New Phytologist

***New Phytologist* Supporting Information**

Article title: Genomic consequences of artificial selection during early domestication of a wood fibre crop
Authors: Marja M. Mostert-O'Neill, Hannah Tate, S. Melissa Reynolds, Makobatjatji M. Mphahlele, Gert van den Berg, Steve D. Verryn, Juan J. Acosta, Justin O. Borevitz and Alexander A. Myburg
Article acceptance date: 20 May 2022

The following Supporting Information is available for this article:

Fig. S1 Population structure in relation to wild *E. grandis* and other species in section *Latoangulatae* based on PCA, DAPC and sNMF

Fig. S2 Breeding *E. grandis* population structure for all breeding samples, those excluding introgressed, and those excluding infused individuals in relation to the wild progenitor populations based on PCA, sNMF and DAPC analyses

Fig. S3 Population differentiation F_{ST} estimates among breeding *E. grandis*, wild *E. grandis*, and other species in section *Latoangulatae*

Fig. S4 Chloroplast (cp) haplotype network based on 24 cp SNPs

Fig. S5 Marker-specific Hardy-Weinberg Equilibrium (HWE) signed R values of wild vs breeding populations

Fig. S6 Genomic outliers and LD plots per chromosome

Fig. S7 Breeding population LD decay over genomic distance in kb

Fig. S8 Outlier detection by *pcadapt* scan

Table S1 Ancestry assignment of chromosomal segments (supplementary Excel file)

Table S2 Cluster assignment of samples using DAPC to identify genetically infused breeding individuals (supplementary Excel file)

Table S3 Summary statistics of genetic diversity using *hierfstat* v. 0.04-22 (Goudet 2005)

Table S4 Wilcoxon signed rank test p-values supporting the alternative hypothesis, that the mean of the outliers was greater than the mean of the rest of the SNPs

Table S5 Gene Ontology (GO) enrichment analysis for genes in LD (within 2 kb) with outlier SNPs against the SNP-captured gene space before excluding organellar-targeting SNPs

Table S6 BLASTn against the organellar genomes (supplementary Excel file)

Table S7 Marker statistics of SNPs with multi-genome targets (supplementary Excel file)

SI References

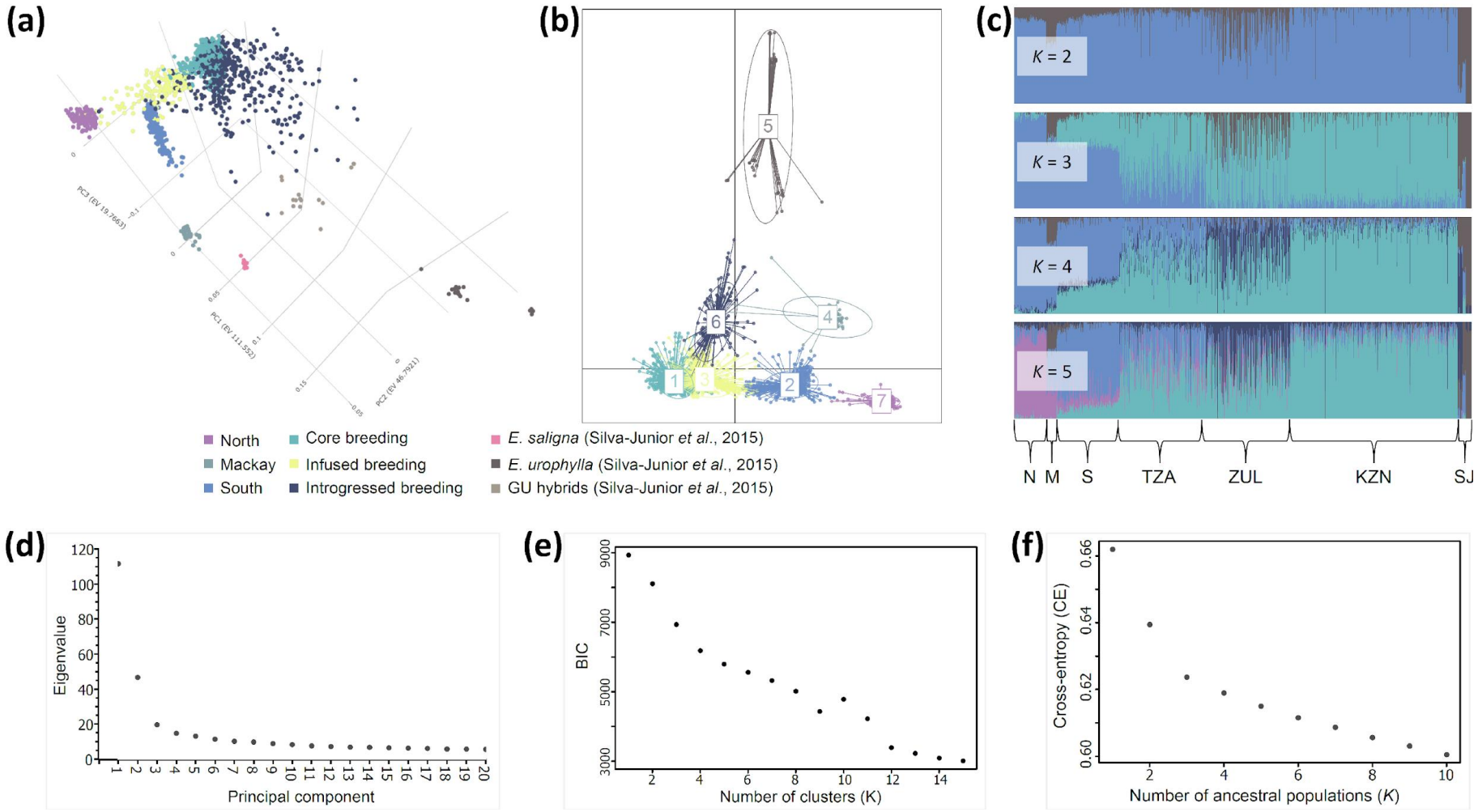

Fig. S1 Population structure in relation to wild *E. grandis* and other species in section *Latoangulatae* based on PCA, DAPC and sNMF. In **a**, the first three principle components of population structure PCA is given with breeding *E. grandis* individuals differentiated as core (turquoise), infused (yellow) and introgressed (dark blue). An interactive version of this plot can be accessed at https://chart-studio.plotly.com/~Marja/119/#/. **b** gives DAPC clustering for *K* = 7 along the first two principal components with 95% inertia ellipses for each cluster. At *K* = 7, the clusters separated breeding samples into core (turquoise), infused (yellow) and introgressed (dark blue) groups. All *E. urophylla*, *E. saligna* and *GU hybrids* grouped in cluster 7 (dark grey). Values of *K* beyond *K* = 7 resulted in poorly resolved clusters that did not allow biological interpretation. **c** – sNMF genomic composition plots for *K* = 2 to *K* = 5 illustrating genomic proportions (y-axes) assigned to ancestral populations for each individual (x-axis) from the North (N), Mackay (M), South (S) wild

progenitor populations (Mostert-O'Neill *et al.*, 2021), the three breeding populations, TZA, ZUL and KZN, and GU hybrids, *E. saligna* and *E. urophylla* (SJ) obtained from Silva-Junior *et al.* (2015). At *K* = 3, represented as a "knee" in the cross-entropy (CE) plot (**f)**, ancestry assignment supports *E. urophylla* (dark grey), wild *E. grandis* (blue) or breeding *E. grandis* (turquoise) ancestral populations, with clear evidence of introgression particularly in breeding population ZUL. At *K* = 5, the presence of genomic segments assigning to the North and South wild subpopulations are evident in several breeding samples, particularly in TZA. *E. saligna* individuals showed genomic composition assigned to *E. urophylla* (dark grey) and to the South wild subpopulation ancestry, supporting its close phylogenetic relation to *E. grandis*. The GU hybrid individuals had genomic composition assignment similar to introgressed breeding samples at *K* = 4 and *K* = 5. The scree- (**d**), Bayesian Information Criterion- (BIC) (**e**) and CE plots (**f**) were used to determine the number of principal components or ancestral populations (value of *K*) to be visualised in the PCA, DAPC and sNMF analyses, respectively. This was given as a clear "knee" in the plot as in **d** and **f** or as the minimal y-axis value that still resulted in well resolved, biologically relevant clustering of samples. For sNMF genomic composition plots, values of *K* up to *K* = 5 were interrogated as it revealed infusion from North and South wild subpopulations. These analyses were based on 24 306 informative SNPs.

New Phytologist

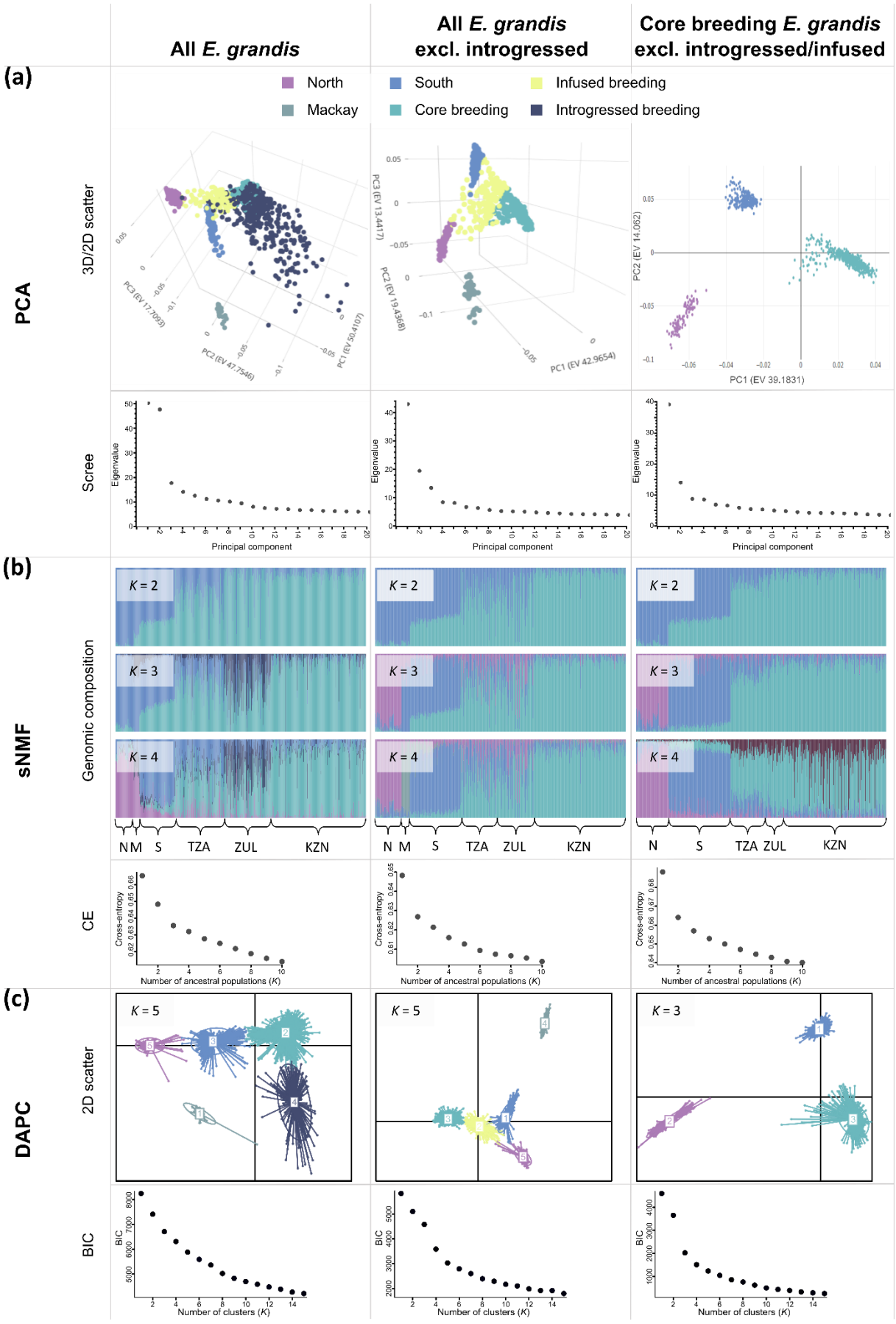

Fig. S2 Breeding *E. grandis* population structure for all breeding samples, those excluding introgressed, and those excluding infused individuals in relation to the wild progenitor populations based on PCA, sNMF and DAPC analyses. Three data sets were analysed; "All *E. grandis*" (using 23 661 informative SNPs), "All *E. grandis* excl. introgressed" (using 23 661 informative SNPs) and "*E. grandis* retained (for outlier detection)", in which introgressed and infused breeding individuals were removed and the Mackay wild subpopulation was excluded (using 21 991 informative SNPs). In the PCA (**a**) and DAPC (**c**) plots, breeding *E. grandis* individuals were differentiated as core (turquoise), infused (yellow) and introgressed (dark blue). Supporting scree- and Bayesian Information Criterion (BIC) plots are given in **a** and **c**, respectively. Interactive versions of the PCA plots can be accessed at https://chart-studio.plotly.com/~Marja/125/#/ for "All *E. grandis*", https://chart-studio.plotly.com/~Marja/127/#/ for "All *E. grandis* excl. introgressed", and https://chart-studio.plotly.com/~Marja/129/#/ for "*E. grandis* retained (for outlier detection)". **b**) sNMF genomic composition plots for $K = 2$ to $K = 4$ illustrating genomic proportions (y-axes) assigned to ancestral populations for each individual (x-axis) from the North (N), Mackay (M), South (S) wild progenitor populations, the three breeding populations, TZA, ZUL and KZN. The supporting Cross Entropy (CE) plots are given below the genomic composition plots. The scree- (**a**), CE- (**b**) and BIC (**c**) plots were used to determine the number of principal components or ancestral populations (value of K) to be visualised in the PCA, sNMF and DAPC analyses, respectively. This was given as a clear "knee" in the plot or as the minimal y-axis value that still resulted in well resolved, biologically relevant clustering of samples. Ancestry assignment supports wild *E. grandis* (blue) or breeding *E. grandis* (turquoise) ancestral populations at $K = 2$ in all three data sets. At $K = 3$, clear evidence of introgression can be observed in the "All *E. grandis*" data set sNMF analysis, particularly in breeding population ZUL (B2). Introgressed individuals also cluster separately from other breeding individuals at $K = 5$ in the DAPC analysis (**c**). The presence of infused genomic segments assigning to the North and South wild subpopulations are evident in several breeding samples, particularly in TZA in the sNMF analysis of "All *E. grandis*" and "All *E. grandis* excl. introgressed" data sets (**b**). Infused breeding individuals also clustered separately in the DAPC analysis of "All *E. grandis* excl. introgressed" at $K = 5$. Wild *E. grandis* subpopulations were described previously in Mostert-O'Neill *et al.* (2021).

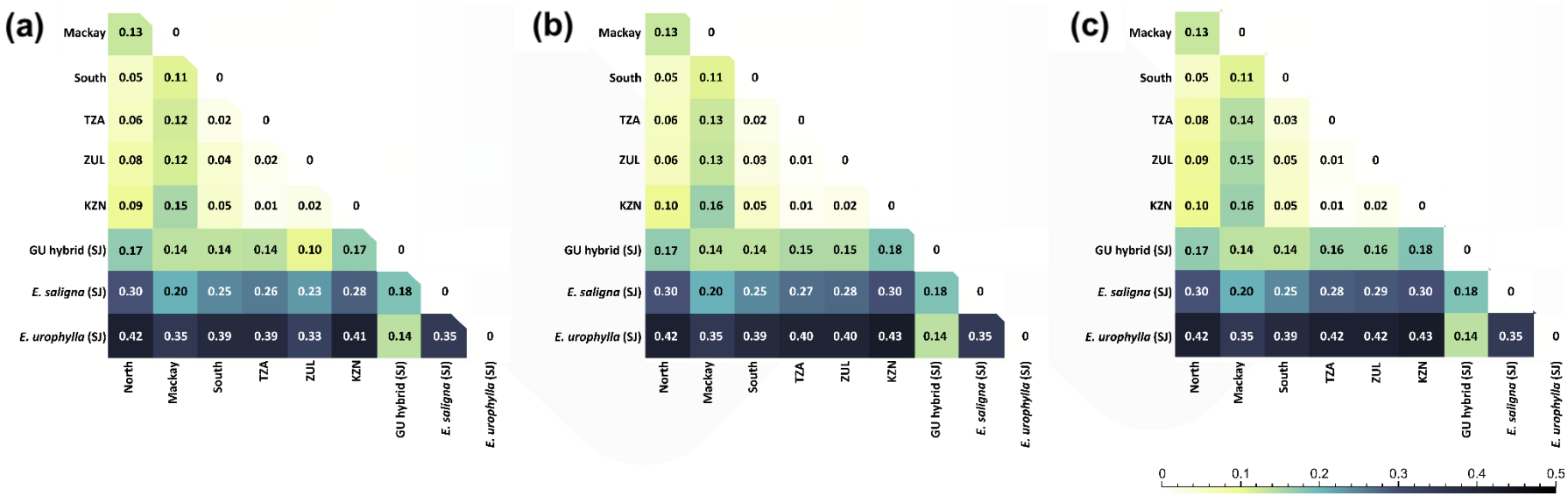

Fig. S3 Population differentiation F_{ST} estimates among breeding *E. grandis*, wild *E. grandis*, and other species in section *Latoangulatae*. Population differentiation was determined for three data sets; including all *E. grandis* breeding material (**a**), excluding introgressed breeding individuals (**b**) and excluding introgressed and infused individuals (**c**). In each analysis, markers were filtered for MAF > 0.05 and call rate > 0.9. *E. urophylla*, *E. saligna* and *E. grandis* x *E. urophylla* (GU) hybrids were obtained from published data (SJ, Silva-Junior *et al.* 2015). Wild *E. grandis* genotypes were described previously in Mostert-O'Neill *et al.* (2021).

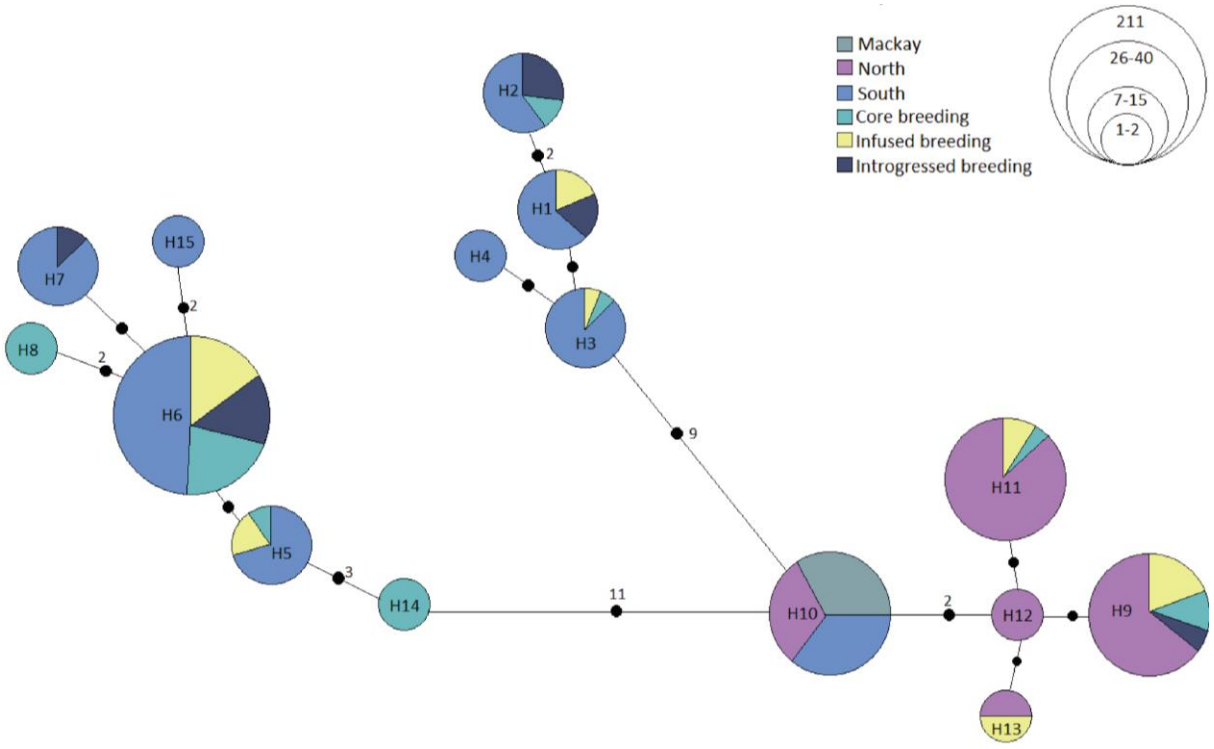

Fig. S4 Chloroplast (cp) haplotype network based on 24 cp SNPs. Each circle represents a unique haplotype with subdivisions (see colour key) based on subpopulation assignments and circle circumference indicative of the number of individuals with that particular haplotype. Mutational changes between haplotypes are indicated in black circles on connecting lines. In total, 17 Mackay, 54 Northern, 104 Southern, 76 core, 56 recently infused, and 54 introgressed breeding individuals, each representing a unique family, were analysed.

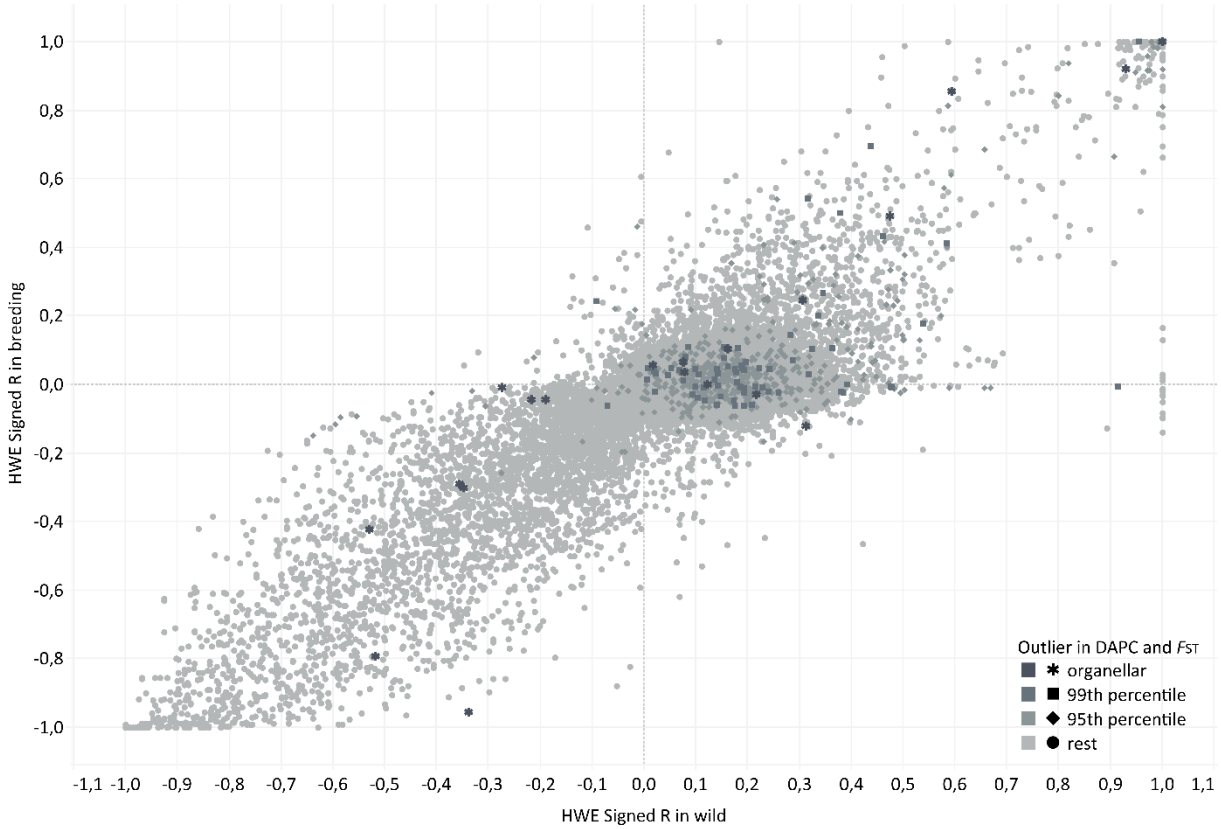

Fig. S5 Marker-specific Hardy-Weinberg Equilibrium (HWE) signed R values of wild vs breeding populations. HWE signed R values were calculated to determine if SNPs were more homozygous (values < 0) or heterozygous (values > 0) in the wild (x-axis, excludes Mackay subpopulation) and breeding populations (y-axis, excludes introgressed and infused individuals). SNPs that had target sequences in nuclear and organellar genomes are indicated as asterisks (these are included for illustration purposes only and were not considered for functional enrichment analysis). SNPs identified as outliers in the 99th and 95th percentiles of F_{ST} and DAPC are indicated as squares and diamonds, respectively.

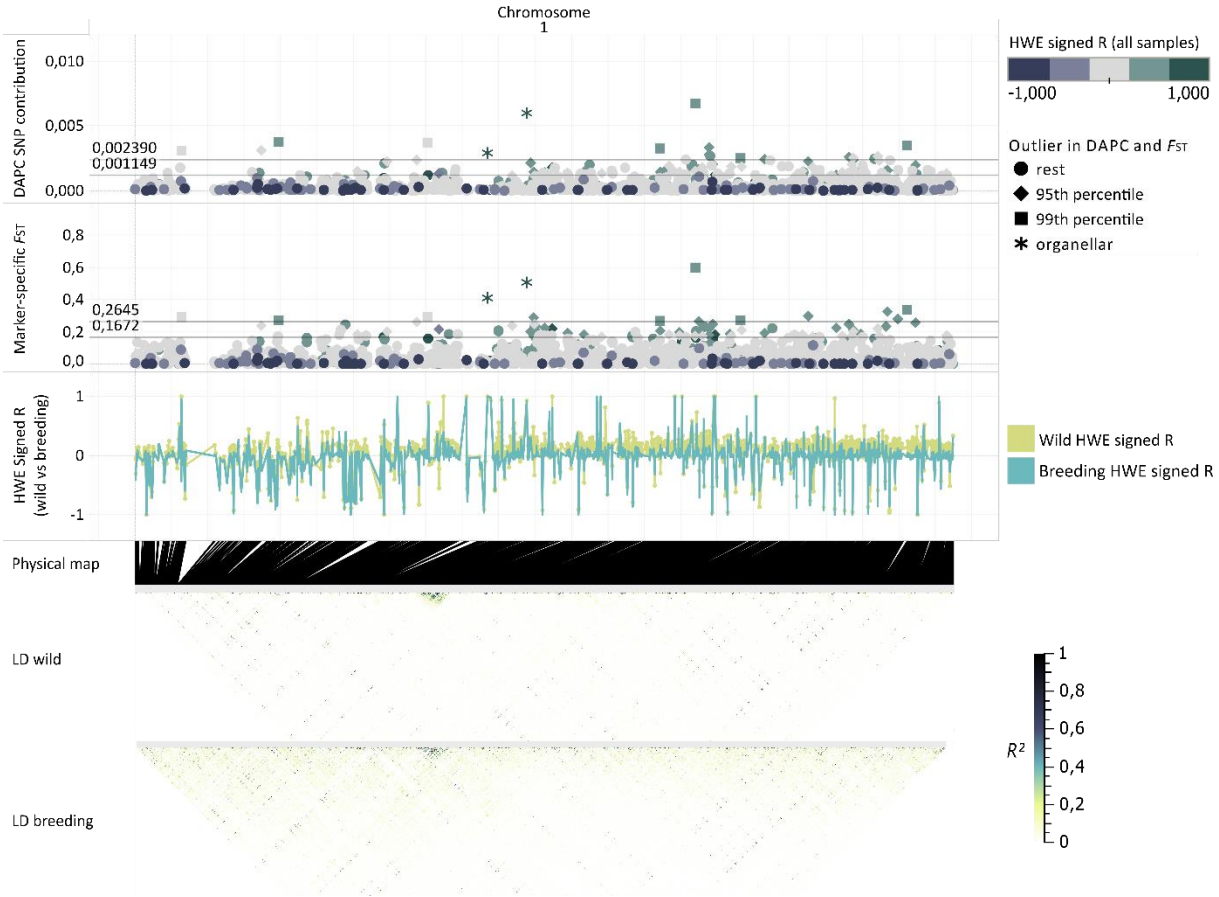

Fig. S6 Genomic outliers and LD plots per chromosome. Discriminant Analysis of Principal Components (DAPC) SNP contribution, indicative of a marker's informativeness in separating breeding and wild samples into $K = 2$ clusters (Jombart *et al.*, 2010), and marker specific F_{ST} values as calculated for breeding (excluding introgressed and infused individuals) vs wild progenitors (Northern and Southern subpopulations) are given for each chromosome with genomic positions given on the x-axis. In each panel, the 95th and 99th percentile values for each of the outlier detection methods are indicated as horizontal lines. Markers identified as differentiated in the 95th and 99th percentile in both analyses are indicated as squares and diamonds, respectively, and markers that target the organellar genomes in addition to the nuclear genome are indicated as asterisks (these are included for illustration purposes only and were not considered for functional enrichment analysis). The colour scale is based on the Hardy-Weinberg Equilibrium (HWE) signed R values of each SNP, indicative of whether a marker is more homozygous (green) or heterozygous (blue) across the breeding and wild populations. The third panel shows HWE signed R values for each marker as calculated in the wild (yellow) and breeding (turquoise) populations to illustrate changes in heterozygosity. Beneath this plot is a physical map of all SNPs and linkage disequilibrium (LD) calculated as the squared correlation (R^2) between alleles at two loci in the wild progenitors (top) and core breeding population (bottom).

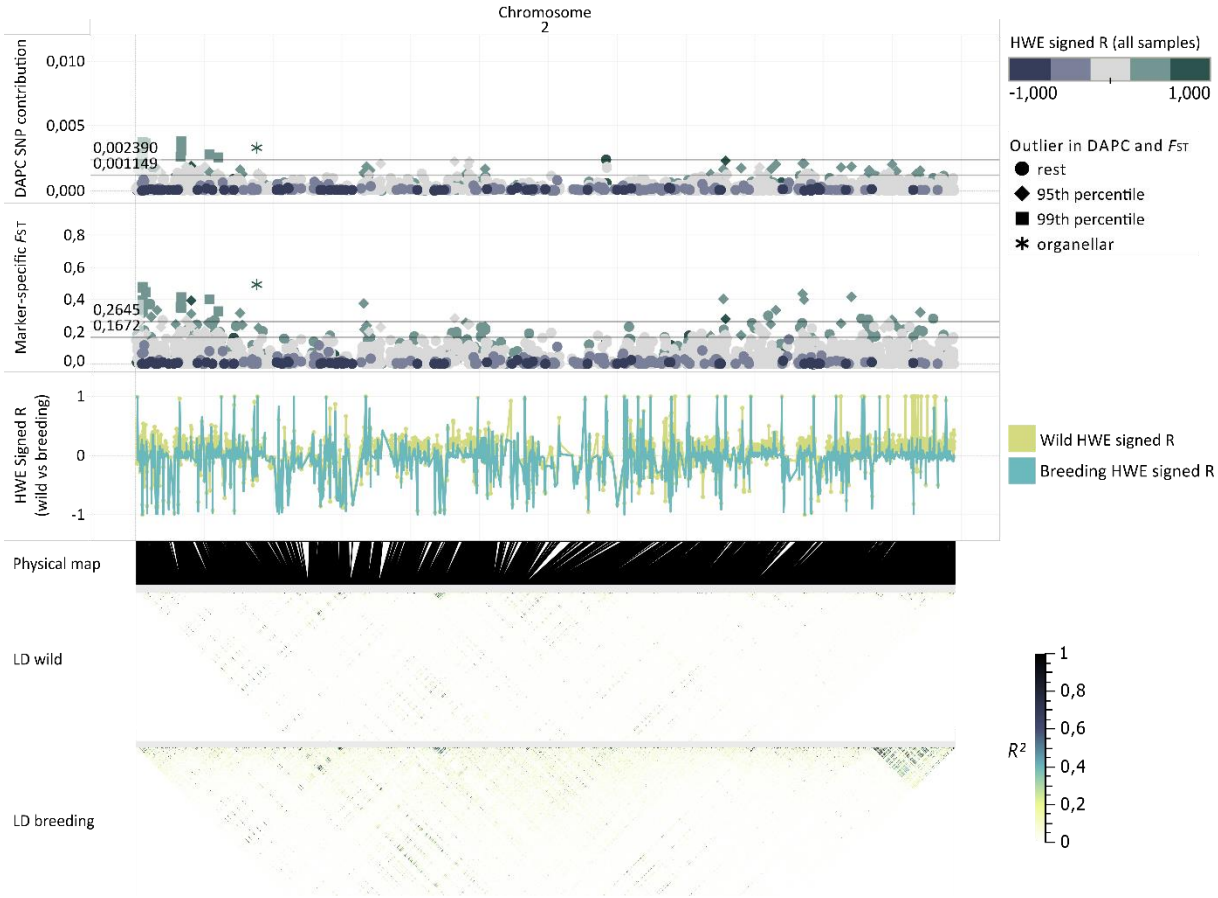

Fig. S6 (cont.) Genomic outliers and LD plots per chromosome. Discriminant Analysis of Principal Components (DAPC) SNP contribution, indicative of a marker's informativeness in separating breeding and wild samples into $K = 2$ clusters (Jombart *et al.*, 2010), and marker specific F_{ST} values as calculated for breeding (excluding introgressed and infused individuals) vs wild progenitors (Northern and Southern subpopulations) are given for each chromosome with genomic positions given on the x-axis. In each panel, the 95th and 99th percentile values for each of the outlier detection methods are indicated as horizontal lines. Markers identified as differentiated in the 95th and 99th percentile in both analyses are indicated as squares and diamonds, respectively, and markers that target the organellar genomes in addition to the nuclear genome are indicated as asterisks (these are included for illustration purposes only and were not considered for functional enrichment analysis). The colour scale is based on the Hardy-Weinberg Equilibrium (HWE) signed R values of each SNP, indicative of whether a marker is more homozygous (green) or heterozygous (blue) across the breeding and wild populations. The third panel shows HWE signed R values for each marker as calculated in the wild (yellow) and breeding (turquoise) populations to illustrate changes in heterozygosity. Beneath this plot is a physical map of all SNPs and linkage disequilibrium (LD) calculated as the squared correlation (R^2) between alleles at two loci in the wild progenitors (top) and core breeding population (bottom).

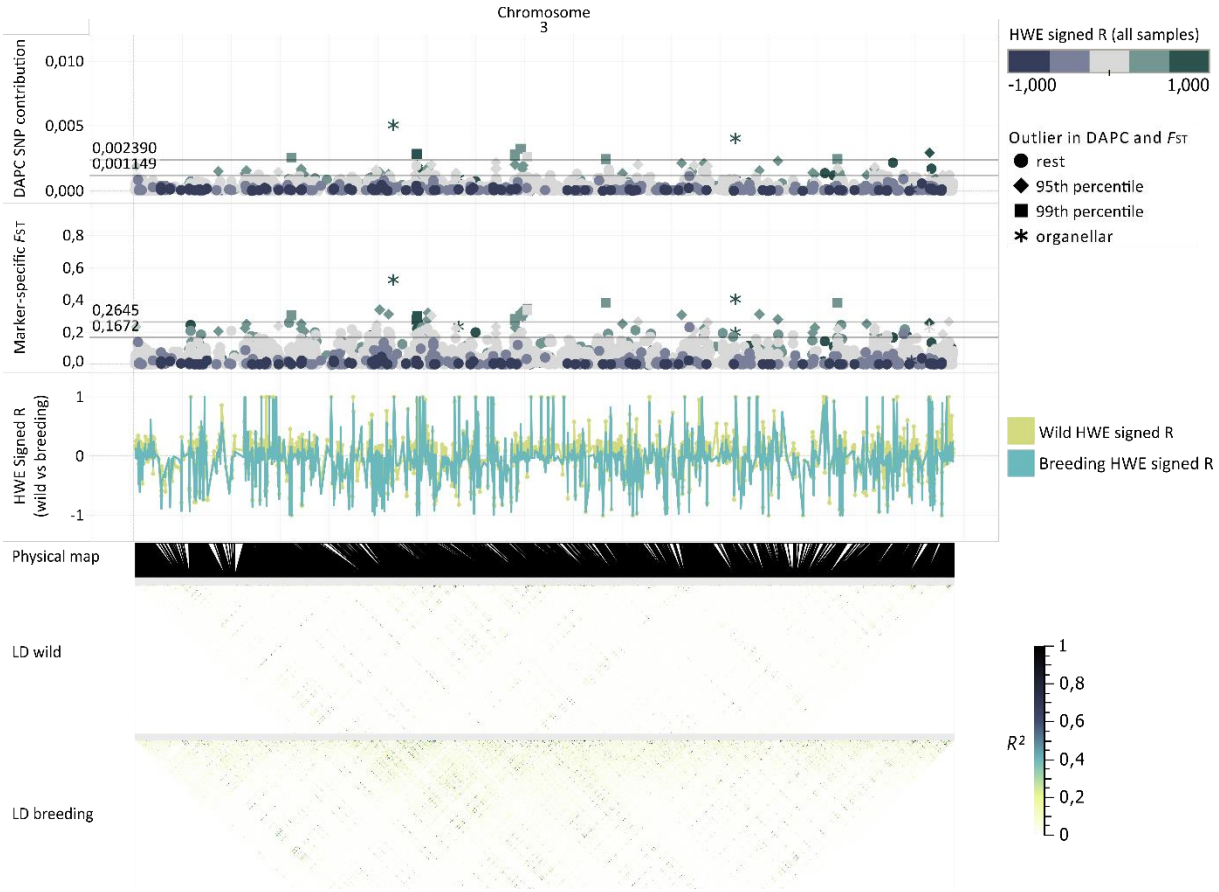

Fig. S6 (cont.) Genomic outliers and LD plots per chromosome. Discriminant Analysis of Principal Components (DAPC) SNP contribution, indicative of a marker's informativeness in separating breeding and wild samples into $K = 2$ clusters (Jombart *et al.*, 2010), and marker specific F_{ST} values as calculated for breeding (excluding introgressed and infused individuals) vs wild progenitors (Northern and Southern subpopulations) are given for each chromosome with genomic positions given on the x-axis. In each panel, the 95th and 99th percentile values for each of the outlier detection methods are indicated as horizontal lines. Markers identified as differentiated in the 95th and 99th percentile in both analyses are indicated as squares and diamonds, respectively, and markers that target the organellar genomes in addition to the nuclear genome are indicated as asterisks (these are included for illustration purposes only and were not considered for functional enrichment analysis). The colour scale is based on the Hardy-Weinberg Equilibrium (HWE) signed R values of each SNP, indicative of whether a marker is more homozygous (green) or heterozygous (blue) across the breeding and wild populations. The third panel shows HWE signed R values for each marker as calculated in the wild (yellow) and breeding (turquoise) populations to illustrate changes in heterozygosity. Beneath this plot is a physical map of all SNPs and linkage disequilibrium (LD) calculated as the squared correlation (R^2) between alleles at two loci in the wild progenitors (top) and core breeding population (bottom).

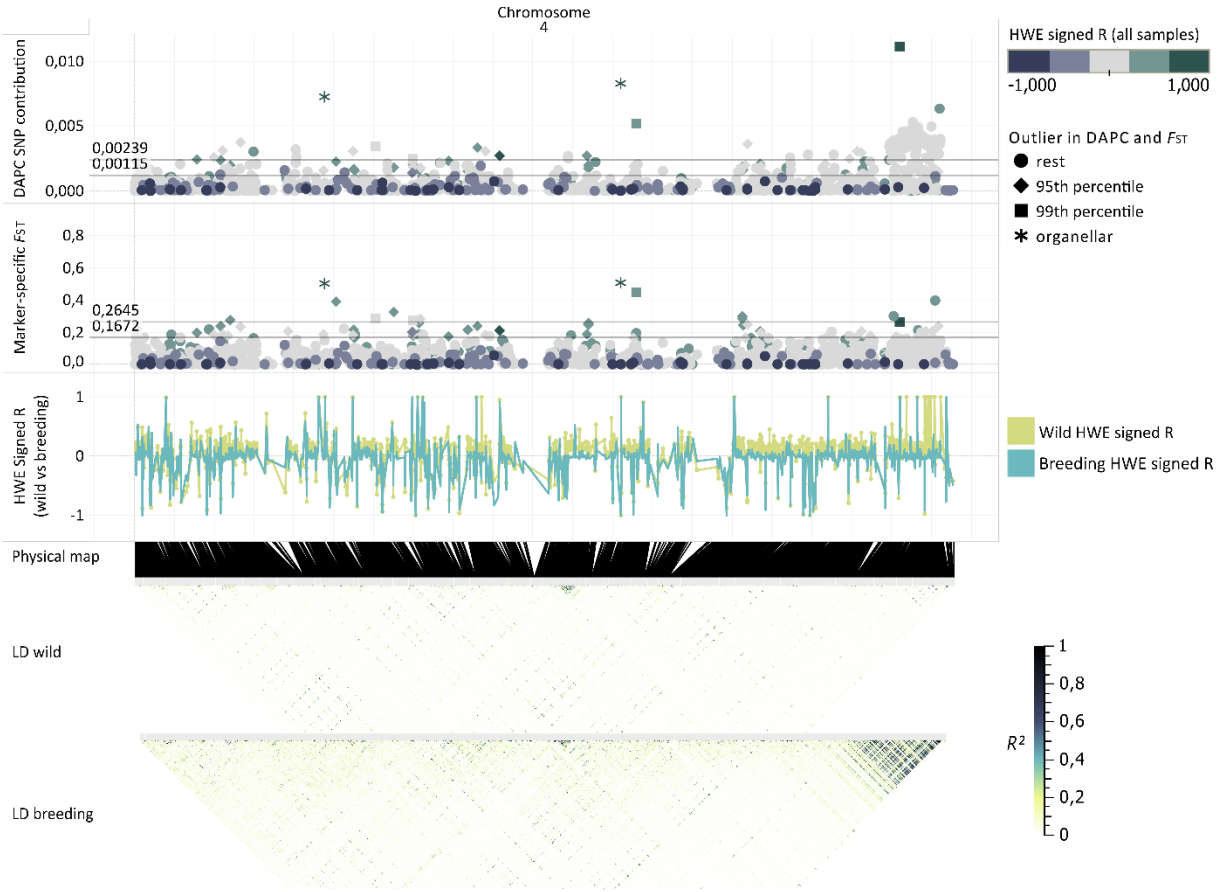

Fig. S6 (cont.) Genomic outliers and LD plots per chromosome. Discriminant Analysis of Principal Components (DAPC) SNP contribution, indicative of a marker's informativeness in separating breeding and wild samples into $K = 2$ clusters (Jombart *et al.*, 2010), and marker specific F_{ST} values as calculated for breeding (excluding introgressed and infused individuals) vs wild progenitors (Northern and Southern subpopulations) are given for each chromosome with genomic positions given on the x-axis. In each panel, the 95th and 99th percentile values for each of the outlier detection methods are indicated as horizontal lines. Markers identified as differentiated in the 95th and 99th percentile in both analyses are indicated as squares and diamonds, respectively, and markers that target the organellar genomes in addition to the nuclear genome are indicated as asterisks (these are included for illustration purposes only and were not considered for functional enrichment analysis). The colour scale is based on the Hardy-Weinberg Equilibrium (HWE) signed R values of each SNP, indicative of whether a marker is more homozygous (green) or heterozygous (blue) across the breeding and wild populations. The third panel shows HWE signed R values for each marker as calculated in the wild (yellow) and breeding (turquoise) populations to illustrate changes in heterozygosity. Beneath this plot is a physical map of all SNPs and linkage disequilibrium (LD) calculated as the squared correlation (R^2) between alleles at two loci in the wild progenitors (top) and core breeding population (bottom).

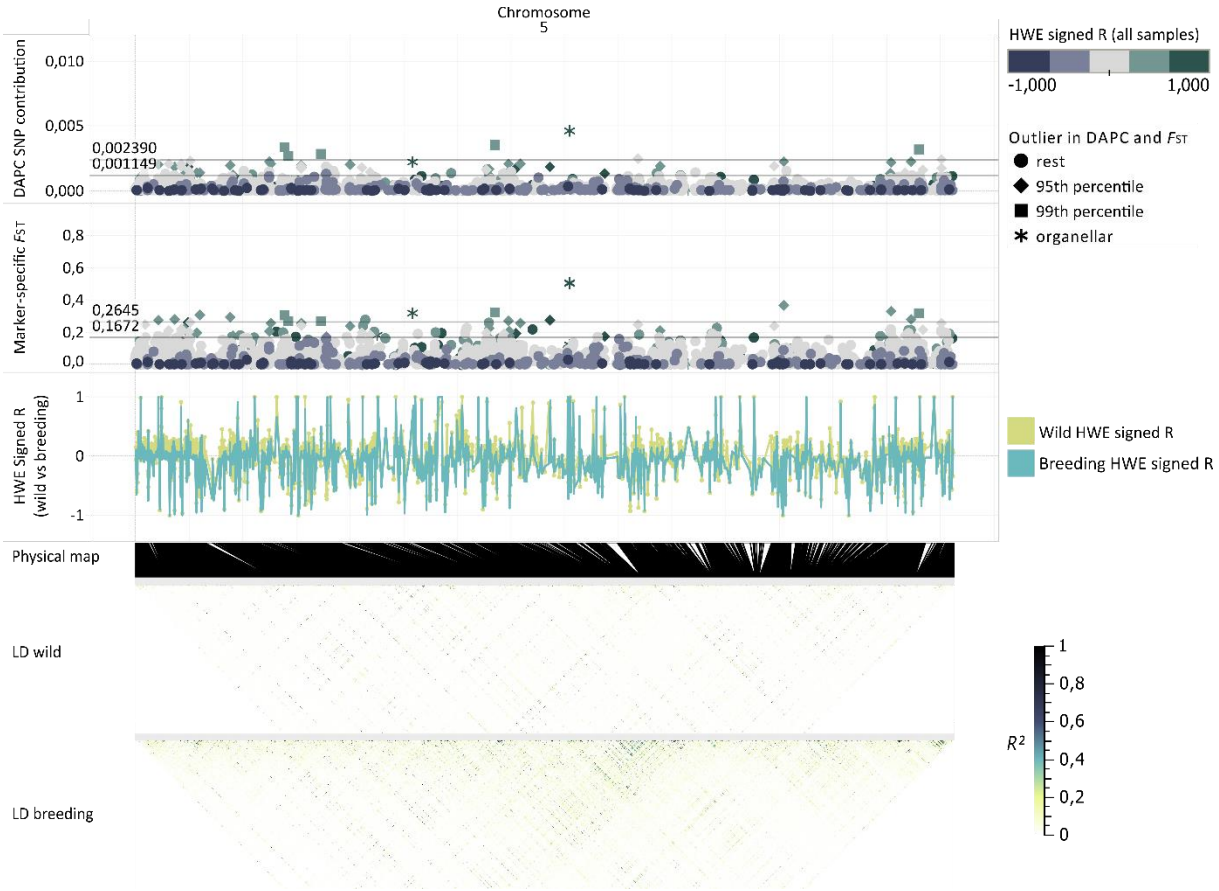

Fig. S6 (cont.) Genomic outliers and LD plots per chromosome. Discriminant Analysis of Principal Components (DAPC) SNP contribution, indicative of a marker's informativeness in separating breeding and wild samples into $K = 2$ clusters (Jombart *et al.*, 2010), and marker specific F_{ST} values as calculated for breeding (excluding introgressed and infused individuals) vs wild progenitors (Northern and Southern subpopulations) are given for each chromosome with genomic positions given on the x-axis. In each panel, the 95th and 99th percentile values for each of the outlier detection methods are indicated as horizontal lines. Markers identified as differentiated in the 95th and 99th percentile in both analyses are indicated as squares and diamonds, respectively, and markers that target the organellar genomes in addition to the nuclear genome are indicated as asterisks (these are included for illustration purposes only and were not considered for functional enrichment analysis). The colour scale is based on the Hardy-Weinberg Equilibrium (HWE) signed R values of each SNP, indicative of whether a marker is more homozygous (green) or heterozygous (blue) across the breeding and wild populations. The third panel shows HWE signed R values for each marker as calculated in the wild (yellow) and breeding (turquoise) populations to illustrate changes in heterozygosity. Beneath this plot is a physical map of all SNPs and linkage disequilibrium (LD) calculated as the squared correlation (R^2) between alleles at two loci in the wild progenitors (top) and core breeding population (bottom).

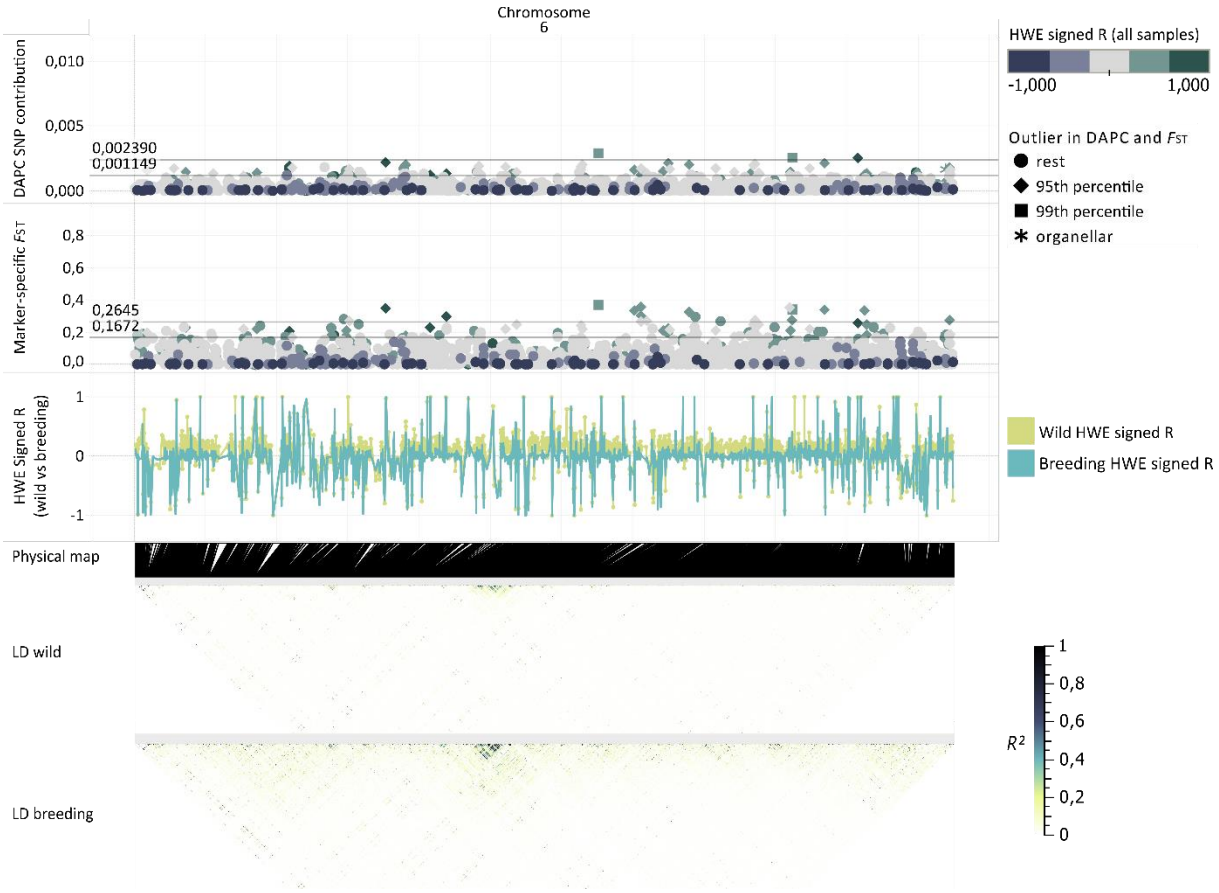

Fig. S6 (cont.) Genomic outliers and LD plots per chromosome. Discriminant Analysis of Principal Components (DAPC) SNP contribution, indicative of a marker's informativeness in separating breeding and wild samples into $K = 2$ clusters (Jombart *et al.*, 2010), and marker specific F_{ST} values as calculated for breeding (excluding introgressed and infused individuals) vs wild progenitors (Northern and Southern subpopulations) are given for each chromosome with genomic positions given on the x-axis. In each panel, the 95th and 99th percentile values for each of the outlier detection methods are indicated as horizontal lines. Markers identified as differentiated in the 95th and 99th percentile in both analyses are indicated as squares and diamonds, respectively, and markers that target the organellar genomes in addition to the nuclear genome are indicated as asterisks (these are included for illustration purposes only and were not considered for functional enrichment analysis). The colour scale is based on the Hardy-Weinberg Equilibrium (HWE) signed R values of each SNP, indicative of whether a marker is more homozygous (green) or heterozygous (blue) across the breeding and wild populations. The third panel shows HWE signed R values for each marker as calculated in the wild (yellow) and breeding (turquoise) populations to illustrate changes in heterozygosity. Beneath this plot is a physical map of all SNPs and linkage disequilibrium (LD) calculated as the squared correlation (R^2) between alleles at two loci in the wild progenitors (top) and core breeding population (bottom).

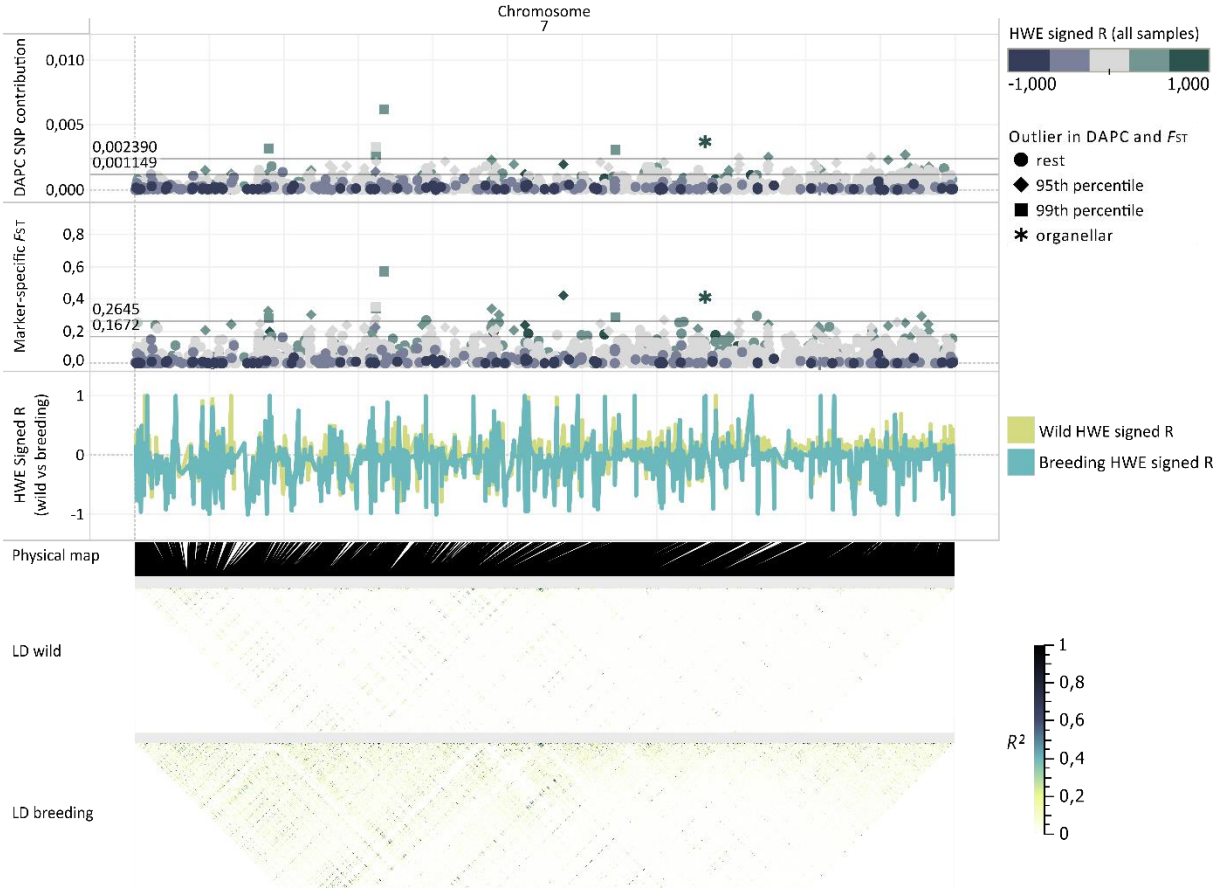

Fig. S6 (cont.) Genomic outliers and LD plots per chromosome. Discriminant Analysis of Principal Components (DAPC) SNP contribution, indicative of a marker's informativeness in separating breeding and wild samples into $K = 2$ clusters (Jombart *et al.*, 2010), and marker specific F_{ST} values as calculated for breeding (excluding introgressed and infused individuals) vs wild progenitors (Northern and Southern subpopulations) are given for each chromosome with genomic positions given on the x-axis. In each panel, the 95th and 99th percentile values for each of the outlier detection methods are indicated as horizontal lines. Markers identified as differentiated in the 95th and 99th percentile in both analyses are indicated as squares and diamonds, respectively, and markers that target the organellar genomes in addition to the nuclear genome are indicated as asterisks (these are included for illustration purposes only and were not considered for functional enrichment analysis). The colour scale is based on the Hardy-Weinberg Equilibrium (HWE) signed R values of each SNP, indicative of whether a marker is more homozygous (green) or heterozygous (blue) across the breeding and wild populations. The third panel shows HWE signed R values for each marker as calculated in the wild (yellow) and breeding (turquoise) populations to illustrate changes in heterozygosity. Beneath this plot is a physical map of all SNPs and linkage disequilibrium (LD) calculated as the squared correlation (R^2) between alleles at two loci in the wild progenitors (top) and core breeding population (bottom).

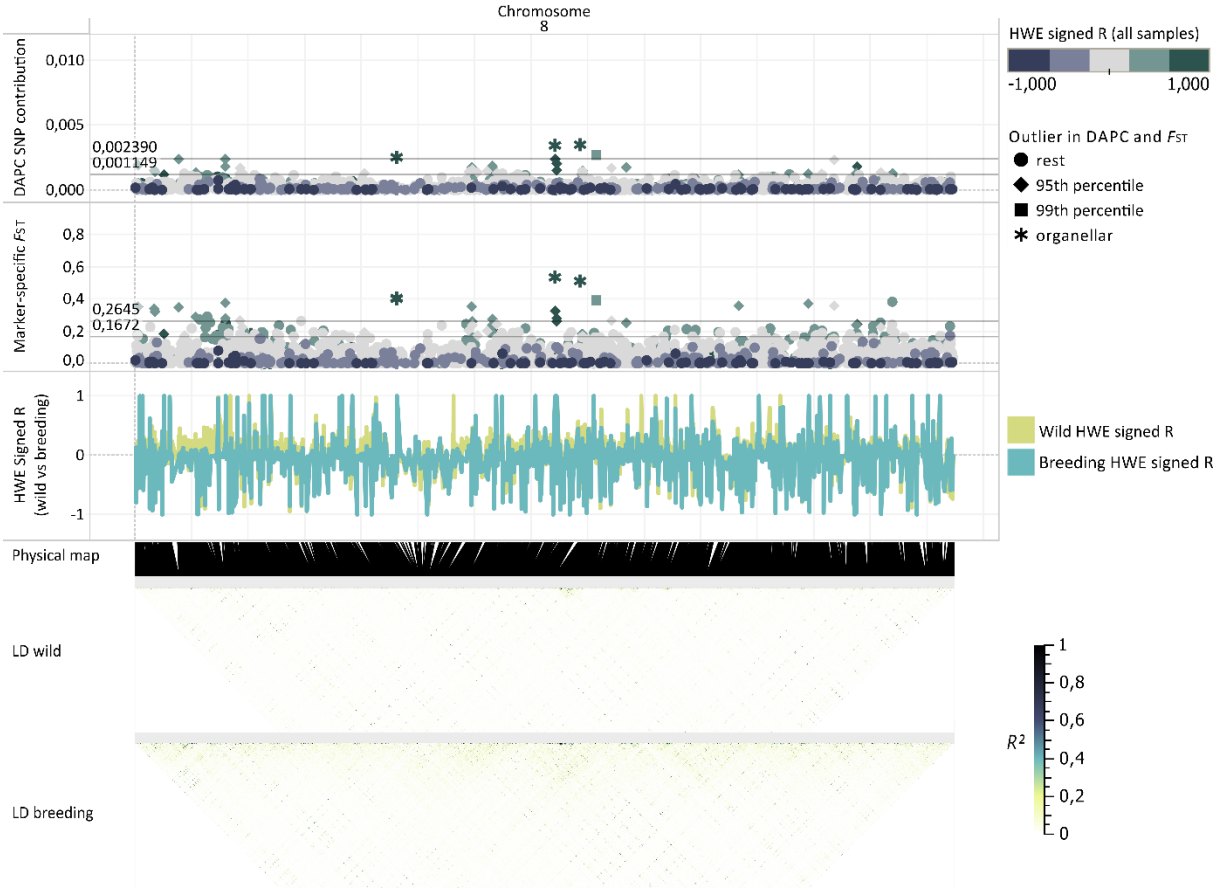

Fig. S6 (cont.) Genomic outliers and LD plots per chromosome. Discriminant Analysis of Principal Components (DAPC) SNP contribution, indicative of a marker's informativeness in separating breeding and wild samples into $K = 2$ clusters (Jombart *et al.*, 2010), and marker specific F_{ST} values as calculated for breeding (excluding introgressed and infused individuals) vs wild progenitors (Northern and Southern subpopulations) are given for each chromosome with genomic positions given on the x-axis. In each panel, the 95th and 99th percentile values for each of the outlier detection methods are indicated as horizontal lines. Markers identified as differentiated in the 95th and 99th percentile in both analyses are indicated as squares and diamonds, respectively, and markers that target the organellar genomes in addition to the nuclear genome are indicated as asterisks (these are included for illustration purposes only and were not considered for functional enrichment analysis). The colour scale is based on the Hardy-Weinberg Equilibrium (HWE) signed R values of each SNP, indicative of whether a marker is more homozygous (green) or heterozygous (blue) across the breeding and wild populations. The third panel shows HWE signed R values for each marker as calculated in the wild (yellow) and breeding (turquoise) populations to illustrate changes in heterozygosity. Beneath this plot is a physical map of all SNPs and linkage disequilibrium (LD) calculated as the squared correlation (R^2) between alleles at two loci in the wild progenitors (top) and core breeding population (bottom).

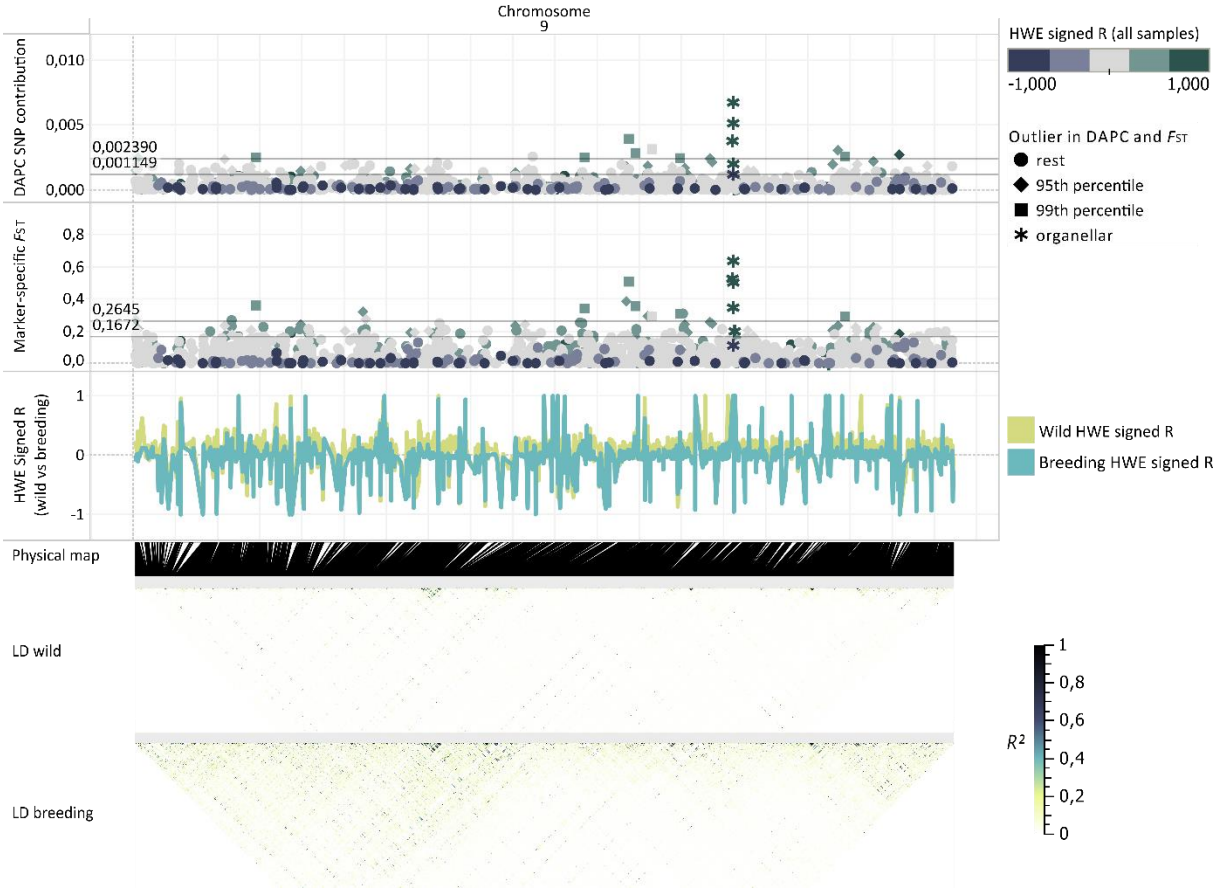

Fig. S6 (cont.) Genomic outliers and LD plots per chromosome. Discriminant Analysis of Principal Components (DAPC) SNP contribution, indicative of a marker's informativeness in separating breeding and wild samples into $K = 2$ clusters (Jombart *et al.*, 2010), and marker specific F_{ST} values as calculated for breeding (excluding introgressed and infused individuals) vs wild progenitors (Northern and Southern subpopulations) are given for each chromosome with genomic positions given on the x-axis. In each panel, the 95th and 99th percentile values for each of the outlier detection methods are indicated as horizontal lines. Markers identified as differentiated in the 95th and 99th percentile in both analyses are indicated as squares and diamonds, respectively, and markers that target the organellar genomes in addition to the nuclear genome are indicated as asterisks (these are included for illustration purposes only and were not considered for functional enrichment analysis). The colour scale is based on the Hardy-Weinberg Equilibrium (HWE) signed R values of each SNP, indicative of whether a marker is more homozygous (green) or heterozygous (blue) across the breeding and wild populations. The third panel shows HWE signed R values for each marker as calculated in the wild (yellow) and breeding (turquoise) populations to illustrate changes in heterozygosity. Beneath this plot is a physical map of all SNPs and linkage disequilibrium (LD) calculated as the squared correlation (R^2) between alleles at two loci in the wild progenitors (top) and core breeding population (bottom).

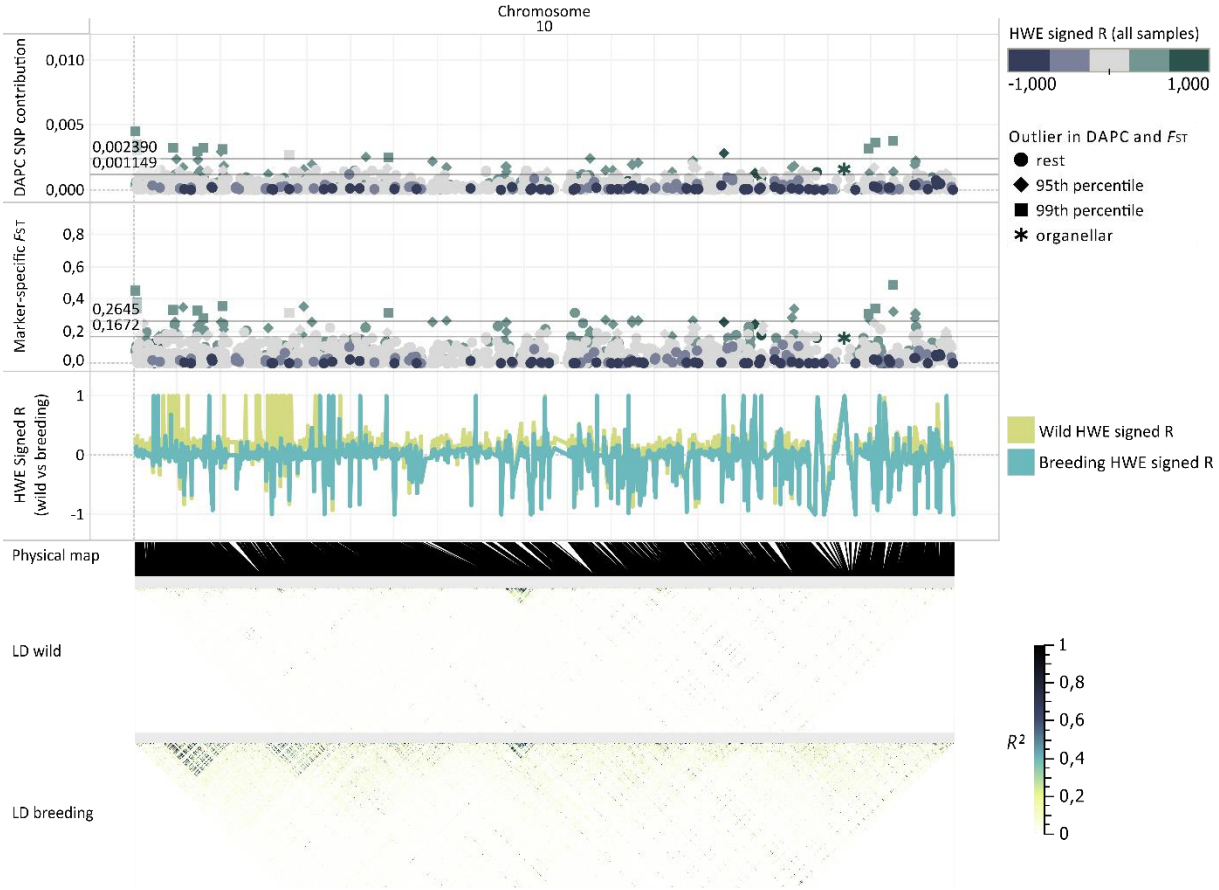

Fig. S6 (cont.) Genomic outliers and LD plots per chromosome. Discriminant Analysis of Principal Components (DAPC) SNP contribution, indicative of a marker's informativeness in separating breeding and wild samples into $K = 2$ clusters (Jombart *et al.*, 2010), and marker specific F_{ST} values as calculated for breeding (excluding introgressed and infused individuals) vs wild progenitors (Northern and Southern subpopulations) are given for each chromosome with genomic positions given on the x-axis. In each panel, the 95th and 99th percentile values for each of the outlier detection methods are indicated as horizontal lines. Markers identified as differentiated in the 95th and 99th percentile in both analyses are indicated as squares and diamonds, respectively, and markers that target the organellar genomes in addition to the nuclear genome are indicated as asterisks (these are included for illustration purposes only and were not considered for functional enrichment analysis). The colour scale is based on the Hardy-Weinberg Equilibrium (HWE) signed R values of each SNP, indicative of whether a marker is more homozygous (green) or heterozygous (blue) across the breeding and wild populations. The third panel shows HWE signed R values for each marker as calculated in the wild (yellow) and breeding (turquoise) populations to illustrate changes in heterozygosity. Beneath this plot is a physical map of all SNPs and linkage disequilibrium (LD) calculated as the squared correlation (R^2) between alleles at two loci in the wild progenitors (top) and core breeding population (bottom).

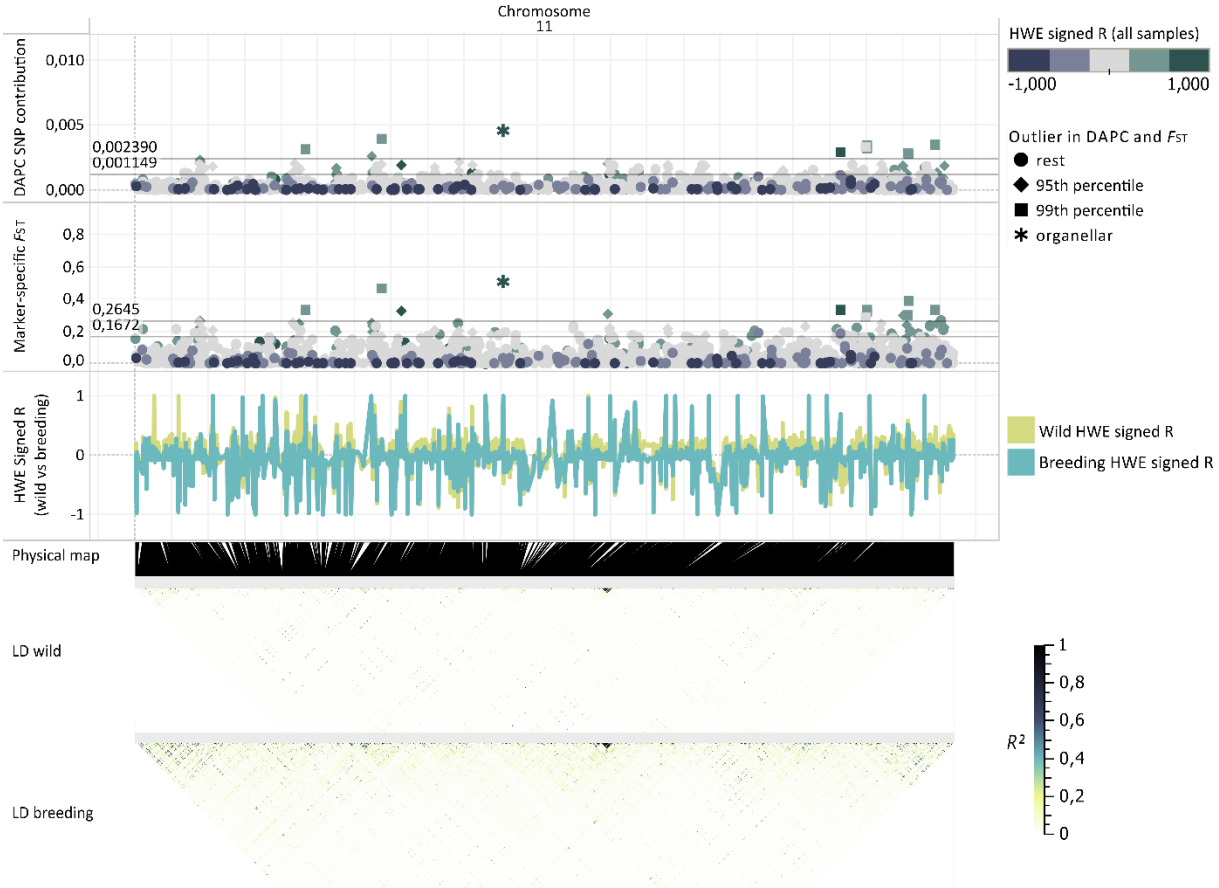

Fig. S6 (cont.) Genomic outliers and LD plots per chromosome. Discriminant Analysis of Principal Components (DAPC) SNP contribution, indicative of a marker's informativeness in separating breeding and wild samples into $K = 2$ clusters (Jombart *et al.*, 2010), and marker specific F_{ST} values as calculated for breeding (excluding introgressed and infused individuals) vs wild progenitors (Northern and Southern subpopulations) are given for each chromosome with genomic positions given on the x-axis. In each panel, the 95th and 99th percentile values for each of the outlier detection methods are indicated as horizontal lines. Markers identified as differentiated in the 95th and 99th percentile in both analyses are indicated as squares and diamonds, respectively, and markers that target the organellar genomes in addition to the nuclear genome are indicated as asterisks (these are included for illustration purposes only and were not considered for functional enrichment analysis). The colour scale is based on the Hardy-Weinberg Equilibrium (HWE) signed R values of each SNP, indicative of whether a marker is more homozygous (green) or heterozygous (blue) across the breeding and wild populations. The third panel shows HWE signed R values for each marker as calculated in the wild (yellow) and breeding (turquoise) populations to illustrate changes in heterozygosity. Beneath this plot is a physical map of all SNPs and linkage disequilibrium (LD) calculated as the squared correlation (R^2) between alleles at two loci in the wild progenitors (top) and core breeding population (bottom).

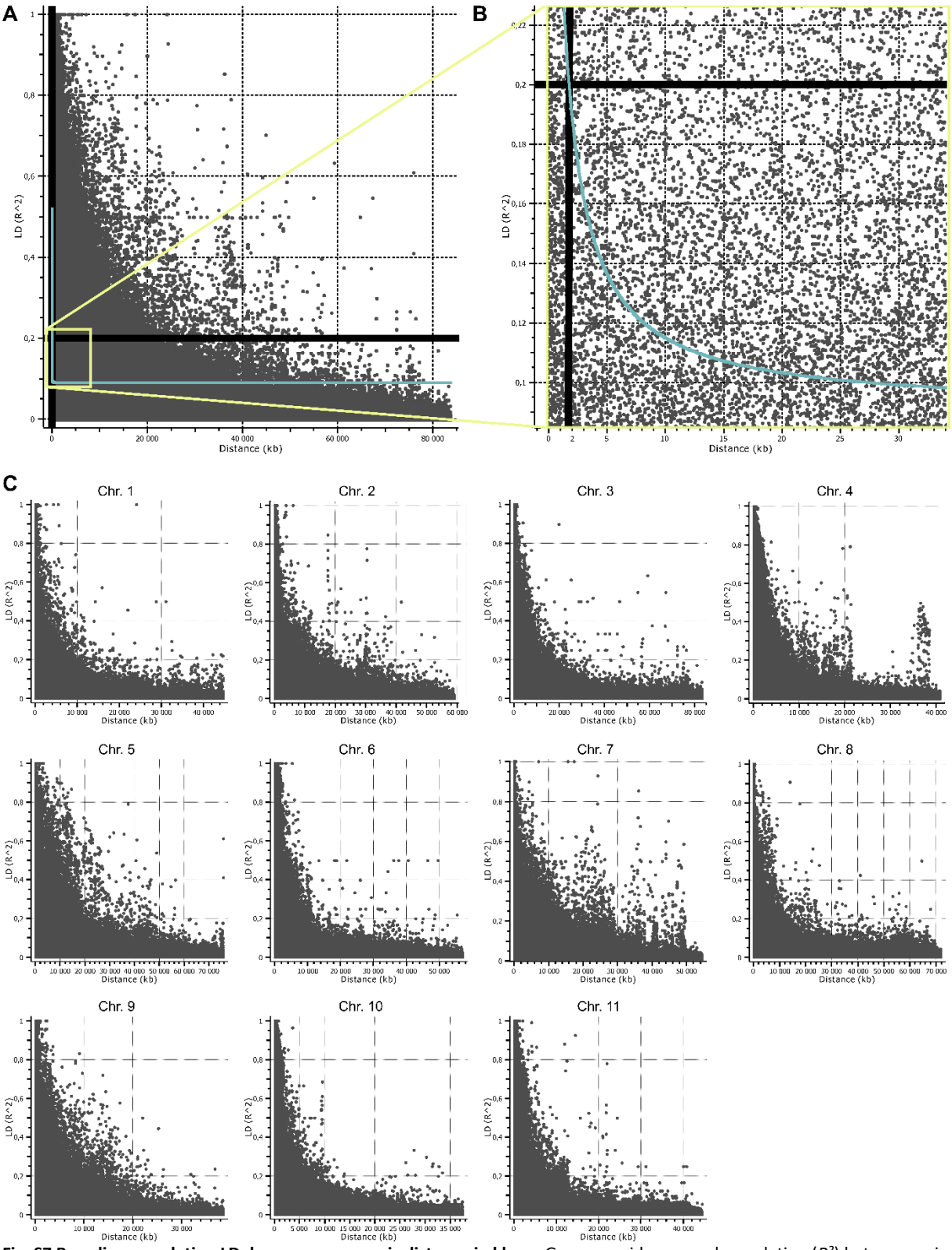

Fig. S7 Breeding population LD decay over genomic distance in kb. a - Genome-wide squared correlation (R^2) between pairs of loci (21 991 SNPs), and average genome-wide LD given by non-linear regression curve (turquoise curve) decay ($R^2 < 0.2$) at 1.75 kb (**b**). **c** - Per chromosome squared correlation (R^2) between pairs of loci. LD was calculated as squared correlation using the composite haplotype method (CHM) in SNP & Variation Suite™ v8.x (SVS8; Golden Helix, Inc., Bozeman, MT, www.goldenhelix.com).

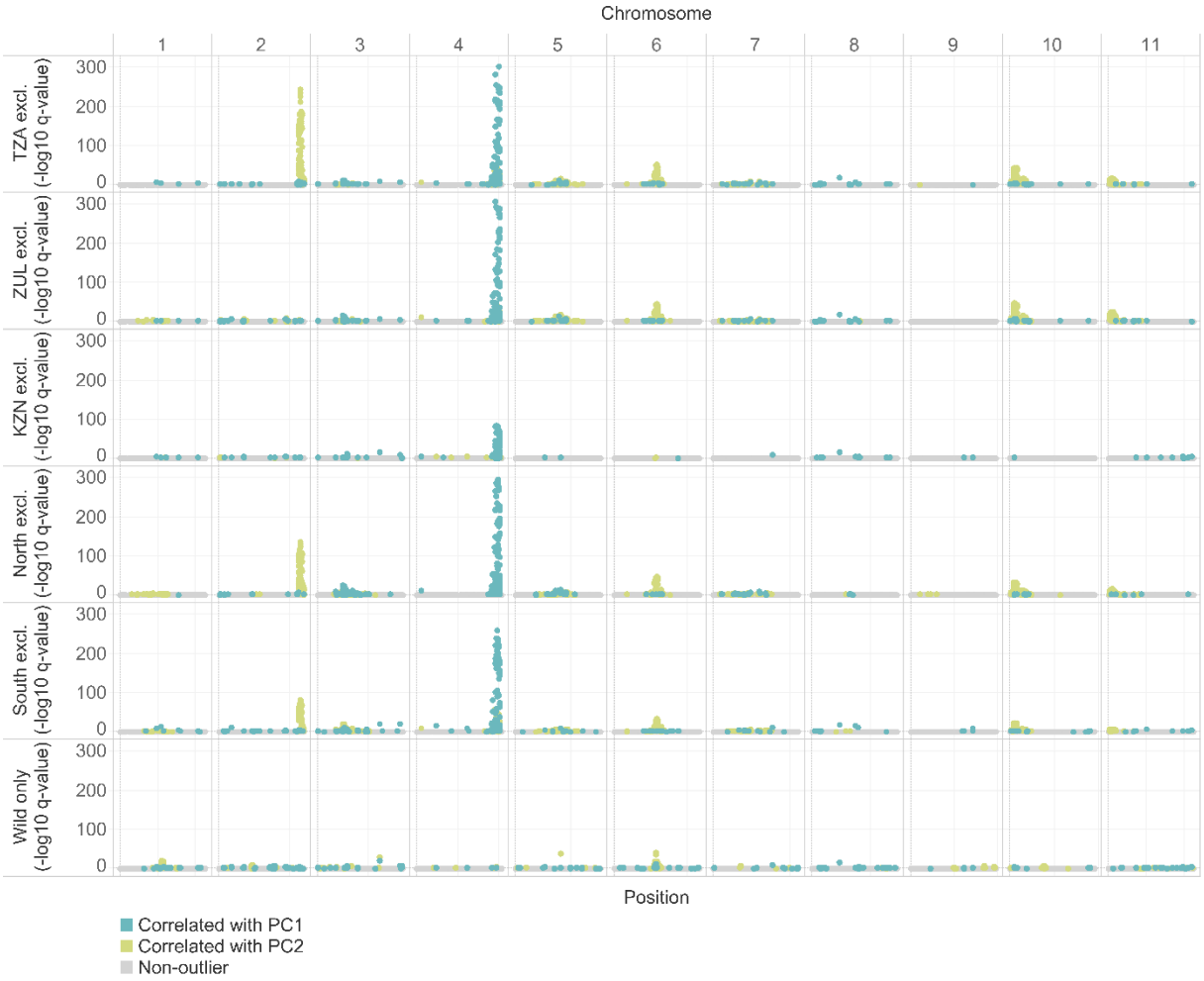

Fig. S8 Outlier detection by *pcadapt* scan. The -log10 q-value (*y*-axis) is given for SNPs (*x*-axis) and the principal component (PC) with which the outliers were most correlated is given in the colour legend. In each panel, a different subpopulation was excluded starting with the three breeding populations (TZA, ZUL and KZN), and followed by the two wild subpopulations retained for outlier detection (North and South). The last panel shows -log10 q-values for outlier detection done on the wild subpopulations only.

Table S3 Summary statistics of genetic diversity using *hierfstat* v. 0.04-22 (Goudet 2005)

Statistic	All *E. grandis* retained†	Wild *E. grandis*	Breeding *E. grandis*
Average observed heterozygosity ($\widehat{H}_O$)	0.2637	0.2496	0.2731
Average gene diversity (expected heterozygosity) within subpopulations ($\widehat{H}_S$)	0.2735	0.2807	0.2686
Average gene diversity (expected heterozygosity) over the total population ($\widehat{H}_T$)	0.2856	0.2888	0.2709
Average inbreeding coefficient within subpopulations ($\widehat{F}_{IS}$)	0.0359	0.1109	-0.0166
Average genetic differentiation among subpopulations ($\widehat{F}_{ST}$)	0.0423	0.028	0.0082

† Includes Northern and Southern wild subpopulations, and three breeding subpopulations (TZA, ZUL and KZN) excluding introgressed and recently infused individuals.

Table S4 Wilcoxon signed rank test p-values supporting the alternative hypothesis, that the mean of the outliers was greater than the mean of the rest of the SNPs

	p-value DAPC SNP contribution		p-value marker-specific F_{ST}		p-value HWE signed R†	
	Incl. oSNPs	Excl. oSNPs	Incl. oSNPs	Excl. oSNPs	Incl. oSNPs	Excl. oSNPs
95th percentile	0	0	0	0	$8.79e^{-243}$	$2.86e^{-219}$
99th percentile	$9.16e^{-57}$	$3.26e^{-50}$	$3.95e^{-57}$	$1.44e^{-50}$	$9.56e^{-47}$	$6.22e^{-39}$

oSNPs – SNPs that had targets in organellar genomes.

† The mean Hardy-Weinberg Equilibrium (HWE) signed R value for outlier loci detected in the 95th and 99th percentile of both outlier detection methods (DAPC SNP contribution and marker-specific F_{ST} estimates) was compared to the mean of the rest of the loci. This was done for the HWE signed R values calculated across all samples (including wild and breeding samples)

Table S5 Gene Ontology (GO) enrichment analysis for genes in LD (within 2 kb) with outlier SNPs against the SNP-captured gene space[†] before excluding organellar-targeting SNPs

Term ID	Description	Genes in category	Query genes in category	Adjusted p-value
GO:0015979	Photosynthesis	134	9	0.000214316
GO:0006354	DNA-dependent transcription, elongation	63	7	0.000196334
GO:0006091	Generation of precursor metabolites and energy	23	6	$7.91e^{-06}$

[†] SNP-captured gene space consisting of 12 071 genes within 2 kb of the 21 991 informative SNPs prior to removal of organellar-targeting SNPs.

SI References

Goudet J. 2005. *hierfstat*, a package for R to compute and test hierarchical *F*-statistics. *Molecular Ecology Notes* **5**(1): 184-186.

Jombart T, Devillard S, Balloux F. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics* **11**(1): 94.

Mostert-O'Neill MM, Reynolds SM, Acosta JJ, Lee DJ, Borevitz JO, Myburg AA. 2021. Genomic evidence of introgression and adaptation in a model subtropical tree species, *Eucalyptus grandis*. *Molecular Ecology* **30**(3): 625-638.

Silva-Junior OB, Faria DA, Grattapaglia D. 2015. A flexible multi-species genome-wide 60K SNP chip developed from pooled resequencing of 240 *Eucalyptus* tree genomes across 12 species. *New Phytologist* **206**(4): 1527-1540.