nature genetics

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41588-018-0158-0

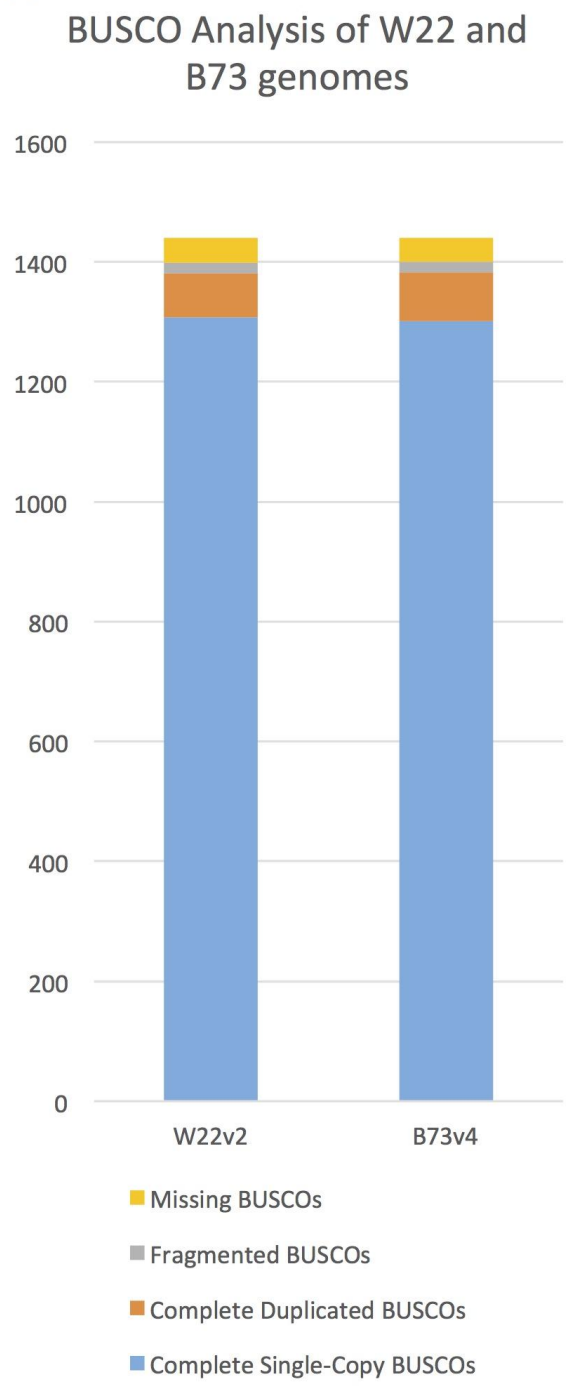
In the format provided by the authors and unedited.

The maize W22 genome provides a foundation for functional genomics and transposon biology

Nathan M. Springer[1], Sarah N. Anderson[1], Carson M. Andorf[2], Kevin R. Ahern[3], Fang Bai[4], Omer Barad[5], W. Brad Barbazuk[6], Hank W. Bass[7], Kobi Baruch[5], Gil Ben-Zvi[5], Edward S. Buckler[8,9], Robert Bukowski[9], Michael S. Campbell[10], Ethalinda K. S. Cannon[2], Paul Chomet[5], R. Kelly Dawe[11], Ruth Davenport[6], Hugo K. Dooner[12,13], Limei He Du[12,13], Chunguang Du[14], Katherine A. Easterling[7], Christine Gault[6], Jiahn-Chou Guan[4], Charles T. Hunter[15], Georg Jander[3], Yinping Jiao[10], Karen E. Koch[4], Guy Kol[5], Tobias G. Köllner[16], Toru Kudo[4,17], Qing Li[1], Fei Lu[9,18,19], Dustin Mayfield-Jones[20], Wenbin Mei[6], Donald R. McCarty[4], Jaclyn M. Noshay[1], John L. Portwood II[2], Gil Ronen[5], A. Mark Settles[4], Doron Shem-Tov[5], Jinghua Shi[21], Ilya Soifer[5], Joshua C. Stein[10], Michelle C. Stitzer[22], Masaharu Suzuki[4], Daniel L. Vera[23], Erik Vollbrecht[24], Julia T. Vrebalov[3], Doreen Ware[8,10,25], Sharon Wei[10], Kokulapalan Wimalanathan[24], Margaret R. Woodhouse[2], Wenwei Xiong[14] and Thomas P. Brutnell[20,26]*

[1]Department of Plant and Microbial Biology, University of Minnesota, Saint Paul, MN, USA. [2]USDA-ARS, Corn Insects and Crop Genetics Research Unit and Iowa State University, Department of Computer Science, Iowa State University, Ames, IA, USA. [3]Boyce Thompson Institute, Ithaca, NY, USA. [4]Horticultural Sciences Department, University of Florida, Gainesville, FL, USA. [5]NRGene Ltd, Ness Ziona, Israel. [6]Department of Biology and the UF Genetics Institute, University of Florida, Cancer & Genetics Research Complex, Gainesville, FL, USA. [7]Department of Biological Science, The Florida State University, Tallahassee, FL, USA. [8]USDA-ARS, Holley Center for Agriculture and Health, Ithaca, NY, USA. [9]Institute for Genomic Diversity, Biotechnology Building, Cornell University, Ithaca, NY, USA. [10]Cold Spring Harbor Laboratory, Cold Springs Harbor, NY, USA. [11]Department of Plant Biology, University of Georgia, Athens, GA, USA. [12]Department of Plant Biology, Rutgers University, New Brunswick, NJ, USA. [13]Waksman Institute, Rutgers University, Piscataway, NJ, USA. [14]Department of Biology, Montclair State University, Montclair, NJ, USA. [15]USDA-ARS Chemistry Research Unit, Gainesville, FL, USA. [16]Department of Biochemistry, Max Planck Institute for Chemical Ecology, Jena, Germany. [17]Metabologenomics, Inc., Tsuruoka, Yamagata, Japan. [18]CAS-JIC Centre of Excellence for Plant and Microbial Science (CEPAMS), Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China. [19]The State Key Laboratory of Plant Cell and Chromosome Engineering, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China. [20]Donald Danforth Plant Science Center, St. Louis, MO, USA. [21]Bionano Genomics, San Diego, CA, USA. [22]Department of Plant Sciences and Center for Population Biology, University of California, Davis, Davis, California, USA. [23]Center for Genomics and Personalized Medicine, The Florida State University, Tallahassee, FL, USA. [24]Department of Genetics, Development and Cell Biology, Iowa State University, Ames, IA, USA. [25]USDA-ARS, NEA Robert W. Holley Center for Agriculture and Health, Cornell University, Ithaca, NY, USA. [26]Present address: College of Agronomic Sciences State Key Laboratory of Crop Biology, Shandong Agricultural University, Shandong, China. *e-mail: brutnell@gmail.com

A



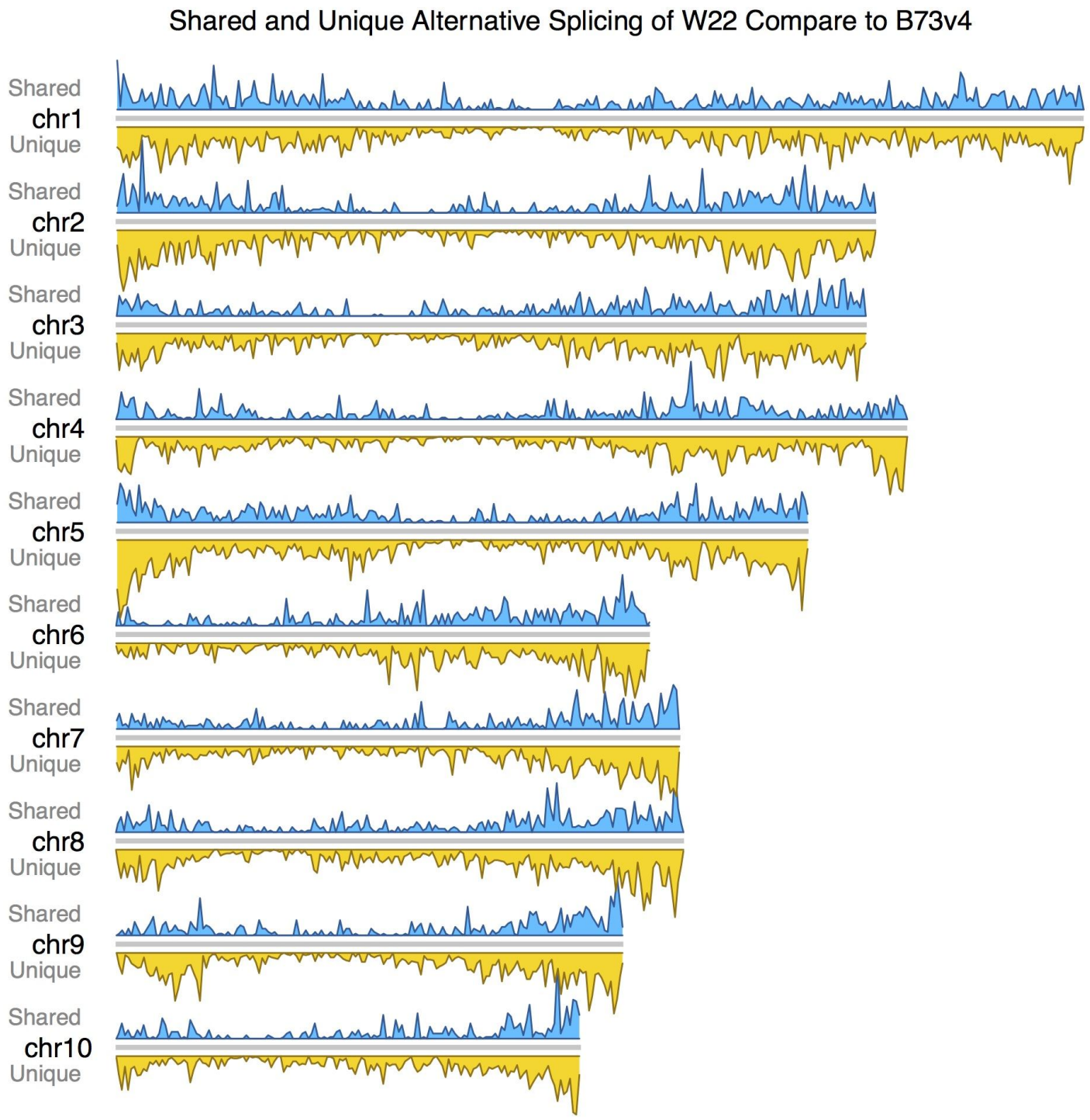
B

Gene Type	W22v2	B73v4
Complete Single-Copy BUSCOs	1307	1302
Complete Duplicated BUSCOs	75	81
Fragmented BUSCOs	16	18
Missing BUSCOs	42	39
Total BUSCO groups searched	1440	1440

Supplementary Figure 1

Comparison of the completeness of the B73v4 and W22v2 genome annotations.

The completeness of the genome assemblies and genome annotations was assessed by benchmarking a universal single-copy orthologous gene set (BUSCO) (Simão et al. 2015). The relative frequencies of complete single-copy genes, duplicated genes, fragmented genes and missing genes were very similar in the B73v4 and W22v2 genomes.



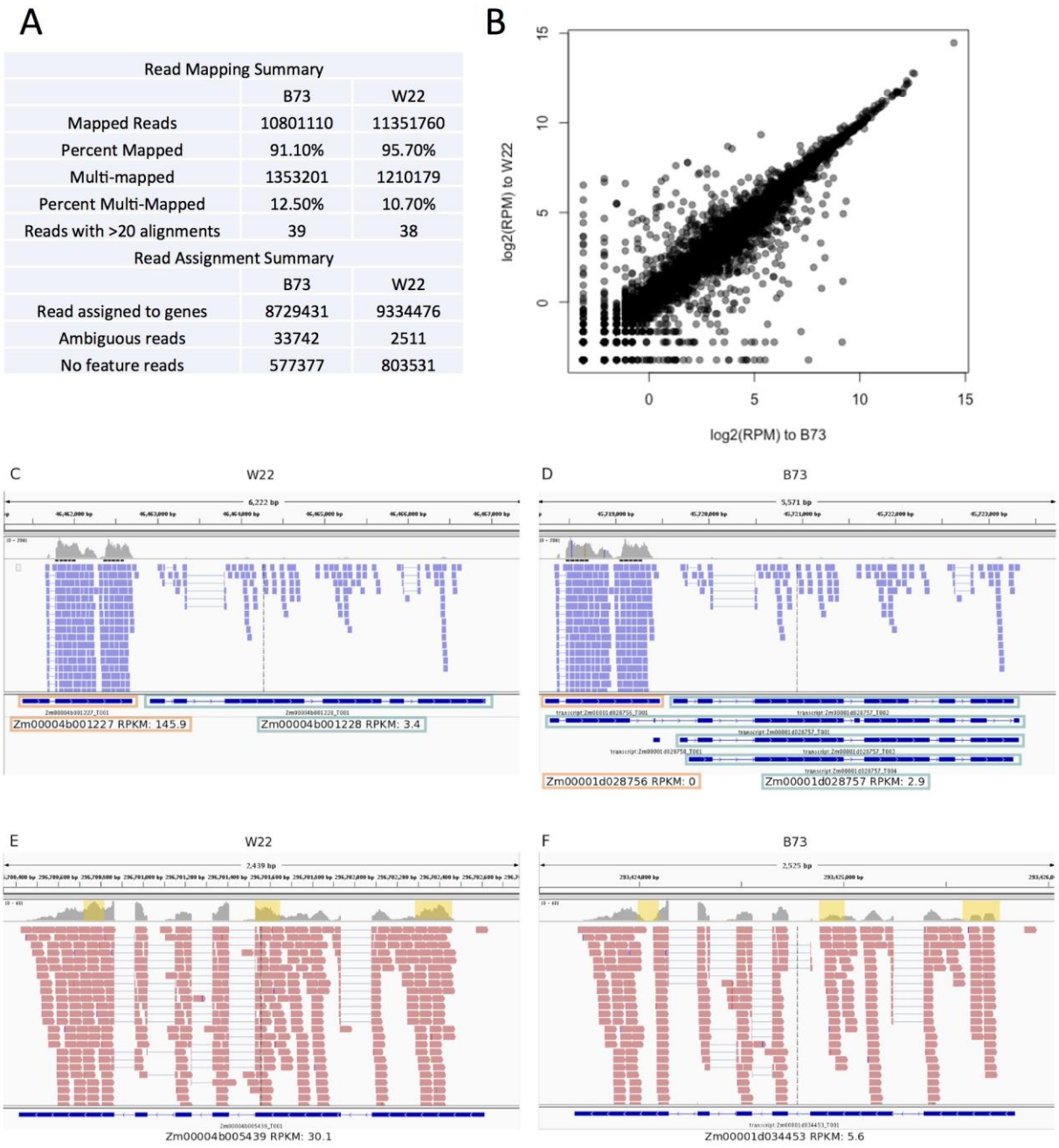
Supplementary Figure 2

Density plot of shared and unique alternative splicing of W22 compared to B73v4.

For the B73v4 alternative splicing events, we mapped the isoforms from B73v4 annotation to the W22 genome and identified the alternative splicing events. We classified the common alternative splicing events based on the coordinates of the alternative region relative to the W22v2 genome.

A

Read Mapping Summary		
	B73	W22
Mapped Reads	10801110	11351760
Percent Mapped	91.10%	95.70%
Multi-mapped	1353201	1210179
Percent Multi-Mapped	12.50%	10.70%
Reads with >20 alignments	39	38
Read Assignment Summary		
	B73	W22
Read assigned to genes	8729431	9334476
Ambiguous reads	33742	2511
No feature reads	577377	803531

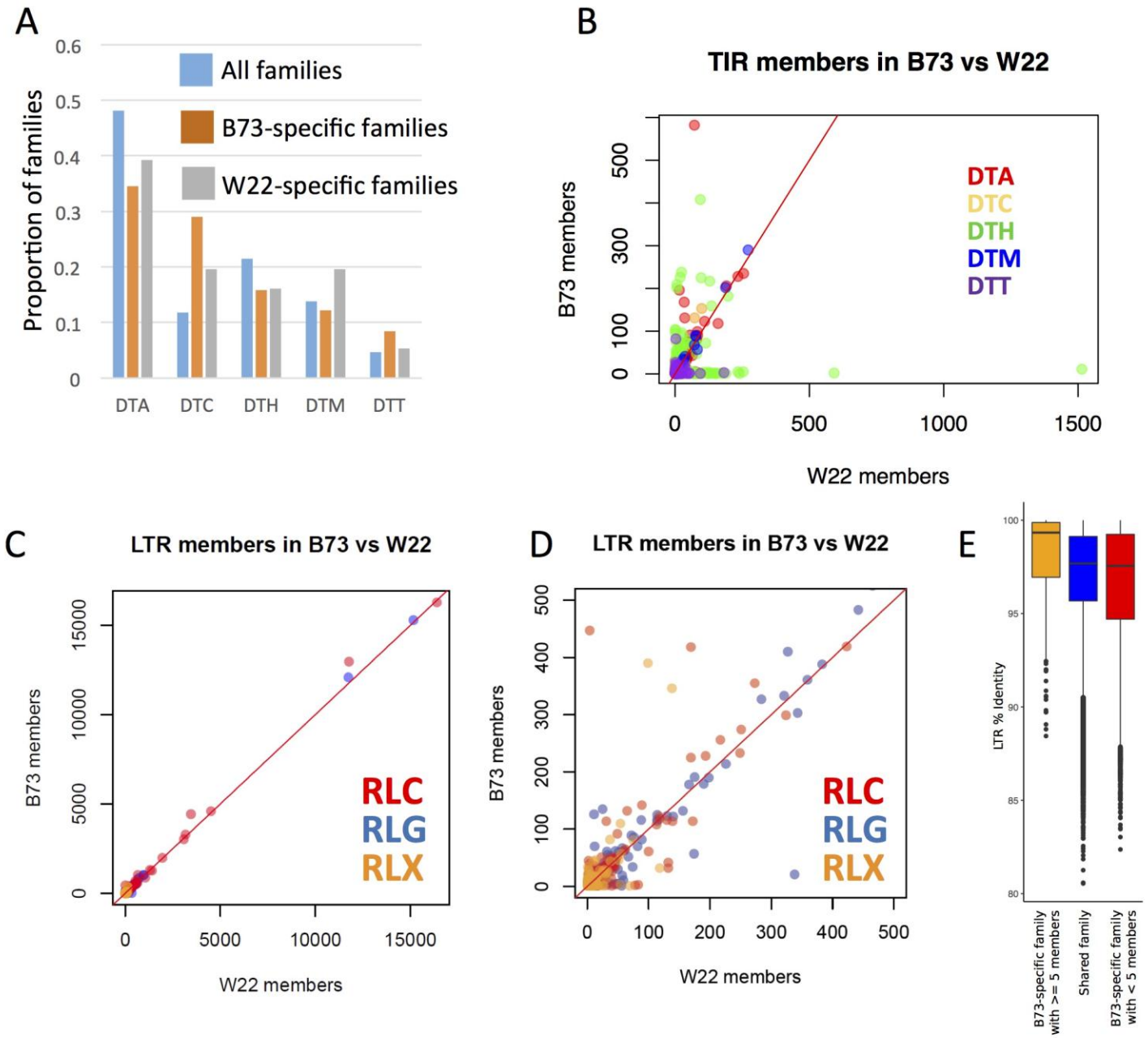


Supplementary Figure 3

Improved transcriptome analyses using the W22v2 genome.

a, RNA-seq data derived from endosperm tissue of W22 (SRA: SRR1986376) were aligned to both B73v4 and W22v2 using the same parameters. The percentage of mapped reads improves by mapping the data to W22. **b**, A set of 20,994 syntenic orthologous genes in B73 and W22 were identified and used for comparison of expression levels in alignments to B73 or W22. A comparison of expression level (reads per million, RPM) shows generally similar estimates in both genotypes with some genes that have differing expression estimates depending upon which genome was used for a reference. **c**–**f**, Differences in RPKM estimates can result from differences in annotation or mapping efficiency. **c**,**d**, Overlapping gene models in B73 result in all reads mapping to Zm00001d028756 (orange transcript) to be called ambiguous, while the corresponding gene in W22, Zm00004b001227 (orange transcript), has reads assigned. The adjacent gene, Zm00004b001228 in W22 and Zm00001d028757 in B73 (teal transcripts), has the same number of reads assigned to each reference, with a lower RPKM value reported in B73 owing to the longer gene model. **e**,**f**, The RPKM value for genes

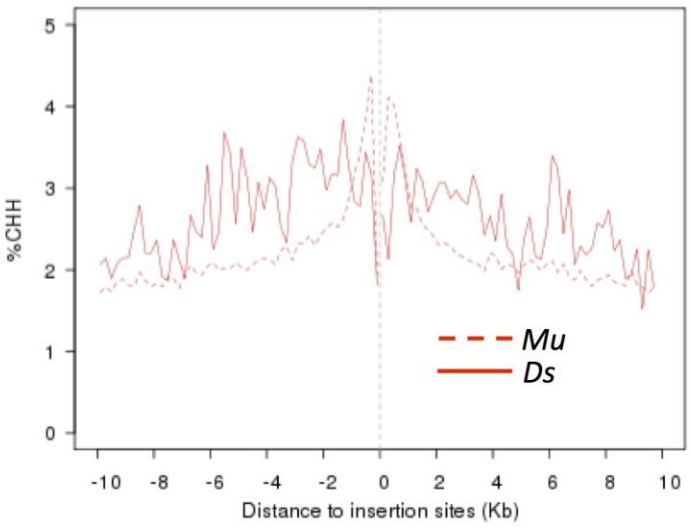
Zm00004b005439 in W22 and Zm00001d034453 in B73 is higher when mapping to W22 (**e**) than to B73 (**f**) owing to improved alignment to several regions of the gene (marked in yellow). Mapped reads are colored by strand: blue, forward; red, reverse.



Supplementary Figure 4

Comparison of TIR copy number in B73 and W22.

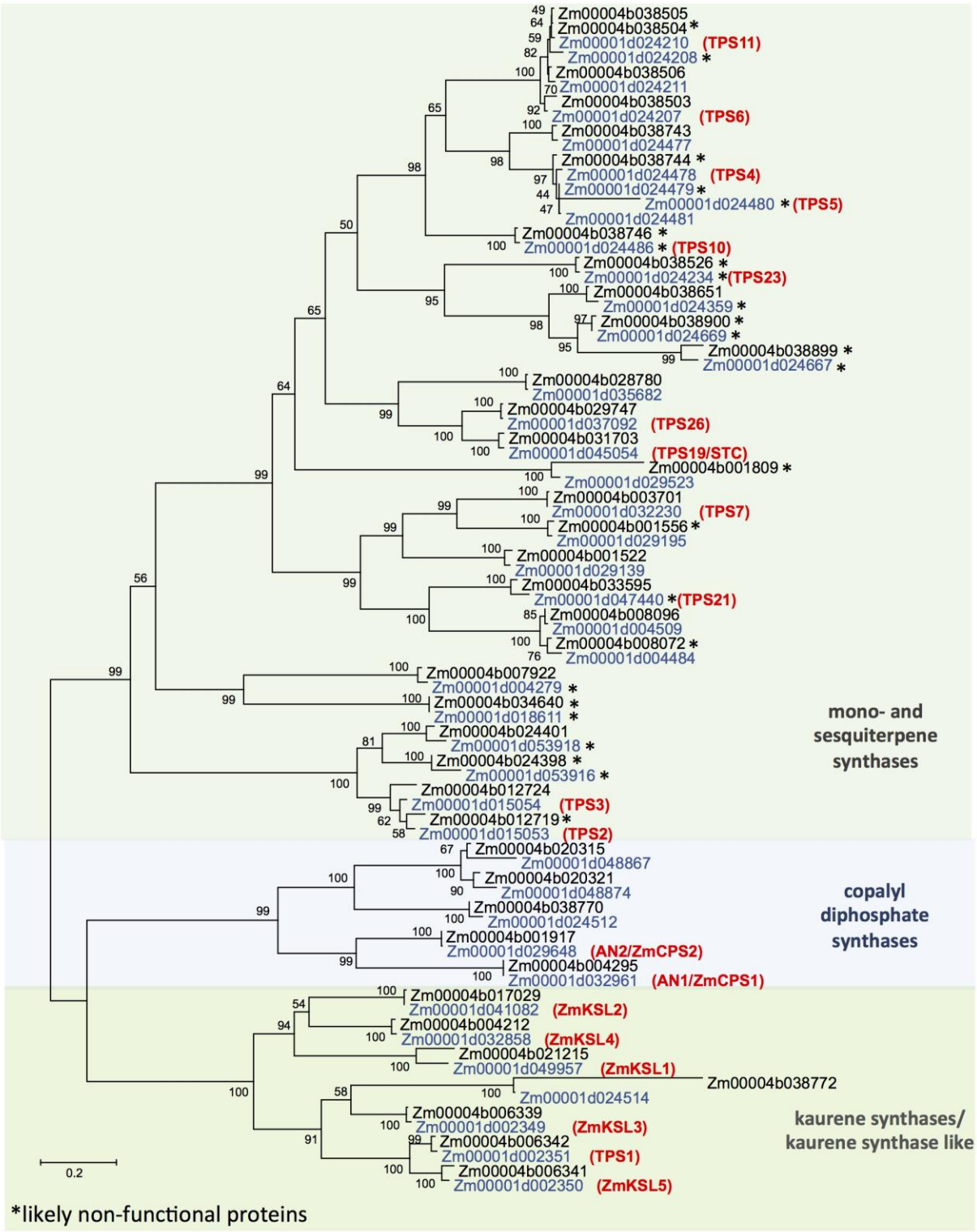
a, The proportion of TIR families in each of the superfamilies (Activator (DTA); CACTA (DTC), PIF/Harbinger (DTH), Mutator (DTM), Tourist (DTT)) was determined for all TIR families (blue) in the B73 and W22 genomes. The proportion of TIR families in these categories was then determined for B73-specific (orange) and W22-specific (gray) families. **b**, The copy number in each TIR TE family is shown for B73 and W22. Color indicates superfamilies. **c**, The relative copy number for each LTR TE family in B73 and W22 is shown. In **d**, only families with <500 copies are shown. **e**, Boxplot of the percent identity of LTR sequences for LTR retrotransposons, demonstrating that elements in B73-specific families with at least five members (orange) are younger than members of shared families (blue) and elements in B73-specific families with fewer than five members (red). Line, median; box limits, first and third quartiles; whiskers, furthest point within 1.5 * IQR; points, outliers.



Supplementary Figure 5

Profiles of CHH methylation surrounding sites targeted by *Ds* (solid lines) or *Mu* (dashed lines).

The level of CHH methylation is shown for the flanking regions (up to 10 kb) near *Ds* or *Mu* insertion sites.

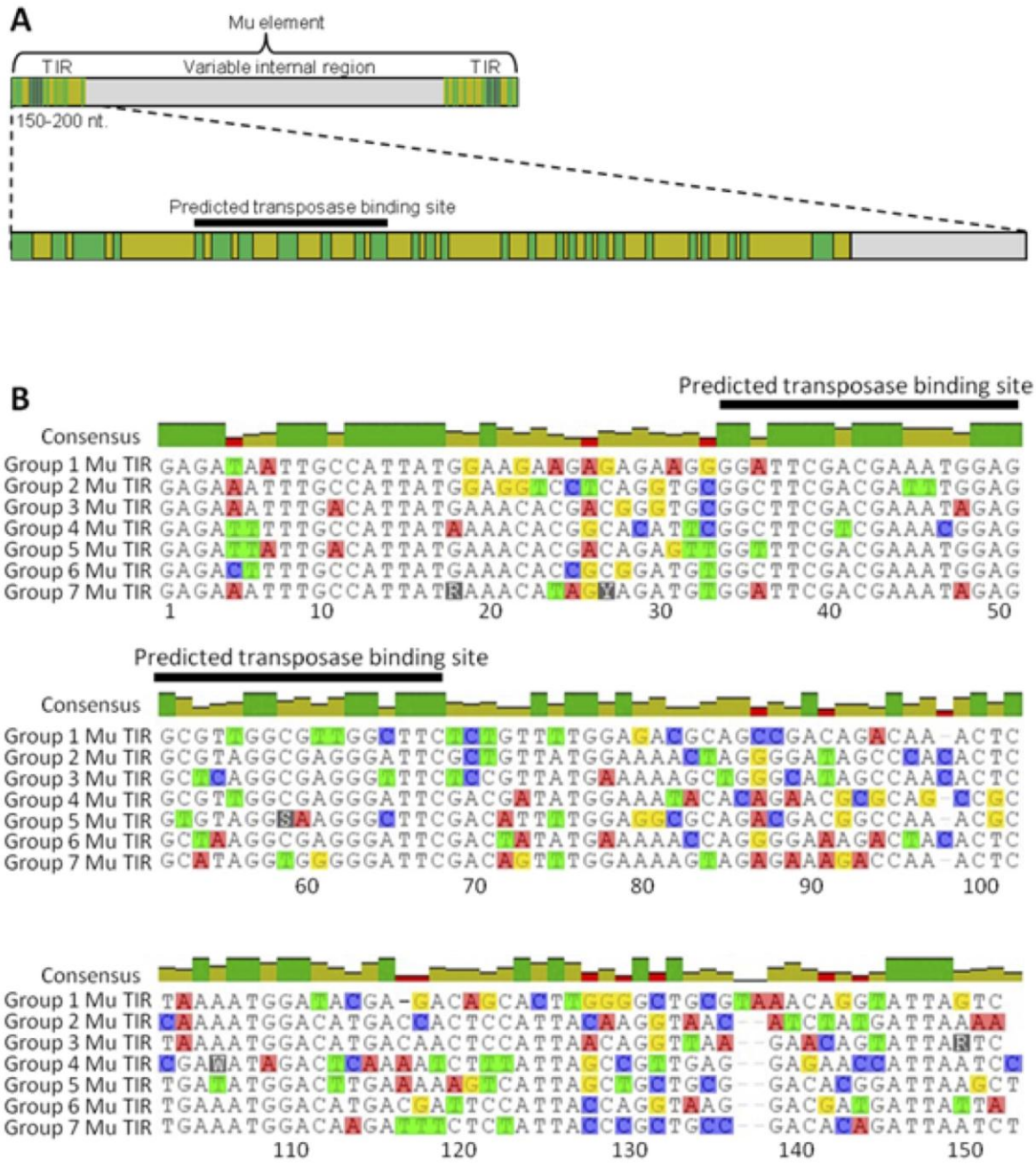


Supplementary Figure 6

Dendrogram analysis (unrooted tree) of terpene synthases in B73 and W22.

A set of 81 amino acid sequences (42 from B73 in blue and 39 from W22 in black) were used to generate a tree based on MUSCLE protein alignment by using the Maximum Likelihood method and a previously described substitution model. Bootstrap values (n = 1,000 replicates) are shown next to each node. The tree is drawn to scale, with branch lengths measured in the number of substitutions per

site. All positions with less than 80% site coverage were eliminated. Evolutionary analyses were conducted in MEGA6. Asterisks are used to mark likely non-functional genes.



Supplementary Figure 7

Sequence conservation among Mutator transposable elements.

a, Diagram of a Mutator transposable element. The highly conserved terminal inverted repeats (TIRs) can be used to identify Mu elements and to categorize them based on phylogenetic groups. **b**, Alignment of consensus sequences for each of the seven Mu TIR groups. Highlighted nucleotides indicate disagreements with the overall consensus sequence. The terminal 20 positions are the most conserved across TIR groups. The predicted transposase-binding site (from position 34 to 68) is also highly conserved (57%) across the seven groups.

Supplementary Figure 8

Analysis of synteny between *Mutator* transposons identified in W22 and B73, including intact elements and orphan TIRs.

Of the 386 Mu elements or orphan TIRs examined, 133 were present in both the W22 and B73 genomes.

Supplementary Note

***Characterization and screening of gene models*:**

The "working set" of gene models (n=40,690 loci) was subjected to several analyses to distinguish high-confidence genes from transposon-encoded loci and other dubious annotations. MAKER-P calculates an annotation evidence distance (AED) for each model that scores how well the model is supported by its evidence (a range between 0 and 1, with lower scores indicating higher support) [1]. To gain greater knowledge of putative function of annotated loci, all predicted proteins were annotated using InterProScan (v5) [2], following default parameters.

Screens for transposon-encoded genes: Probable transposable element (TE) genes were identified using two screens. First, we tagged loci whose longest predicted coding region (CDS) overlapped with repeat-masked regions by more than 40% of length. Such annotations can arise from from evidence that seeded in non-masked regions but subsequently extended into masked regions. Second, loci with the following InterPro domains were tagged as probable TE: IPR000477, IPR004252, IPR004264, IPR004332, IPR005162, IPR007321, IPR008906, IPR009227, IPR013103, IPR013242, IPR018289, IPR025476, IPR026960, IPR027806.

Comparative genomics analysis: Sequence homology and conserved synteny within related species is suggestive of genetic function and can provide a measure of confidence in the validity of predicted genes. We applied the Ensembl Compara phylogenetic gene tree pipeline [3,4] to define homologies within the W22 working set and identify orthologous and paralogous relationships with related grass and other plant species. Additional representative genome annotations included those of Zea mays (B73 RefGen_v4), Oryza sativa (IRGSP-1.0), Sorghum bicolor (JGI v2.0), Setaria italica (JGI v2.0), Brachypodium distachyon (JGI v1.0), and three dicot species, Arabidopsis thaliana (TAIR10), Glycine max (JGI v1.0), and Vitis vinifera (CRIBI

V1). The analysis was performed with Ensembl software release 86; online documentation provides further details of the protocol used ("Protein Trees and Orthologies" 2017). Synteny maps relating collinear or near-collinear orthologous genes were constructed between all pairwise combinations of W22, B73, rice, sorghum, Setaria, and Brachypodium using previously described methods [5,6]. This enabled the categorization of W22 annotations, after excluding probable TE, into the following classes based on evolutionary conservation, 1) syntelogs (having conserved ancestral chromosomal position with orthologs in another grass species), 2) synteny with B73 only (which may include loci from maize-specific families), 3) non-syntenic orthologs (having orthologs at non-conserved position in other grasses), and 4) non-orthologs (including W22-specific and maize-specific loci).

Fragmented loci: Putative fragmented loci, which may represent pseudogenes or artifacts from incorrect annotation or misassembly, were identified in two screens. First, we identified gene models that appeared to lack a complete CDS, by absence of a methionine start codon or a stop codon, in all of its transcript isoforms. Second, for those models having an ortholog in B73 or other grass, we looked for extreme deviations of its predicted longest protein length from the average coding length of its orthologs. Those with a z-score less than -2 (e.g. length greater than two standard deviations shorter than the ortholog mean) were also tagged as putative fragmented loci.

Analysis of local duplications in W22 and B73

The frequency of locally duplicated genes is comparable in B73 and W22, but W22 (14.73%) had slightly more than B73 (~14.08%) (Supplementary Table 3). Both genomes had more duplicated genes in tandem (i.e., no intervening genes) (~63.9% in B73 and ~56.4% in W22) compared to the sum of all other local duplication classes (i.e., with 1 to 20 intervening genes) (Supplementary Table 3). B73 had more tandemly duplicated genes, but W22 had more genes

in other locally duplicated classes. Moreover, the proportional increase in the number of non-tandem locally duplicated genes in W22 genome compared to B73 is positively correlated to the number of intervening genes between local duplication events (Supplementary Table 3). In many cases locally duplicated genes form arrays of similar genes, which indicates that a single ancestral gene has been copied multiple times. The current analysis cannot determine the nature and timing of these multiplication events, but it can delineate the current number of gene copies in each multiplication cluster. The overall number of multiplication clusters is comparable between W22 and B73 (~27% of total multiplication events), although W22 had slightly more (2.3%) clusters (Supplementary Table 3). As the number of gene copies in a cluster increases, the number of clusters decreases (Supplementary Table 4). Only a few clusters have more than 10 copies, and the highest cluster sizes are 21 for W22 and 20 for B73. Tandem duplicated genes were also classified based on the master list of ortholog mappings between W22 and B73, which revealed that B73 had a higher number (133 duplications or ~6.86% more) of tandem duplications (Supplementary Table 5), and that fewer tandem duplications are shared between genomes than are unique to one or the other (Supplementary Table 5). Unique tandem duplications predominate and could reflect PAVs or genes that have diverged beyond recognition by the current methodology. A larger proportion of both the shared and unique tandem duplications were in the same (head-tail) orientation, which means that both genes are in the same strand, whereas the number of divergent (head-head) and divergent (tail-tail) are comparable, and both these orientations mean that the genes are not on the same strand.

Example of functional implications of local duplication for terpene synthase

The terpene synthases of B73 and W22 were assessed in detail (Supplementary Figure 6). The analysis involved 81 amino acid sequences, 42 predicted terpene synthases from B73 and 39 from W22. Protein sequences consisting of less than 500 amino acids, which are likely to be non-functional [7], are included in the analysis and are marked with asterisks in Figure S6. B73

terpene synthases are shown in blue and W22 terpene synthases are shown in black.

There is a one-to-one correspondence of B73 and W22 copalyl diphosphate synthases and kaurene synthases, with no obviously non-functional proteins. However, mono- and sesquiterpene synthases show significant variation between the two inbred lines. Based on having less than 500 amino acids in the predicted protein length, 13 of the 30 B73 proteins and 12 of the 27 W22 proteins may be non-functional. Although some of these shortened proteins may be the result of incorrect annotation, non-functional terpene synthase pseudogenes have been identified previously in maize. Six terpene synthases are present in B73 but are absent or probably non-functional in W22. Three terpene synthases are present in W22 but probably non-functional in B73. Six terpene synthases are likely to be non-functional in both B73 and W22. This relatively large amount of genetic variation between two maize inbred lines is likely reflective of a much greater diversity in the biosynthesis of mono- and sesquiterpenes in maize as a species.

The TPS2/TPS3 sub-tree provides an example where genetic variation facilitated the identification of a knockout mutation for investigating in vivo protein function. B73 has tandem-duplicated TPS2 and TPS3 genes, which encode two proteins with 95% identity at the amino acid sequence level that catalyze the synthesis of linalool, (E)-nerolidol, and (E,E)-geranyllinalool [8]. In contrast, W22 has only one such gene, Zm00004b012724, which is similar to TPS3. The more TPS2-like Zm00004b012719 is a truncated pseudogene in W22. Due to this natural gene knockout in W22, it was possible to identify a *Ds* transposon knockout mutation of Zm00004a053478, thereby confirming not only the in vivo function in terpene production, but also a role for this enzyme activity in maize-insect interactions [9].

Detailed analysis of native Mutator elements in B73 and W22

Mutator (*Mu*) transposable elements are best classified by their highly conserved terminal inverted repeats (TIRs) due to extensive divergence among internal sequences [10-12] (Supplementary Figure 7). Here, we used known TIRs of *Mu* elements to query the B73 (v4) and W22 genomes. Phylogenetic analyses revealed 7 distinct clades of *Mu* TIRs, termed Group 1 through Group 7 (Supplementary Table 7). The Group-1 TIRs (96 in B73 and 99 in W22) included those from the mobile *Mu* elements in *Mu*-active populations derived from Robertson's Mutator [13-16]. Also in Group 1 are all but one of the *Mu* elements previously designated "*Mu1* through *Mu18*" [17-25]. The exception was "*Mu12*" [26], which has TIRs of phylogenetic Group 2. Consensus sequences for each group were generated by MUSCLE alignment [27] and are diagrammed in Supplementary Figure 7B. The predicted transposase binding site [28] is conserved (57% of nucleotides are identical across all seven TIR groups between positions 34 and 68). When these clade-specific TIR consensus sequences were used to query the B73 (v4) and W22 (v2) genomes, the two inbreds were found to have similar numbers of *Mu*-element TIRs across phylogenetic groups, as well as total Mu TIRs (Supplementary Table 7).

Individual TIRs within each genome were manually assigned partners (left and right arms) based on proximity to one another and on the presence of matching target-site duplications (TSDs) produced during *Mu*-element insertion (Supplementary Table 8). The majority (89%) of TIRs could be paired, resulting in intact *Mu* elements with left and right arms. The remaining "orphan" TIRs represent either TIRs that have lost a recognizable partner, or TIRs that occur as tandem duplications within an intact element. Synteny of *Mu*-element insertion sites between W22 and B73 was examined by comparing TSD sequences, TIR group ID's, and chromosome assignments for each *Mu* element and orphan TIR. Of the 257 *Mu* elements in W22, approximately half (133) were syntenic with B73 (Supplementary Figure 8).

Although the abundance and types of *Mutator* (*Mu*) transposable elements in B73 and W22 are similar, the individual identities and locations of *Mu* insertions in these genomes differ substantially. Both B73 and W22 carry comparable numbers of *Mu* transposons and also similar proportions of *Mu* insertions belonging to the seven, phylogenetically-distinct clades or "groups" (Supplementary Table 7, Supplementary Figure 8). Together, this conservation of total *Mu* numbers and their consistent phylogenetic distribution (Supplementary Table 6, Supplementary Figure 8) indicate that the observed pattern predated development of separate inbreds. However, differences in identity and location of individual *Mu* elements (Supplementary Table 8) are consistent with the probable diversity of *Mu* transposons present in the common ancestor of the two inbreds. This inference is consistent with a shared synteny of approximately 50% for specific *Mu*-insertions in both B73 and W22 (Supplementary Figure 8). It is tempting to speculate that the extent of non-syntenic *Mu* sites may correlate with other measures of genome diversity.

Supplementary Note references

1. Law, M. *et al.* Automated update, revision, and quality control of the maize genome annotations using MAKER-P improves the B73 RefGen_v3 gene models and identifies new genes. *Plant Physiol.* **167,** 25–39 (2015).
2. Finn, R. D. *et al.* InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res.* **45,** D190–D199 (2017).
3. Herrero, J. *et al.* Ensembl comparative genomics resources. *Database* **2016,** (2016).
4. Vilella, A. J. *et al.* EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* **19,** 327–335 (2009).
5. Schnable, P. S. *et al.* The B73 maize genome: complexity, diversity, and dynamics. *Science* **326,** 1112–1115 (2009).
6. Youens-Clark, K. *et al.* Gramene database in 2010: updates and extensions. *Nucleic Acids*

Res. **39,** D1085–94 (2011).

7. Degenhardt, J., Köllner, T. G. & Gershenzon, J. Monoterpene and sesquiterpene synthases and the origin of terpene skeletal diversity in plants. *Phytochemistry* **70,** 1621–1637 (2009).
8. Richter, A. *et al.* Characterization of Biosynthetic Pathways for the Production of the Volatile Homoterpenes DMNT and TMTT in Zea mays. *Plant Cell* **28,** 2651–2665 (2016).
9. Tzin, V. *et al.* Dynamic Maize Responses to Aphid Feeding Are Revealed by a Time Series of Transcriptomic and Metabolomic Assays. *Plant Physiol.* **169,** 1727–1743 (2015).
10. Chandler, V. L. & Hardeman, K. J. The Mu elements of Zea mays. *Adv. Genet.* **30,** 77–122 (1992).
11. Bennetzen, J. L. The Mutator transposable element system of maize. *Curr. Top. Microbiol. Immunol.* **204,** 195–229 (1996).
12. Lisch, D. Mutator transposons. *Trends Plant Sci.* **7,** 498–504 (2002).
13. Walbot, V. Saturation mutagenesis using maize transposons. *Curr. Opin. Plant Biol.* **3,** 103–107 (2000).
14. May, B. P. *et al.* Maize-targeted mutagenesis: A knockout resource for maize. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 11541–11546 (2003).
15. Settles, A. M. *et al.* Sequence-indexed mutations in maize using the UniformMu transposon-tagging population. *BMC Genomics* **8,** 116 (2007).
16. Williams-Carrier, R. *et al.* Use of Illumina sequencing to identify transposon insertions underlying mutant phenotypes in high-copy Mutator lines of maize. *Plant J.* **63,** 167–177 (2010).
17. Robertson, D. S. Characterization of a mutator system in maize. *Mutat. Res./Fundam. Mol. Mech. Mutag.* **51,** 21–28 (1978/7).
18. Bennetzen, J. L. Transposable element Mu1 is found in multiple copies only in Robertson's Mutator maize lines. *J. Mol. Appl. Genet.* **2,** 519–524 (1984).
19. Taylor, L. P. & Walbot, V. Isolation and characterization of a 1.7-kb transposable element

from a mutator line of maize. *Genetics* **117,** 297–307 (1987).

20. Oishi, K. K. & Freeling, M. A New Mu Element from a Robertson's Mutator Line. in *Plant Transposable Elements* 289–291 (Springer, Boston, MA, 1988).
21. Talbert, L. E., Patterson, G. I. & Chandler, V. L. Mu transposable elements are structurally diverse and distributed throughout the genusZea. *J. Mol. Evol.* **29,** 28–39 (1989).
22. Fleenor, D., Spell, M., Robertson, D. & Wessler, S. Nucleotide sequence of the maize Mutator element, Mu8. *Nucleic Acids Res.* **18,** 6725 (1990).
23. Chomet, P., Lisch, D., Hardeman, K. J., Chandler, V. L. & Freeling, M. Identification of a regulatory transposon that controls the Mutator transposable element system in maize. *Genetics* **129,** 261–270 (1991).
24. Schnable, P. S., Peterson, P. A. & Saedler, H. The bz-rcy allele of the Cy transposable element system of Zea mays contains a Mu-like element insertion. *Mol. Gen. Genet.* **217,** 459–463 (1989).
25. Hershberger, R. J., Warren, C. A. & Walbot, V. Mutator activity in maize correlates with the presence and expression of the Mu transposable element Mu9. *Proc. Natl. Acad. Sci. U. S. A.* **88,** 10198–10202 (1991).
26. Dietrich, C. R. *et al.* Maize Mu transposons are targeted to the 5' untranslated region of the gl8 gene and sequences flanking Mu target-site duplications exhibit nonrandom nucleotide composition throughout the genome. *Genetics* **160,** 697–716 (2002).
27. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).
28. Benito, M. I. & Walbot, V. Characterization of the maize Mutator transposable element MURA transposase as a DNA-binding protein. *Mol. Cell. Biol.* **17,** 5165–5175 (1997).

Supplementary tables:

Supplementary Table 1. Gap number and sizes in maize genomes

	# gaps (>10Ns)	Mean gap size	Total gap length
B73	2520	12196	30732868
W22	68123	596	40626859
PH207	362647	1219	442114873

Supplementary Table 2. Gene content variation in B73 and W22 relative to sorghum.

B73 as query	Number of genes	% of genes
Number of annotated nuclear genes in study	38254	
Present in W22 and Sorghum	23072	60.3
Present in W22 but not in Sorghum	7861	20.5
Present in Sorghum but not W22	881	2.3
Not in W22 or Sorghum	6440	16.8
W22 as query		
Number of annotated nuclear genes in study	40667	
Present in B73 and Sorghum	24784	60.9
Present in B73 but not in Sorghum	6099	15.0
Present in Sorghum but not B73	1412	3.5
Not in B73 or Sorghum	8372	20.6

Supplementary Table 3. The number of locally duplicated genes by the number of intervening genes. The number of genes determined to be local duplicates when the number of intervening genes varies from zero (tandem duplicates) to a maximum of 20.

# of Intervening Genes	# of Locally Duplicated Genes in	
	W22	B73
0	3,405	3,690
1	1,863	1,821
2	1,360	1,178
3	1,012	766
4	743	574
5	552	491
6-10	1,017	765
11-20	880	546
Total	6,034	5,768

Supplementary Table 4. Number of local multiplication clusters in B73 and W22 genomes. The distribution of local multiplication clusters was classified by the number of gene copies in each cluster. The distribution was determined allowing for a maximum of 20 intervening genes.

	# of Clusters	
# of Copies	W22	B73
2	1,706	1,683
3	344	342
4	130	119
5	62	61
6-10	74	60
>10	13	12
Total	2,329	2,277

Supplementary Table 5. Number of shared and unique tandem duplications between W22 and B73. Among the tandem duplications (two copies, zero intervening genes), the number that are shared between the two inbreds and unique to each inbred. The column descriptions are Head to head (H-H), Head to tail (H-T), Tail to Tail (T-T) and the total duplications for each row.

	# of Tandem Duplications			
	H-H	H-T	T-T	Total
Unique to W22	215	862	184	1,261
Shared (W22)	90	510	78	678
Shared (B73)	85	511	82	
Unique to B73	242	925	227	1,394

Supplementary Table 6: Genome-wide alternative splicing in W22v2 (AltA: alternative acceptor; AltD: alternative donor; AltTE: alternate exon; ExonS: exon skip; IntronR: intron retention).

AStype	Number of AS Genes	Number of AS Isoforms	Number of AS Events
AltA	6,499	29,312	12,377
AltD	4,893	21,869	8,535
AltTE	3,082	12,773	5,033
ExonS	3,420	8,114	4,713
IntronR	10,078	27,258	29,939
Total events	13,591	58,279	60,597

Supplementary Table 7. Numbers of Mu-element TIRs in Group 1 through Group 7 identified in W22 and B73 (TIR, Terminal Inverted Repeat).

Mu-TIR Group	B73	W22
Group1	96	99
Group 2	89	81
Group 3	90	94
Group 4	95	106
Group 5	63	48
Group 6	38	34
Group 7	10	12
Total	481	474

Supplementary Table 8. Mu element TIRs in W22 and B73 (v3) by chromosome (TIR, Terminal Inverted Repeat).

Chromosome	Intact Element		Orphan TIRs		Tandem TIRs		Total TIRs	
	W22	B73	W22	B73	W22	B73	W22	B73
Chr 1	28	27	6	6	2	3	64	65
Chr 2	26	29	18	6	2	3	72	67
Chr 3	24	20	2	3	1	1	51	44
Chr 4	26	23	2	4	0	1	54	51
Chr 5	26	38	4	4	0	0	56	80
Chr 6	9	14	5	8	10*	1	33	37
Chr 7	8	11	3	2	0	0	19	24
Chr 8	22	22	9	6	0	0	53	50
Chr 9	14	9	5	6	0	0	33	24
Chr 10	17	13	3	5	2	0	39	31
Chr Unk	0	4	0	0	0	0	0	8
Total	200	210	56	50	17	9	474	481

* One region on chromosome 6 of W22 contains an array of 10 tandem duplications of a single TIR (scored as one orphan and 9 tandem TIRs).

Supplementary Table 9. TIR families with greater than 10 copies in W22

Family name	Superfamily	# copies in W22 genome	Order in Figure 4A
DTH13942	DTH	21	1
DTH10730	DTH	124	2
DTH11101	DTH	169	3
DTH11209	DTH	74	4
DTH15158	DTH	189	5
DTH12258	DTH	40	6
DTT10101	DTT	94	7
DTH11270	DTH	254	8
DTH11374	DTH	1514	9
DTH11602	DTH	84	10
DTH10268	DTH	122	11
DTH10107	DTH	96	12
DTH12507	DTH	35	13
DTT10927	DTT	34	14
DTH12298	DTH	67	15
DTH12718	DTH	23	16
DTT14784	DTT	20	17
DTH12996	DTH	234	18
DTH11238	DTH	38	19
DTH16100	DTH	26	20
DTH12997	DTH	107	21
DTH16329	DTH	154	22
DTH11541	DTH	139	23
DTH10775	DTH	176	24
DTH12973	DTH	24	25
DTH10445	DTH	149	26
DTH16443	DTH	34	27
DTH13117	DTH	22	28
DTH12864	DTH	31	29
DTH16563	DTH	30	30
DTT15264	DTT	25	31
DTH13439	DTH	78	32
DTH10818	DTH	44	33
DTH10194	DTH	151	34
DTH10187	DTH	58	35
DTH10176	DTH	22	36
DTH13854	DTH	58	37
DTH16233	DTH	56	38
DTH16174	DTH	25	39
DTH13110	DTH	240	40

DTH15132	DTH	39	41
DTH13583	DTH	66	42
DTH10856	DTH	592	43
DTH10855	DTH	134	44
DTA00256	DTA	24	45
DTC00118	DTC	27	46
DTH13261	DTH	23	47
DTH10573	DTH	125	48
DTH10113	DTH	21	49
DTA00306	DTA	24	50
DTH10239	DTH	23	51
DTH11614	DTH	21	52
DTA00295	DTA	36	53
DTH10240	DTH	59	54
DTH14736	DTH	44	55
DTH10672	DTH	100	56
DTH11388	DTH	36	57
DTA00180	DTA	20	58
DTT10062	DTT	45	59
DTA00145	DTA	63	60
DTA00114	DTA	22	61
DTH11594	DTH	76	62
DTT10089	DTT	52	63
DTA00229	DTA	34	64
DTA00234	DTA	255	65
DTH10637	DTH	82	66
DTH14738	DTH	36	67
DTA00291	DTA	21	68
DTH11674	DTH	21	69
DTA00359	DTA	35	70
DTA00100	DTA	23	71
DTA00149	DTA	32	72
DTC00030	DTC	58	73
DTA00199	DTA	23	74
DTH12306	DTH	39	75
DTH10047	DTH	24	76
DTH00410	DTH	45	77
DTH00378	DTH	21	78
DTT11073	DTT	41	79
DTH00429	DTH	80	80
DTH00058	DTH	136	81
DTM00796	DTM	77	82
DTM00473	DTM	188	83
DTH00129	DTH	23	84

DTA00294	DTA	24	85
DTH00194	DTH	76	86
DTH00437	DTH	197	87
DTM01654	DTM	20	88
DTH12389	DTH	22	89
DTH00233	DTH	28	90
DTA00200	DTA	34	91
DTM00555	DTM	29	92
DTT11230	DTT	22	93
DTT11335	DTT	27	94
DTH00127	DTH	22	95
DTH00160	DTH	48	96
DTM00299	DTM	72	97
DTT11056	DTT	182	98
DTH00163	DTH	48	99
DTH11715	DTH	32	100
DTM00257	DTM	272	101
DTH00118	DTH	97	102
DTA00111	DTA	56	103
DTA00267	DTA	49	104
DTH00051	DTH	76	105
DTA00126	DTA	29	106
DTH00434	DTH	24	107
DTM00266	DTM	33	108
DTH12490	DTH	34	109
DTH00276	DTH	38	110
DTH00458	DTH	114	111
DTH00090	DTH	20	112
DTH00460	DTH	26	113
DTH00249	DTH	24	114
DTH15359	DTH	57	115
DTH00102	DTH	129	116
DTH10388	DTH	37	117
DTH00409	DTH	94	118
DTH00489	DTH	21	119
DTA00323	DTA	20	120
DTA00242	DTA	20	121
DTA00364	DTA	24	122
DTH00412	DTH	26	123
DTT10009	DTT	36	124
DTA00383	DTA	30	125
DTA00166	DTA	24	126
DTA00140	DTA	34	127
DTA00139	DTA	80	128

DTA00322	DTA	84	129
DTA00231	DTA	78	130
DTA00110	DTA	22	131
DTA00373	DTA	75	132
DTA12512	DTA	35	133
DTA00040	DTA	379	134
DTC12155	DTC	34	135
DTH13200	DTH	30	136
DTA00346	DTA	30	137
DTA00240	DTA	22	138
DTT11465	DTT	54	139
DTH11592	DTH	31	140
DTA00177	DTA	64	141
DTA00098	DTA	160	142
DTA00334	DTA	83	143
DTH11615	DTH	76	144
DTA00263	DTA	39	145
DTA00179	DTA	42	146
DTA00151	DTA	31	147
DTA00217	DTA	110	148
DTA00133	DTA	26	149
DTA00327	DTA	28	150
DTA00268	DTA	35	151
DTA00104	DTA	32	152
DTA00307	DTA	56	153
DTM00800	DTM	37	154
DTA00204	DTA	42	155
DTA00313	DTA	20	156
DTA00155	DTA	42	157
DTA00333	DTA	20	158
DTA00178	DTA	28	159
DTA00300	DTA	53	160
DTA00261	DTA	49	161
DTA00208	DTA	54	162
DTH10658	DTH	26	163
DTA00368	DTA	47	164
DTH12502	DTH	31	165
DTH10440	DTH	21	166
DTA00073	DTA	28	167
DTH10328	DTH	35	168
DTH16801	DTH	33	169
DTH00175	DTH	34	170
DTH11714	DTH	34	171
DTH12617	DTH	79	172

DTA00252	DTA	40	173
DTA00188	DTA	23	174
DTM00743	DTM	77	175
DTH00327	DTH	26	176
DTA00163	DTA	72	177
DTA00117	DTA	89	178
DTM00460	DTM	30	179
DTH10310	DTH	59	180
DTA00106	DTA	44	181
DTA00165	DTA	21	182
DTA13185	DTA	71	183
DTA00156	DTA	45	184
DTA00153	DTA	33	185
DTH12584	DTH	38	186
DTM00268	DTM	83	187
DTC00122	DTC	73	188
DTA00283	DTA	234	189
DTA00169	DTA	191	190
DTC00119	DTC	99	191