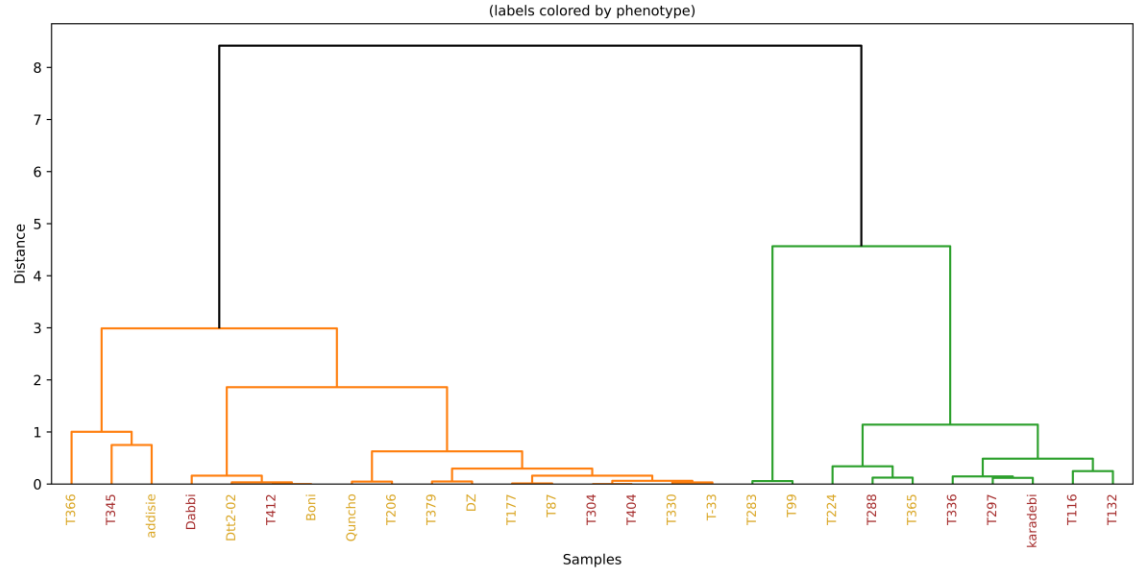
Hierarchical Clustering Dendrogram for Quncho:4B

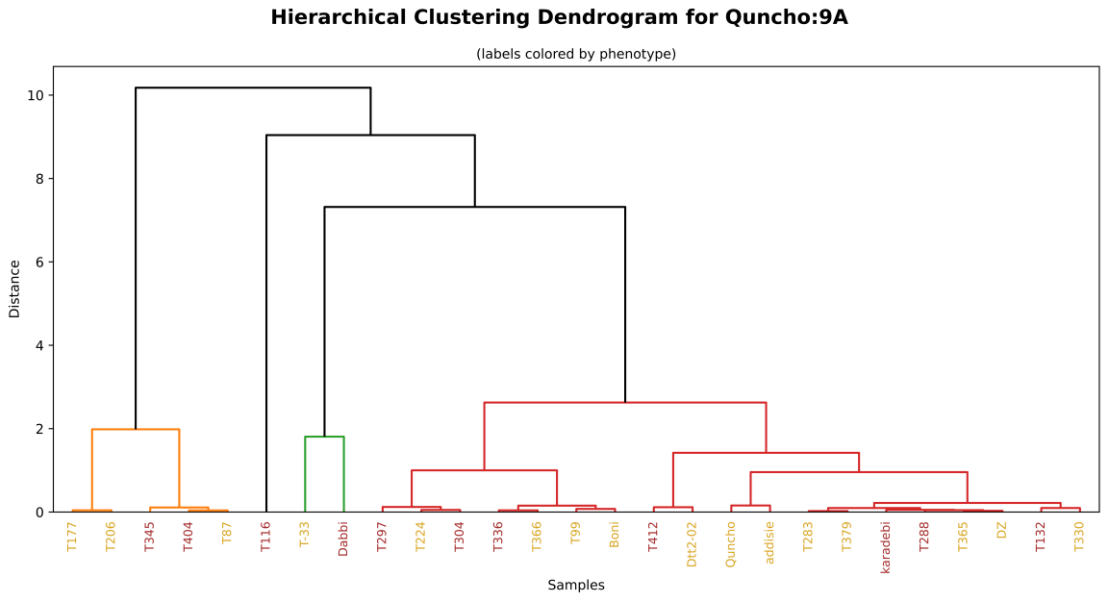


Hierarchical Clustering Dendrogram for Quncho:5A

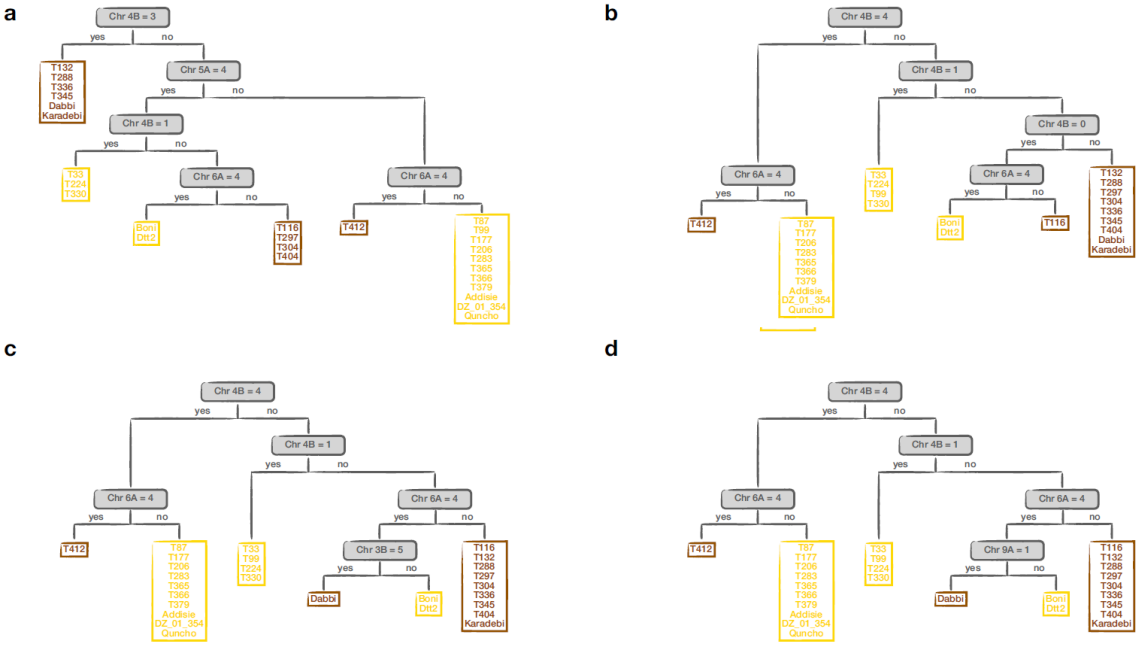


Hierarchical Clustering Dendrogram for Quncho:6A



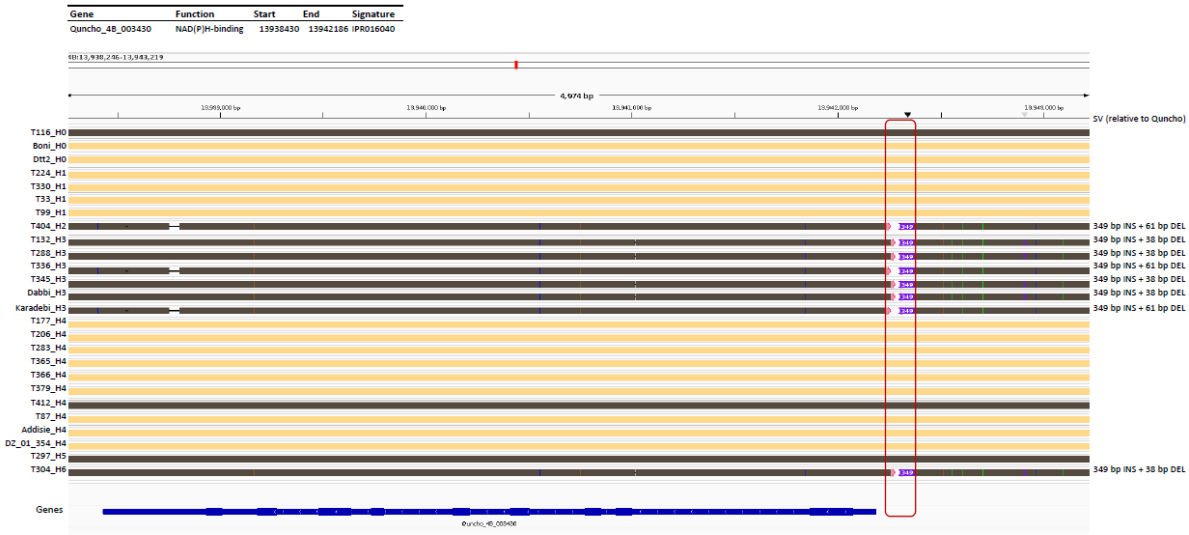


Supplementary Fig. 14 | Hierarchical clustering of genomic intervals corresponding to seed color variation on chromosome 1B, 3B, 4A, 4B, 5A, 6A and 9A. Dendrograms from Ward's hierarchical clustering of PCA scores for each genomic region. Haplotypes were defined by inspecting the dendrogram structure and selecting cut points that reflect stable genetic groupings. Pangenome accessions labels on the x-axis are colored according to observed phenotype.



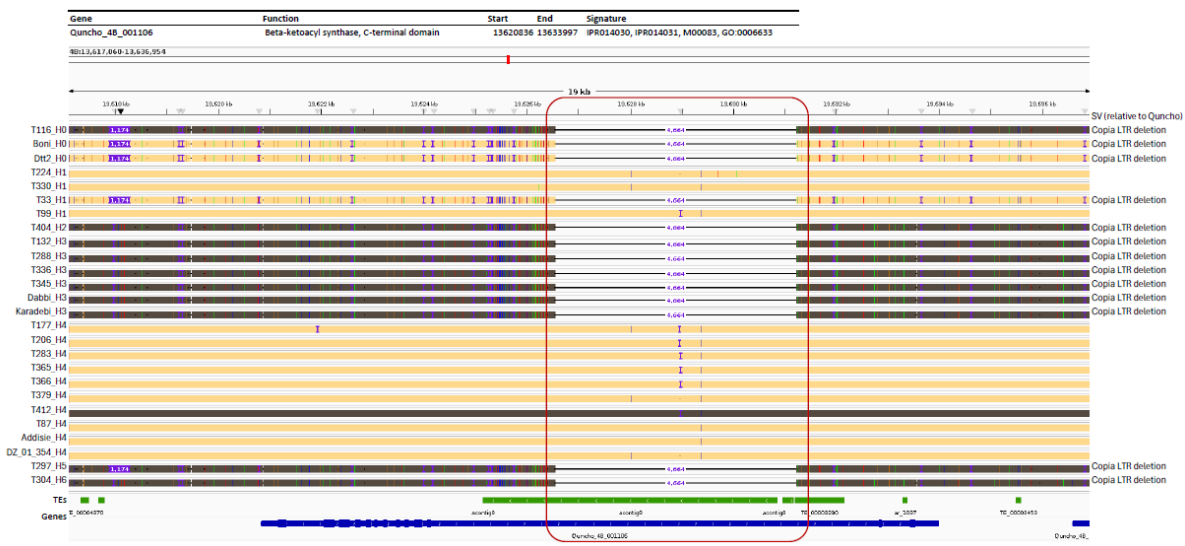
Supplementary Fig 15 | Decision trees for multi-locus haplotype combinations. a–d, Four alternative decision trees representing minimal subsets of haplotype indicators that separate 27 teff pangenome accessions (including Dabbi) into white and brown-seeded. At the top gray node (*e.g.* "Chr 4B = 3"), which displays a binary haplotype-level indicator. "Yes" branch if an accession possesses the haplotype, or the "no" branch if it does not. Continue through subsequent nodes until reaching a terminal box. The terminal box lists the accessions separated by that specific path of rules; yellow outlines for white-seeded and brown outlines for brown-seeded. The trees were generated following an exhaustive combinatorial search on 41 binary indicators to identify minimal subsets yielding perfect separation. To visualize joint rules, decision trees classifiers were grown with information-gain splitting and permissive settings (cp=0, minsplit=2, minbucket=1), serving as an explanatory representation rather than an independent estimate of predictive generalization.

Supplementary Fig. 16 | Integrated Genomics Viewer (IGV) visualization of all genomes aligned to Quncho at the locus on chromosome 4B containing a candidate anthocyanin biosynthesis gene. The IGV panel displays the alignment of all assemblies against Quncho, whose coordinates system is used as reference across the genomic interval 13,683,596– 13,685,475 bp. Accessions with brown seeds are indicated by dark-brown horizontal bars, whereas light-yellow bars represent white-seeded accessions. The bottom annotation green blocks display transposable element(s), the blue block represents the annotated gene model (*Quncho_4B_001110*). Labels H0 to H6 appended to accession names indicate the assigned haplotype groups. The red vertical box highlights SVs relative to Quncho detailed on the right.

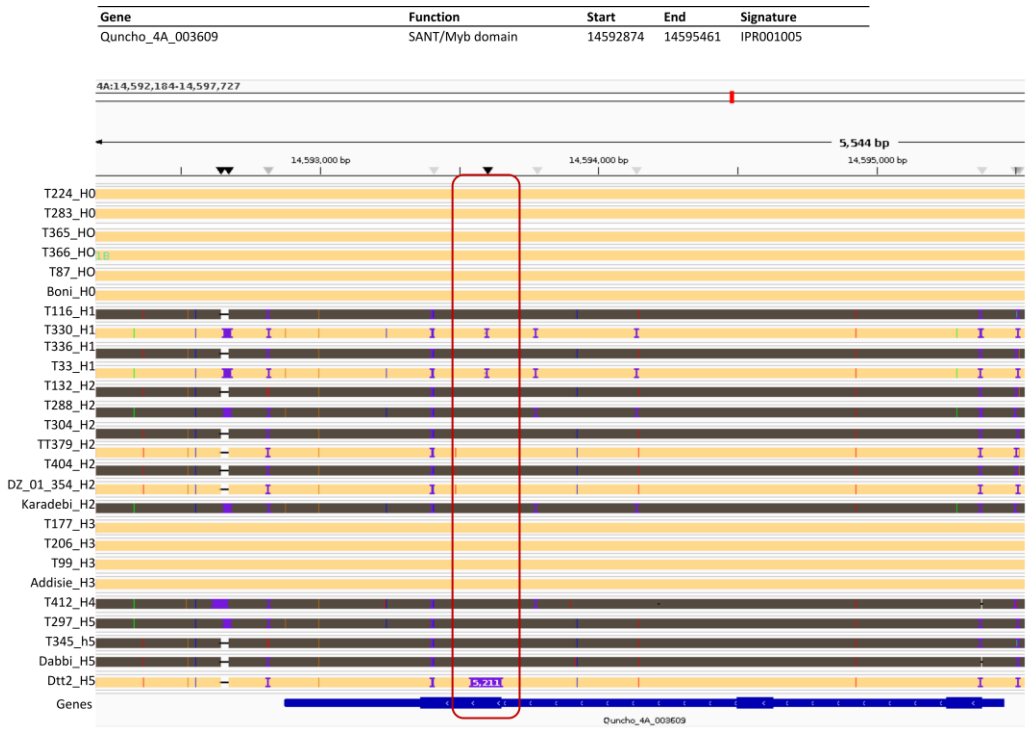


Supplementary Fig. 17 | Integrated Genomics Viewer (IGV) visualization of all genomes aligned to Quncho at the locus on chromosome 4B containing a candidate anthocyanin biosynthesis gene. The IGV panel displays the alignment of all assemblies against Quncho, whose coordinates system is used as reference across the genomic interval 13,938,430 – 13,9421,86 bp. Accessions with brown seeds are indicated by dark-brown horizontal bars, whereas light-yellow bars represent white-seeded accessions. The bottom annotation blue block represents the annotated gene model (*Quncho_4B_003430*). Labels H0 to H6 appended to accession names indicate the assigned haplotype group. The red vertical box highlights SVs relative to Quncho detailed on the right.

Supplementary Fig. 18 | Integrated Genomics Viewer (IGV) visualization of all genomes aligned to Quncho at the locus on chromosome 4B containing a candidate anthocyanin biosynthesis gene. The IGV panel displays the alignment of all assemblies against Quncho, whose coordinates system is used as reference across the genomic interval 13,917,425 – 13,925,620 bp. Accessions with brown seeds are indicated by dark-brown horizontal bars, whereas light-yellow bars represent white-seeded accessions. The bottom annotation green blocks display transposable elements (Helitron, DNA transposon and LTR /Copia), the blue blocks represent the annotated gene models (*Quncho_4B_003433* and *Quncho_4B_003432*). Labels H0 to H6 appended to accession names indicate the assigned haplotype groups. The red vertical box marks SVs relative to Quncho, which are detailed on the right.



Supplementary Fig.19 | Integrated Genomics Viewer (IGV) visualization of all genomes aligned to Quncho at the locus on chromosome 4B containing a candidate fatty acid biosynthesis gene. The IGV panel displays the alignment of all assemblies against Quncho, whose coordinates system is used as reference across the genomic interval 13,620,836 – 13,633,997 bp. Accessions with brown seeds are indicated by dark-brown horizontal bars, whereas light-yellow bars represent white-seeded accessions. The bottom annotation green blocks display transposable element (LTR /Copia), the blue block represents the annotated gene model (*Quncho_4B_001106).* Labels H0 to H6 appended to accession names indicate the assigned haplotype groups. The red vertical box highlights structural variants (SVs) relative to Quncho detailed on the right.



Supplementary Fig. 20 | Integrated Genomics Viewer (IGV) visualization of all genomes aligned to Quncho at the locus on chromosome 4A containing a candidate fatty acid biosynthesis gene. The IGV panel displays the alignment of all assemblies against Quncho, whose coordinates system is used as reference across the genomic interval 14,592,874 – 14,595,461 bp. Accessions with brown seeds are indicated by dark-brown horizontal bars, whereas light-yellow bars represent white-seeded accessions. The bottom annotation blue block represents the annotated gene model (*Quncho_4A_003609).* Labels H0 to H6 appended to accession names indicate the assigned haplotype groups. The red vertical box highlights the region with an insertion of 5,211 bp specific of Dtt2.

Supplementary Tables 1-16

Supplementary Table 1. Sequencing, assembly and annotation statistics for teff pangenome accessions
Supplementary Table 2. Telomeric repeat (AAACCCT) organization across 20 chromosomes in 23 teff accessions
Supplementary Table 3. Passport information for teff pangenome accessions
Supplementary Table 4. Synteny and structural rearrangements between Quncho A and B subgenomes
Supplementary Table 5. BUSCO assessment of gene annotations
Supplementary Table 6. Core, dispensable and unique gene count per genome
Supplementary Table 7. Gene Ontology functional annotations of core, dispensable and unique gene sets
Supplementary Table 8. Repetitive element annotations across teff pangenome accessions
Supplementary Table 9. Structural variants (≥ 50 bp) in teff accessions relative to the Quncho reference genome
Supplementary Table 10. Length distribution of structural variants (≥ 50 bp) by type across the teff pangenome
Supplementary Table 11. Chromosomal distribution of structural variant (≥ 50 bp) counts
Supplementary Table 12. Syntenic regions across teff accessions relative to the Quncho reference genome
Supplementary Table 13. GWAS-identified chromosomal intervals associated with teff seed color
Supplementary Table 14. Predictor standardization on classification accuracy and model complexity of seed color haplotypes
Supplementary Table 15. Chromosomal distribution of repetitive elements in the Quncho genome
Supplementary Table 16. Functional annotation of candidate genes associated with seed color pigmentation