

Supplementary Information

Whole genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum

Emma S Mace^{1,a}, Shuaishuai Tai^{2,a}, Edward K Gilding³, Yanhong Li², Peter J Prentis⁴, Lianle Bian², Bradley C Campbell³, Wushu Hu², David J Innes⁵, Xuelian Han², Alan Cruickshank¹, Changming Dai², Céline Frère³, Haikuan Zhang², Colleen H Hunt¹, Xianyuan Wang², Tracey Shatte¹, Miao Wang², Zhe Su², Jun Li², Xiaozhen Lin², Ian D Godwin³, David R Jordan⁶, Jun Wang^{2,7,8}

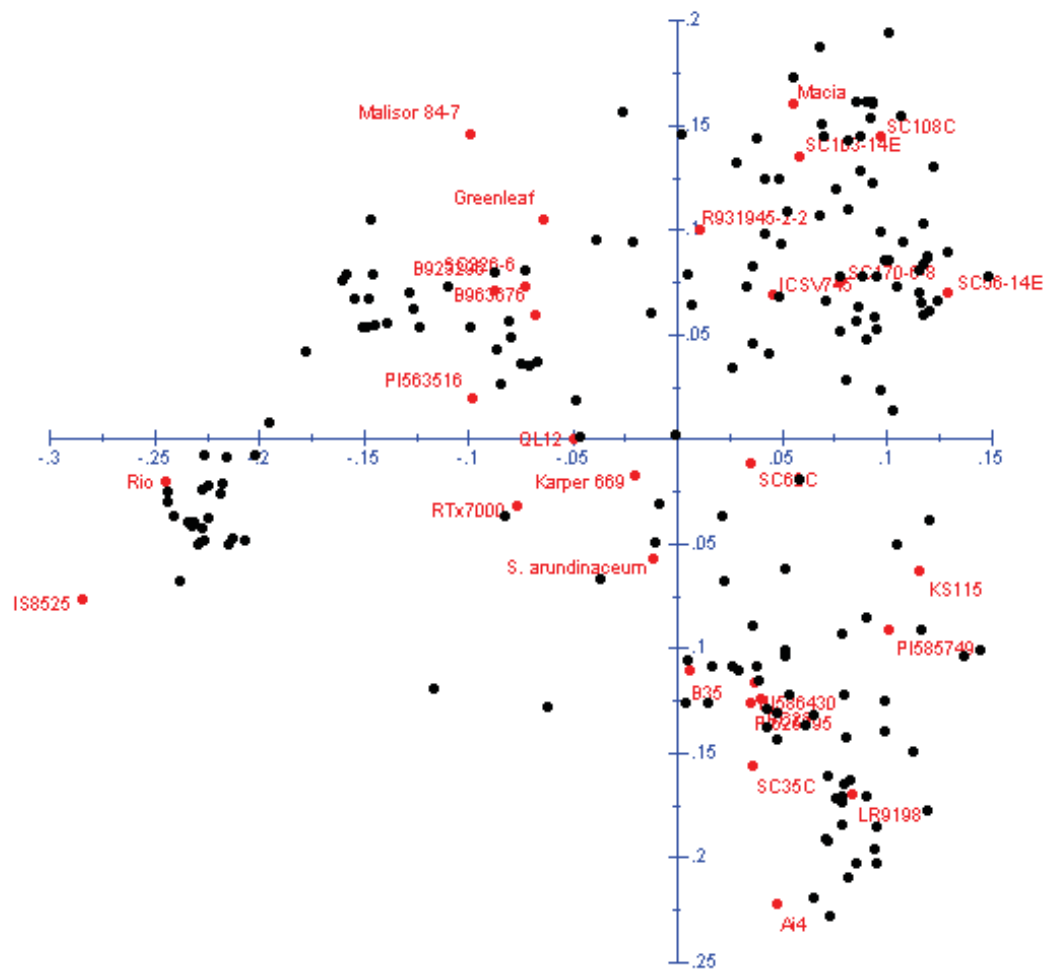
¹Department of Agriculture, Fisheries & Forestry Queensland (DAFFQ), Warwick, QLD, Australia.

²BGI-Shenzhen, Shenzhen, China. ³The University of Queensland, School of Agriculture and Food Sciences, Brisbane, QLD, Australia. ⁴Queensland University of Technology, Brisbane, QLD, Australia. ⁵DAFFQ, Cooper's Plains, Brisbane, QLD, Australia. ⁶Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, Warwick, QLD, Australia. ⁷Dept of Biology, University of Copenhagen, DK-2200 Copenhagen, Denmark. ⁸The Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, DK-2200 Copenhagen, Denmark.

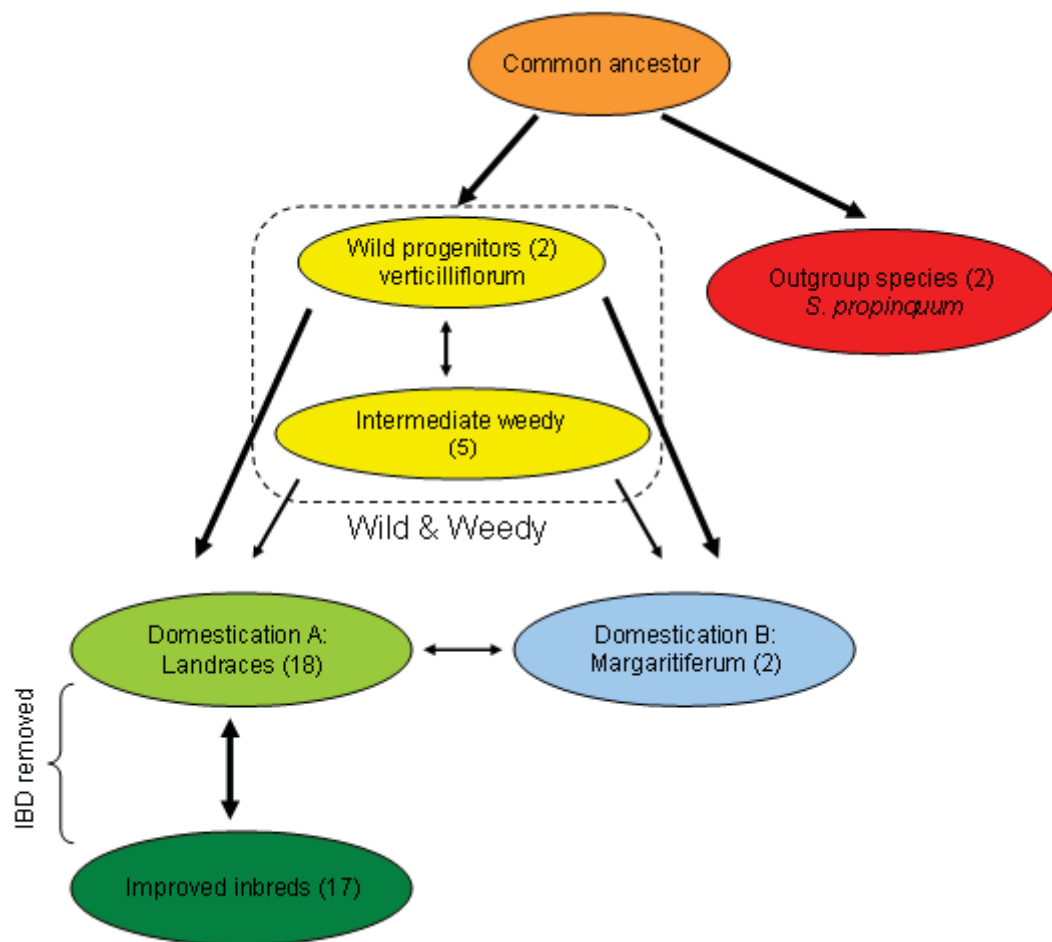
Correspondence should be addressed to D.R.J. (david.jordan@uq.edu.au) or J.W.

(wangj@genomics.org.cn)

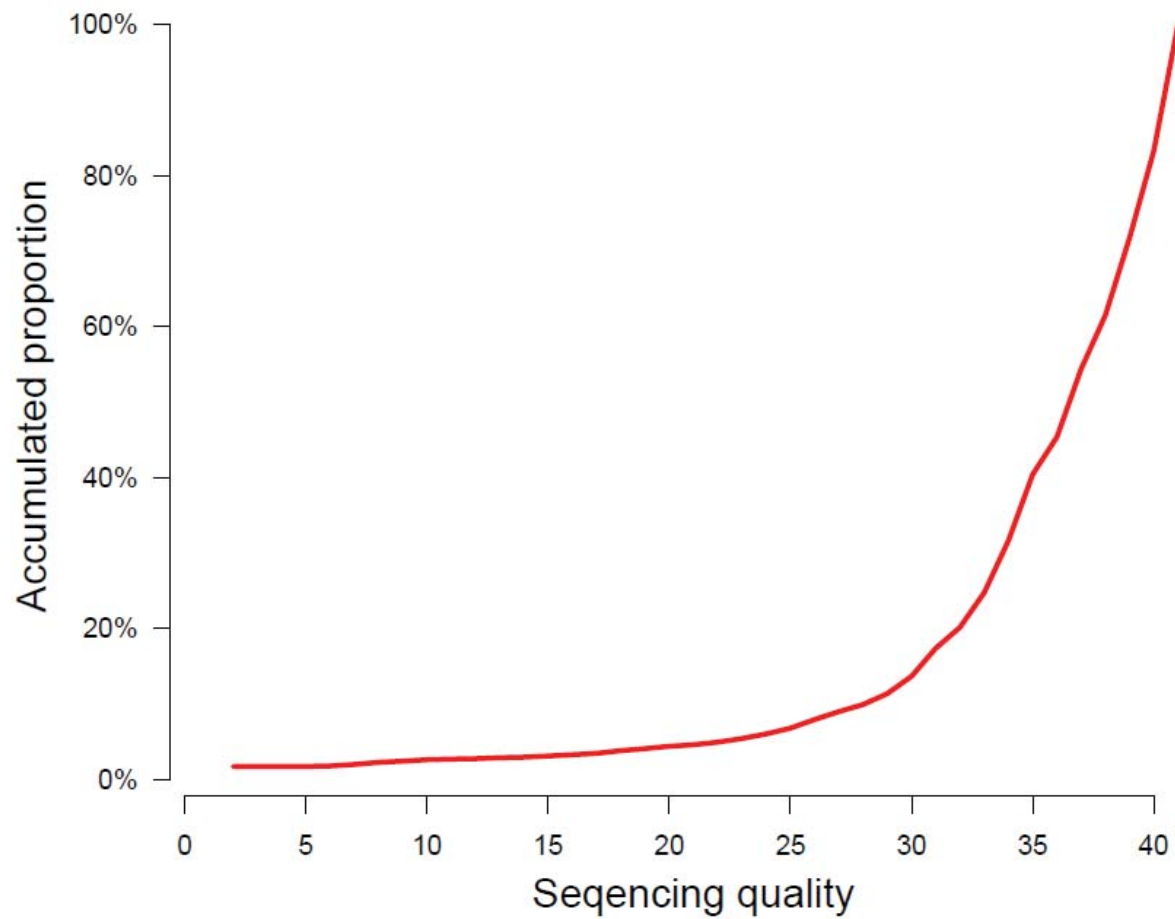
^aThese authors contributed equally to the work



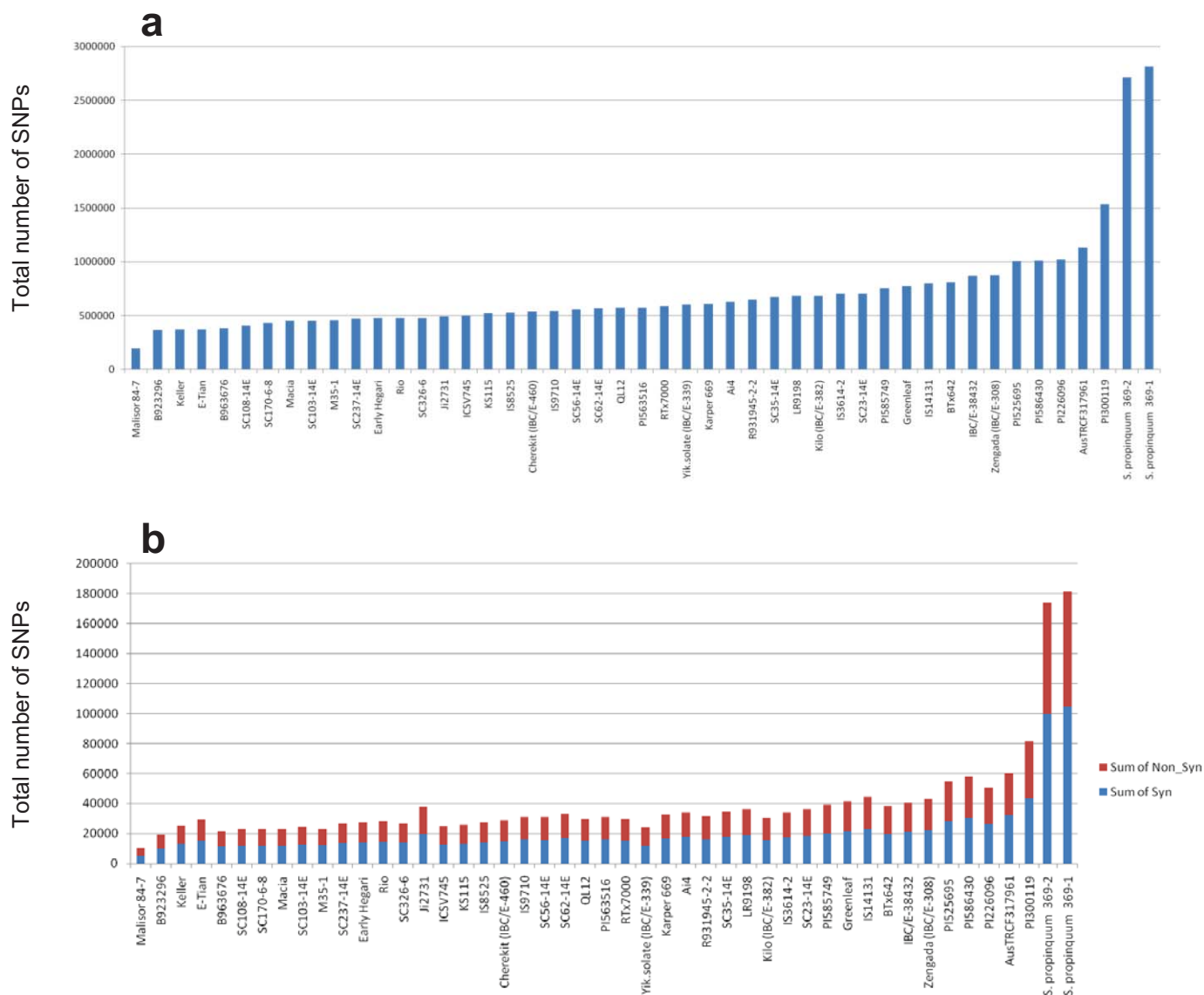
Supplementary Figure S1. Diversity analysis on a subset of sorghums in current study (in red) integrated with 178 genotypes from the Generation Challenge Program sorghum reference set (http://www.icrisat.org/what-we-do/crops/sorghum/Sorghum_Reference.htm). This figure presents the Principal Components Analysis based on a set of 326 common DArT markers. A total of 19.7% variance is explained (Axis 1 10.5%; Axis 2 9.2%)



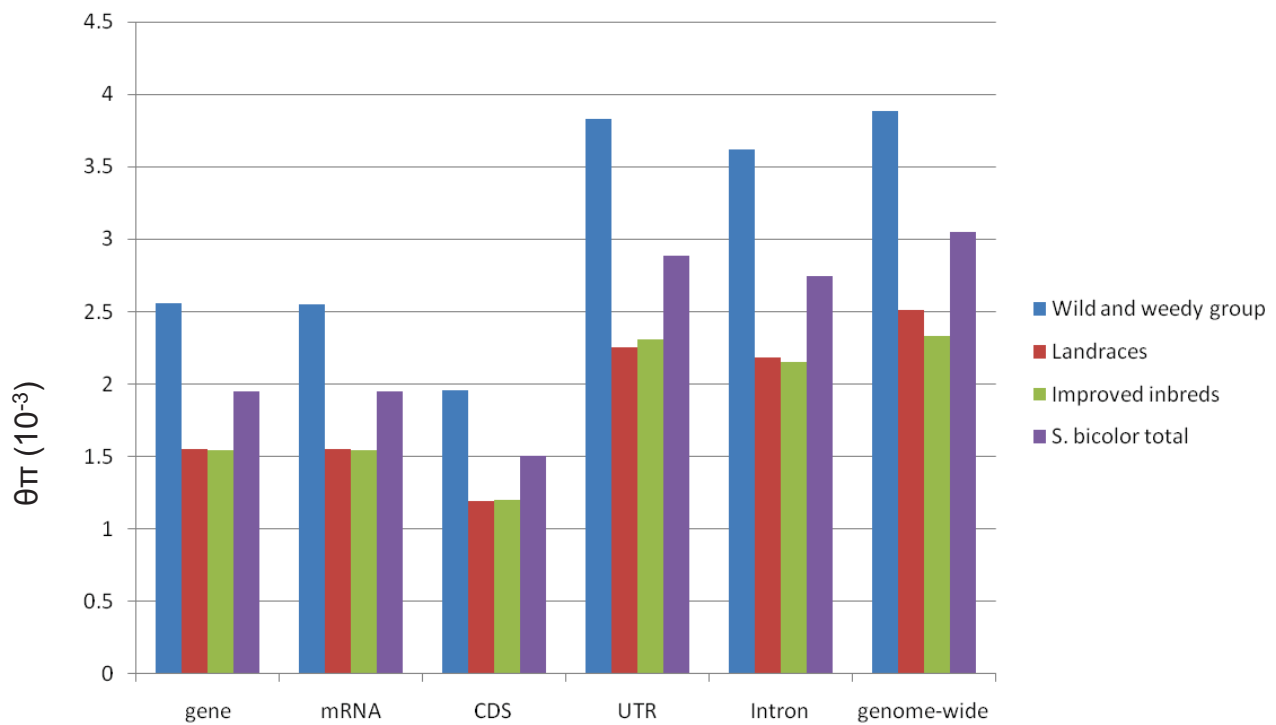
Supplementary Figure S2. Sorghum domestication flow-chart . The key groups represented in this study (numbers in each group in brackets) are detailed.



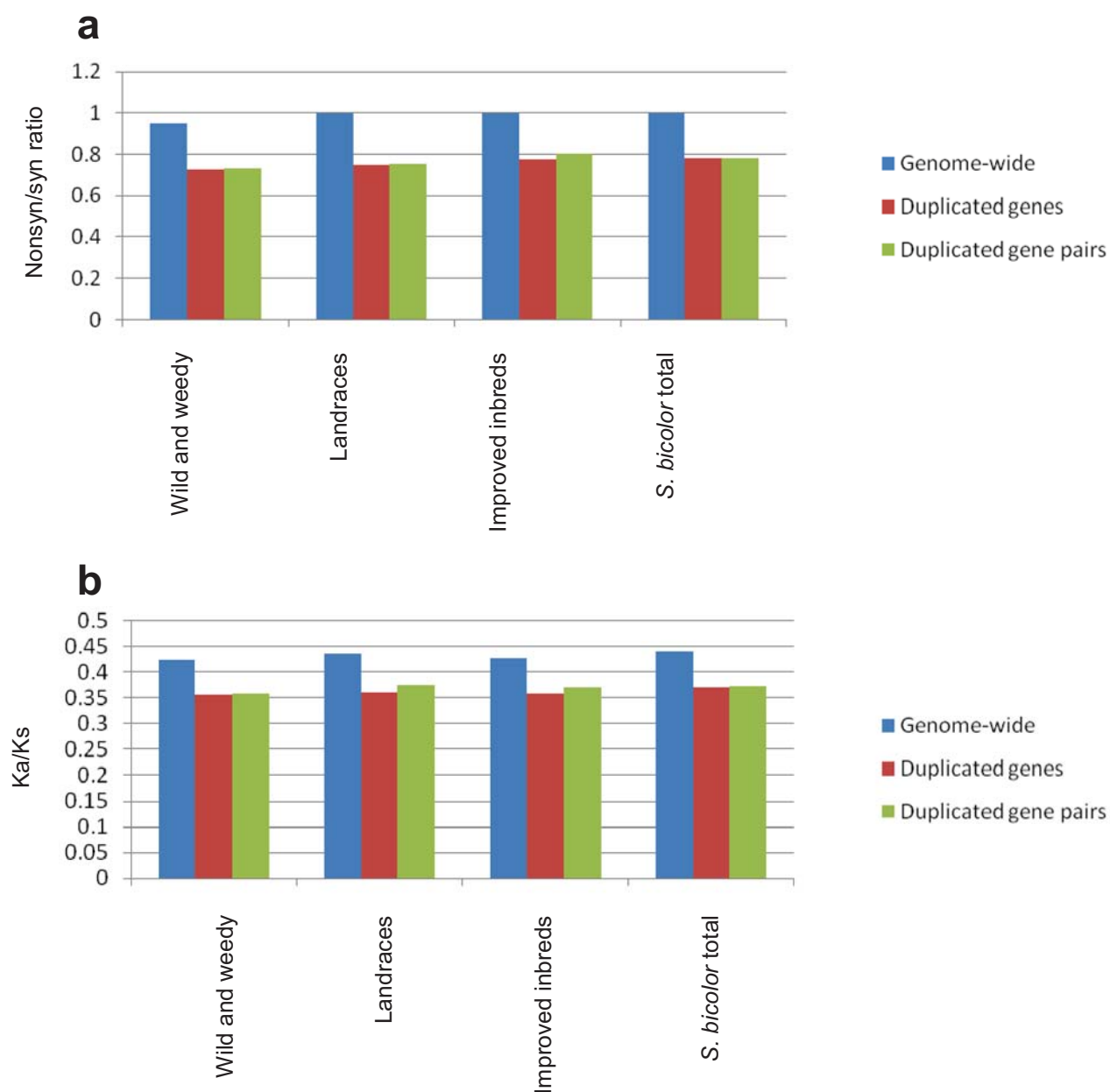
Supplementary Figure S3. Phred quality score of bases in all the reads. The y-axis is the accumulated proportion of bases, which indicates the proportion of bases with phred quality less than a given value (x-axis). More than 90% of bases have phred quality score greater than 29, indicating high sequence quality.



Supplementary Figure S4. The number of SNPs per genotype **(a)** Total number of SNPs (y-axis) per genotype (x-axis); **(b)** Total number of non-synonymous and synonymous SNPs (y-axis) per genotype (x-axis)

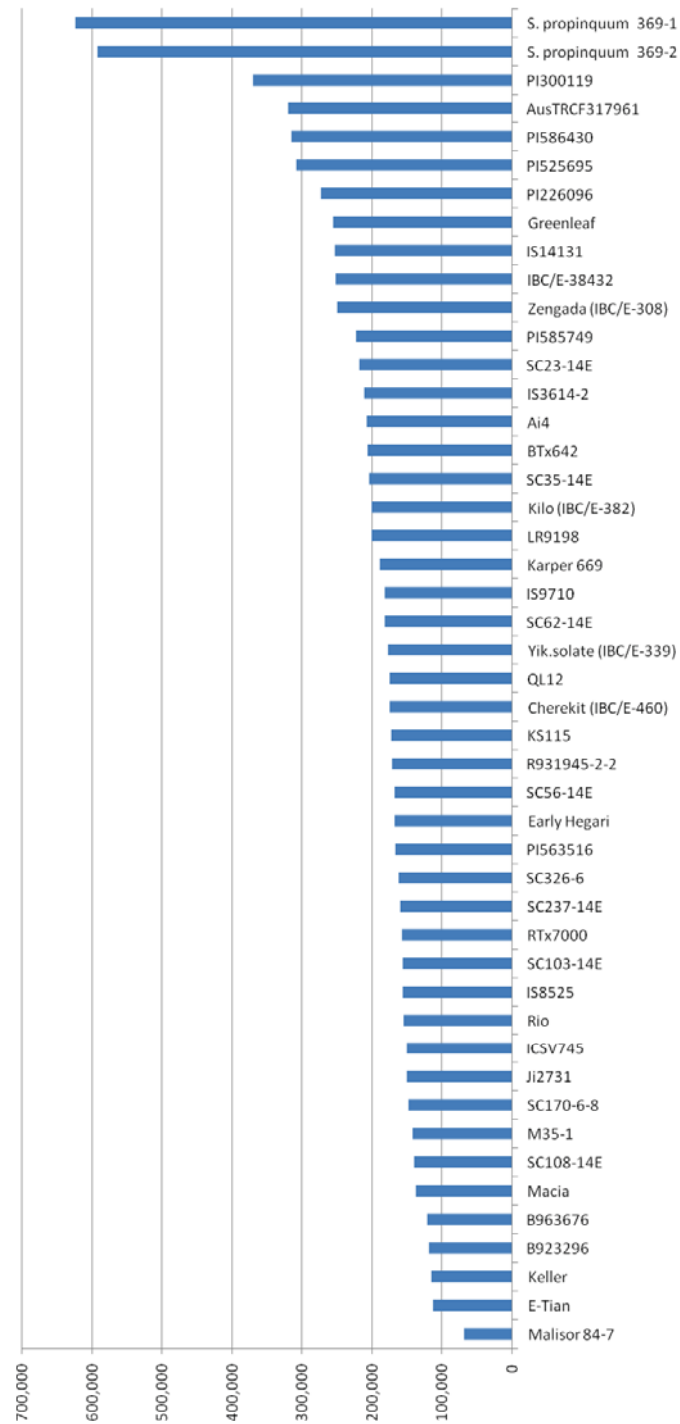


Supplementary Figure S5. Comparison of the diversity ($\theta\pi \times 10^{-3}$) levels across genic and non-genic regions. The genic regions include both coding and non-coding regions of the genes (gene, mRNA, CDS, UTR and intron), in comparison to the genic and non-genic regions genome-wide.

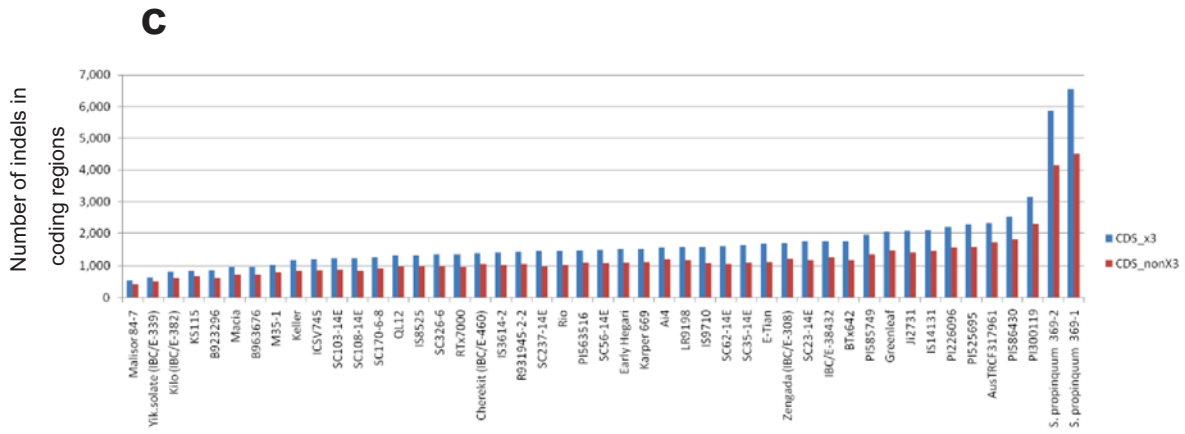
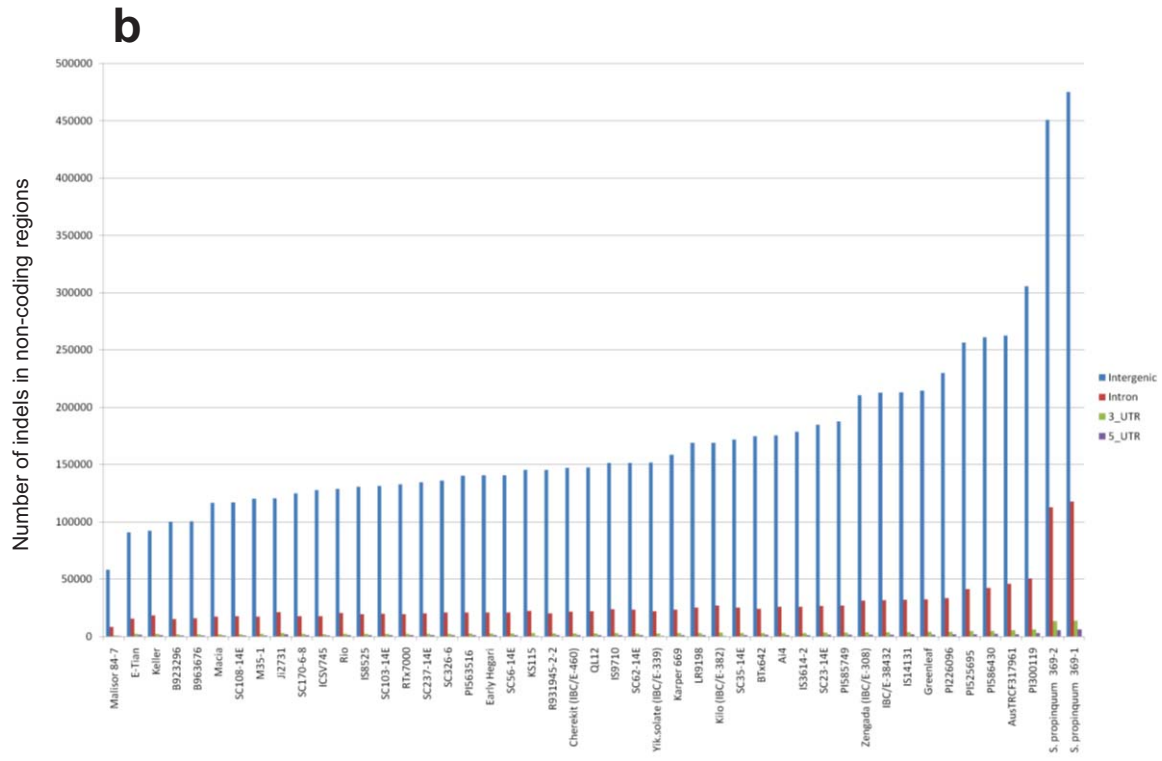


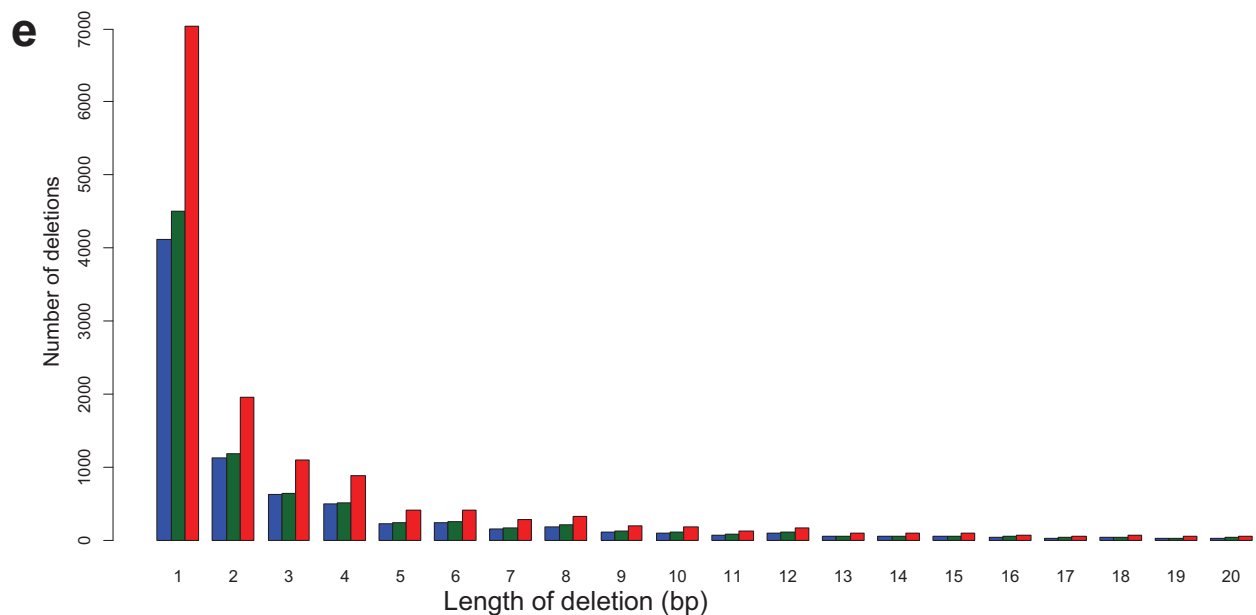
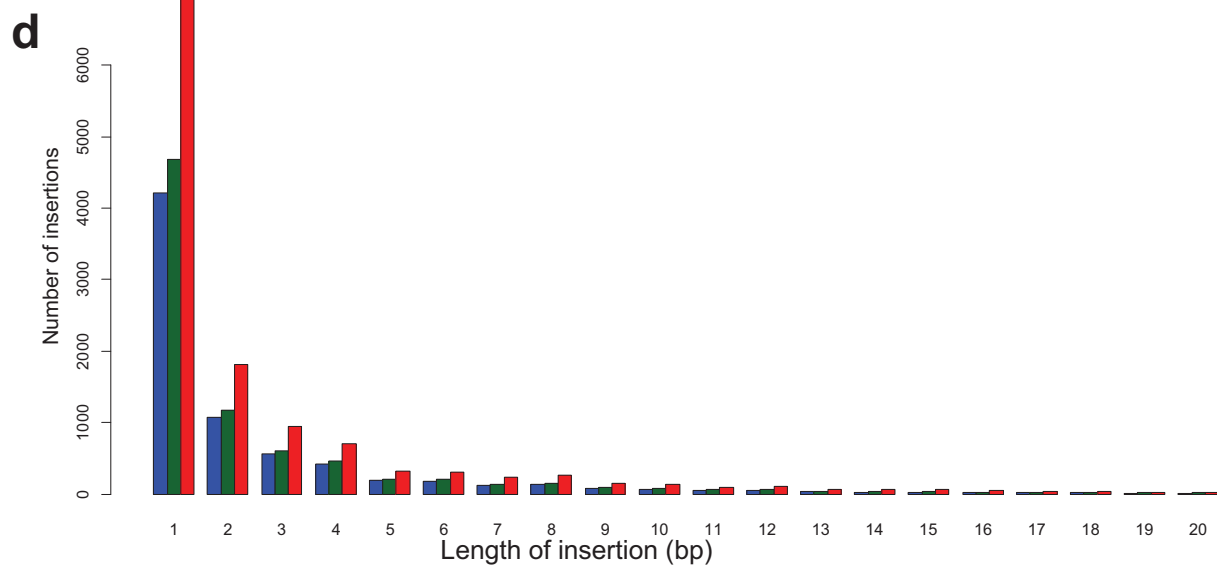
Supplementary Figure S6. Comparison of the number of synonymous versus non-synonymous SNPs; (a) Genome-wide non-synonymous/synonymous substitution ratio comparisons across groups; (b) Genome-wide Ka/Ks comparisons across groups

a

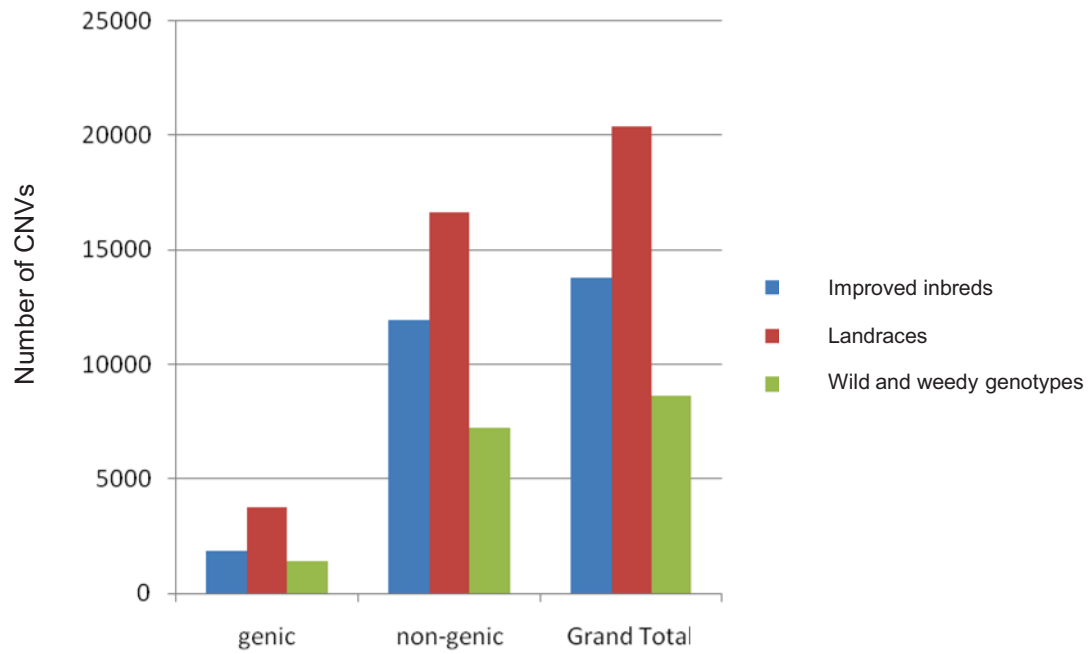


Number of indels

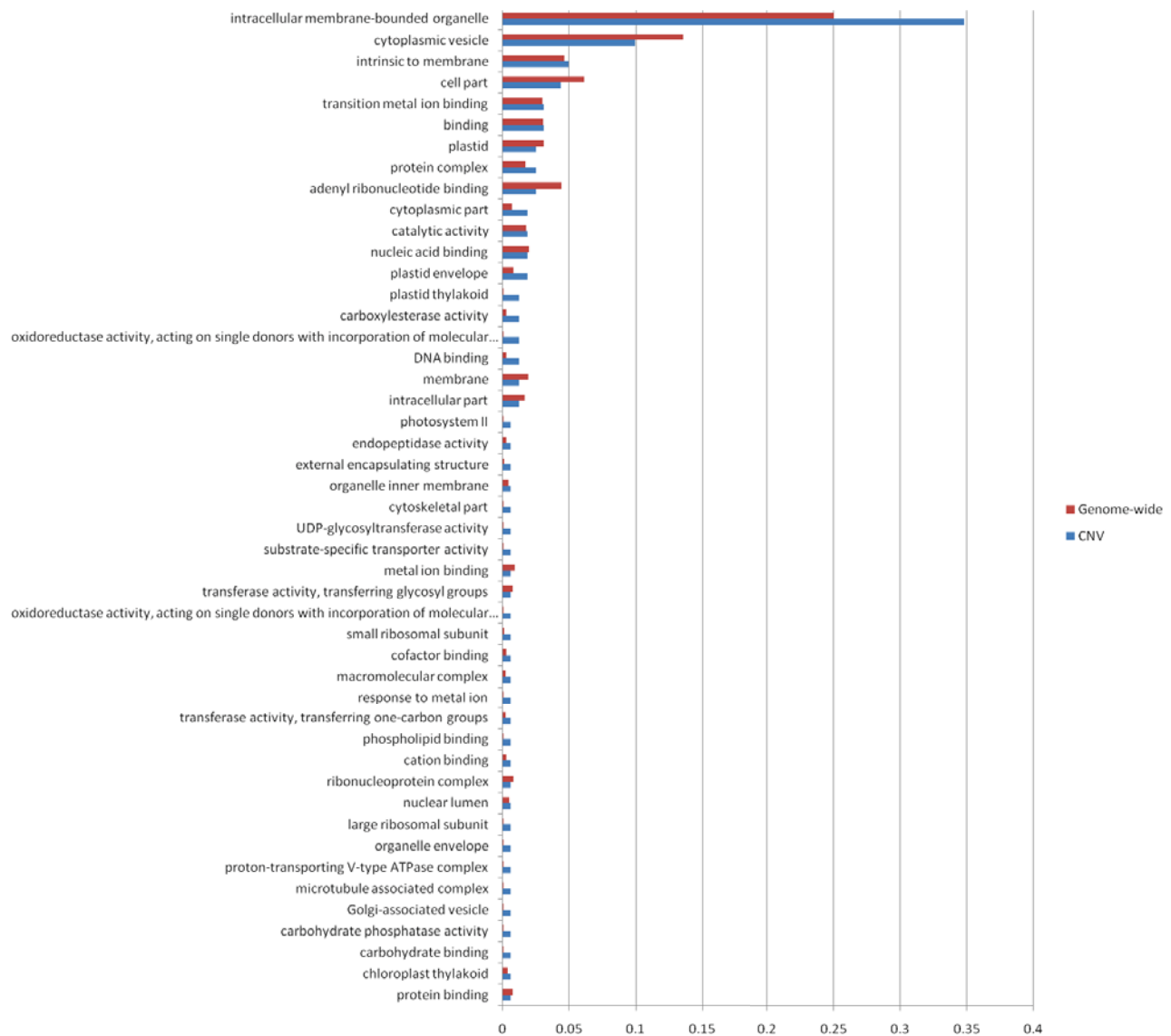




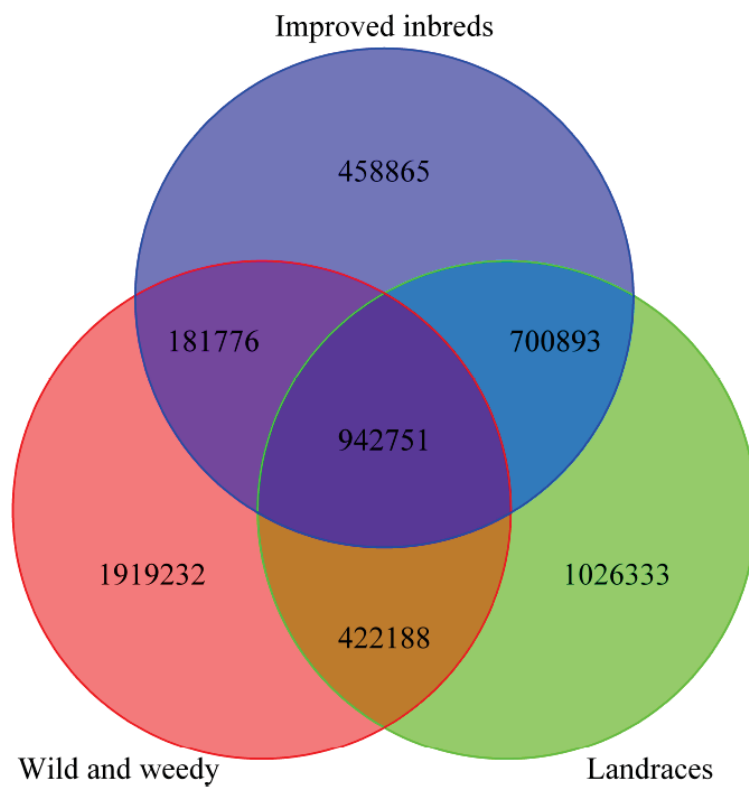
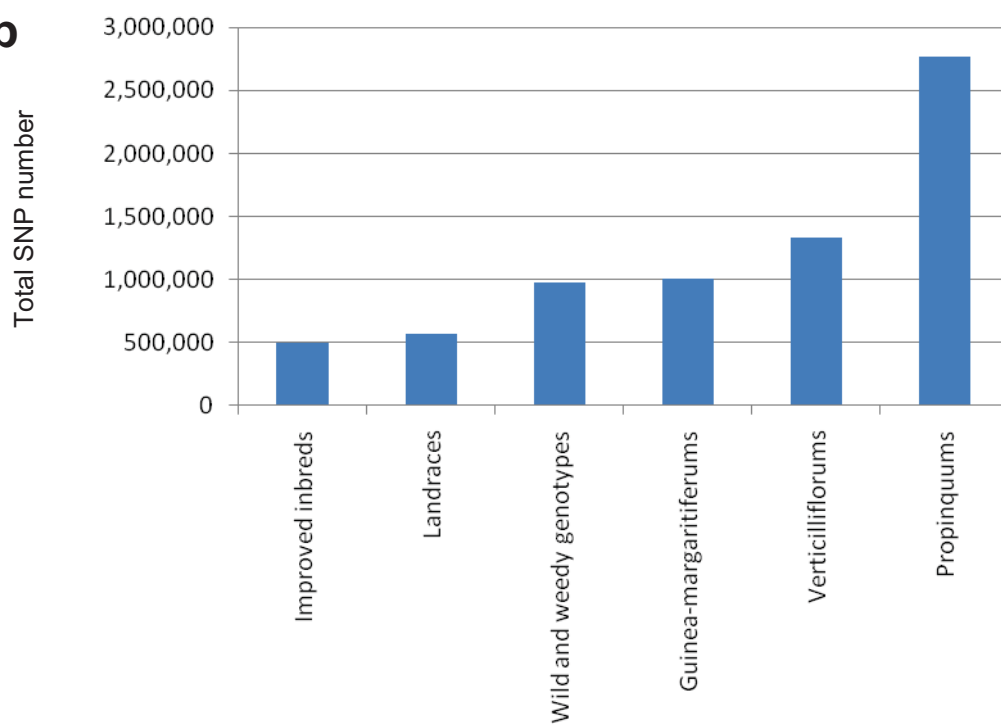
Supplementary Figure S7. The number and length of indels identified **(a)** Total number of indels (y-axis) per genotype (x-axis); **(b)** Total number of indels in non-coding regions per genotype; **(c)** Total number of indels in coding regions per genotype; **(d)** Distribution of small to medium insertions (1-20bp; x-axis) within the wild and weedy group (red), landraces (green) and improved inbreds (blue). **(e)** Distribution of small to medium deletions (1-20bp; x-axis) within the wild and weedy group (red), landraces (green) and improved inbreds (blue).

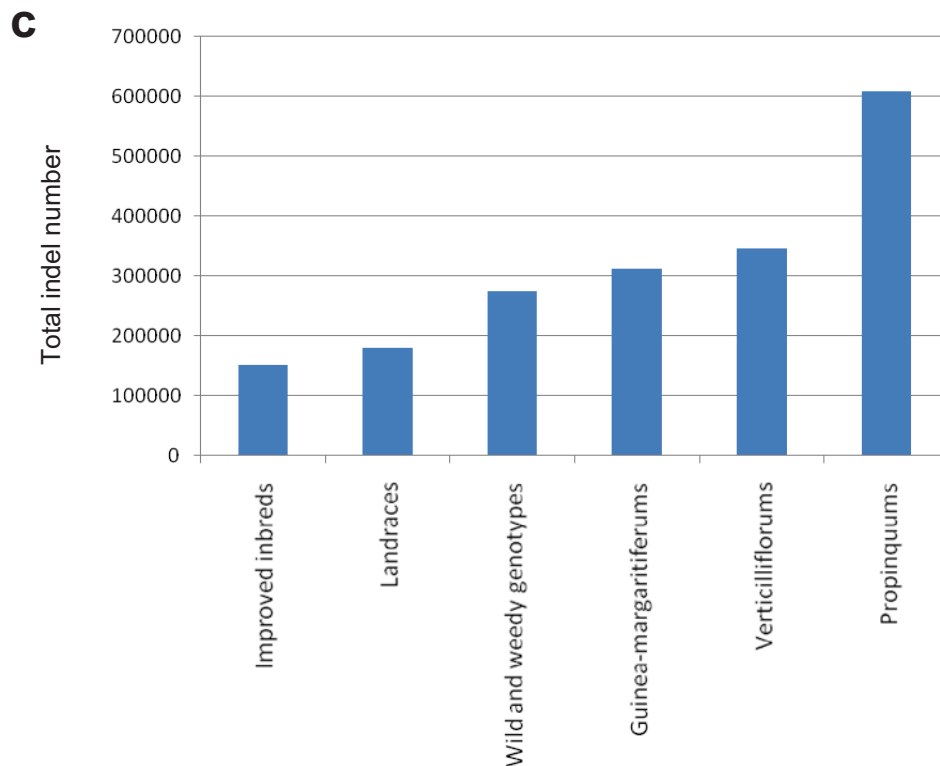


Supplementary Figure S8. Distribution of copy number variations (CNVs) across genotype groups and genomic regions. The genomic regions are grouped into genic and non-genic regions.

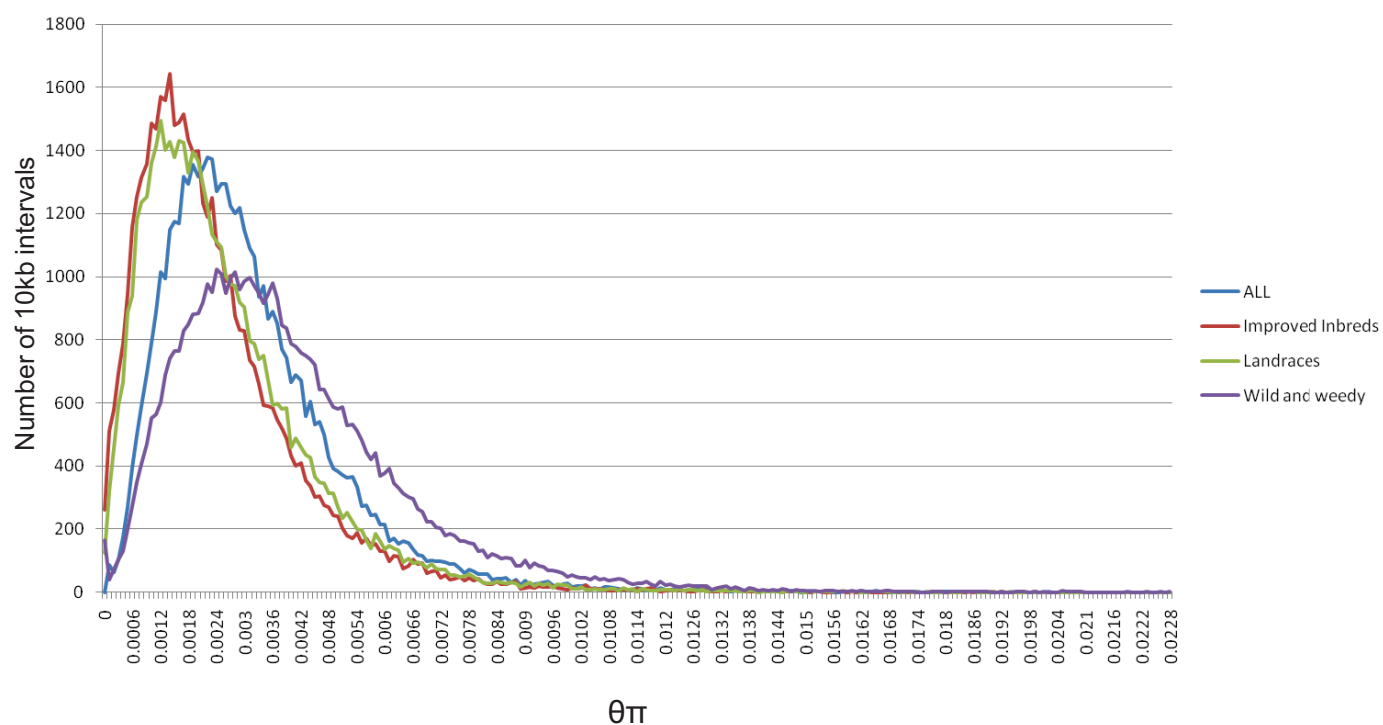


Supplementary Figure S9. GO-SLIM enrichment analysis⁴⁴ of 224 genes with CNVs across the wild and weedy genotypes, landraces and improved inbreds. The genome-wide frequencies across genotype groups are presented in red and the frequencies of the genes with CNVs are presented in blue.

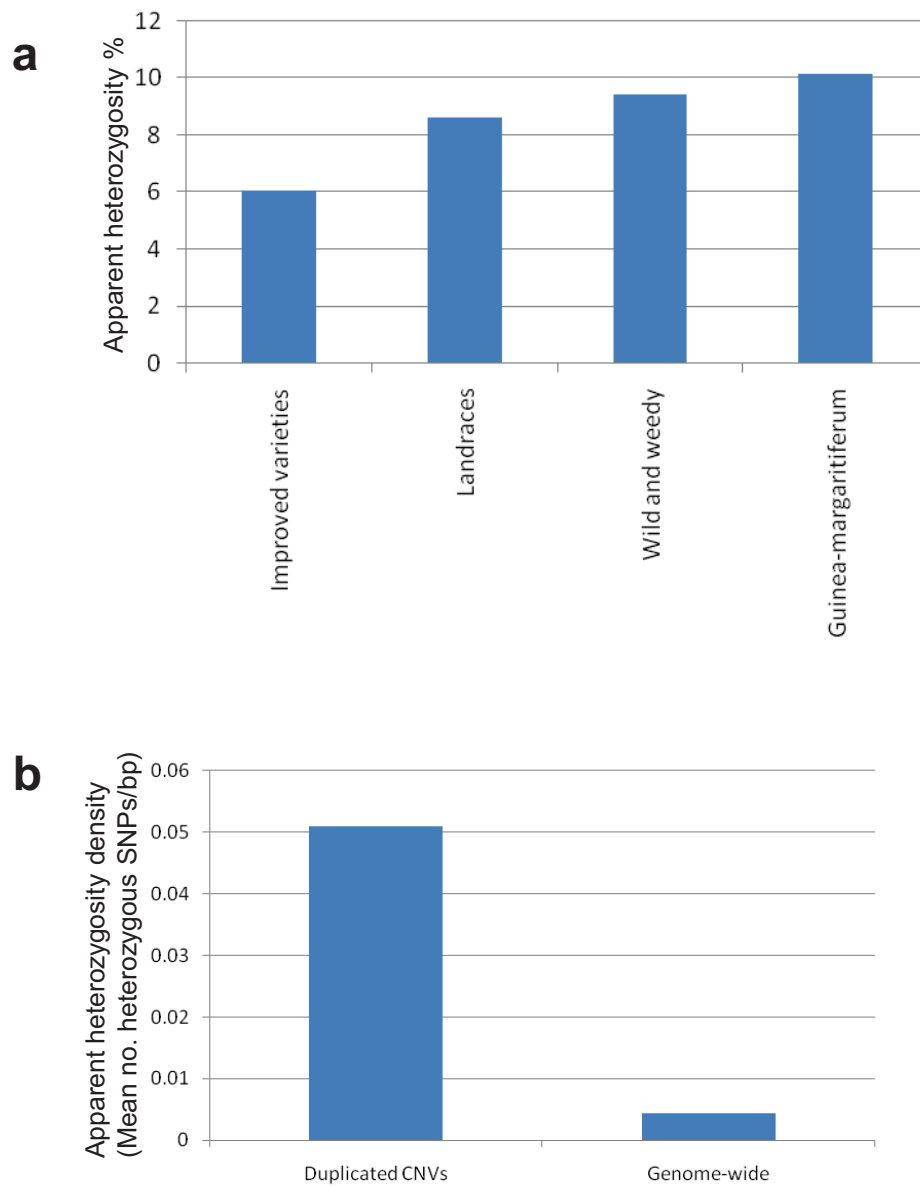
a**b**



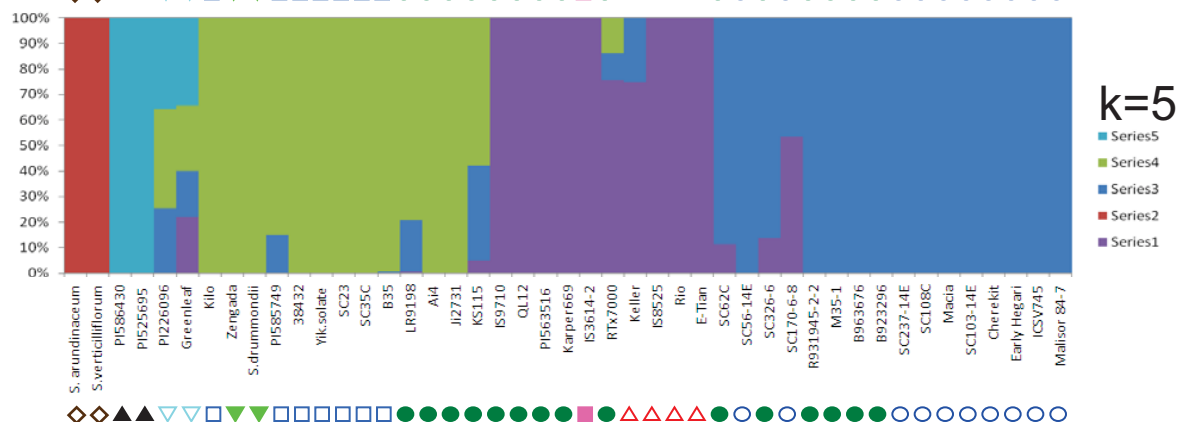
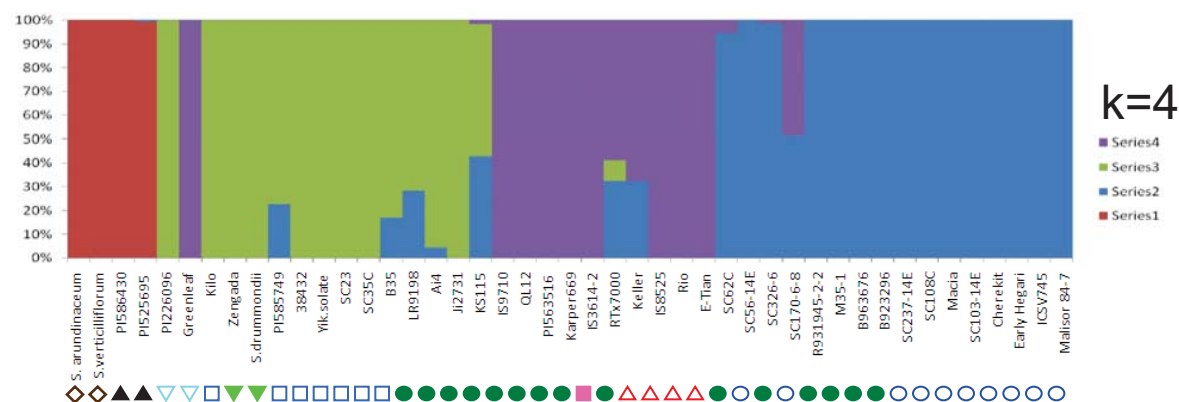
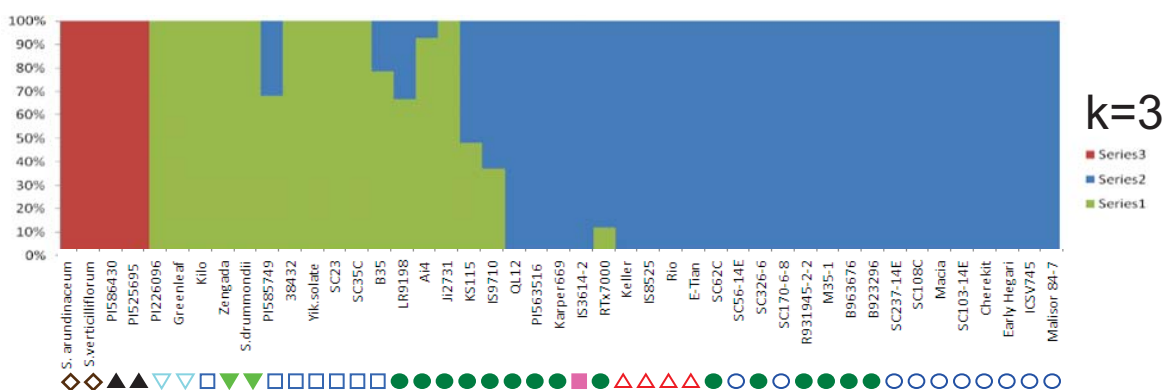
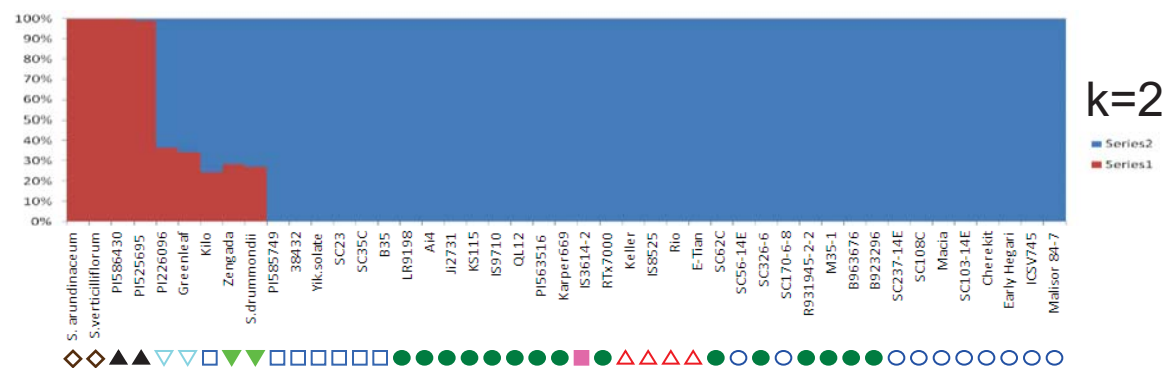
Supplementary Figure S10. Summaries of the total number of sequence variants across groups **(a)** Total number of SNPs within and amongst the wild and weedy genotypes, the landraces and improved inbreds groups; **(b)** Average total number of SNPs per genotype per group; **(c)** Average total number of indels per genotype per group

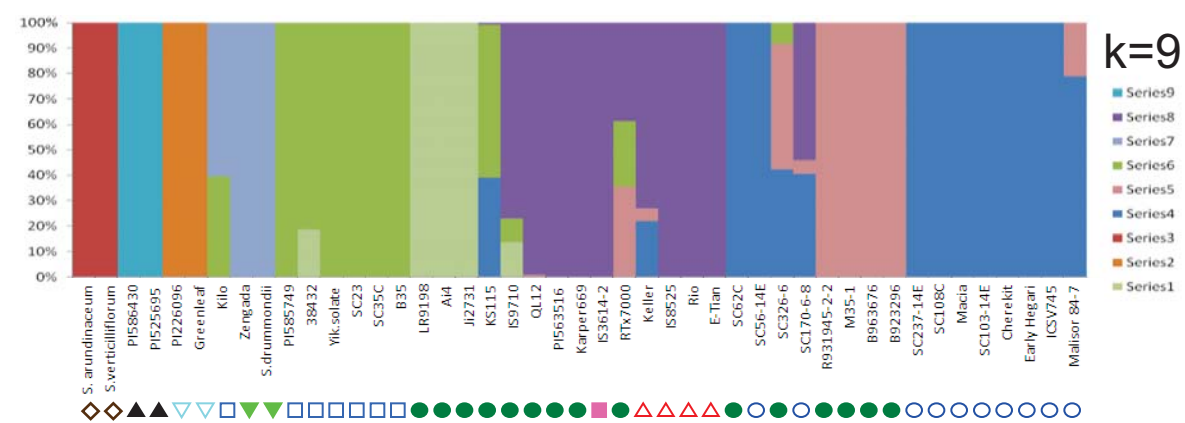
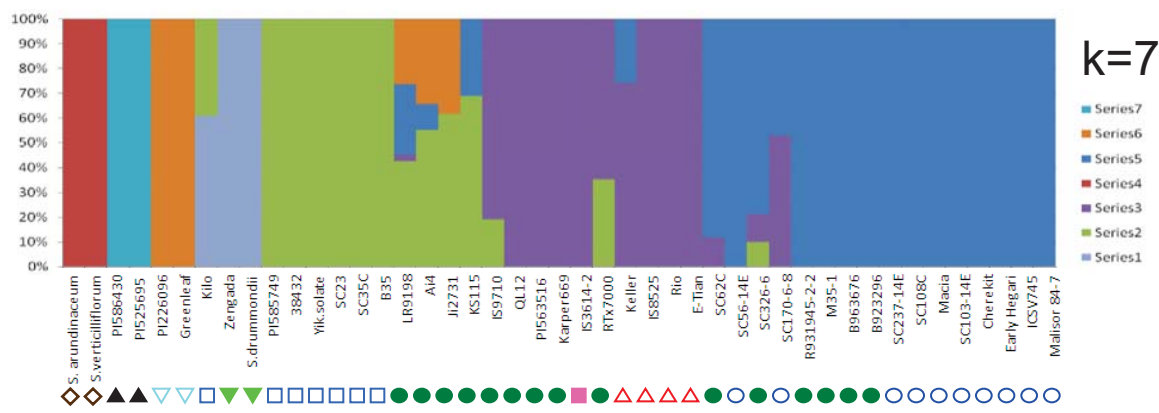
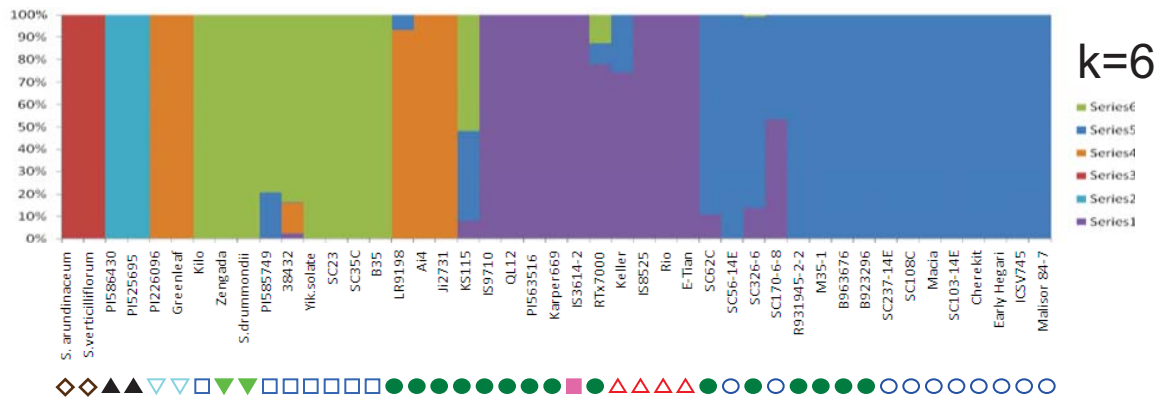


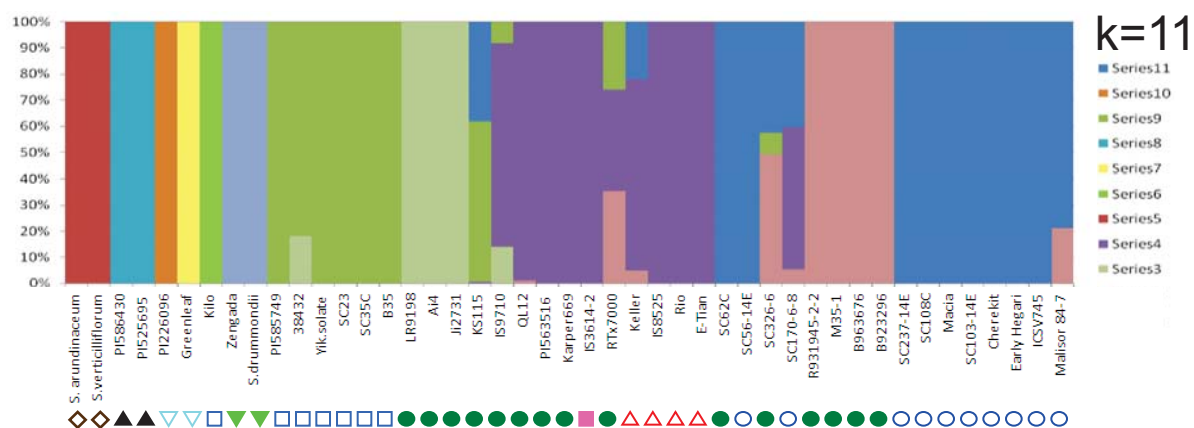
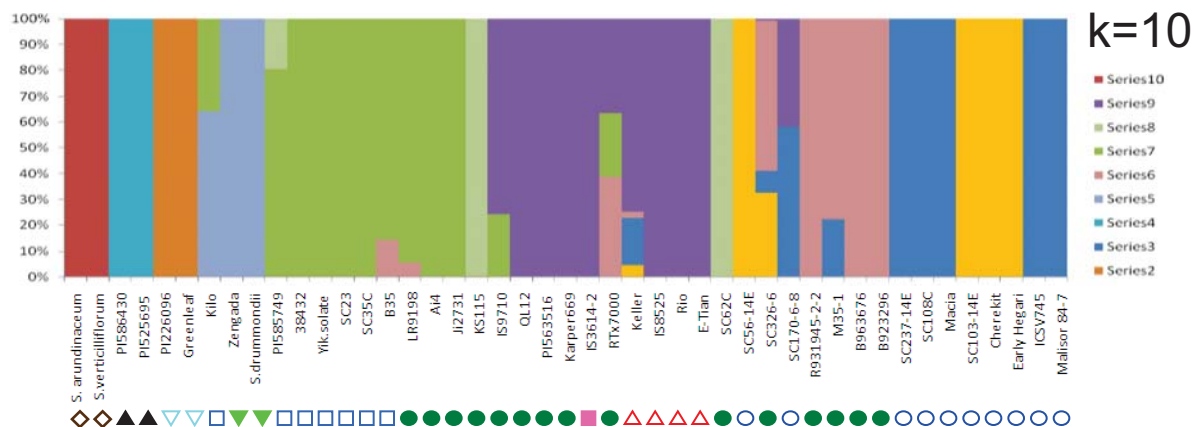
Supplementary Figure S11. Frequency distribution of diversity ($\theta\pi$) statistic (x-axis) within the genotype groups. Colour-coding is as follows; wild and weedy group (purple), landrace group (green), the improved inbred group (red), and all genotypes (blue).



Supplementary Figure S12. Summaries of apparent heterozygosity rates
(a). Apparent mean heterozygosity rate across groups, calculated based on the number of heterozygous SNPs relative to the total number of SNPs identified within the cultivated gene-pool; (b) Density of apparent heterozygous SNPs in identified duplicated CNVs compared to the density of apparent heterozygous SNPs genome-wide

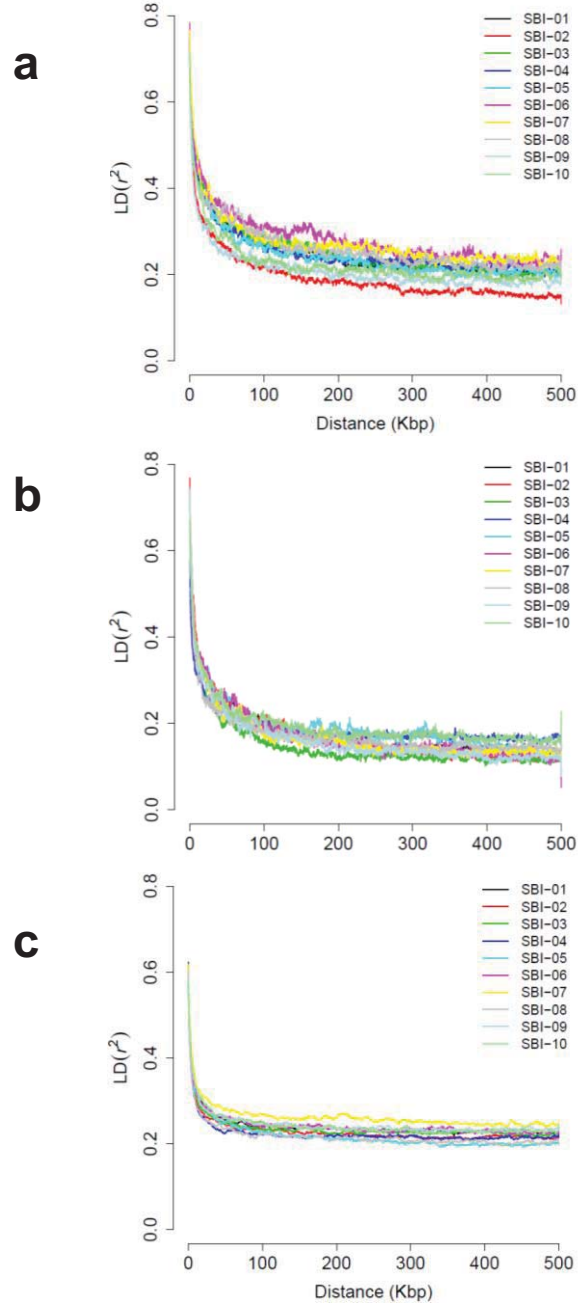




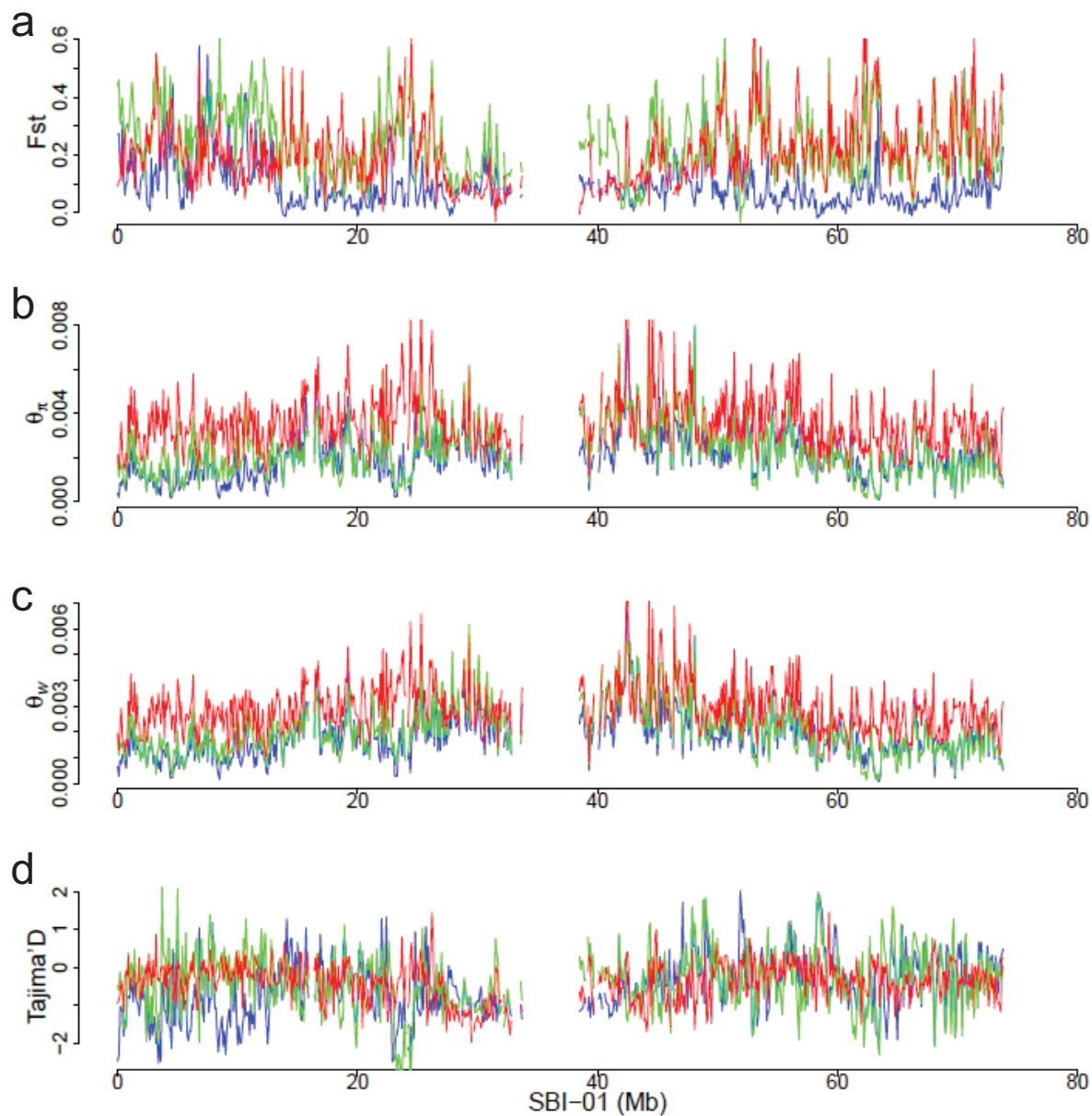


Supplementary Figure S13. Population structure analysis using FRAPPE⁵⁵. Each genotype is represented by a vertical bar, and the length of the coloured segment in each vertical bar represents the proportion contributed by ancestral populations. The genotypes are coded as follows.

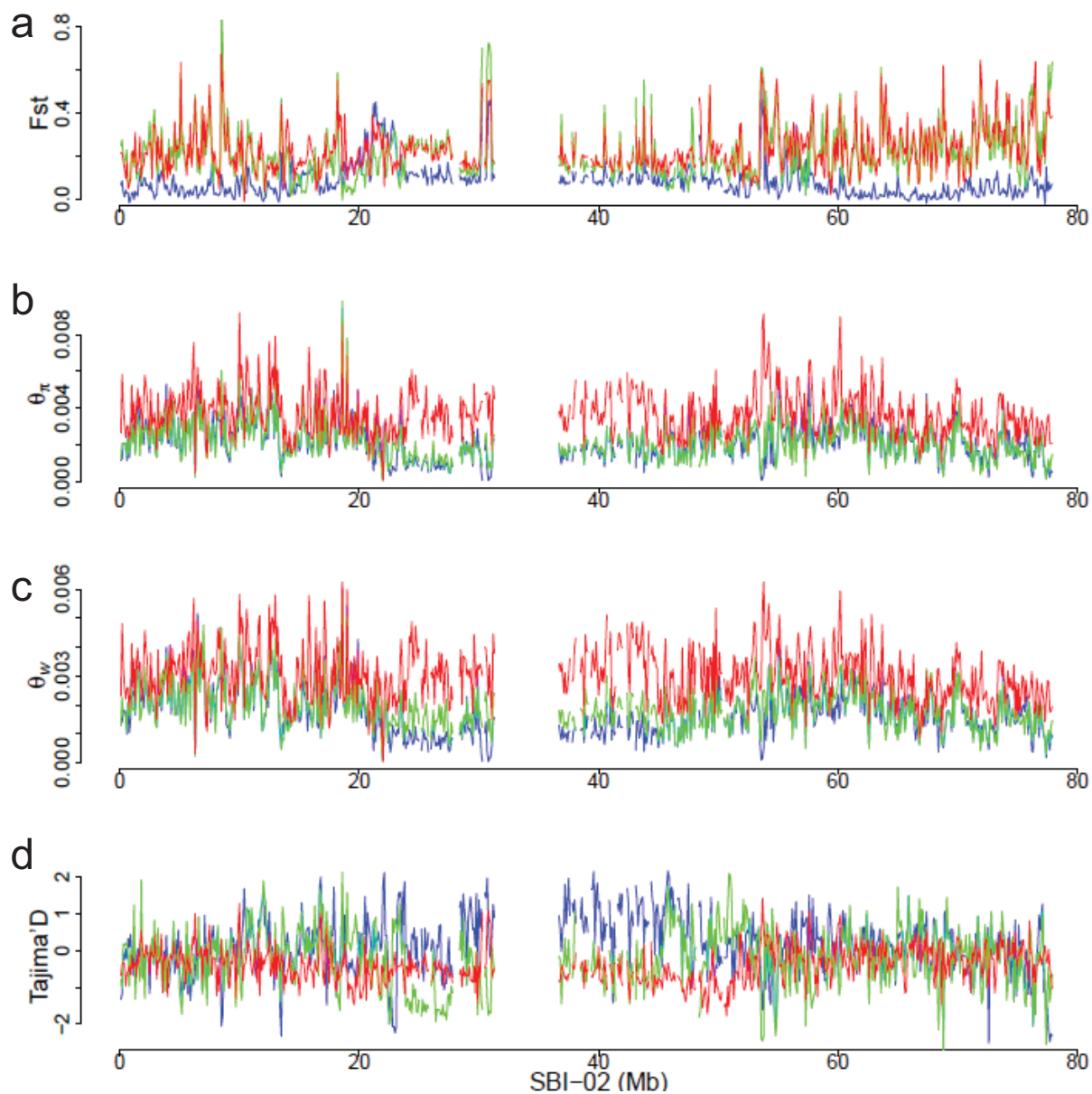
○ Caudatum; △ Kafir; □ Durra; ■ Guinea; ● Complex; ▲ Guinea-margaritiflorum; ▽ Weedy bicolor; ▼ drummondii; ◇ Wild; ◆ Propinquum



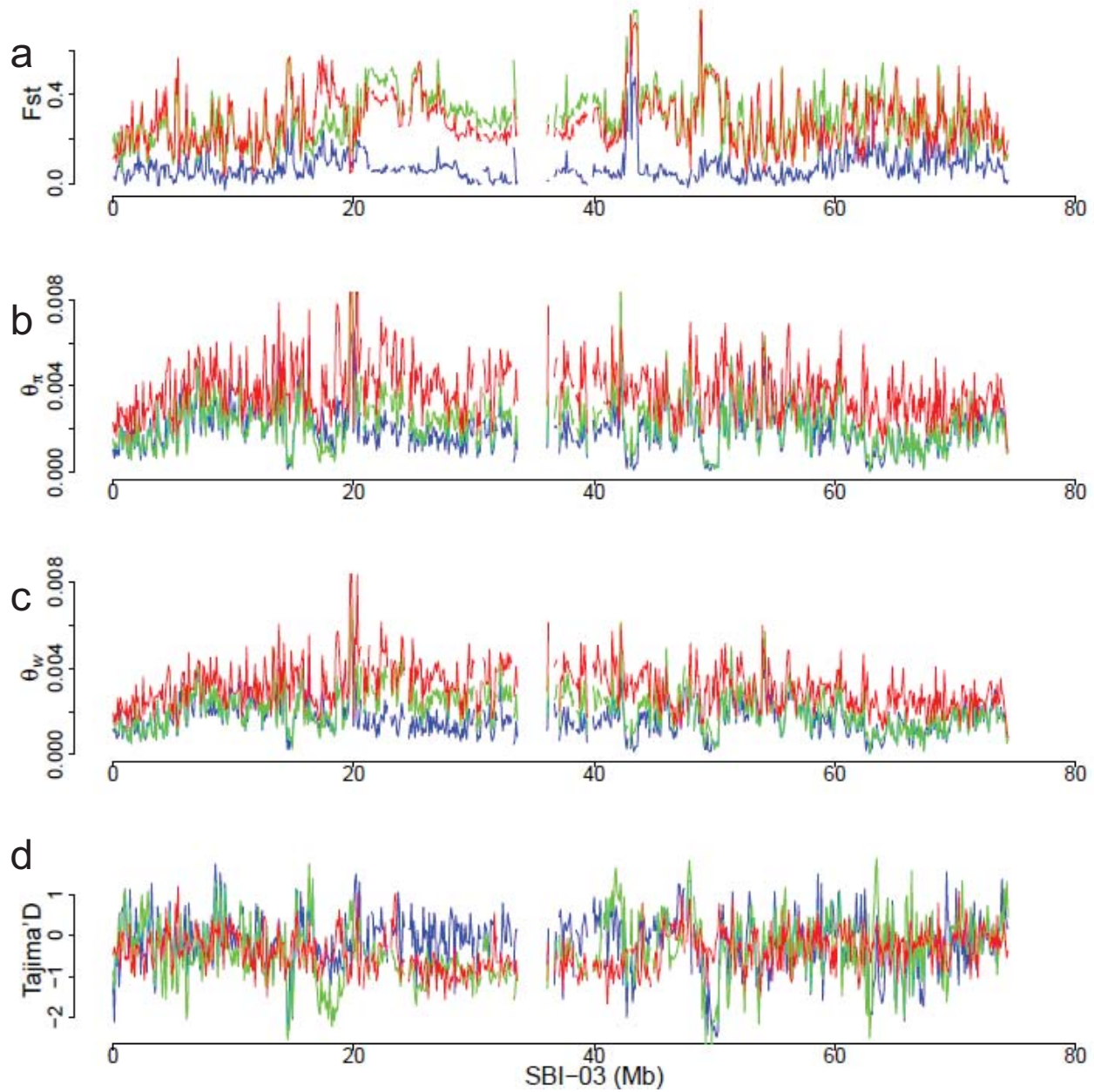
Supplementary Figure S14. Linkage Disequilibrium plots per chromosome: **(a)** Improved inbreds; **(b)** Landraces; **(c)** Wild and weedy genotypes



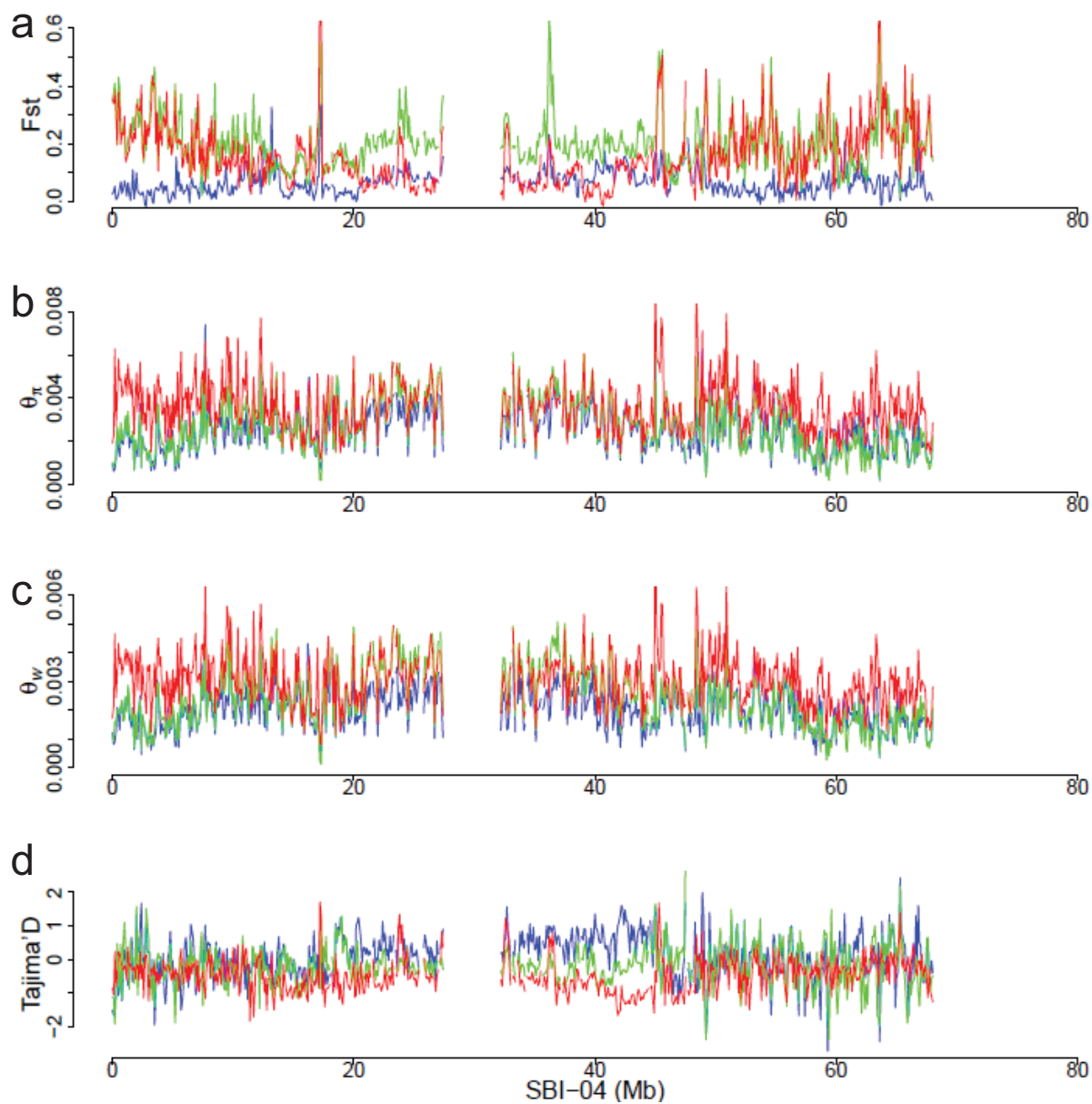
Supplementary Figure S15. Distribution of population summary statistics within and between groups for SBI-01: **(a)** group differentiation (F_{ST}); **(b)** diversity θ_{π} ; **(c)** diversity θ_w ; **(d)** allele frequency spectra (Tajima's D). Groups colour-coded as follows: Improved inbreds (blue), landraces (green), wild and weedy genotypes (red).



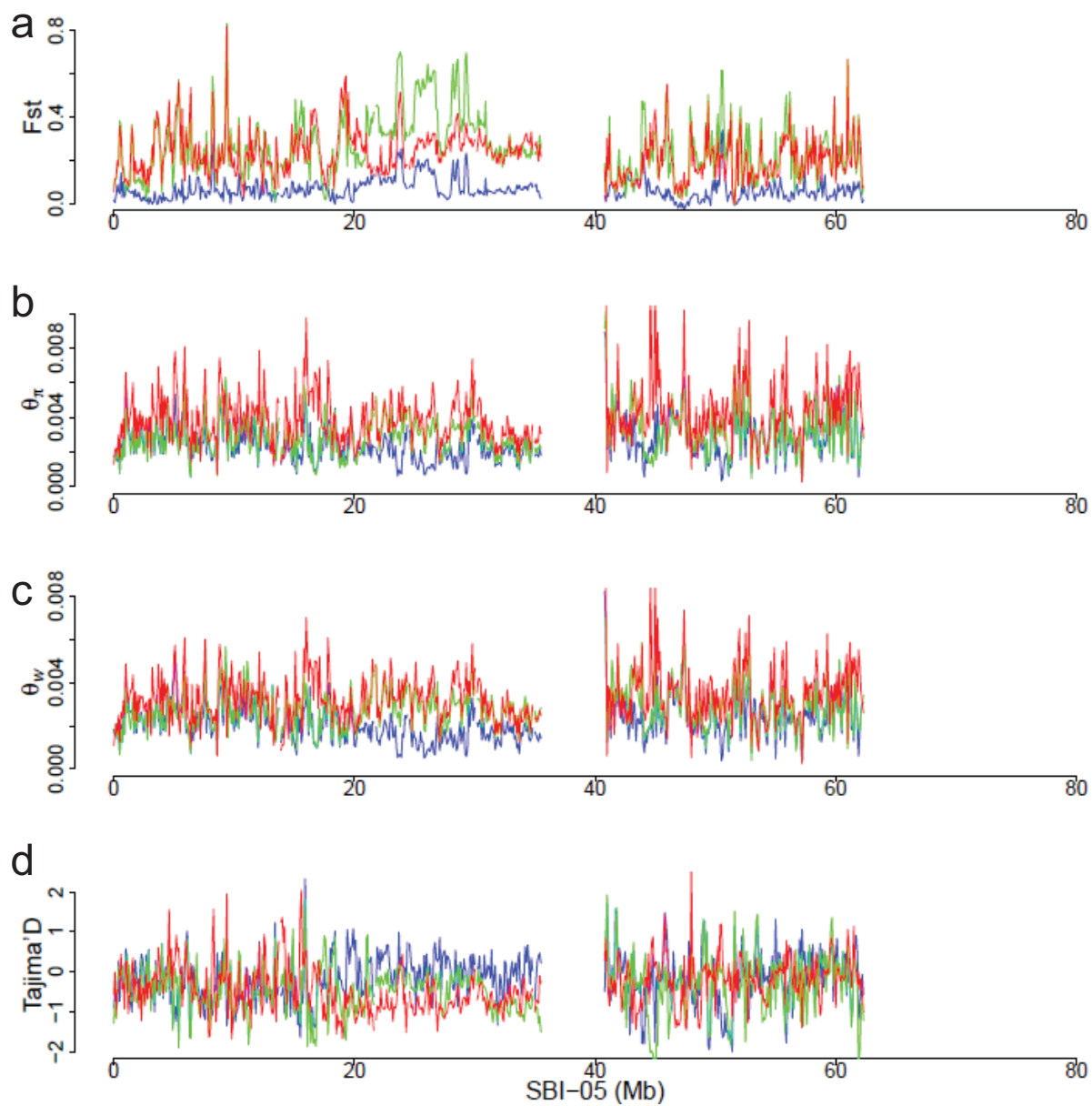
Supplementary Figure S16. Distribution of population summary statistics within and between groups for SBI-02: **(a)** group differentiation (F_{ST}); **(b)** diversity θ_π ; **(c)** diversity θ_w ; **(d)** allele frequency spectra (Tajima's D). Groups colour-coded as follows: Improved inbreds (blue), landraces (green), wild and weedy genotypes (red).



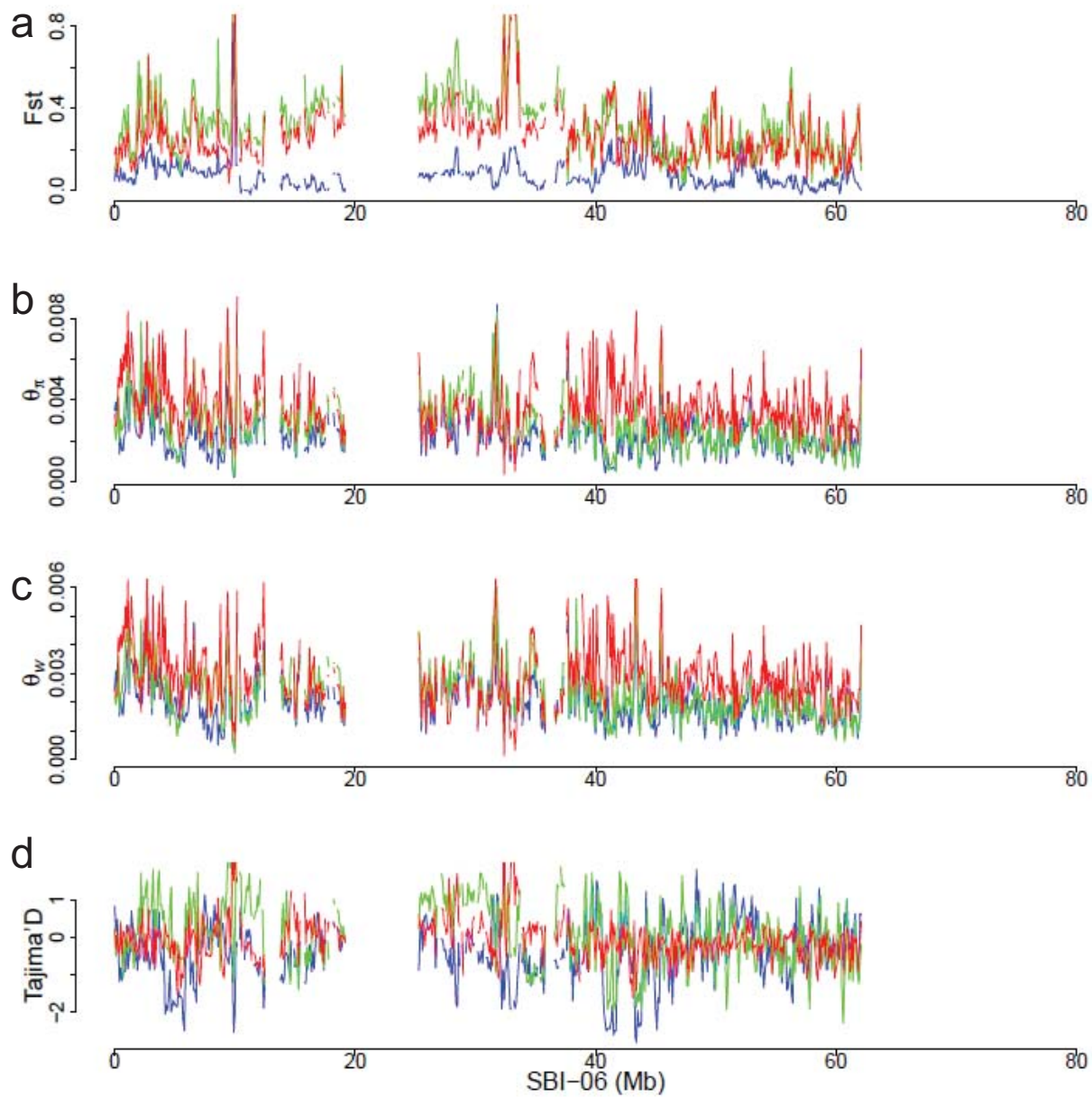
Supplementary Figure S17. Distribution of population summary statistics within and between groups for SBI-03: **(a)** group differentiation (F_{ST}); **(b)** diversity θ_{π} ; **(c)** diversity θ_w ; **(d)** allele frequency spectra (Tajima's D). Groups colour-coded as follows: Improved inbreds (blue), landraces (green), wild and weedy genotypes (red).



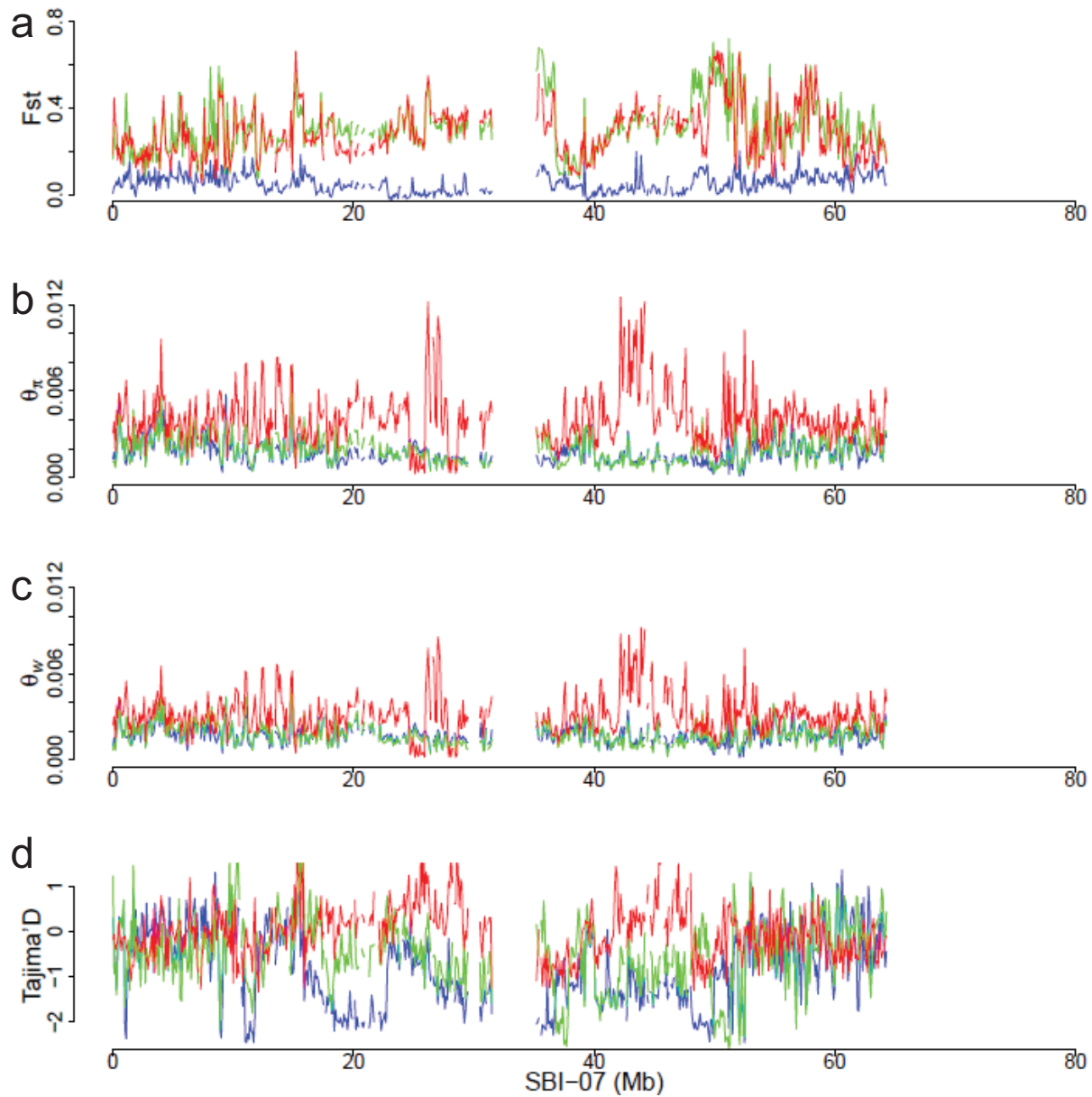
Supplementary Figure S18. Distribution of population summary statistics within and between groups for SBI-04: **(a)** group differentiation (F_{ST}); **(b)** diversity θ_{π} ; **(c)** diversity θ_w ; **(d)** allele frequency spectra (Tajima's D). Groups colour-coded as follows: Improved inbreds (blue), landraces (green), wild and weedy genotypes (red).



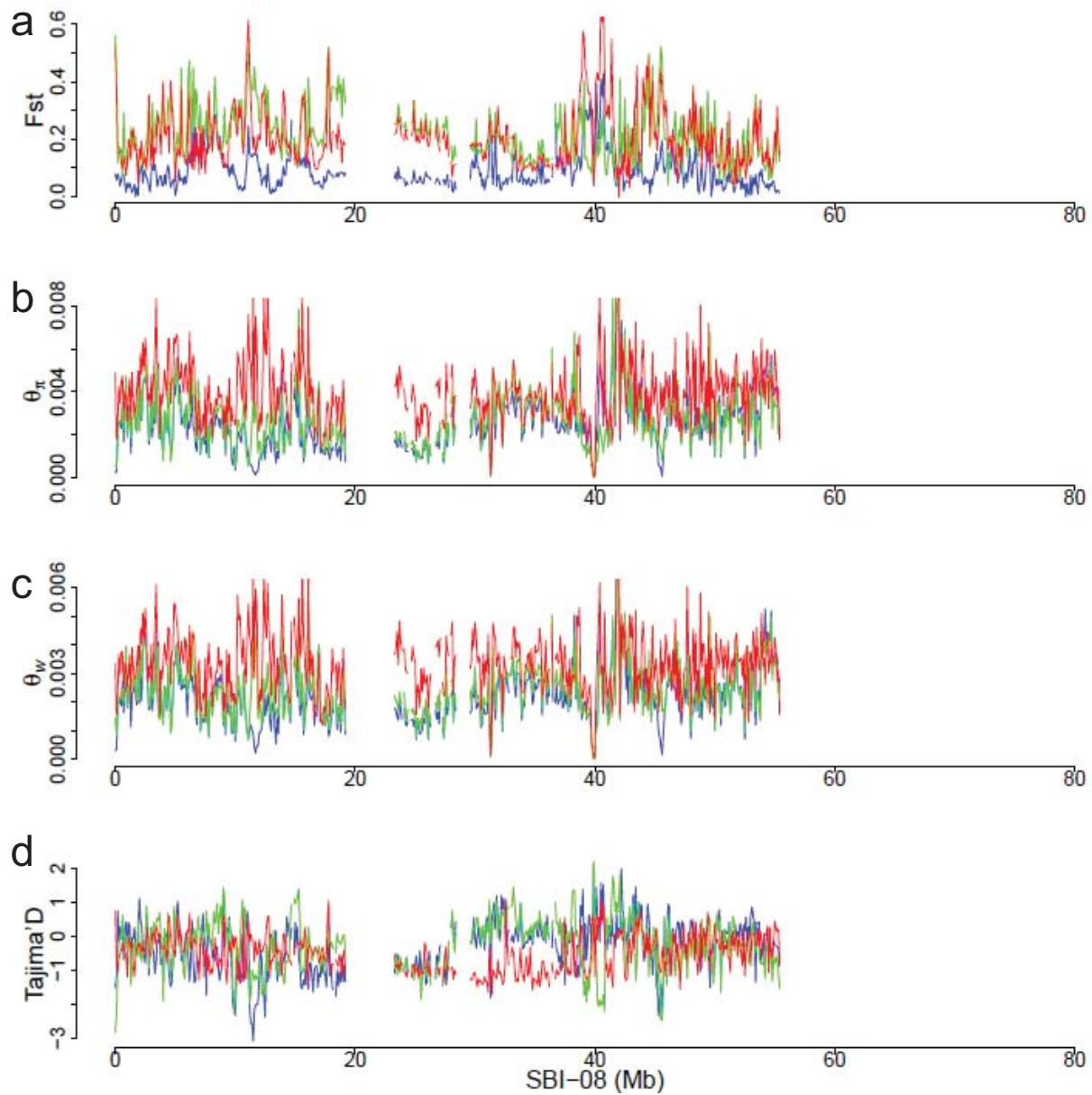
Supplementary Figure S19. Distribution of population summary statistics within and between groups for SBI-05: **(a)** group differentiation (F_{ST}); **(b)** diversity θ_{π} ; **(c)** diversity θ_w ; **(d)** allele frequency spectra (Tajima's D). Groups colour-coded as follows: Improved inbreds (blue), landraces (green), wild and weedy genotypes (red).



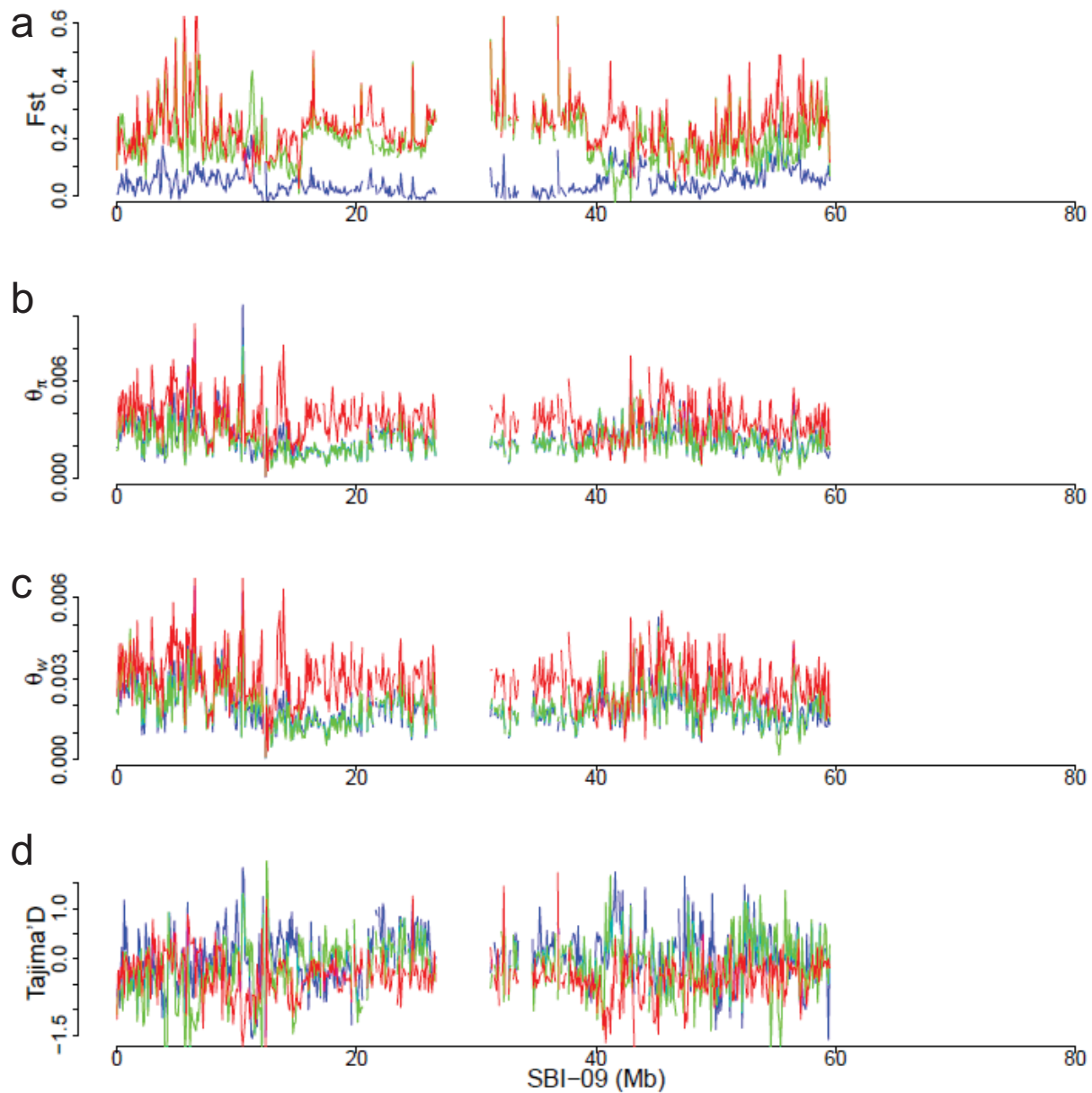
Supplementary Figure S20. Distribution of population summary statistics within and between groups for SBI-06: **(a)** group differentiation (F_{ST}); **(b)** diversity θ_{π} ; **(c)** diversity θ_w ; **(d)** allele frequency spectra (Tajima's D). Groups colour-coded as follows: Improved inbreds (blue), landraces (green), wild and weedy genotypes (red).



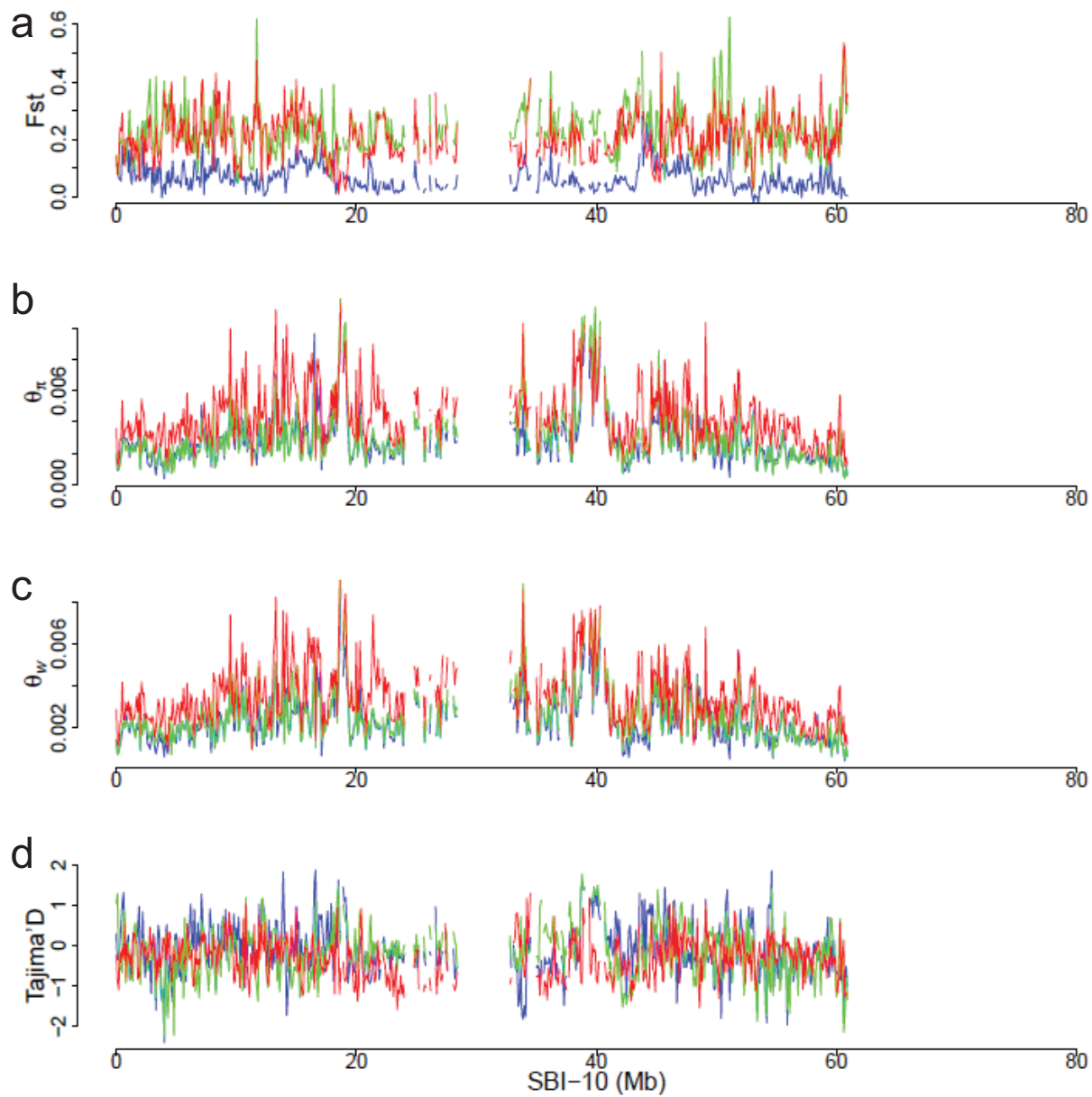
Supplementary Figure S21. Distribution of population summary statistics within and between groups for SBI-07: **(a)** group differentiation (F_{ST}); **(b)** diversity θ_{π} ; **(c)** diversity θ_w ; **(d)** allele frequency spectra (Tajima's D). Groups colour-coded as follows: Improved inbreds (blue), landraces (green), wild and weedy genotypes (red).



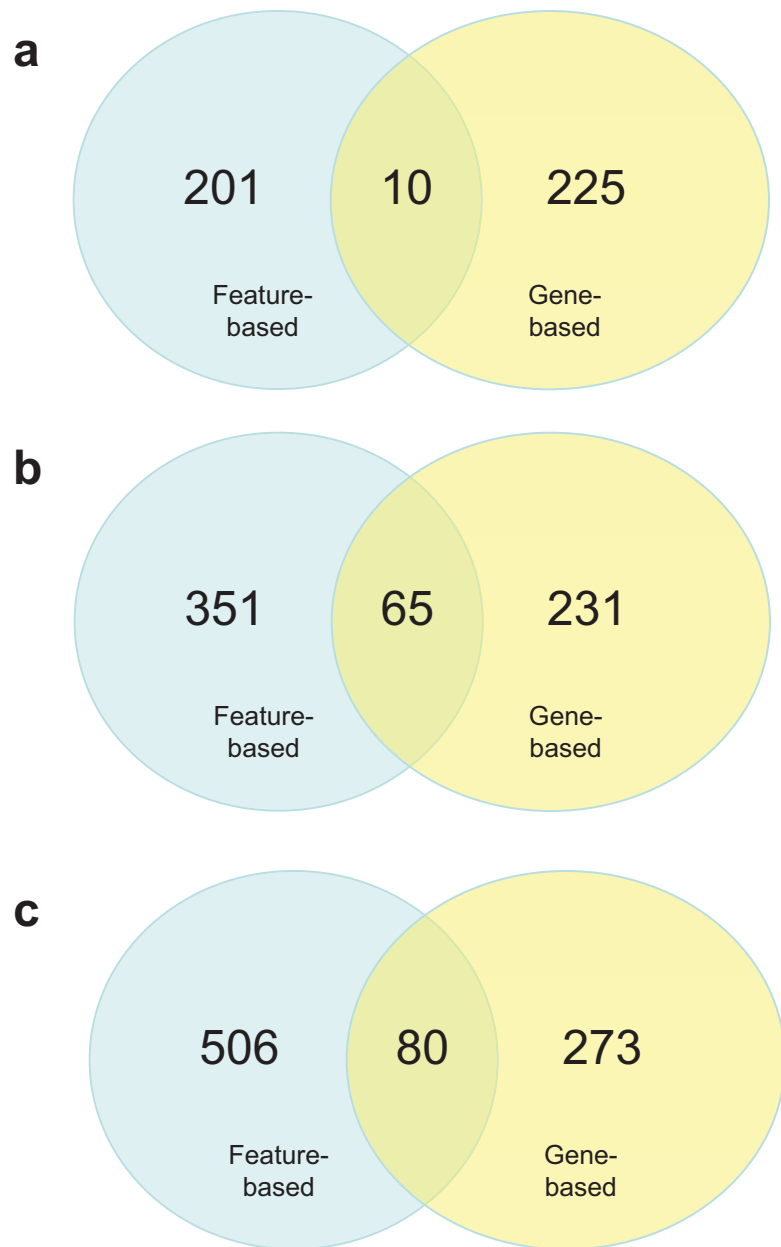
Supplementary Figure S22. Distribution of population summary statistics within and between groups for SBI-08: **(a)** group differentiation (F_{ST}); **(b)** diversity θ_{π} ; **(c)** diversity θ_w ; **(d)** allele frequency spectra (Tajima's D). Groups colour-coded as follows: Improved inbreds (blue), landraces (green), wild and weedy genotypes (red).



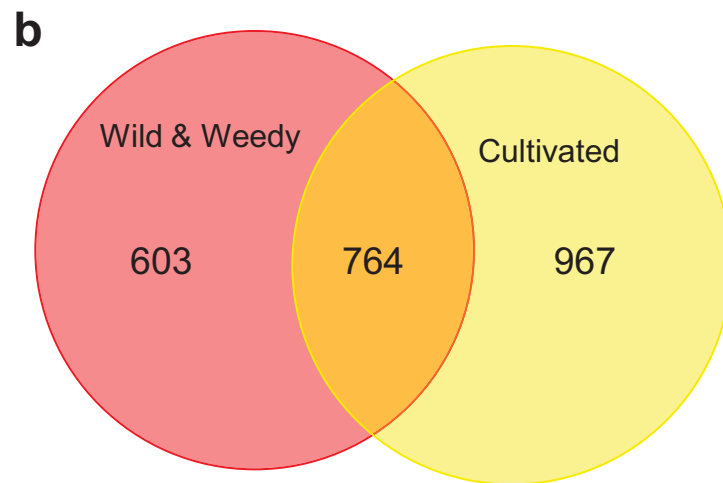
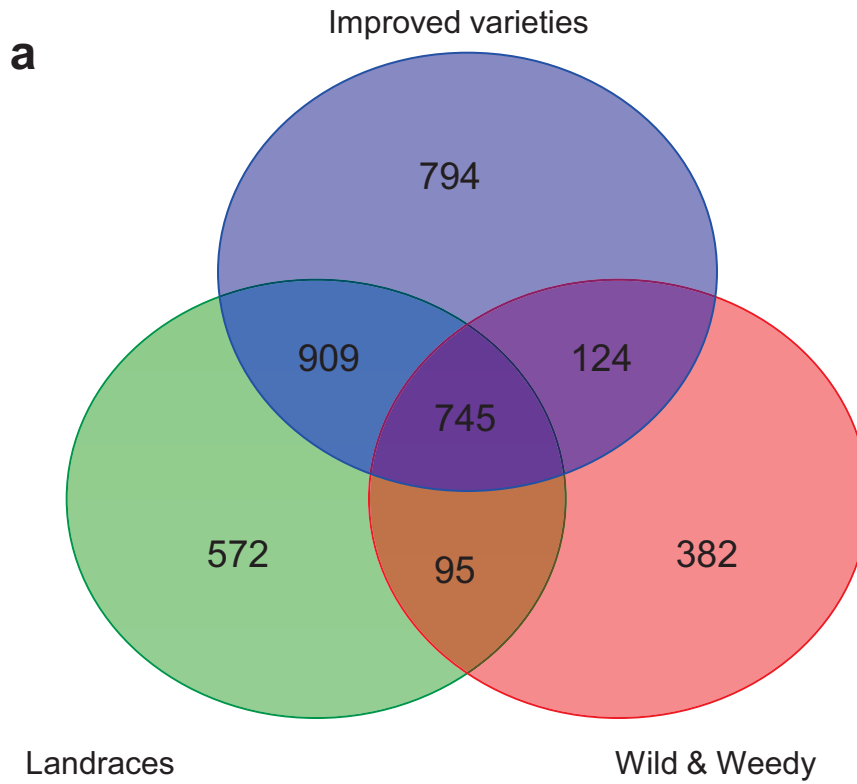
Supplementary Figure S23. Distribution of population summary statistics within and between groups for SBI-09: **(a)** group differentiation (F_{ST}); **(b)** diversity θ_{π} ; **(c)** diversity θ_w ; **(d)** allele frequency spectra (Tajima's D). Groups colour-coded as follows: Improved inbreds (blue), landraces (green), wild and weedy genotypes (red).



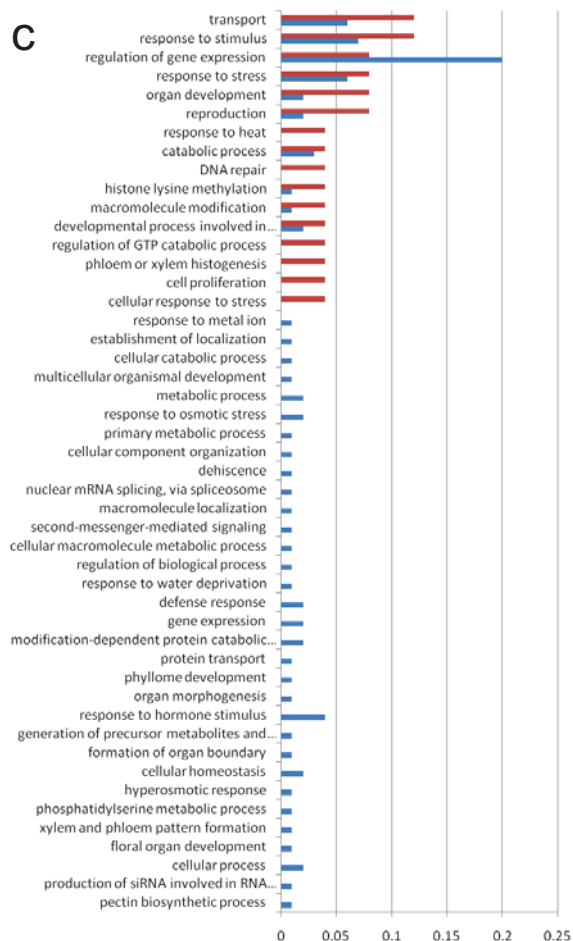
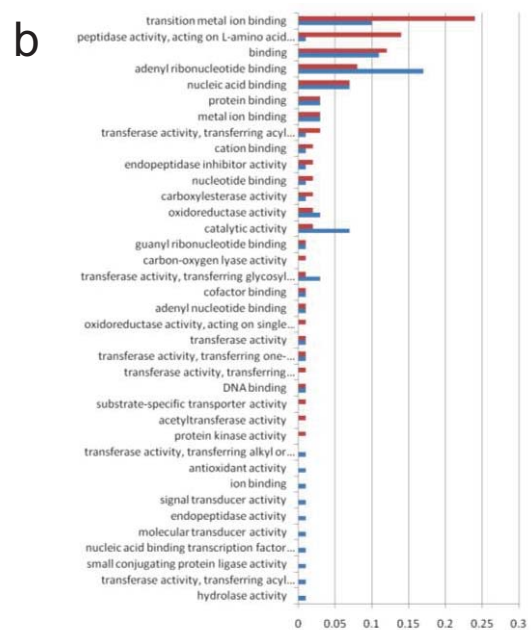
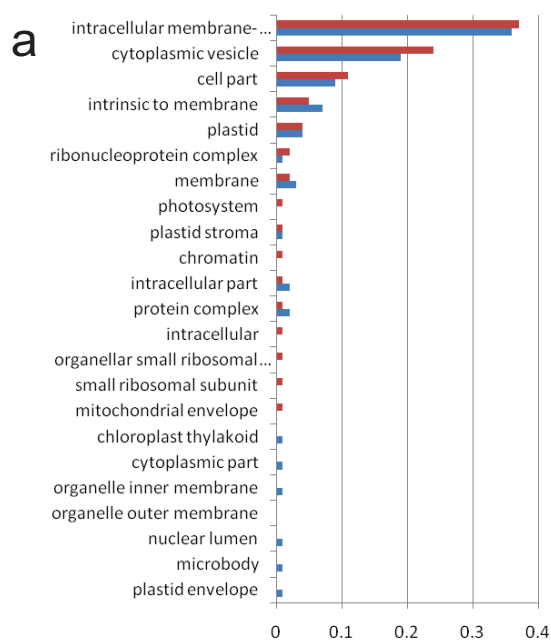
Supplementary Figure S24. Distribution of population summary statistics within and between groups for SBI-10: **(a)** group differentiation (F_{ST}); **(b)** diversity θ_{π} ; **(c)** diversity θ_w ; **(d)** allele frequency spectra (Tajima's D). Groups colour-coded as follows: Improved inbreds (blue), landraces (green), wild and weedy genotypes (red).



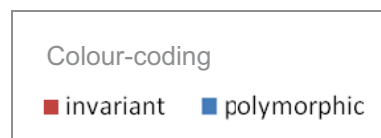
Supplementary Figure S25. Number of candidate genes under selection identified via the 10kb window (feature-based) versus predicted gene model (gene-based) scales for the following three pairwise comparisons: **(a)** Improved inbreds vs landraces; **(b)** Improved inbreds vs wild and weedy genotypes; **(c)** Landraces vs. wild and weedy genotypes

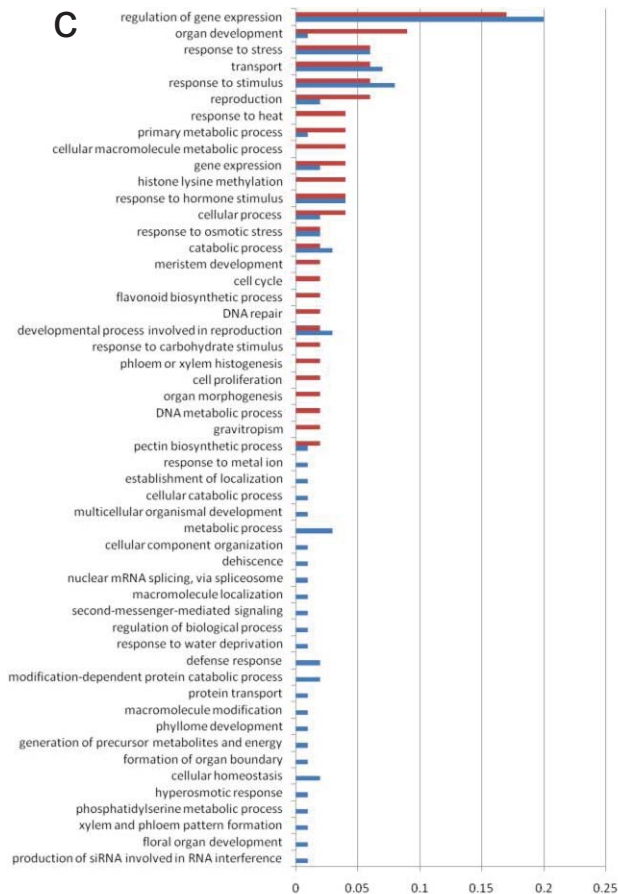
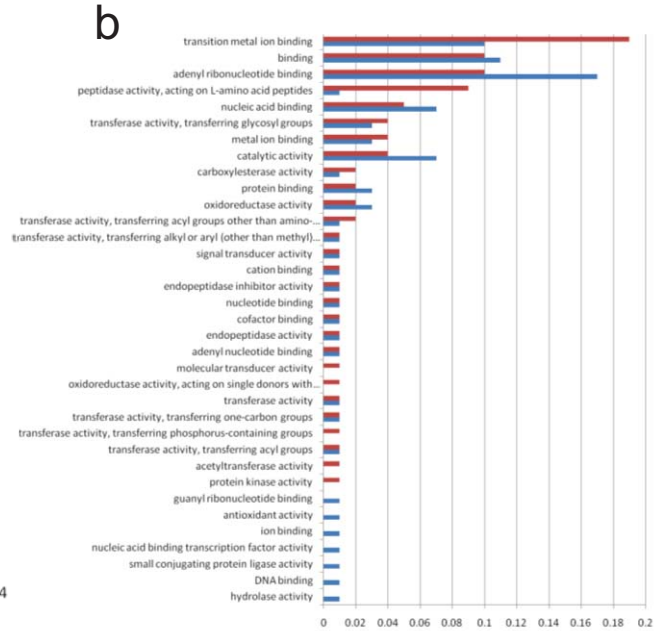
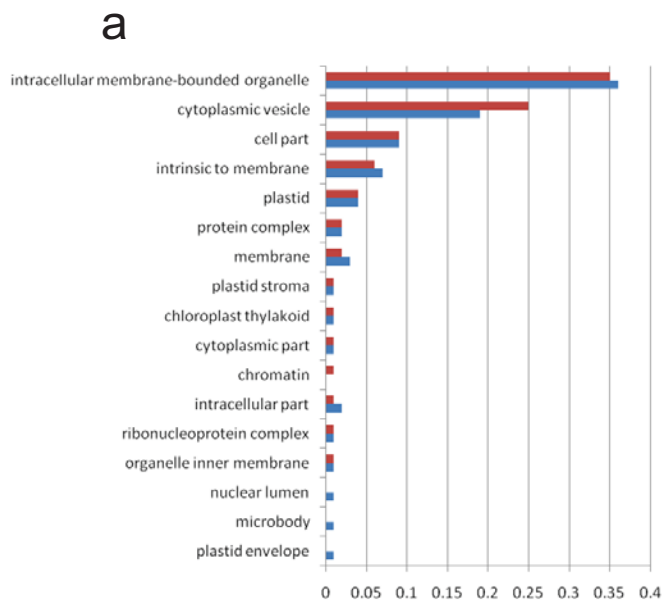


Supplementary Figure S26. Summaries of the number of invariant genes across group (a). Total number of invariant genes within and amongst the improved inbreds, landraces and wild and weedy genotypes; (b) Total number of invariant genes within and amongst the cultivated (improved inbreds and landraces combined) and wild and weedy genotypes.



Supplementary Figure S27. Proportion of invariant genes in GO-SLIM categories⁴⁴ in the wild and weedy genotypes (**a**) Cellular components; (**b**) Molecular function; (**c**) Biological processes

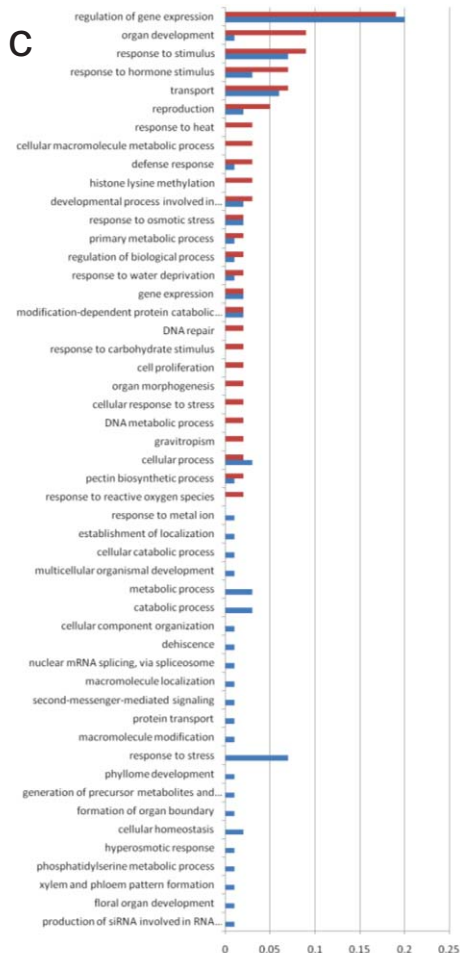
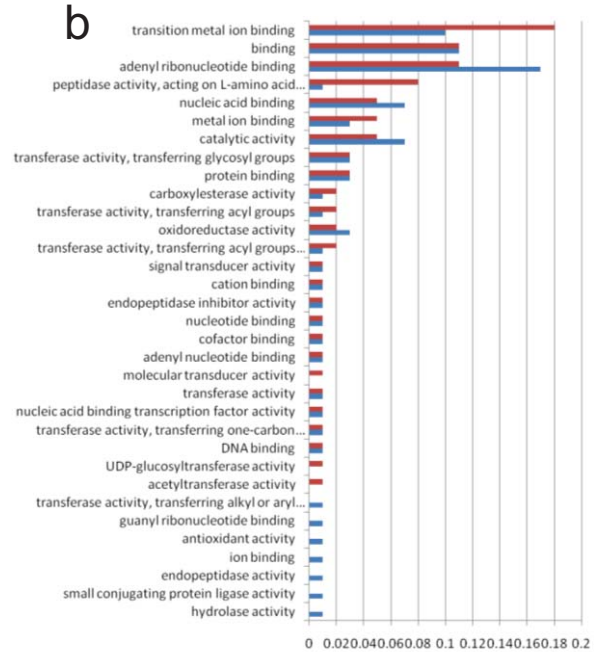
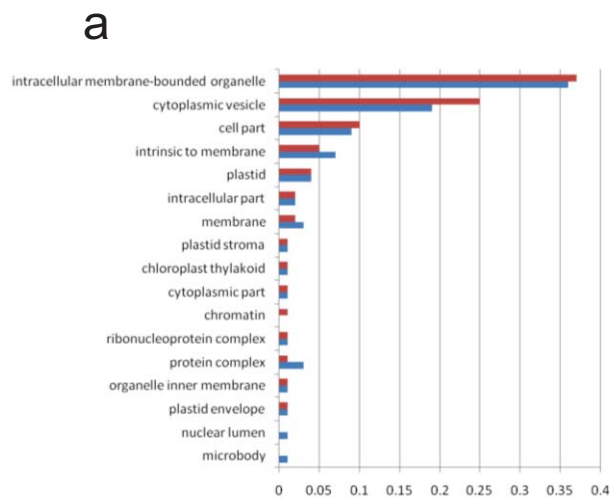




Supplementary Figure S28. Proportion of invariant genes in GO-SLIM categories⁴⁴ in the landrace genotypes **(a)** Cellular components; **(b)** Molecular function; **(c)** Biological processes

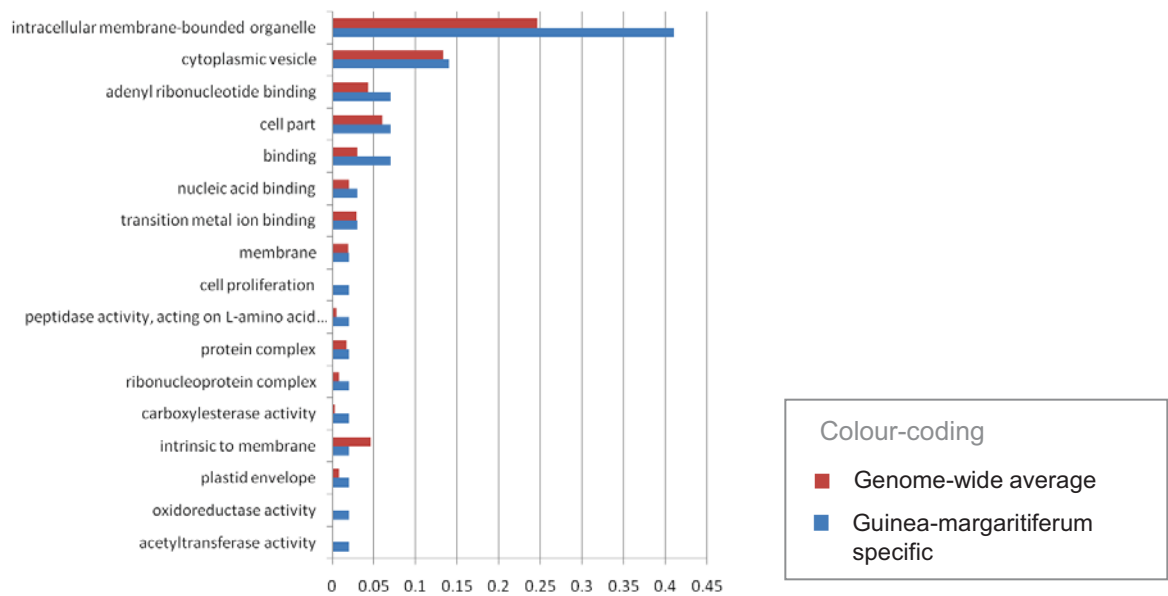
Colour-coding

■ invariant ■ polymorphic

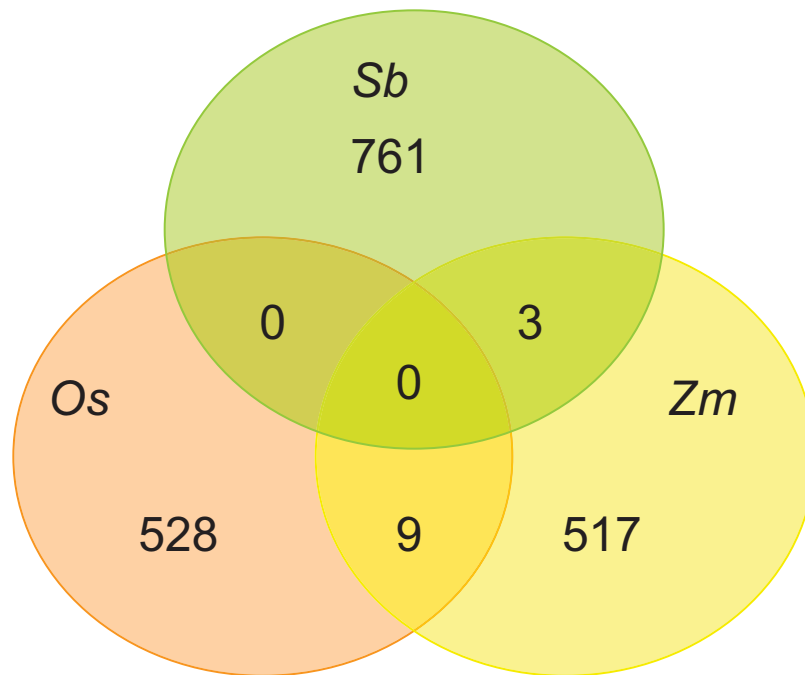


Supplementary Figure S29. Proportion of invariant genes in GO-SLIM categories⁴⁴ in the improved inbred genotypes **(a)** Cellular components; **(b)** Molecular function; **(c)** Biological processes

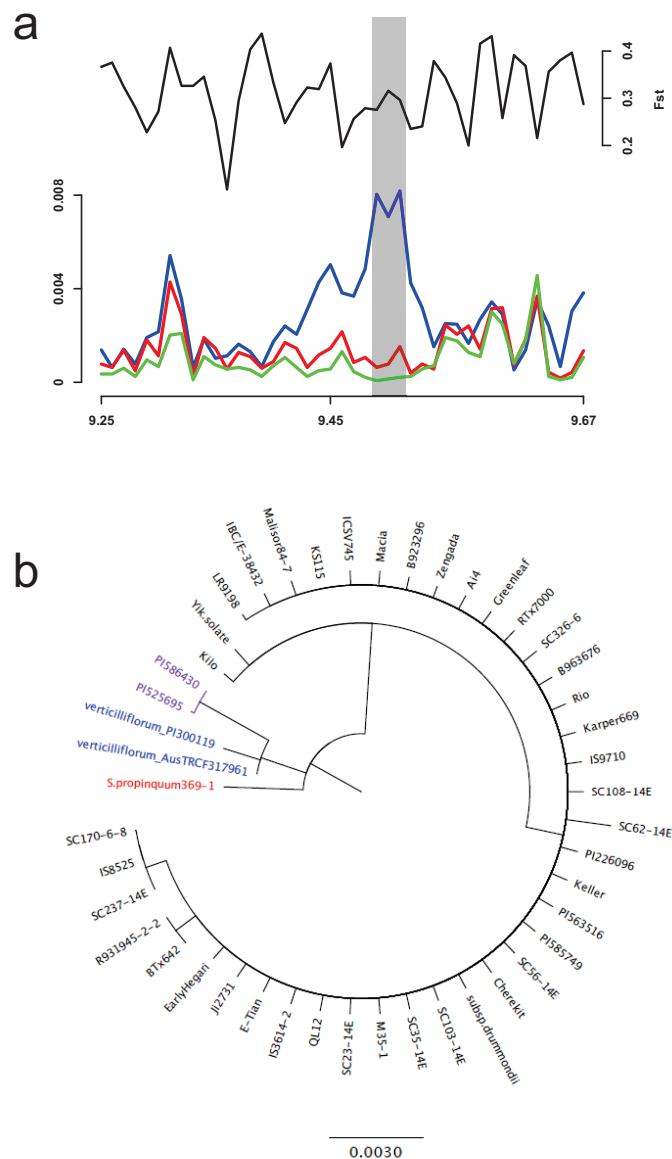




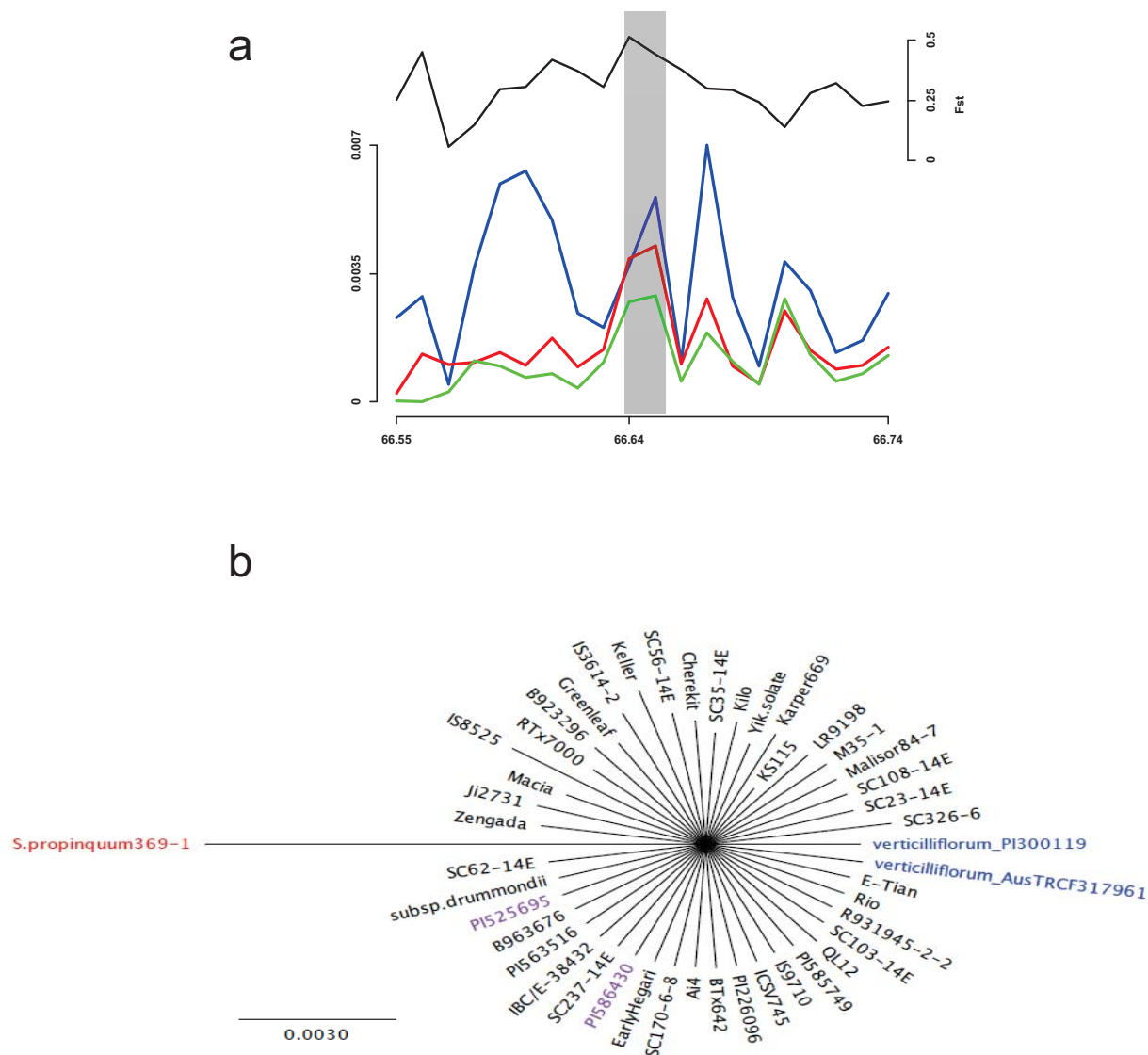
Supplementary Figure S30. Proportion of invariant genes in combined GO-Slim categories⁴⁴ in the guinea-margaritiferum genotypes.



Supplementary Figure S31. A venn diagram comparing the numbers of invariant genes identified within and across species. The species included in this comparison are sorghum (Sb), rice (Os)²⁶ and maize (Zm)^{34,35}.

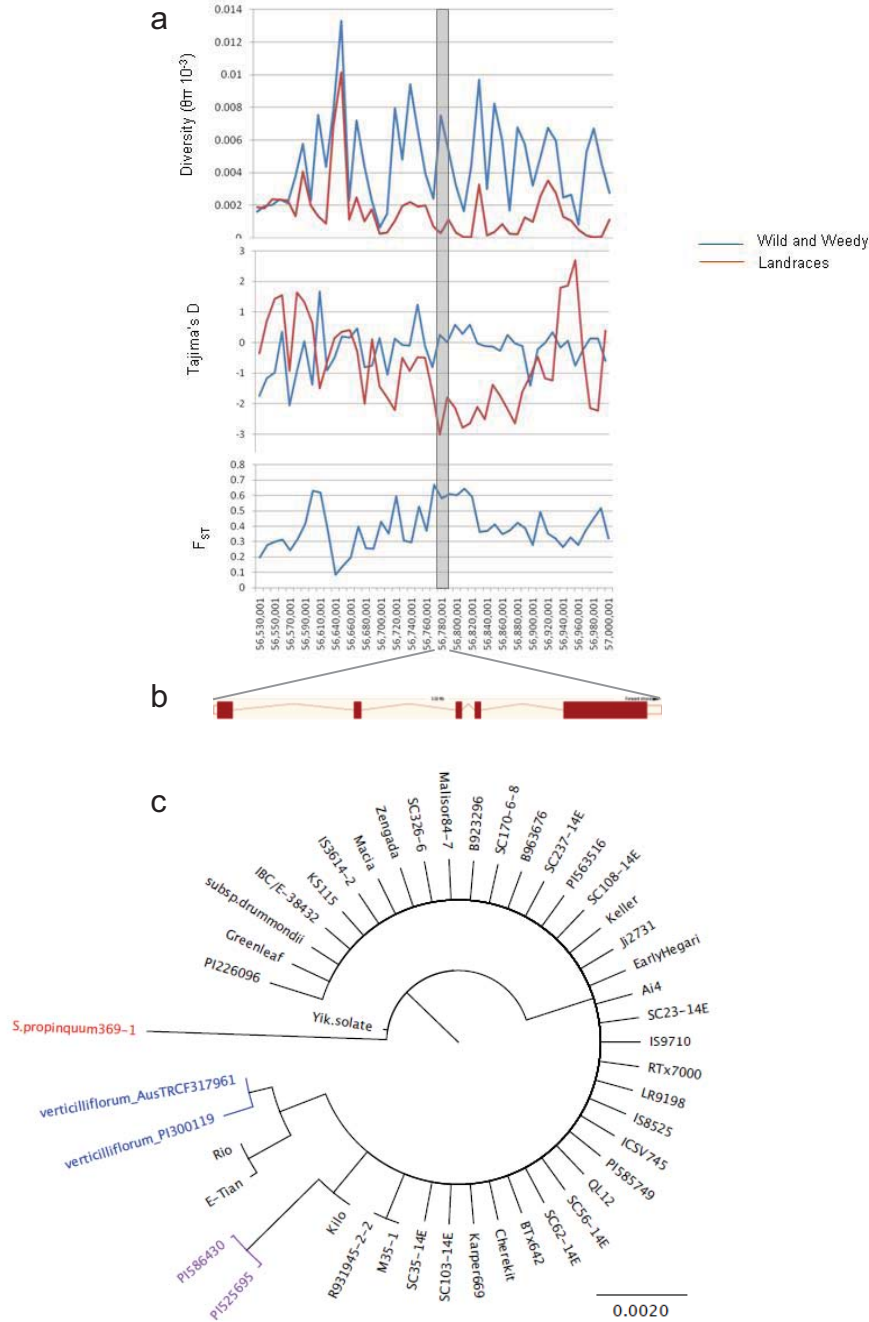


Supplementary Figure S32. Reductions in diversity around the key domestication gene *Tb1* (Sb01g010690) (a) Diversity ($\theta\pi$) (wild and weedy group in red; landraces in blue, improved inbreds in green) and F_{ST} values (black line) in 10kb intervals, with *Tb1* interval highlighted in grey; (b) Consensus gene tree for *Tb1* constructed using the neighbor-joining method with the HKY genetic distance model (Geneious version 5.6.5.); with the two guinea-margaritifera genotypes coloured purple, the two *S. bicolor* subsp. *verticilliflorum* genotypes coloured blue and a single *S. propinquum*, as an outgroup, in red.

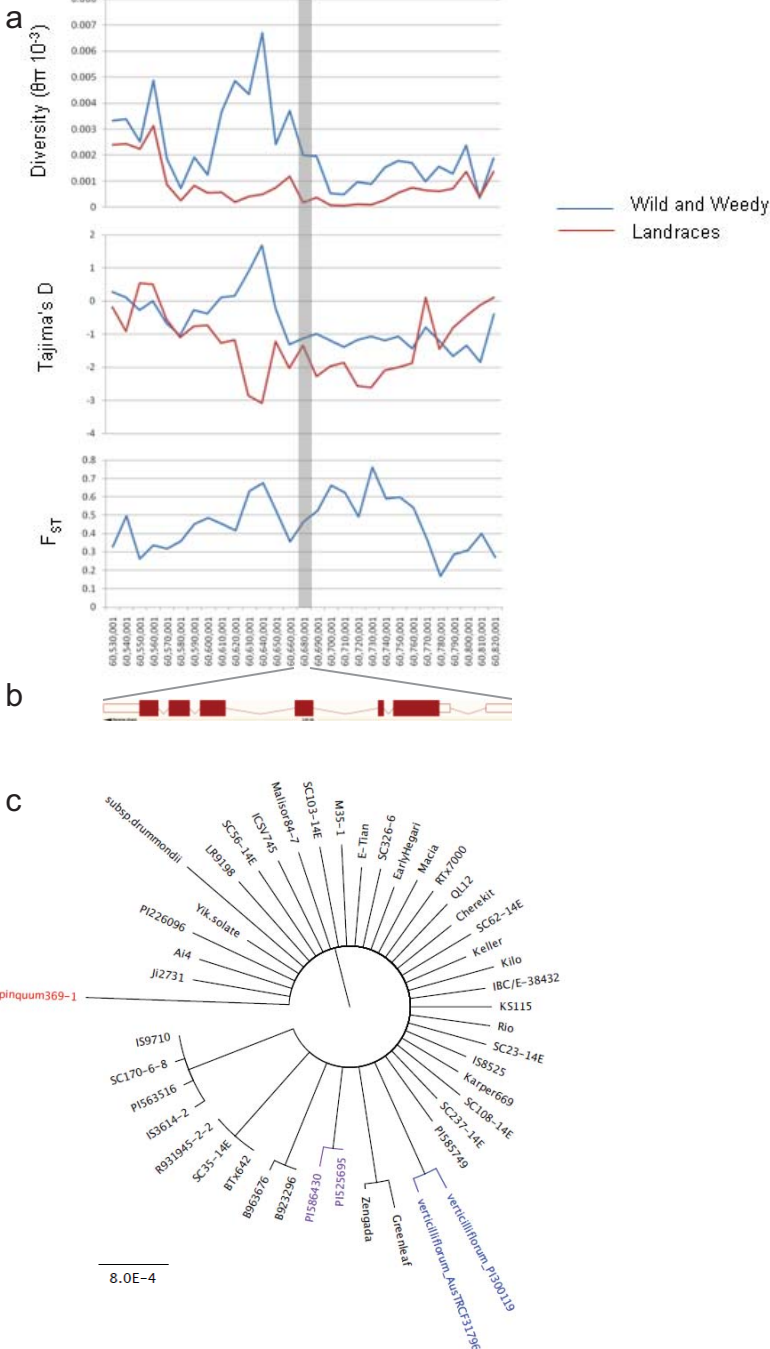


Supplementary Figure S33. Reductions in diversity around the key domestication gene *ba1* (Sb03g038820) (**a**) Diversity ($\theta\pi$) (wild and weedy group in red; landraces in blue, improved inbreds in green) and F_{ST} values (black line) in 10kb intervals; (**b**) Consensus gene tree for *ba1* constructed using the neighbor-joining method with the HKY genetic distance model (Geneious version 5.6.5.); with the two guinea-margaritiflorum genotypes coloured purple, the two *S. bicolor* subsp. *verticilliflorum* genotypes coloured blue and a single *S. propinquum*, as an outgroup, in red.

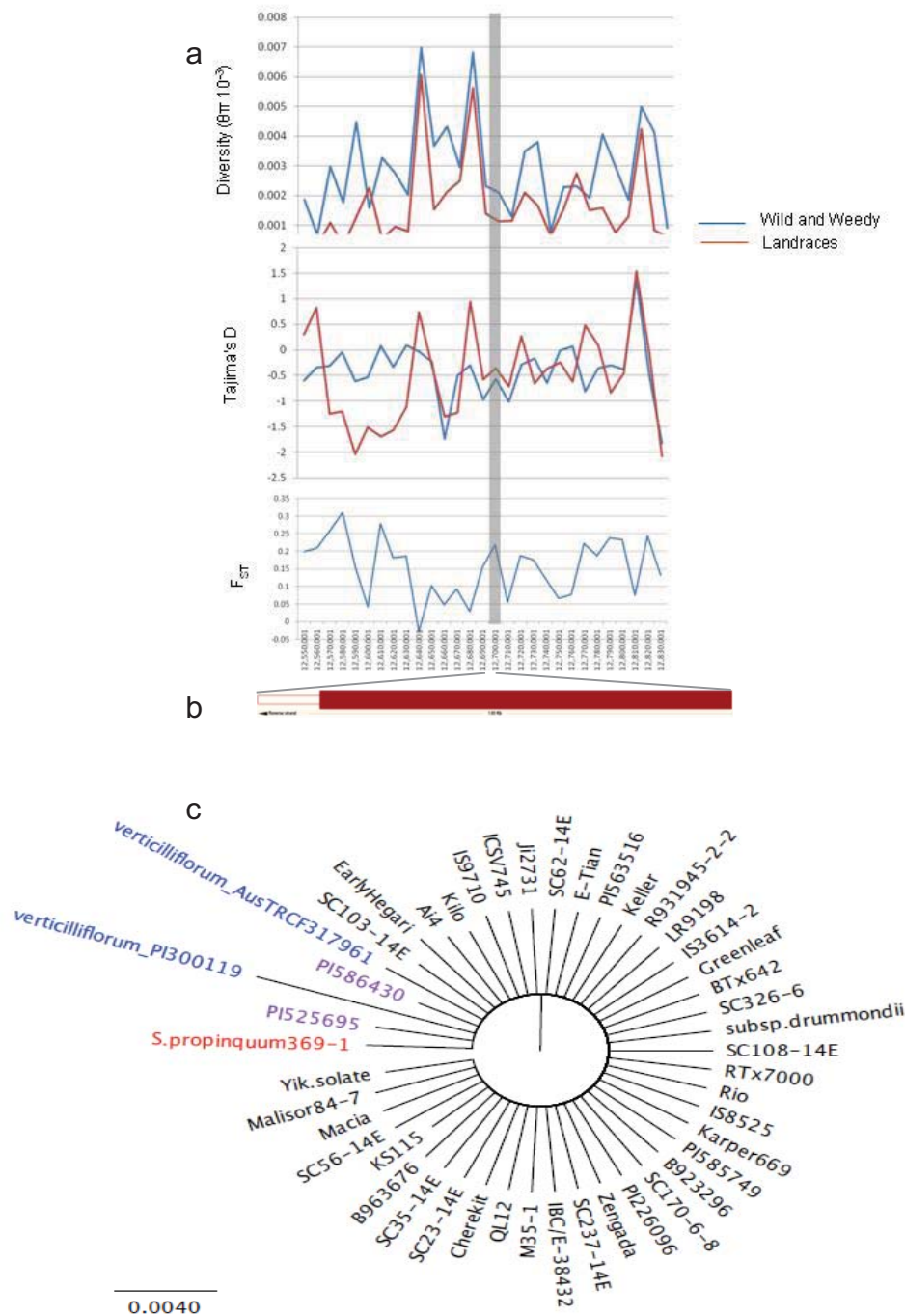
Sb07g022330 (dep1)



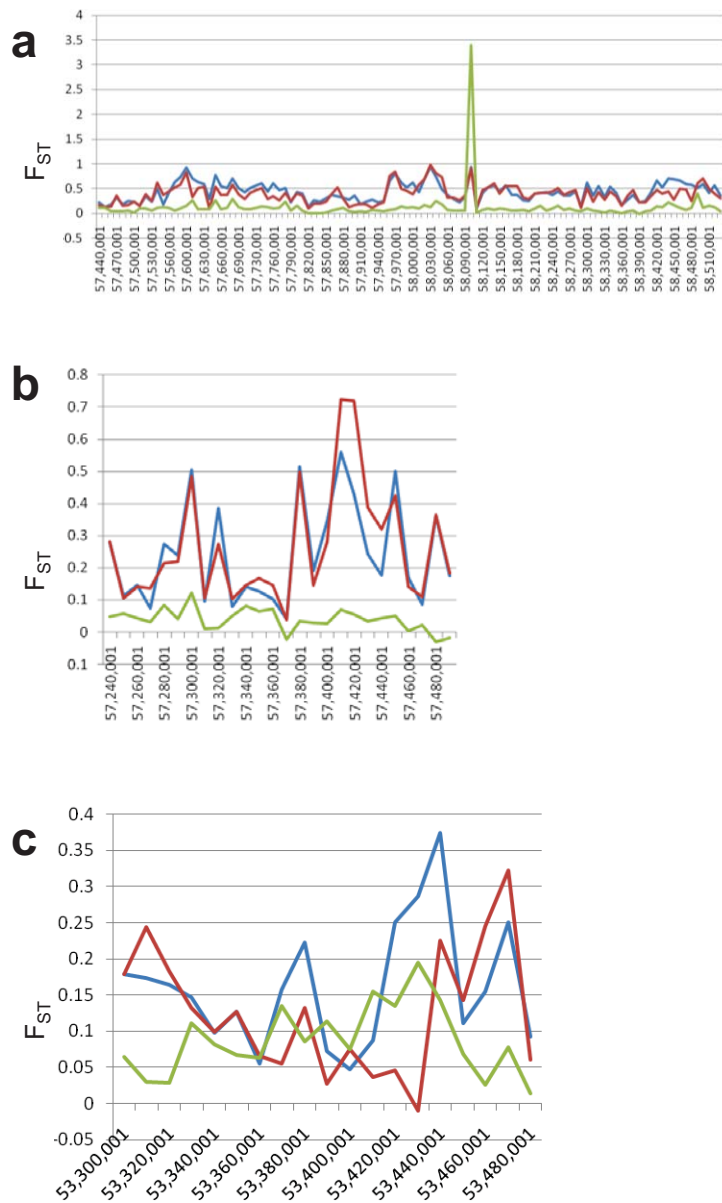
Supplementary Figure S34. Strong signatures of selective sweeps around Sb07g022330 (*dep1*), a gene which enhances grain yield in rice³⁸. **(a)** Diversity ($\theta\pi$), Tajima's D and F_{ST} values per 10kb window; **(b)** Predicted gene model structure; **(c)** Consensus gene tree constructed using the neighbor-joining method with the HKY genetic distance model (Geneious version 5.6.5.); with the two guinea-margaritiferum genotypes coloured purple, the two *S. bicolor* subsp. *verticilliflorum* genotypes coloured blue and a single *S. propinquum*, as an outgroup, in red.

Sb10g031020 (Psy1)

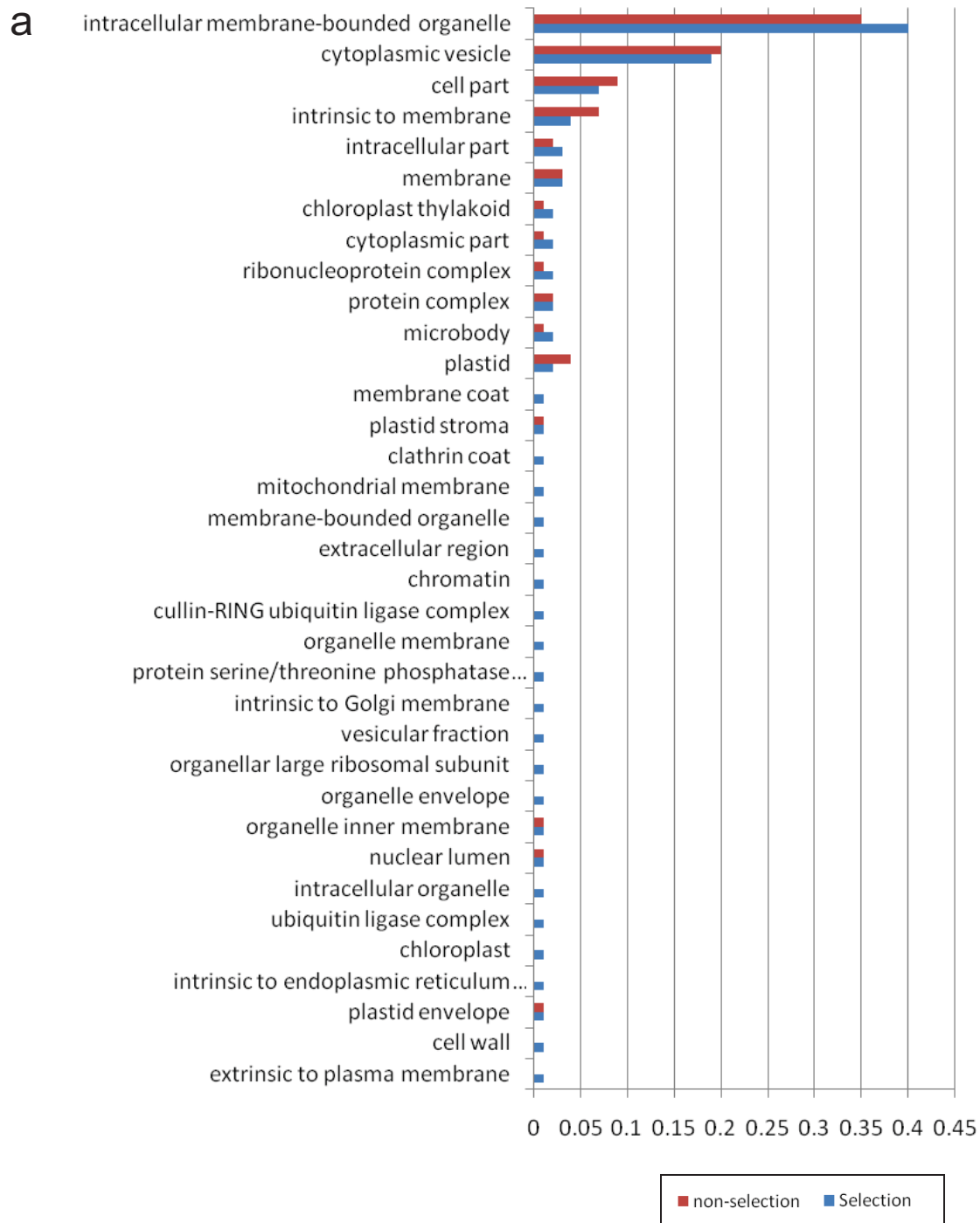
Supplementary Figure S35. Strong signatures of selective sweeps around Sb10g031020 (*Psy1*), a gene controlling grain colour⁴⁰; **(a)** Diversity ($\theta\pi$). Tajima's D and F_{ST} values per 10kb window; **(b)** Predicted gene model structure; **(c)** Consensus gene tree constructed using the neighbor-joining method with the HKY genetic distance model (Geneious version 5.6.5.); with the two guinea-margaritifera genotypes coloured purple, the two *S. bicolor* subsp. *verticilliflorum* genotypes coloured blue and a single *S. propinquum*, as an outgroup, in red.



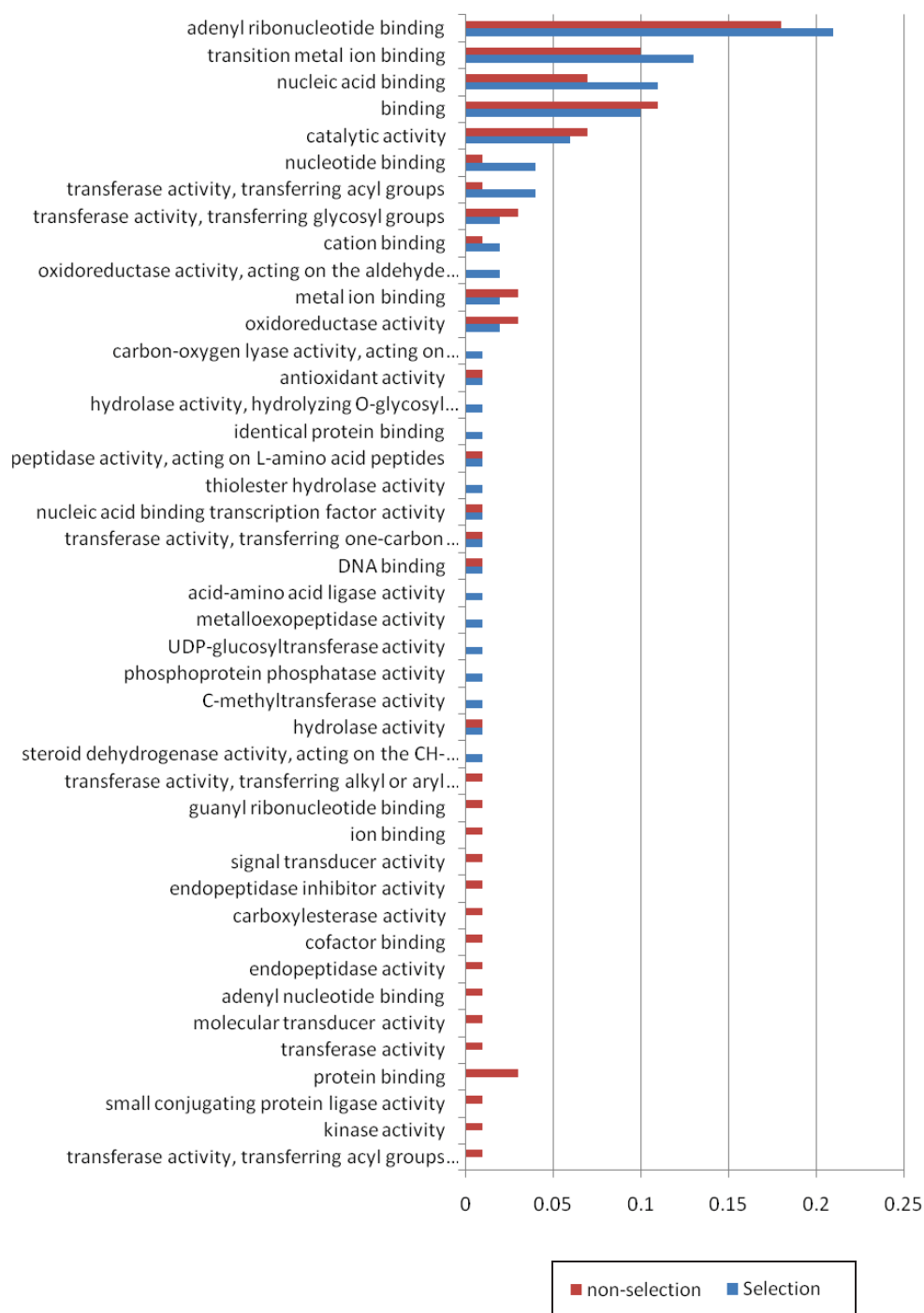
Supplementary Figure S36. Strong signatures of selective sweeps around Sb01g013640 (*Bif2*; Barren Inflorescence2⁵⁹); **(a)** Diversity ($\theta\pi$), Tajima's D and F_{ST} values per 10kb window; **(b)** Predicted gene model structure, **(c)** Consensus gene tree constructed using the neighbor-joining method with the HKY genetic distance model (Geneious version 5.6.5.); with the two guinea-margaritiferum genotypes coloured purple, the two *S. bicolor* subsp. *verticilliflorum* genotypes coloured blue and a single *S. propinquum*, as an outgroup, in red.



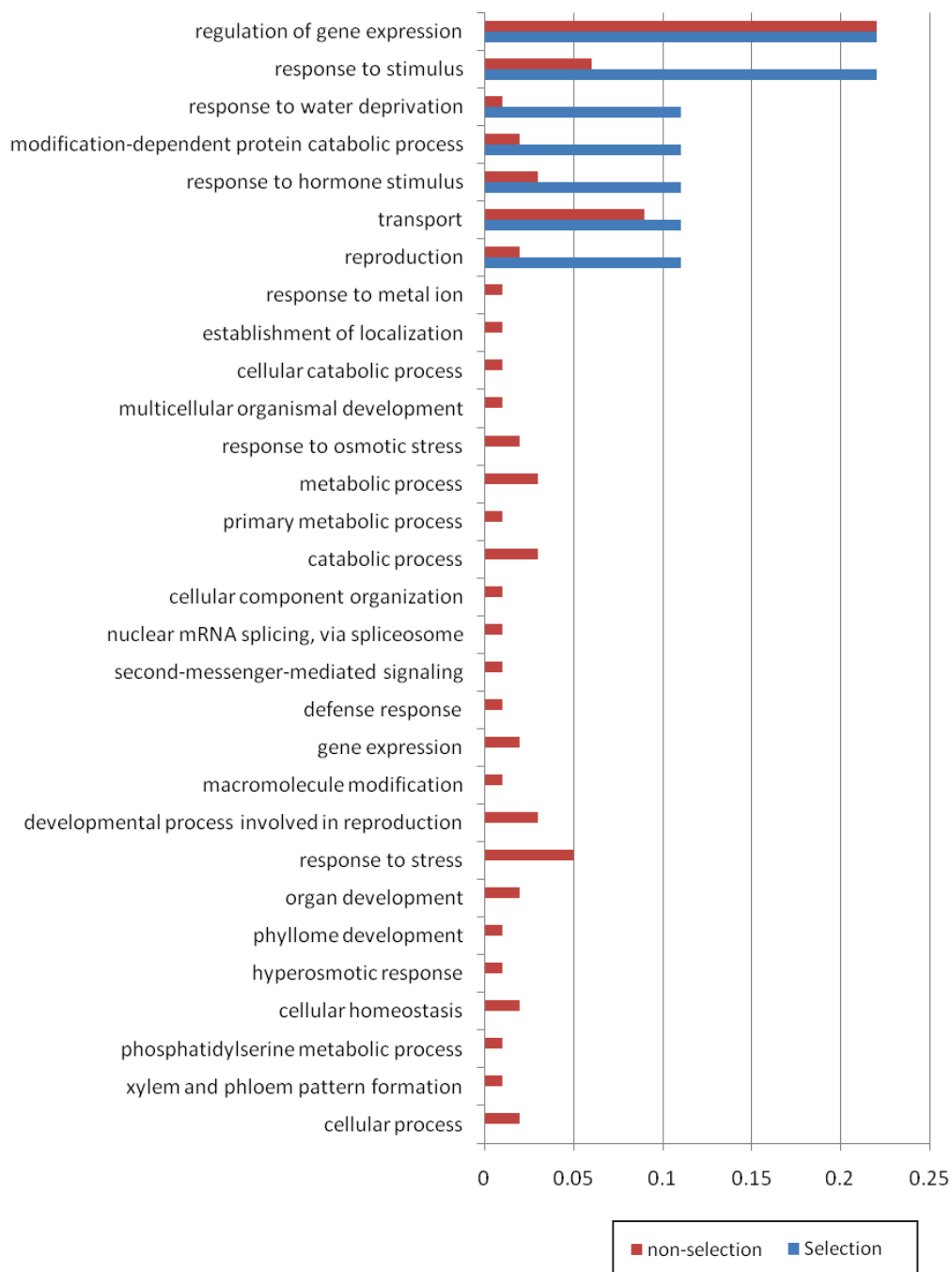
Supplementary Figure S37. F_{ST} signatures across three major effect genes⁴¹ under selection: **(a)** *I* (pericarp colour) on SBI-07; **(b)** *R* (grain colour) on SBI-03; **(c)** *Rs1* (seedling colour) on SBI-06. Blue line indicates F_{ST} values between the wild and weedy and landrace groups; red line indicates F_{ST} values between the wild and weedy and improved inbred groups; green line indicates F_{ST} values between landraces and improved inbred groups.



b

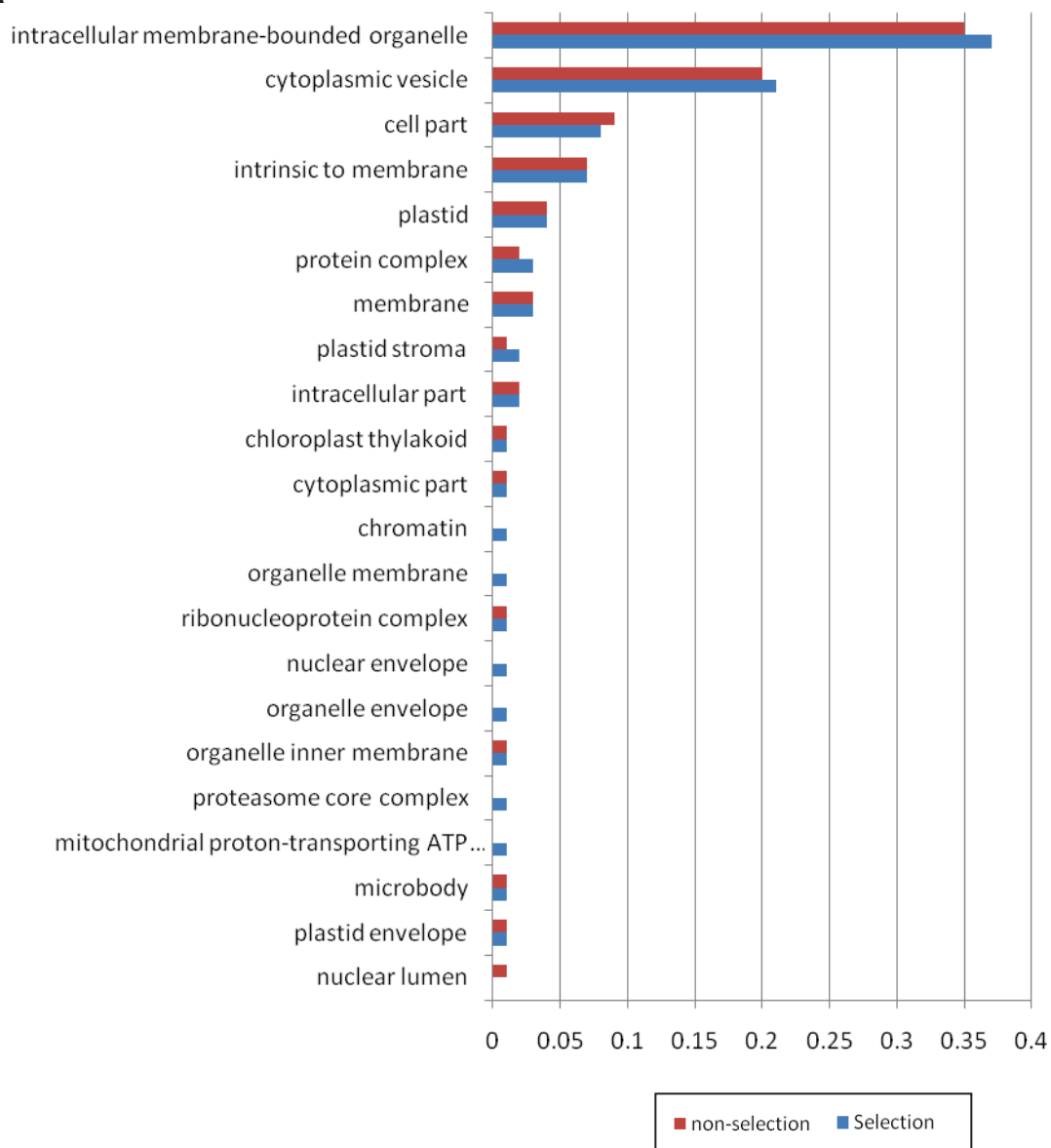


C



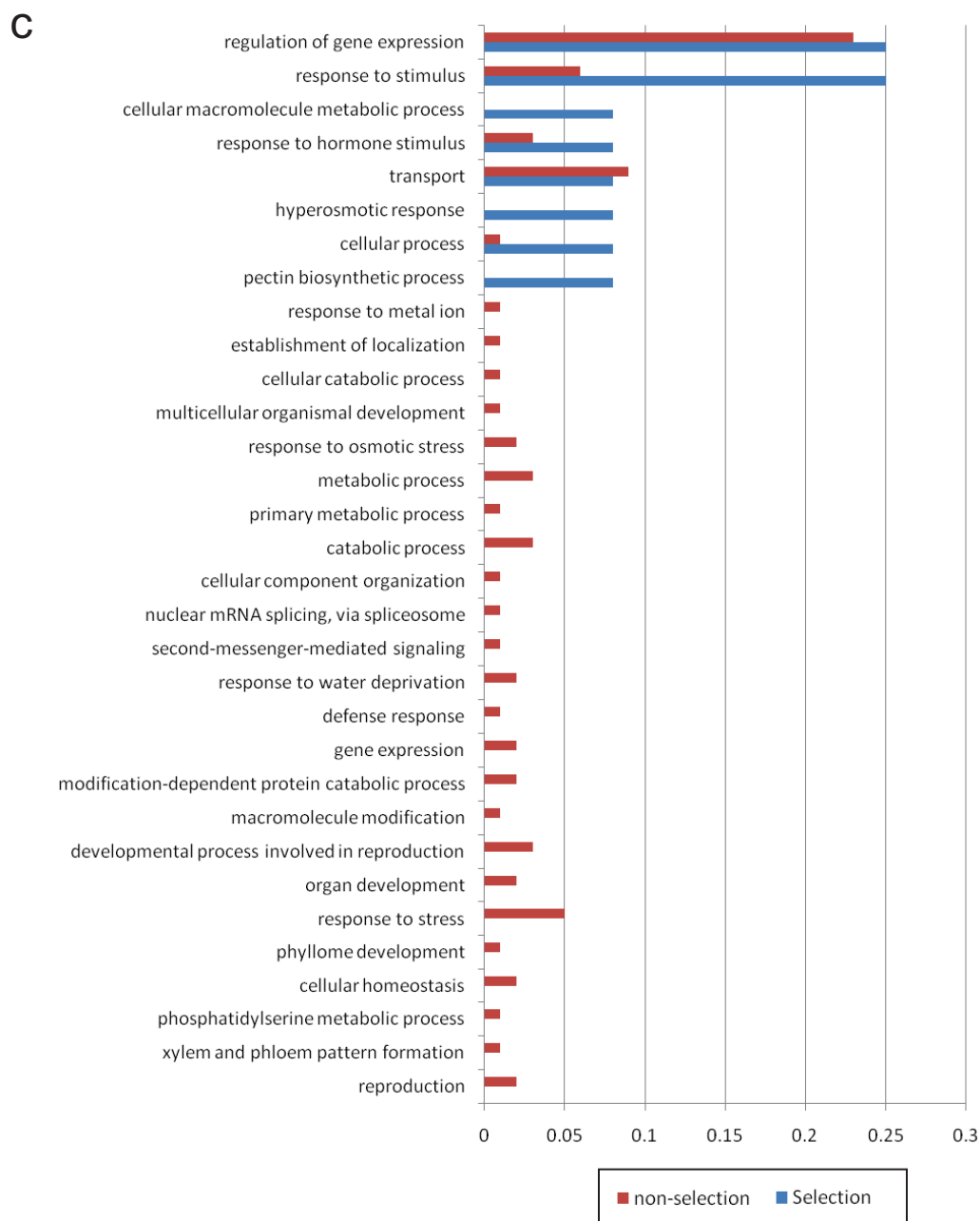
Supplementary Figure S38. Proportion of candidate improvement genes under selection in GO-SLIM categories⁴⁴ relative to non-candidate genes; (a) Cellular Components; (b) Molecular Function; (c) Biological Processes.

a



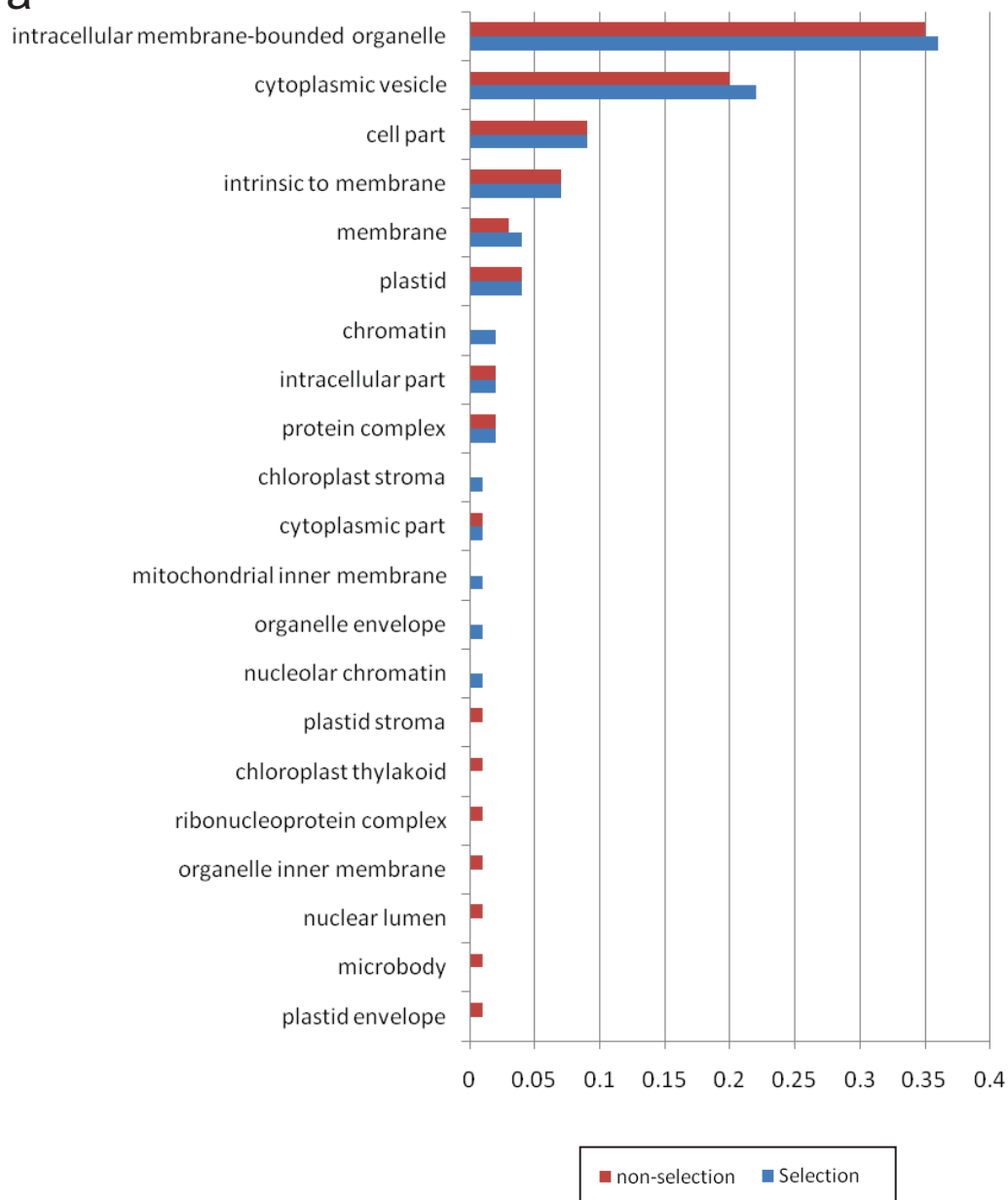
b





Supplementary Figure S39. Proportion of candidate domestication genes under selection in GO-SLIM categories⁴⁴ relative to non-candidate genes; (a) Cellular Components; (b) Molecular Function; (c) Biological processes

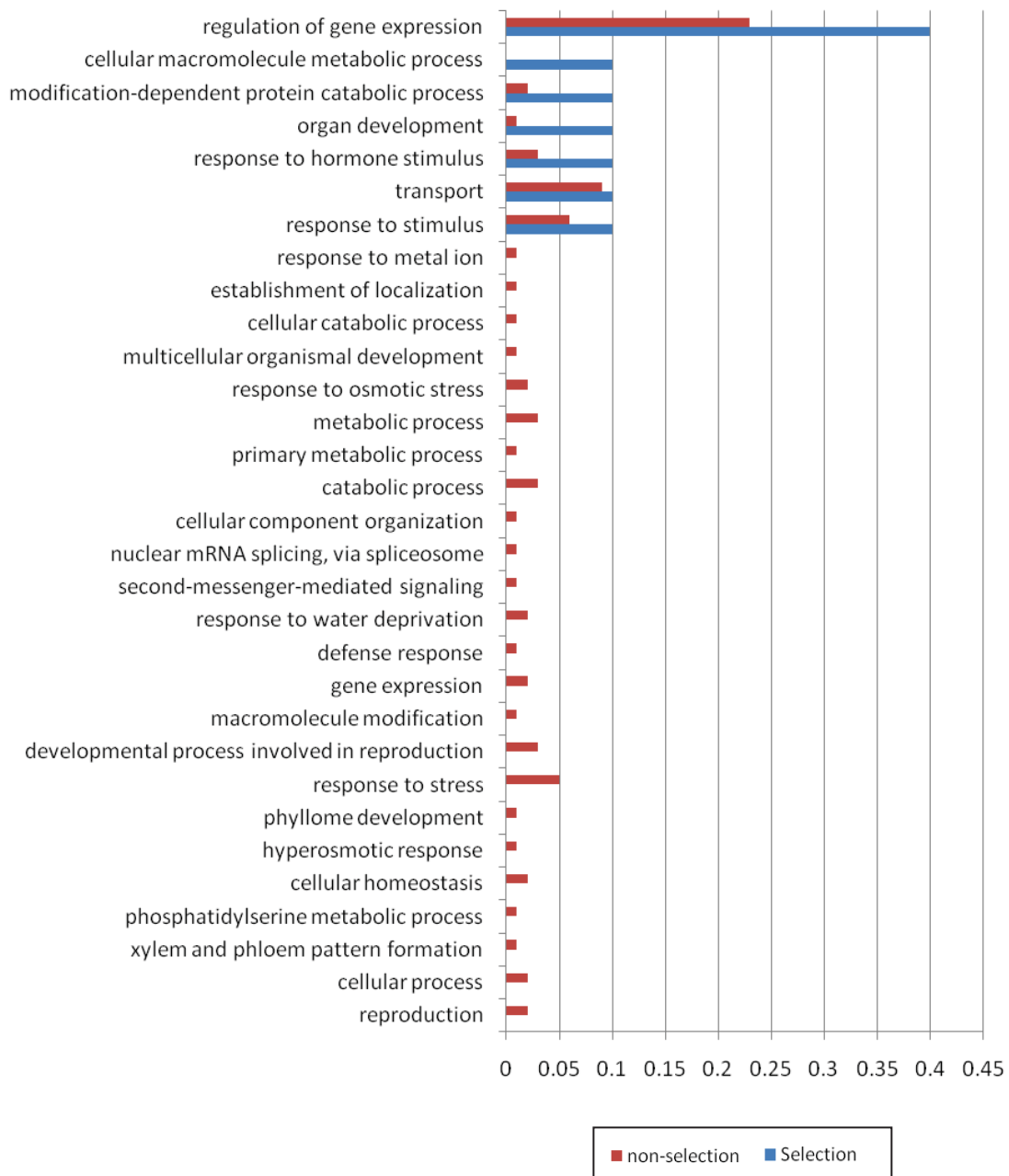
a



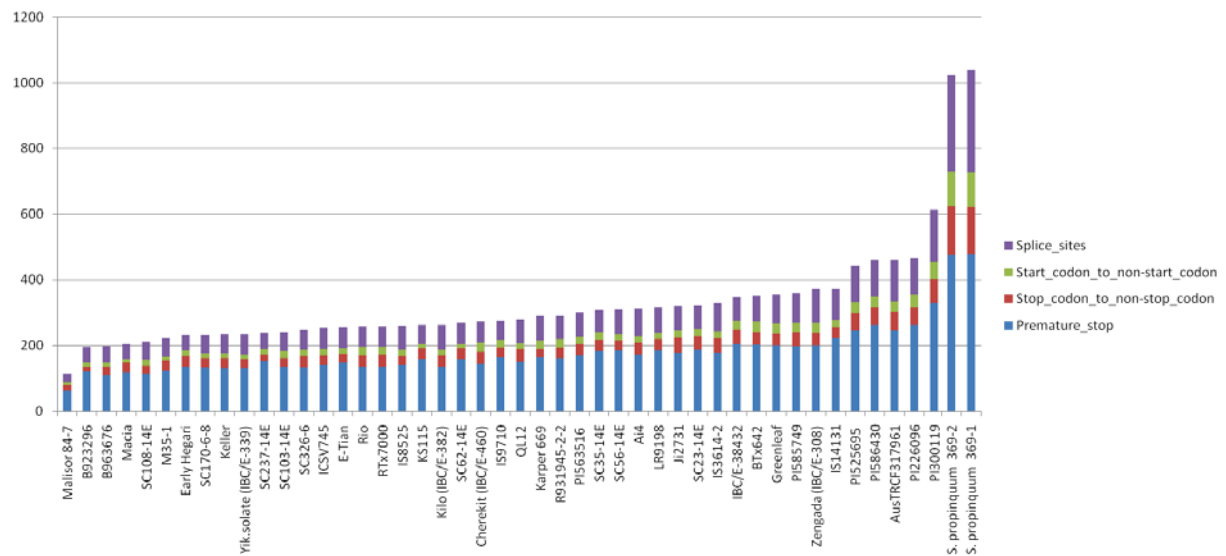
b



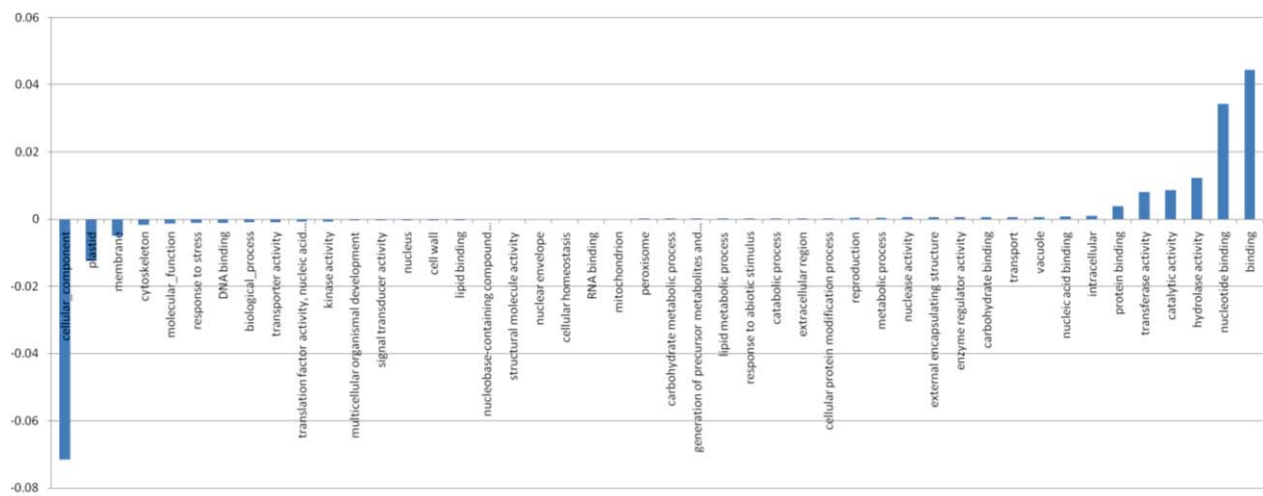
C



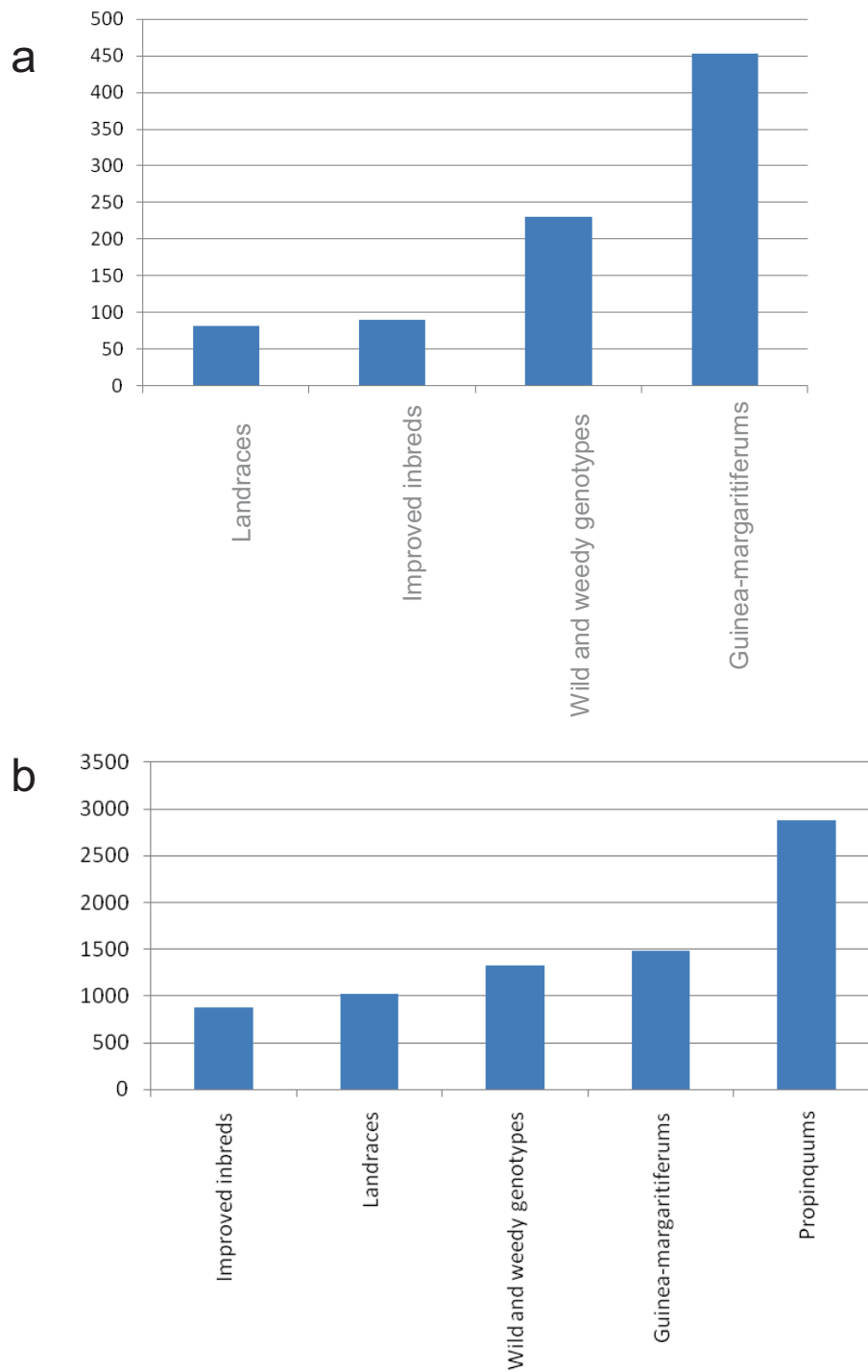
Supplementary Figure S40. Proportion of candidate domestication and improvement genes under selection in GO-SLIM categories⁴⁴ relative to non-candidate genes; (a) Cellular components; (b) Molecular function; (c) Biological processes



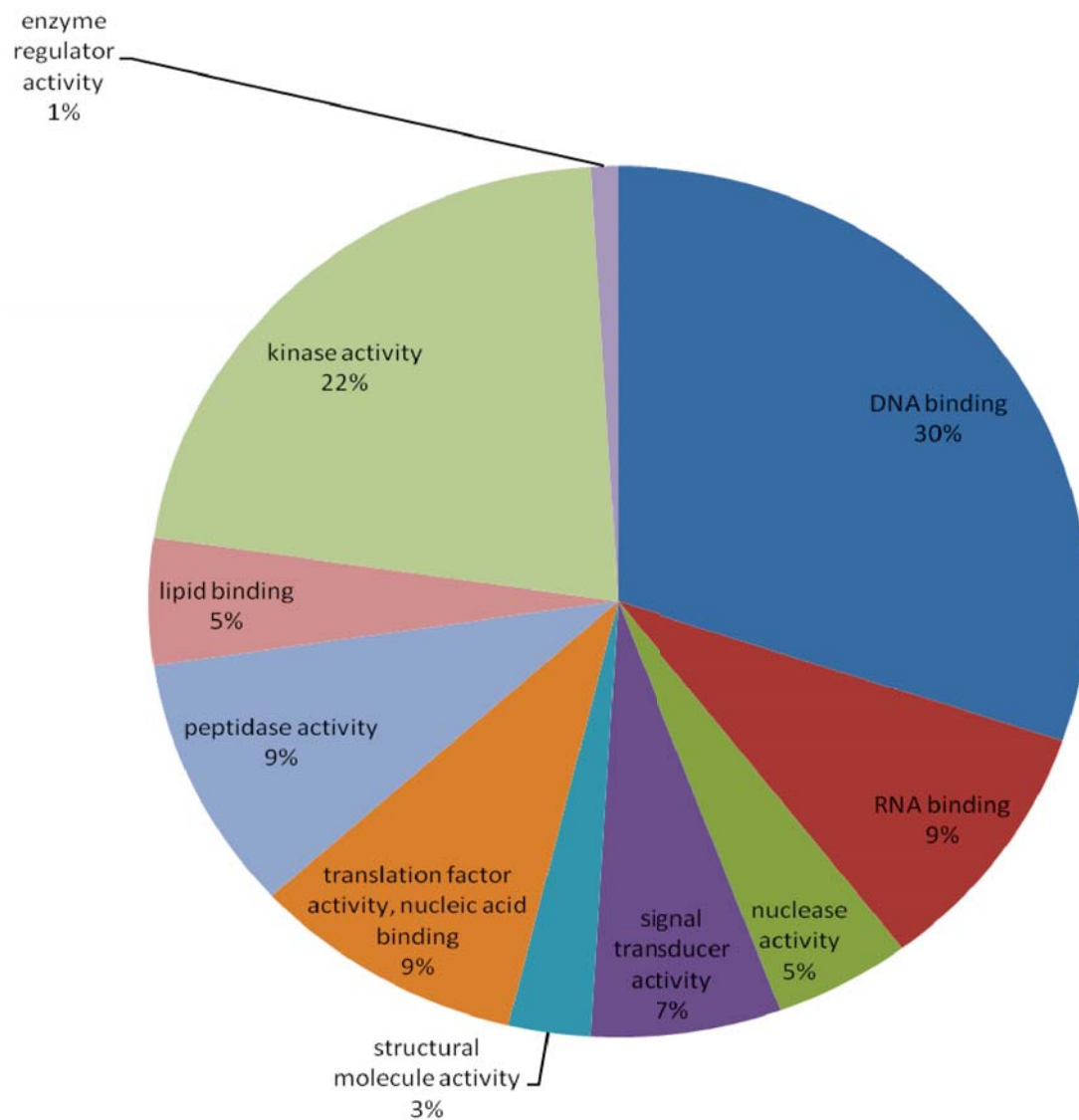
Supplementary Figure S41. Total number and proportion of large effect (LE) SNP types across genotypes. Four LE SNP types are detailed; splice sites (purple); start codon to non-start codon (green); stop codon to non-stop codon (red); premature stop (blue).



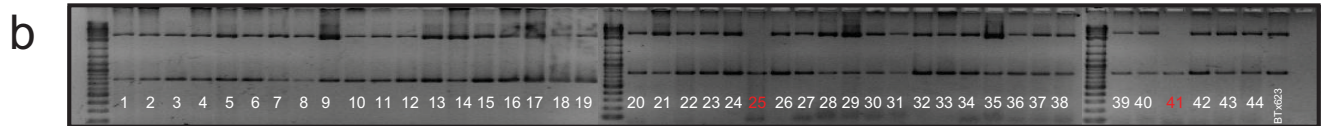
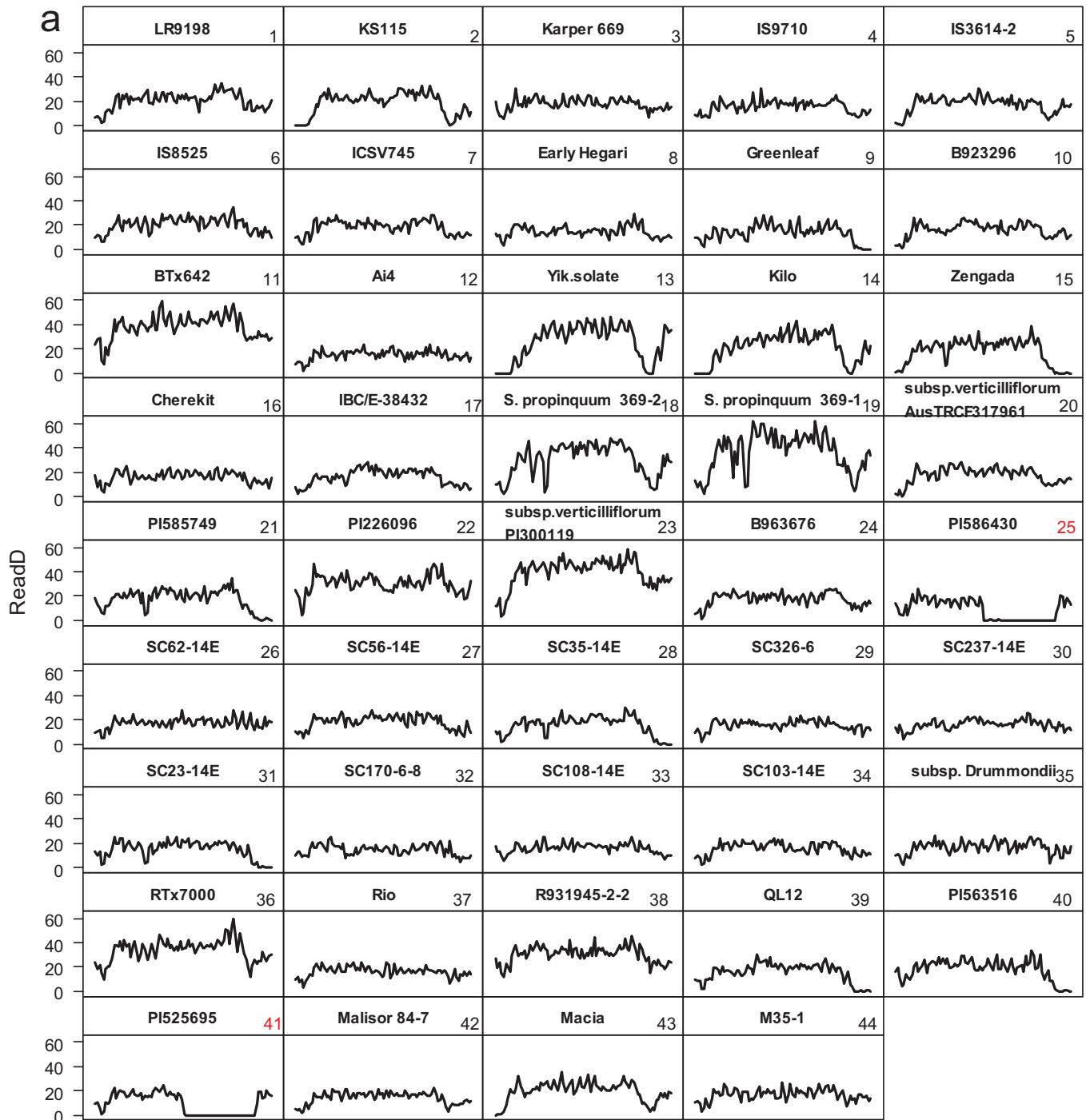
Supplementary Figure S42. Difference in frequency in major GO-SLIM types⁴⁴ between genes containing frame-shifting indels and their genome-wide average. A positive number indicates a higher proportion in the genes with frame-shifting indels, whereas a negative number indicates a lower proportion in the genes with frame-shifting indels compared to the genome-wide average.



Supplementary Figure S43. Summaries of large effect SNPs and indels across genotype groups; **(a)** Average number of LE-SNPs per genotype per group; **(b)** Average number of frame-shifting indels per genotype per group

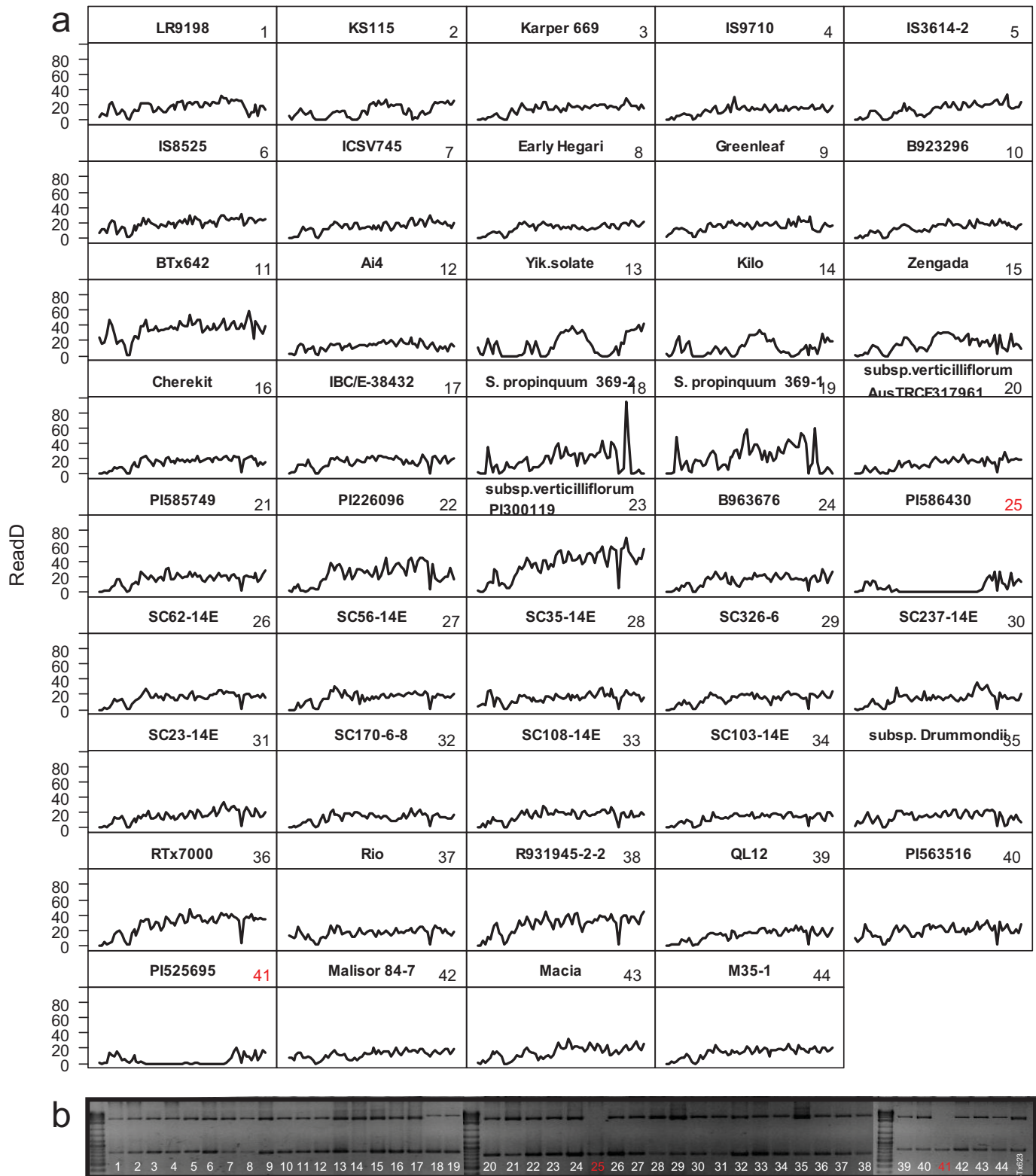


Supplementary Figure S44. GO-SLIM function analysis⁴⁴: Molecular function of LE-SNPs



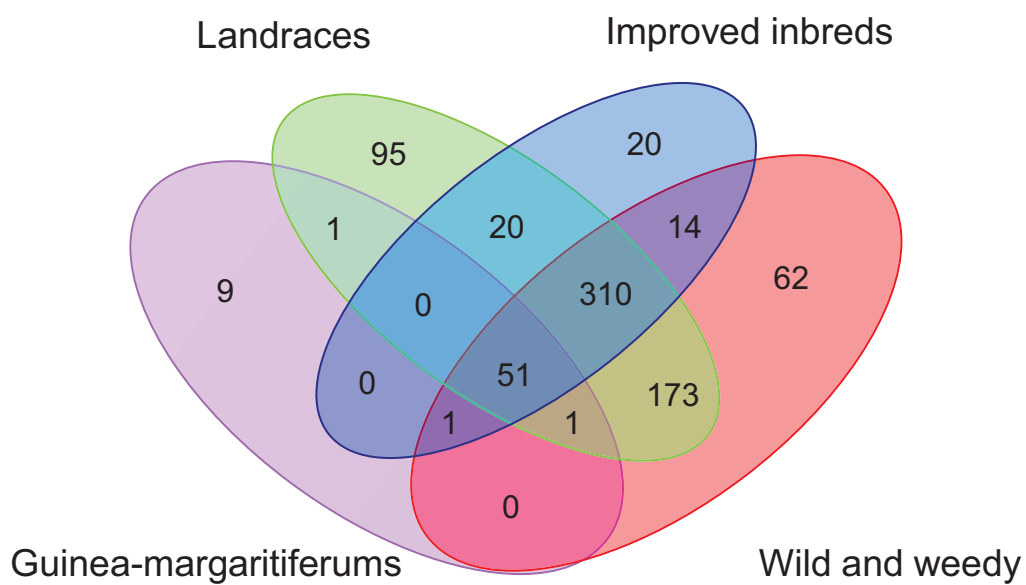
Supplementary Figure S45. Validation of gene loss events for **Sb09g018590**. (a) 100bp read depth plots within gene and 1kb 5' and 3' across 44 genotypes. (b) PCR amplification of genotypes, in same order as plots, with gene loss events coloured in red*.

*Gel image shows results of multiplex PCR; DNA ladder (2-Log DNA Ladder, 0.1–10.0 kb); upper amplification product Sb03g0145670 (product size 3933bp); lower amplification product gamma-kafirin, Sb02g025510 (product size 791bp).

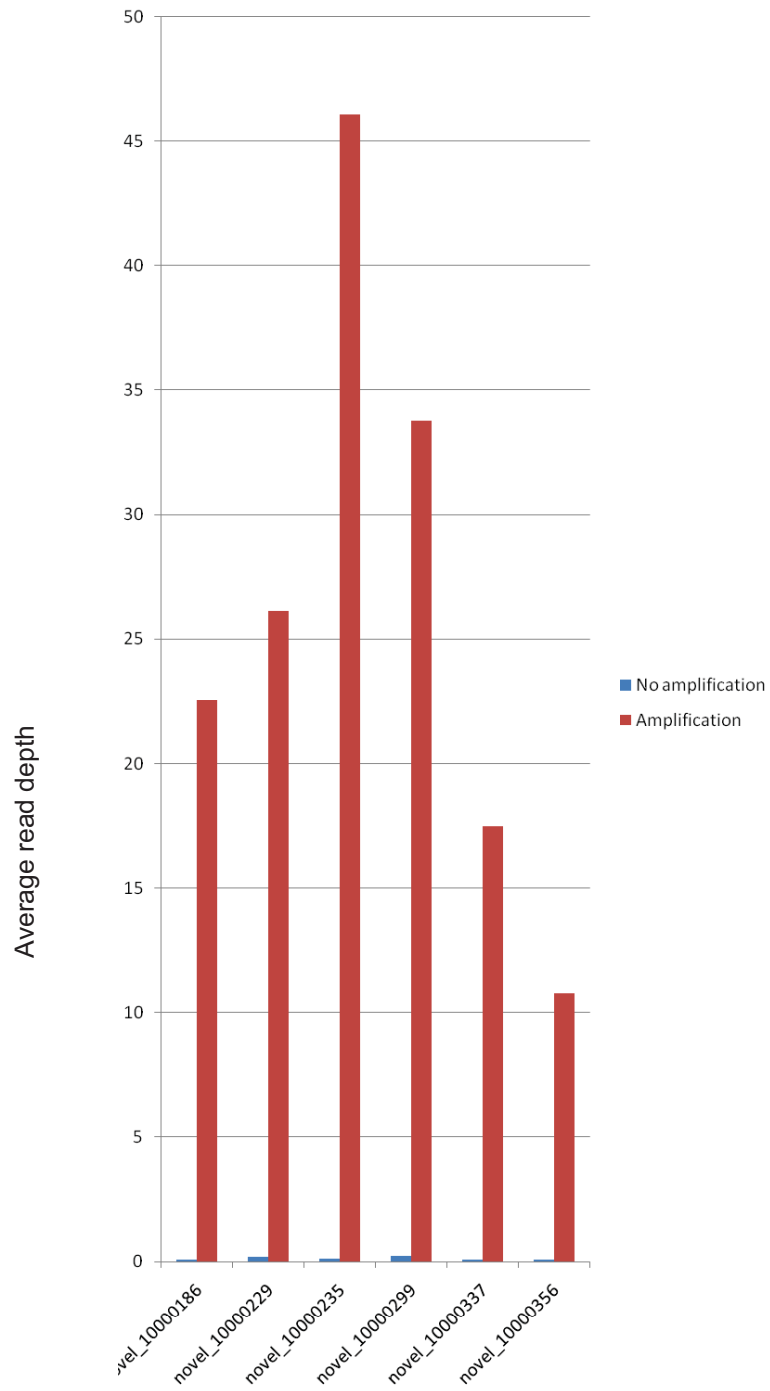


Supplementary Figure S46. Validation of gene loss events for **Sb03g045670**. **(a)** 100bp read depth plots within gene and 1kb 5' and 3' across 44 genotypes. **(b)** PCR amplification of genotypes, in same order as plots, with gene loss events coloured in red*.

*Gel image shows results of multiplex PCR; DNA ladder (2-Log DNA Ladder, 0.1–10.0 kb); upper amplification product Sb03g0145670 (product size 3933bp); lower amplification product gamma-kafirin, Sb02g025510 (product size 791bp).

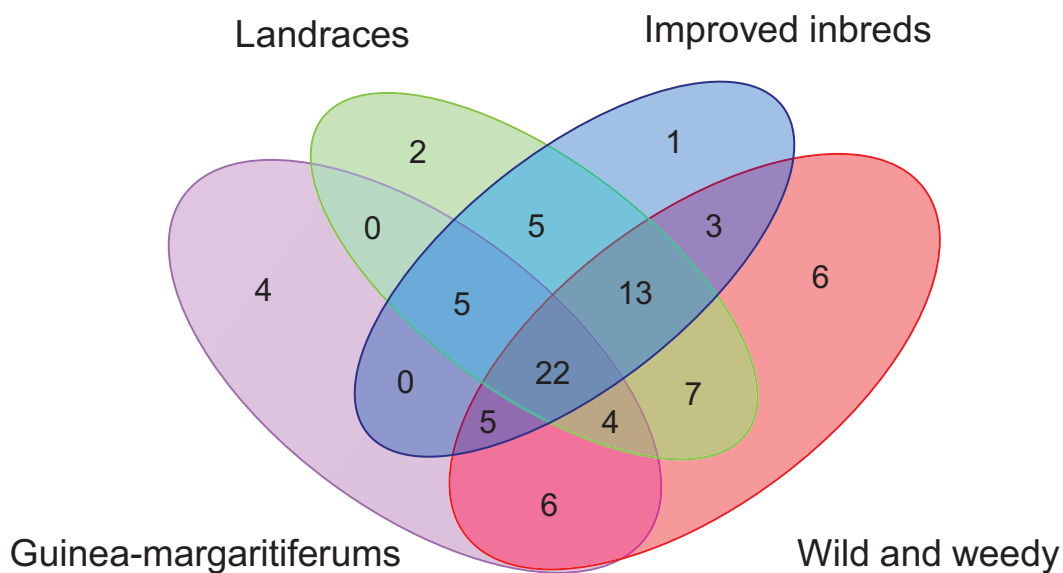


Supplementary Figure S47. Total number of gene losses within and amongst improved inbreds, landraces, guinea-margaritiferums and the wild and weedy genotypes

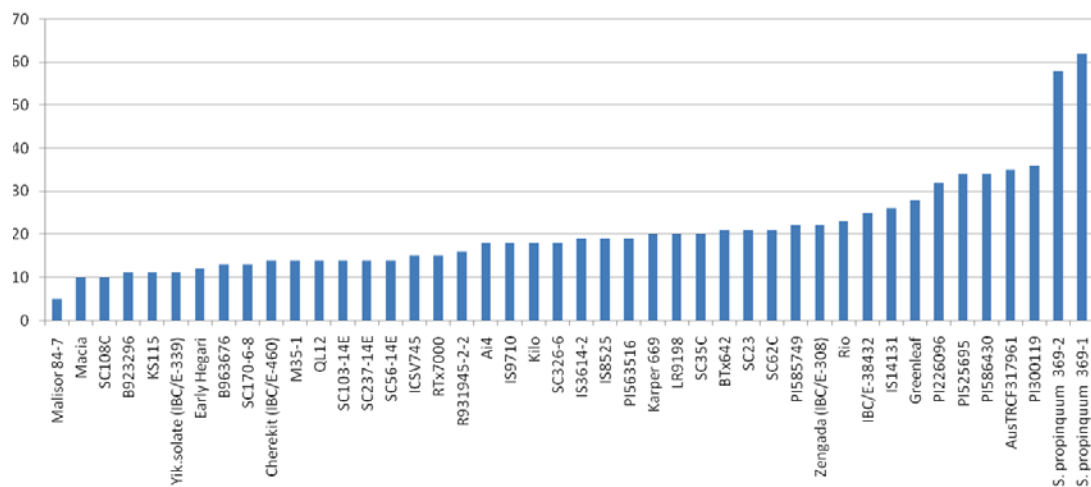


Supplementary Figure S48. PCR validation of six gene gain events; average read depth of genotypes (y-axis) with zero amplification (in blue) versus amplification (in red) for six novel genes (novel_10000186, novel_10000229, novel_10000235, novel_10000299, novel_10000337 and novel_10000356).

a



b



Supplementary Figure S49. Summaries of the total number of novel genes across groups and genotypes; (a) Total number of novel genes within and amongst improved inbreds, landraces, guinea-margaritiferums and the wild and weedy genotypes; (b) Total number of novel genes per genotype

Supplementary Tables

Supplementary Table S1. Total number of SNPs across chromosomes per group

Group	LG	Gene	mRNA	5'UTR	CDS	Intron	3'UTR	Intergenic	Total
All samples*	SBI-01	131,091	130,636	2,243	32,647	87,240	8,506	469,044	600,135
All samples*	SBI-02	100,993	100,782	1,497	28,154	65,098	6,033	456,185	557,178
All samples*	SBI-03	110,736	110,416	1,784	28,622	73,090	6,920	461,429	572,165
All samples*	SBI-04	89,480	89,222	1,366	23,709	58,570	5,577	409,875	499,355
All samples*	SBI-05	59,393	59,377	689	21,020	34,909	2,759	424,390	483,783
All samples*	SBI-06	71,633	71,498	1,051	19,438	47,072	3,937	331,415	403,048
All samples*	SBI-07	56,083	55,937	858	16,924	35,066	3,089	364,815	420,898
All samples*	SBI-08	55,389	55,320	730	18,283	33,490	2,817	342,102	397,491
All samples*	SBI-09	66,307	66,215	1,021	16,758	44,418	4,018	334,359	400,666
All samples*	SBI-10	64,065	63,910	965	17,629	41,383	3,933	387,728	451,793
Wild and Weedy	SBI-01	95,907	95,572	1,405	22,386	65,543	6,238	332,427	428,334
Wild and Weedy	SBI-02	72,364	72,188	965	19,060	47,819	4,344	332,053	404,417
Wild and Weedy	SBI-03	80,633	80,417	1,077	19,467	54,964	4,909	324,743	405,376
Wild and Weedy	SBI-04	67,095	66,893	816	16,803	45,194	4,080	288,819	355,914
Wild and Weedy	SBI-05	41,030	41,025	438	14,226	24,464	1,897	280,248	321,278
Wild and Weedy	SBI-06	51,022	50,894	639	13,475	33,857	2,923	219,621	270,643
Wild and Weedy	SBI-07	39,734	39,628	507	11,620	25,317	2,184	245,302	285,036
Wild and Weedy	SBI-08	39,288	39,249	399	12,508	24,305	2,037	244,740	284,028
Wild and Weedy	SBI-09	47,042	46,987	651	11,465	31,996	2,875	232,689	279,731
Wild and Weedy	SBI-10	44,442	44,286	567	11,582	29,420	2,717	270,928	315,370
Landraces	SBI-01	70,026	69,761	1,082	17,056	47,302	4,321	280,885	350,911
Landraces	SBI-02	58,199	58,115	803	16,355	37,521	3,436	275,696	333,895
Landraces	SBI-03	61,830	61,649	932	16,023	40,830	3,864	282,240	344,070
Landraces	SBI-04	48,144	47,965	709	12,542	31,688	3,026	278,162	326,306
Landraces	SBI-05	38,947	38,937	384	13,771	23,006	1,776	293,778	332,725
Landraces	SBI-06	43,202	43,143	558	11,594	28,573	2,418	223,215	266,417
Landraces	SBI-07	30,661	30,551	429	9,457	19,006	1,659	200,216	230,877
Landraces	SBI-08	35,178	35,139	431	11,653	21,318	1,737	221,312	256,490
Landraces	SBI-09	37,984	37,924	531	9,458	25,740	2,195	206,052	244,036
Landraces	SBI-10	36,232	36,170	506	9,821	23,701	2,142	250,729	286,961
Improved inbreds	SBI-01	62,094	61,858	1,091	15,739	40,997	4,031	219,452	281,546
Improved inbreds	SBI-02	53,583	53,504	806	15,557	33,844	3,297	212,923	266,506
Improved inbreds	SBI-03	55,693	55,540	893	15,129	35,947	3,571	191,700	247,393
Improved inbreds	SBI-04	44,149	44,007	707	12,256	28,238	2,806	190,164	234,313
Improved inbreds	SBI-05	34,226	34,219	386	12,810	19,415	1,608	198,634	232,860
Improved inbreds	SBI-06	38,411	38,351	529	10,766	24,903	2,153	157,038	195,449
Improved inbreds	SBI-07	26,242	26,160	394	8,124	16,284	1,358	143,463	169,705
Improved inbreds	SBI-08	31,618	31,575	447	10,922	18,622	1,584	167,406	199,024
Improved inbreds	SBI-09	35,632	35,564	519	9,513	23,451	2,081	157,383	193,015
Improved inbreds	SBI-10	32,476	32,414	529	9,052	20,823	2,010	180,128	212,604

* excluding *S. propinquum*

Supplementary Table S2. Non-synonymous/synonymous SNP and Ka/Ks ratios across groups

	Non-synonymous SNP	Synonymous SNPs	Nonsyn/Syn	Ka	Ks	Ka/Ks
Wild and weedy	74,587	78,714	0.948	0.00505	0.0179	0.425
Landraces	64,244	64,181	1.001	0.00487	0.0166	0.436
Improved inbreds	60,112	60,304	0.997	0.00479	0.0166	0.427
All *	112,108	112,255	0.999	0.00608	0.0211	0.441

* excluding *S. propinquum*

Supplementary Table S3. Distribution of Ka/Ks ratios across groups

	<0.95	>=0.95 & <=1.05	>1.05
Wild and weedy	11,918	238	984
Landraces	10,273	201	911
Improved inbreds	9,698	189	806
All *	15,651	332	1421

* excluding *S. propinquum*

Supplementary Table S4. Contrasting Ka/Ks values between heterochromatic and euchromatic regions of the genome

	Mean Ka/Ks	Mean Ka	Mean Ks
Euchromatin	0.42346	0.00549	0.02012
Heterochromatin	0.50290	0.00582	0.0169

Supplementary Table S5. Total number of indels genome-wide across all 47 genotypes

Chromosome	# of Indels	# of Insertions	# of Deletions
SBI-01	271388	123199	148189
SBI-02	236827	106232	130595
SBI-03	240449	107577	132872
SBI-04	207091	90762	116329
SBI-05	166453	72153	94300
SBI-06	167280	73611	93669
SBI-07	156522	68412	88110
SBI-08	143641	63082	80559
SBI-09	160202	70471	89731
SBI-10	173343	75065	98278
Super-contigs	59775	21516	38259
Total	1982971	872080	1110891

Supplementary Table S6. Distribution of indels across LGs

LG	Gene	mRNA	5'UTR	CDS	Intron	3'UTR	Intergenic	Total
SBI-01	56183	55888	2615	4236	43711	5326	215205	271388
SBI-02	41633	41472	1666	4028	32158	3620	195194	236827
SBI-03	46005	45814	2055	3662	35864	4233	194444	240449
SBI-04	36807	36665	1532	3001	28634	3498	170284	207091
SBI-05	19991	19977	695	3018	14754	1510	146462	166453
SBI-06	28997	28905	1211	2452	22776	2466	138283	167280
SBI-07	21474	21395	948	2205	16369	1873	135048	156522
SBI-08	19227	19173	801	2312	14505	1555	124414	143641
SBI-09	26162	26065	1062	2134	20442	2427	134040	160202
SBI-10	26297	26205	1127	2458	20256	2364	147046	173343
Super-contigs	1552	1523	52	191	1137	143	58223	59775
Totals	324328	323082	13764	29697	250606	29015	1658643	1982971

Supplementary Table S7. Characteristics of candidate domestication and improvement features under selection

	No. features	Mean feature size (bp)	Mean No. genes/feature	Total No. genes	No. features with 0 genes
Landrace vs Improved inbreds	280	14,502	0.75	211	167
Wild and weedy vs Improved inbreds	462	16,586	0.91	416	243
Wild and weedy vs Landraces	486	17,150	1.20	586	211
Grand total	1228	16,496	0.99	1213	621

Supplementary Table S8. Maximum-likelihood analysis of nucleotide polymorphism in a subset of 422 candidate genes which showed evidence of selective sweeps to determine whether a model of neutral or adaptive evolution best explained the patterns of nucleotide polymorphism. The 422 loci were broken into five input files and run 9 times with the same 10 neutral reference loci used in each run.

Input file	ML statistic (selection)	ML statistic (neutral)	DF	Likelihood ratio statistic	P value*
1	-2399.35	-566.939	85	3664.822	0
1	-2197.3	-566.308	85	3261.984	0
1	-2328.18	-564.633	85	3527.094	0
1	-2465.81	-574.29	85	3783.04	0
1	-2211.3	-577.189	85	3268.222	0
1	-2172.03	-574.413	85	3195.234	0
1	-2988.8	-600.476	85	4776.648	0
1	-3046.37	-597.864	85	4897.012	0
1	-2751.98	-600.073	85	4303.814	0
2	-1816.68	-512.76	85	2607.84	0
2	-1772.02	-517.16	85	2509.72	0
2	-1826.68	-516.556	85	2620.248	0
2	-1816.95	-525.234	85	2583.432	0
2	-1656.27	-526.649	85	2259.242	0
2	-1819.5	-522.12	85	2594.76	0
2	-2287.77	-549.057	85	3477.426	0
2	-2291.57	-550.122	85	3482.896	0
2	-2399.5	-549.5	85	3700	0
3	-1894.7	-510.862	85	2767.676	0
3	-1900.05	-513.284	85	2773.532	0
3	-1893.08	-515.724	85	2754.712	0
3	-1908.71	-522.6	85	2772.22	0
3	-1766.47	-522.715	85	2487.51	0
3	-1918.03	-526.72	85	2782.62	0
3	-2488.66	-547.053	85	3883.214	0
3	-2435.87	-546.979	85	3777.782	0
3	-2478.09	-549.074	85	3858.032	0
4	-1539.21	-462.134	85	2154.152	0
4	-1435.92	-464.056	85	1943.728	0
4	-1523.26	-462.969	85	2120.582	0
4	-1544.6	-472.4	85	2144.4	0
4	-1450.81	-472.984	85	1955.652	0
4	-1421.24	-471.773	85	1898.934	0
4	-2016.64	-495.554	85	3042.172	0
4	-2112.2	-494.622	85	3235.156	0
4	-1620.2	-496.399	85	2247.602	0
5	-1372.05	-400.944	67	1942.212	0
5	-1164.95	-398.489	67	1532.922	5.21E-276
5	-1401.13	-399.894	67	2002.472	0
5	-1385.57	-407.078	67	1956.984	0
5	-1304.42	-407.568	67	1793.704	0
5	-1412.57	-405.877	67	2013.386	0
5	-1976.99	-428.75	67	3096.48	0
5	-1544.63	-429.523	67	2230.214	0
5	-1982.57	-428.777	67	3107.586	0

* P values calculated using a log-likelihood ratio test

Supplementary Table S9. Characteristics of invariant features

	No. features	Mean feature size (bp)	Mean No. genes/feature	Total No. genes	No. features with 0 genes
Landrace vs Improved inbreds	81	10,494	0.38	31	60
Wild and weedy vs Improved inbreds	101	11,683	0.37	37	76
Wild and weedy vs Landraces	77	10,649	0.39	30	57
Grand total	259	11,004	0.38	98	193

Supplementary Table S10. Summary of LE-SNP type per group

Location	Type	Total*	Wild and Weedy	Landrace	Improved inbreds	Guinea-margaritiferums
CDS	Premature stop	1,553	996	915	831	507
	Stop codon to non-stop codon	207	130	119	123	108
	Start codon to non-start codon	178	115	89	89	66
Intron	Splice sites	579	374	334	308	223

* Total excluding propinquum

Supplementary Table S11. Summary of *de novo* assembly of unmapped reads and novel gene prediction

	No. contigs	No. contigs (length >5kb)	No. contigs (Length >10kb)	Total length	Average length (bp)	Longest (bp)	GC content
Novel sequence	4,768	1,201	249	20.8M	4,362	26,807	41.40%
Novel sequence with genes	217	98	28	1.3M	5,991	26,807	43.60%