Supplementary Information:

Comparative population genomics of maize domestication and improvement

Matthew B. Hufford, Xun Xu, Joost van Heerwaarden, Tanja Pyhäjärvi, Jer-Ming Chia, Reed A. Cartwright, Robert J. Elshire, Jeffrey C. Glaubitz, Kate E. Guill, Shawn M. Kaeppler, Jinsheng Lai, Peter L. Morrell, Laura M. Shannon, Chi Song, Nathan M. Springer, Ruth A. Swanson-Wagner, Peter Tiffin, Jun Wang, Gengyun Zhang, John Doebley, Michael D. McMullen, Doreen Ware, Edward S. Buckler, Shuang Yang, Jeffrey Ross-Ibarra

Supplementary Note:

Additional Methods:

Identification of a Putative Inversion on Chromosome 1

A comparison of large-scale patterns of linkage disequilibrium (LD) based on the Infinium 55K SNP data was performed using TASSEL[1] and revealed a striking pattern of long-range LD on chromosome 1 in the *parviglumis* samples (Supplementary Fig. 7c). Combined with our observation of haplotype structure at the locus (Supplementary Fig. 7d), we conclude that this finding represents a previously unknown inversion on the short arm of chromosome 1. The putative inversion, ~50 Mb in size, is polymorphic in *parviglumis* but absent from both landraces and improved lines (further characterized in Ref. 2). Because of the strong differentiation seen between the two haplotypes at the inversion locus -- and thus between *parviglumis* and landraces -- we chose to mask this region from our genome scan. While it is likely that some loci in this region were under selection during domestication, removal of *parviglumis* lines with the inversion reduces our sample size and power to detect selection, making comparison with the rest of the genome difficult.

Rescaling of Site Frequency Spectra

Site frequency spectra were statistically rescaled to a common sample size to handle missing data and the differences in sample size of improved lines, landraces, and *parviglumis*. For each SNP, rescaling predicted a distribution of the sample number of derived alleles under a different sample size, given the currently observed sample. Rescaled spectra were then produced from these probabilities. Source code is available from http://www.rilab.org.

For a SNP, let k be number of derived alleles and, a be the number of ancestral alleles, and $n = a + k$. (Unidentifiable alleles and SNPs with unknown ancestry were dropped.) Let m be the sample size after rescaling. If the sample was already at the rescaling size ($n = m$), no rescaling was done: $P(x \mid a,k,m) = 1$ if $x = k$ and 0 otherwise, where $0 \leq x \leq m$. If a SNP had more samples than the rescaling size ($n > m$), its sample size was reduced using the hypergeometric distribution (sampling without replacement)[3]:

$$P(x \mid a,k,m) = \frac{\binom{k}{x}\binom{a}{m-x}}{\binom{a+k}{m}}$$

If a SNP had fewer observations than the rescaling size ($n < m$), the site needed to be scaled upwards by predicting what additional samples would look like at that site. This was done by putting a prior, $f(p)$, on the underlying allele frequency, p, and then integrating over the prior:

$$P(x,k \mid m,n) = \int_0^1 \binom{n}{k}\binom{m-n}{x-k} p^x (1-p)^{m-x} \times f(p)dp$$

where $k \leq x \leq m - a$. We assumed a uniform prior and found the marginal probability of x:

$$P(x \mid m,k,a) = \frac{P(x,k \mid m,n)}{P(k \mid m,n)} = \frac{P(x,k \mid m,n)}{\sum_{y=k}^{k+m-n} P(y,k \mid m,n)}$$

$$= \frac{a+k+1}{m+1} \frac{\binom{a+k}{k}\binom{m-a-k}{x-k}}{\binom{m}{x}}$$

Sequence-Depth Correlations with Population Genetic Statistics

To investigate the possible effects of sequence-depth biases on our results, we calculated coverage in *parviglumis*, landraces, and improved lines per 10-kb window across sequence with ≤ 50% missing data (*i.e.,* data included in our population genetic analyses). Sequence depth was then correlated with several summary statistics for each group. Across all three taxonomic groups we observed higher coverage in more genic regions and lower coverage in regions rich in transposable elements. Median sequence depth was also lower in 10-kb windows in the lowest quantiles of π. Additionally, we observed higher sequence depth in windows with low Tajima's D and Fay and Wu's H, perhaps due to higher coverage necessary for calling rare variants. The distribution of sequence depth, however, was not correlated with XP-CLR values, our test statistic used to identify features likely subject to selection.

Evidence for *mexicana* introgression

We compared F_{ST}[4] values in 10-kb windows across the genome between the 23 landrace genomes and both *parviglumis* and *mexicana*. Across all windows, F_{ST}

landraces-*parviglumis* and F_{ST} landraces-*mexicana* were highly correlated (R=0.63). We identified putative *mexicana* introgression into landraces by scanning for regions of the genome with multiple consecutive windows with F_{ST} landraces-*parviglumis* in the top 10% quantile and F_{ST} landraces-*mexicana* ≤ 0 (Supplementary Fig. 4).

Transposable element abundance in improved lines and *parviglumis*

We made use of Reads Per Kilobase of Million mapped reads (RPKM) values[5] estimated for each of the 1495 transposable elements (TEs) in the maize UTE database[6] for each of the genomes included in our study. Because of a significant effect of library preparation[5] we only made comparisons between *parviglumis* and improved lines sequenced at the same facility. For each TE, we compared RPKM between *parviglumis* and improved maize with a t-test, and identified TEs as significantly different using a false discovery rate of 1% (Supplementary Table 3 and Supplementary Fig. 6).

Gene Ontology

The GO Slim gene ontology classifications of candidates were determined through annotations obtained from www.maizesequence.org and relative proportions within categories were compared to the entire list of non-candidate genes in the FGS. Enrichment of candidates for gene ontology classes was determined using the resources developed by Du *et al.*[7] by applying a hypergeometric test with a Hochberg (FDR=0.05) correction to account for multiple tests.

QTL comparison

We reanalyzed data from Briggs *et al.*[8] using the r/qtl package[9] for R. Initial scans for QTL were performed for each trait using standard interval mapping with a 1-cM step[10]. The position of the highest significant LOD score on each chromosome was recorded. Significance was determined at a 5% threshold using a permutation test with 10,000 permutations[11]. The positions of significant QTL from the first scan were confirmed using a drop-one ANOVA analysis taking into account the complete model for each trait. Positions of QTL were then refined using multiple imputation[12]; 2.5-cM steps, 300 joint genotype distribution imputations, and an assumed genotyping error rate of 0.01. The "add.qtl" command was used to test for additional QTL. Potentially significant QTL were added in a stepwise manner and subject to both the drop-one ANOVA analysis and the refine step before being considered part of the model. Once the complete model was reached it was refined a final time to determine the precise position of all QTL in the model. Approximate confidence intervals for the locations of QTL were determined using the 1.5 LOD interval. The physical positions of the nearest markers outside the 1.5-LOD interval were identified on AGPv1 of the B73 reference genome.

We tested for overlap of our candidate regions with published QTL intervals for traits implicated in domestication[7] and improvement[13,14]. Domestication QTL were defined as those affecting trait differences between *parviglumis* and maize. Improvement QTL were defined as those affecting traits that are known to have changed during the improvement process such as tassel branch number, and leaf angle[15] and those associated with barrenness and yield under high planting density[13]. Overlapping QTL were fused into joint intervals. We retained joint QTL intervals that were less than 70 Mb in size. Our test statistic consisted of the total proportion of candidate genes overlapping with the set of joint QTL intervals. Significance was evaluated by 1000 independent assignments of random chromosomal positions to each joint QTL interval.

Additional Results and Discussion:

Patterns of diversity

The genome sequence consists of 6% genic sequence, with 18% of 10-kb windows overlapping one or more genes. Mean TE content is 83% with 100% of windows overlapping with one or more TEs. Diversity statistics per 10-kb windows for *parviglumis*, landraces, and improved lines are presented in Supplementary Table 2. Mean nucleotide diversity, π (an estimate of $4N_eu$ where N_e is effective population size) , is reduced by 17% in landraces with respect to *parviglumis*, less of a reduction than reported by Wright *et al.*[16] and Tenaillon *et al.*[17]. F_{ST} between *parviglumis* and landraces is 0.11, similar to previous estimates[18]. Improved lines were chosen to maximize genetic diversity[19], which is reflected in the fact that they retain 98.4% of landrace diversity (Supplementary Table 2). F_{ST} is only 0.02 between the two maize categories. Nucleotide diversity in these lines is similar to previous estimates for genic regions[20], though lower in the genome overall. The population recombination rate ρ ($4N_er$) in *parviglumis* is 1.5 times higher than π at 0.0088. Reduction in ρ is 75% in landraces and a further 29% in improved lines, consistent with previous observations that domestication has caused a stronger reduction in ρ compared to π[16]. The removal of singletons from our data may have caused our values of Tajima's D, a measure of the allele frequency distribution, to be more positive than expected. In genic regions, we observe higher values than reported by Wright *et al.*[16], while confirming their result of a lack of a shift to lower values in maize. We do observe a shift to negative values in landraces and improved lines in the genome overall, however, again suggesting a difference in diversity characteristics between genic and non-genic portions of the genome. That this may in part be due to differential recovery after the domestication bottleneck is suggested by the difference in proportion of segregating sites in maize which are unique (median 24.6% in windows with $\geq$ 1 genic bp, 35.6% in nongenic regions). Normalized Fay and Wu's H', a measure of the distribution of derived alleles, is positive in all groups, indicating a deficit of high frequency derived alleles. Slightly lower values are observed in landraces and improved lines for both genic regions and the genome overall.

From the 30,187 filtered genes for which we had sequence data we identify 177,888 synonymous and 140,185 nonsynonymous SNPs in the coding regions of *parviglumis*, 188,818/156,176 in landraces and 179,277/144,145 in improved lines. Diversity statistics for both synonymous and nonsynonymous sites are similar to those observed for genome-wide genic regions (Supplementary Table 3). Mean synonymous/nonsynonymous F_{ST} per gene between *parviglumis* and landraces is 0.076/0.077 and between landraces and improved lines 0.022/0.018. H' is lower in exons compared to non-coding parts of genes and to the genome as a whole (Supplementary Table 2) but landraces and improved lines again show lower values than *parviglumis* (Supplementary Table 3). The median ratio of nonsynonymous and synonymous nucleotide diversity (π_a/π_s) in all three taxa is ~0.2.

Genome scan

Of the 21,141,953 total SNPs, our genome scan used 71 and 69 percent of SNPs that were polymorphic in the respective reference populations in the domestication and improvement contrasts, resulting in coverage of 94% of 10-kb windows. Of these, ~14.5 M SNPs (98%) were sampled in the overlapping XP-CLR windows.

The domestication contrast resulted in 484 features above the 10% XP-CLR cutoff of 59.2, with an average per feature maximum of 83.3 and the highest feature reaching a value of 185.4. The distribution of feature widths is exponential with a mean of 322 kb and a median of 170 kb. There is a moderate correlation between XP-CLR scores and the width of candidate features (R^2=0.08, p=6.8e-10). In total, 3.6% of all 10-kb windows have values above the cutoff, and 7.6% of the genome overlaps with one of the potentially selected features. The mean estimated selection coefficient, *s*, corresponding to the maximum XP-CLR windows within candidate features is 0.015 compared to 0.0011 for the genome as a whole.

The signal for selection in the improvement contrast is weaker in absolute terms. The 10% XP-CLR cutoff is 12.5, while the average maximum value per feature is 19.1 (maximum: 75.4). A total of 695 features, and 1.6% of the 10-kb windows, exceed the cutoff value. Mean and median feature width are 176 and 110 kb respectively, with 6.0% of the genome contained in candidate features. The correlation between feature width and XP-CLR is lower than for the domestication contrast (R^2=0.008, p=0.0187). Average estimated *s* in features is low at 0.003 but higher than the 0.0003 genome-wide average. Only three peaks, on chromosomes 4, 5 and 7, have values of *s* exceeding the average domestication estimate of 0.015.

A total of 124 improvement features (18%), containing 360 genes, overlap with domestication features. F_{ST} between landraces and improved lines is lower in overlapping peaks (mean F_{ST} = 0.040 vs 0.061 in non-overlapping peaks), suggesting that some of these features may be artifacts caused by the strong diversity reductions during domestication in the landrace population. Nonetheless, 107 of these features are in the lower 50th percentile of $\pi_{(\text{improved lines})} / \pi_{(\text{landraces})}$ indicating that there has been further reduction in diversity from landraces to improved lines and some genes have been subject to ongoing selection.

Domestication and improvement features contain 1,669 and 1,565 filtered genes. Although the largest feature contains 50 genes, 74 and 87 percent of the domestication and improvement features contain less than five genes, 21 and 23 percent contain a single gene while 21 and 28 percent contain no filtered genes.

The 484 domestication features include *tga1*, a classic domestication gene [21]. Interestingly, this gene is not one of the strongest signals of selection; 335 features have a higher maximum value of XP-CLR than the *tga1*-associated feature. Another classic domestication gene, *tb1*[22], was initially not found to be associated with our candidate features. However, inspection of our *parviglumis* sample revealed three lines containing the maize allele. To test if this affected the strength of the selection signal at *tb1* we reanalyzed our data without these three lines. This analysis indeed yields evidence of selection at *tb1* (Supplementary Fig. 14), yet 252 out of 475 domestication peaks have a higher maximum value of XP-CLR than the peak corresponding to *tb1*.

There is a greater density of peaks with high XP-CLR values in more genic regions. For both comparisons, density of XP-CLR features positively correlates with genic content at the 5-Mb level (domestication features: R=0.13, p<1e-12; improvement features: R=0.14, p<1e-12) while the location of both domestication and improvement QTL also broadly overlap with regions of high genic content. This suggests that caution should be taken when evaluating the correspondence of selected features to QTL by resampling random sections of the genome. Bootstrapping of random samples of genes shows no significantly higher QTL overlap of domestication candidate genes (p = 0.645). Improvement candidates show a slightly higher overlap than expected (23% vs 19% of genes, p = 0.016). This result is mostly due to a single barrenness QTL on chromosome 3 that overlaps with a higher than expected number of candidate genes. Excluding this QTL removes the overall significant results (18 vs 17%, p = 0.191).

Diversity and divergence patterns in candidates

Both domestication and improvement features have higher F_{ST} with respect to their ancestral populations (F_{ST} 0.221/0.057, p<0.001, Supplementary Fig. 8) while domestication features have reduced F_{ST} between landraces and improved lines (F_{ST} =0.0003, p<0.001). Nucleotide diversity, Tajima's D, and H′ in landraces and improved lines are reduced in both types of features compared to *parviglumis*. Nucleotide diversity in domestication and improvement features is also significantly reduced in *parviglumis* relative to non-candidate features suggesting that these candidates could be of functional importance in the wild ancestor. Improvement features show an unexpected strong reduction of diversity and lower D and H' in landraces but these reductions are much stronger in features overlapping with domestication features (Supplementary Fig. 8). Both domestication and improvement features have lower coverage than the rest of the genome (Supplementary Fig. 8), but correlation with XP-CLR is very low (R^2=0.006, 0.027 at the feature level, 0.002, 0.016 for the genome as a whole) suggesting coverage has a negligible effect on the identification of candidate regions.

The effects of domestication and improvement are similar for synonymous and nonsynonymous diversity in candidate genes compared to our estimates based on features (Supplementary Fig. 9). For domestication and improvement candidate genes respectively, F_{ST} for synonymous sites is 0.196 between teosinte and landraces and 0.058 between landraces and improved lines, significantly higher than non-candidates (two-sample Wilcoxon test, p<0.001 in both contrasts). In contrast to genome-wide patterns, we find no increase of low-frequency variants within genes in landraces relative to *parviglumis*, (synonymous Tajima's D *parviglumis* = 0.39 and landrace = 0.53), a result that echoes earlier analyses of maize genes[16,17] and is suggestive of the action of purifying selection on linked sites slowing the recovery of diversity post-domestication. Tajima's D in synonymous sites in domestication candidates is positive in *parviglumis* (0.46) and close to zero for landraces and improved lines (-0.02 and 0.08). In the improvement candidates, improved lines have the lowest value of synonymous Tajima's D (0.10 vs. 0.45 in landraces and 0.42 in *parviglumis*).

Synonymous Tajima's D is significantly lower in domestication candidates compared to non-candidates (t-test, p<0.001) in landraces and improved lines but slightly higher (p=0.09) in domestication candidates compared to non-candidate genes in *parviglumis*. D is lower in improvement candidates compared to non-candidates in improved lines (p<0.001). In landraces and *parviglumis*, the improvement candidates do not have significantly (p=0.07, p=0.34) different Tajima's D compared to non-candidate genes. Synonymous and nonsynonymous H′ was similarly reduced during domestication and improvement (Supplementary Fig. 9) indicating that candidate genes have more high frequency derived alleles compared to non-candidates.

Evidence for *mexicana* introgression

Introgression between *mexicana* and maize is frequent, especially in the higher altitudes of the Central Plateau[23]. Because of difficulties creating advanced inbred lines, our sample of maize landraces does not include highland Mexican material and thus should show little evidence for introgression from *mexicana*. Across the genome, F_{ST} between landraces and *mexicana* (mean = 0.077) is lower than F_{ST} between landraces and *parviglumis* (mean = 0.11), likely due to the small sample size for *mexicana*. Nonetheless, the two values of F_{ST} are well correlated (r=0.63), and differentiation between *parviglumis* and *mexicana* is low overall (mean F_{ST} = 0.0308). There are a number of regions across the genome, however, that are suggestive of introgression between maize and *mexicana* (e.g., Supplementary Fig. 4). Of the 18,254 windows in the upper 10% quantile of F_{ST} between *parviglumis* and landraces, 1,299 have an F_{ST} of 0 between *mexicana* and landraces.

Characterization of Domestication/Improvement Candidates

After filtering features based on reduction of π between the reference and object populations, the list of candidates includes 468 genes for domestication (1.4% of the FGS) (Supplementary Table 6) and 571 genes for improvement (1.8% of the FGS) (Supplementary Table 7). Of these candidates, 45 genes are found in both the domestication and improvement comparisons, significantly more overlap than is expected by chance (p=0.001).

Some domestication candidates such as *yabby14* (GRMZM2G054795), *zag1* (GRMZM2G052890), *zag2* (GRMZM2G160687), *bif2* (*barren inflorescence2*, GRMZM2G171822) and *zfl2* (*zea floricaula/leafy2*, GRMZM2G180190) are relatively well known genes in maize. *yabby14* is expressed on the adaxial side of developing lateral organs and associated with regions of lateral blade outgrowth in maize[24]. *zag1* and *zag2* are MADS box genes previously identified as targets of selection during domestication[25]. *bif2* interacts with *barren stalk1* in inflorescence development in maize[26-28], whereas the well known gene *zfl2* is associated with flowering time and plant architecture and has been previously associated with domestication in maize[29]. Of 13 genes previously associated with maize domestication, five are identified as domestication candidates and 11 are located within domestication features (Supplementary Table 9).

Several domestication candidates are interesting based on their homology even though their exact function in maize is not known. GRMZM2G359564 has protein homology to FACT complex subunit SPT16 that affects vegetative and reproductive development in *Arabidopsis*[30]. AC207628.4_FG006, which is highly expressed in maize leaves, is orthologous to *Arabidopsis* AMY3 that is involved in starch degradation[31]. GRMZM2G073044 is likely to be part of the LBD gene family that is important in lateral organ development in plants[32].

There are also a few known classical maize genes among improvement candidates. Among them is *tb1* (*teosinte branched1*, AC233950.1_FG002) a known domestication gene that affects plant architecture [22,33,34]. Other known genes are *zmm2* (*Zea mays MADS2*, GRMZM2G359952), a MADS box gene strongly expressed in tassels[35], *su2* (*sugary2*, GRMZM2G348551), a gene for starch synthase IIa[36], and *viviparous15* (GRMZM2G121468) that encodes a molybdopterin (MPT) synthase subunit acting in the ABA-biosynthetic pathway[37].

Several improvement candidates have putative functions in floral development and timing. GRMZM2G089640 is an ortholog of AUXIN RESPONSE FACTOR 6, a gene that regulates *Arabidopsis* ovule and anther development[38]. Another candidate, GRMZM2G103666 (*zcn12*), an *FT*-like gene, is found in a flowering-time QTL on chromosome 3 and is highly expressed in maize during reproductive stages of development[39-41]. GRMZM2G107101 is an ortholog of GIGANTEA, which is a well-known gene in the circadian-clock-dependent flowering pathway of *Arabidopsis*[42]. Additionally, GRMZM2G310069 is an ortholog of *Arabidopsis* LAF3 that participates in light receptor phyA signaling[43]. Other examples of potentially interesting improvement candidate genes are orthologs of *Arabidopsis* SUGAR-DEPENDENT1 (GRMZM2G087612), DOWNY MILDEW RESISTANT 6 (GRMZM2G475380) and PLANT GLYCOGENIN-LIKE STARCH INITIATION PROTEIN 2 (GRMZM2G058472).

Gene Ontology

The enrichment of both domestication and improvement candidates for Gene Ontology GOSlim categories was determined relative to annotated non-candidates in the FGS of maize (Supplementary Fig. 11). Domestication candidates are marginally enriched for genes involved in bio-synthetic processes whereas improvement candidates are marginally enriched for genes associated with membrane-bound organelles, cellular processes, enzyme regulator activity and the nucleus. However, the significance of these observed enrichments does not survive correction for multiple tests.

Expression

The degree of variation in expression among lines, patterns of tissue-specific expression, and dominance were assessed in candidate and non-candidate genes based on full-transcriptome expression data. In our dataset of gene expression from a subset of 25 maize and seven *parviglumis* lines, we retained data for 271/468 domestication candidates and for 336/571 improvement candidates after filtering for quality. Significance was determined by bootstrap resampling. Taking into account both the coefficient of variation (CoV) and the difference between the maximum and minimum expression values across lines, domestication candidates show significantly lower variation in expression among maize lines than random sets of genes (p=0.006, 10% reduction in CoV in candidates compared to non-candidates). Variation in expression is also significantly reduced in maize improvement candidates versus random sets of genes (p=0.019, 8% reduction in CoV in candidates compared to non-candidates). Lower variation in expression in maize relative to *parviglumis* in both domestication and improvement candidates could be due to the loss of additive genetic variation during the two bouts of selection that likely stabilized levels of cis-regulated expression.

In addition to analyses regarding variation in expression across lines, we also investigated differences in the magnitude of expression of our candidates versus non-candidates. Neither domestication nor improvement candidates show significant differences in overall expression levels when compared to non-candidates. However, while the overall percentage of domestication candidates with increased expression in maize (55%) is not significantly higher (p=0.065) than seen in non-candidates (50%), the percentage of domestication candidates with a 1.5-fold increase in expression in maize is significantly higher (p=0.001, 11% in domestication candidates versus 7% in non-candidates). A significant up- or down-regulation in expression is not observed in improvement candidates. Likewise, in domestication candidates, the magnitude of change in expression (increased or decreased) from *parviglumis* to maize is significantly greater (p=0.004) than non-candidates (22% increase in magnitude of change in candidates relative to non-candidates). Significant differences in the magnitude of expression change are not observed in improvement candidates. These results indicate that substantial up- or down-regulation in expression in maize relative to *parviglumis* was more prevalent in domestication than improvement.

Interestingly, trends observed for candidate genes and all genes within features were qualitatively similar for variation in expression and the magnitude of change in expression from teosinte to maize for both domestication and improvement. These results indicate that the effects of selection may not be limited to single candidates but may also extend to linked sites. While this may prove problematic for identifying candidates based on transcriptome data alone, it demonstrates a meaningful effect of selection on cis-acting variation across wide genomic regions.

Recently, expression data for over 60 different tissues in B73 have been made available for most genes in the FGS[41], making it possible to characterize tissue-specific patterns of expression within a great number of our candidates (339/468 domestication candidates and 428/571 improvement candidates). These data also reveal a substantial bias toward the inclusion of constitutively expressed genes in loci previously scanned for selection during maize domestication[16,44,45], likely due to their ascertainment from EST libraries that are enriched for genes expressed in multiple tissues. In contrast, both our domestication and improvement candidates appear to be a representative cross-section of expression profiles (Supplementary Table 6). Using these data, we were able to test for changes in the magnitude of expression of candidates in 11 different tissue types in maize (Supplementary Fig. 13). After correcting for multiple tests (Benjamini and Hochberg FDR = 0.05) domestication candidates do not show significantly higher expression than non-candidates in any given tissue (Supplementary Fig. 13a). In stark contrast, improvement candidates are significantly more highly expressed in all but one of the tissue types tested (p=0.024-0.044, Supplementary Fig. 13b), a result indicating that crop improvement has targeted more highly and perhaps more constitutively (Supplementary Table 8) expressed genes.

Dominance was assessed by comparing hybrid to midparent expression levels in each of five crosses (three intra-heterotic group crosses and two inter-heterotic group crosses). After filtering for quality, data were available for 125/468 domestication candidates and 131/571 improvement candidates. Dominance is significantly higher in domestication candidates when compared to non-candidates (p=0.001). When domestication dominance estimates are split into intra- and inter-heterotic groups, no significant difference is detected between the two groups (t-test, p=0.74). Patterns of dominance differ substantially in improvement candidates: like domestication candidates, these genes show a significant elevation in dominance (p=0.007), however, when estimates are split into inter- and intra-heterotic group crosses, inter-heterotic group crosses show markedly elevated dominance in comparison to intra-heterotic group crosses (t-test, p=0.013). In fact, the significant elevation in dominance seen when considering all crosses is largely attributable to crosses between heterotic groups (p=0.001, within heterotic group elevated dominance p=0.060). These trends suggest dominance and complementation have played a role in maize improvement in crosses between genetic groups.

References

1. Bradbury, P. J. et al. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633-2635 (2007).
2. Fang, Z. et al. Megabase-scale inversion polymorphism in the wild ancestor of maize. *Genetics.* published online, doi:10.1534/genetics.112.138578 (27 April 2012).
3. Nielsen, R. et al. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol* **3**, 976-985-e170 (2005).
4. Hudson, R. R., Boos, D. D. & Kaplan, N. L. A statistical test for detecting geographic subdivision. *Mol Biol Evol* **9**, 138-151 (1992).
5. Chia, J.-M. et al. Maize HapMap2 identifies extant variation from a genome in flux**.** *Nat Genet* (Accept in Principle).
6. Tenaillon, M. I., Hufford, M. B., Gaut, B. S. & Ross-Ibarra, J. Genome size and transposable element content as determined by high-throughput sequencing in maize and *Zea luxurians*. *Genome Biol Evol* **3**, 219-229 (2011).
7. Du, Z., Zhou, X., Ling, Y., Zhang, Z. H. & Su, Z. agriGO: a GO analysis toolkit for the agricultural community. *Nucleic Acids Res* **38**, W64-W70 (2010).
8. Briggs, W. H., McMullen, M. D., Gaut, B. S. & Doebley, J. Linkage mapping of domestication loci in a large maize-teosinte backcross resource. *Genetics* **177**, 1915-1928 (2007).
9. Broman, K. W., Wu, H., Sen, S. & Churchill, G. A. R/qtl: QTL mapping in experimental crosses. *Bioinformatics* **19**, 889-890 (2003).
10. Lander, E. S. & Botstein, D. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185-199 (1989).
11. Doerge, R. W. & Churchill, G. A. Permutation tests for multiple loci affecting a quantitative character. *Genetics* **142**, 285-294 (1996).
12. Sen, S. & Churchill, G. A. A statistical framework for quantitative trait mapping. *Genetics* **159**, 371-387 (2001).
13. Gonzalo, M., Holland, J. B., Vyn, T. J. & McIntyre, L. M. Direct mapping of density response in a population of B73 x Mo17 recombinant inbred lines of maize (*Zea mays* L.). *Heredity* **104**, 583-599 (2010).
14. Mickelson, S. M., Stuber, C. S., Senior, L. & Kaeppler, S. M. Quantitative trait loci controlling leaf and tassel traits in a B73 x MO17 population of maize. *Crop Sci* **42**, 1902-1909 (2002).
15. Duvick, D. N. & Cassman, K. G. Post-green revolution trends in yield potential of temperate maize in the north-central United States. *Crop Sci* **39**, 1622-1630 (1999).
16. Wright, S. I. et al. The effects of artificial selection of the maize genome. *Science* **308**, 1310-1314 (2005).
17. Tenaillon, M. I., U'Ren, J., Tenaillon, O. & Gaut, B. S. Selection versus demography: A multilocus investigation of the domestication process in maize. *Mol Biol Evol* **21**, 1214-1225 (2004).
18. Ross-Ibarra, J., Tenaillon, M. & Gaut, B. S. Historical divergence and gene flow in the genus *Zea*. *Genetics* **181**, 1397-1409 (2009).
19. Yu, J. M., Holland, J. B., McMullen, M. D. & Buckler, E. S. Genetic design and statistical power of nested association mapping in maize. *Genetics* **178**, 539-551 (2008).
20. Gore, M. A. et al. A first-generation haplotype map of maize. *Science* **326**, 1115-1117 (2009).
21. Wang, H. et al. The origin of the naked grains of maize. *Nature* **436**, 714-719 (2005).
22. Doebley, J., Stec, A. & Hubbard, L. The evolution of apical dominance in maize. *Nature* **386**, 485-488 (1997).

23. van Heerwaarden, J. et al. Genetic signals of origin, spread, and introgression in a large sample of maize landraces. *Proc Natl Acad Sci U S A* **108**, 1088-1092 (2011).

24. Juarez, M. T., Twigg, R. W. & Timmermans, M. C. Specification of adaxial cell fate during maize leaf development. *Development* **131**, 4533-4544 (2004).

25. Zhao, Q., Weber, A. L., McMullen, M. D., Guill, K. & Doebley, J. MADS-box genes of maize: frequent targets of selection during domestication. *Genet Res* **93**, 65-75 (2011).

26. McSteen, P. & Hake, S. barren inflorescence2 regulates axillary meristem development in the maize inflorescence. *Development* **128**, 2881-2891 (2001).

27. Pressoir, G. et al. Natural variation in maize architecture is mediated by allelic differences at the PINOID co-ortholog barren inflorescence2. *Plant J* **58**, 618-628 (2009).

28. Skirpan, A., Wu, X. & McSteen, P. Genetic and physical interaction suggest that BARREN STALK 1 is a target of BARREN INFLORESCENCE2 in maize inflorescence development. *Plant J* **55**, 787-797 (2008).

29. Bomblies, K. & Doebley, J. F. Pleiotropic effects of the duplicate maize FLORICAULA/LEAFY genes *zfl1* and *zfl2* on traits under selection during maize domestication. *Genetics* **172**, 519-531 (2006).

30. Lolas, I. B. et al. The transcript elongation factor FACT affects *Arabidopsis* vegetative and reproductive development and genetically interacts with HUB1/2. *Plant J* **61**, 686-697 (2010).

31. Kotting, O. et al. STARCH-EXCESS4 is a laforin-like Phosphoglucan phosphatase required for starch degradation in Arabidopsis thaliana. *Plant Cell* **21**, 334-346 (2009).

32. Majer, C. & Hochholdinger, F. Defining the boundaries: structure and function of LOB domain proteins. *Trends Plant Sci* **16**, 47-52 (2011).

33. Hubbard, L., McSteen, P., Doebley, J. & Hake, S. Expression patterns and mutant phenotype of teosinte branched1 correlate with growth suppression in maize and teosinte. *Genetics* **162**, 1927-1935 (2002).

34. Wang, R. L., Stec, A., Hey, J., Lukens, L. & Doebley, J. The limits of selection during maize domestication. *Nature* **398**, 236-239 (1999).

35. Mena, M. et al. Diversification of C-function activity in maize flower development. *Science* **274**, 1537-1540 (1996).

36. Zhang, X. L. et al. Molecular characterization demonstrates that the *Zea mays* gene sugary2 codes for the starch synthase isoform SSIIa. *Plant Mol Biol* **54**, 865-879 (2004).

37. Suzuki, M. et al. The maize viviparous15 locus encodes the molybdopterin synthase small subunit. *Plant J* **45**, 264-274 (2006).

38. Wu, M. F., Tian, Q. & Reed, J. W. *Arabidopsis* microRNA167 controls patterns of ARF6 and ARF8 expression, and regulates both female and male reproduction. *Development* **133**, 4211-4218 (2006).

39. Buckler, E. S. et al. The genetic architecture of maize flowering time. *Science* **325**, 714-718 (2009).

40. Danilevskaya, O. N., Meng, X., Hou, Z. L., Ananiev, E. V. & Simmons, C. R. A genomic and expression compendium of the expanded PEBP gene family from maize. *Plant Physiol* **146**, 250-264 (2008).

41. Sekhon, R. S. et al. Genome-wide atlas of transcription during maize development. *Plant J* **66,** 553-563 (2011).

42. Park, D. H. et al. Control of circadian rhythms and photoperiodic flowering by the Arabidopsis GIGANTEA gene. *Science* **285**, 1579-1582 (1999).

43. Hare, P. D., Moller, S. G., Huang, L. F. & Chua, N. H. LAF3, a novel factor required for normal phytochrome A signaling. *Plant Physiol* **133**, 1592-1604 (2003).

44. Hufford, K. M., Canaran, P., Ware, D. H., McMullen, M. D. & Gaut, B. S. Patterns of selection and tissue-specific expression among maize domestication and crop improvement loci. *Plant Physiol* **144**, 1642-1653 (2007).

45. Yamasaki, M. et al. A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. *Plant Cell* **17**, 2859-2872 (2005).

46. Sigmon, B. & Vollbrecht, E. Evidence of selection at the ramosa1 locus during maize domestication. *Mol Ecol* **19**, 1296-1311 (2010).

47. Whitt, S. R., Wilson, L. M., Tenaillon, M. I., Gaut, B. S. & Buckler, E. S. Genetic diversity and selection in the maize starch pathway. *Proc Natl Acad Sci U S A* **99**, 12959-12962 (2002).

48. Jaenicke-Despres, V. et al. Early allelic selection in maize as revealed by ancient DNA. *Science* **302**, 1206-1208 (2003).

49. Ueda, T., Wang, Z. D., Pham, N. & Messing, J. Identification of a transcriptional activator-binding element in the 27-kilodalton zein promoter, the -300-element. *Mol Cell Biol* **14**, 4350-4359 (1994).

50. Wang, Z. D., Ueda, T. & Messing, J. Characterization of the maize prolamin box-binding factor-1 (PBF-1) and its role in the developmental regulation of the zein multigene family. *Gene* **223**, 321-332 (1998).

51. Whipple, C. J. et al. grassy tillers1 promotes apical dominance in maize and responds to shade signals in the grasses. *Proc Natl Acad Sci U S A* **108**, E506-E512 (2011).

52. Gallavotti, A. et al. The role of barren stalk1 in the architecture of maize. *Nature* **432**, 630-635 (2004).

53. Jiang, C. et al. Genetic analysis of adaptation differences between highland and lowland tropical maize using molecular markers. *Theor Appl Genet* **99**, 1106-1119 (1999).

Supplementary Tables

Supplementary Table 1. Information on lines sequenced and depth of sequencing and depth of mapped reads. Asterisks denote lines included in the expression analysis.

Category	Line ID	Description	USDA Germplasm ID	Locality	Sequence Depth
Improved	B73*	Temperate	PI 550473	Iowa, USA	2.83
Improved	B97*	Temperate	PI 564682	Iowa, USA	5.38
Improved	CAU178	Chinese-Temperate	--	China	6.81
Improved	CAU478	Chinese-Temperate	--	China	6.82
Improved	CAU5003	Chinese-Temperate	--	China	7.2
Improved	CAUCHANG72	Chinese-Temperate	--	China	7.02
Improved	CAUMO17	Chinese-Temperate	--	China	8.36
Improved	CAUZHENG58	Chinese-Temperate	--	China	6.97
Improved	CML103*	Tropical	Ames 27081	Federal District, Mexico	3.02
Improved	CML228*	Tropical	Ames 27088	Federal District, Mexico	3.02
Improved	CML247*	Tropical	PI 595541	Federal District, Mexico	2.68
Improved	CML277*	Tropical	PI 595550	Federal District, Mexico	2.6
Improved	CML322*	Tropical	Ames 27096	Federal District, Mexico	2.86
Improved	CML333*	Tropical	Ames 27101	Federal District, Mexico	3.86
Improved	CML52*	Tropical	PI 595561	Federal District, Mexico	4.01
Improved	CML69	Tropical	Ames 28184	Federal District, Mexico	4.58
Improved	HP301*	Temperate	PI 587131	Indiana, USA	3.73
Improved	IL14H*	Temperate	Ames 27118	Illinois, USA	4.18
Improved	KI11*	Tropical	Ames 27124	Thailand	2.6
Improved	KI3*	Tropical	Ames 27123	Thailand	4.77
Improved	KY21*	Temperate	Ames 27130	Kentucky, USA	2.4
Improved	M162W*	Temperate	Ames 27134	North Carolina, USA	3.22
Improved	M37W*	Mixed	Ames 27133	Natal, South Africa	3.04
Improved	MO17*	Temperate	PI 558532	Missouri, USA	2.92
Improved	MO18W*	Mixed	PI 550441	Missouri, USA	3.1

Improved	NC350	Tropical	Ames 27171	North Carolina	2.37
Improved	NC358*	Tropical	Ames 27175	North Carolina, USA	2.32
Improved	OH43*	Temperate	Ames 19288	Ohio, USA	19.94
Improved	OH7B*	Temperate	Ames 19323	Ohio, USA	3.93
Improved	P39*	Temperate	PI 587133	Indiana, USA	4.5
Improved	TX303*	Mixed	Ames 19327	Texas, USA	5.42
Improved	TZI8*	Tropical	PI 506246	Oyo, Nigeria	2.18
Improved	W22*	Temperate	NSL 30053	Wisconsin, USA	4.52
Improved	MS71	Temperate	PI 587137	Michigan, USA	4.26
Improved	W64A	Temperate	PI 587152	Wisconsin, USA	19.05
Landrace	MR01	Araguito	Ames 30522	Anzoategui, Venezuela	5.51
Landrace	MR02	Assiniboine	Ames 30523	Bismark, ND, USA	5.59
Landrace	MR03	Bolita	--	Nochixtlan, Mexico	5.63
Landrace	MR05	Cateto	Ames 30524	Canamina/La Paz, Bolivia	5.15
Landrace	MR06	Chapalote	--	Culiacan, Mexico	5.34
Landrace	MR07	Comiteco	Ames 30525	Comitan, Mexico	5.44
Landrace	MR08	Costeno	Ames 30526	Las Colmenas, Venezuela	5.18
Landrace	MR09	Cravo Riogranense	Ames 30527	Brazil	5.1
Landrace	MR10	Cristalino Norteno	Ames 30528	Cautin, Chile	5.62
Landrace	MR11	Cuban Flint	Ames 30529	Havana, Cuba	5.55
Landrace	MR12	Havasupai	Ames 30530	Supai, Arizona, USA	5.75
Landrace	MR13	Hickory King	--	Virginia, USA	2.75
Landrace	MR14	Longfellow Flint	Ames 30531	Northeast, USA	5.4
Landrace	MR17	Pisankalla	--	Tarija, Bolivia	5.51
Landrace	MR18	Reventador	Ames 30532	Las Penitas, Mexico	5.58
Landrace	MR19	Santa Domingo	Ames 30533	Santa Domingo Pueblo, New Mexico, USA	5.41
Landrace	MR20	Shoe Peg	Ames 30534	Dora, Missouri, USA	5.25
Landrace	MR21	Tabloncillo	--	Santa Ana, Mexico	5.26
Landrace	MR22	Tuxpeno	Ames 30535	Ursulo Galvan, Mexico	5.49
Landrace	MR23	Zapalote Chico	Ames 30536	Tehuantepec, Mexico	5.57
Landrace	MR24	Chullpi	--	Huanta, Peru	5.57
Landrace	MR25	Pororo	Ames 30537	San Ignacio/Velasco, Bolivia	5.43

Landrace	MR26	Pollo	--	Tiribita, Columbia	5.68
parviglumis	TIL15*	*Zea mays ssp. parviglumis*	Ames 28407	Palo Blanco, Guerrero, Mexico	5.98
parviglumis	TIL01*	*Zea mays ssp. parviglumis*	Ames 28399	Tzitzio, Michoacan, Mexico	3.93
parviglumis	TIL03	*Zea mays ssp. parviglumis*	Ames 28400	Toliman, Jalisco, Mexico	3.26
parviglumis	TIL04	*Zea mays ssp. parviglumis*	--	Teloloapan, Guerrero, Mexico	4.56
parviglumis	TIL05	*Zea mays ssp. parviglumis*	--	San Pedro Juchatengo, Oaxaca, Mexico	5.74
parviglumis	TIL06*	*Zea mays ssp. parviglumis*	Ames 28401	Chilpancingo, Guerrero, Mexico	3.19
parviglumis	TIL07	*Zea mays ssp. parviglumis*	Ames 28402	Tierra Colorada, Guerrero, Mexico	3.79
parviglumis	TIL09*	*Zea mays ssp. parviglumis*	Ames 28403	Tejupilco, Mexico, Mexico	4.55
parviglumis	TIL10	*Zea mays ssp. parviglumis*	Ames 28404	Teloloapan, Guerrero, Mexico	4.11
parviglumis	TIL11*	*Zea mays ssp. parviglumis*	Ames 28405	Amatlan De Canas, Nayarit, Mexico	2.11
parviglumis	TIL12	*Zea mays ssp. parviglumis*	--	Huitzuco, Guerrero, Mexico	2.11
parviglumis	TIL14*	*Zea mays ssp. parviglumis*	Ames 28406	El Rodeo, Jalisco, Mexico	4.06
parviglumis	TIL16	*Zea mays ssp. parviglumis*	Ames 28408	Palo Blanco, Guerrero, Mexico	2.97
parviglumis	TIL17*	*Zea mays ssp. parviglumis*	Ames 28409	Teloloapan, Guerrero, Mexico	3
mexicana	TIL08	*Zea mays ssp. mexicana*	--	Tepoztlan, Morelos, Mexico	4.56
mexicana	TIL25	*Zea mays ssp. mexicana*	Ames 28398	Degollado, Jalisco, Mexico	9.1
Tripsacum	MIA34597	*Tripsacum dactyloides var. meridionalis*	MIA 34597	Santander, Colombia	12.62
* lines included in expression analyses				**Average Sequence/Coverage Depth**	5.05

Supplementary Table 2. Population genetic summary statistics across 10-kb windows. Statistics are provided for all 10-kb windows (white) and genic 10-kb windows (grey). Shown are values of nucleotide diversity (π), Tajima's D (TajD), Fay and Wu's H' (H'), and estimates of the population recombination rate (ρ). Values of ρ and π are reported per bp.

Statistic	*parviglumis*	landrace	improved
π	0.0059	0.0049	0.0048
π_{genic}	0.0083	0.0072	0.0071
TajD	0.0412	-0.0716	-0.2132
TajD_{genic}	0.4475	0.4543	0.4129
H'	3.0815	2.7637	2.5368
H'_{genic}	2.9554	2.6847	2.4665
ρ	0.0088	0.0022	0.0016
ρ_{genic}	0.0139	0.0040	0.0024

Supplementary Table 3. Population genetic summary statistics in candidate genes. Statistics are provided for synonymous (white) and non-synonymous sites (grey) as well as in all genes (all) and candidate genes (cand). ρ and π values are reported per bp. Abbreviations are as in Supplementary Table 2.

Domestication

	parviglumis		landrace		improved	
Statistic	all	cand	all	cand	all	cand
π_{syn}	0.0078	0.0067	0.0074	0.0043	0.0072	0.0046
π_{nonsyn}	0.0022	0.0015	0.0020	0.0010	0.0020	0.0012
TajD_{syn}	0.3943	0.4628	0.5285	-0.0230	0.5335	0.0814
TajD_{nonsyn}	0.2808	0.2761	0.3844	-0.1729	0.3630	-0.1467
H'_{syn}	0.4907	0.3939	0.3179	-0.5305	-0.0170	-0.7934
H'_{nonsyn}	0.5940	0.5633	0.4415	-0.4021	0.1483	-0.6695

Improvement

	parviglumis		landrace		improved	
Statistic	all	cand	all	cand	all	cand
π_{syn}	0.0078	0.0071	0.0074	0.0063	0.0072	0.0053
π_{nonsyn}	0.0022	0.0019	0.0020	0.0019	0.0020	0.0015
TajD_{syn}	0.3943	0.4246	0.5285	0.4496	0.5335	0.1016
TajD_{nonsyn}	0.2808	0.2913	0.3844	0.3082	0.3630	0.0014
H'_{syn}	0.4907	0.4188	0.3179	0.2038	-0.0170	-0.5994
H'_{nonsyn}	0.5940	0.5462	0.4415	0.2841	0.1483	-0.3212

Supplementary Table 4. Transposable element families significantly differing in abundance in improved lines and *parviglumis*. Family names follow nomenclature in the maize TE database (www.maizetedb.org). Shown are ratios of mean abundance (in reads per kb per million mapped reads) and p-values from a two-tailed t-test. A false discovery rate of 1% was used to determine significance.

Table S4A. Transposable element families with significantly lower copy number in improved lines relative to parviglumis.

Class	Order	Super-family	Family	Sequence Reference	Ratio (Improved line/parviglumis)	p-value
1	long terminal repeat (LTR)	Unknown LTR (RLX)	avahi	AC191363-3084	0.394881648	4.90018E-15
1	long terminal repeat (LTR)	Unknown LTR (RLX)	CRM3	AC200048-6717	0.798003211	2.11131E-06
1	long terminal repeat (LTR)	Unknown LTR (RLX)	sela	AC195130-4415	0.505740577	3.43872E-06
1	long terminal repeat (LTR)	Unknown LTR (RLX)	guali	AC190263-78	0.315362265	5.02424E-06
1	long terminal repeat (LTR)	Gypsy (RLG)	CRM1	AC207803-9728	0.830250898	7.52176E-06
1	long terminal repeat (LTR)	Unknown LTR (RLX)	osed	AC191084-2931	0.514084242	1.28905E-05
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00279	consensus	0.732767878	2.69658E-05
1	long terminal repeat (LTR)	Gypsy (RLG)	CRM1	AC208678-9984	0.863712286	4.08886E-05
1	long terminal repeat (LTR)	Gypsy (RLG)	aneas	AC203312-7773	0.639486127	0.00044777
1	long terminal repeat (LTR)	Unknown LTR (RLX)	wihov	AC205351-8739	0.863826605	0.000669374

Table S4B. Transposable element families with significantly higher copy number in improved lines relative to parviglumis.

Class	Order	Super-family	Family	Sequence Reference	Ratio (Improved line/parviglumis)	p-value
1	long terminal repeat (LTR)	Gypsy (RLG)	cinful-zeon	AC192460-3502	1.768188659	2.47803E-14
1	long terminal repeat (LTR)	Unknown LTR (RLX)	CRM2	AC206920-9397	1.755280906	7.72914E-13
1	long terminal repeat (LTR)	Gypsy (RLG)	cinful-zeon	AC206171-9091	1.66697475	1.81181E-12
1	long terminal repeat (LTR)	Copia (RLC)	opie	AC198924-6206	1.535939237	5.70921E-12
2	terminal inverted repeat (TIR)	Cacta (DTC)	ZM00001	consensus	1.355162201	3.80559E-11
1	long terminal repeat (LTR)	Gypsy (RLG)	cinful-zeon	AC194954-4372	1.842914227	3.98012E-11
2	terminal inverted repeat (TIR)	Cacta (DTC)	ZM00084	consensus	1.317778219	1.52833E-09
2	terminal inverted repeat (TIR)	Cacta (DTC)	ZM00026	consensus	1.331575272	1.73535E-09
2	terminal inverted repeat (TIR)	Cacta (DTC)	ZM00100	consensus	1.318369147	1.75307E-09
2	terminal inverted repeat (TIR)	PIF/Harbinger (DTH)	ZM00071	consensus	2.617576428	4.38279E-09
2	terminal inverted repeat (TIR)	Cacta (DTC)	ZM00010	consensus	1.307497285	5.75234E-09
2	terminal inverted repeat (TIR)	Cacta (DTC)	ZM00002	consensus	1.299418323	8.73902E-09
2	terminal inverted repeat (TIR)	Cacta (DTC)	ZM00078	consensus	1.44546117	9.07447E-09
1	long terminal repeat (LTR)	Unknown LTR (RLX)	milt	AC194936-4356	1.197186482	1.19523E-08
1	long terminal repeat (LTR)	Copia (RLC)	totu	AC194464-4142	1.201631684	3.45909E-08
1	long terminal repeat (LTR)	Gypsy (RLG)	cinful-zeon	AC205118-8623	1.427333087	4.86236E-08
1	long terminal repeat (LTR)	Unknown LTR (RLX)	uluil	AC185306-17	2.018357124	9.13711E-08
2	terminal inverted repeat (TIR)	Cacta (DTC)	ZM00061	consensus	1.292092852	1.06222E-07
1	long terminal repeat (LTR)	Unknown LTR (RLX)	bawigu	AC208532-9902	1.722197471	1.3961E-07
1	long terminal repeat (LTR)	Gypsy (RLG)	fuved	AC204055-8151	1.835137945	1.42612E-07
1	long terminal repeat (LTR)	Copia (RLC)	fipi	AC195215-4479	1.206348011	1.60517E-07
1	long terminal repeat (LTR)	Copia (RLC)	pifo	AC209023-10096	1.357123885	1.72161E-07
1	long terminal repeat (LTR)	Unknown LTR (RLX)	wiru	AC210670-10789	1.58960149	2.15488E-07

2	terminal inverted repeat (TIR)	Cacta (DTC)	ZM00022	consensus	1.239314382	2.2668E-07
2	terminal inverted repeat (TIR)	Cacta (DTC)	ZM00007	consensus	1.261151944	3.15723E-07
1	long terminal repeat (LTR)	Copia (RLC)	tufe	AC211502-11228	1.187956238	3.45392E-07
1	long terminal repeat (LTR)	Copia (RLC)	jelat	AC194217-3960	1.208306797	3.71205E-07
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00293	consensus	1.723174685	5.17825E-07
1	long terminal repeat (LTR)	Unknown LTR (RLX)	milt	AC198975-6250	1.189957873	7.10577E-07
1	long terminal repeat (LTR)	Gypsy (RLG)	bosohe	AC206838-9357	1.202826261	8.92076E-07
1	long terminal repeat (LTR)	Copia (RLC)	giream	AC204000-8116	1.28408841	9.93962E-07
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00366	consensus	1.497325196	1.20297E-06
1	long terminal repeat (LTR)	Copia (RLC)	opie	AC201793-7083	1.167479028	1.27899E-06
2	terminal inverted repeat (TIR)	PIF/Harbinger (DTH)	ZM00045	consensus	1.422241085	1.4996E-06
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00108	consensus	1.560695569	1.75056E-06
1	long terminal repeat (LTR)	Copia (RLC)	ebel	AC213044-12072	1.201813591	2.03829E-06
1	long terminal repeat (LTR)	Copia (RLC)	janoov	AC209690-10294	1.306110195	2.26417E-06
1	long interspersed element (LINE)	L1 (RIL)	etiti	AC211734-0	1.210696767	2.29572E-06
1	long terminal repeat (LTR)	Copia (RLC)	opie	AC185480-1137	1.205515158	2.32955E-06
1	long terminal repeat (LTR)	Gypsy (RLG)	uwew	AC187787-1947	1.495408109	2.57048E-06
2	terminal inverted repeat (TIR)	Cacta (DTC)	ZM00021	consensus	1.237953837	2.73307E-06
1	long terminal repeat (LTR)	Gypsy (RLG)	bosohe	AC185471-1125	1.31773195	2.82639E-06
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00190	consensus	1.205409173	2.95943E-06
1	long terminal repeat (LTR)	Gypsy (RLG)	upus	AC200875-7012	1.545727341	3.32838E-06
1	long terminal repeat (LTR)	Unknown LTR (RLX)	petopi	AC195376-4582	1.377659984	3.38359E-06
1	long terminal repeat (LTR)	Unknown LTR (RLX)	uwub	AC195372-161	1.168461608	3.48649E-06
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00240	consensus	1.390406516	3.49532E-06
1	long terminal repeat (LTR)	Gypsy (RLG)	okopam	AC187789-1948	1.527228507	3.67783E-06
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00102	consensus	1.182549873	4.22043E-06
1	long interspersed element (LINE)	L1 (RIL)	cin4	AC210188-0	1.468417084	4.64812E-06
1	long interspersed element (LINE)	L1 (RIL)	odoif	AC210308-0	1.354649424	5.47589E-06
1	long terminal repeat (LTR)	Copia (RLC)	finaij	AC194312-4026	1.804659442	5.57068E-06
1	long terminal repeat (LTR)	Copia (RLC)	pute	AC197188-5467	1.166539606	5.57731E-06
1	long terminal repeat (LTR)	Unknown LTR (RLX)	anim	AC206032-183	1.230755173	5.72521E-06
2	terminal inverted repeat (TIR)	PIF/Harbinger (DTH)	ZM00280	consensus	1.198902264	6.5447E-06
1	long interspersed element (LINE)	RTE (RIT)	jare	AC204843-0	1.287153558	6.56704E-06
2	terminal inverted repeat (TIR)	PIF/Harbinger (DTH)	ZM00418	consensus	1.139963213	7.11943E-06
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00197	consensus	1.327674636	8.88288E-06
1	long terminal repeat (LTR)	Gypsy (RLG)	ahoru	AC187284-1845	1.267842258	9.14295E-06
1	long terminal repeat (LTR)	Copia (RLC)	raider	AC209705-10304	1.168407578	9.17127E-06
1	long terminal repeat (LTR)	Copia (RLC)	victim	AC182414-456	1.177119656	1.03774E-05
1	long terminal repeat (LTR)	Copia (RLC)	afad	AC199807-6594	1.331416255	1.16288E-05
2	terminal inverted repeat (TIR)	Cacta (DTC)	ZM00003	consensus	1.16737023	1.16322E-05
2	terminal inverted repeat (TIR)	Cacta (DTC)	ZM00059	consensus	1.198075545	1.21977E-05
1	long terminal repeat (LTR)	Unknown LTR (RLX)	kahowu	AC205018-178	2.030959129	1.26983E-05
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00163	consensus	1.364000466	1.33989E-05
1	long terminal repeat (LTR)	Copia (RLC)	ebel	AC210216-10670	1.190402743	1.34587E-05
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00355	consensus	1.249345022	1.37294E-05
2	terminal inverted repeat (TIR)	PIF/Harbinger (DTH)	ZM00351	consensus	1.197704153	1.41845E-05

1	long terminal repeat (LTR)	Copia (RLC)	ajipe	AC183372-580	1.153483309	1.47577E-05
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00239	consensus	1.173091365	1.57472E-05
1	long terminal repeat (LTR)	Unknown LTR (RLX)	bs1	AC208724-10016	1.508404647	1.59538E-05
1	long terminal repeat (LTR)	Gypsy (RLG)	bosohe	AC205330-8724	1.353774438	1.60936E-05
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00209	consensus	1.436470305	1.64069E-05
1	long terminal repeat (LTR)	Copia (RLC)	ebel	AC188777-2128	1.204113472	1.65188E-05
1	long interspersed element (LINE)	L1 (RIL)	leijoh	AC212369-0	1.149433462	1.67127E-05
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00374	consensus	1.234606541	1.68902E-05
1	long terminal repeat (LTR)	Copia (RLC)	ji	AC210731-10832	1.204775757	1.7081E-05
1	long terminal repeat (LTR)	Gypsy (RLG)	kubi	AC196180-5016	1.342953983	2.0072E-05
1	long terminal repeat (LTR)	Copia (RLC)	opie	AC197201-5474	1.159219954	2.02104E-05
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00258	consensus	1.236460805	2.10084E-05
1	long terminal repeat (LTR)	Copia (RLC)	ubat	AC212211-11652	1.183244436	2.1701E-05
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00236	consensus	1.15032233	2.26741E-05
1	long terminal repeat (LTR)	Gypsy (RLG)	epohi	AC205903-9050	1.174224023	2.36312E-05
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00154	consensus	1.505512558	2.42072E-05
1	long terminal repeat (LTR)	Gypsy (RLG)	huck	AC208546-9913	1.199096303	2.50113E-05
1	long terminal repeat (LTR)	Copia (RLC)	opie	AC202033-7274	1.227532451	2.66351E-05
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00281	consensus	1.187023729	2.77262E-05
1	long terminal repeat (LTR)	Copia (RLC)	waneer	AC195414-4606	1.780257343	2.79565E-05
1	long terminal repeat (LTR)	Gypsy (RLG)	olepo	AC212207-11645	1.726844626	2.90667E-05
1	long terminal repeat (LTR)	Unknown LTR (RLX)	bene	AC198379-5969	1.161621396	3.02977E-05
1	long terminal repeat (LTR)	Copia (RLC)	ebel	AC195143-4427	1.189227107	3.04807E-05
2	terminal inverted repeat (TIR)	Mutator (DTM)	Zm01980	AC186668-1	1.202360289	3.30587E-05
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00327	consensus	1.22146568	3.52626E-05
1	short interspersed element (SINE)	tRNA (RST)	AU	consensus-0	1.183404456	3.54311E-05
1	long terminal repeat (LTR)	Unknown LTR (RLX)	nabu	AC187882-57	1.302062404	3.62656E-05
1	long terminal repeat (LTR)	Unknown LTR (RLX)	tuteh	AC183372-584	1.219272604	3.64175E-05
1	long terminal repeat (LTR)	Copia (RLC)	ji	AC192600-3526	1.23926888	3.76503E-05
1	long terminal repeat (LTR)	Unknown LTR (RLX)	ywely	AC190897-98	1.128918712	3.85589E-05
1	long terminal repeat (LTR)	Gypsy (RLG)	oguod	AC209724-10327	1.571166324	3.8698E-05
1	long terminal repeat (LTR)	Gypsy (RLG)	xilon-diguus	AC203313-7774	1.158002548	4.04082E-05
1	long interspersed element (LINE)	L1 (RIL)	totyru	AC203014-0	1.369061517	4.48333E-05
1	long terminal repeat (LTR)	Gypsy (RLG)	puck	AC215312-13067	1.187751827	4.50497E-05
1	long terminal repeat (LTR)	Unknown LTR (RLX)	mibaab	AC205139-8652	1.463110593	4.73578E-05
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00226	consensus	1.287392093	5.20751E-05
1	long terminal repeat (LTR)	Copia (RLC)	opie	AC188002-2029	1.177045279	5.87171E-05
2	terminal inverted repeat (TIR)	Cacta (DTC)	ZM00098	consensus	1.155418886	6.02381E-05
1	long terminal repeat (LTR)	Copia (RLC)	opie	AC187149-1780	1.170658486	6.3201E-05
1	long terminal repeat (LTR)	Gypsy (RLG)	laiwa	AC214288-12602	1.331063842	6.61519E-05
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00103	consensus	1.245440465	6.96918E-05
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00385	consensus	1.229047166	7.42446E-05
2	terminal inverted repeat (TIR)	Cacta (DTC)	ZM00030	consensus	1.168925428	7.45431E-05
2	terminal inverted repeat (TIR)	Cacta (DTC)	ZM00039	consensus	1.325127615	7.60195E-05
1	long terminal repeat (LTR)	Copia (RLC)	kuvi	AC207313-9512	1.326171975	7.63558E-05
2	terminal inverted repeat (TIR)	Mutator (DTM)	Zm00884	AC214130-1	1.471614067	7.76692E-05

1	long terminal repeat (LTR)	Copia (RLC)	victim	AC183319-577	1.16710442	8.04186E-05
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00280	consensus	1.157189204	8.77419E-05
1	long terminal repeat (LTR)	Unknown LTR (RLX)	fanuab	AC193594-3712	1.530821901	8.94249E-05
1	long terminal repeat (LTR)	Gypsy (RLG)	ytub	AC187411-1880	1.661749714	8.97153E-05
2	terminal inverted repeat (TIR)	PIF/Harbinger (DTH)	ZM00022	consensus	1.310567013	9.07702E-05
1	long terminal repeat (LTR)	Gypsy (RLG)	huck	AC210804-10865	1.16847531	9.2985E-05
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00288	consensus	1.17444791	9.47101E-05
1	long terminal repeat (LTR)	Gypsy (RLG)	bygum	AC188125-2053	1.161081103	9.99515E-05
1	long terminal repeat (LTR)	Gypsy (RLG)	huck	AC214833-12913	1.179272234	0.00010106
1	long terminal repeat (LTR)	Copia (RLC)	machiavelli	AC200490-6883	1.18962389	0.000105221
1	long terminal repeat (LTR)	Unknown LTR (RLX)	kaise	AC203928-8059	1.237437047	0.000106353
1	long terminal repeat (LTR)	Copia (RLC)	homy	AC197914-5822	1.222175721	0.000111136
1	long terminal repeat (LTR)	Copia (RLC)	ebel	AC211737-11397	1.212404669	0.000112236
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00125	consensus	1.493750443	0.000114343
1	long terminal repeat (LTR)	Gypsy (RLG)	ewib	AC198384-5975	1.132112899	0.00011635
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00111	consensus	1.258457129	0.000118024
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00354	consensus	1.281600505	0.00012019
1	long terminal repeat (LTR)	Gypsy (RLG)	fege	AC205532-8884	1.294714924	0.000123042
2	terminal inverted repeat (TIR)	Mutator (DTM)	Zm02785	AC190936-1	1.151478194	0.000125455
1	long terminal repeat (LTR)	Copia (RLC)	ajajog	AC191578-3186	1.267843052	0.000126989
1	long terminal repeat (LTR)	Unknown LTR (RLX)	fate	AC194466-4144	1.427303039	0.00013036
1	long terminal repeat (LTR)	Gypsy (RLG)	loba	AC194942-4364	1.621264618	0.00014172
2	terminal inverted repeat (TIR)	PIF/Harbinger (DTH)	ZM00453	consensus	1.230725228	0.000143505
1	long terminal repeat (LTR)	Gypsy (RLG)	ywyt	AC209975-10517	1.163691838	0.000148307
1	long terminal repeat (LTR)	Gypsy (RLG)	anysaf	AC203052-7637	1.409428933	0.000149207
1	long terminal repeat (LTR)	Gypsy (RLG)	bosohe	AC191654-3248	1.247276084	0.000153445
1	long terminal repeat (LTR)	Gypsy (RLG)	huck	AC208842-10038	1.17772699	0.000156317
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00003	consensus	1.405369773	0.000158305
1	long terminal repeat (LTR)	Copia (RLC)	raider	AC197426-5583	1.178774239	0.000161473
1	long terminal repeat (LTR)	Gypsy (RLG)	huck	AC213612-12218	1.168607957	0.000165775
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00356	consensus	1.144010703	0.000165892
1	long terminal repeat (LTR)	Copia (RLC)	ji	AC190978-2799	1.153264122	0.000192274
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00253	consensus	1.277998571	0.000193556
1	long terminal repeat (LTR)	Unknown LTR (RLX)	lamyab	AC208713-10008	1.167227783	0.000195284
2	terminal inverted repeat (TIR)	Mutator (DTM)	Zm00800	consensus	1.18507087	0.000203706
1	long terminal repeat (LTR)	Gypsy (RLG)	doke	AC197224-5479	1.16109018	0.000207128
2	terminal inverted repeat (TIR)	Cacta (DTC)	ZM00085	consensus	1.148309454	0.00022264
1	long terminal repeat (LTR)	Gypsy (RLG)	grande	AC190611-2432	1.129965672	0.000230591
1	long terminal repeat (LTR)	Gypsy (RLG)	grande	AC200214-6803	1.130988069	0.00023096
1	long terminal repeat (LTR)	Gypsy (RLG)	puck	AC208456-9876	1.153396492	0.000232906
2	terminal inverted repeat (TIR)	Cacta (DTC)	ZM00005	consensus	1.327647451	0.00024019
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00148	consensus	1.159337659	0.000244967
2	terminal inverted repeat (TIR)	Mutator (DTM)	Zm17242	AC197682-1	1.330532568	0.000249455
1	long interspersed element (LINE)	L1 (RIL)	edaej	AC215611-0	1.206186699	0.000255332
2	Helitron	Helitron	Hip2	2	1.118399059	0.000257609
1	long terminal repeat (LTR)	Unknown LTR (RLX)	ruda	AC195952-4839	1.184340928	0.000275128

1	long terminal repeat (LTR)	Copia (RLC)	dijap	AC211466-11198	1.125452536	0.000278657
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00259	consensus	1.1897262	0.000284918
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00265	consensus	1.149966527	0.000285863
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00271	consensus	1.169222632	0.000303004
1	long terminal repeat (LTR)	Gypsy (RLG)	dagaf	AC195302-4533	1.130039229	0.000312878
1	long terminal repeat (LTR)	Gypsy (RLG)	ivuk	AC194103-3869	1.161337048	0.000341447
1	long terminal repeat (LTR)	Gypsy (RLG)	sokiit	AC210743-10848	1.290909982	0.000344463
2	terminal inverted repeat (TIR)	PIF/Harbinger (DTH)	ZM00012	consensus	1.263864833	0.000346087
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00384	consensus	1.209150531	0.00038024
1	long terminal repeat (LTR)	Gypsy (RLG)	dagaf	AC182835-556	1.114005474	0.000381475
2	terminal inverted repeat (TIR)	Mutator (DTM)	Zm02656	AC189800-1	1.246591676	0.000384491
1	long terminal repeat (LTR)	Gypsy (RLG)	puck	AC211927-11478	1.138105005	0.000385312
1	long terminal repeat (LTR)	Copia (RLC)	hera	AC214536-12775	1.23311791	0.000389209
2	terminal inverted repeat (TIR)	PIF/Harbinger (DTH)	ZM00246	consensus	1.233351992	0.000392743
1	long terminal repeat (LTR)	Unknown LTR (RLX)	vafim	AC196676-5176	1.316973037	0.000407839
1	long terminal repeat (LTR)	Copia (RLC)	opie	AC202020-7258	1.177442598	0.000409517
1	long terminal repeat (LTR)	Unknown LTR (RLX)	wuywu	AC190718-2536	1.376379344	0.000414415
2	terminal inverted repeat (TIR)	Cacta (DTC)	ZM00066	consensus	1.131996007	0.000425674
2	terminal inverted repeat (TIR)	PIF/Harbinger (DTH)	ZM00025	consensus	1.145868165	0.000428104
1	long terminal repeat (LTR)	Gypsy (RLG)	huck	AC194973-4393	1.176173172	0.000433169
1	long terminal repeat (LTR)	Unknown LTR (RLX)	ifab	AC209906-10473	1.327459618	0.000448654
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00246	consensus	1.283029675	0.000462141
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00115	consensus	1.124020996	0.000467787
1	long terminal repeat (LTR)	Gypsy (RLG)	bosohe	AC190752-2561	1.277823751	0.000477935
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00076	consensus	1.161487976	0.00047858
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00235	consensus	1.117035852	0.000496042
1	long terminal repeat (LTR)	Copia (RLC)	ogiv	AC205856-9034	1.36469704	0.000506689
1	long terminal repeat (LTR)	Copia (RLC)	wamenu	AC191287-3028	1.19319201	0.000517846
1	long terminal repeat (LTR)	Copia (RLC)	anar	AC206985-9422	1.167551207	0.000526048
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00159	consensus	1.228561072	0.000542095
1	long terminal repeat (LTR)	Copia (RLC)	doke	AC186158-1307	1.151375846	0.000550249
1	long terminal repeat (LTR)	Gypsy (RLG)	huck	AC212331-11708	1.171626837	0.000567832
1	long terminal repeat (LTR)	Gypsy (RLG)	udav	AC196188-5022	1.146624171	0.00059633
1	long terminal repeat (LTR)	Gypsy (RLG)	puck	AC208673-9982	1.148623168	0.00061055
1	long terminal repeat (LTR)	Gypsy (RLG)	huck	AC210079-10574	1.181125833	0.000626304
2	terminal inverted repeat (TIR)	Cacta (DTC)	ZM00089	consensus	1.161016754	0.000628503
1	long terminal repeat (LTR)	Gypsy (RLG)	bosohe	AC212057-11553	1.1899588	0.000631009
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00341	consensus	1.123604564	0.000659799
2	terminal inverted repeat (TIR)	Mutator (DTM)	Zm10271	AC186904-1	1.274952222	0.000698142
1	long terminal repeat (LTR)	Gypsy (RLG)	huck	AC199418-6452	1.174637059	0.000701617
2	terminal inverted repeat (TIR)	PIF/Harbinger (DTH)	ZM00124	consensus	1.239294287	0.000717287
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00329	consensus	1.274071261	0.000734349
1	long terminal repeat (LTR)	Unknown LTR (RLX)	jeli	AC200611-6933	1.348883247	0.000767086
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00112	consensus	1.552114492	0.00080943
2	terminal inverted repeat (TIR)	Cacta (DTC)	ZM00102	consensus	1.108534056	0.000852479
1	long terminal repeat (LTR)	Unknown LTR (RLX)	mafogo	AC199961-6695	1.296581067	0.00085971

2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00245	consensus	1.164520762	0.000861939
2	terminal inverted repeat (TIR)	PIF/Harbinger (DTH)	ZM00038	consensus	1.232419699	0.000862465
2	terminal inverted repeat (TIR)	Cacta (DTC)	ZM00031	consensus	1.167672211	0.000872494
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00298	consensus	1.163876498	0.000875643
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00022	consensus	1.190277805	0.000965648
1	long terminal repeat (LTR)	Copia (RLC)	opie	AC198173-5898	1.113903396	0.000978396
2	terminal inverted repeat (TIR)	PIF/Harbinger (DTH)	ZM00050	consensus	1.18177874	0.001026708
1	long terminal repeat (LTR)	Gypsy (RLG)	lyruom	AC185669-1267	1.234370769	0.001103976
1	long terminal repeat (LTR)	Copia (RLC)	hani	AC186285-1359	1.191721144	0.001115826
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00382	consensus	1.195681459	0.001116742
1	long terminal repeat (LTR)	Unknown LTR (RLX)	vufe	AC194263-159	1.362624579	0.001120154
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00138	consensus	1.197314901	0.001123225
1	long interspersed element (LINE)	L1 (RIL)	jikeuf	AC211152-0	1.15690779	0.001150821
1	long terminal repeat (LTR)	Unknown LTR (RLX)	milt	AC209648-10275	1.117892297	0.001152723
1	long terminal repeat (LTR)	Gypsy (RLG)	dagaf	AC208646-9966	1.112303907	0.001163677
1	long terminal repeat (LTR)	Unknown LTR (RLX)	fajy	AC190874-2682	1.120576673	0.001166363
1	long terminal repeat (LTR)	Unknown LTR (RLX)	daju	AC190492-79	1.256924502	0.001186603
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00059	consensus	1.124494629	0.001208503
1	long terminal repeat (LTR)	Gypsy (RLG)	labu	AC188126-2057	1.452173276	0.00125006
2	Helitron	Helitron	Hip1	26	1.139396972	0.001253962
1	long terminal repeat (LTR)	Gypsy (RLG)	CRM4	AC201761-7053	1.113829423	0.00125692
1	long terminal repeat (LTR)	Gypsy (RLG)	huck	AC216048-13250	1.161179233	0.001272728
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00213	consensus	1.256359501	0.001318996
2	terminal inverted repeat (TIR)	Mutator (DTM)	Zm00555	consensus	1.173251085	0.001328993
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00095	consensus	1.362325701	0.001343722
2	terminal inverted repeat (TIR)	PIF/Harbinger (DTH)	ZM00005	consensus	1.177643867	0.001350215
1	long terminal repeat (LTR)	Unknown LTR (RLX)	panen	AC192606-115	1.141862446	0.001394678
2	terminal inverted repeat (TIR)	Cacta (DTC)	ZM00045	consensus	1.215448541	0.001403994
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00030	consensus	1.190897072	0.001418274
2	terminal inverted repeat (TIR)	Cacta (DTC)	ZM00012	consensus	1.109185009	0.001471474
1	long terminal repeat (LTR)	Copia (RLC)	opie	AC196469-5133	1.125885599	0.00147654
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00162	consensus	1.312766482	0.001488195
1	long terminal repeat (LTR)	Copia (RLC)	ji	AC193479-3665	1.113849631	0.00155333
2	terminal inverted repeat (TIR)	PIF/Harbinger (DTH)	ZM00018	consensus	1.233598497	0.001575389
2	terminal inverted repeat (TIR)	hAT (DTA)	ZM00362	consensus	1.221277537	0.00164852
2	terminal inverted repeat (TIR)	Mutator (DTM)	Zm26908	AC183312-1	1.441490678	0.001688853

Supplementary Table 5. Population genetic summary statistics for domestication and improvement features in *parviglumis*, landraces, and improved lines. ρ and π values are reported per bp. Abbreviations are as in Supplementary Table 2.

Domestication

Statistic	*parviglumis*	landrace	improved
π	0.005	0.002	0.002
TajD	-0.022	-1.075	-1.032
H'	2.967	1.435	1.426
ρ	0.013	0.003	0.002

Improvement

Statistic	*parviglumis*	landrace	improved
π	0.005	0.004	0.003
TajD	-0.105	-0.388	-0.799
H'	2.844	2.299	1.657
ρ	0.013	0.003	0.002

Supplementary Table 6. Domestication candidates and genes falling within selected features. Features are ranked from highest to lowest evidence for selection during domestication. Table available as an excel file in online supplementary information.

Supplementary Table 7. Improvement candidates and genes falling within selected features. Features are ranked from highest to lowest evidence for selection during improvement. Table available as an excel file in online supplementary information.

Supplementary Table 8. Percentage of genes in tissue-specific expression categories based on the maize expression atlas[40].

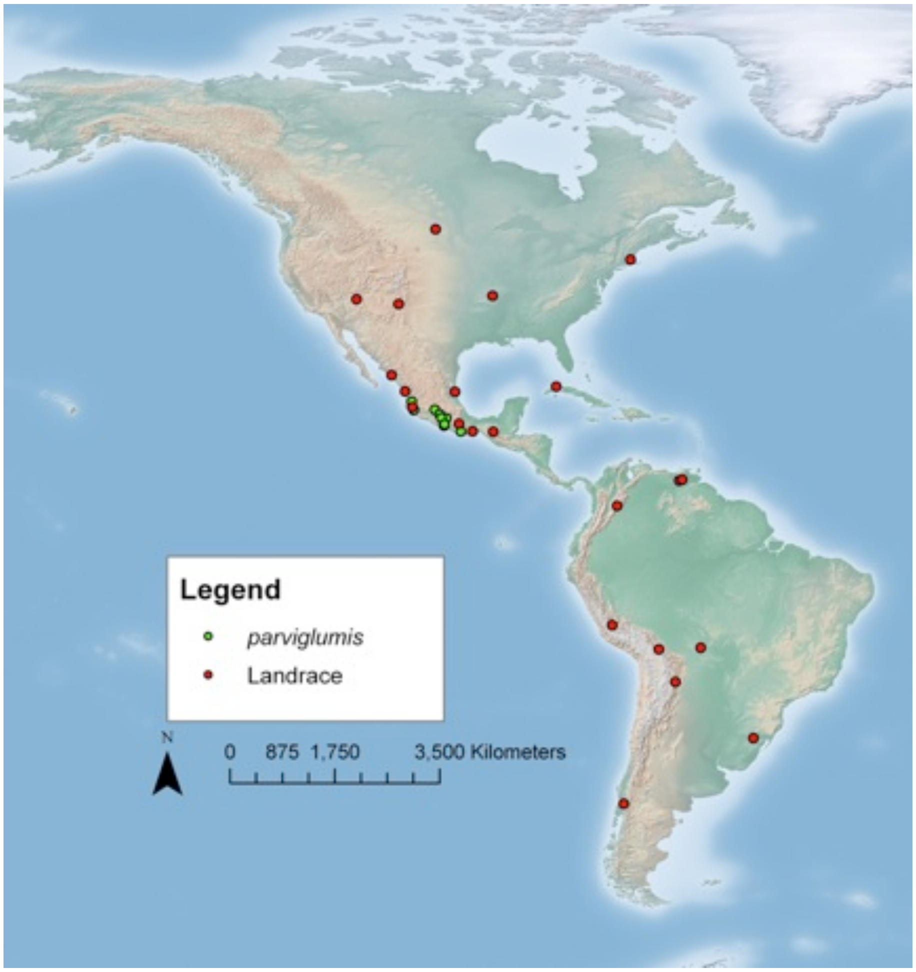
Expression Profile	All FGS Genes	Wright *et al.* 2005 Overall	Wright *et al.* 2005 Candidate	Domestication Candidates	Improvement Candidates
Absent	10.24	0.33	0	8.85	10.75
Constitutive	41.32	68.9	73.91	42.18	43.93
Mixed	45.48	29.93	26.09	45.72	43.46
cob	0.01	0	0	0	0
Embryo	0.17	0	0	0.59	0.47
Endosperm	0.53	0.17	0	0.29	0.47
Internode	0.03	0	0	0	0.23
Leaf	1.16	0.33	0	0.88	0.47
Pericarp	0.02	0	0	0	0
Root	0.54	0.33	0	0.59	0
Silk	0.05	0	0	0.29	0
Tassel	0.4	0	0	0.59	0.23
Tissue Specific Overall	2.96	0.84	0	3.24	1.87

Supplementary Table 9. Previously identified domestication genes and status in current scan.

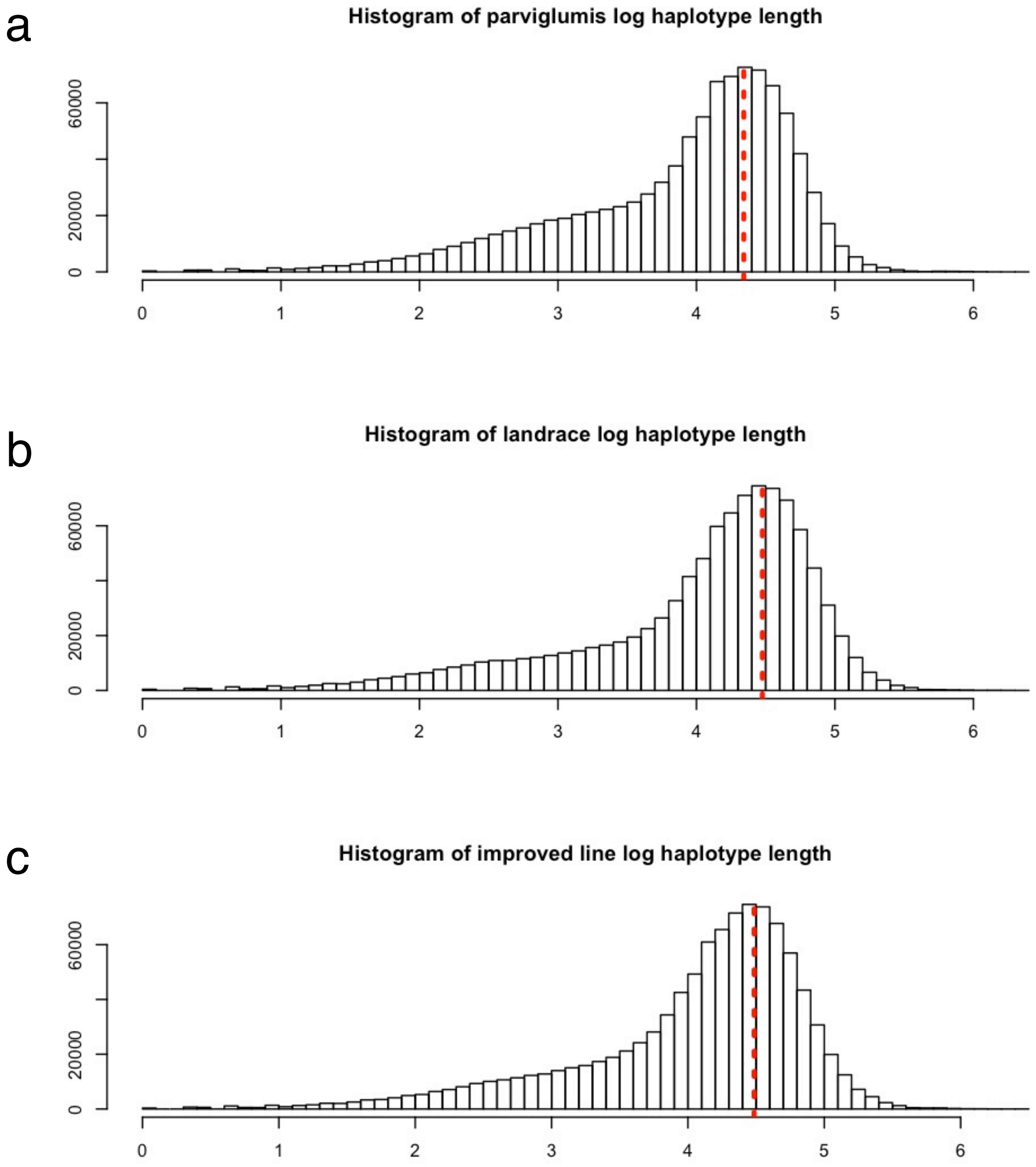
FGS Gene ID	Gene Symbol	Gene Name	candidate	in candidate feature	XP-CLR feature percentile	reference
AC233950.1_FG002	*tb1*	*teosinte branched1*	yes*	yes*	95th	22
GRMZM2G101511	*tga1*	*teosinte glume architecture1*	no	yes	93rd	21
GRMZM2G003927	*ra1*	*ramosa1*	no	yes	96th	46
GRMZM2G138060	*su1*	*sugary1*	no	yes	97th	47
GRMZM2G146283	*pbf1*	*prolamine-box binding factor1*	no	yes	97th	48-50
GRMZM2G160687	*zag2*	*zea agamous homolog2*	yes	yes	99th	25
GRMZM2G052890	*zag1*	*zea agamous-like1*	yes	yes	94th	25
GRMZM2G171822	*bif2*	*barren inflorescence2*	yes	yes	98th	26-28
GRMZM2G005624	*gt1*	*grassy tillers1*	no	no	not in feature	51
GRMZM2G397518	*ba1*	*barren stalk1*	no	yes	99th	52
GRMZM2G068506	*bt2*	*brittle endosperm2*	no	yes	99th	47
GRMZM2G180190	*zfl2*	*Zea floricaula leafy2*	yes	yes	95th	29
GRMZM2G032628	*ae1*	*amylose extender1*	no	no	not in feature	47

* not a candidate in preliminary scan, but candidate after introgressed *parviglumis* with maize allele removed

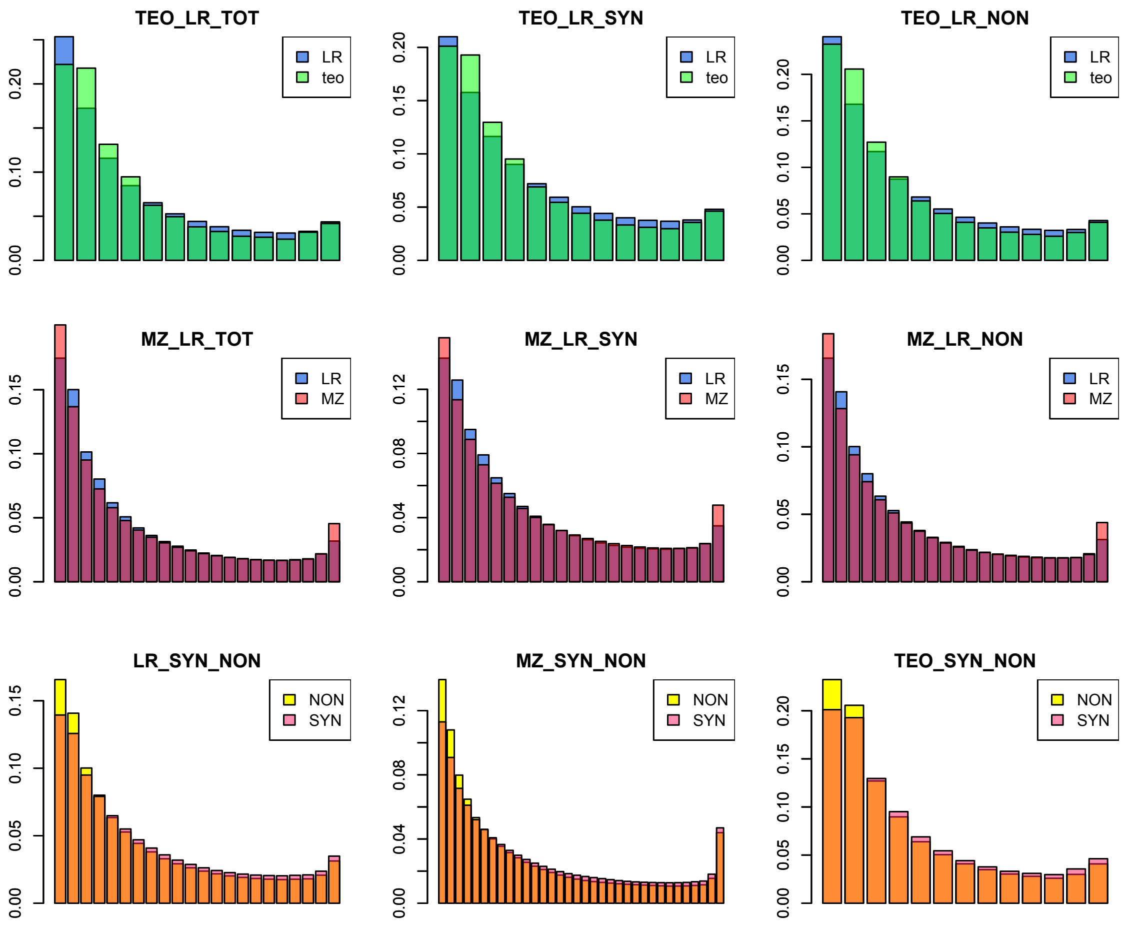
Supplementary Figures:



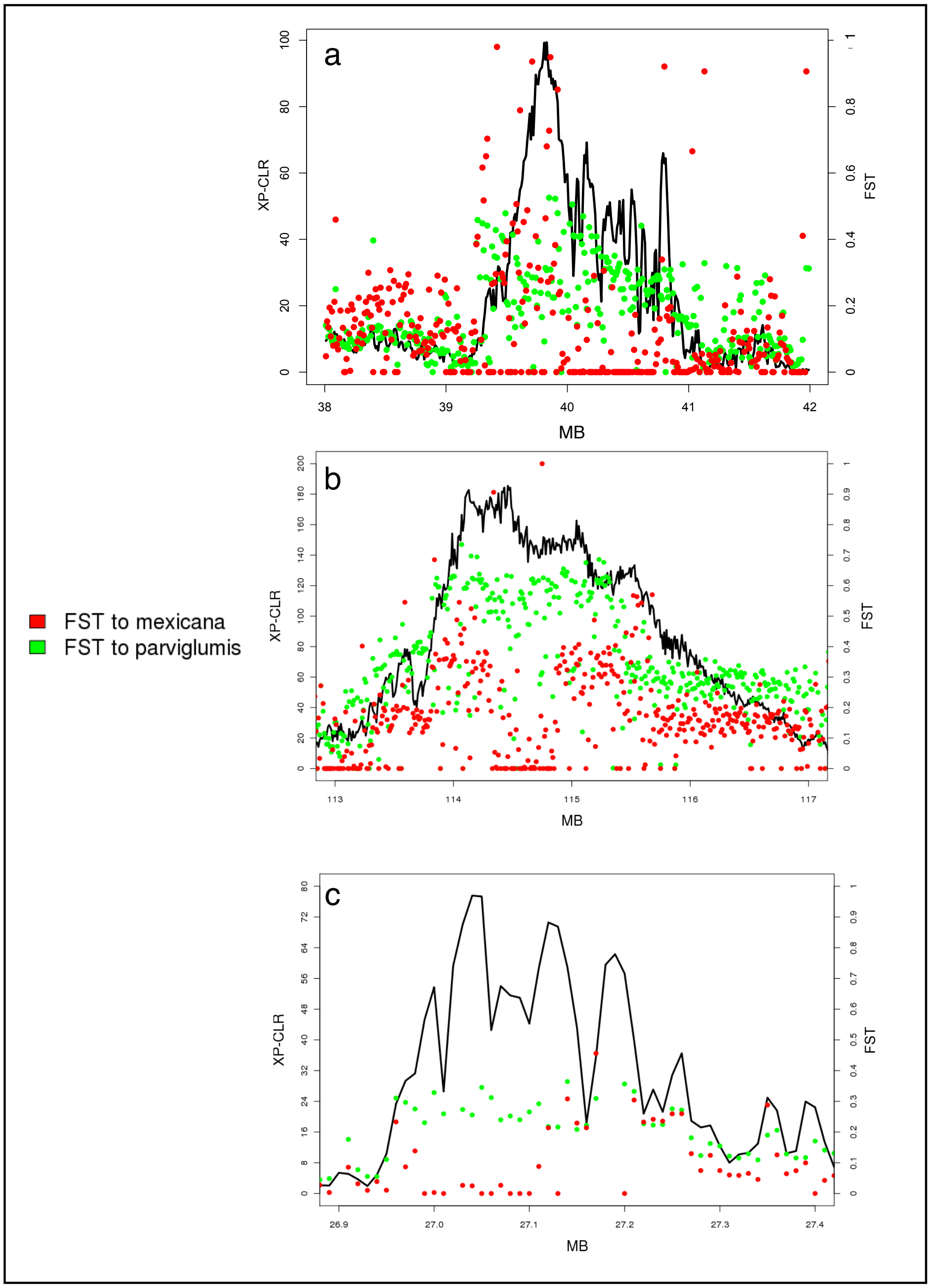
Supplementary Figure 1. Map of the Americas with sequenced landrace lines in red and sequenced *parviglumis* lines in green.



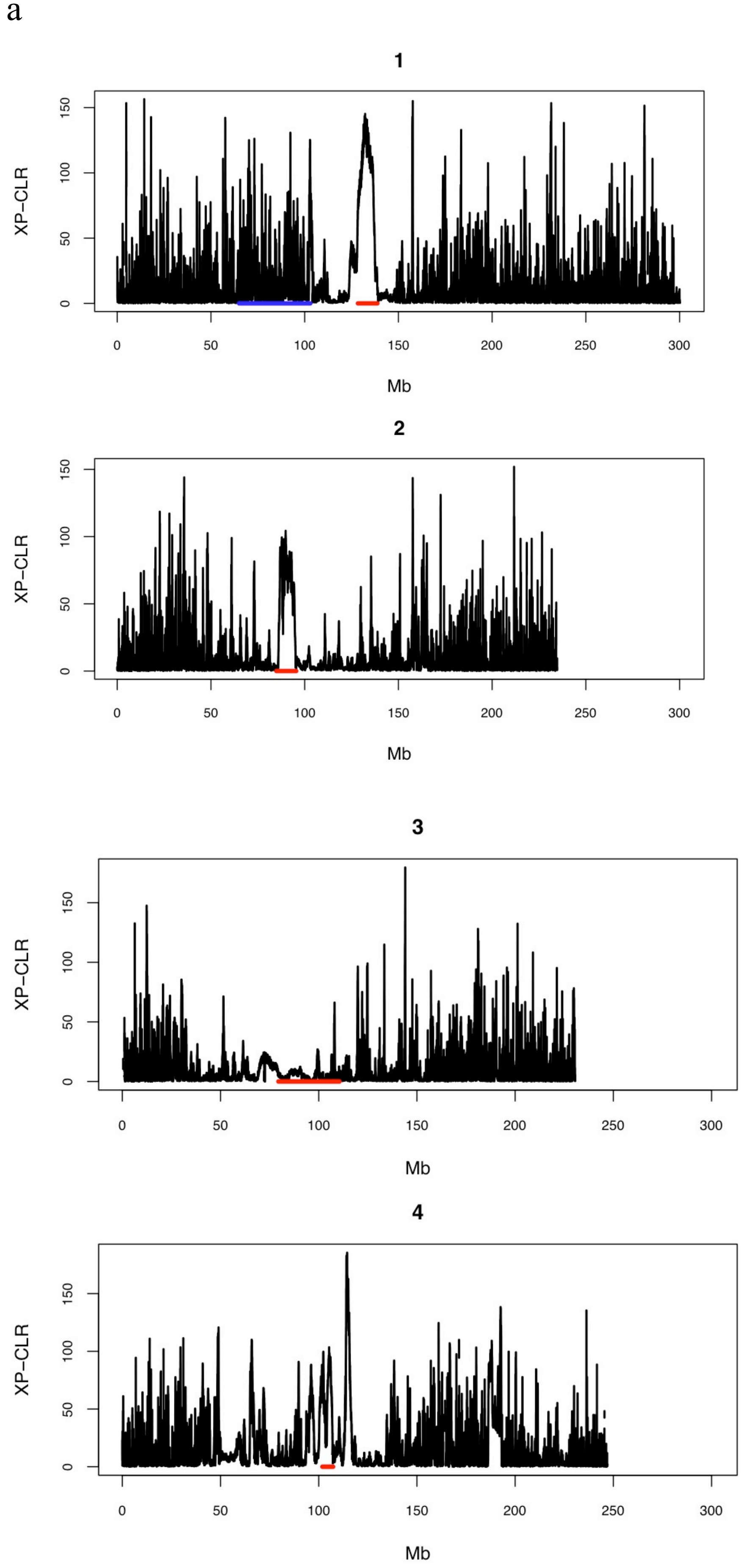
Supplementary Figure 2. Distribution of haplotype lengths across taxa. Histograms of the log10 length of haplotypes in *parviglumis* (a), landraces (b), and improved lines (c) based on estimates from one million random sites in each taxon.

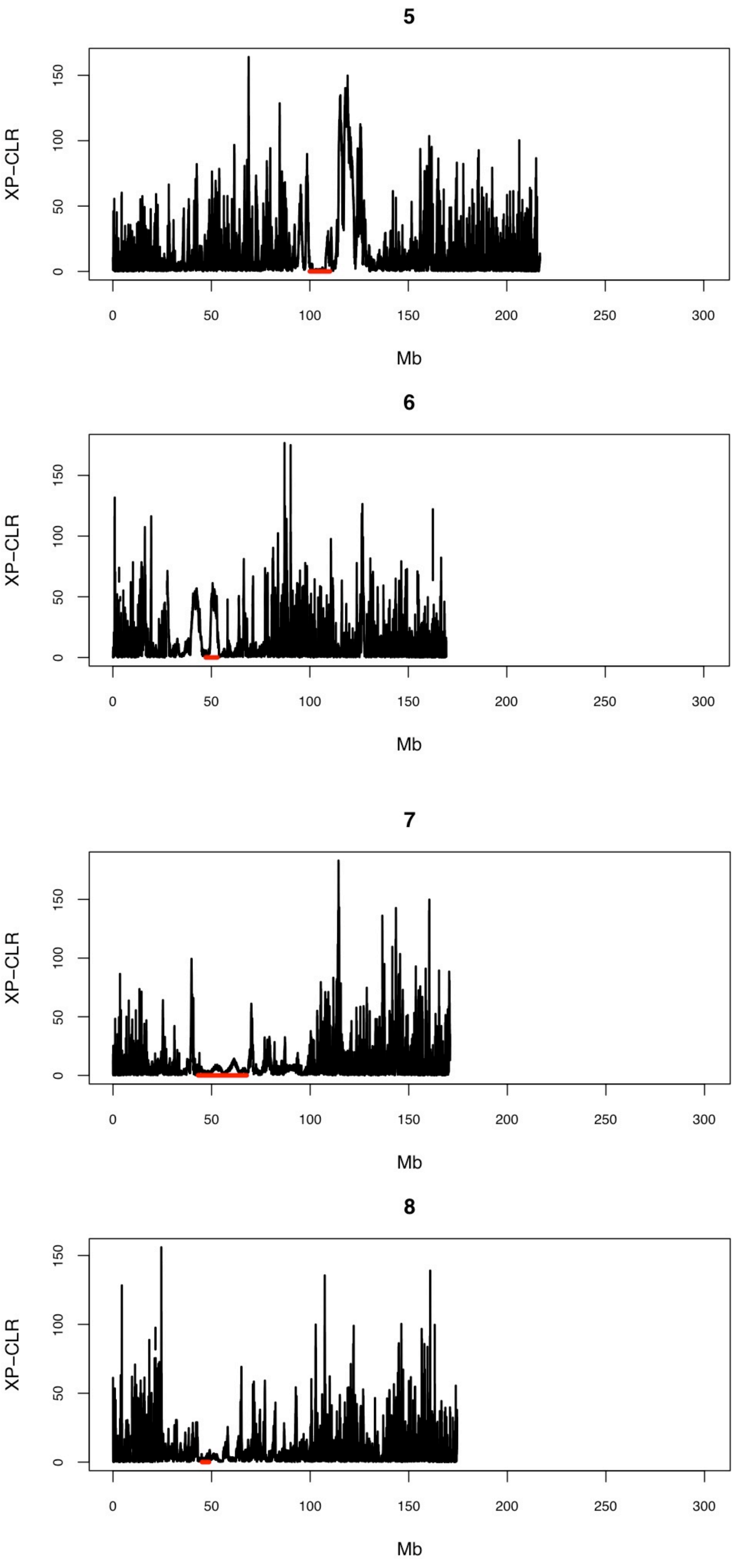


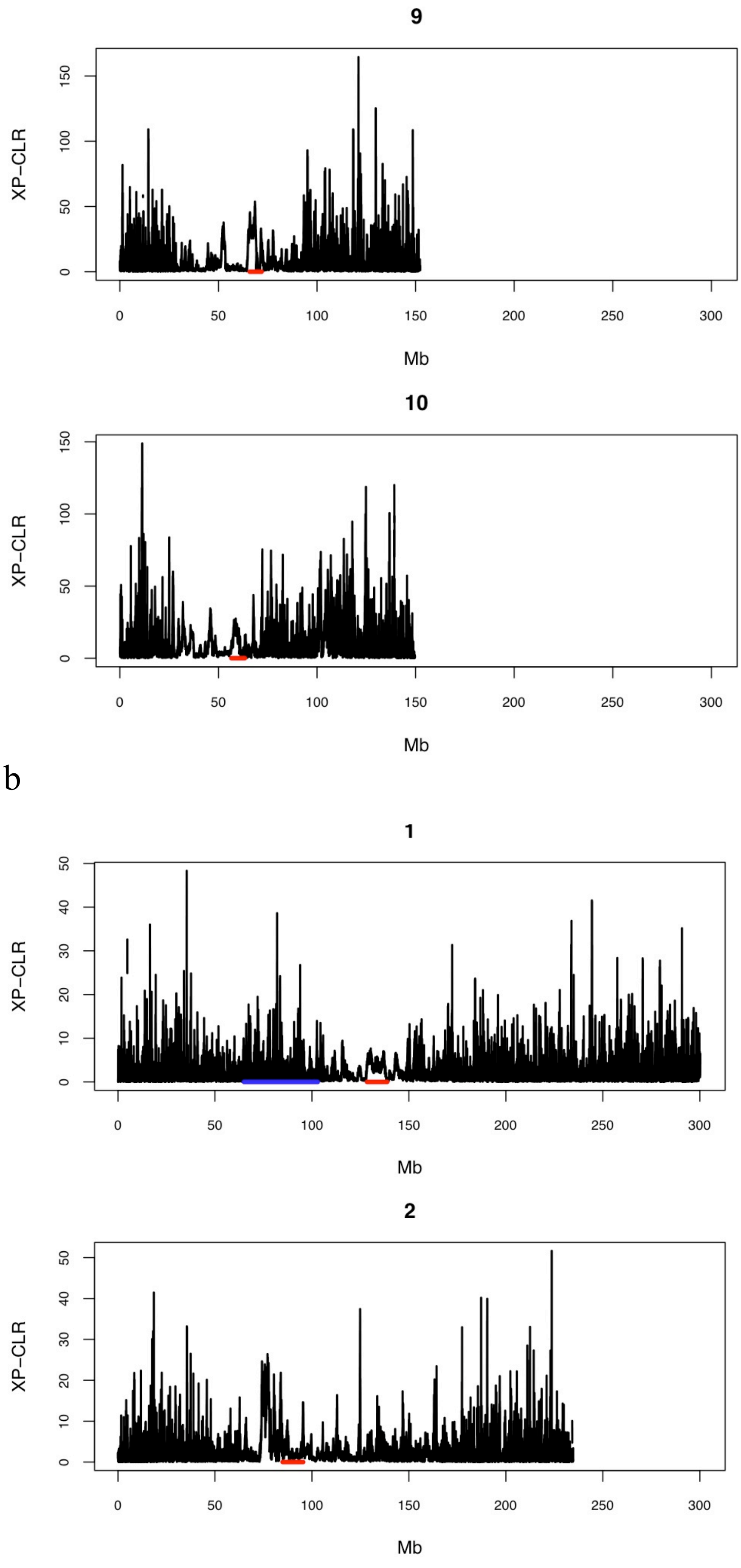
Supplementary Figure 3. Synonymous and non-synonymous unfolded site-frequency spectra; determination of derived state used *Tripsacum* as an outgroup. LR: landraces, MZ: improved lines, TEO: *parviglumis*, NON: non-synonymous sites, SYN: synonymous sites, TOT: all SNPs. The number of classes depicted for LR, MZ, and TEO are 22, 34, and 13.

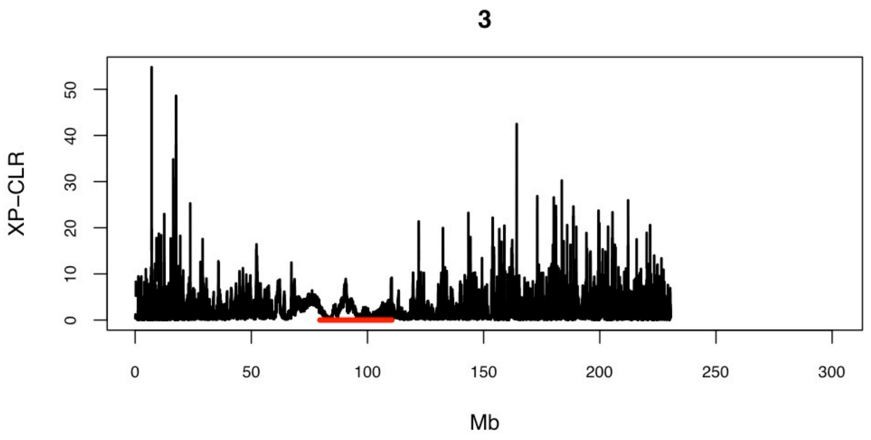


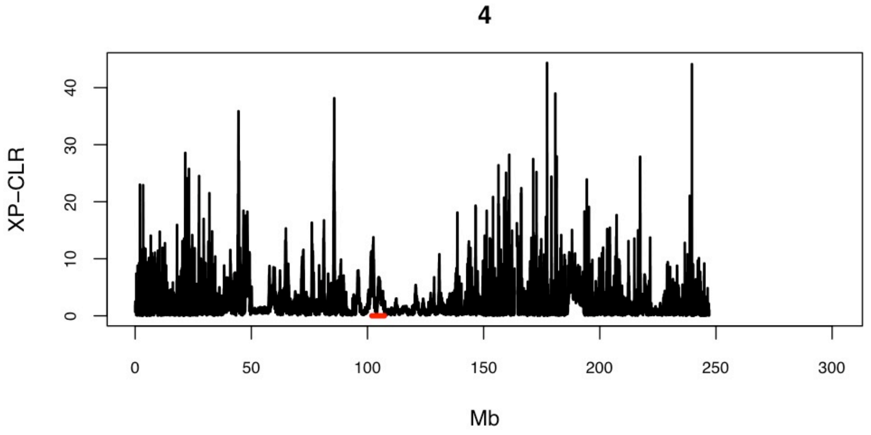
Supplementary Figure 4. Evidence for introgression of *mexicana* germplasm into landraces. These features show elevated XP-CLR for the domestication scan (in black), low F_{ST} of landraces (≤0) relative to *mexicana* (in red) and elevated F_{ST} (≥90th quantile) of landraces relative to *parviglumis* (in green). (a) A 2Mb region of putative introgression on chromosome 7 that falls within the largest QTL for tassel branch number found in a study of highland and lowland maize[53]. (b) A 600kb region on chromosome 4 that occurs within a feature found to have the highest signal of XP-CLR in the genome in the domestication scan and includes a gene (GRMZM2G059167) orthologous to a MYB transcription factor in *Arabidopsis thaliana*. (c) A 100kb region of introgression on chromosome 4.

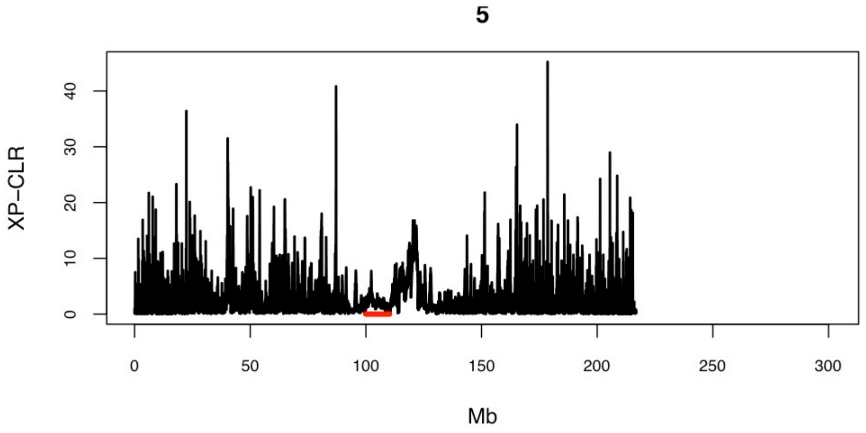

a
1
XP−CLR
0
50
100
150
0
50
100
150
200
250
300
Mb
2
XP−CLR
0
50
100
150
0
50
100
150
200
250
300
Mb
3
XP−CLR
0
50
100
150
0
50
100
150
200
250
300
Mb
4
XP−CLR
0
50
100
150
0
50
100
150
200
250
300
Mb

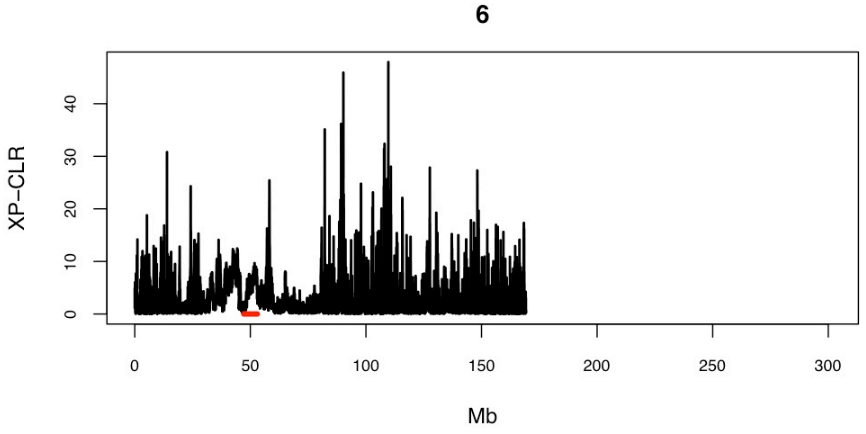

5
6
7
8
XP−CLR
0
50
100
150
Mb
0
50
100
150
200
250
300

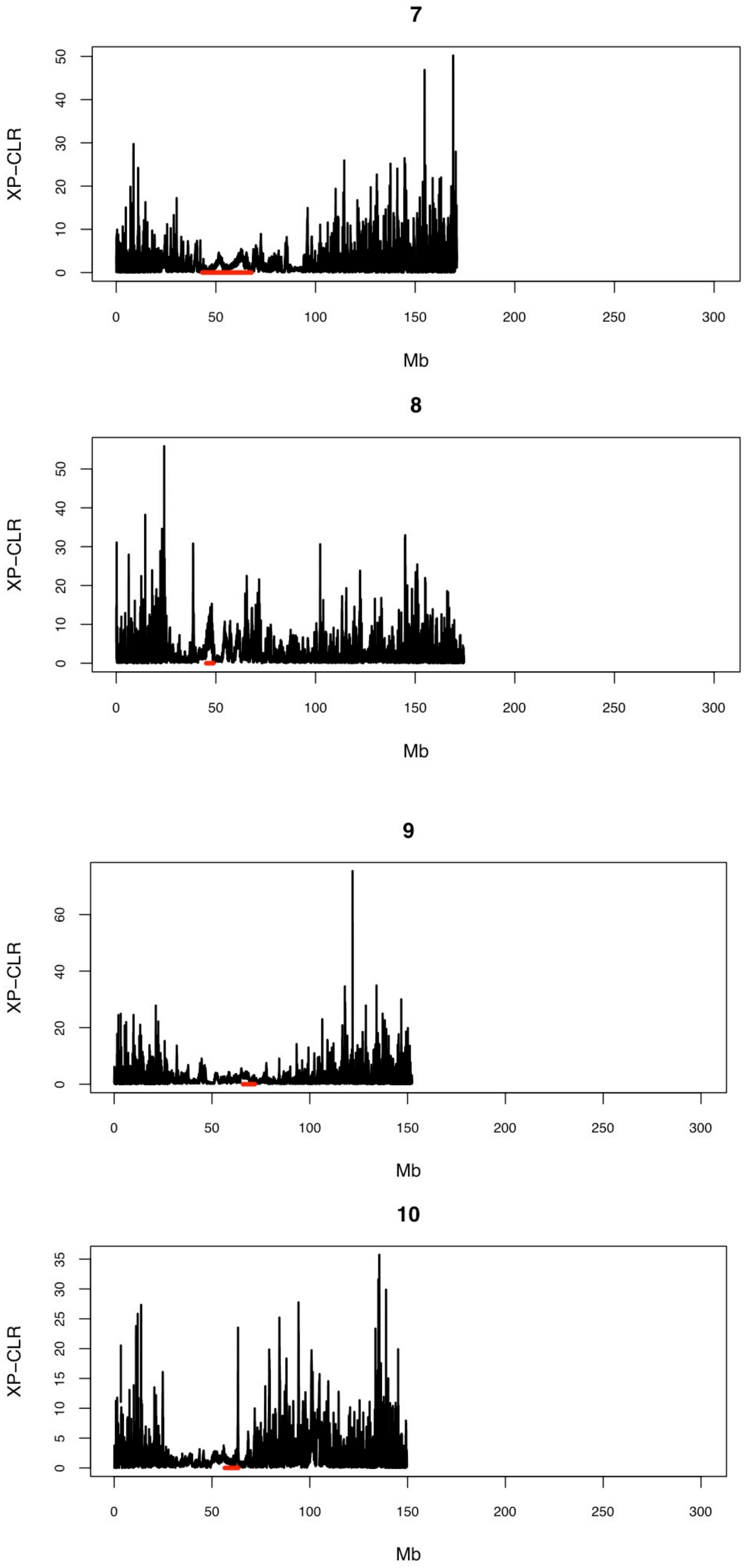

9
XP−CLR
0
50
100
150
0
50
100
150
200
250
300
Mb
10
XP−CLR
0
50
100
150
0
50
100
150
200
250
300
Mb
b
1
XP−CLR
0
10
20
30
40
50
0
50
100
150
200
250
300
Mb
2
XP−CLR
0
10
20
30
40
50
0
50
100
150
200
250
300
Mb


3
XP−CLR
0
10
20
30
40
50
0
50
100
150
200
250
300
Mb

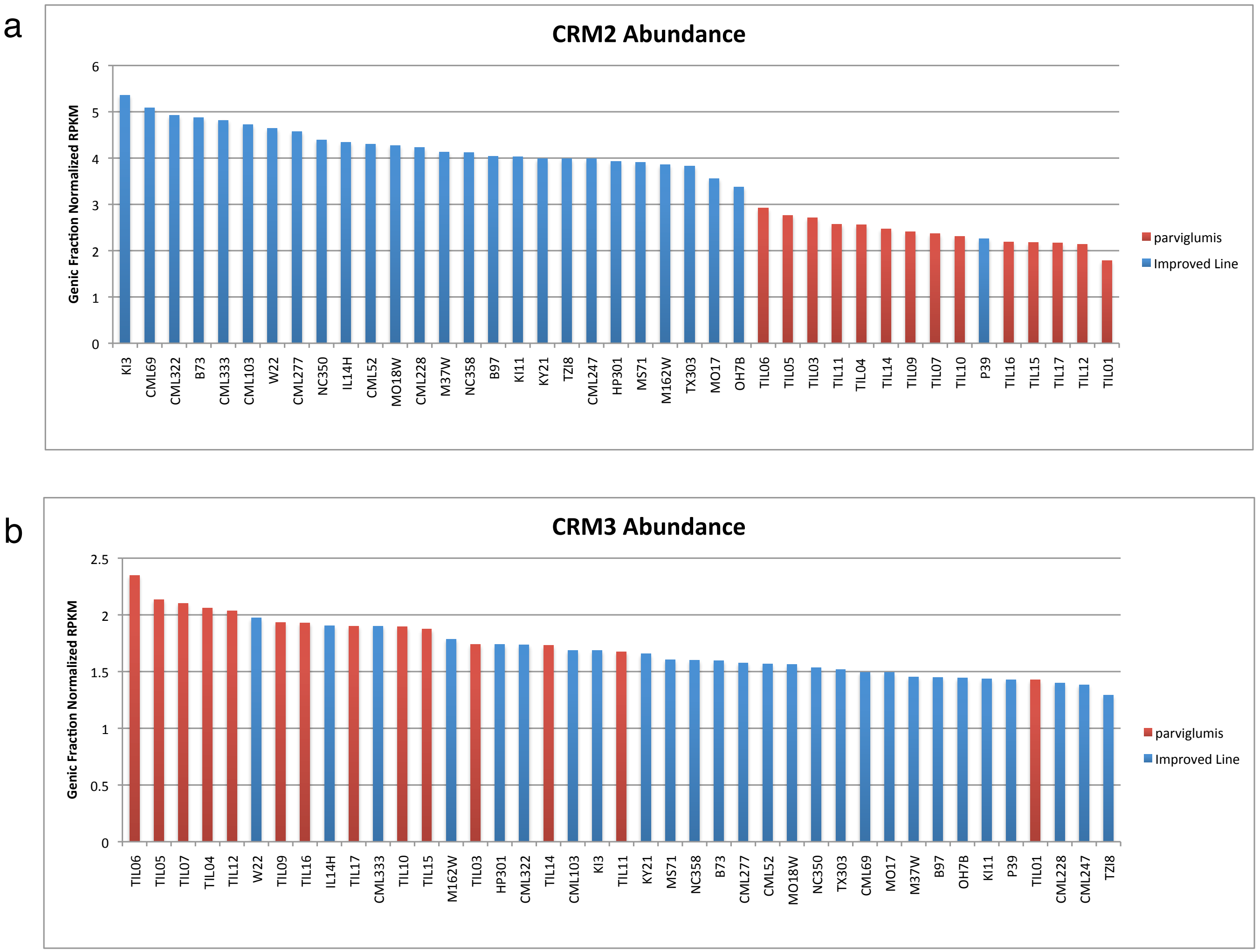

4
XP−CLR
0
10
20
30
40
0
50
100
150
200
250
300
Mb


5
XP−CLR
0
10
20
30
40
0
50
100
150
200
250
300
Mb

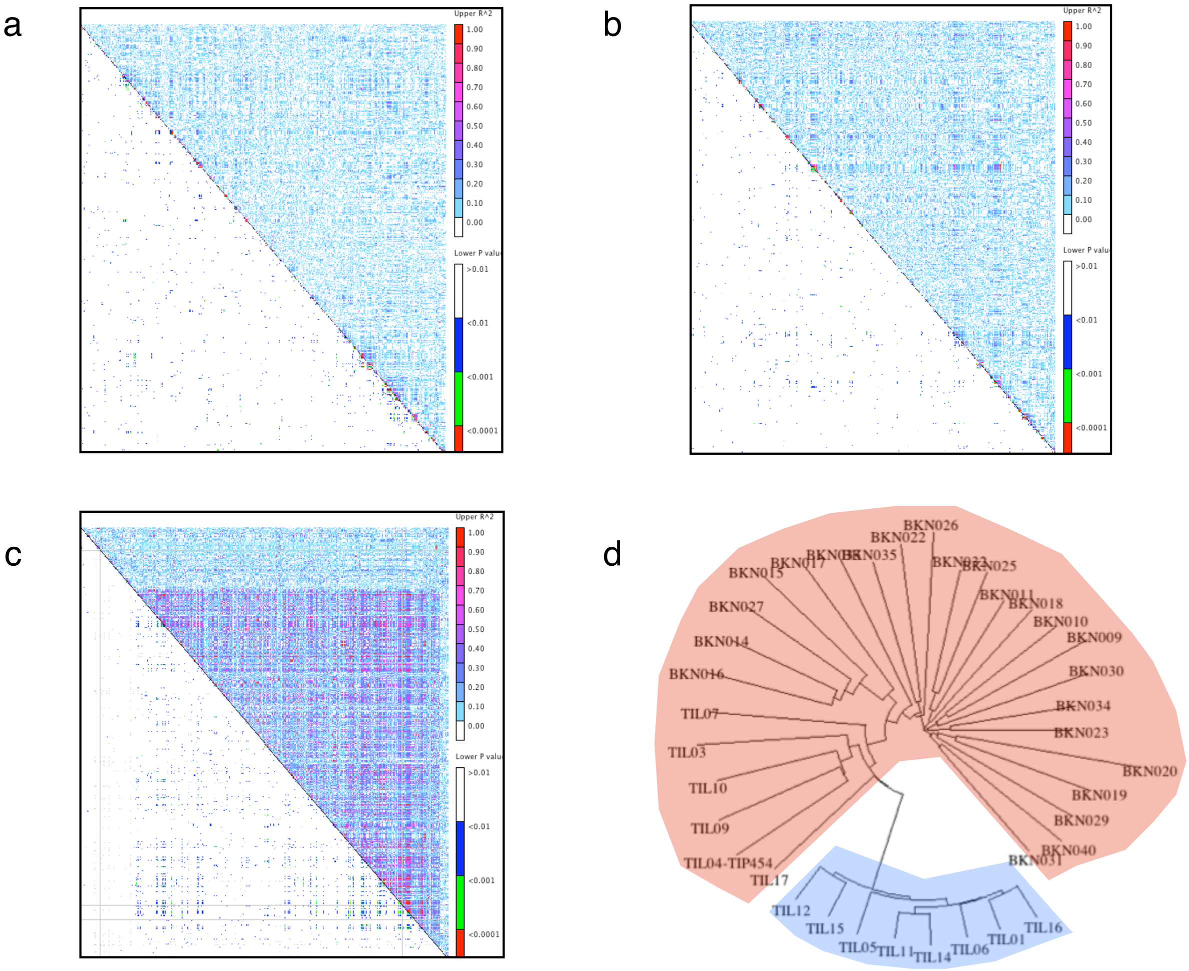

6
XP−CLR
0
10
20
30
40
0
50
100
150
200
250
300
Mb



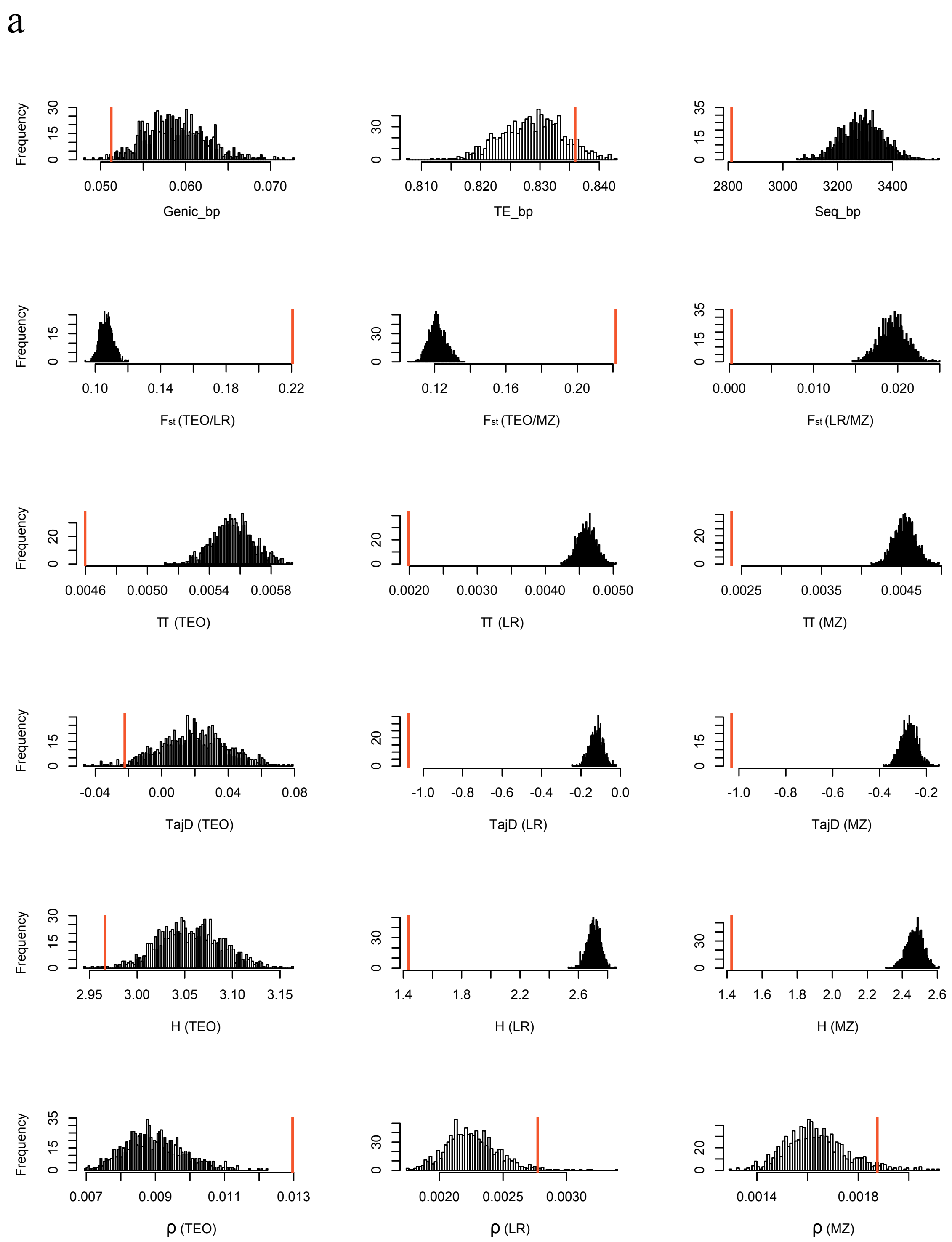
Supplementary Figure 5. XP-CLR across the 10 maize chromosomes. The domestication comparison (a) uses *parviglumis* as a reference and landraces as the object. The improvement comparison (b) uses landraces as a reference and improved lines as the object. Red underscores denote centromeric regions, blue underscore denotes a putative inversion on chromosome 1, these regions are masked from analyses.

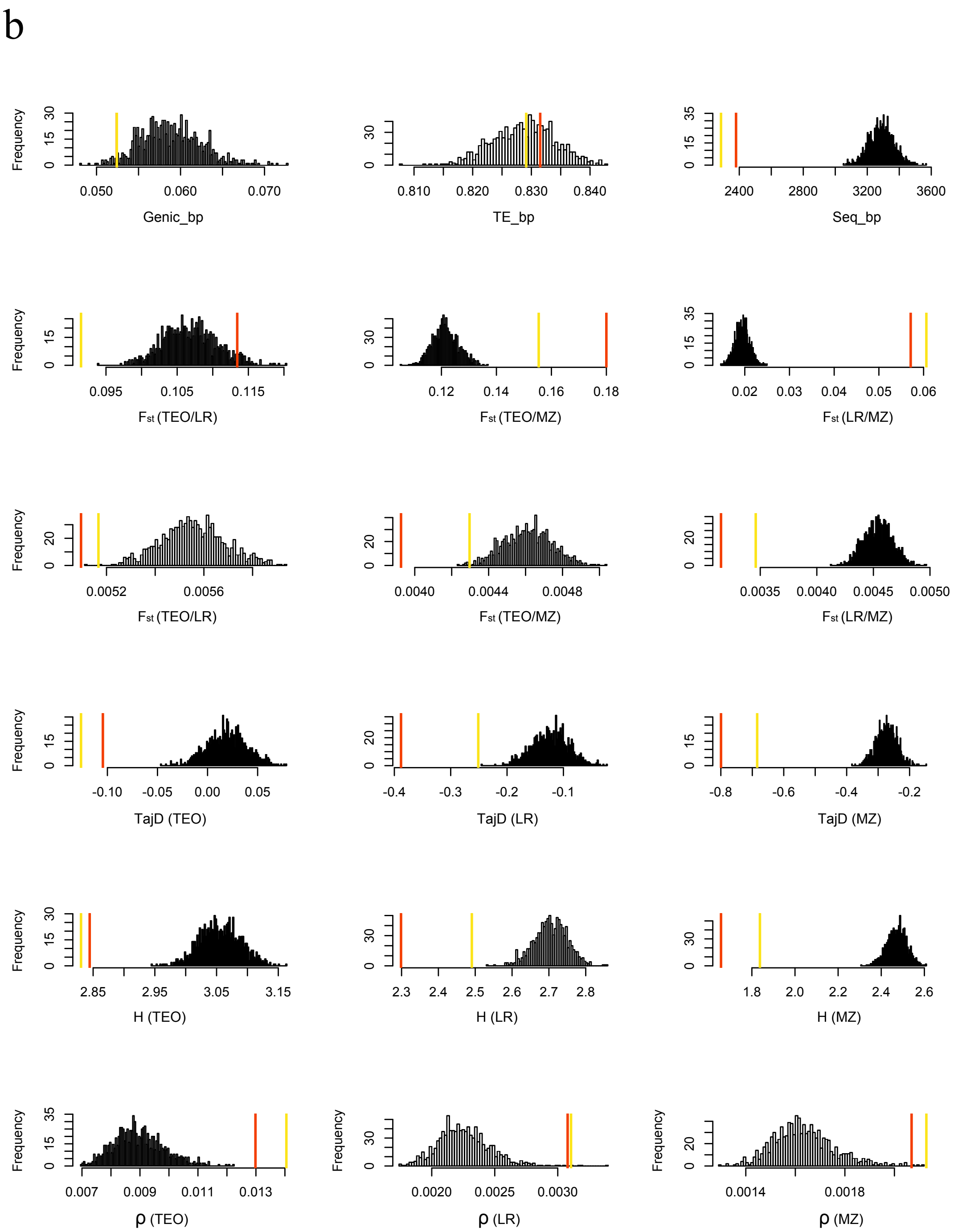


Supplementary Figure 6. Abundance of centromeric retrotransposons CRM2 (a) and CRM3 (b) in modern improved lines and *parviglumis*. Both families have significantly different copy number between the two groups (t-test, FDR = 0.01).

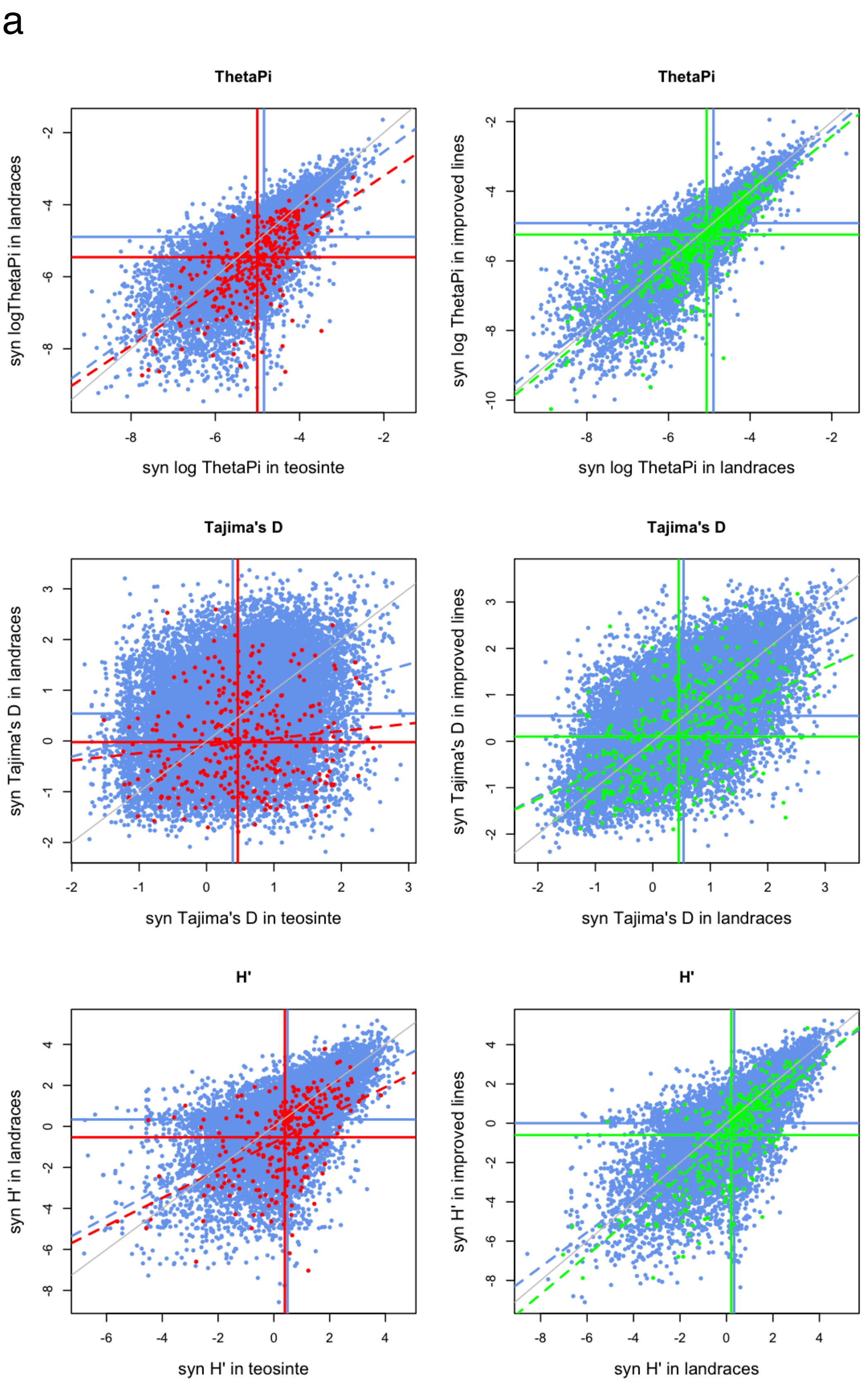


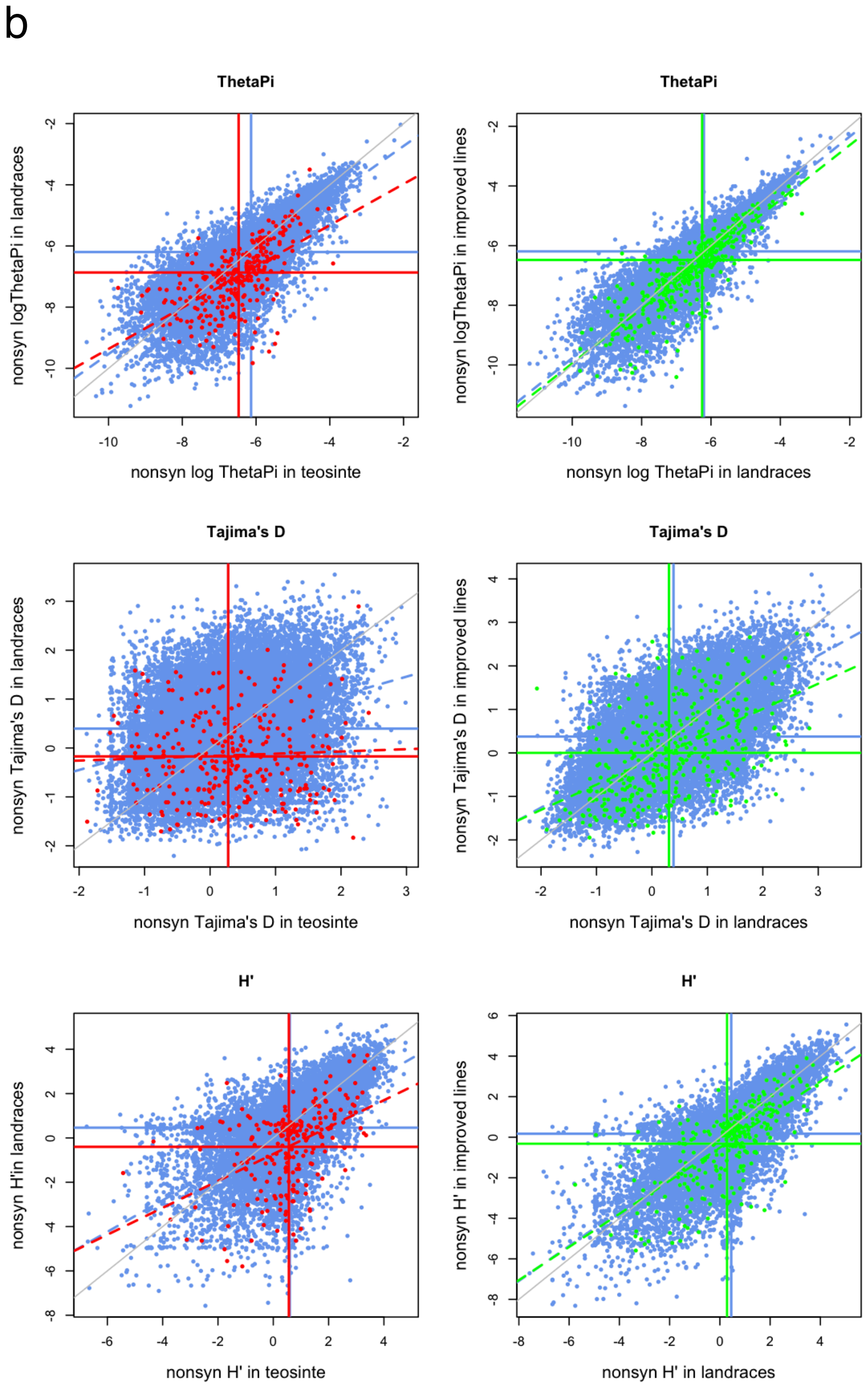
Supplementary Figure 7. Evidence for an inversion on chromosome 1. A-C. Linkage disequilibrium in improved maize (a), landraces (b), and *parviglumis* (c), across a 50-Mb region on chromosome 1. R^2 is shown above the diagonal, and permutation p-values below. (d) A neighbor-joining tree of 64 of the lines using Illumina 55K SNP data from inside the putative inversion. The standard arrangement is on a red background and the inverted arrangement is on a blue background.


a
Frequency
Genic_bp
0.050
0.060
0.070
TE_bp
0.810
0.820
0.830
0.840
Seq_bp
2800
3000
3200
3400
Fst (TEO/LR)
0.10
0.14
0.18
0.22
Fst (TEO/MZ)
0.12
0.16
0.20
Fst (LR/MZ)
0.000
0.010
0.020
π (TEO)
0.0046
0.0050
0.0054
0.0058
π (LR)
0.0020
0.0030
0.0040
0.0050
π (MZ)
0.0025
0.0035
0.0045
TajD (TEO)
-0.04
0.00
0.04
0.08
TajD (LR)
-1.0
-0.8
-0.6
-0.4
-0.2
0.0
TajD (MZ)
-1.0
-0.8
-0.6
-0.4
-0.2
H (TEO)
2.95
3.00
3.05
3.10
3.15
H (LR)
1.4
1.8
2.2
2.6
H (MZ)
1.4
1.6
1.8
2.0
2.2
2.4
2.6
ρ (TEO)
0.007
0.009
0.011
0.013
ρ (LR)
0.0020
0.0025
0.0030
ρ (MZ)
0.0014
0.0018
0
15
20
30
35

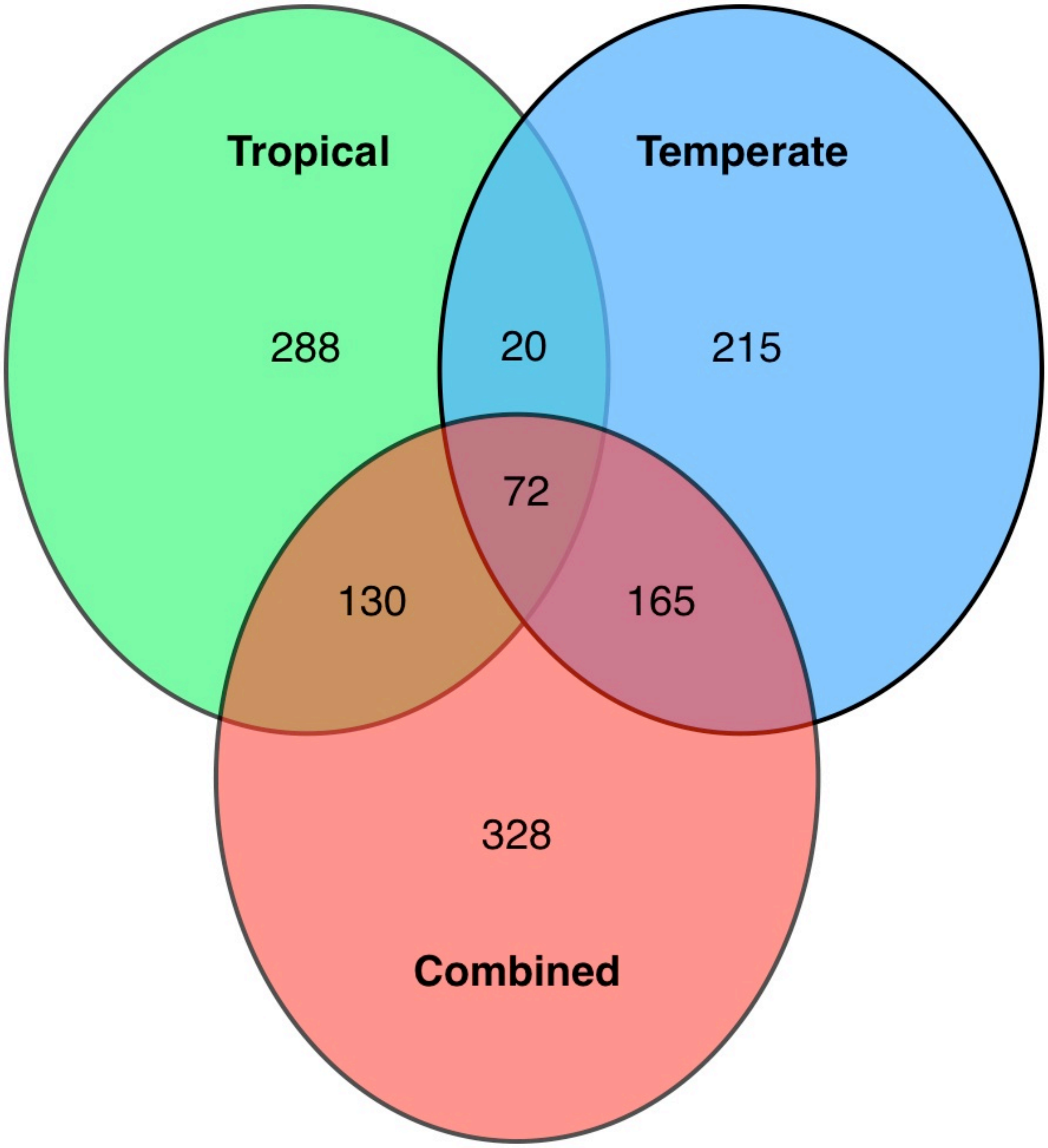


Supplementary Figure 8. Bootstrap results for domestication and improvement features. (a) Domestication: red lines indicate observed values in candidate features. TEO: *parviglumis*, LR: landraces, MZ: improved lines. Genic_bp: proportion of genic basepairs. TE_bp: proportion transposable elements, Seq_bp: average number of basepairs covered per 10kb. F_{st}: F_{st}, π: average nucleotide diversity. TajD: Tajima's D. H: Fay and Wu's H', ρ: population recombination rate. (b) Improvement: Red lines indicate observed values in features. Yellow lines are observed values for candidate features excluding those overlapping with domestication features. Abbreviations are as in (a).


a
ThetaPi
syn logThetaPi in landraces
-2
-4
-6
-8
-8
-6
-4
-2
syn log ThetaPi in teosinte
ThetaPi
syn log ThetaPi in improved lines
-2
-4
-6
-8
-10
-8
-6
-4
-2
syn log ThetaPi in landraces
Tajima's D
syn Tajima's D in landraces
3
2
1
0
-1
-2
-2
-1
0
1
2
3
syn Tajima's D in teosinte
Tajima's D
syn Tajima's D in improved lines
3
2
1
0
-1
-2
-2
-1
0
1
2
3
syn Tajima's D in landraces
H'
syn H' in landraces
4
2
0
-2
-4
-6
-8
-6
-4
-2
0
2
4
syn H' in teosinte
H'
syn H' in improved lines
4
2
0
-2
-4
-6
-8
-8
-6
-4
-2
0
2
4
syn H' in landraces

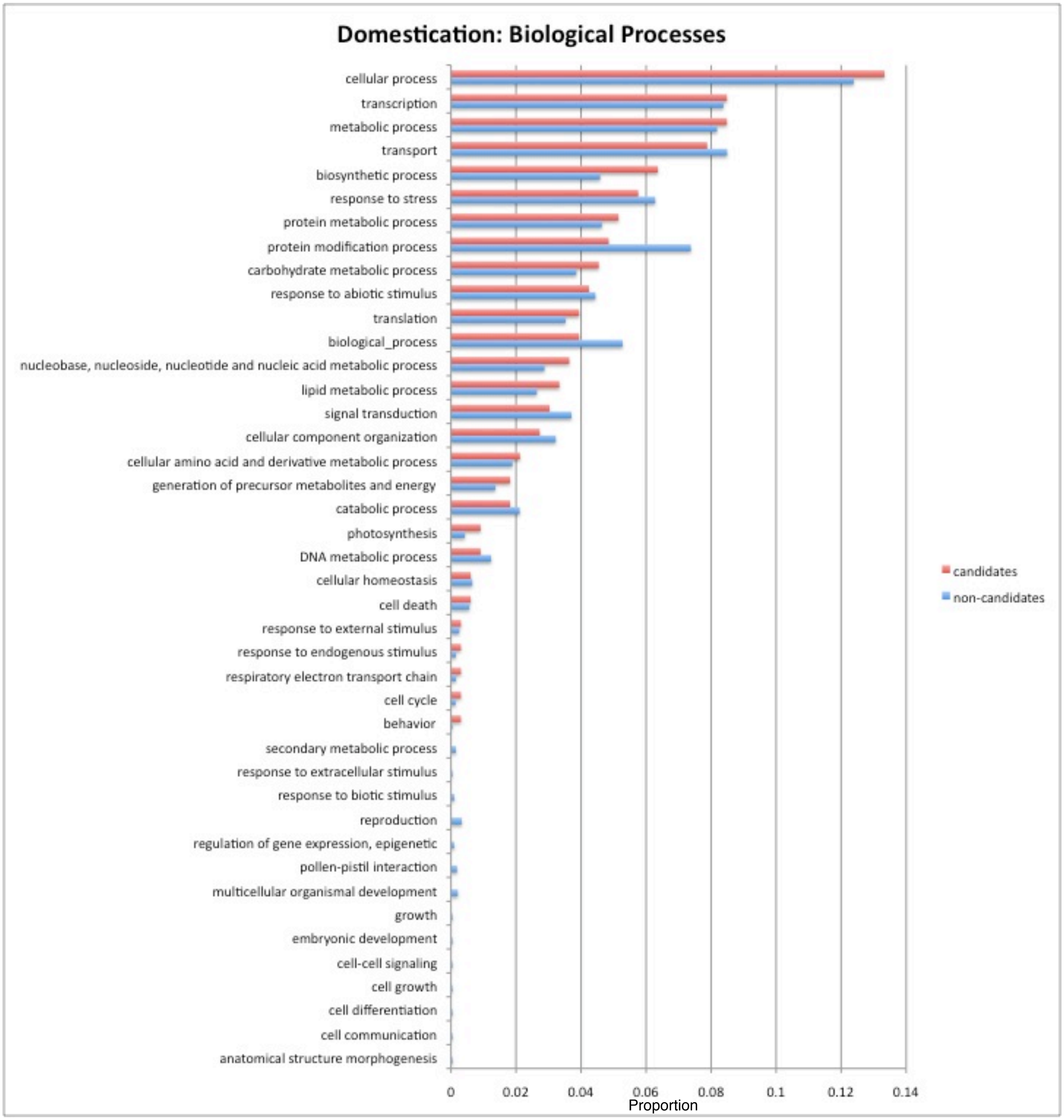


Supplementary Figure 9. Relationship of π (ThetaPi), Tajima's D, and Fay and Wu's normalized H value, H' per gene between parviglumis and landraces (domestication candidates in red) and between landraces and improved lines (improvement candidates in green) for synonymous (syn) (a) and non-synonymous (nonsyn) (b) sites. 1:1 line in grey. Non-candidates in blue. Horizontal and vertical lines indicate mean values.

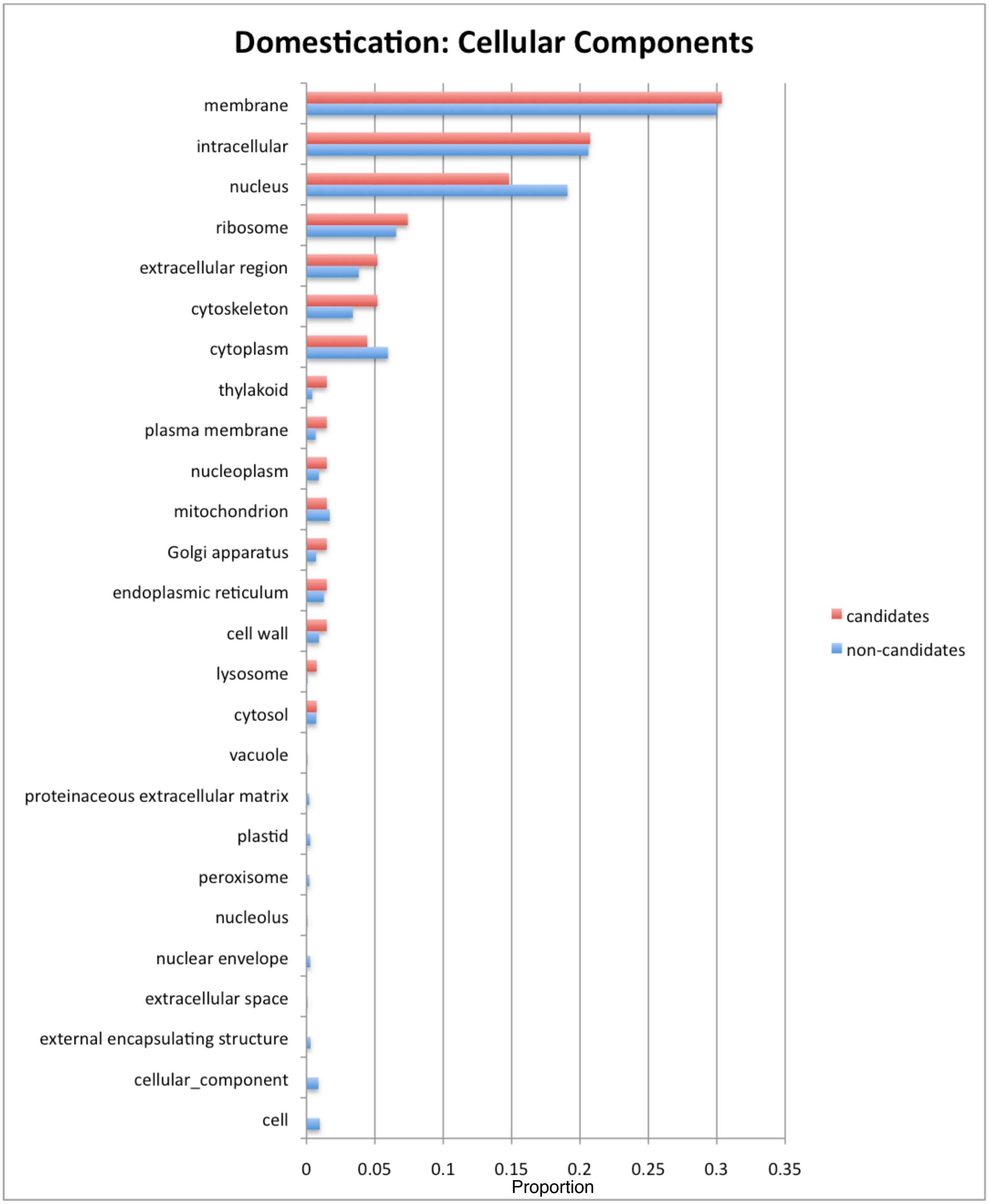


Supplementary Figure 10. Venn diagram illustrating the overlap of features in the top 10% of XP-CLR in the combined improvement scan and in separate scans including only tropical and temperate improved lines.

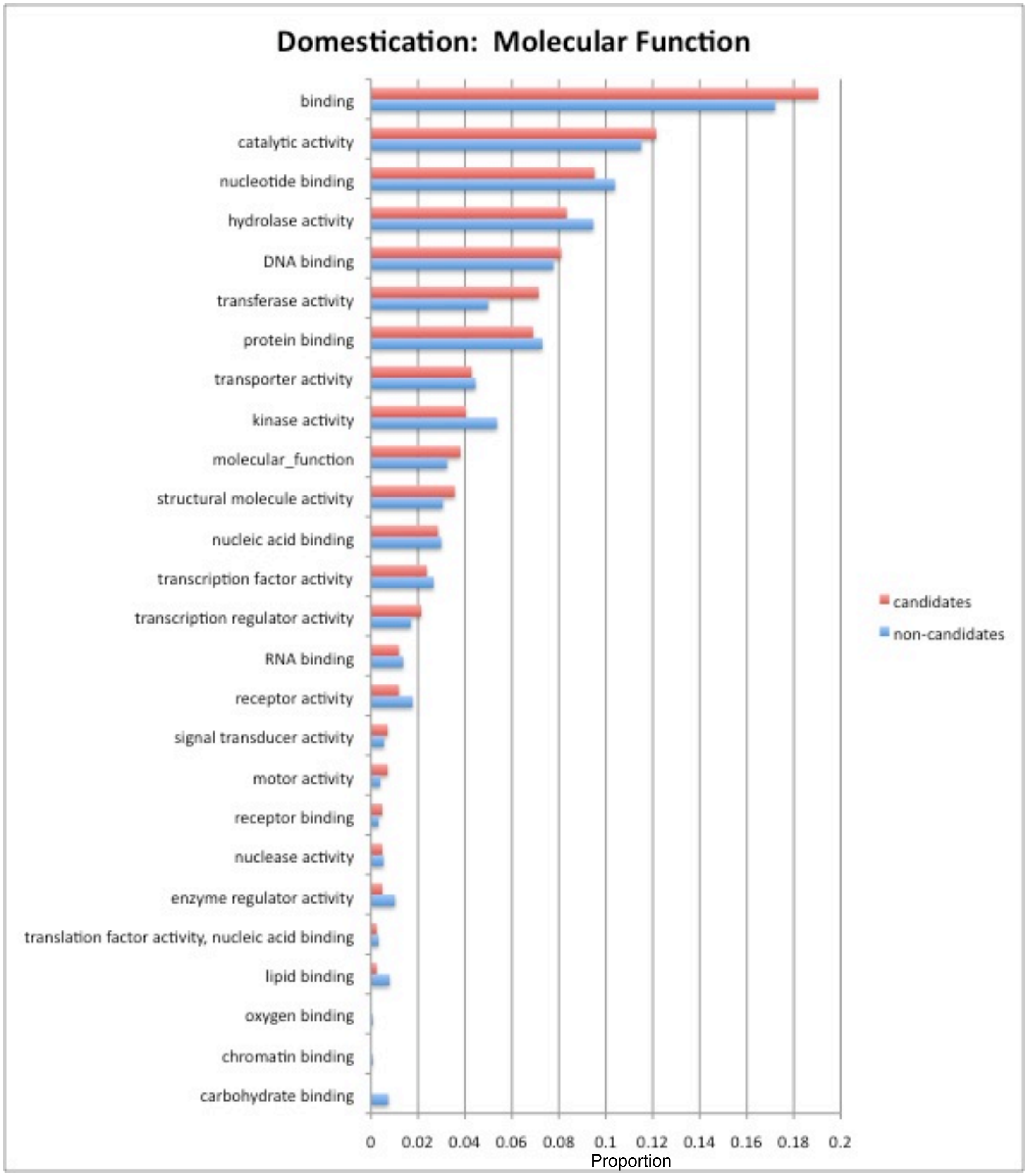
a


Domestication: Biological Processes
cellular process
transcription
metabolic process
transport
biosynthetic process
response to stress
protein metabolic process
protein modification process
carbohydrate metabolic process
response to abiotic stimulus
translation
biological_process
nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
lipid metabolic process
signal transduction
cellular component organization
cellular amino acid and derivative metabolic process
generation of precursor metabolites and energy
catabolic process
photosynthesis
DNA metabolic process
cellular homeostasis
cell death
response to external stimulus
response to endogenous stimulus
respiratory electron transport chain
cell cycle
behavior
secondary metabolic process
response to extracellular stimulus
response to biotic stimulus
reproduction
regulation of gene expression, epigenetic
pollen-pistil interaction
multicellular organismal development
growth
embryonic development
cell-cell signaling
cell growth
cell differentiation
cell communication
anatomical structure morphogenesis
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14
Proportion
candidates
non-candidates

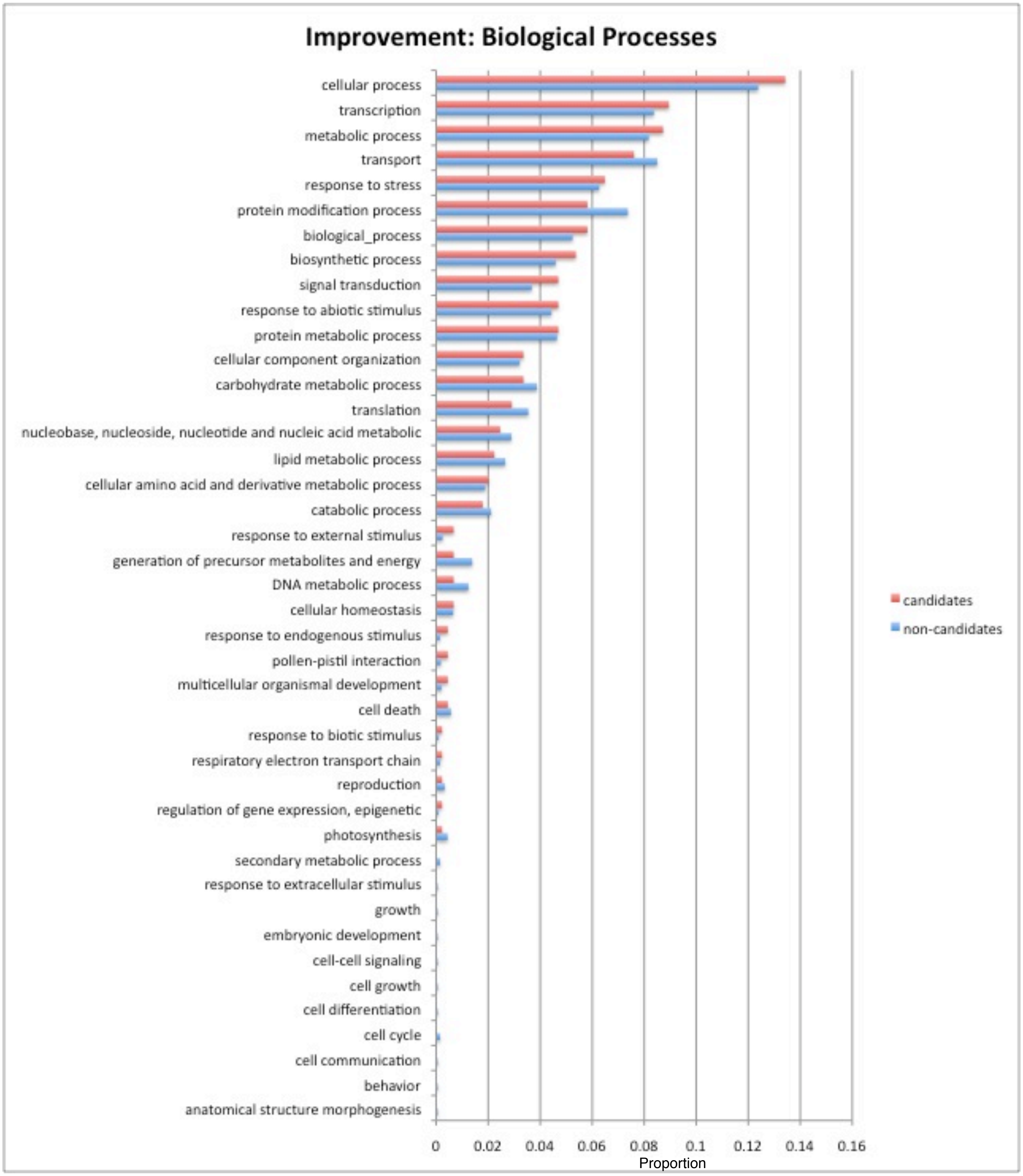
b


Domestication: Cellular Components
membrane
intracellular
nucleus
ribosome
extracellular region
cytoskeleton
cytoplasm
thylakoid
plasma membrane
nucleoplasm
mitochondrion
Golgi apparatus
endoplasmic reticulum
cell wall
lysosome
cytosol
vacuole
proteinaceous extracellular matrix
plastid
peroxisome
nucleolus
nuclear envelope
extracellular space
external encapsulating structure
cellular_component
cell
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
Proportion
🟥 candidates
🟦 non-candidates

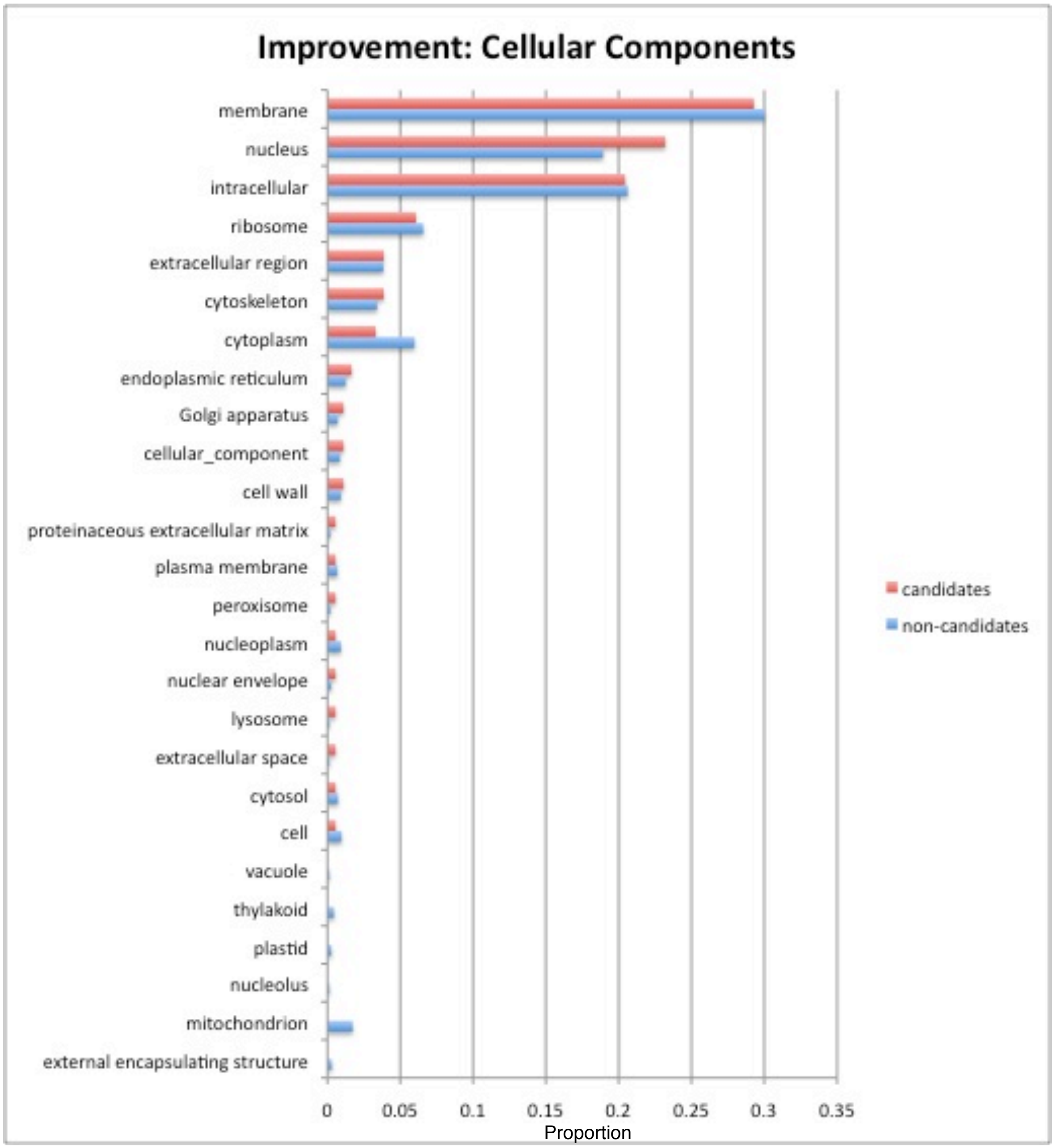
C


Domestication: Molecular Function
binding
catalytic activity
nucleotide binding
hydrolase activity
DNA binding
transferase activity
protein binding
transporter activity
kinase activity
molecular_function
structural molecule activity
nucleic acid binding
transcription factor activity
transcription regulator activity
RNA binding
receptor activity
signal transducer activity
motor activity
receptor binding
nuclease activity
enzyme regulator activity
translation factor activity, nucleic acid binding
lipid binding
oxygen binding
chromatin binding
carbohydrate binding
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14
0.16
0.18
0.2
Proportion
🟥 candidates
🟦 non-candidates

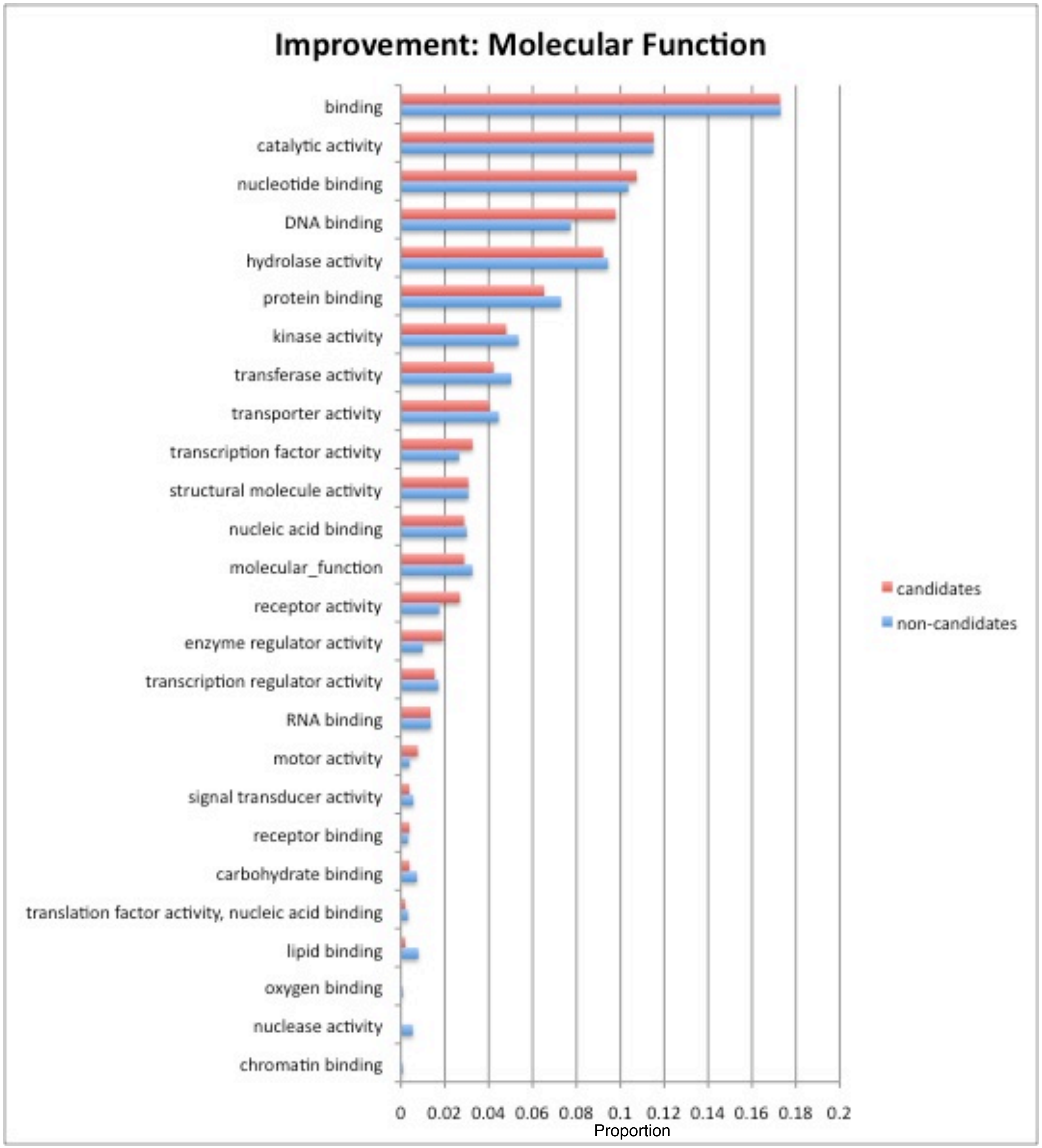
d


Improvement: Biological Processes
cellular process
transcription
metabolic process
transport
response to stress
protein modification process
biological_process
biosynthetic process
signal transduction
response to abiotic stimulus
protein metabolic process
cellular component organization
carbohydrate metabolic process
translation
nucleobase, nucleoside, nucleotide and nucleic acid metabolic
lipid metabolic process
cellular amino acid and derivative metabolic process
catabolic process
response to external stimulus
generation of precursor metabolites and energy
DNA metabolic process
cellular homeostasis
response to endogenous stimulus
pollen-pistil interaction
multicellular organismal development
cell death
response to biotic stimulus
respiratory electron transport chain
reproduction
regulation of gene expression, epigenetic
photosynthesis
secondary metabolic process
response to extracellular stimulus
growth
embryonic development
cell-cell signaling
cell growth
cell differentiation
cell cycle
cell communication
behavior
anatomical structure morphogenesis
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14
0.16
Proportion
🟥 candidates
🟦 non-candidates

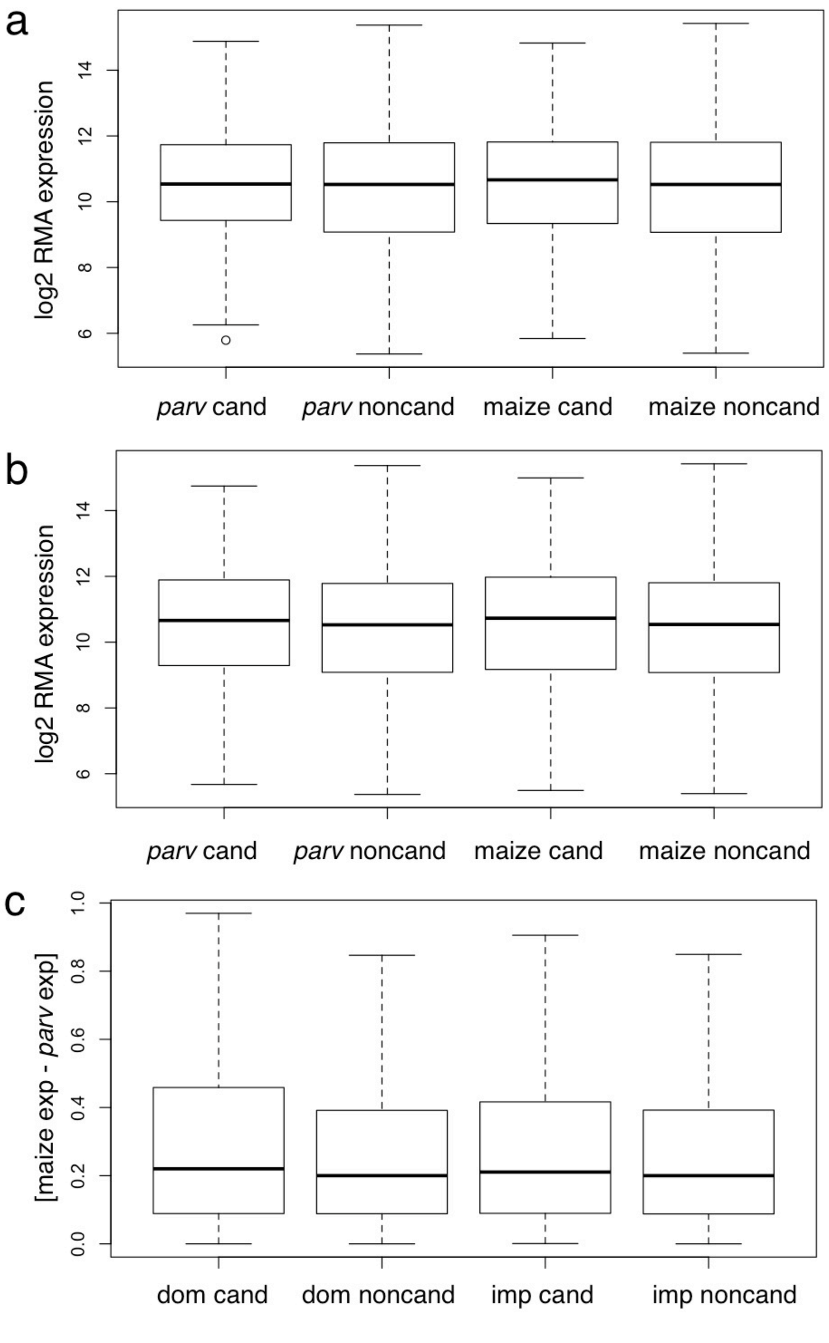
e


Improvement: Cellular Components
membrane
nucleus
intracellular
ribosome
extracellular region
cytoskeleton
cytoplasm
endoplasmic reticulum
Golgi apparatus
cellular_component
cell wall
proteinaceous extracellular matrix
plasma membrane
peroxisome
nucleoplasm
nuclear envelope
lysosome
extracellular space
cytosol
cell
vacuole
thylakoid
plastid
nucleolus
mitochondrion
external encapsulating structure
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
Proportion
candidates
non-candidates

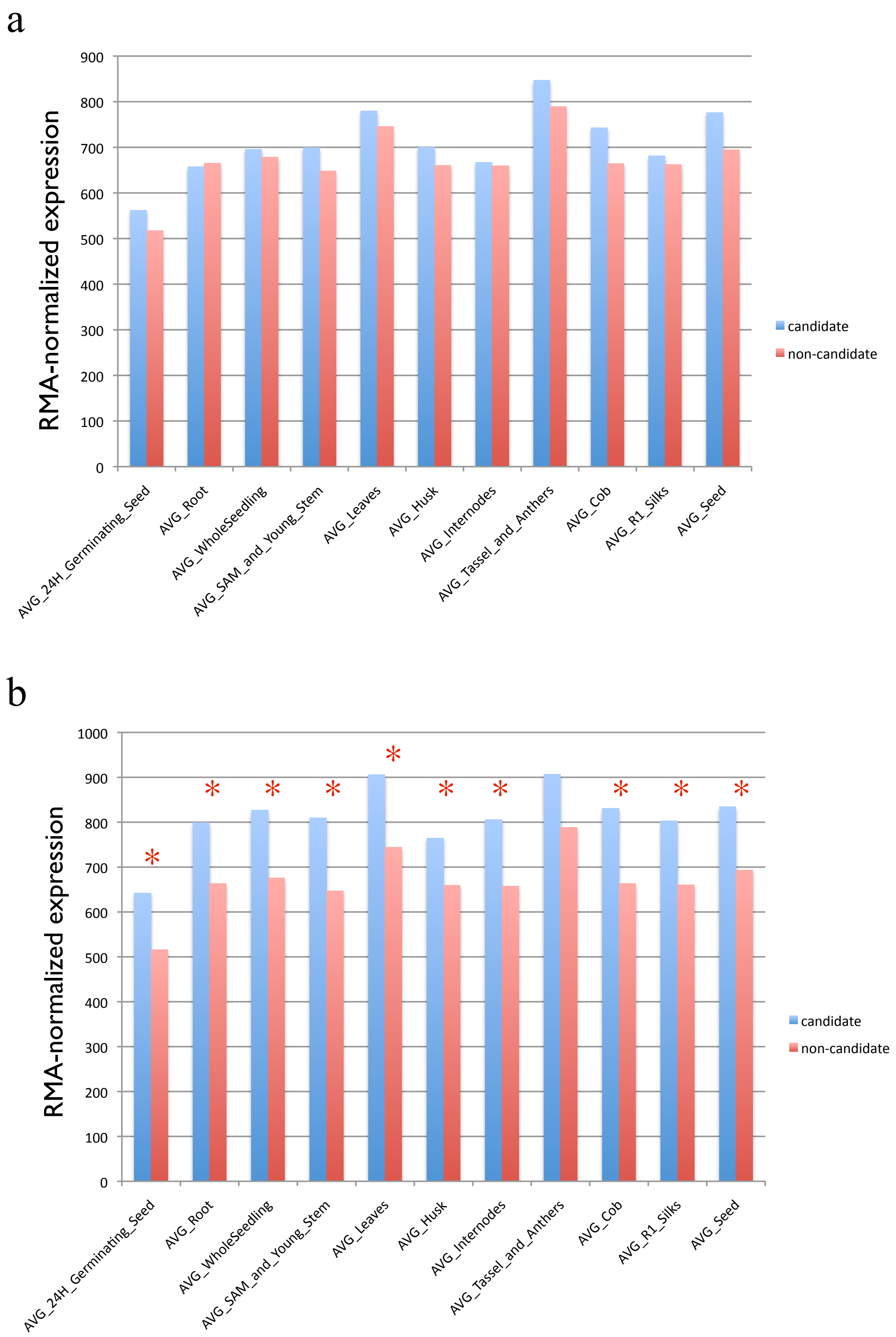
f



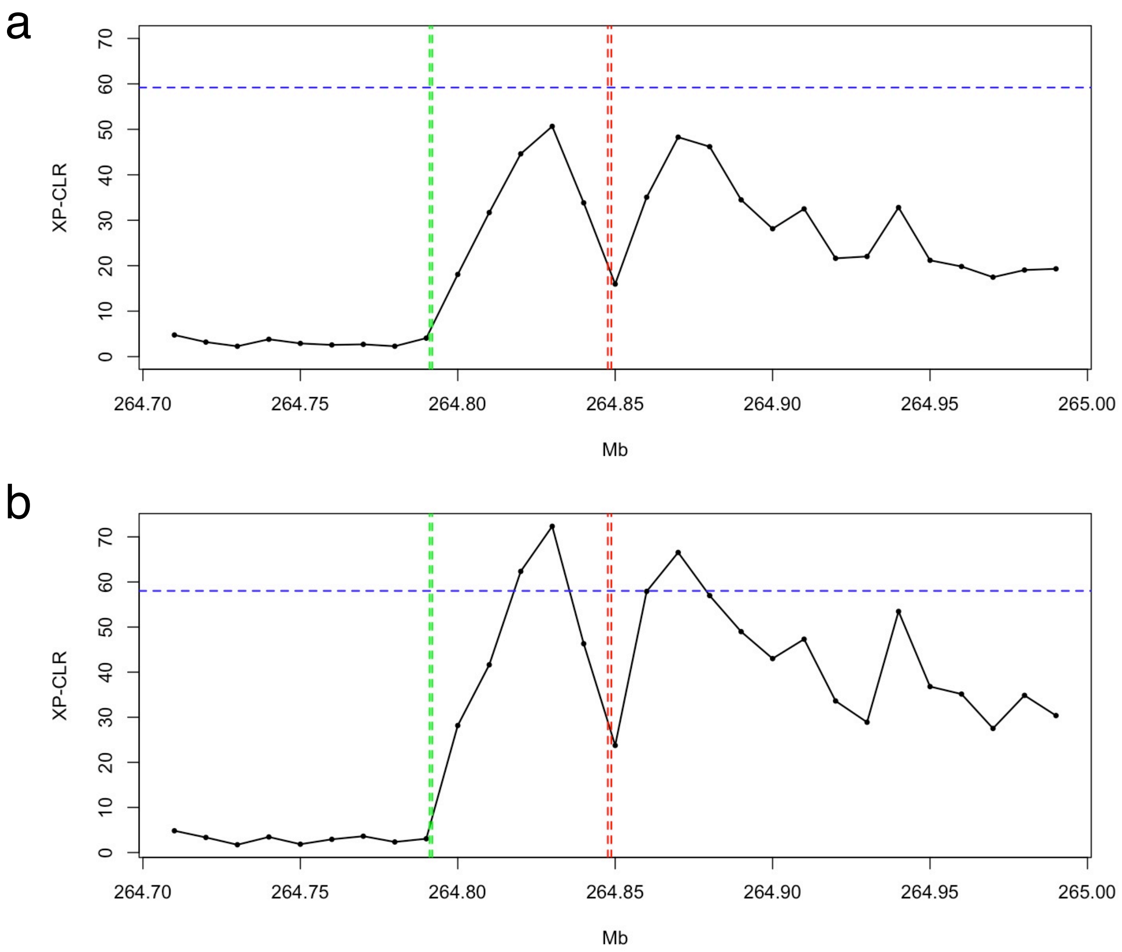
Supplementary Figure 11. Proportion of domestication (a-c) and improvement (d-f) candidates in GoSlim categories relative to non-candidate genes.



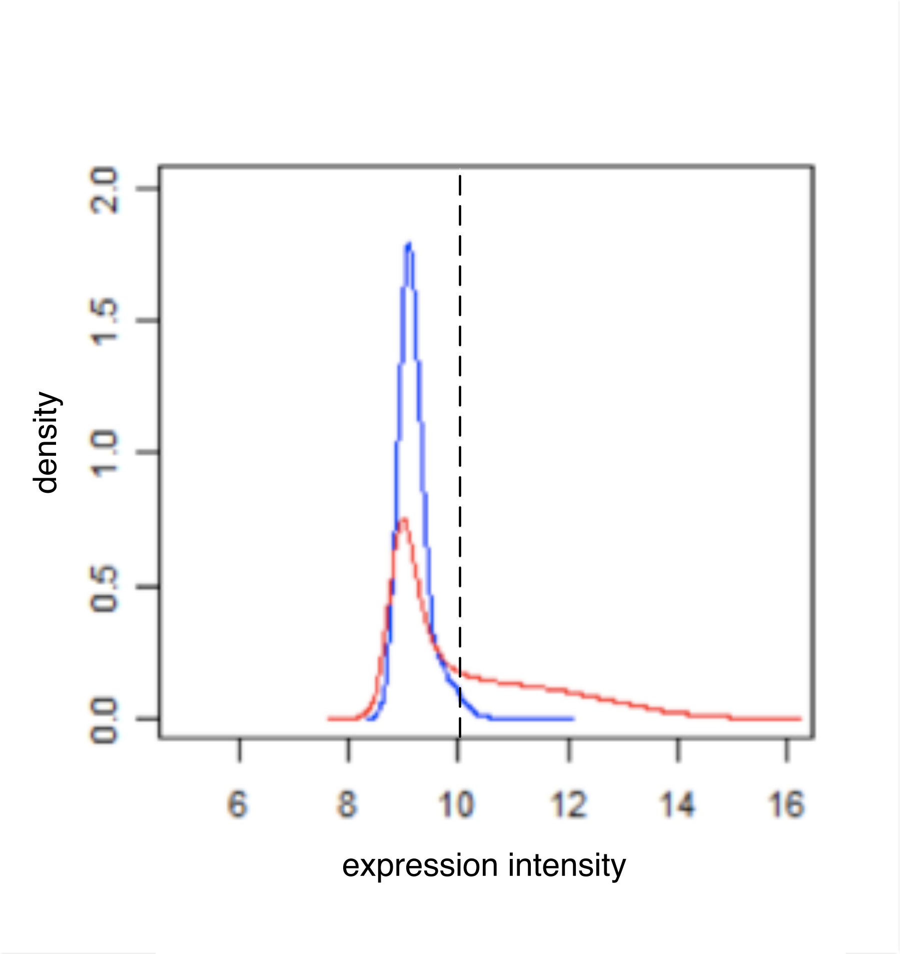
Supplementary Figure 12. Distributions of expression values in candidates (cand) and non-candidates (noncand) in *parviglumis* (*parv*) and maize. (a) Log2 RMA-normalized expression values in domestication candidates relative to non-candidates. (b) Log2 RMA-normalized expression values in improvement candidates relative to non-candidates. (c) Absolute value of the difference in expression between maize and *parviglumis* in domestication (dom) and improvement (imp) candidates and non-candidates. Outliers have been removed in (c).



Supplementary Figure 13. RMA-normalized expression in 11 tissue types in domestication (a) and improvement (b) candidates relative to non-candidates. Asterisks denote significant differences in expression levels.



Supplementary Figure 14. Domestication scan statistic (XP-CLR) in the region surrounding the *teosinte branched1* (*tb1*) locus including the full set of *parviglumis* lines (a) and the subset of *parviglumis* lines without the maize allele (b). Green dashed lines delimit the *Hopscotch* insertion found in domesticated maize, red dashed lines delimit the coding region of *tb1*, and the blue dashed line indicates the cut-off for the top 10% of XP-CLR features in the two scans.



Supplementary Figure 15. Density plots of signal intensity for random sequence controls (blue) and experimental gene probes (red). Dashed line indicates the threshold at which genes were considered expressed.