nature genetics

Comparative population genomics of maize domestication and improvement

Matthew B Hufford[1,17], Xun Xu[2,17], Joost van Heerwaarden[1,17], Tanja Pyhäjärvi[1,17], Jer-Ming Chia[3], Reed A Cartwright[4,5], Robert J Elshire[6], Jeffrey C Glaubitz[6], Kate E Guill[7,8], Shawn M Kaeppler[9,10], Jinsheng Lai[11], Peter L Morrell[12], Laura M Shannon[13], Chi Song[2], Nathan M Springer[14], Ruth A Swanson-Wagner[14], Peter Tiffin[14], Jun Wang[2], Gengyun Zhang[2], John Doebley[13], Michael D McMullen[7,8], Doreen Ware[3,7], Edward S Buckler[6,7], Shuang Yang[2] & Jeffrey Ross-Ibarra[1,15,16]

Domestication and plant breeding are ongoing 10,000-year-old evolutionary experiments that have radically altered wild species to meet human needs. Maize has undergone a particularly striking transformation. Researchers have sought for decades to identify the genes underlying maize evolution[1,2], but these efforts have been limited in scope. Here, we report a comprehensive assessment of the evolution of modern maize based on the genome-wide resequencing of 75 wild, landrace and improved maize lines[3]. We find evidence of recovery of diversity after domestication, likely introgression from wild relatives, and evidence for stronger selection during domestication than improvement. We identify a number of genes with stronger signals of selection than those previously shown to underlie major morphological changes[4,5]. Finally, through transcriptome-wide analysis of gene expression, we find evidence both consistent with removal of *cis*-acting variation during maize domestication and improvement and suggestive of modern breeding having increased dominance in expression while targeting highly expressed genes.

Archaeological[6] and genetic[7,8] evidence indicate that maize (*Zea mays* ssp. *mays*) was domesticated approximately 10,000 years ago in the Balsas River Basin of southwestern Mexico. Domestication involved a radical phenotypic transformation from the wild progenitor, *Zea mays* ssp. *parviglumis* (hereafter, *parviglumis*; **Fig. 1**), resulting in an unbranched plant with seed attached to a cob and thereby making maize entirely dependent on humans for propagation. Subsequent to domestication, maize has been subject to intensive improvement efforts, culminating in the development of hybrid maize lines that are highly adapted to modern agricultural practices. We present a population genomics analysis of maize evolution based on the resequencing of 75 genomes of maize and its wild relatives (**Fig. 1**, **Supplementary Fig. 1** and **Supplementary Table 1**). We generated 781 Gb of sequence from 35 improved maize lines, 23 traditional landraces and 17 wild relatives (14 *parviglumis*; 2 *Zea mays* ssp. *mexicana*, hereafter *mexicana*; and 1 *Tripsacum dactyloides* var. *meridionale*) using short-read technology, sequencing each line to an average depth of more than 5× (**Supplementary Table 1**)[3]. Reads were mapped to the maize reference genome (release 4a.53), and analyses are based on a final set of 21,141,953 high-quality SNPs.

Maize landraces retain more nucleotide diversity (83%; **Fig. 2a**) and show lower genetic differentiation from their wild progenitor (F_{ST} = 0.11) than other crop species[9,10]. This is likely due to a large census size and an outcrossing mating system in maize landraces. Linkage disequilibrium (LD) has increased markedly as a result of domestication, with the genome-wide population recombination rate ρ in landraces estimated to be 25% of the rate in *parviglumis* and with average haplotype length increasing from 22 kb to 30 kb (**Supplementary Fig. 2**). These results are consistent with the effects of a domestication bottleneck, but an excess of rare SNPs (**Supplementary Fig. 3** and **Supplementary Table 2**) suggests that variation has begun to recover across most of the genome. Gene-rich regions, however, have fewer SNPs unique to landraces (*t* test, $P < 0.001$) and no excess of rare SNPs (**Supplementary Tables 2** and **3**), a difference that is likely due to the effects of background selection against deleterious mutations slowing the post-domestication recovery of variation at linked sites. Modern breeding seems to have had negligible effects on genome-wide diversity or mean haplotype lengths in our broad sample of

[1]Department of Plant Sciences, University of California, Davis, California, USA. [2]BGI-Shenzhen, Shenzhen, China. [3]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA. [4]Center for Evolutionary Medicine and Informatics, Biodesign Institute, Arizona State University, Tempe, Arizona, USA. [5]School of Life Sciences, Arizona State University, Tempe, Arizona, USA. [6]Institute for Genomic Diversity, Cornell University, Ithaca, New York, USA. [7]US Department of Agriculture–Agriculture Research Service (USDA-ARS). [8]Division of Plant Sciences, University of Missouri, Columbia, Missouri, USA. [9]Department of Energy (DOE) Great Lakes Bioenergy Research Center, University of Wisconsin, Madison, Wisconsin, USA. [10]Department of Agronomy, University of Wisconsin, Madison, Wisconsin, USA. [11]State Key Laboratory of Agrobiotechnology, China Agricultural University, Beijing, China. [12]Department of Agronomy & Plant Genetics, University of Minnesota, St Paul, Minnesota, USA. [13]Department of Genetics, University of Wisconsin, Madison, Wisconsin, USA. [14]Department of Plant Biology, University of Minnesota, St Paul, Minnesota, USA. [15]The Genome Center, University of California, Davis, California, USA. [16]The Center for Population Biology, University of California, Davis, California, USA. [17]These authors contributed equally to this work. Correspondence should be addressed to J.R.-I. (rossibarra@ucdavis.edu), E.S.B. (esb33@cornell.edu) or S.Y. (yangsh@genomics.org.cn).

Received 8 August 2011; accepted 4 May 2012; published online 3 June 2012; doi:10.1038/ng.2309

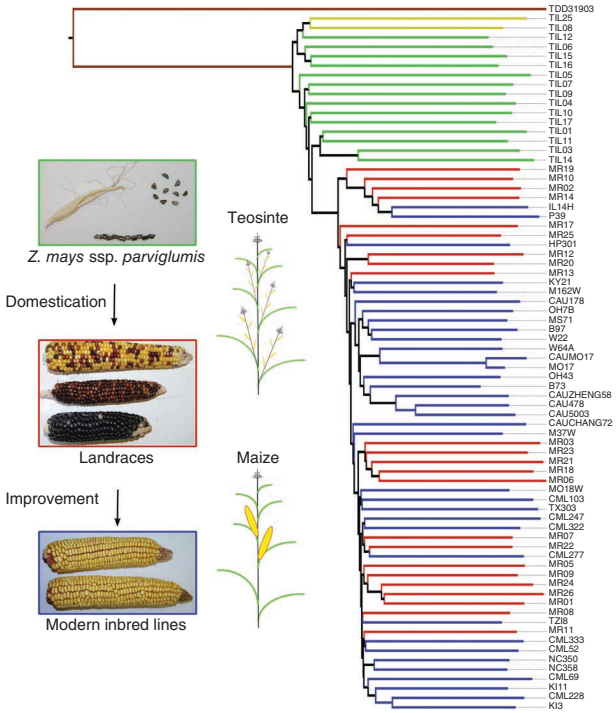

Figure 1 Neighbor-joining tree and changing morphology of domesticated maize and its wild relatives. Taxa in the neighbor-joining tree (right) are represented by different colors: *parviglumis* (green), landraces (red), improved lines (blue), *mexicana* (yellow) and *Tripsacum* (brown). Morphological changes (left) are shown for female inflorescences and plant architecture during domestication and improvement.

modern lines (**Fig. 2a**, **Supplementary Fig. 2** and **Supplementary Table 2**). Although our estimates of nucleotide diversity in improved lines may be inflated by the diverse inbred lines chosen, the relationships between inbred and landrace lines (**Fig. 1**) suggest a weaker genome-wide bottleneck during improvement. Finally, comparison of maize landraces to the two *mexicana* genomes identifies several extended regions of high genetic similarity (**Supplementary Fig. 4**), consistent with previous observations of admixture between these taxa[8,11] and raising the possibility that *mexicana* may have contributed alleles important for maize evolution.

To identify regions of the genome most affected by selection during maize evolution, we used a likelihood method (the cross-population composite likelihood ratio, XP-CLR)[12] to scan for extreme allele frequency differentiation over extended linked regions (**Fig. 2b**,**c**). Adjacent windows of high XP-CLR were grouped into 'features', with each likely representing the effect of a single selective sweep. Features in multiple centromeres showed high XP-CLR values (**Fig. 2b**,**c** and **Supplementary Fig. 5**). Combined with evidence for change in abundance of centromeric retroelements (**Supplementary Fig. 6** and **Supplementary Table 4**)[3], this finding suggests rapid centromere evolution. However, because centromeres harbor few genes and because our genetic map may underestimate the extended LD in these regions, we masked centromeres from further analysis. We also masked a newly discovered ~50-Mb inversion polymorphism on chromosome 1 (**Supplementary Fig. 7**)[13].

We focused analyses on the 484 domestication and 695 improvement features in the highest 10% of XP-CLR values (**Fig. 2b**,**c**). Domestication features contained an average of 3.4 genes, had a mean size of 322 kb (**Fig. 2d**,**e**), covered approximately 7.6% of the maize genome and showed multiple signatures of selection, including elevated differentiation, low nucleotide diversity and an excess of high-frequency derived SNPs (**Supplementary Figs. 8** and **9** and **Supplementary Table 5**). We estimate the mean strength of selection in these features as $s = 0.015$, which is within the range of estimates determined on the basis of archaeological data from other domesticates[14] and more than an order of magnitude higher than the mean value of 0.0011 across the rest of the genome.

Whereas selection during maize improvement can be strong[15,16], XP-CLR values and estimated selection coefficients (mean $s = 0.003$) from our improvement scan were substantially lower than those observed for domestication (**Fig. 2b**,**c**). Consistent with this finding, improvement features had smaller average size (**Fig. 2d**) and contained fewer genes (**Fig. 2e**) than domestication features. One explanation for these results may be that the diverse tropical and temperate lines analyzed derive from distinct landrace founders (**Fig. 1**) and have been subject to different selective pressures[17]. Indeed, independent scans of temperate and tropical lines found stronger evidence of selection and little overlap of selected features (**Supplementary Fig. 10**). However, previous estimates of effect size for loci involved in domestication and improvement traits provide some independent evidence of stronger selection during domestication[18]. We found that 23% of domestication features (107) showed additional evidence of selection during improvement, indicating that a subset of domestication loci may contribute to phenotypes of continued agronomic importance.

Individual features likely result from a single selective event, and we assigned the gene closest to the 10-kb window with the maximum XP-CLR score in each feature as the most likely candidate (**Supplementary Tables 6** and **7**). Our domestication and improvement candidate lists, each including 1–2% of the maize filtered gene set (FGS), represent our best estimate of the direct targets of selection within features, but linked genes have also been affected by selection, limiting the diversity available for modern improvement for many of the 3,040 genes found

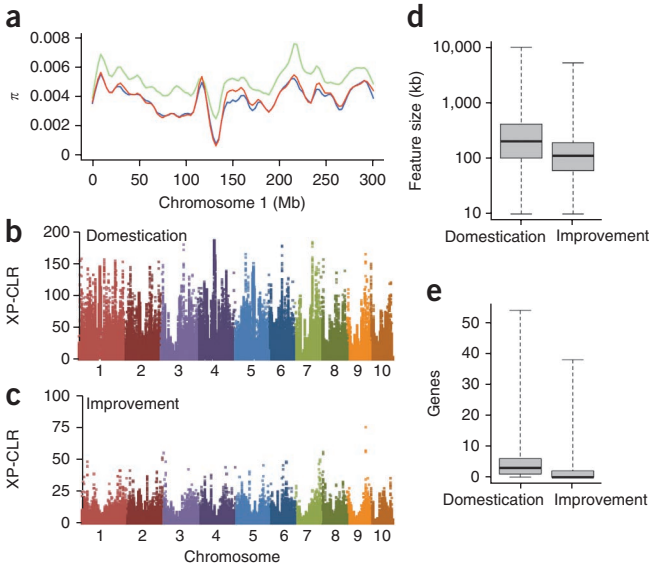

Figure 2 Genome-wide analysis of nucleotide diversity and selection. (**a**) LOWESS curves of nucleotide diversity (π) along chromosome 1 in *parviglumis* (green), landraces (red) and improved lines (blue). (**b**,**c**) Genome-wide likelihood (XP-CLR) values for selection during domestication (**b**) and improvement (**c**), with chromosome number indicated along the *x* axis. (**d**,**e**) Distributions of feature size (**d**) and gene counts within features (**e**) in domestication and improvement scans. Whiskers indicate maximum and minimum of data, boxes span the interquartile range, and the solid line indicates the median.

within features. Thus, although our candidates are of the most interest for understanding genes directly related to maize evolution, breeding programs would likely benefit from efforts to incorporate diversity from exotic germplasm in these genomic regions.

A sizeable fraction of the domestication and improvement features contained no annotated sequence (6% and 11%, respectively), a finding that could implicate regulatory variants in the process of maize evolution. However, the majority of the features identified contained genes in the high-confidence FGS and should prove useful, both in dissecting existing quantitative trait loci (QTLs) and identifying novel candidate genes. For example, the domestication candidate GRMZM2G448355, an ortholog of the rice gene *OsMADS56* that delays flowering under long day conditions (**Fig. 3a–c**), is found within a flowering time QTL on chromosome 9 (ref. 19), and two improvement candidates implicated in nitrogen metabolism, GRMZM2G036464 (encoding glutamine synthetase) and GRMZM2G428027 (encoding nitrate reductase), both reside in a QTL for multiple traits, including thousand kernel weight and nitrogen mobilization[20]. Only a fraction of the newly identified candidate genes have been functionally characterized in maize; one example is the domestication candidate *abph1* (GRMZM2G035688) that is known to affect phyllotaxy[21]. However, function can often be inferred from orthology (**Supplementary Fig. 11**); the protein encoded by the domestication candidate GRMZM2G010290 has no known function but shows close sequence identity to the DAG1 and DAG2 proteins in *Arabidopsis* that affect seed germination[22]. Two improvement candidates, gibberellin 2–oxidase (encoded by GRMZM2G152354) and gibberellin 3–oxidase (encoded by GRMZM2G036340, *dwarf1*) are found in the plant growth hormone gibberellin biosynthesis pathway (**Fig. 3d–f**) upstream and downstream of the 'green revolution' gene encoding gibberellin 20–oxidase[23]. Other notable improvement candidates include GRMZM2G082468, a homolog of the *Arabidopsis* gene encoding farnesyltransferase, which has been engineered as a drought tolerance transgene in canola[24], and GRMZM2G087612, whose *Arabidopsis* ortholog *SDP1* initiates storage oil breakdown in seed[25].

To further characterize the genomic impact of domestication, we used long oligonucleotide array hybridization to survey expression of 18,242 genes in the FGS in seedling tissue of a subset of 25 improved maize and 7 *parviglumis* lines (**Supplementary Table 1**). Compared to non-candidates, the domestication candidates showed greater absolute changes in expression between *parviglumis* and maize (29% versus 22% in non-candidates, $P = 0.004$; **Supplementary Fig. 12**), upregulation in maize relative to *parviglumis* (11.4% of candidates upregulated versus 6.5% of non-candidates, $P = 0.001$; **Supplementary Fig. 12**) and a 10% lower coefficient of variation (CoV) in expression among maize lines ($P = 0.006$). Reduced variation in expression was observed throughout genes in candidate features ($P < 0.001$) suggesting the removal of *cis* variation at sites linked to the target of selection. Improvement candidates also showed decreased variation in expression in maize relative to *parviglumis* (8% reduction in CoV, $P = 0.019$) but did not show a significant change in the magnitude of expression. Although the lower variation in expression in maize could be due to selection on linked sites, the directional change seen in domestication candidates suggests selection on *cis*-acting regulatory regions.

Although changes in expression during domestication are unlikely to be limited to seedling tissue, our domestication candidates showed no tissue-specific patterns of expression (**Supplementary Fig. 13a**)[26]. Improvement candidates also showed no tissue specificity but were more highly expressed than non-candidates in all but one of the tissue groups evaluated ($P = 0.025$–0.044; **Supplementary Fig. 13b**). Because improvement candidates showed no significant difference in expression between teosinte and modern inbreds, this latter result suggests that modern maize improvement may have targeted loci that

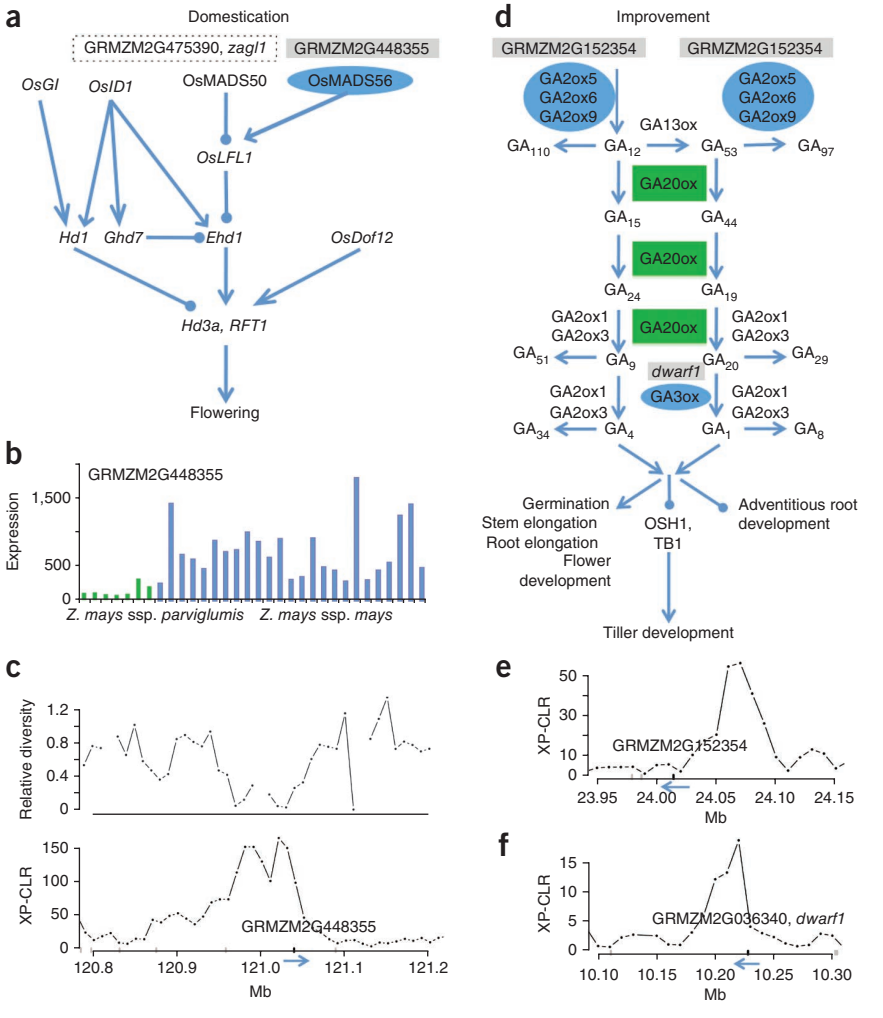

Figure 3 Domestication and improvement candidate genes in relation to two pathways in rice. *Z. mays* genes are shown in boxes above the proteins encoded by their rice orthologs (filled blue circles). Candidate genes are on a gray background, and genes within a selected feature are in a dotted box. Negative regulation is indicated by a circle at the end of an arrow. (**a**–**c**) Domestication candidate genes. (**a**) The flowering time pathway[30], including GRMZM2G448355 and *zagl1*. (**b**) Seedling expression pattern of GRMZM2G448355 in *parviglumis* and maize inbreds. (**c**) XP-CLR and relative diversity near GRMZM2G448355; gene orientation is indicated by the arrow. (**d**) The gibberellin (GA) biosynthesis pathway[31]. The high-yielding rice variety IR8 has a mutation in *GA20ox* (shown on a green background)[23]. (**e**,**f**) XP-CLR values near the improvement candidates GRMZM2G152354 (**e**) and *dwarf1* (**f**); gene orientation is indicated by the arrows.

were already highly expressed. Comparison to the full FGS gave no evidence for an overall bias toward constitutive expression in the candidates, in contrast to what was observed in previous resequencing scans[2,27] (**Supplementary Table 8**).

Finally, we took advantage of expression data from crosses between inbred lines to evaluate levels of dominance in the candidate genes[28]. Domestication candidates showed elevated dominance ($P = 0.001$) but no significant difference in dominance between crosses from the same or different genetic (heterotic) groups (*t* test, $P = 0.74$), a result that can be explained simply by the loss of additive *cis*-regulatory variation. Improvement candidates, in contrast, showed higher dominance of expression than non-candidates ($P = 0.007$), mostly due to higher dominance in crosses between heterotic groups ($P = 0.001$), likely reflecting the important role that complementation between heterotic groups has had in maize improvement.

Our comparative genomic analysis of wild, landrace and modern maize sheds light on the complexities of crop evolution and offers guidance to modern breeding. Earlier work[4,5,29] has shown that a few genes (for example, *tb1* and *tga1*; **Supplementary Fig. 14** and **Supplementary Table 9**) radically altered some aspects of morphology during domestication. The majority of domestication features we identify show stronger evidence of selection than these canonical domestication genes, implying that domestication targeted hundreds of genes of diverse biological function that likely affected unstudied aspects of phenotype. The loss of diversity at sites linked to selection and the observed enrichment of improvement candidates for highly expressed genes suggest that modern breeding has mostly worked with the low-hanging fruit of the genome and that much could be gained by focusing breeding efforts on the effective incorporation of diversity at other loci.

URLs. Gene Expression Omnibus (GEO), http://www.ncbi.nlm.nih.gov/geo/; maize reference genome, http://www.maizesequence.org/; Panzea, http://www.panzea.org/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Expression data are available from the Gene Expression Omnibus (GEO) under the series accession GSE30036.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

The authors would like to thank T. Kono, S. Watson and M. Watson for photographs of inflorescences, P. Brown for help with QTL delineation, B.S. Gaut, A.M. Gonzales and two anonymous reviewers for comments on an earlier version of the manuscript and M. Grote for statistical advice. This work was supported by funding to the maize diversity project from the US National Science Foundation (NSF; IOS-0820619 to E.S.B., J.D. and M.D.M.) and USDA-ARS (to E.S.B., M.D.M. and D.W.), as well as from USDA Hatch Funds (to P.T. and N.M.S.), the Chinese 973 program (2007CB815701 to J.W.), the Chinese Ministry of Agriculture 984 program (2010-Z13 to G.Z.), the Shenzhen Municipal Government Basic Research Program (to J.W.), the US DOE Great Lakes Bioenergy Research Center (DOE Office of Science; BER DE-FC02-07ER64494), the Office of Science of the US DOE (contract DE-AC02-05CH11231 to the US DOE Joint Genome Institute) and by grants from the US NSF (IOS-0922703 to J.R.-I.) and the USDA–National Institute of Food and Agriculture (2009-01864 to J.R.-I.).

AUTHOR CONTRIBUTIONS

J.D., M.D.M., E.S.B., D.W. and J.R.-I. designed the project. M.B.H., J.v.H., T.P. and J.R.-I. performed most data analyses. J.D. developed wild and landrace inbred lines. E.S.B., S.M.K., J.L., M.D.M. and D.W. contributed sequence data for inbred maize and *parviglumis*. K.E.G. and R.J.E. developed libraries and managed sequencing for inbred maize and *parviglumis*. X.X., S.Y., J.W. and G.Z. directed sequencing for landrace maize, *mexicana* and *Tripsacum*. E.S.B., J.R.-I., D.W. and X.X. directed bioinformatics analyses. J.-M.C. and C.S. performed read mapping, SNP calling and annotation, and analysis of coding sequence. E.S.B., J.-M.C. and J.C.G. performed quality control filtering of SNPs. N.M.S., R.A.S.-W. and P.T. generated Nimblegen expression data for maize and *parviglumis*. S.M.K. provided early access expression data. L.M.S. reanalyzed QTL data for domestication traits. R.A.C. analyzed site frequency spectra. M.B.H., J.v.H., T.P., P.L.M. and J.R.-I. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/ng.2309.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. 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).
2. Wright, S.I. *et al.* The effects of artificial selection of the maize genome. *Science* **308**, 1310–1314 (2005).
3. Chia, J.-M. *et al.* Maize HapMap2 identifies extant variation from a genome in flux. *Nat. Genet.* published online, doi:10.1038/ng.2313 (3 June 2012).
4. Doebley, J., Stec, A. & Hubbard, L. The evolution of apical dominance in maize. *Nature* **386**, 485–488 (1997).
5. Wang, H. *et al.* The origin of the naked grains of maize. *Nature* **436**, 714–719 (2005).
6. Piperno, D.R., Ranere, A.J., Holst, I., Iriarte, J. & Dickau, R. Starch grain and phytolith evidence for early ninth millennium BP maize from the Central Balsas River Valley, Mexico. *Proc. Natl. Acad. Sci. USA* **106**, 5019–5024 (2009).
7. Matsuoka, Y. *et al.* A single domestication for maize shown by multilocus microsatellite genotyping. *Proc. Natl. Acad. Sci. USA* **99**, 6080–6084 (2002).
8. van Heerwaarden, J. *et al.* Genetic signals of origin, spread, and introgression in a large sample of maize landraces. *Proc. Natl. Acad. Sci. USA* **108**, 1088–1092 (2011).
9. Caicedo, A.L. *et al.* Genome-wide patterns of nucleotide polymorphism in domesticated rice. *PLoS Genet.* **3**, 1745–1756 (2007).
10. Lam, H.M. *et al.* Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat. Genet.* **42**, 1053–1059 (2010).
11. Wilkes, H.G. *Teosinte: The Closest Relative of Maize* (The Bussey Institute of Harvard University, Cambridge, Massachusetts, 1967).
12. Chen, H., Patterson, N. & Reich, D. Population differentiation as a test for selective sweeps. *Genome Res.* **20**, 393–402 (2010).
13. 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).
14. Purugganan, M.D. & Fuller, D.Q. Archaeological data reveal slow rates of evolution during plant domestication. *Evolution* **65**, 171–183 (2011).
15. Olsen, K.M. *et al.* Selection under domestication: evidence for a sweep in the rice *waxy* genomic region. *Genetics* **173**, 975–983 (2006).
16. Palaisa, K., Morgante, M., Tingey, S. & Rafalski, A. Long-range patterns of diversity and linkage disequilibrium surrounding the maize *Y1* gene are indicative of an asymmetric selective sweep. *Proc. Natl. Acad. Sci. USA* **101**, 9885–9890 (2004).
17. Camus-Kulandaivelu, L. *et al.* Maize adaptation to temperate climate: relationship between population structure and polymorphism in the *Dwarf8* gene. *Genetics* **172**, 2449–2463 (2006).
18. Brown, P.J. *et al.* Distinct genetic architectures for male and female inflorescence traits of maize. *PLoS Genet.* **7**, e1002383 (2011).
19. Buckler, E.S. *et al.* The genetic architecture of maize flowering time. *Science* **325**, 714–718 (2009).
20. Gallais, A. & Hirel, B. An approach to the genetics of nitrogen use efficiency in maize. *J. Exp. Bot.* **55**, 295–306 (2004).
21. Jackson, D. & Hake, S. Control of phyllotaxy in maize by the *abphyl1* gene. *Development* **126**, 315–323 (1999).
22. Gualberti, G. *et al.* Mutations in the Dof zinc finger genes *DAG2* and *DAG1* influence with opposite effects the germination of *Arabidopsis* seeds. *Plant Cell* **14**, 1253–1263 (2002).
23. Sasaki, A. *et al.* Green revolution: a mutant gibberellin-synthesis gene in rice—new insight into the rice variant that helped to avert famine over thirty years ago. *Nature* **416**, 701–702 (2002).
24. Wang, Y. *et al.* Molecular tailoring of farnesylation for plant drought tolerance and yield protection. *Plant J.* **43**, 413–424 (2005).
25. Eastmond, P.J. *SUGAR-DEPENDENT1* encodes a patatin domain triacylglycerol lipase that initiates storage oil breakdown in germinating *Arabidopsis* seeds. *Plant Cell* **18**, 665–675 (2006).
26. Sekhon, R.S. *et al.* Genome-wide atlas of transcription during maize development. *Plant J.* **66**, 553–563 (2011).
27. 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).
28. Stupar, R.M. *et al.* Gene expression analyses in maize inbreds and hybrids with varying levels of heterosis. *BMC Plant Biol.* **8**, 33 (2008).
29. Beadle, G.W. Teosinte and the origin of maize. *J. Hered.* **30**, 245–247 (1939).
30. Ryu, C.H. *et al. OsMADS50* and *OsMADS56* function antagonistically in regulating long day (LD)-dependent flowering in rice. *Plant Cell Environ.* **32**, 1412–1427 (2009).
31. Lo, S.F. *et al.* A novel class of gibberellin 2–oxidases control semidwarfism, tillering, and root development in rice. *Plant Cell* **20**, 2603–2618 (2008).

ONLINE METHODS

Sequencing and read mapping. As part of the maize HapMap2 project, sequence was generated and paired-end libraries were prepared in a subset of 75 of the 103 lines described in an accompanying paper[3]. Sampled accessions here included 35 improved maize lines, 23 traditional landraces and 17 wild relatives but excluded 28 tropical inbreds from CIMMYT sampled by Chia *et al.*[3]. Sequence depth for each line is detailed in **Supplementary Table 1**. Lines were sequenced to a mean depth of 5.05× (median of 4.58×), with mean coverage in analyzed regions of 3.74× (median of 3.70×). Paired-end sequence was aligned to the B73 reference genome (release 4a.53)[32], and SNPs were called using algorithms described in Chia *et al.*[3]. We further filtered SNPs by retaining only non-singleton, biallelic SNPs with ≤50% missing data across all three groups of interest (*parviglumis*, landraces and improved lines), resulting in a final data set of 21,141,953 SNPs.

Error rate estimation. Per-nucleotide error rates for the HapMap2 data set were estimated to be ~0.1% compared to BACs that were sequenced with Sanger sequencing, on par with the ~0.12% error rate at loci across the genome in Sanger resequencing data[3]. To estimate the genotypic error rate in our subset of genotyped SNPs, SNP calls from the maize HapMap2 and maize Illumina Infinium 55K chip data sets were compared for a subset of 66 lines (SNP data for the lines genotyped with Infinium 55K are available at PanZea; see URLs). The Infinium 55K data set comprised 51,584 SNPs, of which 37,279 were in common with the subset of HapMap2 SNPs used in our study. Out of 2.2 million comparable genotypes, 2.45% had different SNP calls in the two data sets. Most of the genotypic discrepancies (88%) consisted of a heterozygous versus homozygous call; hence, if we assume no errors in the Infinium 55K data set, the mean and median homozygous error rates in our data set per line (0.31% and 0.20% respectively) were much lower than overall genotypic error rates, and the actual per-allele error rate was ~1%. Most SNPs (59.1%) had a genotypic error rate of 2% or less, 30.2% of SNPs had no errors, and only 1.5% of SNPs had an error rate higher than 10%. Both the genotypic and homozygous error rates followed a unimodal distribution with a mode of zero, and there was no obvious second class of SNPs with poor performance with respect to error rates. B73 had the lowest genotypic error rate (0.87%) and the third lowest homozygous error rate (0.12%); this suggests that most of the genotyping errors in the HapMap2 data resulted from alignment issues rather than raw sequencing errors. The four teosinte lines with the highest homozygous error rates (TIL01, TIL07, TIL08 and TIL09) were represented in the two data sets by individuals of different selfing generations (for example, S5 versus S8 for TIL08).

Genome scan for selection. We performed a genome scan using a composite likelihood approach (XP-CLR)[12] modified to incorporate missing data (code available on request). Evidence for selection across the genome during domestication and improvement was evaluated in two contrasts: landraces versus *parviglumis* for domestication and improved lines versus landraces for improvement. Our scan used a 0.05-cM sliding window with 100-bp steps across the whole genome. Individual SNPs were assigned a position along the genetic map[33] by assuming uniform recombination between mapped markers. To ensure comparability of the composite likelihood score in each window, we fixed the number of SNPs assayed in each window to 50. Following the described protocol[12], we down-weighted pairs of SNPs in high LD ($r^2 > 0.70$) to minimize the effect of dependence on the composite likelihood score. Final estimates were tabulated in nonoverlapping 10-kb windows across the genome, assigning each 10-kb window the mean likelihood score (XP-CLR) and selection coefficient (s) estimated by the method[12]. To partially account for the non-independence of XP-CLR scores along the physical map, we grouped regions into putatively selected features. Features were defined as groups of 10-kb windows with XP-CLR values above the genome-wide 80th percentile not containing more than one window below this threshold. Features falling within 0.05 cM of functional centromeres[34] and an inversion on chromosome 1 (**Supplementary Note**) were masked from subsequent analyses. Our analyses of regions selected during domestication and improvement focused on features in the highest 10th percentile of mean feature-wise XP-CLR. However, we applied a more stringent criterion for identifying candidate genes, drawing only from features in which the observed reduction in nucleotide diversity during domestication or improvement was lower than the median observed from 1,000 random windows of similar width and nucleotide diversity. We assigned the maize FGS gene closest to the window with the maximum XP-CLR value as the most likely candidate.

Population genetics analyses. Individual SNPs for each gene were classified as noncoding, synonymous coding or nonsynonymous coding on the basis of annotations of the first transcript in the FGS. Standard population genetics summary statistics (π, ρ, F_{ST}, Tajima's D and normalized Fay and Wu's H) were calculated for nonoverlapping 10-kb windows across the genome and separately for individual genes in the FGS using a combination of custom scripts, programs written using the libsequence C++ library[35] and SAMtools[36]. In addition to statistics within groups, we calculated weighted F_{ST} (ref. 37) between groups and net pairwise divergence between *parviglumis* and the *Tripsacum* outgroup. We tested for outliers for summary statistics in candidate regions by comparing average values to a distribution calculated for randomly sampled nonoverlapping genomic regions of identical width. Site frequency spectra were rescaled to address missing data and differing sample sizes (**Supplementary Note**).

To estimate mean haplotype lengths in each group, we used a custom Perl script to choose 1 million starting points uniformly across the genome. At each point, we chose two random lines from within a group (*parviglumis*, landraces or improved lines). We compared the two lines' genotypes at the focal point, extending outward in both directions until we found different genotypes. Missing data and heterozygous SNPs were ignored.

Historical recombination rates ($\rho = 4N_e$) were estimated using a described composite likelihood approach[38]. In estimating recombination, we treated the data as haploid, coding all heterozygous sites as missing data. We then removed all SNPs with minor allele counts of ≤2. For windows with ≥10 remaining SNPs, values of ρ per basepair were estimated across a grid of values from 1×10^{-4} to 0.2, assuming no homologous gene conversion.

Expression analyses. Three separate data sets were used to assess patterns of gene expression in maize and *parviglumis* transcriptomes. First, using a custom long oligonucleotide microarray[39] designed by NimbleGen (GPL10846), we characterized variation in gene expression in a subset of 25 improved maize and 7 *parviglumis* inbred lines (**Supplementary Table 1**). Multiple replicates of the maize inbred lines B73 and Mo17 were included to assess consistency. Plants were grown, and seedling leaf tissue was harvested 8 d after germination. RNA was isolated using TRIzol (Invitrogen, 15596026) from above-ground tissue and purified by lithium chloride treatment and precipitation with 3 M sodium acetate (0.1 vol) and 95% ethanol (2.5 vol). Purified RNA (10 µg) was reverse transcribed and labeled according to the array manufacturer's protocol. For each sample, ~20 µg of Cy3- or Cy5-labeled RNA was hybridized to the array slide for 16–20 h at 42 °C using the NimbleGen Hybridization System. After hybridization, slides were washed (NimbleGen Wash Buffer Kit) and dried for 2 min by centrifugation. Slides were immediately scanned using the GenePix 4000B Scanner (Molecular Devices) according to the array manufacturer's protocol.

Array images and data were processed using NimbleScan software. Briefly, images from each slide were separated into 12 subarrays and aligned to a grid to extract signal intensity for each feature on the array. Experimental integrity was verified by evaluation of the signal intensities of the sample tracking control features for each subarray. In addition, metrics reports were produced for each array to describe signal uniformity across the array and the intensities of known empty features, random probes and experimental probes. Signal-to-noise ratios were estimated by dividing the average signal intensity of experimental gene probes by that of the control probes. Only slides with a signal-to-noise ratio of ≤2 were retained. NimbleScan was used to generate robust multi-array average (RMA)[40] gene expression values from the spatially corrected probe signal intensities on a per-probe and per-gene basis. Normalized gene expression values across multiple replications (technical or biological) of the same genotype were averaged when possible. Comparisons of the distributions of signal intensity for control and experimental probes (**Supplementary Fig. 15**) were used to determine a reasonable signal threshold for positive expression across all slides. Genes with average probe $\log_2$ signals of >10 in at least 3 arrays were retained as expressed (N = 19,792 expressed genes).

doi:10.1038/ng.2309

Nucleotide polymorphism may contribute to differences in hybridization between transcripts from divergent genotypes, although previous results suggest that as many as four or five SNPs are needed to strongly affect probe hybridization[41]. To minimize the impact of polymorphism on hybridization, we further filtered the probe set on the basis of a previous comparative genomic hybridization (CGH) data set developed using many of the same genotypes and the same array platform[39]. Probes that showed substantially reduced CGH values for at least 3 genotypes (26,937 probes) were removed, resulting in a set of 46,167 probes that detected expression of 18,242 genes with 1–4 probes per gene. This subset of the data was used for all subsequent analyses. Finally, a Spearman's rank correlation showed a weakly positive correlation between *parviglumis* expression and F_{ST} *parviglumis*-maize ($\rho = 0.043$), providing no evidence for hybridization bias.

The second data set comprised expression estimates in 60 different tissues of the inbred B73 line from a NimbleGen array of 23,740 genes in the FGS[26]. These data were curated and categorized following described conventions[26].

The third data set consisted of expression data from five improved lines and their F_1 hybrids characterized using an Affymetrix GeneChip Maize Genome Array (GMGA)[28]. Probe sets with fewer than two biological replicates were removed, and the remaining probe sets were annotated by mapping array probes to the B73 reference genome (release 4a.53)[42]. Probe sets were included in the analysis only if they mapped to a single gene. When multiple probe sets mapped to a single gene, expression data from all probe sets were averaged for further analysis.

Significant differences between candidate and non-candidate expression values were determined by bootstrap resampling of $\log_2$-transformed, RMA-normalized data from non-candidate genes. The validity of a bootstrap approach was assessed by plotting the mean of each bootstrap sample against its variance, in order to confirm that calculated test statistics were pivots[43]. For tissue-specific expression, bootstrap significance values were adjusted for multiple tests with a Benjamini-Hochberg false discovery rate correction at 0.05. Dominance was assessed as (hybrid expression signal – mid-parent expression signal) / (range of parental expression / 2).

Finally, in order to ensure that the relatively low coefficient of variation in expression observed in candidates did not result in an inflation of estimates of dominance, an analysis of covariance was conducted with the coefficient of variation included as a covariate.

32. Schnable, P.S. *et al.* The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112–1115 (2009).
33. McMullen, M.D. *et al.* Genetic properties of the maize nested association mapping population. *Science* **325**, 737–740 (2009).
34. Wolfgruber, T.K. *et al.* Maize centromere structure and evolution: sequence analysis of centromeres 2 and 5 reveals dynamic loci shaped primarily by retrotransposons. *PLoS Genet.* **5**, e1000743 (2009).
35. Thornton, K. libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics* **19**, 2325–2327 (2003).
36. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
37. Hudson, R.R., Boos, D.D. & Kaplan, N.L. A statistical test for detecting geographic subdivision. *Mol. Biol. Evol.* **9**, 138–151 (1992).
38. Hudson, R.R. Two-locus sampling distributions and their application. *Genetics* **159**, 1805–1817 (2001).
39. Swanson-Wagner, R.A. *et al.* Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Res.* **20**, 1689–1699 (2010).
40. Irizarry, R.A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264 (2003).
41. Springer, N.M. *et al.* Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genet.* **5**, e1000734 (2010).
42. Lawrence, C.J., Dong, O.F., Polacco, M.L., Seigfried, T.E. & Brendel, V. MaizeGDB, the community database for maize genetics and genomics. *Nucleic Acids Res.* **32**, D393–D397 (2004).
43. Davison, A.C. & Hinkley, D.V. *Bootstrap Methods and Their Application* (Cambridge University Press, New York, 1997).