Lin et al. Genome Biology (2021) 22:175
https://doi.org/10.1186/s13059-021-02396-x

Genome Biology

RESEARCH Open Access

Chromosome-level genome assembly of a regenerable maize inbred line A188

Guifang Lin[1†], Cheng He[1†], Jun Zheng[2†], Dal-Hoe Koo[1], Ha Le[1], Huakun Zheng[1], Tej Man Tamang[3], Jinguang Lin[1,4], Yan Liu[2], Mingxia Zhao[1], Yangfan Hao[1], Frank McFraland[5], Bo Wang[6], Yang Qin[2], Haibao Tang[7], Donald R. McCarty[8], Hairong Wei[9], Myeong-Je Cho[10], Sunghun Park[3], Heidi Kaeppler[5], Shawn M. Kaeppler[5], Yunjun Liu[2], Nathan Springer[11], Patrick S. Schnable[12], Guoying Wang[2], Frank F. White[13] and Sanzhen Liu[1*]

* Correspondence: liu3zhen@ksu.edu
†Guifang Lin, Cheng He and Jun Zheng contributed equally to this work.
[1]Department of Plant Pathology, Kansas State University, 4024 Throckmorton Center, Manhattan, KS 66506-5502, USA
Full list of author information is available at the end of the article

Abstract

Background: The maize inbred line A188 is an attractive model for elucidation of gene function and improvement due to its high embryogenic capacity and many contrasting traits to the first maize reference genome, B73, and other elite lines. The lack of a genome assembly of A188 limits its use as a model for functional studies.

Results: Here, we present a chromosome-level genome assembly of A188 using long reads and optical maps. Comparison of A188 with B73 using both whole-genome alignments and read depths from sequencing reads identify approximately 1.1 Gb of syntenic sequences as well as extensive structural variation, including a 1.8-Mb duplication containing the Gametophyte factor1 locus for unilateral cross-incompatibility, and six inversions of 0.7 Mb or greater. Increased copy number of carotenoid cleavage dioxygenase 1 (*ccd1*) in A188 is associated with elevated expression during seed development. High *ccd1* expression in seeds together with low expression of yellow endosperm 1 (*y1*) reduces carotenoid accumulation, accounting for the white seed phenotype of A188. Furthermore, transcriptome and epigenome analyses reveal enhanced expression of defense pathways and altered DNA methylation patterns of the embryonic callus.

Conclusions: The A188 genome assembly provides a high-resolution sequence for a complex genome species and a foundational resource for analyses of genome variation and gene function in maize. The genome, in comparison to B73, contains extensive intra-species structural variations and other genetic differences. Expression and network analyses identify discrete profiles for embryonic callus and other tissues.

Keywords: Maize, Genome assembly, Long reads, Structural variation, Kernel color

Background

The maize inbred line A188 was derived from a line related to the commercial maize variety Silver King and a northwestern dent line [1]. A188 has a mixed origin and belongs to neither of the two major heterotic breeding groups [2, 3]. A188 is amenable to somatic embryogenic culture and regeneration and was the first maize line used to

produce genetically modified plants [4]. A popular maize transformation line, Hi-II, was isolated from offspring of a cross between A188 and B73, an elite maize reference inbred line [5, 6]. Although highly valuable for plant regeneration and transformation, A188 is not agronomically desirable, having small ears and low grain yield. The line also exhibits a high degree response to environmental conditions, including sensitivity to abiotic and biotic stresses, such as drought, heat, and bacterial and fungal diseases, in comparison to elite maize lines [7]. A188, therefore, in addition to traits related to transformability, can serve as a model inbred line for the genetic dissection of many important agronomic traits, heterosis, and plant-environment interactions.

Efforts have been pursued to develop efficiency and quality strategies for maize genome sequencing and assemblies. The first maize reference genome for B73 was sequenced and assembled using bacterial artificial chromosomes (BACs) [8]. Since then, additional assemblies have been produced using so-called next-generation high-throughput sequencing, including both short- and long-read technologies [9–15]. Recently, long-read technologies were combined with optical DNA mapping to produce high-continuity maize assemblies, including Nested Association Mapping (NAM) founder lines [16–18]. Here, we used Nanopore long reads and optical DNA mapping to construct a chromosome-level maize genome of A188 for the discovery of structural variation as well as performed transcriptome and DNA methylome analyses of embryogenic callus.

Results

Phenotype of A188

Inbred line A188 (PI 693339) has smaller ears and lower grain yield in comparison to the community reference line B73 and is amenable to plant transformation due to abundant production of the callus favorable to regeneration (Fig. 1, Additional file 1: Figure S1, Additional file 2: Table S1). Hybrids of A188 and B73 (PI 550473) exhibit extensive heterosis (Additional file 1: Figure S2).

Chromosome-level A188 assembly

Long reads, representing a 90X coverage, were generated from A188 genomic DNA using the Oxford Nanopore sequencing platform. The N50 of read lengths is 23.9 kb, and the longest read is 270.6 kb (Additional file 1: Figure S3). Genome assembly, performed using Canu, resulted in 1830 contigs, comprising approximately 2.2 Gb of total sequences. The N50 of contigs is 5.99 Mb (Additional file 2: Table S2).

Read depths for contigs were assessed using Illumina short reads generated independently from seedling and immature ear DNAs to identify potential contamination from organelle genomes or extraneous microbial DNA. Contigs from organelle genomes or extraneous microbial DNA were expected to have differential read depths between the two tissues. Based on this strategy, contigs identified as the chloroplast or mitochondrion sequences were replaced respectively with the previously complete assemblies of A188 organelle genomes [19, 20] and contigs from extraneous contamination were discarded (Additional file 1: Figure S4). The remaining contigs were polished using raw Nanopore data and 80X PCR-free Illumina 2x250 paired-end whole-genome sequencing reads (Additional file 2: Table S3), followed by the scaffolding with 113 A188

Fig. 1 Ear photos of A188 and B73

Bionano Genome (BNG) optical maps (Additional file 2: Table S4), for which the total length is 2.17 Gb and the N50 is 103.4 Mb. The BNG aided assembly placed 875 contigs into 39 scaffolds, which consist of 2.15 Gb. Chromosome pseudomolecules were then generated using a genetic map constructed from 100 B73xA188 double haploid (DH) lines (Additional file 2: Table S5). The final assembly (A188Ref1) consists of 2.25 Gb, including 10 chromosomal pseudomolecules, a mitochondrial genome, a chloroplast genome, and 986 scaffolds or contigs (Table 1).

The base accuracy of the A188Ref1 assembly was estimated at approximately 99.82% using the KAD pipeline [21]. Approximately 96.4% of the potential errors are in transposons or other repetitive sequences. The estimated accuracy of genic sequences was >99.97%. The completeness of the A188 assembly was assessed using the BUSCO software [22] and found to contain 97.25% (3189/3278) of the Liliopsida core gene set, similar to the 97.36% (3193/3278) in the B73 reference genome (B73Ref4) [9].

Presence of complex repeats and nuclear organelle sequences in A188Ref1

In total, 86.3% of the A188 genome sequence is annotated as repetitive elements. The long terminal repeat (LTR) retrotransposons *Gypsy* and *Copia* were the most prevalent elements, consisting of 44% and 23.9% of A188Ref1, respectively (Fig. 2, circos plot [23]). LTR centromere retrotransposon of maize (CRM) were largely co-localized with centromere-specific satellite repeat CentC, both of which were largely syntenic to the B73 centromeres [9]. Approximately 8.3% of A188Ref1 is annotated as DNA transposable elements (TEs), including helitron and Miniature Inverted-repeat Transposable

Table 1 Summary of A188Ref1 assembly and annotation

Chromosome	Length (bp)	# genes	# transcripts
1	307,989,483	6034	9265
2	251,027,758	4873	7384
3	243,219,806	4313	6619
4	255,421,021	4315	6640
5	229,324,730	4613	7194
6	181,596,323	3412	5134
7	183,343,242	3208	4864
8	182,018,909	3653	5472
9	165,494,689	3082	4704
10	153,829,095	2824	4254
mt	525,405	40	40
pt	140,437	39	41
Scaffolds (N = 986)	9,2920,514	341	531
Sum	**2,246,851,412**	**40,747***	**62,142**

*Filtered from 46,009 gene models produced by Maker

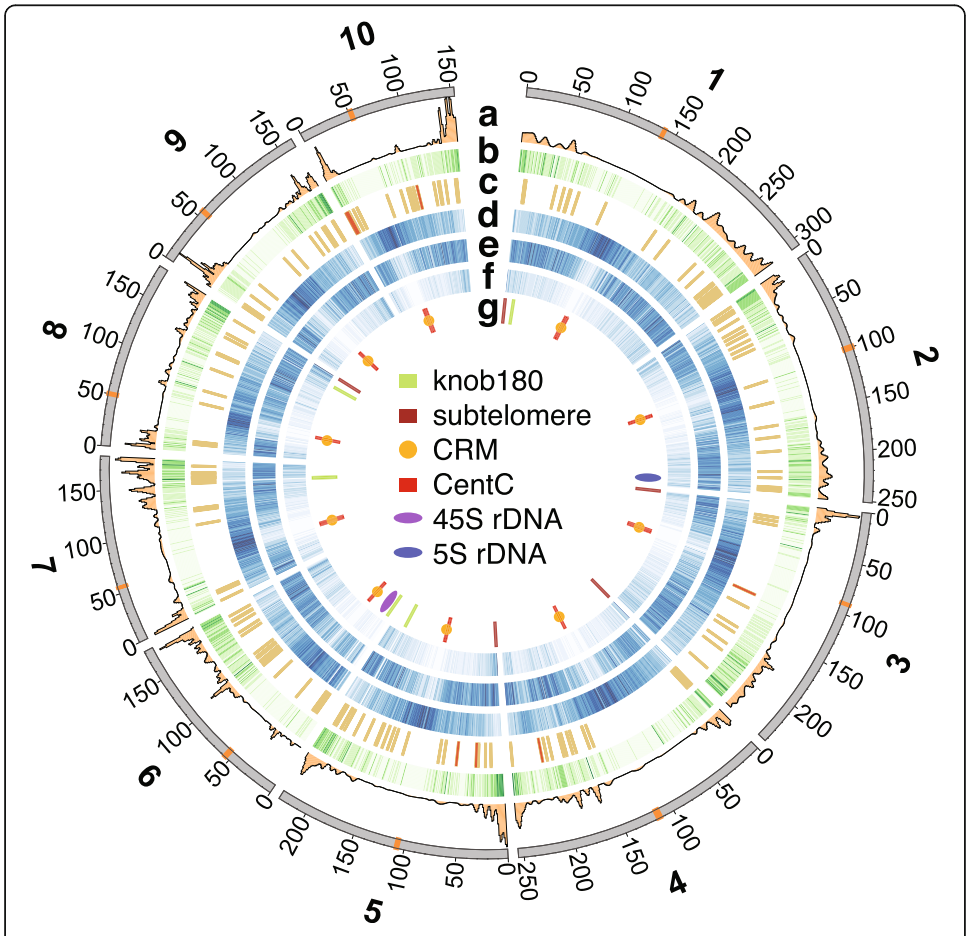

Fig. 2 Circos plot of genomic features. Features on chromosomes are (**a**) recombination rate (cM/Mb); (**b**) gene density per Mb; (**c**) gene clusters; (**d**) number of *Gypsy* per Mb; (**e**) number of *Copia* per Mb; (**f**) number of MITEs per Mb; (**g**) high-copy repetitive elements. The central inset is the legend for the track of **g**. Tracks of **b**, **d**, **e**, and **f** are intensity-coded. The higher the intensity, the higher the frequency of each element. Centromeres are in orange on the outmost chromosome track, on which numbers are coordinates in Mb

Elements (MITEs) (Fig. 2). Major knob clusters were found on the long arm of chromosomes 5 (5L), the short arm of chromosome 6 (6S), 7L, and 8L, and major subtelomeric repeats (4-12-1) were clustered on the distal regions of 1S, 3S, 4S, 5S, and 8L (Fig. 2). Through similarity alignments, the 45S and 5S ribosomal DNA (rDNA) clusters were localized on 6S and 2L, respectively (Fig. 2). Knob and rDNA locations were in agreement with previously reported A188 fluorescent in situ hybridization (FISH) data [24]. Most repetitive components were located in regions of low-recombination contexts except the 5S rDNA locus and subtelomeric clusters (Additional file 1: Figure S5).

Nuclear mitochondrial DNA (NUMT) and nuclear plastid (chloroplast) DNA (NUPT) were identified at 10 and 21 genomic loci, respectively (Fig. 3a, Additional file 1: Figure S6). The largest nuclear organelle-like sequence (~136 kb) is a NUMT locus on the short arm of chromosome 8, which contains an array of DNA transposons likely inserted subsequent to the NUMT integration. FISH analysis corroborated the chromosome 8 location and confirmed a homolog on the distal end of 10S (Fig. 3b, c). In summary, the genomic locations of repetitive sequences and nuclear organelle sequences

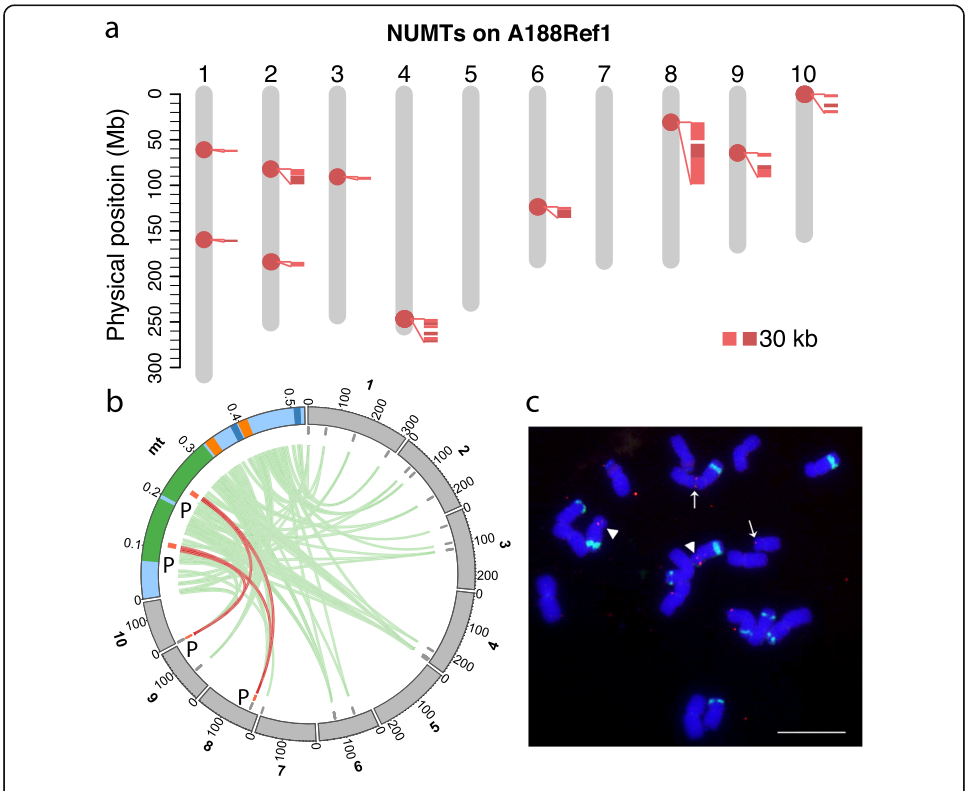

Fig. 3 NUMT on A188 nuclear genomes. **a** NUMT sequence on 10 chromosomes of A188Ref1. Each dot on chromosomes designates a potential NUMT integration. Close-up alignments with the mitochondrion (mt) genome are shown along NUMTs. Each alignment requires at least 5 kb match and 95% identity. **b** Circos plot of alignments between the mt genome and 10 chromosomes. The same colors of green, orange, dark blue label duplicated regions in mt. "P" regions match the probe sequence used for FISH. Brown links highlight alignments on chromosomes 8 and 10. Note that the chromosomal scale is different from the mt scale. Numbers on the track are in Mb. **c** Physical mapping of a mt DNA (mtDNA) and knob repeats on the mitotic metaphase chromosomes of maize A188. The knob repeat probe (green signals) was used to identify the chromosomes. Two FISH sites of the mtDNA insertion on the chromosomes were detected: arrowheads, chromosome 8; arrows, chromosome 10. Bar = 10 μm

are largely consistent with previous findings by FISH [25, 26], supporting the large-scale correctness of A188Ref1.

Gene annotation

Annotation of A188Ref1 was performed using the Maker pipeline with evidence from transcripts assembled with A188 long Nanopore direct cDNA sequencing data, A188 RNA-Seq Illumina short reads, and transcripts from other maize lines, as well as protein sequences from closely related plant species. The Maker genome annotation resulted in 40,747 high-confidence gene models with 62,142 transcripts (A188Ref1a1) (Table 1). BUSCO evaluation showed that 97.8% of Liliopsida conserved genes were in A188Ref1a1. Comparison of protein sequences identified 52,971 orthologous pairs between A188 and B73, consisting of 27,273 A188 genes and 27,529 B73 genes. We also identified 178 gene clusters in A188 each of which contains at least three paralogous genes, comprising in total 694 genes. The clusters of genes encoding pectin methylesterase (PME) were identified on an unanchored scaffold c04_002 (two clusters with 25 and 18 genes), on chromosome 4 (one cluster with nine genes), and on chromosome 5 (one cluster with five genes) (Fig. 4a). Gene clusters also include eight clusters of 42 nucleotide-binding leucine-rich repeat (NLR) disease resistance (R) genes (Fig. 4a). One NLR gene cluster on chromosome 10 has 16 genes homologous to the *rp1* gene that confers resistance to common rust [27] and was associated with Goss's wilt resistance [28] (Fig. 4b). Most paralogous clusters were not located in regions with high recombination (Fig. 4c). Exceptions include the *rp1* locus, which has a high level of haplotype instability through frequent recombination among *rp1* paralogs [29–31]. Divergent *rp1* haplotypes were observed between A188 and B73 that contains 11 *rp1* homologs at the syntenic locus (Fig. 4b).

We identified 2259 paralogous gene pairs of which one member was located in a high-recombination chromosomal compartment and the other in a low-recombination compartment ("Methods" section). Comparison of DNA methylation of A188 seedlings found that, on average, both CG and CHG methylation, where H represents A, C, or T, were higher near and within low-recombination paralogous genes as compared to high-recombination genes. No obvious differences were observed in CHH methylation (Fig. 4d–f). Comparison of gene expression between members of the paralogous pairs using seedling RNA-Seq data showed most paralogs had similar expression levels and no expression bias to either high- or low-recombination genes was observed for those paralogs that did exhibit differential expression (Additional file 1: Figure S7). The result indicated that the genomic context of genes is a driver for a certain epigenomic modification but not a major driver for gene expression.

High-level structural variation between A188 and B73

Structural variation between the A188 and B73 genomes was identified through comparisons of whole-genome assemblies of both genomes using SyRI software (Additional file 3) [32] and through the analysis of whole-genome Illumina sequencing reads with Comparative Genomic Read Depth (CGRD) that is based on quantitative comparison of depths of short reads (Additional files 4 and 5) [33]. SyRI revealed ~1.1 Gb of syntenic regions, 2302 translocations, and 4083 duplications in B73 and 2333 duplications

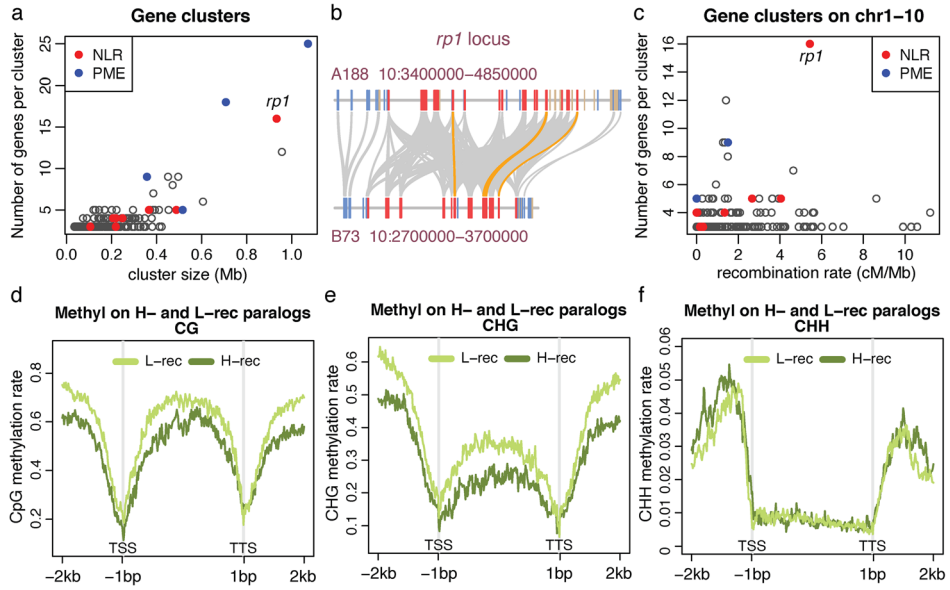

Fig. 4 Gene clusters and paralogs in low- and high-recombination regions. **a** The scatter plot of numbers of genes per cluster versus their cluster size. **b** Example of an NLR gene (*rp1*) cluster in A188 and their alignments with the B73 *rp1* locus. Each rectangle box represents a gene with blue, tan, and red colors indicating plus, minus orientation, and *rp1* homologous genes, respectively. All *rp1* homologs are in the same minus orientation. Gray bands connect orthologs and orange bands highlight the top *rp1* alignments with at least 98.5% identity and a 2500-bp match. **c** The scatter plot of numbers of genes per cluster versus the recombination rate estimated 1 Mb around the midpoint of the cluster. All clusters plotted are on 10 chromosomes. **d**–**f** Distribution of cytosine methylation in sequence contexts of CG, CHG, and CHH around paralogous genes. An average methylation rate per window across all examined genes from two replicates of seedling samples was determined and plotted versus the window order. A window in the gene body, from translation start site (TSS) to translation termination site (TSS), is 1/200 of the gene body in length. A window outside of the gene body is 20 bp

in A188 using a minimum cutoff of 10 kb for each translocation or duplication event (Additional file 2: Table S6). In addition, SyRI identified 441.9 Mb of B73 and 543.8 Mb of A188 DNA sequences that were not aligned with the other respective genome. Further filtering with CGRD that compared read depths between the two genomes revealed 381.3 Mb of B73-specific sequences and 409.2 Mb of A188-specific sequences that represent presence and absence variance (PAV) or highly divergent sequences (HDS). These PAV/HDS regions contain 6728 genes in B73 and 7301 genes in A188 (Additional files 6 and 7). Gene ontology enrichment analysis indicated that genes related to endopeptidase inhibitor activity and extracellular activities are enriched in both PAV/HDS gene sets (Additional file 1: Figure S8).

Seventeen large inversions of 0.5 Mb or greater were identified between the two genomes (Fig. 5, Additional file 1: Figures S9-S17, Additional file 2: Table S7). Nine of the seventeen inversions are likely errors in B73Ref4 as the newly released B73Ref5 showed the same orientation as A188Ref1, including the largest inversion region (INV37083 on B73Ref4, 97.8–103.9 Mb on chromosome 4). FISH analysis of A188 and B73 corroborated the absence of inversion INV37083 (Additional file 1: Figure S18). Recombination and pairwise linkage disequilibrium (LD, R^2) values among single nucleotide polymorphisms (SNPs) within each inversion were determined, and out of eight remaining inversion candidates, six have recombination frequencies close to 0 and a high mean LD ranging from 0.56 to 0.79 of all pairs of SNPs that are separated by 0.2–0.3 Mb within

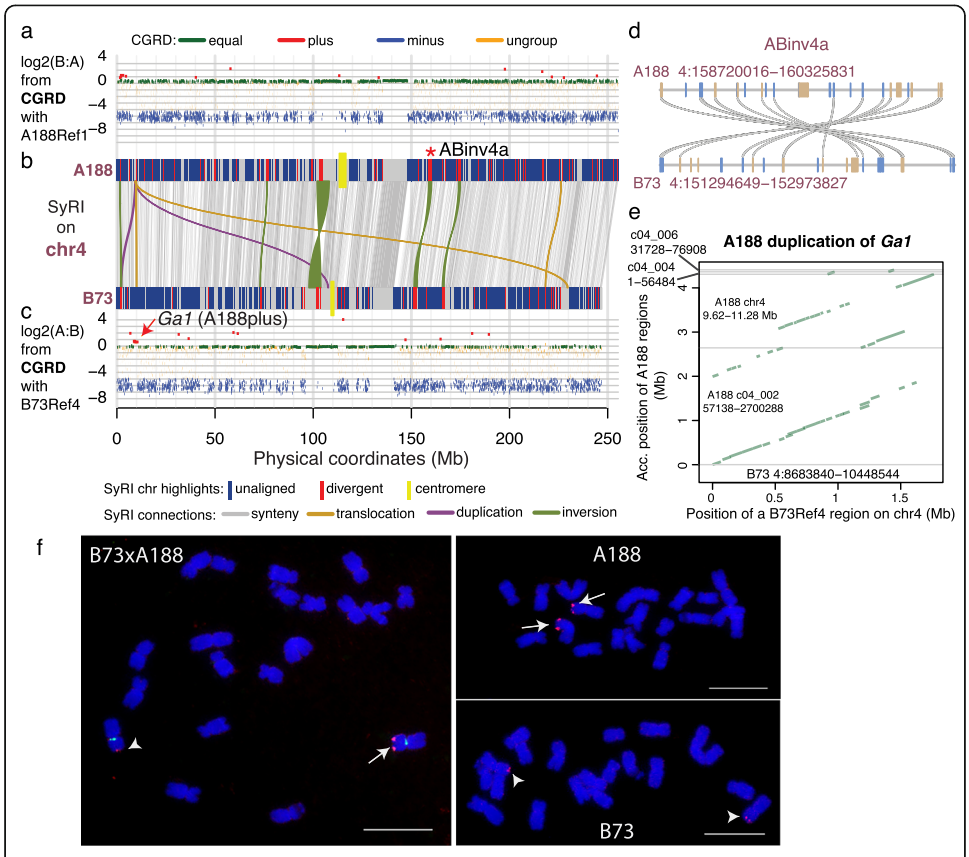

Fig. 5 Megabase-level duplication and inversion on chromosome 4. **a**–**c** SyRI and CGRD results on chromosome 4. **a** The CGRD result using A188Ref1 as the reference genome. The Y-axis represents log2 values of ratios of read depths of B73 to A188, log2(B:A), signifying copy number variation (CNV). Regions with higher and lower sequence depths of B73 versus A188 were B73 plus (red) and B73 minus (blue), respectively. Green and orange represent conserved and ungrouped regions, respectively. **b** The SyRI result is displayed. Alignments of syntenic blocks larger than 10 kb and alignments of other rearrangements larger than 0.5 Mb are plotted. On each A188 and B73 chromosome, segments that were not aligned to the other genome or highly divergent with the other genome are highlighted. The red * labels a well-evidenced inversion. **c** The CGRD result using B73Ref4 as the reference genome. The similar color scheme to that in **a** is used. **d** Synteny of genes (rectangle blocks) in the well-evidenced inversion (ABinv4a) regions between A188 and B73. Blue and tan colors stand for plus and minus gene orientations. **e** A dot plot between the 1.8-Mb B73 region that was duplicated in A188 and its aligned regions in A188Ref1. **f** FISH of the PME probe on A188, B73, and F_1 (B73xA188). Cent4 probe (green) that is specifically from chromosome 4 centromere was used in F_1 FISH. Arrows and arrowheads point at PME signals of A188 and B73 chromosomes, respectively. Bar = 10 μm

an inversion, which are much higher than the genome-wide average LD of 0.2 between SNPs in separated by 0.2 Mb (Additional file 2: Table S7). These six inversions exhibiting marked recombination suppression characteristic of inversion [34], therefore, are strongly supported. The six inversions range from 0.7 to 2.1 Mb in size, of which two are located close to the centromere of chromosome 2 and four are on 3L, 4L, 5L, and 9L (Additional file 2: Table S7). Each of these six inversions can be identified between A188Ref1 and many other maize genomes, including the genomes of NAM founder lines (Additional file 1: Figure S19-S23, Additional file 2: Table S8). In total, the six inversion sequences harbor 69 genes in B73 and 75 genes in A188. The syntenic relationships of these genes were largely maintained between inverted sequences in the two genomes (example in Fig. 5d), although the gene sequences are divergent in a high

degree from each other (Fig. 5b, Additional file 1: Figures S10, S11, S12, S16). The divergence of these inversions indicated that the inversions were not recent events established in modern maize populations. Admixture structure analysis showed that both A188 and B73 haplotypes of 3/6 inversions (ABinv2a, ABinv2b, and ABinv3a) exist in teosinte, the maize wild ancestor (Additional file 1: Figure S24), and there is no clear evidence of the haplotype of the remaining three inversions (ABinv4a, ABinv5a, and ABinv9a) existed in teosinte (Additional file 1: Figure S25). Among landraces, both the A188 and B73 haplotypes of each of the six inversions could be identified based on SNP genotyping data, supporting that all these six inversions exist in landrace maize lines (Additional file 1: Figures S24 and S25).

CGRD analysis also identified an A188 duplication of a 1.8-Mb region from 8.68 to 10.45 Mb on chromosome 4 of B73Ref4 (Fig. 5c). In A188, a portion of the duplication was found in the unanchored scaffold c04_002 while most of the remaining duplicated sequences can be found in chromosome 4 (Fig. 5e). The duplication region overlapped with the *Gametophyte factor1* (*Ga1*) locus conferring unilateral cross-incompatibility [35]. The underlying causal gene of B73, Zm00001d048936, encodes a PME, which is a wildtype allele. We designed a PME DNA probe that is from the duplication and repeatedly matches 35 loci in B73Ref4 and 78 loci in both the region on chromosome 4 and the scaffold c04_002 on A188Ref1. FISH using this probe resulted in strong hybridization signals on A188 chromosome 4S and weak signals on B73 chromosome 4S, indicating that the duplication occurred locally on 4S (Fig. 5f). The B73 Zm00001d048936 gene has no additional homologous copies in B73Ref4 but five homologous sequences can be identified on the duplicated sequence of A188Ref1, including the syntenic gene Zm00056a022745 that is identical to Zm00001d048936. Collectively, the result documented the complexity and the potential dynamic of the *Ga1* locus of maize.

Associating structural variation with phenotypic variation

The CGRD result indicated that A188 had many more copies (A188plus) at a region from 155.23 to 155.24 Mb of chromosome 9 in B73Ref4 (Additional file 1: Figure S16). This region includes the *carotenoid cleavage dioxygenase 1* (*ccd1*) gene catalyzing the cleavage of carotenoids to apocarotenoid products, which is located at the White cap locus (*Wc1*) conditioning kernel colors [36]. SyRI analysis supported a duplication of this region but failed to find a number of copies in A188. SyRI analysis also indicated the duplicated region is embedded in A188-specific sequences (Fig. 6a). Comparison of A188Ref1 with an A188 BNG optical map aligned to the duplication region indicated the incomplete assembly of the region. Previously, tandem repeats of an ~27 kb sequence at the *Wc1* locus were discovered [37]. Each repeat exhibits four discernible sites that can be detected via Bionano analysis, referred to as Type A repeat. Analysis of A188 sequences revealed a repeat variant containing an additional site, referred to as Type B repeat. Based on the BNG map, the A188 genome contains 13 intact tandem copies of the 27-kb sequence, consisting of 9 copies of Type A and 4 copies of Type B repeats, as well as partial copies of the 27-kb sequence on both ends of the array. Each repeat copy contains a *ccd1* gene, indicating at least 13 copies of *ccd1* in A188 (Fig. 6b), consistent with the A188plus result from the CGRD analysis. Neither intact Type A nor B repeat exists in B73, which, however, does contain a *ccd1* gene.

A188 seeds are white, whereas B73 seeds are yellow (Fig. 1). Analysis of quantitative trait locus (QTL) of kernel colors of B73xA188 DH lines resulted in two major QTLs on chromosomes 6 and 9, both of which were discovered in a previous genome-wide association study [38], as well as a weaker peak on chromosome 2 (Fig. 6c, Additional file 1: Figure S26). Two known genes *y1* and *ccd1* in the major peaks are responsible for kernel colors (Fig. 6d) [37, 39]. The dominant *Y1* allele conditions yellow kernels [39]. Several variants exist between the A188 *y1* (Zm00056a032392) and B73 *Y1* (Zm00001d036345) alleles, including one amino acid polymorphism (Ser258Thr) in the

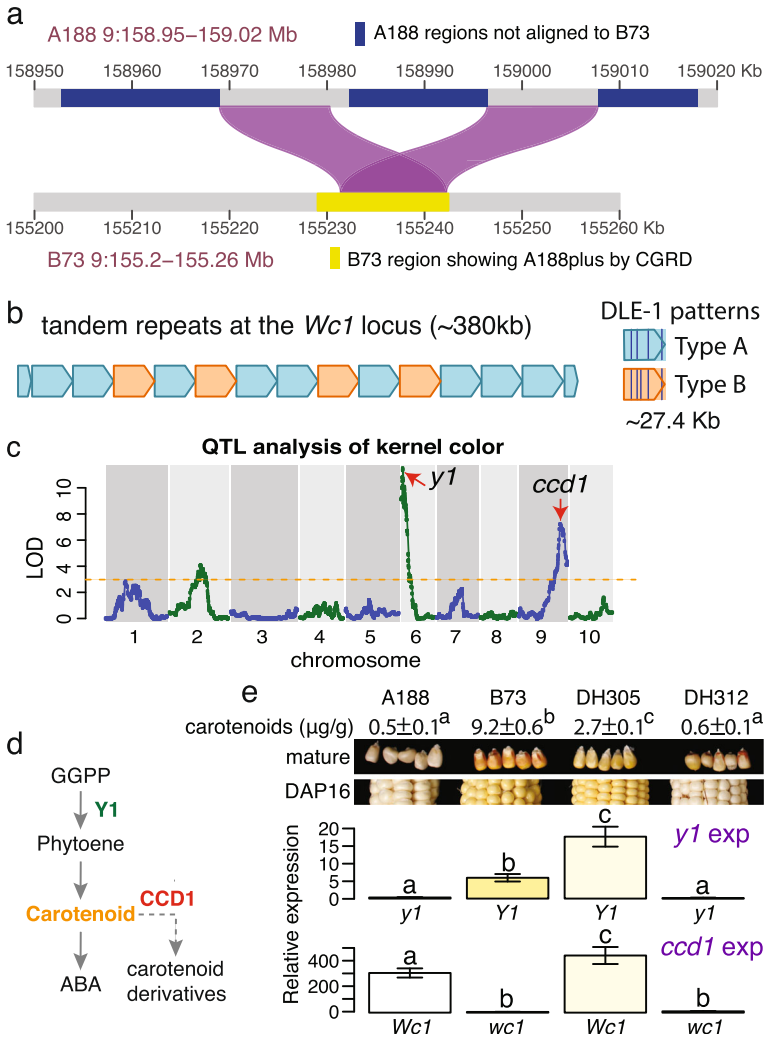

Fig. 6 Structural variation and genetic analysis of the *Wc1* locus. **a** Duplication alignments between A188 and a B73 region identified as A188plus by CGRD. **b** Tandem repeats of 13 intact copies of 27.4-kb sequences. Two DLE-1 restriction patterns in repeat units: Types A and B were identified. **c** The QTL result of kernel color using the DH population. Arrows point at locations of known causal genes. **d** A simplified carotenoid pathway. GGPP stands for geranylgeranyl diphosphate. **e** Seeds at 16 days after pollination (DAP16) were collected and used for quantifying gene expression (exp) of *y1* and *ccd1*. Three biological replicates were used. Bars are color-coded based on the colors of mature seeds. Error bars represent standard deviation (SD). Letters on top of bars are statistical groups determined by Tukey tests. *Y1* (*wc1*) and *y1* (*Wc1*) stand for B73 and A188 alleles, respectively. Mature seeds from the same lines show slightly different colors from seeds of DAP16. Total carotenoids of mature seeds were measured and the values of mean ± SD are listed. Superscript letters are statistical groups determined by Tukey tests

coding region (Additional file 1: Figure S27) and polymorphisms found in 5′ upstream and 3′ downstream regions, including a $(CCA)_n$ microsatellite variation in the 5′ untranslated region [40] (Additional file 1: Figure S28). Quantitative reverse transcription PCR (qRT-PCR) reveals higher expression of the *y1* gene in B73 relative to A188 (Fig. 6e). In contrast, the B73 *ccd1* expression was much lower than that of A188, presumably due to the differences in copy number (Fig. 6e). Because higher expression of functional alleles of the *ccd1* and *y1* genes is expected to reduce and increase the accumulation of carotenoids, respectively, the differences in the expression of the *ccd1* and *y1* genes in B73 and A188 explain yellow kernels of B73 and white kernels of A188 (Fig. 6d, e). Consistently, B73 mature seeds contained much high carotenoid contents as compared to A188 mature seeds (Fig. 6e). The expression levels of the alleles in two DH lines with different allele combinations of these two loci were similar, and the allele combination of *y1* and *ccd1* largely determined the level of carotenoids and seed colors (Fig. 6e).

In addition to kernel color, QTL analysis of cob glume color of which A188 is white and B73 is red mapped a single strong peak on chromosome 1S (LOD = 23.8) (Additional file 1: Figure S29). *Pericarp color 1* (*P1*) encoding a MYB transcription factor located in the QTL peak was known to regulate pigment genes [41]. The CGRD result indicated that B73 had more copies of the *P1* gene than A188, presenting another structural variation event associated with a phenotypic trait (Additional file 1: Figure S29).

Distinct gene expression and hypermethylation in calli relative to seedlings

Transcriptomic data were generated for 11 diverse tissues with three biological replicates each. Both principal component analysis and clustering of these tissue samples based on their genome-wide gene expression showed that the callus from tissue culture were closely related to root, leaf base, embryo, and ear, but distinct from middle leaf, leaf tip, and seedlings (Additional file 1: Figure S25, Additional file 8). A set of 734 callus featured genes were identified that exhibited at least 2-fold up-regulation in the callus as compared to any other tissues (Additional file 2: Table S9). Genes involved in cell wall biosynthesis, defense activity, heme binding, transmembrane transport, and transcription regulation are enriched in these featured genes (Additional file 1: Figure S30). For example, a number of NLR and defense-related genes, including *Pathogenesis-related protein 1* (*PR1*) (Zm00056a001451), were activated in the callus. The top six enriched transcription factor families are WOX, AUX/IAA, LBD, AP2, WRKY, and NAC, which included homologs of *Baby boom* (AP2) and *Wuschel2* (Wox) genes relevant to cell division and expansion (Additional file 1: Figure S31) [42]. Of these callus featured genes, the homologs of *Baby boom* (Zm00056a020360), *Wuschel2* (Zm00056a020673), and LBD (Zm00056a020860) are located in the vicinity of the chromosome 3 locus affecting callus development [43]. Three callus featured genes were located in the previously mapped regions (~3 Mb) [43], including Zm00056a020765 encoding protein upstream of flowering locus C (FLC), Zm00056a020767 encoding a zinc finger protein, and Zm00056a020775 encoding a caffeoyl shikimate esterase (CSE). The B73 CSE syntenic gene, Zm00001d042944, contains a 120-bp MITE insertion at 141 bp upstream of the gene start, while the A188 CSE gene does not (Additional file 1: Figure S32).

The callus and seedling tissues were selected for examination of genome-wide DNA methylation levels. The callus exhibited elevated methylation for all three sequence contexts as compared to the seedling, 89.3% vs 85.2% on CG, 74.5% vs 71.9% on CHG, and 3.2% vs 1.5% on CHH (Additional file 2: Table S10). The analysis of CG and CHG methylation over all genes did not find major differences between callus and seedling tissue (Fig. 7a, b). However, there were major differences in the level of CHH methylation (Fig. 7c). On average, there were no major changes in the level of CG or CHG methylation over repetitive elements but there was a consistent trend for slightly higher CG methylation callus for most classes of repetitive elements ($p < 0.0001$ from paired t-tests, Fig. 7d, e). Similarly, CHH methylation was slightly higher for most classes of repetitive elements with the most notable increase observed at MITE elements ($p < 0.0001$ from paired t-tests, Fig. 7f).

Differentially methylated regions (DMRs) were identified through comparison of the DNA methylation profiles of callus and seedling. In total, 6927 CG DMRs, 9631 CHG DMRs, and 11,275 CHH DMRs were identified (Additional file 2: Table S11, Additional files 9, 10, and 11). Hypermethylation in callus relative to seedling was the predominant type of DMRs for both CG and CHH methylation in both genic and intergenic regions while CHG exhibited roughly equal proportions of hypermethylation and hypomethylation DMRs with more hypermethylation in genic regions and more hypomethylation in intergenic regions (Fig. 7g). The analysis of the distribution of DMRs relative to genes revealed that the CG DMRs were enriched near TSS regions while CHH DMRs tended to be found in regions just upstream or downstream of genes, mirroring CHH island distributions (Fig. 7h) [44, 45]. CHG DMRs exhibited different trends for localization for hypermethylation and hypomethylated DMRs with hypermethylation DNAs enriched at TSS and TTS regions and hypomethylated DMRs enriched in gene bodies (Fig. 7h). The high frequency of some types of DMRs near the TSS led us to assess whether these DMRs may be contributing to differential expression in callus relative to seedling tissue. Genes with DMRs were enriched for being differential expression (DE) in seedling relative to callus compared to genes without DMRs ($\chi^2 = 20.9$, p-value = 4.9e−6). Based on prior studies in maize, we expected that gains of CG or CHG methylation near the TSS would be associated with down-regulation of expression while gains of CHH upstream of the promoter might be associated with up-regulation of expression [46, 47]. We found that the DE genes with hypomethylated or hypermethylated DMRs at most regions exhibited roughly similar numbers of up- and down-regulated with exception at CG hypomethylation at 5′ upstream regions of genes and CHG hypomethylation in the gene body, both of which were associated with up-regulation of gene expression in the callus (Fig. 7i, Additional file 2: Table S12). These results reveal dynamic changes in some types of DNA methylation in callus relative to seedling and a marginal association of DNA methylation with gene expression changes.

Discussion

Here, the A188 genome assembly capitalized on long-read technologies, including Nanopore single molecule reads and long-range optical mapping, which adds a new high-quality reference genome to the collection of sequenced maize genomes [8–16]. The quality of the assembly was enhanced by the strategy of comparing read depths of short read data from two independent DNA sources to filter contigs before the

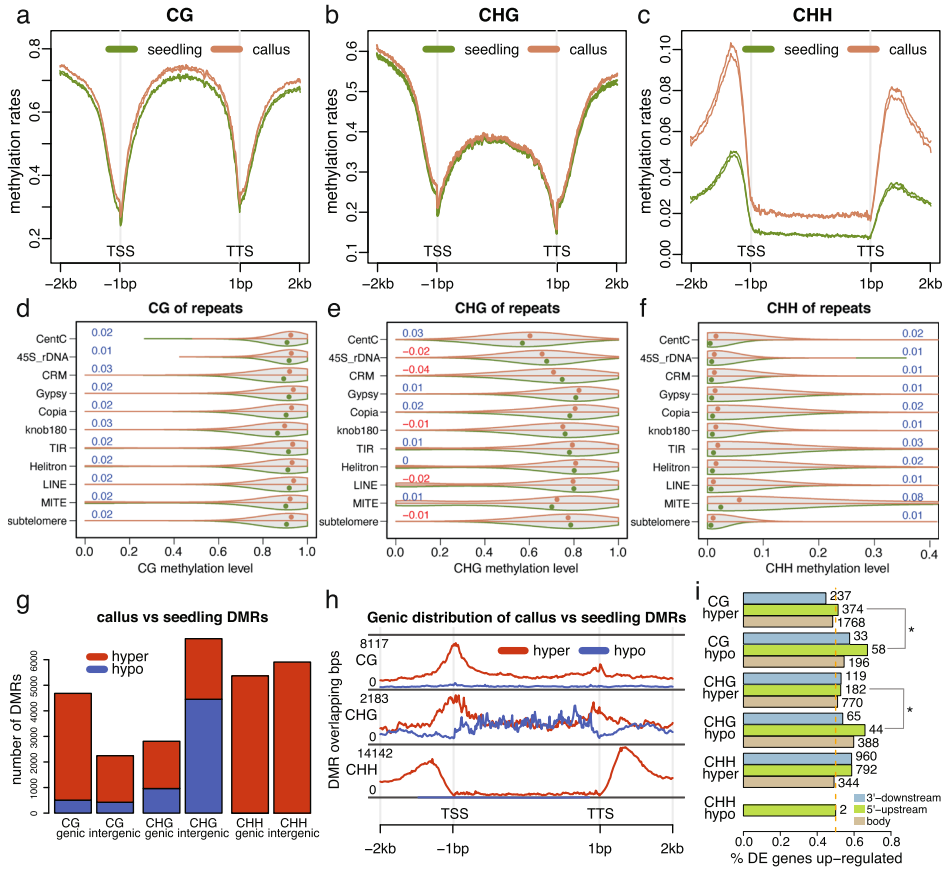

Fig. 7 DNA methylation in callus and seedling tissues. **a**–**c** Distribution of cytosine methylation in three sequence contexts (CG, CHG, and CHH) around genes in two biological replicates of the callus (orange) and two biological replicates of the seedling (green). An average methylation rate per window across all examined genes was determined and plotted versus the window order. A window in the gene body is 1/200 of the gene body. A window outside of the gene body is 20 bp. **d**–**f** Violin plots of methylation on repetitive sequences. For each violin plot, the top half is the distribution of methylation in the callus and the bottom half is the distribution of methylation in the seedling. Each dot represents the median of methylation rates. Numbers stand for the mean methylation differences between the callus and the seedling, which are color-coded with blue and red to represent increased methylation and decreased methylation in the callus, respectively. All differences are significant (p-value < 0.0001) by paired t-test. **g** Barplots of DMRs on genic regions, including 2 kb beyond each of TSS and TTS, and the rest of the genome (intergenic regions). **h** Distribution of DMR sequences around genes. The definition of the gene body is the same as described in **a**–**c**. **i** Proportions of DE genes up-regulated in hyper DMR and hypo DMR regions (gene body, 1 kb 5′ upstream and 1 kb 3′ downstream regions). Numbers on top of bars are numbers of DE genes. Stars indicate significance ($p < 0.05$) from χ^2 tests for the independence of the DMR and DE changing directions. In **g**, **h**, and **i**, hyper and hypo stand for increased and decreased methylation in the callus relative to the seedling, respectively

scaffolding. The use of the independent DNA sources reduced contamination from DNA sequences of organelle genomes and microorganisms while preserving nuclear-integrated organellar sequences. In addition, a novel approach for the discovery of genome structural variation based on quantitative comparison of depths of sequencing reads, here named Comparative Genomic Read Depth or CGRD, was introduced. Detailed characterization of genomic structural variation in complex genomes such as maize is challenging. Comparisons using complete genome sequences based on their alignments would be an ideal method to reveal copy number variation and rearrangements. However, technically, alignment-based methods still suffer from a low ability of

confident alignments of repetitive sequences. More critically, finding structural variation with assembled genome sequences is subject to the quality of assemblies. Unfortunately, assemblies of most plant genomes or other large complex genomes are generally not complete or error-free. B73Ref4, for example, is missing the topmost region of the short arm sequence of chromosome 6 (Additional file 1: Figure S13) and includes multiple assembly inversion errors. CGRD based on the comparison of depths of short reads complements the approaches that rely on whole-genome alignments, including SyRI [32]. In particular, the CGRD pipeline can detect copy number variation missed by SyRI due to incomplete assembly at structurally complex regions. CGRD identified a 1.8-Mb duplication at the *Ga1* locus and a high-copy tandem duplication of *Wc1* in A188, both of which were missed by SyRI. The two methods are complementary in that CGRD captures unbalanced structural variation due to copy number variation rather than balanced structural variation that SyRI can detect. Therefore, the combination of SyRI and CGRD provides an optimal strategy for the discovery of genomic structural variation, which is critical for further characterization of their impacts on gene expression and phenotypes.

Analysis of structural variation elucidated a repetitive structure of the *ccd1* gene, which, in A188, consists of 13 copies. The high copy number of *ccd1* corresponds to the high expression level of *ccd1*, which was previously observed and presumably leads to a high activity of the carotenoid cleavage enzyme and enhanced carotenoid degradation [37]. Furthermore, the expression of *y1*, which encodes for phytoene synthase and the entry reaction to the carotenoid pathway, in A188 is low during seed development, while the *y1* expression in B73 seeds is relatively high [48]. Both alleles were highly expressed in some non-seed tissues, including leaves. The A188 *y1* allele is likely functional since a low but perceptible level of carotenoids is produced at seed development. An additional minor kernel color QTL was identified from the DH lines and concordant with the QTLs from multiple other B73-derived biparental populations [49]. The underlying candidate gene, *zep1* (Zm00001d003513) encoding zeaxanthin epoxidase, was also identified in an earlier association study [50]. However, functional validation for the involvement of *zep1* in seed color is needed. At the same time, the three QTL loci are not sufficient to fully determine the kernel color variation of DH lines. Analysis with a larger B73xA188-derived population may reveal additional loci influencing kernel colors as not all DH lines shared the expected color based on the associated QTLs. In any event, carotenoid levels were expected based on the allele types of *y1* and *ccd1* and supported the hypothesis that higher levels of *ccd1* and low *y1* levels contribute to the differences in seed color of A188 and B73.

An important goal of A188 characterization is to gain insight into plant tissue culture. Development from a highly differentiated tissue to the callus for genetic engineering purposes involves a process of dedifferentiation to gain pluripotency [51]. The transition of differentiation status is, physiologically, stressful [52]. Somaclonal variation, including sterility, in plants produced through tissue culture may be the product of DNA damaging stress responses [53]. Transcriptomic data from this study revealed that, in fact, defense response genes were enriched among the callus featured genes that were up-regulated in the callus as compared to any other tissues. Hypermethylation is considered to be a protection mechanism against

stresses, which enhances genome stability and safeguards genome integrity [54]. Our comparison between the callus and the seedling uncovered globally elevated methylation in the callus in all three sequence contexts. Consistently, hypermethylation in the callus relative to the immature embryo was found in a study that used another maize inbred line and methylated DNA immunoprecipitation sequencing (MeDIP-seq) [55]. In this study, 24 nt small RNA was shown to be positively correlated with DNA methylation. In rice, CG hypermethylation was seen in 1- and 3-year callus relative to the shoot in the rice mutant *MET1-2*, which encodes a DNA methyltransferase with a major role in maintaining CG methylation. In wildtype rice, however, only CHH hypermethylation was observed [52]. In our study, the callus versus seedling comparison showed that the A188 *MET1-2* homolog (Zm00056a035610) was ~2× up-regulated in the callus, and *mop1* (Zm00056a013519), a homolog of RNA-dependent RNA polymerase 2 that is involved in the production of 24 nt small RNA [56], was 5–6× up-regulated in the callus, indicating that the transcriptomic machinery was regulated to enhance global DNA methylation in the callus. In plants regenerated from calli, CG and CHG methylation tended to be lost as compared to non-regenerated plants and many methylation events were heritable [57]. Heritable hypomethylation in regenerated plants was observed in an earlier maize study [58]. In rice, as compared to non-regenerated plants in rice, pronounced hypomethylation was found in regenerated plants from tissue culture [59]. The discrepant DNA methylation levels between regenerated plants and calli indicated that most methylation gained from tissue culture is not stable or heritable. Collectively, DNA methylation was elevated during the formation of the callus, likely due to the cellular defense responses. The majority of DNA methylation gained appears to be demethylated during re-differentiation, resulting in hypomethylated regenerated plants.

Conclusions

The genome of a regenerable maize inbred line A188 was assembled with long reads and optical maps, producing a reference-quality genome sequence. Comparison of the A188 genome with the reference B73 genome identified structural variants, including those responsible for phenotypic discrepancies between A188 and B73. Examination of DNA methylation and gene expression with the newly generated A188 reference genome found hypermethylation in the callus as compared with the seedling and the activation of defense genes in the callus, indicative of the defensive state for cellular protection in the embryogenic callus.

Methods

Genetic materials

A188 (PI 693339) seeds were obtained from the North Central Regional Plant Introduction Station in Ames, IA. The A188 inbred line was derived from a cross between the inbred line 4-29 and the inbred line 64, also named A48, followed by four generations of backcross with 4-29. The line 4-29 was derived from the commercial variety Silver King and the line 64 was from a northwestern dent line [1]. Double haploid lines were

developed from the F_1 of B73 (PI 550473) x A188 at the Doubled Haploid Facility at Iowa State University.

Nanopore A188 whole-genome sequencing

A188 were grown in the greenhouse at 28°C and 23°C day/night, with a photoperiod of 14:10 h (light:dark). Nuclei were isolated from seedling leaves using a modified nucleus isolation protocol [60] and dissolved in buffer G2 (Qiagen). The lysate was used for DNA isolation with Qiagen DNeasy Plant Mini Kit (Qiagen) following the manufacturer protocol. A188 genomic DNA was size selected for 15–30 kb and above with the BluePippin cassette kit BLF7510 (Sage Science) high-pass-filtering protocol, followed by a library preparation with the SQK-LSK109 kit (Oxford Nanopore). Each DNA library was loaded on an FLO-MIN106D flowcell and sequenced on MinION (Oxford Nanopore). The basecaller Guppy (version 3.4.4) was used to convert FAST5 raw data to FASTQ data with default parameters.

Illumina A188 whole-genome sequencing

Three independent A188 leaf samples were collected for extracting nuclear DNAs. Two were used for PCR-free paired-end 2x125 bp Illumina sequencing and one was used for PCR-free paired-end 2x250 bp Illumina sequencing on Hiseq2500 at Novogene. In addition, genomic DNA was extracted from A188 immature ears for additional PCR-free paired-end 2x250 bp Illumina sequencing. Therefore, comparable 2x250 bp data were generated from the leaf and ear tissue samples. The 2x125 bp Illumina sequencing data were comparable with the previously generated 2x125 bp B73 whole-genome sequencing data (SRR4039069 and SRR4039070) [61], both of which were used for CGRD analysis.

Assembly of Nanopore data via Canu

FASTQ Nanopore data were assembled with Canu (1.9) [62] with the following options: "'corMhapOptions=--threshold 0.8 --ordered-sketch-size 1000 --ordered-kmer-size 14' correctedErrorRate=0.105 genomeSize=2.4g minReadLength=10000 minOverlapLength=800 corOutCoverage=60".

Contigs filtering

Leaf and ear 2x250 bp data were aligned to the contigs with the "mem" module in bwa (0.7.12-r1039) [63]. Uniquely mapped reads with less than 15% mismatches were used to determine read count per contig with the "intersect" module of BEDTools (v2.29.2) [64]. The log2 of the ratio of read counts normalized by using total reads of leaf and ear samples was calculated for each contig. The contigs with a log2 value larger than 0.5 were considered as the contigs with variable counts from leaf and ear samples. The contigs (N = 21) that had variable counts and less than 100 kb and were not anchored to B73Ref4 via Ragoo (version 1.2) [65] were discarded. In addition, the contigs (N = 16) smaller than 15 kb were also discarded.

Through analysis of read counts, the contigs that had variable counts and matched with the previously sequenced mitochondrion genome sequence (Genbank accession: DQ490952.1) and the chloroplast genome sequence (Genbank accession: KF241980.1)

were identified. One chloroplast contig and 13 mitochondrion contigs were found. The chloroplast contig had almost identical sequences to the Genbank accession KF241980.1. The failure of assembling mitochondrial contigs into one was likely due to heterogeneous forms of mitochondria. In A188Ref1, the previously assembled DQ490952.1 and KF241980.1, were used to represent the mitochondrion and chroloplast genomes, respectively.

Sequence polishing of assembled contigs

After filtering contigs that were derived from organelles or contamination, the remaining contigs were first polished with raw Nanopore reads that contained signal information using Nanopolish (0.11.0) (github.com/jts/nanopolish). Briefly, Nanopore reads were aligned with the contigs using the aligner Minimap2 (2.14-r892) [66]. Polymorphisms, including small insertions and deletions as well as single nucleotide polymorphisms, were called and corrected. The Nanopolish polishing was performed twice, followed by twice polishing with Illumina sequencing data using Pilon (version 1.23) [67]. In each Pilon polishing, reads were aligned to contigs with the module of "mem" in bwa (0.7.12-r1039) [63]. Contigs were corrected with the parameters of "--minmq 40 --minqual 15" using Pilon.

Hybrid scaffolding with Bionano data and polished contigs

Bionano raw molecules were filtered to remove molecules less than 100 kb. The remaining molecules were assembled into Bionano maps with the assembly module in the software Bionano Tools (v1.0). Five times extension and merge iterations and noise parameters were automatically determined by using the parameters of "-i 5 -y". The hybrid scaffolding module from the Bionano Tools was used for scaffolding polished contigs. The conflict filter level for both genome maps and sequences were set to 2 by using the parameters of "-B 2 -N 2".

Construction of a B73xA188 genetic map

Genomic DNA of DH lines was extracted by using BioSprint 96 DNA Plant Kit (Qiagen) and normalized to 10 ng/μL for Genotyping-By-Sequencing (GBS) modified from tunable GBS [68]. Briefly, for each genomic DNA sample, the restriction enzyme *Bsp1286I* (NEB) was used for DNA digestion for 3 h at 37°C, followed by ligation with a barcoded single-stranded oligo with T4 DNA ligase (NEB) for 1 h at 16°C. Enzymatic activity was inactivated at 65°C for 20 min and all samples of ligated DNA were pooled, followed by purification with Qiagen PCR purification kit (Qiagen). The purified ligated DNA was subject to PCR amplification with Q5 high-fidelity DNA polymerase (NEB), followed by purification with Agencourt AMPure XP (Beckman Coulter). The final sequencing library product was prepared by size selection at the range of 200 to 400 bp by a Pippin Prep run with 2% agarose gel cassettes (Sage Science). Illumina sequencing was performed on a HiseqX 10 at Novogene (USA).

Raw FASTQ data were demultiplexed to multiple samples and trimmed to remove barcode sequences and low-quality bases with Trimmomatic (version 0.38). Clean reads were aligned to polished contigs with the "mem" module of bwa and uniquely mapped reads with less than 8% mismatches were used for SNP analysis. SNPs were discovered

by HaplotypeCaller of GATK (version 4.1.0.0) and filtered by SelectVariants of GATK to select biallelic variants [69]. SNP sites with at most 80% missing data, at least 10% minor allele frequency and at most 5% heterozygous rates remained. A segmentation (or binning) algorithm was implemented to determine genotypes of chromosomal segments in each DH line [68]. Genotypes of bin markers of 100 DH lines were used to construct a genetic map (BAgm.v01) with MSTmap [70].

Another genetic map was built using A188Ref1 as the reference genome with 137 DH lines (100 DH lines for BAgm.v01 and 37 additional DH lines). Recombination data was inferred from the genetic map (BAgm.v02) based on A188Ref1 (Additional file 2: Table S13).

ALLMaps to build pseudomolecules

The genetic map that was built with GBS markers was used for further scaffolding. The GBS markers were developed using polished contigs as the reference genome. Each scaffold harbored more than 10 markers. In total, 29 scaffolds were on the map. Scaffolds were aligned to B73Ref4 via NUCMer [71]. Based on the orientation of scaffolds relative to B73Ref4 chromosomes, the order of markers in each linkage group was either kept the same order or flipped to match their orders in B73Ref4. The software ALLMaps (JCVI utility libraries v1.0.6) [72] was run with default parameters, constructing 10 pseudomolecules corresponding to ten A188 chromosomes.

BUSCO assessment

Benchmarking Universal Single-Copy Orthologs (BUSCO) [22] was run in a mode of "genome" to assess the completeness of the assembly with default parameters. BUSCO was run in a mode of "transcriptome" to assess the completeness of the gene annotation with default parameters. Both assessments used the Liliopsida database (liliopsida_odb10) that consisted of 3278 conserved core genes.

Estimation of base errors using KAD analysis

The module "KADprofile.pl" in the KAD tool (version 0.1.7) [21] was used to estimate errors in A188Ref1. The input read data were the merged trimmed Illumina 2x250 bp reads from leaf and immature ears. The k-mer length of 47 mer was used.

Estimation of recombination rates

Genetic distances of non-overlapping 1-Mb windows were estimated. Non-overlapping 1-Mb windows were generated by the module of "makewindows" in BEDTools (v2.29.2) [64]. The last window of each chromosome was discarded due to the smaller size than 1 Mb. The prediction of genetic distance per window utilized a method developed previously [73]. Briefly, a generalized additive model (GAM) was used for the prediction of the genetic distance of any physical interval.

The similar method was used to estimate recombination rates around each gene and repetitive element. For example, for a given element, we first find the midpoint of the element. The genetic positions were then predicted, by GAM, for the position 0.5 Mb upstream and the position 0.5 Mb downstream. The distance of the genetic positions was then used to represent the recombination context of the element.

The recombination rates that are lower than 0.6 cM/Mb and higher than 3 cM/Mb were categorized to low recombination and high recombination, respectively.

Callus induction from immature embryos

A188 ears were harvested at 11 days after pollination (DAP11), and surface-sterilized for 30 min in 50% (v/v) bleach (6% sodium hypochlorite) that contains 3–4 drops of Tween 20 followed by three washes in sterile distilled water. Immature embryos of size 1.0–1.5 mm were isolated and cultured on callus induction medium (CIM) media [74]. CIM was composed of Chu N6 basal medium with vitamins [75] supplemented with 2.3 g/L L-proline, 200 mg/L casein hydrolysate, 3% sucrose, 1 mg/L 2,4-dichlorophenoxyacetic acid, 3 g/L gelrite, pH 5.8. Subculture was conducted every 14 days. The 39-day callus samples were collected for methylome and transcriptome analysis.

Illumina RNA-Seq, transcriptomic assembly, and differential expression

Thirty-three RNA samples were extracted from 11 diverse tissue types of A188 with three biological replicates using RNeasy Plant Mini Kit (Qiagen) (Additional file 2: Table S14). Briefly, the 11 tissues included the root and the above-ground of 10-day-old seedling, three different parts of the 11th leaf tissue at V12, the meiotic tassel, anther, and immature ear at V18, the endosperm and embryo 16 days after pollination, and the callus after 39 days culture of DAP11 immature embryos. RNA quality control, library preparation, and sequencing were performed on an Illumina Novaseq 6000 platform at Novogene. Trimmomatic (version 0.38) [76] was used to trim the adaptor sequence and low-quality bases of RNA-Seq raw reads. The parameters used for the trimming is "ILLUMINACLIP:trimming_db:3:20:10:1:true LEADING:3 TRAILING:3 SLIDINGWINDOW:4:13 MINLEN:40". The trimming adaptor database (trimming_db) includes the sequences: adaptor1, TACACTCTTTCCCTACACGACGCTCTTCCGATCT; adaptor2, GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT. Only paired reads both of which were at least 40 bp after trimming were retained for further analysis.

Trimmed reads were aligned to A188 (A188Ref1) using HISAT2 (version 2.1.0) with the parameters of "-p 8 --dta --no-mixed --no-discordant -k 5 -x" [77]. Alignments whose paired reads were concordantly paired were kept. The software StringTie2 (version 2.1.0) [78] was used to assemble the transcriptome with alignments from a dataset of each A188 sample with the default parameters. In total, 33 transcriptome assemblies from 33 samples were generated. All transcriptome assemblies were merged to build an A188 Illumina transcriptome assembly with the merge function in StringTie2.

Differential expression of the callus relative to other tissue types

Trimmed reads were aligned to A188Ref1 with STAR (2.7.3a) [79]. Uniquely mapped reads with at least 96% coverage and 96% identity were used for determining read counts per gene. DESeq2 (version 1.26.0) [80] was used to identify differential expression between the callus and each of other tissue types. Multiple tests were corrected with the FDR (false discovery rate) approach [81]. The FDR of 5% was set as the threshold.

Nanopore A188 cDNA direct sequencing

Three biological replicates of the seedling and callus samples from the same tissue samples used for Illumina RNA-Seq were sequenced by the Nanopore direct cDNA sequencing protocol. Briefly, mRNA was first isolated from 10 μg total RNA with Poly(A) RNA Selection Kit (Lexogen), followed by direct cDNA library preparation with SQK-DCS109 kit (Oxford Nanopore). The protocol version for library preparation was DCS_9090_v109_revB_04Feb2019. The cDNA library was loaded onto a FLO-MIN106D R9 flowcell and sequenced on MinION (Oxford Nanopore). FAST5 raw data was converted to FASTQ data using the basecaller Guppy version 3.4.5 (Oxford Nanopore) with default parameters. Two trimming steps were employed. Adapter sequence was first trimmed by porechop (version 0.2.4) (https://github.com/rrwick/Porechop) with parameters "--check_reads 10000 --adapter_threshold 100 --end_size 100 --min_trim_size 5 --end_threshold 80 --extra_end_trim 1 --middle_threshold 100 --extra_middle_trim_good_side 5 --extra_middle_trim_bad_side 50", and then poly A was trimmed by the software cutadapt (version 2.6) [82] with the options of " -g T{12} -e 0.1 -a A{12} -n 100". Trimmed reads were aligned to A188Ref1 as unstranded spliced long reads using MiniMap2 (version 2.14) [83] with the parameter "-ax splice". Merged alignments from three replicates were input to StingTie2 for generating assembled transcripts.

Genome annotation

The Maker (2.31.10) pipeline was used for genome annotation [84]. The genome was masked by using Repeatmasker (4.0.7) [85] with the A188 repeat library built by the Extensive de novo TE Annotator (EDTA, v1.8.4) [86]. Two rounds of the maker prediction were performed. At the first round, the A188 assembled transcripts and B73Ref4 protein data were used as EST and protein evidence, respectively. The parameters "est2genome=1" and "protein2genome=1" were set to directly produce gene models from transcripts and proteins. At this round, no ab initio gene predictors were used. Prior to the second maker round, a snap model was trained using the confident gene set from the first round. Gene models produced from round 1 were input as one of the predicted gene models. These gene models were competed with gene models predicted by three gene predictors: snap (2013_11_29) [87], augustus (3.3.3) [88], and fgenesh (v.8.0.0) (softberry.com). ESTs from relative maize genotypes and proteins from closely related species were provided as additional evidence. Gene model output from Maker were further filtered. First genes matched with the following criteria "-evalue 1e-50 -qcov_hsp_perc 60" to the transposon database in Maker were filtered. Second, a transcript retained if it carried Pfam domains from the result of InterProScan (version 5.39-77.0) and/or had an annotation edit distance (AED) less than 0.4, which measured the level of discrepancy of an annotation from supporting evidence.

Functional annotation of transcripts

BLASTP was used to map all proteins to the SWISS-Prot database (https://www.uniprot.org/) with the e-value cutoff of 1e−6. Gene ontology (GO) was extracted from InterProScan.

Identification of a major transcript per gene

For a gene containing multiple transcripts, a major transcript per gene was selected if a transcript had the highest non-zero FPKM (Fragments Per Kilobase of transcript per Million mapped reads) determined from Illumina RNA-Seq datasets of diverse tissues by Cufflink (v2.2.1) [89], and/or the lowest BLASTP e-value to the SWISS-Prot database, and/or the longest transcript length. The BLASTP e-value had a priority relative to the transcript length. If data were not sufficient to make a decision, the one with the longest length was selected.

Syntenic genes between A188 and B73

Syntenic genes were identified with MCscan (JCVI utility libraries v1.0.6) [90]. Major transcripts were used as the input and the parameter "--cscore=.99" was used to find 1-to-1 syntenic gene relationships.

Paralogs in A188 and orthologs between A188 and B73

Paralogs in A188 and orthologs between A188 and B73 were identified with OrthoMCL [91]. Briefly, protein sequences of major transcripts with at least 20 amino acids were used for all-to-all BLASTP with the e-value cutoff of 1e−5. The BLASTP result was input to OrthoMCL to identify paralogous and orthologous groups.

Identification of gene clusters

A gene cluster was defined if at least three genes from a group of A188 paralogs identified by OrthoMCL were physically closely located on a chromosome. The maximum distance is 250 kb for two neighboring genes in a cluster.

Annotation of NLR genes

The NLR genes of A188Ref1 were annotated using the NLR-Annotator pipeline [92].

Repeat annotation

EDTA (v1.8.4) [86] was used for repeat annotation with, maize as the "species" input, the curated maize transposable element database from https://github.com/oushujun/MTEC as the "curatedlib" input, and B73 coding sequences as the "cds' input.

Analysis of NUMT and NUPT

The "nucmer" command from the software MUMmer 4 [93] was used to align the A188 mitochondrion or chloroplast genomes to A188Ref1. For mitochondrial alignments, each required at least 5 kb and 95% identity. For chloroplast DNA alignments, each required at least 3 kb and 95% identity based on the minimal requirement for a sufficient FISH signal [26]. Multiple alignments with the distance less than 100 kb were clustered into a block, considered to be a nuclear integration event.

Comparative genomic analysis via SyRI and CGRD

The "nucmer" command was used for whole-genome alignment of 10 chromosomal pseudomolecules between A188Ref1 and B73Ref4. The parameter of "--maxmatch -c 500 -b 500 -l 50" was used in the command "nucmer" and the parameter of "-i

95 -l 1000 -m" in the command of delta-filter, which resulted in best alignments with at least 1-kb matches and at least 95% identity between the two assembled genomes. The "show-coords" command with the parameter of "-THrd" was run to convert alignments to a tab-delimited flat text format. Alignment results were then used for identifying genomic structural variation and nucleotide polymorphisms through SyRI (v1.2) [32] with the parameter of "--allow-offset 100". Syri analysis discovered genome duplication, translocation, inversion, as well as syntenic, unaligned, divergent sequences. SNPs, small insertions, and deletions were identified as well. SyRI analyses using minimap2 alignments were also performed between A188Ref1 and each of the newly assembled genomes of NAM founders, including version 5 of B73Ref5 [18], as well as genome assemblies of Mo17 [10] and SK [11]. The parameter used for minimap2 alignments is "-ax asm10 -t 16 -K 800M -f 500 --eqx" [83].

The CGRD pipeline (v0.1) (github.com/liu3zhenlab/CGRD) was employed to find copy number variation (CNV) through comparing depths of Illumina reads from A188 and B73 with the default parameters [33]. A value of the log2 read depth ratio per sequence segment (*LogRD*) is the indication for CNV. For a segment, the *LogRD* is close to zero if sequences of two genotypes are identical and no CNVs. The sufficient deviation of the mean of *LogRD* from zero is likely due to CNV or a high level of divergence. CGRD was performed using A188Ref1 as the reference genome and identified sequences of A188Ref1 showing conserved (B73 = A188), copy number plus (B73 > A188), and copy number minus (B73 < A188) in B73 relative to A188. When B73Ref4 was used as the reference, the analysis found sequences of B73Ref4 showing conserved (A188 = B73), copy number plus (A188 > B73), and copy number minus (A188 < B73) in A188 relative to B73.

Identification of PAV or highly divergent sequences (HDS)

SyRI analysis listed B73Ref4 sequences that were not aligned to A188Ref1, and vice versa, as well as insertion/deletion polymorphisms between the two chromosomal sequences. Unaligned sequences or insertion/deletion polymorphisms identified by SyRI were compared with CGRD segments. For each SyRI event, a supporting score of read depth data from CGRD was determined by using the formula of $\sum_i^n \frac{-LogRD_i \times O_i}{L}$, where i represents the ith overlap between a CGRD segment and a SyRI event; *LogRD* stands for the *LogRD* of the CGRD segment and only negative values were taken into calculation; O is the overlapping length in bp; L is the length in bp of the SyRI event; and n is the total number of overlaps. The resulting value from the formula represents the degree of the differentiation in read depth between the two genotypes for the SyRI event. The higher the number, the more confidence the PAV or HDS event. A SyRI event is considered to be a PAV or HDS if a supporting score is larger than 3.

Identification of large inversion events

Inversion between A188Ref1 and B73Ref4 were revealed by SyRI. Large events with both A188 and B73 sequences larger than 0.5 Mb were extracted. First, the inversion sequences of B73Ref4 were aligned to B73Ref5 to confirm the inverted orientation

relative to A188Ref1. For a given inversion, if >80% B73Ref4 sequences were aligned to B73Ref5 in the plus orientation, the inversion was supported by B73Ref5. If <20% B73Ref4 sequences aligned to B73Ref5 were in the plus orientation, the inversion was considered to not be supported by B73Ref5. Second, the recombination frequency between the start and the end of an inversion event was estimated and adjusted to cM per Mb. Third, SNPs between the two genomes and located on the inversion were identified. The common SNPs genotyped in the maize 282 [94] population were extracted for determining linkage disequilibrium (LD) between SNPs at a distance of 0.2–0.3 Mb. Vcftools (v0.1.17) [95] was employed to calculate LD. The genome-wide LDs between SNPs at a distance of 0.2 Mb were determined as the control.

Structure analysis of inversions in maize HapMap2 population

The software STRUCTURE (v2.3.4) [96] was used to analyze the inference of population structure for A188 inversions in maize HapMap2 population [97]. A188 and B73 SNPs between inversion regions were discovered by SyRI. HapMap2 genotyping data overlapping with inversion SNPs were extracted and the subset of SNPs with the missing rate less than 20% were input for STRUCTURE analysis. The major alleles, minor allele, and missing locus in SNP dataset were converted to 0, 1, and −1, respectively. K = 2 as the cluster number and 10 replicate runs of the admixture model were used, with a burn-in of 10,000 iterations and a run length of 20,000 steps.

Fluorescence in situ hybridization (FISH)

Mitotic and meiotic chromosomes were prepared as described by Koo and Jiang (2009) with minor modifications [98]. Root tips were collected from seedling plants and treated in a nitrous oxide gas chamber for 1.5 h, fixed overnight in ethanol:glacial acetic acid (3:1), and then squashed in a drop of 45% acetic acid. Anthers were squashed in 45% acetic acid on a slide and checked under a phase microscope. All preparations were stored at −70°C until use.

DNA probes of the CentC, Knob, Cent4 [99], and the probes for examining NUMTs, the PME cluster, and a potential large inversion on chromosome 4 (Additional file 2: Table S15) were labeled with digoxigenin-11-dUTP (Roche, Indianapolis, IN), biotin-16-dUTP (Roche), and/or DNP-11-dUTP (PerkinElmer), depending on whether two or three probes were used in the FISH experiment [99]. The FISH hybridization procedure was according to a previously published protocol [100]. After post-hybridization washes, the probes were detected with Alexa Fluor 488 streptavidin (Invitrogen) for biotin-labeled probes and rhodamine-conjugated anti-digoxigenin for dig-labeled probe (Roche). The DNP-labeled probe was detected with rabbit anti-DNP, followed by amplification with a chicken anti-rabbit Alexa Fluor 647 antibody (Invitrogen). Chromosomes were counterstained with 4′,6-diamidino-2-phenylindole (DAPI) in Vectashield antifade solution (Vector Laboratories). The images were captured with a Zeiss Axioplan 2 microscope (Carl Zeiss Microscopy LLC) using a cooled CCD camera CoolSNAP HQ2 (Photometrics) and AxioVision 4.8 software. The final contrast of the images was processed using Adobe Photoshop CS5 software (Adobe).

QTL mapping

Kernel colors of 125 B73xA188 DH lines were scored as 1 to 6 (1 = white, 6 = yellow, and 2 to 5 indicated colors between white and yellow). QTL mapping of kernel color was performed by using scanone function with the Haley-Knott regression method in the R package rqtl [101]. The LOD cutoff was the 5% highest LOD value from 1000 permutations of phenotypic data.

qRT-PCR

qRT-PCR was used to measure the gene expression of *ccd1* and *y1* gene in genotypes of A188 and B73 as well as two DH lines DH305 and DH312. Immature ears of the four genotypes were harvested from the summer nursery in Manhattan Kansas at 16 days after pollination (DAP16). Fifteen kernels were randomly sampled from the middle of an ear, five kernels of which were pooled as a biological replication for RNA isolation. cDNA was synthesized with Verso cDNA Kit (Thermo Scientific) following the manufacturer's protocol. qRT-PCR was performed in a reaction of 10 μL with the IQTM SYBR Green Supermix reagent (BioRad) on the CFX96 Real-Time PCR System (BioRad). The thermocycling conditions for the PCR included an initial denaturation at 95°C for 3 min, followed by 40 cycles of denature at 95°C for 15 s, annealing, and extension at 60°C for 30 s. The housekeeping reference gene *actin1* was used as the internal control. Cycle threshold values (*Ct*) of two technical replicates were averaged and used to quantify relative gene expression levels. The relative expression levels of each of *ccd1* and *y1* genes in each sample were calculated using the formula $100 \times 2^{actinCt\text{-}geneC}$, where *actinCt* and *geneCt* stand for the *Ct* values of *actin1* and *ccd1* (or *y1*), respectively. The primers used are as follows: *actin1*: act1_qrt_2F and act1_qrt_2R; *ccd1*: ccd1_qrt_5F and ccd1_qrt_5R; *y1*: y1_qrt_4F and y1_qrt_4R. Sequences of primers are in Additional file 2: Table S15.

Kernel carotenoid content measurement

The seed total carotenoids of different genotypes with three biological replicates were measured following the protocol described by Mishra and Singh [102]. Briefly, 10 maize seeds from each replicate were milled into fine flour using mortar and pestle. After the flour was filtered using muslin cloth, 0.5 g of the fine powder was dissolved into 6 mL of 1% butylated hydroxytoluene. The well-mixed samples were treated following the protocol [102], and the optical density at 450 nm of the treated sample was measured using Genesys 20 spectrophotometer (Thermo scientific). The concentration of total carotenoids was calculated using the formula described by Mishra and Singh [102].

Whole-genome bisulfite sequencing (WGBS)

Genomic DNA from two biological replicates of the seedling and callus samples that were used for RNA-Seq were subjected to WGBS on a Novaseq 6000 at Novogene (USA). A Bismark pipeline (v0.22.1) was adapted to process bisulfite sequencing DNA methylation data [103]. Briefly, raw reads were subjected to Trimmomatic trimming (v0.38) [76] to remove adaptor and poor-quality sequences. Bowtie2 (v2.3.5.1) [104] built in Bismark was used for the alignment and alignments of duplicated reads were removed before methylation calling. The methylation levels per cytosine site of all three

sequence contexts (CG, CHG, and CHH) were determined, which were used for identifying differentially methylated regions (DMRs) with the DSS R package (v2.34.0) (github.com/haowulab/DSS).

DNA methylation around genes and on repetitive sequences

Genomic regions (gene body) from the translation start site (TSS) to the translation termination site (TTS), which were based on genomic locations of major transcripts, were equally divided into 200 windows. For each gene, the 2 kb in the 5′ upstream region and the 2 kb in the 3′ downstream region of each gene were also extracted. The DNA methylation rate in three sequence contexts (CG, CHG, and CHH) on each window of the gene body or each 20 bp in upstream and downstream regions was separately determined to examine the distribution of DNA methylation on and around genes.

DMRs were located in the three regions, 5′ upstream 1 kb, gene body, 3′ downstream 1 kb. For each region, the independence between changes of DNA methylation, increased or decreased in the callus versus the seedling, and regulation in gene expression, up- or down-regulated in the callus versus the seedling from DE analysis, was examined through χ^2 statistical test. Tests were performed for all three methylation types: CG, CHG, and CHH.

DNA methylation rates per 100 bp of repetitive sequences were determined. Annotation of repetitive types was from EDTA and additional 45S rDNA alignment analysis. Paired t-test was performed between the two tissues: callus and seedling.

Tissue network and principal component analyses of A188 tissues

The A188 tissue network was constructed with the R package WGCNA (version 1.66) [105] using the expression of 29,222 genes in 33 RNA-Seq datasets from 11 A188 tissue types. WGCNA was performed to cluster A188 tissue samples with the parameters minModuleSize = 6 and soft-thresholding power = 9. The Gephi software (version 0.9.2) [106] was used to visualize tissue networks with the module and connectivity information from the WGCNA result. Principal component analysis (PCA) was also performed using the R functions *prcomp* with the expression per gene averaged from three replicates per tissue type.

Gene ontology (GO) enrichment analysis

The enrichment analyses were performed to determine if a certain GO was over-represented in a selected group of genes. The resampling method in GOSeq [107] was employed.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-021-02396-x.

Additional file 1. Supplementary Figures S1-S32.
Additional file 2. Supplemental Tables S1-S15.
Additional file 3. SyRI output.
Additional file 4. CGRD output with B73Ref4 as the reference.
Additional file 5. CGRD output with A188Ref1 as the reference.
Additional file 6. A188 genes overlapped with A188-specific sequences.

Additional file 7. B73 genes overlapped with B73-specific sequences.

Additional file 8. Normalized read counts from RNA-Seq of 11 tissues.

Additional file 9. CG DMRs between seedling and callus.

Additional file 10. CHG DMRs between seedling and callus.

Additional file 11. CHH DMRs between seedling and callus.

Additional file 12. Review history.

Acknowledgements

We thank Drs. Candice A.C. Gardner and Mark J. Millard at the North Central Regional Plant Introduction Station for distributing A188 seeds, Dr. Candice Hirsch from the University of Minnesota to provide guidance of tissue collection for RNA-Seq, Dr. Kathleen J. Newton from the University of Missouri for the discussion of NUMT and NUPT in maize genomes, Dr. Bill Tracy for the information about the A188 origin, and Drs. Christopher Toomajian and Guihua Bai for comments on the manuscript before publication. We thank the Doubled Haploid Facility at Iowa State University for producing DH lines and computational support from the Beocat High-Performance Computing Facility at Kansas State University. We thank funding support from the US National Science Foundation (awards No. 1741090 and 1656006), the USDA's National Institute of Food and Agriculture (award no. 2018-67013-28511), and the Agricultural Science and Technology Innovation Program of the Chinese Academy of Agricultural Sciences. This is the contribution number 21-040-J from the Kansas Agricultural Experiment Station.

Peer review information

Kevin Pang was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history

The review history is available as Additional file 12.

Authors' contributions

S.L., F.F.W, S.P. M.C., and H.W. conceived and designed experiments. G.L., D.K., H.Z., T.M.T., Y.L. (Yan L.), M.Z., Y.H., Y.Q., Y.L. (Yunjun L.) performed experiments and collected data. G.L., C.H., J.L., H.T., and S.L. analyzed data. H.T. participated in the overall design of the strategies of genome assembly and genome comparison. B.W. assisted in the genome annotation. F.M., H.K., and S.M.K. participated in the maintenance of DH lines. J.Z. and G.W. provided Bionano data. S.L. and P.S.S conceived the CGRD approach. S.L. and N.S. designed and interpreted DNA methylation experiments. D.R.M. contributes ideas and discussion on the kernel color experiment. G.L., C.H., J.Z, H.W., F.F.W., N.S., P.S.S, G.W., and S.L. wrote the manuscript with comments from other authors. The authors read and approved the final manuscript.

Availability of data and materials

The A188 genome assembly described in this paper is version JABWIA010000000 at NCBI [108]. The annotation (A188Ref1a1) is available at MaizeGDB.org [109]. Raw Nanopore whole-genome sequencing data, Illumina whole-genome sequencing data, Nanopore cDNA sequencing data, Illumina RNA-Seq data, and whole-genome bisulfite are available at NCBI SRA under the project of PRJNA635654 [110]. Essential scripts related to the manuscript are available at Zenodo [111] and Github [112]. The CGRD pipeline can be downloaded from Zenodo [113] and Github [114]. All scripts are distributed under the MIT license.

Competing interests

P.S.S is a co-lead of the Genomes to Fields Initiative and a Changjiang Scholar at China Agriculture University. He is co-founder and managing partner of Data2Bio, LLC; Dryland Genetics, LLC; EnGeniousAg, LLC; and LookAhead Breeding, LLC. He is a member of the scientific advisory board and a shareholder of Hi-Fidelity Genetics, Inc., and a member of the scientific advisory boards of Kemin Industries and Centro de Tecnologia Canavieira. S.L is co-founder of Data2Bio, LLC. The authors affirm that they have no other conflicts of interest.

Author details

[1]Department of Plant Pathology, Kansas State University, 4024 Throckmorton Center, Manhattan, KS 66506-5502, USA. [2]Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, China. [3]Department of Horticulture and Natural Resources, Kansas State University, Manhattan, KS 66506-5502, USA. [4]Present Address, Corvallis, OR 97330, USA. [5]Department of Agronomy, University of Wisconsin-Madison, Madison, WI 53706, USA. [6]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA. [7]Center for Genomics and Biotechnology and Fujian Provincial Key Laboratory of Haixia Applied Plant Systems Biology, Fujian Agriculture and Forestry University, Fuzhou 350002, Fujian, China. [8]Department of Horticulture, University of Florida, Gainesville, FL 32611-0680, USA. [9]College of Forest Resources and Environmental Science, Michigan Technological University, Houghton, MI 49931, USA. [10]Innovative Genomics Institute, University of California-Berkeley, Sunnyvale, CA 94704, USA. [11]Department of Plant Biology, University of Minnesota, Saint Paul, MN 55108, USA. [12]Department of Agronomy, Iowa State University, Ames, IA 50011-3605, USA. [13]Department of Plant Pathology, University of Florida, Gainesville, FL 32611-0680, USA.

Received: 21 November 2020 Accepted: 28 May 2021
Published online: 09 June 2021

References

1. Gerdes JT, Behr CF, Coors JG, Tracy WF. Compilation of North American maize breeding germplasm. Madison: Crop Science Society of America; 1993.

2. Wang Q, Dooner HK. Remarkable variation in maize genome structure inferred from haplotype diversity at the bz locus. Proc Natl Acad Sci U S A. 2006;103:17644–9.
3. Liu K, Goodman M, Muse S, Smith JS, Buckler E, Doebley J. Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. Genetics. 2003;165:2117–28.
4. Rhodes CA, Pierce DA, Mettler IJ, Mascarenhas D, Detmer JJ. Genetically transformed maize plants from protoplasts. Science. 1988;240:204–7.
5. Vega JM, Yu W, Kennon AR, Chen X, Zhang ZJ. Improvement of Agrobacterium-mediated transformation in Hi-II maize (Zea mays) using standard binary vectors. Plant Cell Rep. 2008;27:297–305.
6. Armstrong CL, Green CE, Phillips RL. Development and availability of germplasm with high type II culture formation response. Maize Genet Coop News Lett. 1991;65:92–3.
7. Wisser RJ, Kolkman JM, Patzoldt ME, Holland JB, Yu J, Krakowsky M, et al. Multivariate analysis of maize disease resistances suggests a pleiotropic genetic basis and implicates a GST gene. Proc Natl Acad Sci U S A. 2011;108:7339–44.
8. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 maize genome: complexity, diversity, and dynamics. Science. 2009;326:1112–5.
9. Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, et al. Improved maize reference genome with single-molecule technologies. Nature. 2017;546:524–7.
10. Sun S, Zhou Y, Chen J, Shi J, Zhao H, Zhao H, et al. Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. Nat Genet. 2018;50:1289–95.
11. Yang N, Liu J, Gao Q, Gui S, Chen L, Yang L, et al. Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. Nat Genet. 2019;51:1052–9.
12. Ou S, Liu J, Chougule KM, Fungtammasan A, Seetharam AS, Stein JC, et al. Effect of sequence depth and length in long-read assembly of the maize inbred NC358. Nat Commun. 2020;11:2288.
13. Springer NM, Anderson SN, Andorf CM, Ahern KR, Bai F, Barad O, et al. The maize W22 genome provides a foundation for functional genomics and transposon biology. Nat Genet. 2018;50:1282–8.
14. Hirsch CN, Hirsch CD, Brohammer AB, Bowman MJ, Soifer I, Barad O, et al. Draft assembly of elite inbred line PH207 provides insights into genomic and transcriptome diversity in maize. Plant Cell. 2016;28:2700–14.
15. Haberer G, Kamal N, Bauer E, Gundlach H, Fischer I, Seidel MA, et al. European maize genomes highlight intraspecies variation in repeat and gene content. Nat Genet. 2020;52:950–7.
16. Liu J, Seetharam AS, Chougule K, Ou S, Swentowsky KW, Gent JI, et al. Gapless assembly of maize chromosomes using long-read technologies. Genome Biol. 2020;21:121.
17. Hu Y, Colantonio V, Müller BSF, Leach KA, Nanni A, Finegan C, et al. Genome assembly and population genomic analysis provide insights into the evolution of modern sweet corn. Nat Commun. 2021;12:1227.
18. Hufford MB, Seetharam AS, Woodhouse MR. De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. bioRxiv. 2021; biorxiv.org. https://doi.org/10.1101/2021.01.14.426684.
19. Clifton SW, Minx P, Fauron CM-R, Gibson M, Allen JO, Sun H, et al. Sequence and comparative analysis of the maize NB mitochondrial genome. Plant Physiol. 2004;136:3486–503.
20. Bosacchi M, Gurdon C, Maliga P. Plastid genotyping reveals the uniformity of cytoplasmic male sterile-T maize cytoplasms. Plant Physiol. 2015;169:2129–37.
21. He C, Lin G, Wei H, Tang H, White FF, Valent B, et al. Factorial estimating assembly base errors using k-mer abundance difference (KAD) between short reads and genome assembled sequences. NAR Genomics Bioinformatics. 2020;2: lqaa075.
22. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2.
23. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009;19:1639–45.
24. Kato A, Lamb JC, Birchler JA. Chromosome painting using repetitive DNA sequences as probes for somatic chromosome identification in maize. Proc Natl Acad Sci U S A. 2004;101:13554–9.
25. Lough AN, Roark LM, Kato A, Ream TS, Lamb JC, Birchler JA, et al. Mitochondrial DNA transfer to the nucleus generates extensive insertion site variation in maize. Genetics. 2008;178:47–55.
26. Roark LM, Hui AY, Donnelly L, Birchler JA, Newton KJ. Recent and frequent insertions of chloroplast DNA into maize nuclear chromosomes. Cytogenet Genome Res. 2010;129:17–23.
27. Hulbert SH. Structure and evolution of the rp1 complex conferring rust resistance in maize. Annu Rev Phytopathol. 1997;35:293–310.
28. Hu Y, Ren J, Peng Z, Umana AA, Le H, Danilova T, et al. Analysis of extreme phenotype bulk copy number variation (XP-CNV) identified the association of rp1 with resistance to Goss's wilt of maize. Front Plant Sci. Frontiers. 2018;9:110.
29. Smith SM, Pryor AJ, Hulbert SH. Allelic and haplotypic diversity at the rp1 rust resistance locus of maize. Genetics. 2004; 167:1939–47.
30. Sun Q, Collins NC, Ayliffe M, Smith SM, Drake J, Pryor T, et al. Recombination between paralogues at the Rp1 rust resistance locus in maize. Genetics. 2001;158:423–38.
31. Bennetzen JL, Qin M-M, Ingels S, Ellingboe AH. Allele-specific and Mutator-associated instability at the Rpl disease-resistance locus of maize. Nature. Nature Publishing Group. 1988;332:369–70.
32. Goel M, Sun H, Jiao W-B, Schneeberger K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome Biol. 2019;20:277.
33. Peng Z, Oliveira-Garcia E, Lin G, Hu Y, Dalby M, Migeon P, et al. Effector gene reshuffling involves dispensable mini-chromosomes in the wheat blast fungus. PLoS Genet. Public Library of Science San Francisco, CA USA. 2019;15: e1008272.
34. Morgan DT. A cytogenetic study of inversions in Zea mays. Genetics. 1950;35:153–74.
35. Zhang Z, Zhang B, Chen Z, Zhang D, Zhang H, Wang H, et al. A PECTIN METHYLESTERASE gene at the maize Ga1 locus confers male function in unilateral cross-incompatibility. Nat Commun. 2018;9:3678.
36. Vogel JT, Tan B-C, McCarty DR, Klee HJ. The carotenoid cleavage dioxygenase 1 enzyme has broad substrate specificity, cleaving multiple carotenoids at two different bond positions. J Biol Chem. 2008;283:11364–73.

37. Tan B-C, Guan J-C, Ding S, Wu S, Saunders JW, Koch KE, et al. Structure and origin of the white cap locus and its role in evolution of grain color in maize. Genetics. 2017;206:135–50.
38. Romay MC, Millard MJ, Glaubitz JC, Peiffer JA, Swarts KL, Casstevens TM, et al. Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol. 2013;14:R55.
39. Buckner B, Kelson TL, Robertson DS. Cloning of the y1 locus of maize, a gene involved in the biosynthesis of carotenoids. Plant Cell Am Soc Plant Biol. 1990;2:867–76.
40. Phelps TL, Hall AE, Buckner B. Microsatellite repeat variation within the y1 gene of maize and teosinte. J Hered. 1996;87:396–9.
41. Grotewold E, Drummond BJ, Bowen B, Peterson T. The myb-homologous P gene controls phlobaphene pigmentation in maize floral organs by directly activating a flavonoid biosynthetic gene subset. Cell. 1994;76:543–53.
42. Lowe K, Wu E, Wang N, Hoerster G, Hastings C, Cho M-J, et al. Morphogenic regulators Baby boom and Wuschel improve monocot transformation. Plant Cell. 2016;28:1998–2015.
43. Salvo S, Cook J, Carlson AR, Hirsch CN, Kaeppler SM, Kaeppler HF. Genetic fine-mapping of a quantitative trait locus (QTL) associated with embryogenic tissue culture response and plant regeneration ability in maize (Zea mays L.). Plant Genome. 2018;11:170111.
44. Li Q, Gent JI, Zynda G, Song J, Makarevitch I, Hirsch CD, et al. RNA-directed DNA methylation enforces boundaries between heterochromatin and euchromatin in the maize genome. Proc Natl Acad Sci U S A. 2015;112:14728–33.
45. Gent JI, Ellis NA, Guo L, Harkess AE, Yao Y, Zhang X, et al. CHH islands: de novo DNA methylation in near-gene chromatin regulation in maize. Genome Res. 2013;23:628–37.
46. Sartor RC, Noshay J, Springer NM, Briggs SP. Identification of the expressome by machine learning on omics data. Proc Natl Acad Sci U S A. 2019;116:18119–25.
47. Li Q, Song J, West PT, Zynda G, Eichten SR, Vaughn MW, et al. Examining the causes and consequences of context-specific differential DNA methylation in maize. Plant Physiol. 2015;168:1262–74.
48. Stelpflug SC, Sekhon RS, Vaillancourt B, Hirsch CN, Robin Buell C, de Leon N, et al. An expanded maize gene expression atlas based on RNA sequencing and its use to explore root development. Plant Genome. 2016;9:lantgenome2015.04.0025.
49. Chandler K, Lipka AE, Owens BF, Li H, Buckler ES, Rocheford T, et al. Genetic analysis of visually scored orange kernel color in maize. Crop Sci. 2013;53:189–200.
50. Owens BF, Mathew D, Diepenbrock CH, Tiede T, Wu D, Mateos-Hernandez M, et al. Genome-wide association study and pathway-level analysis of kernel color in maize. G3. 2019;1:183–6.
51. Ikeuchi M, Sugimoto K, Iwase A. Plant callus: mechanisms of induction and repression. Plant Cell. 2013;25:3159–73.
52. Hu L, Li N, Zhang Z, Meng X, Dong Q, Xu C, et al. CG hypomethylation leads to complex changes in DNA methylation and transpositional burst of diverse transposable elements in callus cultures of rice. Plant J. 2020;101:188–203.
53. Lee M, Phillips RL. The chromosomal basis of somaclonal variation. Annu Rev Plant Physiol Plant Mol Biol. Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA. 1988;39:413–37.
54. Boyko A, Kathiria P, Zemp FJ, Yao Y, Pogribny I, Kovalchuk I. Transgenerational changes in the genome stability and methylation in pathogen-infected plants: (virus-induced plant genome instability). Nucleic Acids Res. 2007;35:1714–25.
55. Liu H, Ma L, Yang X, Zhang L, Zeng X, Xie S, et al. Integrative analysis of DNA methylation, mRNAs, and small RNAs during maize embryo dedifferentiation. BMC Plant Biol. 2017;17:105.
56. Nobuta K, Lu C, Shrivastava R, Pillay M, De Paoli E, Accerbi M, et al. Distinct size distribution of endogeneous siRNAs in maize: evidence from deep sequencing in the mop1-1 mutant. Proc Natl Acad Sci U S A. 2008;105:14958–63.
57. Han Z, Crisp PA, Stelpflug S, Kaeppler SM, Li Q, Springer NM. Heritable epigenomic changes to the maize methylome resulting from tissue culture. Genetics. 2018;209:983–95.
58. Kaeppler SM, Phillips RL. Tissue culture-induced DNA methylation variation in maize. Proc Natl Acad Sci U S A. 1993;90:8773–6.
59. Stroud H, Ding B, Simon SA, Feng S, Bellizzi M, Pellegrini M, et al. Plants regenerated from tissue culture contain stable epigenome changes in rice. Elife. 2013;2:e00354.
60. Zhang M, Zhang Y, Scheuring CF, Wu C-C, Dong JJ, Zhang H-B. Preparation of megabase-sized DNA from a variety of organisms using the nuclei method for advanced genomics research. Nat Protoc. 2012;7:467–78.
61. Liu S, Zheng J, Migeon P, Ren J, Hu Y, He C, et al. Unbiased k-mer analysis reveals changes in copy number of highly repetitive sequences during maize domestication and improvement. Sci Rep. 2017;7:42444.
62. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27:722–36.
63. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics. Narnia. 2010;26:589–95.
64. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
65. Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, et al. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. Genome Biol. 2019;20:224.
66. Li H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics. 2016;32:2103–10.
67. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014;9:e112963.
68. Ott A, Liu S, Schnable JC, Yeh CTE. tGBS® genotyping-by-sequencing enables reliable genotyping of heterozygous loci. Nucleic Acids. academic.oup.com. 2017;45:e178.
69. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.
70. Wu Y, Bhat PR, Close TJ, Lonardi S. Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. PLoS Genet. 2008;4:e1000212.
71. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: a fast and versatile genome alignment system. PLoS Comput Biol. 2018;14:e1005944.

72. Tang H, Zhang X, Miao C, Zhang J, Ming R, Schnable JC, et al. ALLMAPS: robust scaffold ordering based on multiple maps. Genome Biol. 2015;16:3.
73. Liu S, Yeh C-T, Ji T, Ying K, Wu H, Tang HM, et al. Mu transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome. PLoS Genet. 2009;5:e1000733.
74. Rakshit S, Rashid Z, Sekhar JC, Fatma T, Dass S. Callus induction and whole plant regeneration in elite Indian maize (Zea mays L.) inbreds. Plant Cell Tissue Organ Cult. Springer. 2010;100:31–7.
75. Chu C. Establishment of an efficient medium for another culture of rice through comparative experiments on the nitrogen sources. Sci Sin. 1975;18:223–31.
76. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
77. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37:907–15.
78. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol. 2019;20:278.
79. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.
80. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
81. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol. 1995;57:289–300.
82. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17:10–2.
83. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.
84. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 2011;12:491.
85. Smit AFA, Hubley R, Green P. RepeatMasker. Open-40. 2013-2015.
86. Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 2019;20:275.
87. Korf I. Gene finding in novel genomes. BMC Bioinformatics. 2004;5:59.
88. Stanke M, Waack S. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics. 2003;19(Suppl 2):ii215–25.
89. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28:511–5.
90. Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH. Synteny and collinearity in plant genomes. Science. 2008;320:486–8.
91. Li L, Stoeckert CJ Jr, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13:2178–89.
92. Steuernagel B, Witek K, Krattinger SG, Ramirez-Gonzalez RH, Schoonbeek H-J, Yu G, et al. The NLR-Annotator tool enables annotation of the intracellular immune receptor repertoire. Plant Physiol. 2020;183:468–82.
93. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004;5:R12.
94. Bukowski R, Guo X, Lu Y, Zou C, He B, Rong Z, et al. Construction of the third-generation Zea mays haplotype map. Gigascience. 2018;7:1–12.
95. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.
96. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155:945–59.
97. Hufford MB, Xu X, van Heerwaarden J, Pyhäjärvi T, Chia J-M, Cartwright RA, et al. Comparative population genomics of maize domestication and improvement. Nat Genet. 2012;44:808–11.
98. Koo D-H, Jiang J. Super-stretched pachytene chromosomes for fluorescence in situ hybridization mapping and immunodetection of DNA methylation. Plant J. Wiley Online Library. 2009;59:509–16.
99. Koo D-H, Zhao H, Jiang J. Chromatin-associated transcripts of tandemly repetitive DNA sequences revealed by RNA-FISH. Chromosome Res. 2016;24:467–80.
100. Koo D-H, Han F, Birchler JA, Jiang J. Distinct DNA methylation patterns associated with active and inactive centromeres of the maize B chromosome. Genome Res. 2011;21:908–14.
101. Broman KW, Wu H, Sen S, Churchill GA. R/qtl: QTL mapping in experimental crosses. Bioinformatics. 2003;19:889–90.
102. Mishra P, Singh NK. Spectrophotometric and TLC based characterization of kernel carotenoids in short duration maize. Maydica. 2010;55:95.
103. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics. 2011;27:1571–2.
104. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
105. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.
106. Bastian M, Heymann S, Jacomy M. Others. Gephi: an open source software for exploring and manipulating networks. Icwsm. San Jose, California. 2009;8:361–2.
107. Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. 2010;11:R14.
108. Lin G, He C, Zheng J, Koo D-H, Le H, Zheng H, et al. Chromosome-level genome assembly of a regenerable maize inbred line A188. NCBI GenBank. https://www.ncbi.nlm.nih.gov/nuccore/JABWIA000000000.1. 2020.
109. Lin G, He C, Zheng J, Koo D-H, Le H, Zheng H, et al. Chromosome-level genome assembly of a regenerable maize inbred line A188. MaizeGDB. https://download.maizegdb.org/Zm-A188-REFERENCE-KSU-1.0. 2020.

110. Lin G, He C, Zheng J, Koo D-H, Le H, Zheng H, et al. Chromosome-level genome assembly of a regenerable maize inbred line A188. NCBI SRA. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA635654. 2020.
111. Lin G, He C, Zheng J, Koo D-H, Le H, Zheng H, et al. Script documents for genome assembly of the maize inbred line A188. Zenodo. 2021. https://doi.org/10.5281/zenodo.4758770.
112. Lin G, He C, Zheng J, Koo D-H, Le H, Zheng H, et al. Chromosome-level genome assembly of a regenerable maize inbred line A188. GitHub. https://github.com/liu3zhenlab/A188Ref1. 2020.
113. Lin G, Liu S. Comparative genomic read depth for studying genomic copy number variation. Zenodo. 2021. https://doi.org/10.5281/zenodo.4758649.
114. Lin G, He C, Zheng J, Koo D-H, Le H, Zheng H, et al. Chromosome-level genome assembly of a regenerable maize inbred line A188. GitHub. https://github.com/liu3zhenlab/CGRD. 2020.

Publisher's Note

Supplementary Data

Chromosome-level Genome Assembly of a Regenerable Maize Inbred Line A188

Lin el al.

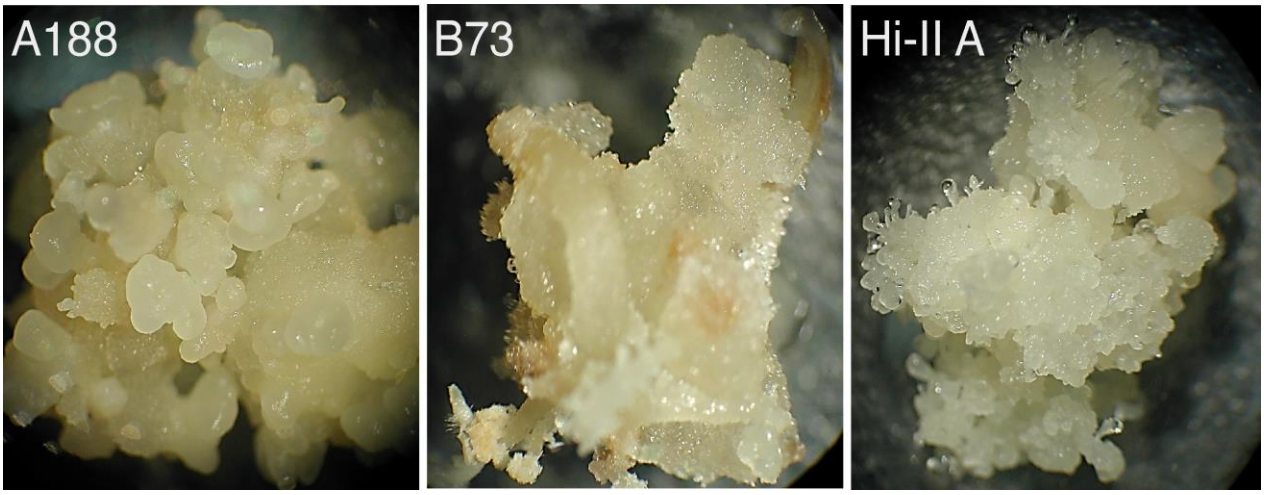

Figure S1. Calli from immature embryos of A188, B73, and Hi-II A. a, **b**, **c**) White, compact, and nodulated somatic embryos and embryogenic callus from A188, brownish, loose and non-embryogenic callus from B73, and white, friable and embryogenic callus from Hi-II A after 28 days of culture in callus induction media.

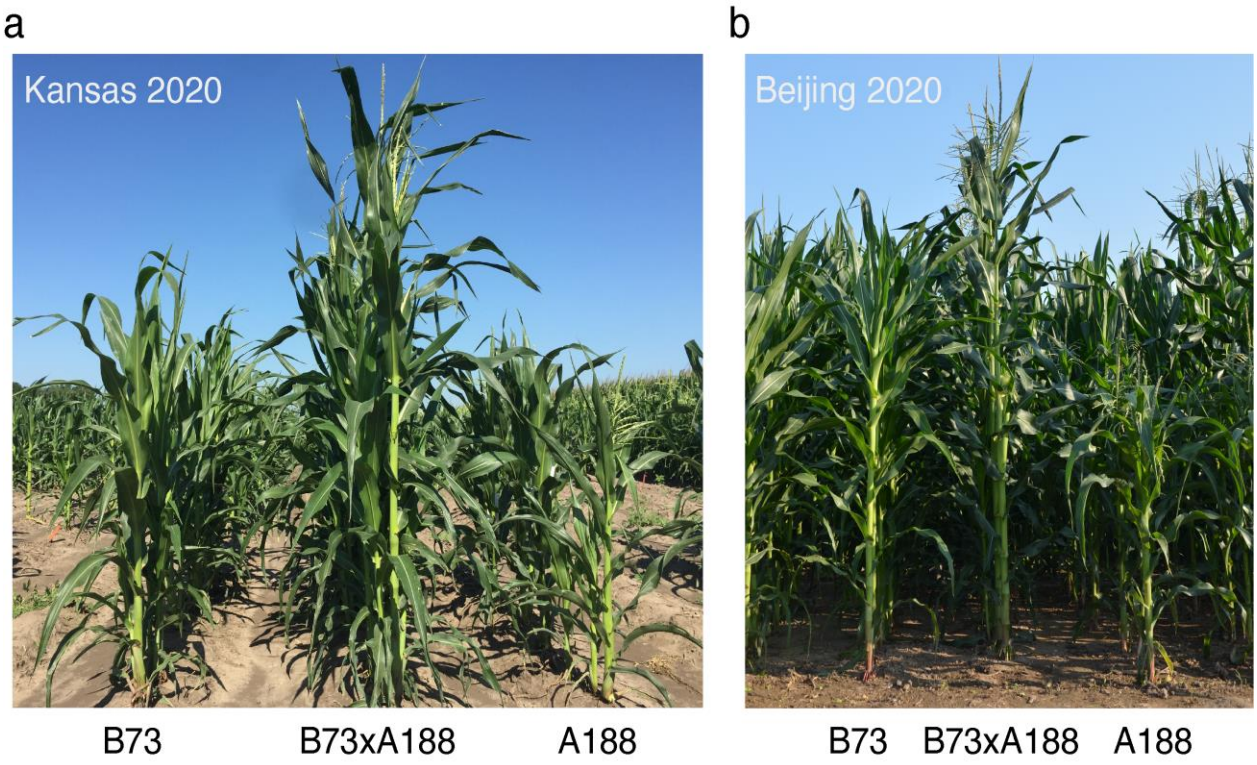

Figure S2. Phenotypes of B73, F1, and A188. a,b). Plants from 2020 summer nursery in Kansas (**a**) and Beijing (**b**).

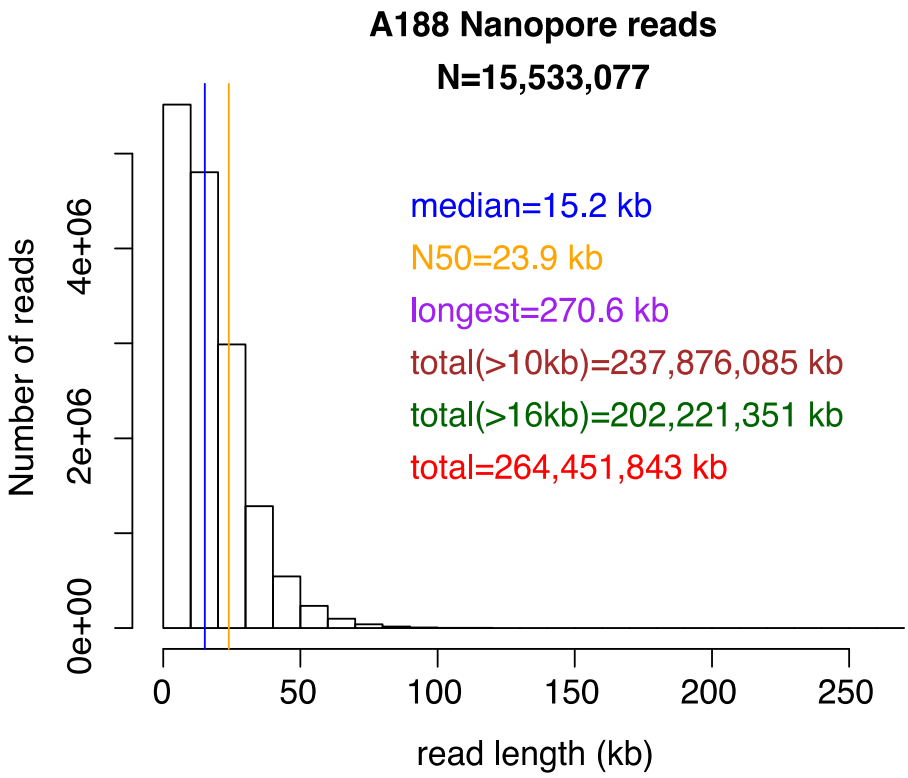

Figure S3. Histogram of lengths of Nanopore raw reads. MinION flowcells (N=31) were used to produce Nanopore reads for the A188 genome assembly, producing >264 Gb total sequences. The median and N50 of read lengths are indicated by blue and red vertical lines, respectively.

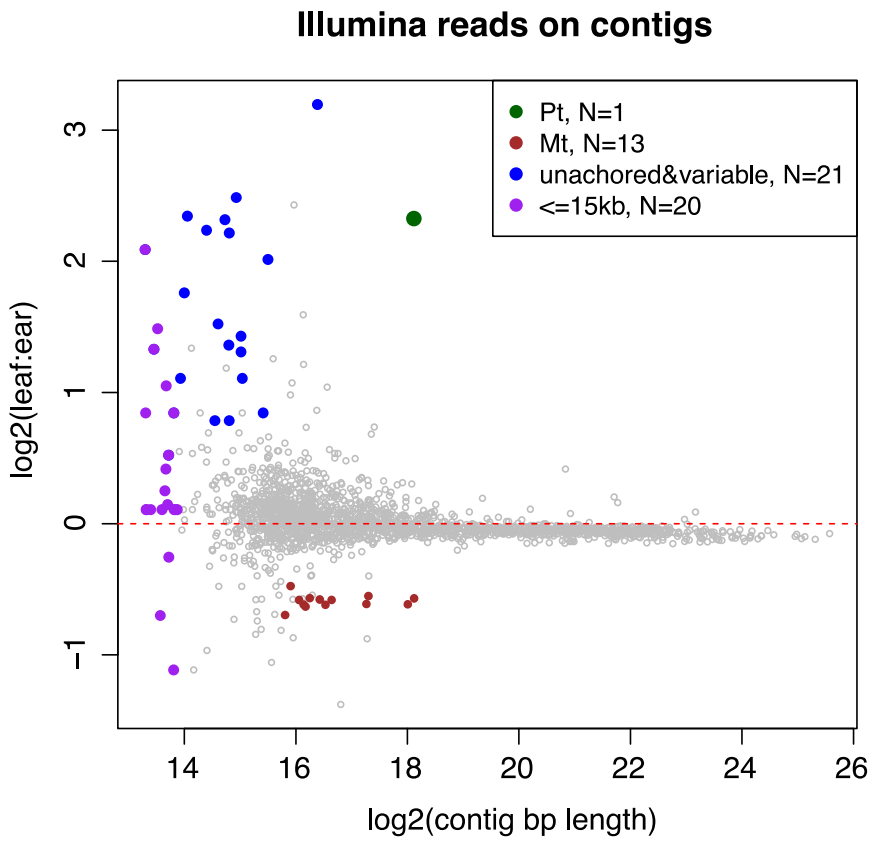

Figure S4. Contig filtering based on read depths. Log2 values of Illumina read depth ratios of seedling leaf to ear samples, log2(leaf:ear), were determined for each contig. Contigs that were not anchored to B73Ref4 and showed a high variability from 0 of log2(leaf:ear) and contigs less than 15 kb were discarded. The chloroplast contig (pt) and mitochondrial contigs (mt) were replaced by the A188 chloroplast complete sequence (Genbank accession KF241980.1) and the A188 mitochondrion complete sequence (Genbank accession DQ490952.1), respectively.

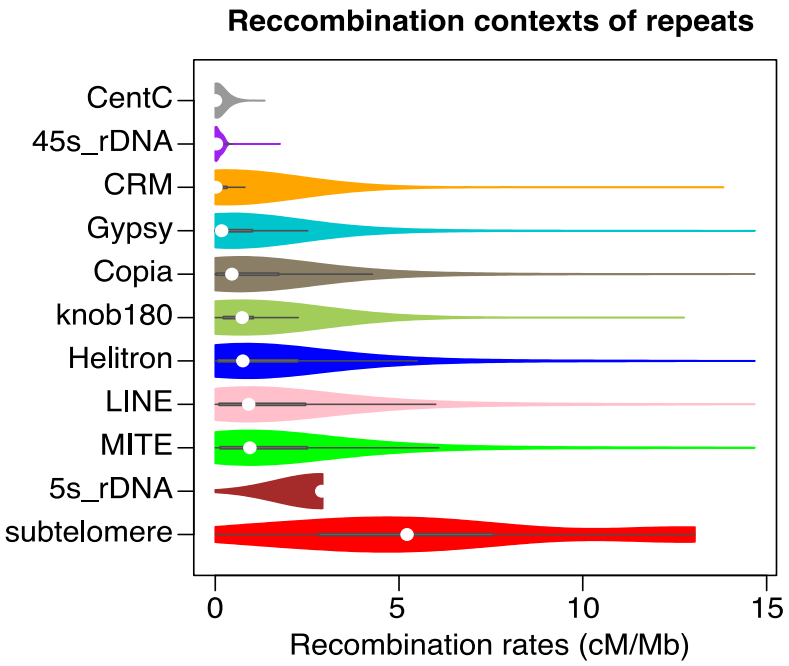

Figure S5. Recombination of contexts around repeats. For each repeat type, the recombination rate (cM/Mb) of the surrounding 1 Mb context was estimated for each sequence on chromosomes. A violin plot of all sequences of each type was plotted with a dot to represent the median.

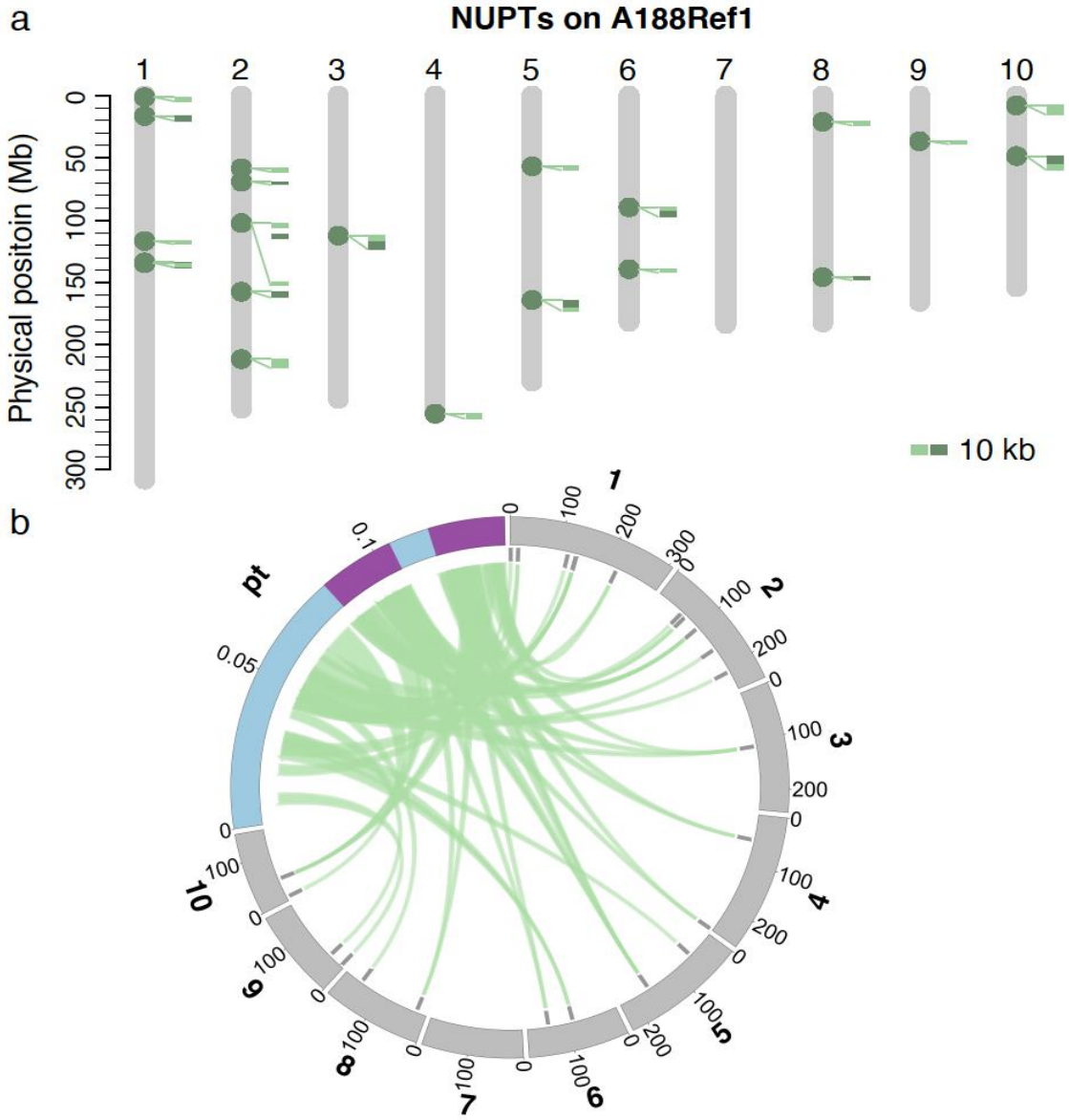

Figure S6. Nuclear integration of chloroplast DNA. a). NUPT sequence on 10 chromosomes of A188Ref1. Each dot on chromosomes designates a potential NUPT integration. Close-up alignments with the chloroplast genome are shown along NUPTs. Each alignment requires at least 3 kb match and 95% identity. **b**). Circos plot of alignments between the chloroplast (pt) genome and ten chromosomes. Purple highlights the large duplicated regions on pt. Gray bars locate NUPT positions. Note that the chromosomal scale is different from the pt scale. Numbers on the track are in Mb.

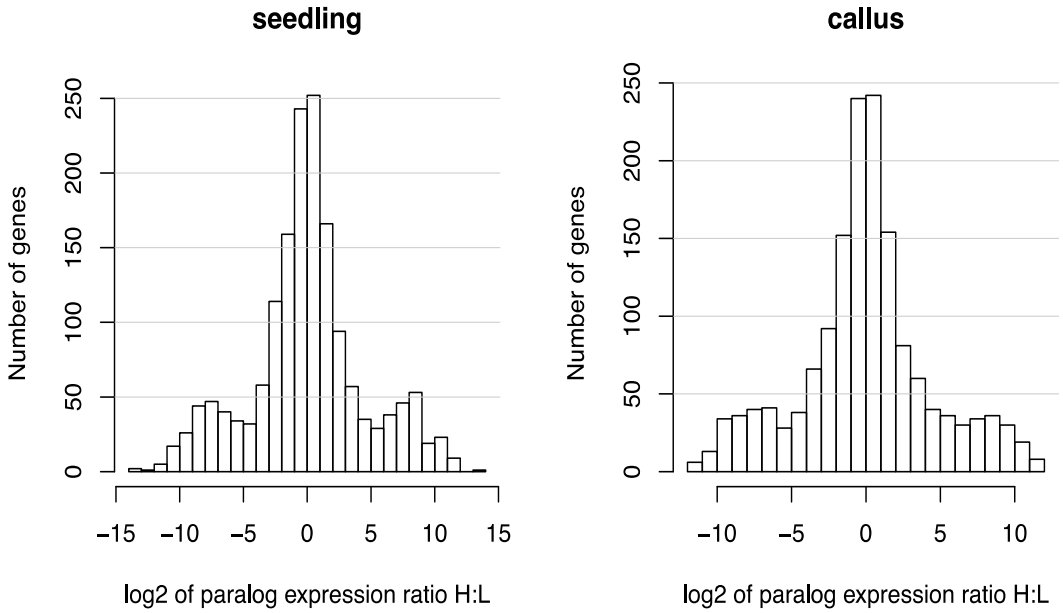

Figure S7. Expression comparison of paralogs in high- and low-recombination regions. Pairs of paralogs of which one is located at a high-recombination region (H) and the other is located at a low-recombination region (L) were compared for their expression. Histograms of the log2 values of read counts ratio of H to L were plotted. Relatively symmetric distributions between positive and negative log2 values indicated that the genomic context of the gene location was not a major driver for gene expression.

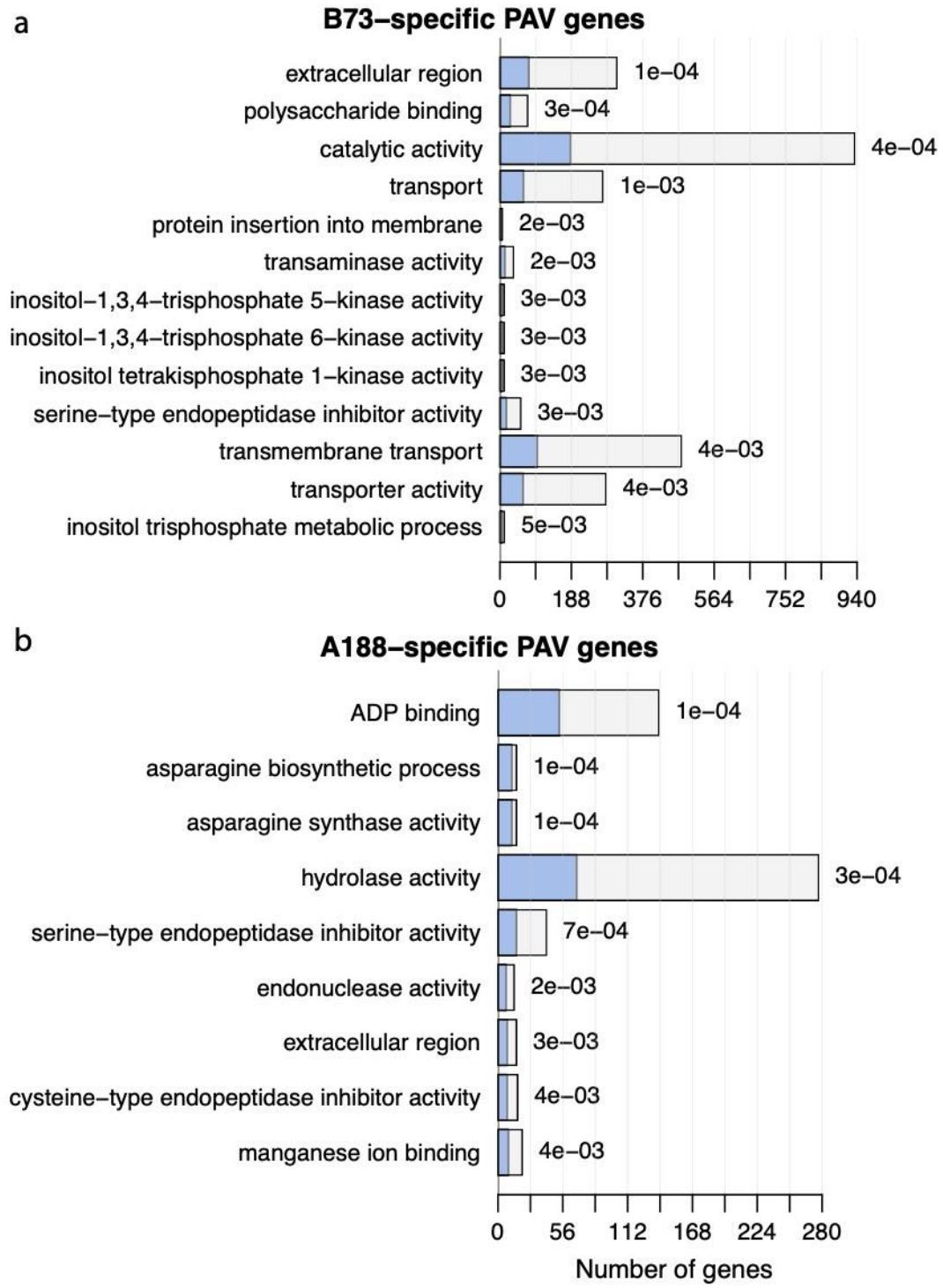

Figure S8. GO enrichments of PAV/HDS genes. Enriched GO terms in B73-specific PAV/HDS genes (**a**) and in A188-specific PAV/HDS genes (**b**). In each barplot, a blue bar stands for the number genes in the PAV/HDS gene set and the whole bar (blue and empty) stands for the total number of genes of the associated GO term. P-values are labeled on the top of each bar. Only the GO terms with the p-value smaller than 0.005 and containing at least five PAV genes were plotted.

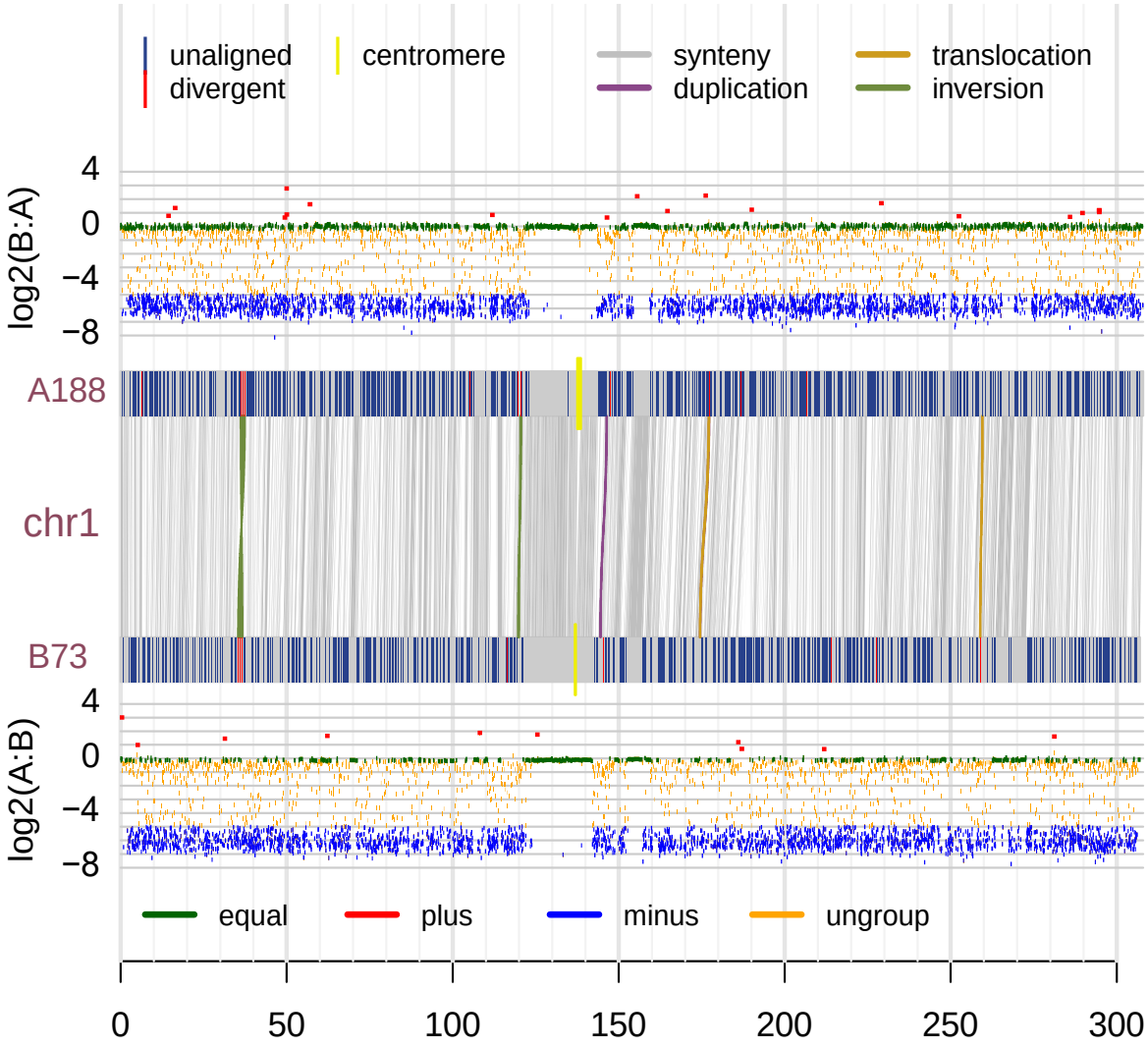

Figure S9. SyRI and CGRD results on chromosome 1. CGRD results using A188Ref1 and B73Ref4 as the reference genomes were plotted on the top and bottom, respectively. Y-axis represents log2 values of ratios of read depths of B73 to A188, log2(B:A), or log2 values of ratios of read depths of A188 to B73, log2(A:B), signifying copy number variation (CNV). The SyRI result is displayed in between two CGRD results. Alignments of syntenic blocks larger than 10 kb and alignments of other rearrangements larger than 0.5 Mb are plotted. On each A188 and B73 chromosome, segments not aligned to the other genome (unaligned), segments divergent with the other genome in a high degree (divergent), and centromeres are highlighted. The same plotting strategy was applied to chromosome 2, 3, 5, 6, 7, 8, 9, and 10.

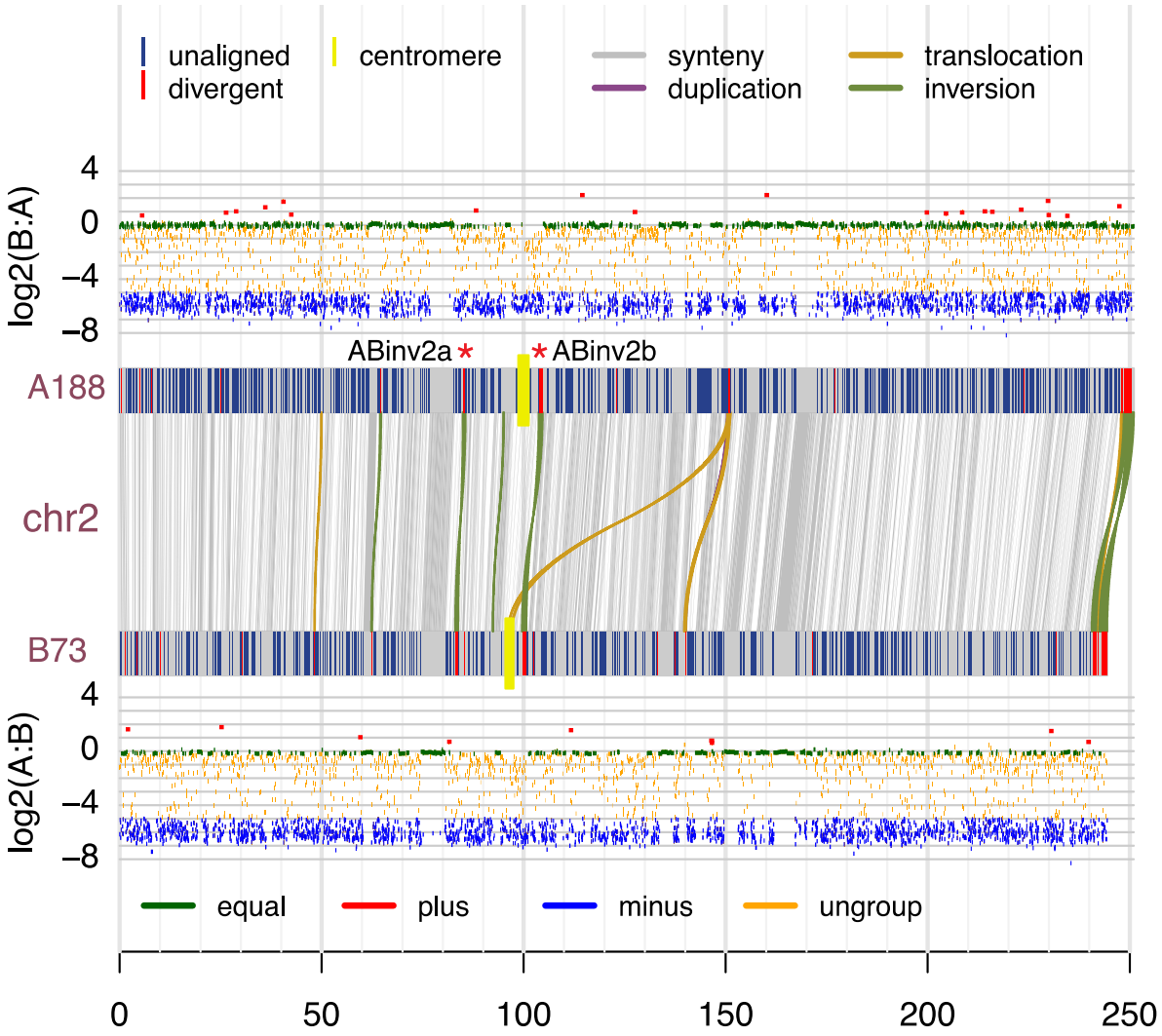

Figure S10. SyRI and CGRD results on chromosome 2. The red * labels well-evidenced inversions.

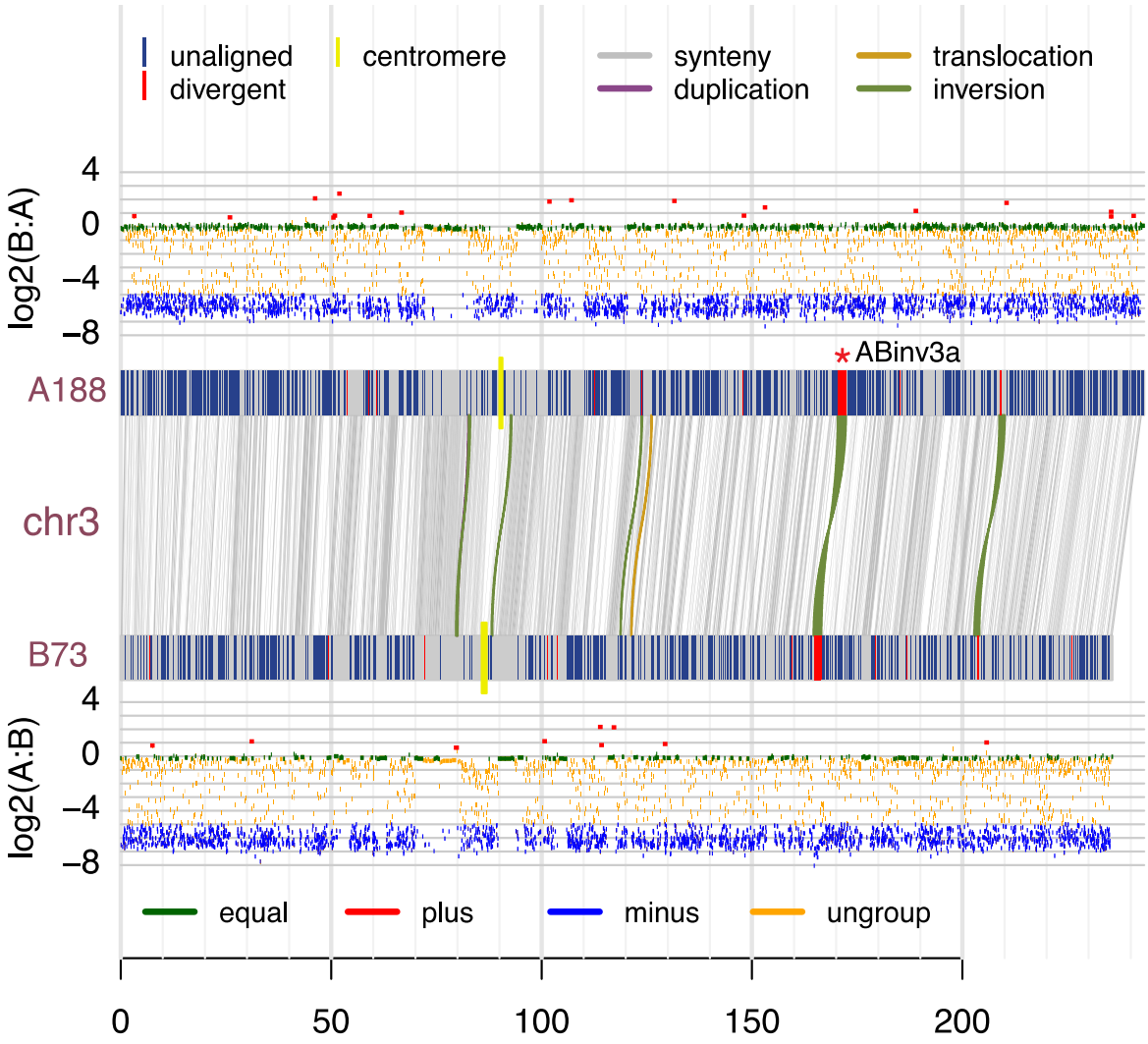

Figure S11. SyRI and CGRD results on chromosome 3. The red * labels a well-evidenced inversion.

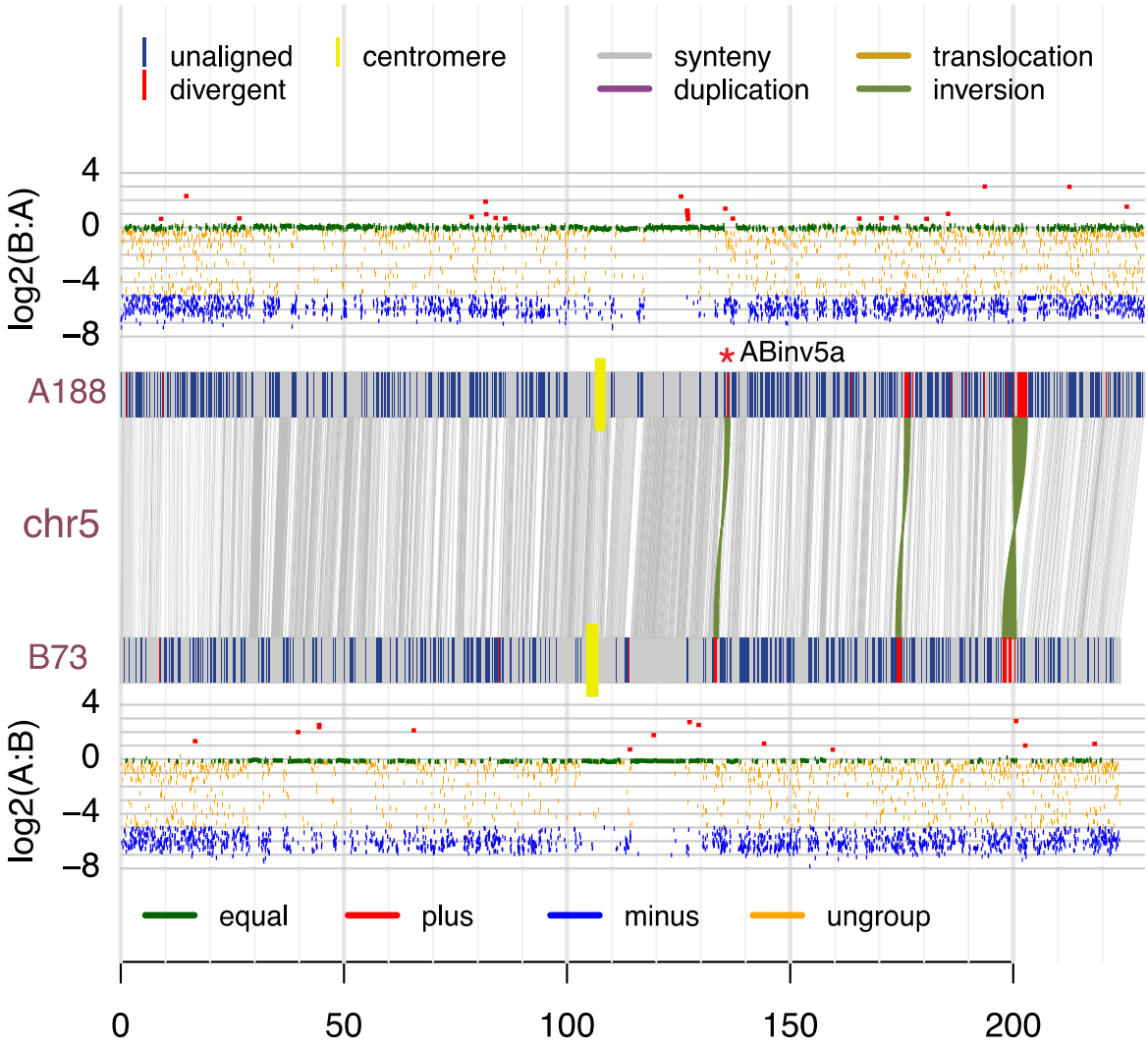

Figure S12. SyRI and CGRD results on chromosome 5. The red * labels a well-evidenced inversion.

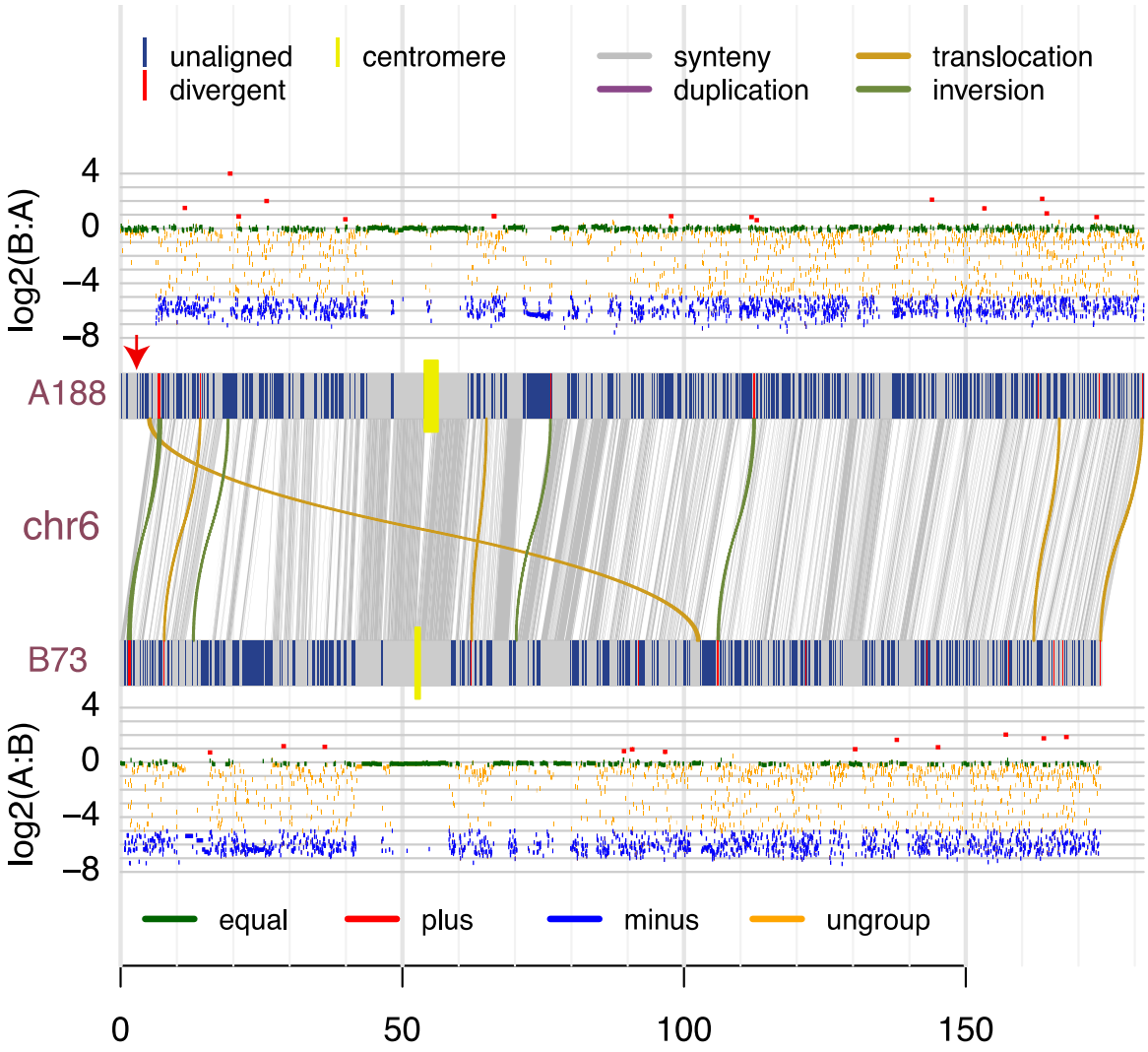

Figure S13. SyRI and CGRD results on chromosome 6. The arrow points at a relative conserved region (equal) between the two genomes. However, the region is missed in the B73Ref4. The newly assembled B73Ref5 has the region at the beginning of chromosome 6.

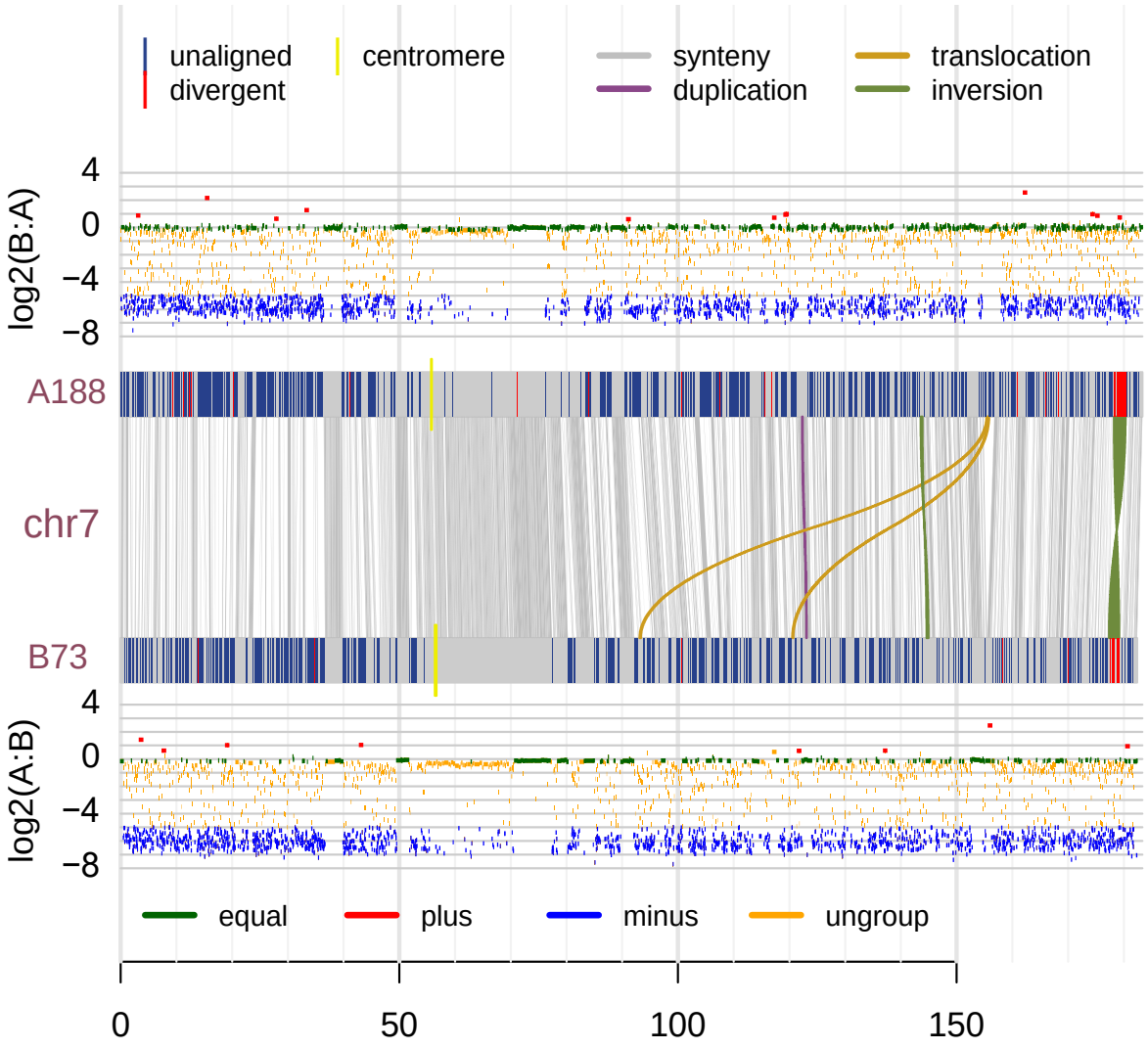

Figure S14. SyRI and CGRD results on chromosome 7.

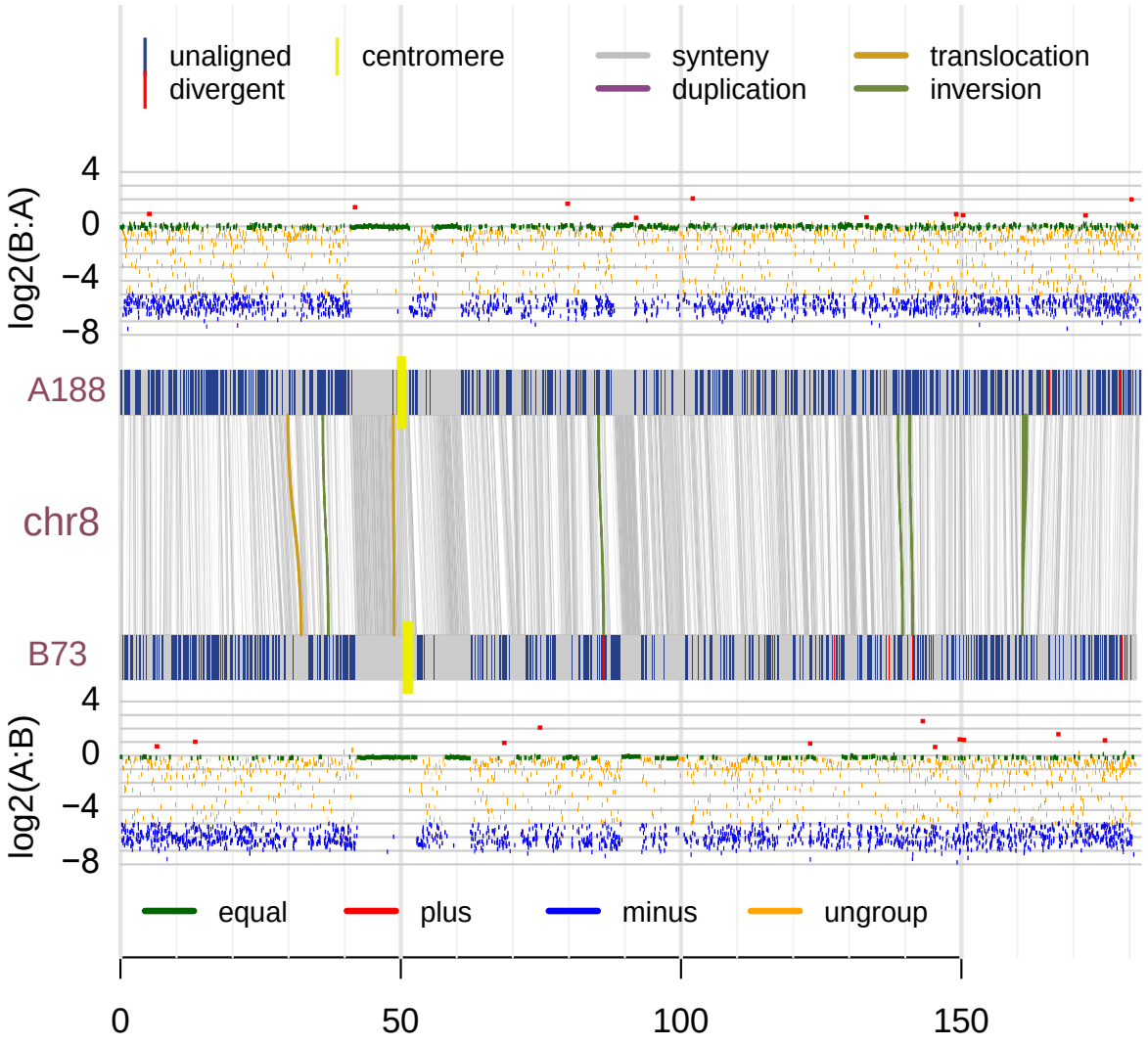

Figure S15. SyRI and CGRD results on chromosome 8.

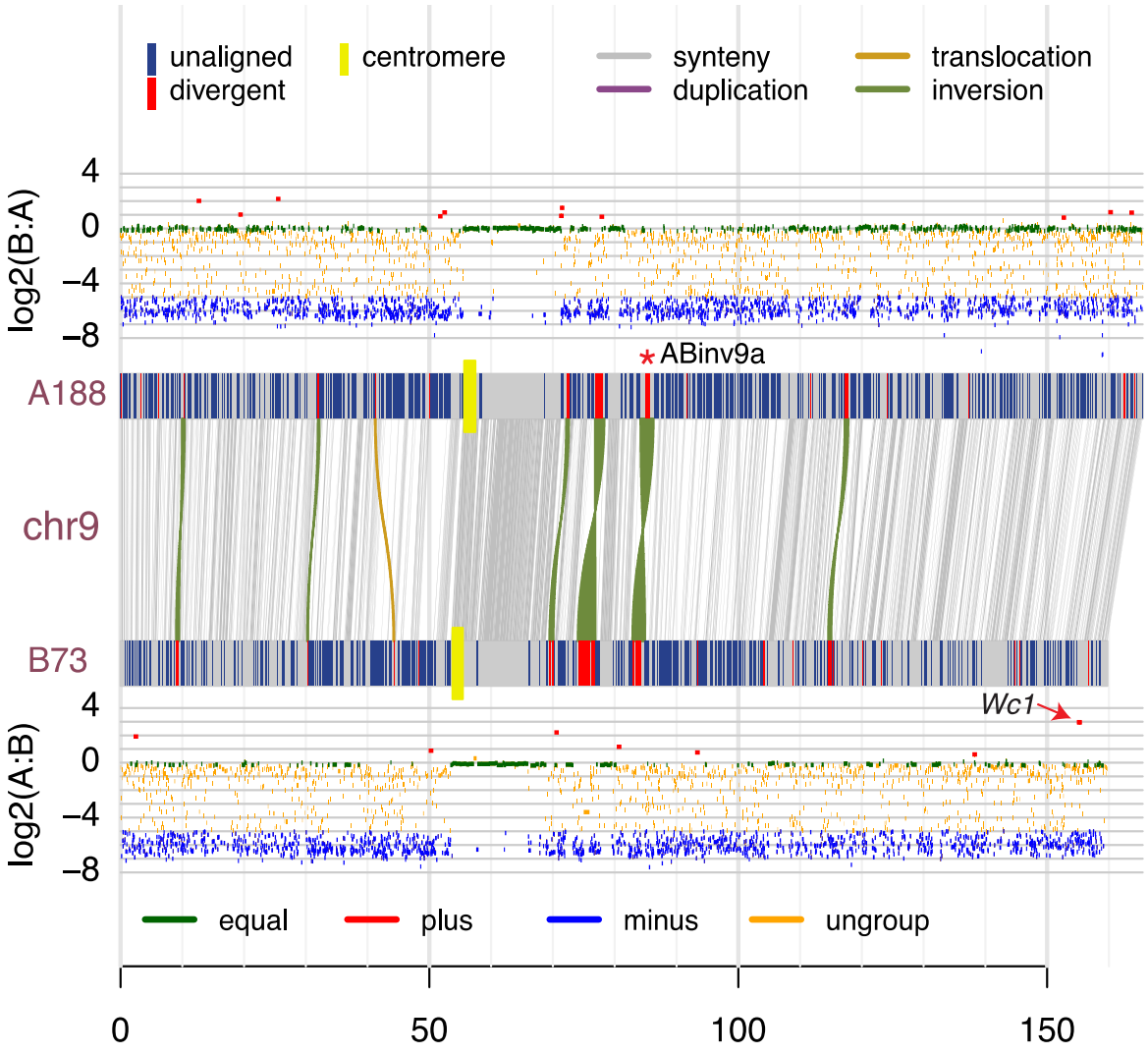

Figure S16. SyRI and CGRD results on chromosome 9. The red * labels a well-evidenced inversion. The arrow points at the B73 *Wc1* region showing A188plus, which A188 has higher copy number as compared to B73.

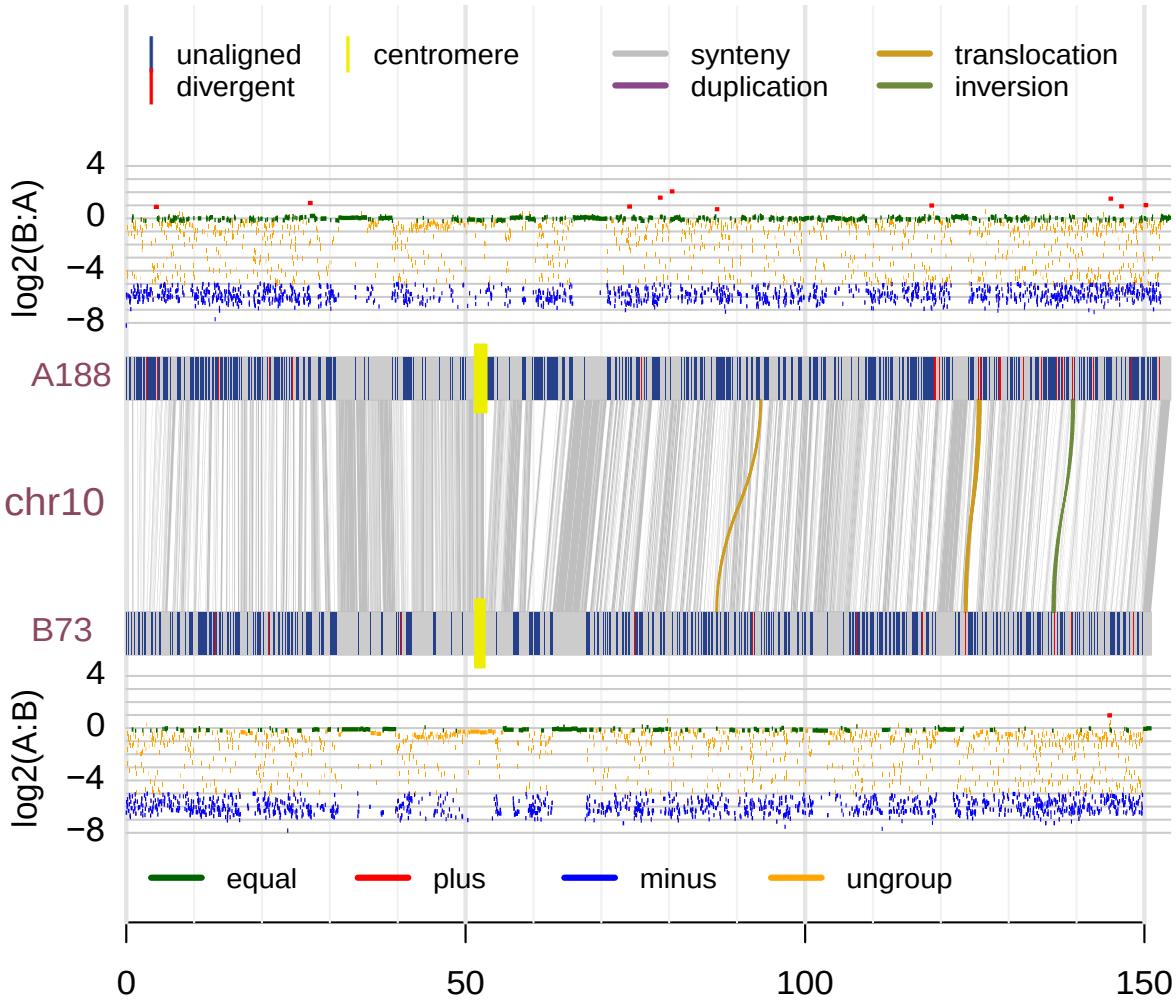

Figure S17. SyRI and CGRD results on chromosome 10.

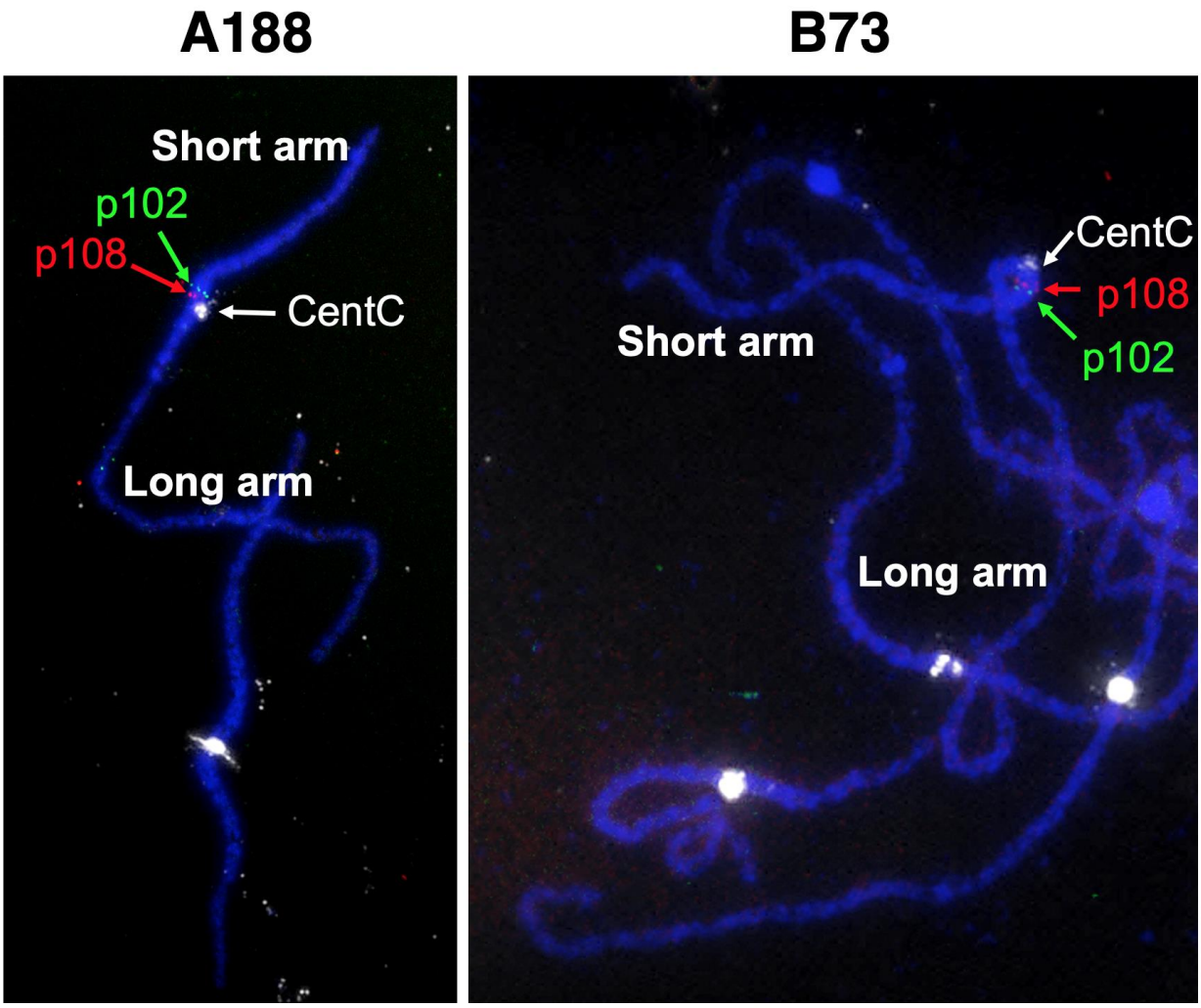

Figure S18. FISH on an inversion candidate. Two probes (p102 and p108) were designed on the potential inversion and labeled with green and red fluorescent colors. The CentC probe was used to locate the centromere. The result indicated that both A188 and B73 had the same order of p102 and p108, which did not support the inversion.

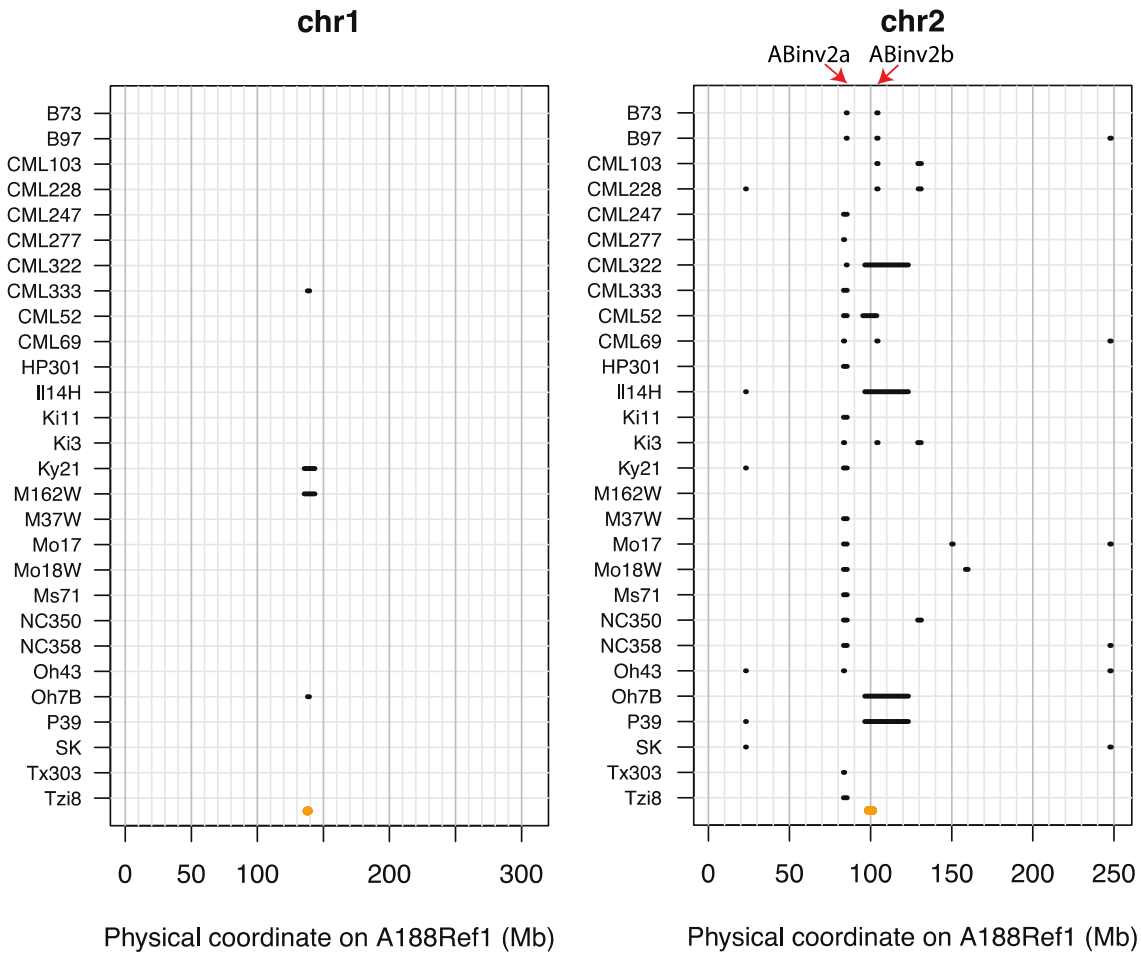

Figure S19. Large inversions on chromosomes 1 and 2. Inversions with at least 50 kb between A188Ref1 and each of genome assemblies were plotted. B73 stands for the genome assembly of B73Ref5 and other NAM founder genome assemblies are in version 1. The Mo17 genome assembly is the version of CAU1.0 and the SK genome assembly is in version 1. Orange bars represent centromere positions. Arrows point at two well-supported inversions: ABinv2a and ABinv2b.

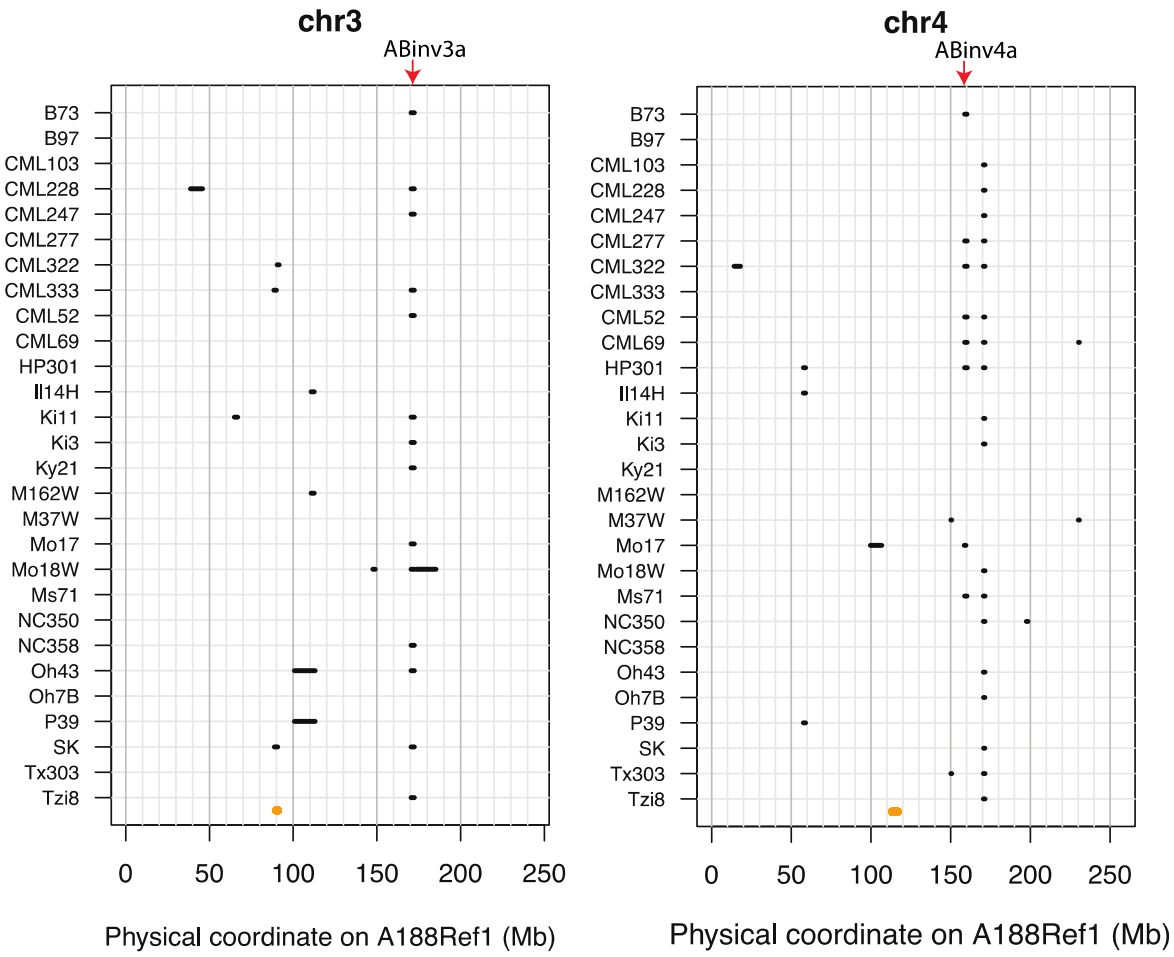

Figure S20. Large inversions on chromosomes 3 and 4. Inversions with at least 50 kb between A188Ref1 and each of genome assemblies were plotted. B73 stands for the genome assembly of B73Ref5 and other NAM founder genome assemblies are in version 1. The Mo17 genome assembly is the version of CAU1.0 and the SK genome assembly is in version 1. Orange bars represent centromere positions. Arrows point at two well-supported inversions: ABinv3a and ABinv4a.

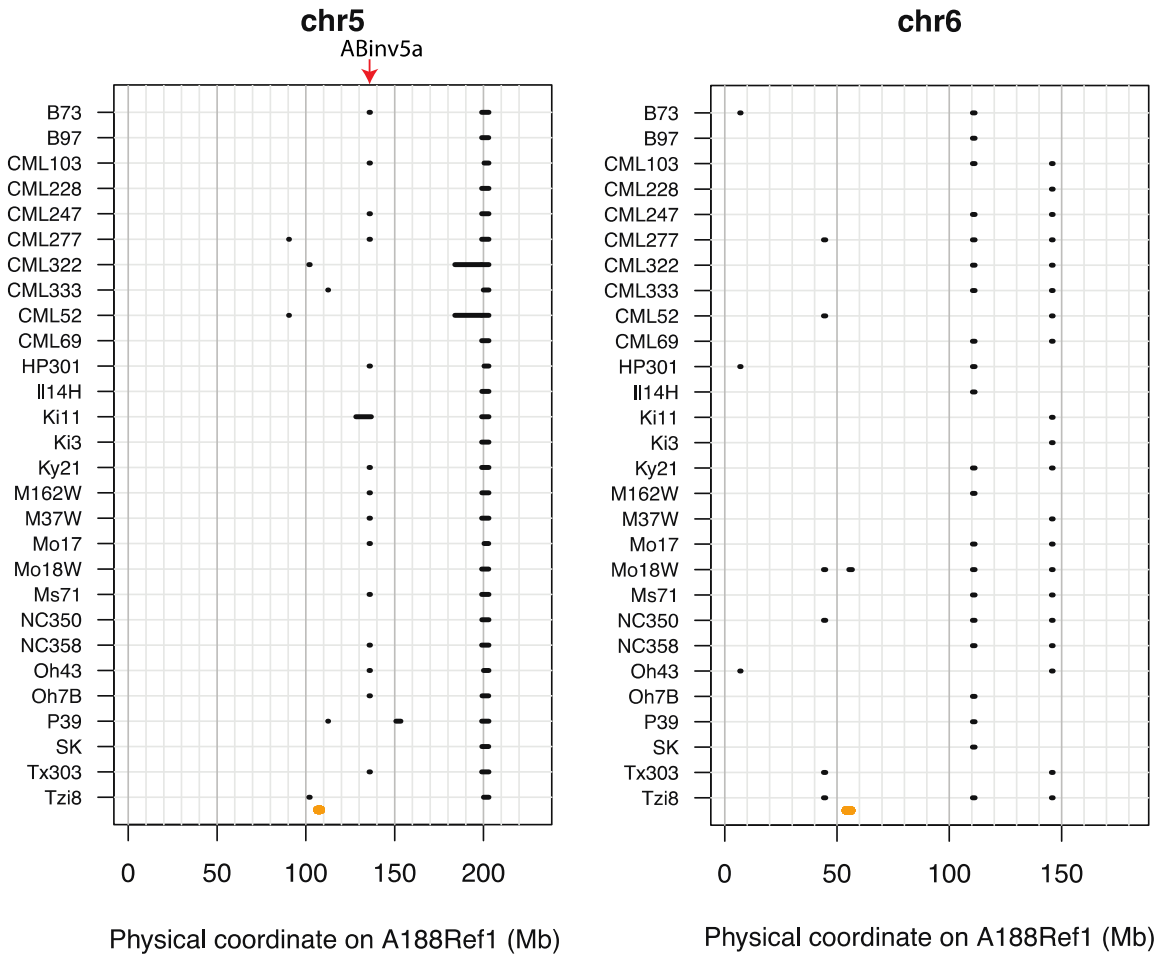

Figure S21. Large inversions on chromosomes 5 and 6. Inversions with at least 50 kb between A188Ref1 and each of genome assemblies were plotted. B73 stands for the genome assembly of B73Ref5 and other NAM founder genome assemblies are in version 1. The Mo17 genome assembly is the version of CAU1.0 and the SK genome assembly is in version 1. Orange bars represent centromere positions. The arrow points at a well-supported inversion: ABinv5a.

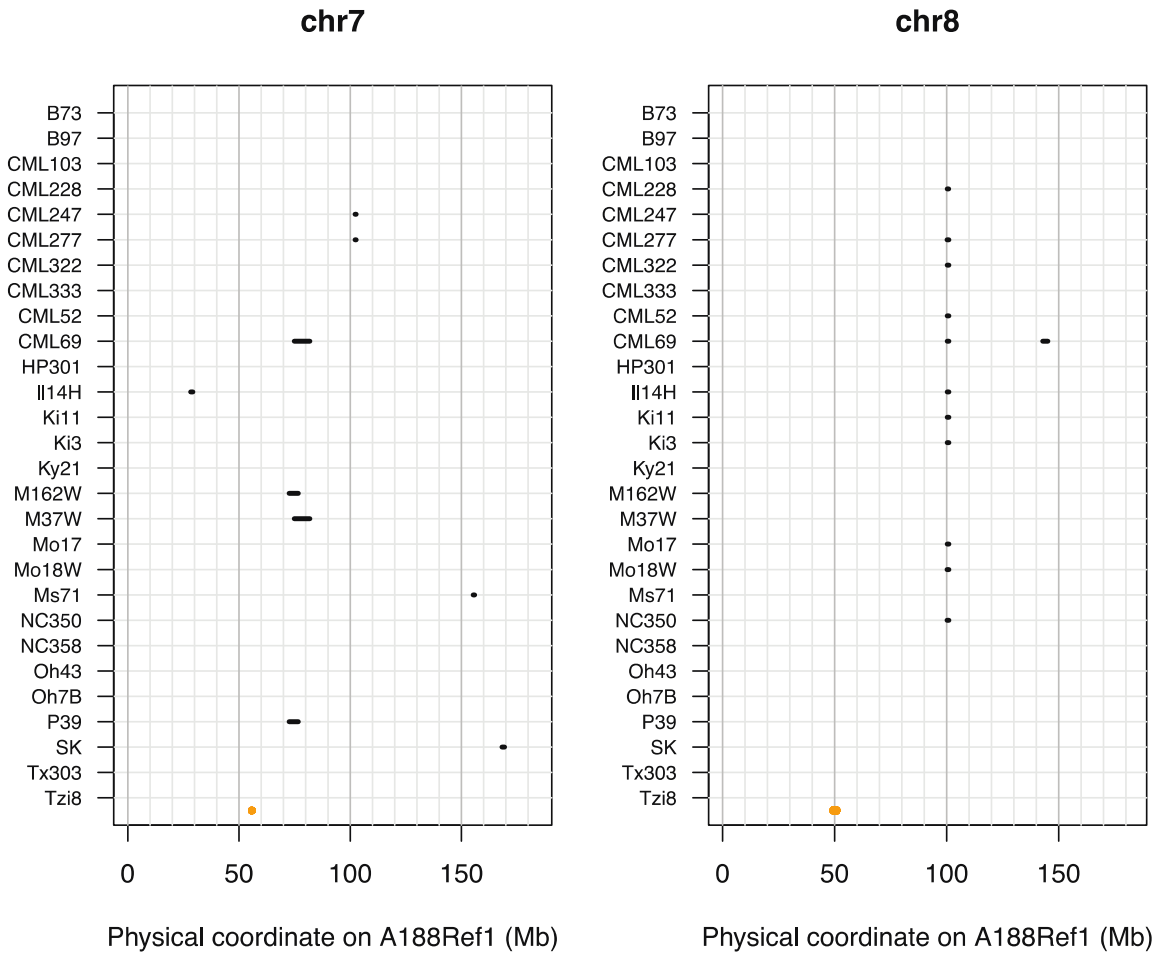

Figure S22. Large inversions on chromosomes 7 and 8. Inversions with at least 50 kb between A188Ref1 and each of genome assemblies were plotted. B73 stands for the genome assembly of B73Ref5 and other NAM founder genome assemblies are in version 1. The Mo17 genome assembly is the version of CAU1.0 and the SK genome assembly is in version 1. Orange bars represent centromere positions.

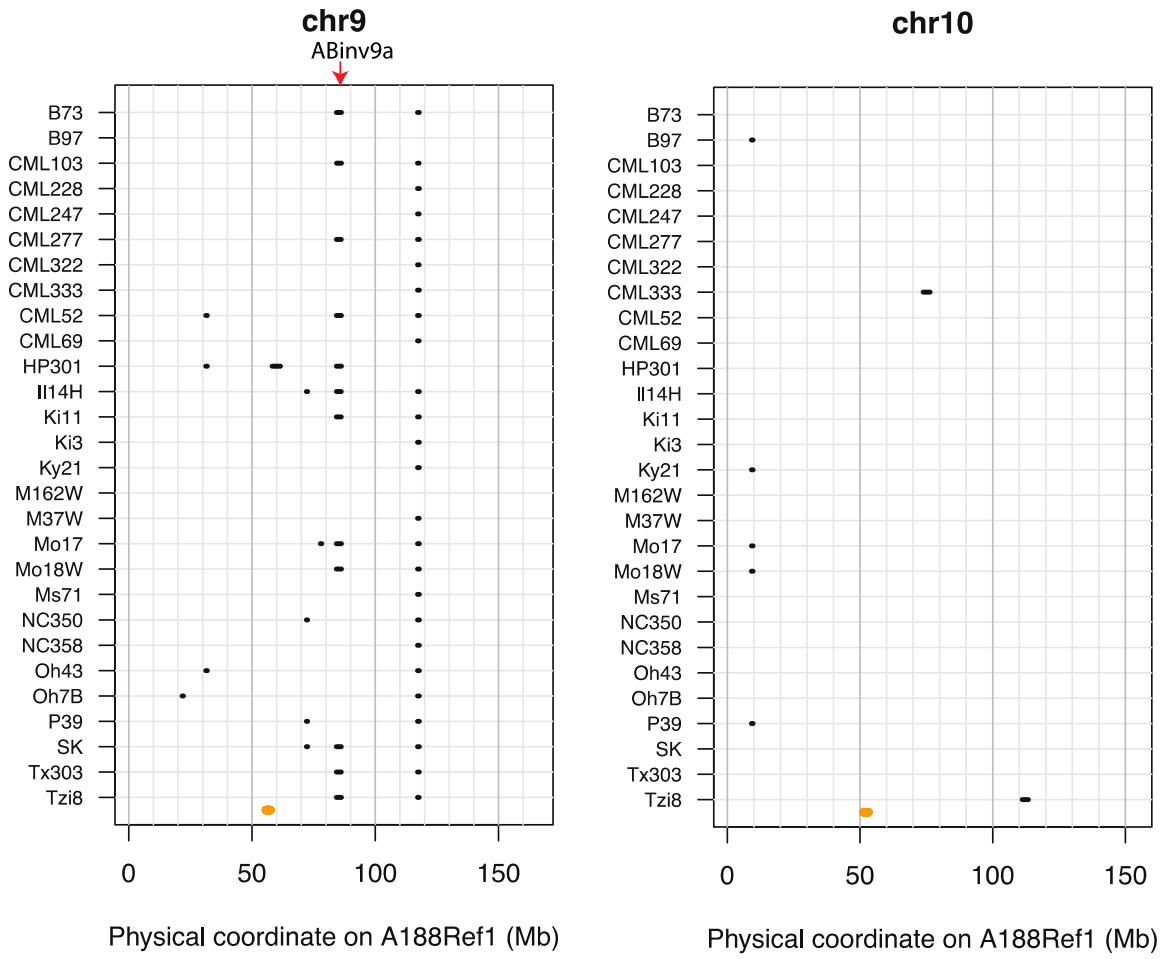

Figure S23. Large inversions on chromosomes 9 and 10. Inversions with at least 50 kb between A188Ref1 and each of genome assemblies were plotted. B73 stands for the genome assembly of B73Ref5 and other NAM founder genome assemblies are in version 1. The Mo17 genome assembly is the version of CAU1.0 and the SK genome assembly is in version 1. Orange bars represent centromere positions. The arrow points at a well-supported inversion: ABinv9a.

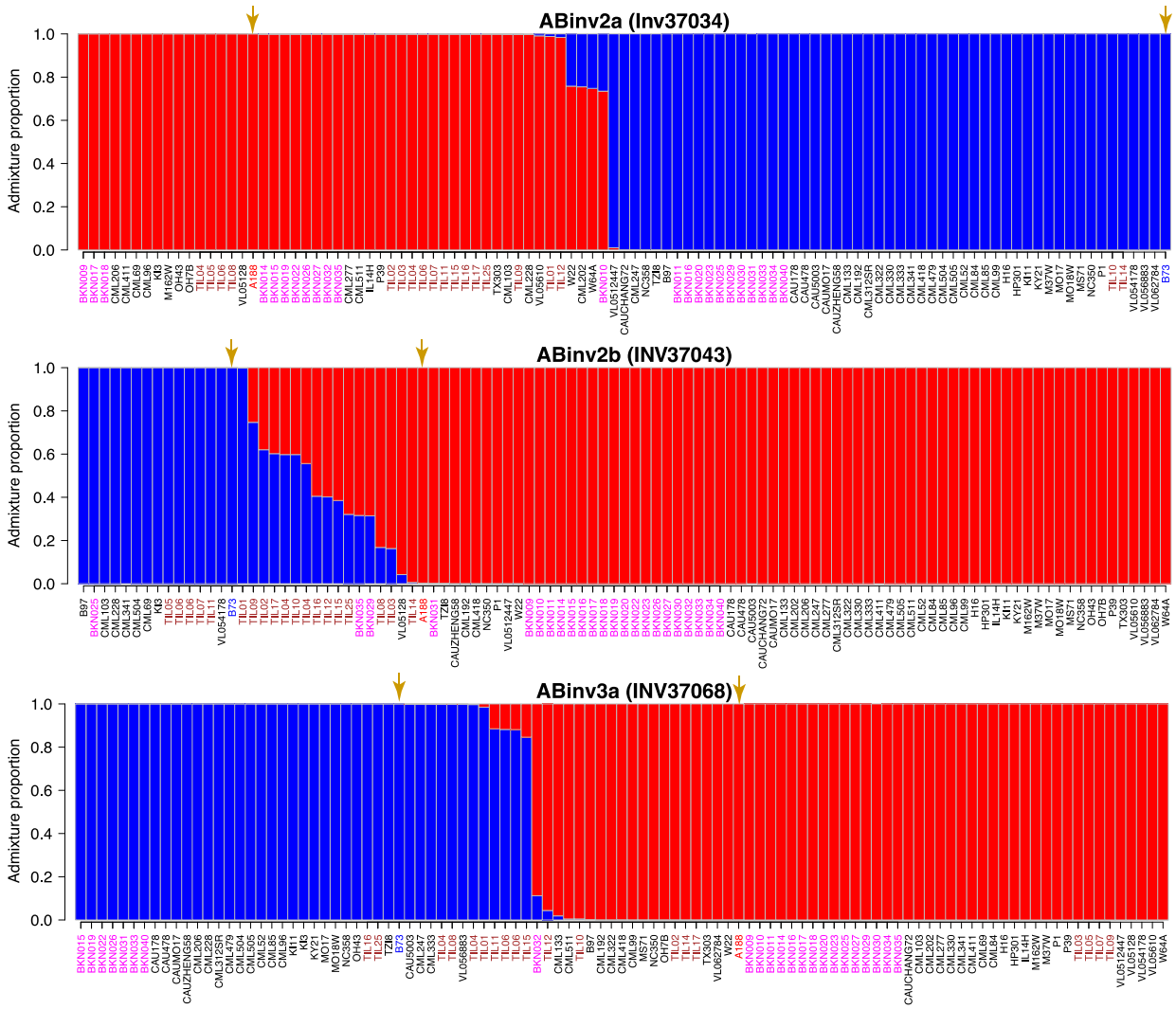

Figure S24. Structure analysis of three inversions in the maize Hapmap2 population. The x-axis represents the maize or teosinte lines. The y-axis represents the admixture proportion of two sub-populations for each line. Arrows point at A188 and B73. The maize wild ancestors, teosinte lines, are highlighted in brown, and landrace lines are highlighted in magenta.

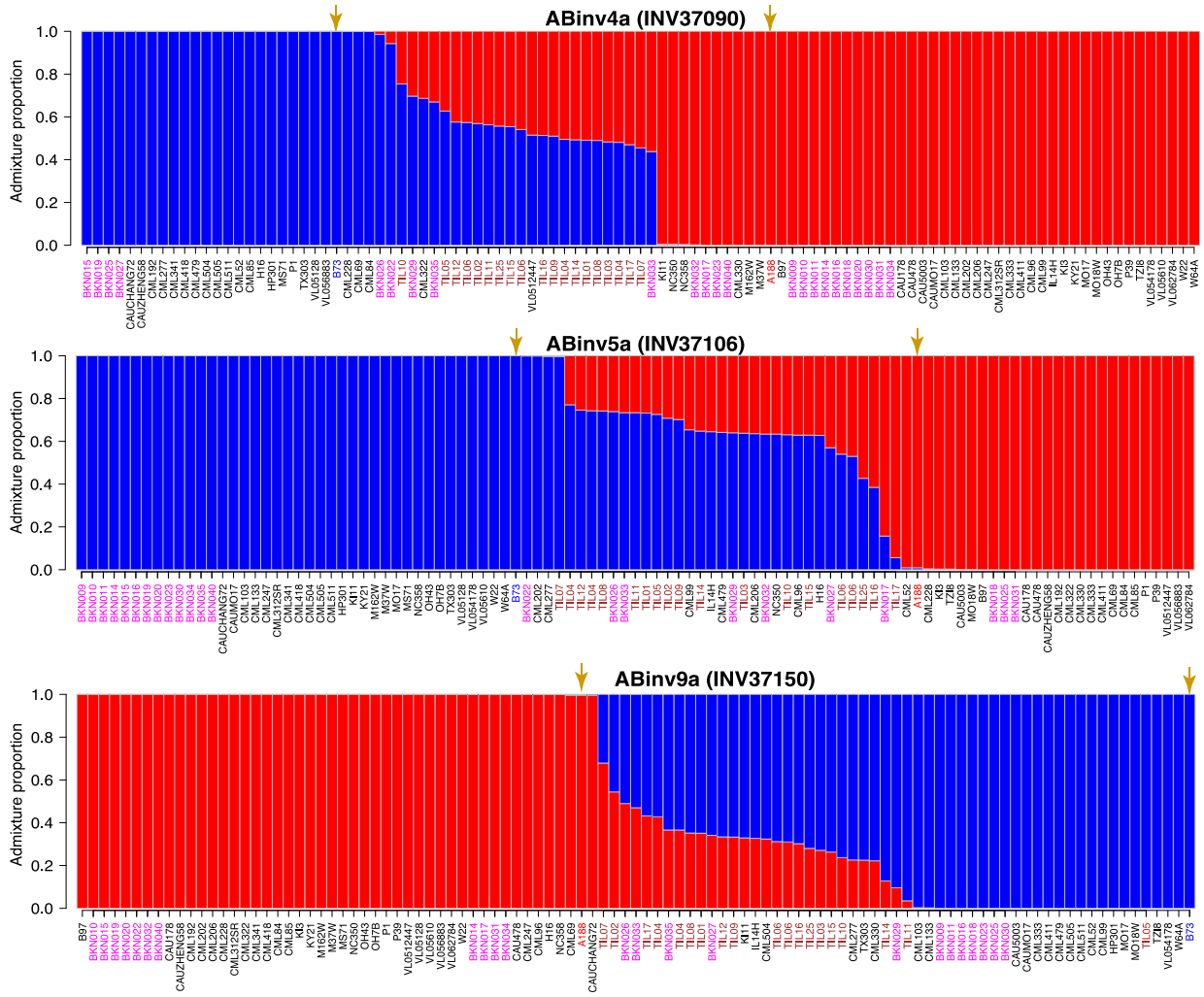

Figure S25. Structure analysis of other three inversions in the maize Hapmap2 population - II. The x-axis represents the maize or teosinte lines. The y-axis represents the admixture proportion of two sub-populations for each line. Arrows point at A188 and B73. The maize wild ancestors, teosinte lines, are highlighted in brown, and landrace lines are highlighted in magenta.

a

b

Genotype combination of DHs			DH count of color codes from 1-6					
QTL_chr2	QTL_chr6	QTL_chr9	1	2	3	4	5	6
A	A	A	13	1	0	0	0	0
B	A	A	7	6	5	0	0	0
A	A	B	3	0	4	0	0	0
B	A	B	5	1	5	0	1	1
A	B	A	5	4	4	0	1	0
B	B	A	0	3	11	6	2	1
A	B	B	0	0	1	0	4	0
B	B	B	0	2	1	1	13	8

Figure S26. Phenotypic and genotypic data of DH lines. a) A standard of kernel colors coded from 1-6. **b**) Counts of DH lines with kernel colors matching to standard codes for each genotype combination of three kernel color QTLs.

```
A188 1   MAIILVRAASPGLSAADSISHQGTLQCSTLLKTKRPAARRWMPCSLLGLHPWEAGRPSPA 60
B73  1   MAIILVRAASPGLSAADSISHQGTLQCSTLLKTKRPAARRWMPCSLLGLHPWEAGRPSPA 60

A188 61  VYSSLAVNPAGEAVVSSEQKVYDVVLKQAALLKRQLRTPVLDARPQDMDMPRNGLKEAYD 120
B73  61  VYSSLAVNPAGEAVVSSEQKVYDVVLKQAALLKRQLRTPVLDARPQDMDMPRNGLKEAYD 120

A188 121 RCGEICEEYAKTFYLGTMLMTEERRRAIWAIYVWCRRTDELVDGPNANYITPTALDRWEK 180
B73  121 RCGEICEEYAKTFYLGTMLMTEERRRAIWAIYVWCRRTDELVDGPNANYITPTALDRWEK 180

A188 181 RLEDLFTGRPYDMLDAALSDTISRFPIDIQPFRDMIEGMRSDLRKTRYNNFDELYMYCYY 240
B73  181 RLEDLFTGRPYDMLDAALSDTISRFPIDIQPFRDMIEGMRSDLRKTRYNNFDELYMYCYY 240

A188 241 VAGTVGLMSVPVMGIASESKATTESVYSAALALGIANQLTNILRDVGEDARRGRIYLPQD 300
B73  241 VAGTVGLMSVPVMGIATESKATTESVYSAALALGIANQLTNILRDVGEDARRGRIYLPQD 300

A188 301 ELAQAGLSDEDIFKGVVTNRWRNFMKRQIKRARMFFEEAERGVTELSQASRWPVWASLLL 360
B73  301 ELAQAGLSDEDIFKGVVTNRWRNFMKRQIKRARMFFEEAERGVTELSQASRWPVWASLLL 360

A188 361 YRQILDEIEANDYNNFTKRAYVGKGKKLLALPVAYGKSLLLPCSLRNGQT 410
B73  361 YRQILDEIEANDYNNFTKRAYVGKGKKLLALPVAYGKSLLLPCSLRNGQT 410
```

Figure S27. Alignment between Y1 protein sequences of A188 and B73. Protein sequences of the transcript Zm00056a032392_T003 (A188) and the transcript Zm00001d036345_T001 (B73) were compared. Of 410 amino acids, 409 were identical. The polymorphic site is highlighted in red.

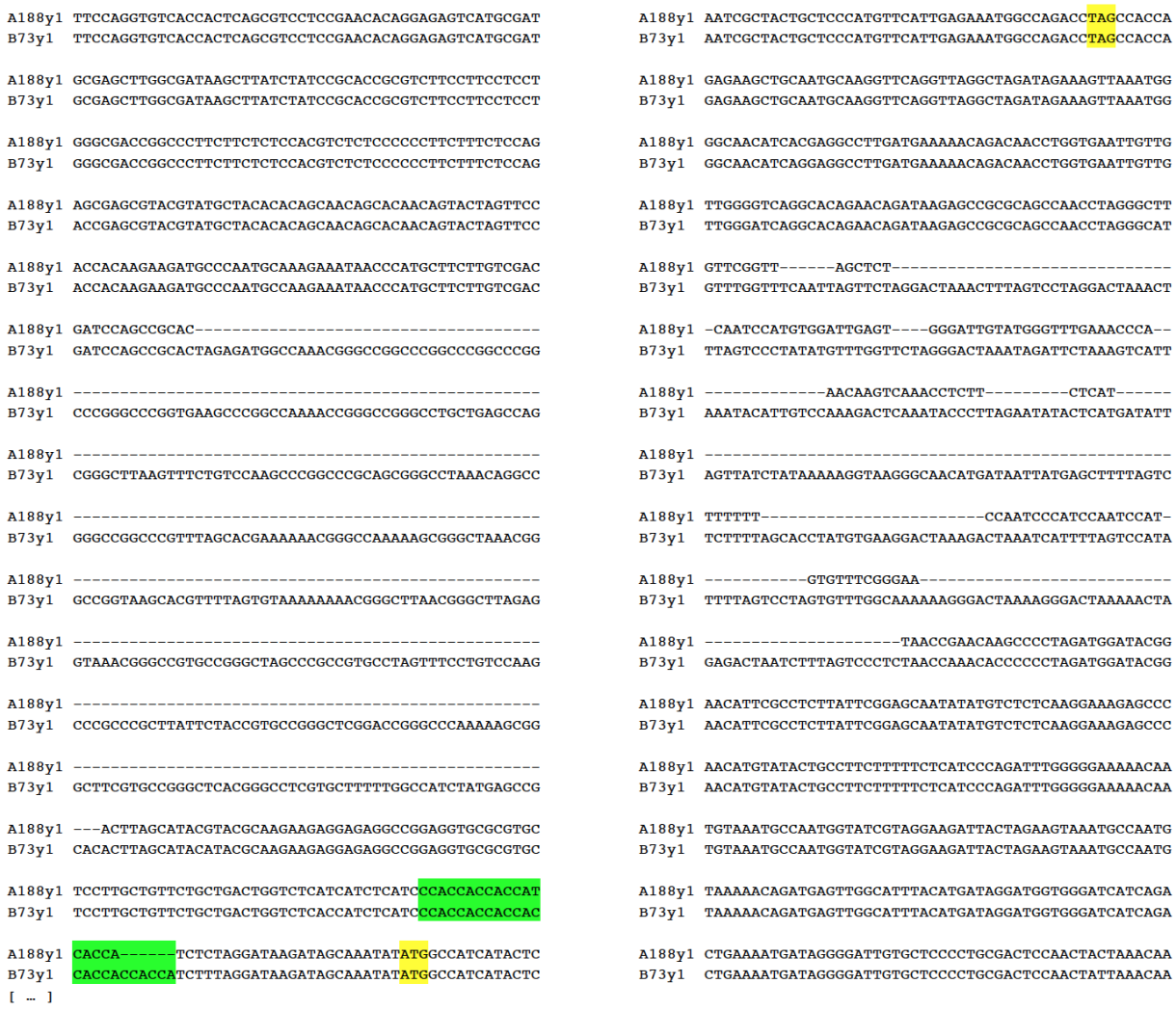

Figure S28. Alignment of 5' and 3' flanking sequences of A188 *y1* and B73 *Y1* alleles. Translation start sites and translation termination sites are highlighted in yellow. A (CCA)n microsatellite variation at the 5’ untranslated region is highlighted in green. Most gene body sequences are skipped.

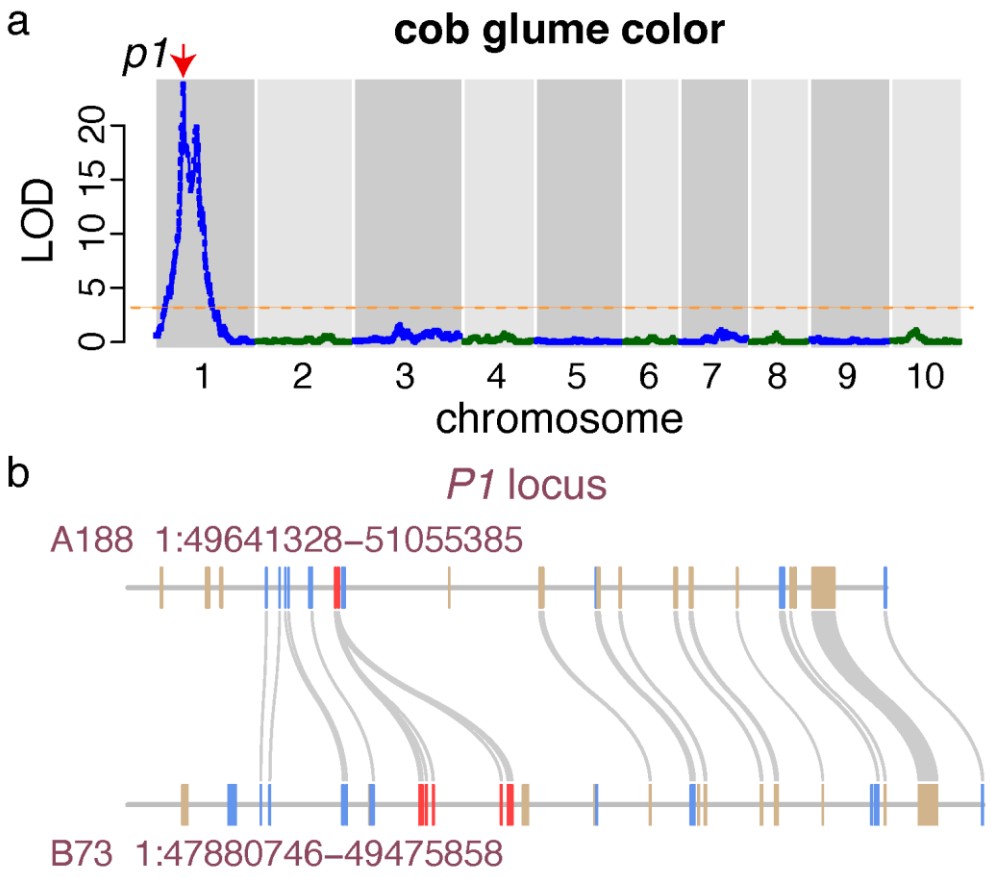

Figure S29. Genetic analysis of cob color. a) Plot of LODs from QTL analysis versus genetic positions of markers along ten chromosomes. The orange horizontal line indicates the LOD threshold from 1,000 permutation test. The arrow points at the genetic location of the *P1* gene. **b**). Each rectangle box represents a gene with blue, tan, and red colors indicating plus, minus orientation, and *P1* homologous genes.

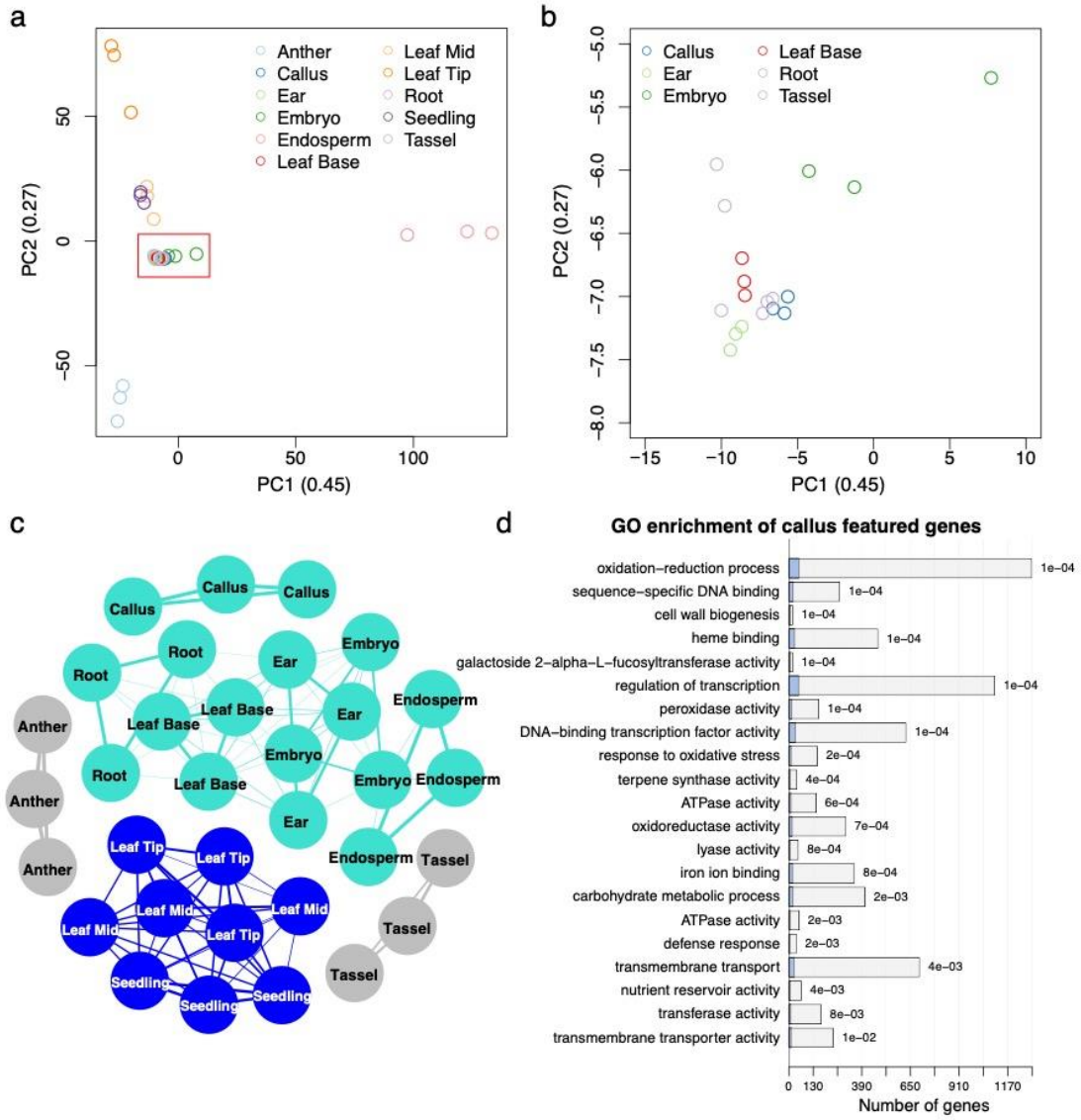

Figure S30. sample clustering and callus-featured genes a) Principal component analysis (PCA) results of gene expressions in 11 A188 tissues. The x-axis and y-axis represent the first component (PC1) and the second component (PC2), respectively. The numbers within the parentheses stand for the proportions of the variation of gene expressions explained by either PC1 or PC2. **b**) Enlarged PCA plot of the red box in **a**. **c**) The network of 33 RNA-Seq samples from 11 tissue types constructed based on their gene expression. Two major clusters were identified. One cluster (turquoise) includes callus, root, leaf base, ear, embryo, and endosperm; the other cluster (blue) includes leaf tip, leaf middle, and seedling. The leaf base, middle, and tip are three parts from base to tip from the same leaf. **d**) GO enrichment of callus featured genes. In each barplot, a blue bar stands for the number callus featured genes and the whole bar (blue and empty) stands for the total number of genes of the associated GO term. P-values are labeled on the top of each bar. Only the GO terms with the p-value smaller than 0.01 and containing at least five callus featured genes were plotted.

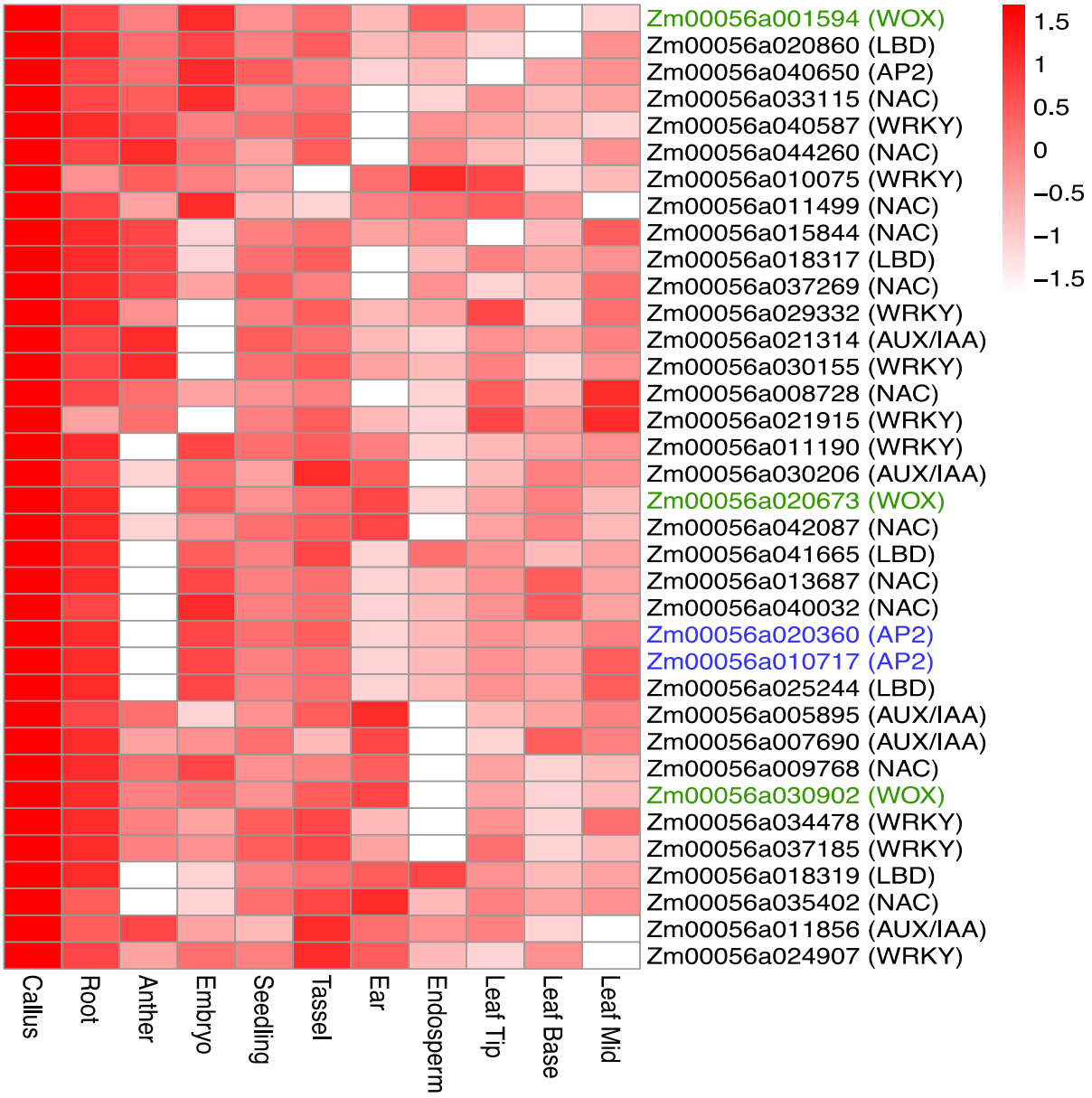

Figure S31. Heatmap plots of expression of callus-featured TF genes. The row and column represent the callus-featured TFs and tissues, respectively. The values of gene-wise quantile-quantile normalized (qqnorm) gene expressions are color-coded. The qqnorm implemented by an R package “qqnorm” that normalizes gene expressions to a Gaussian distribution. Genes homologous to *Baby boom* and *Wuschel2* are colored in blue and green, respectively.

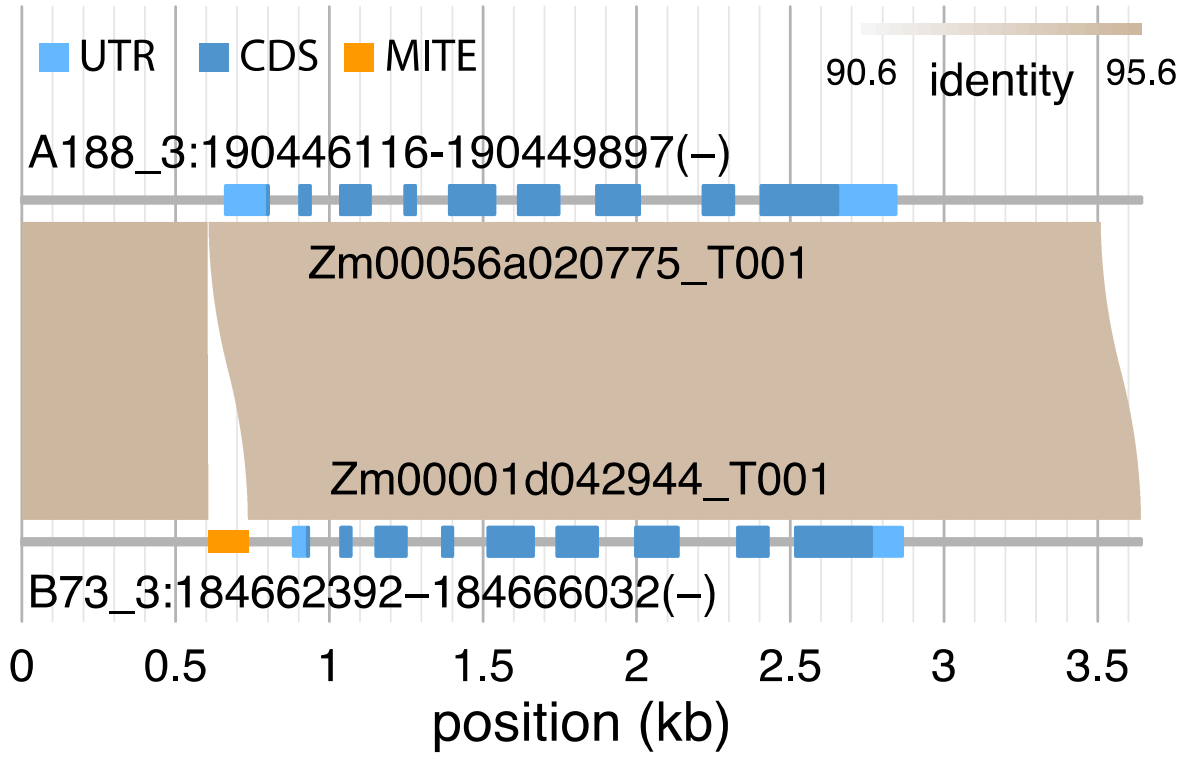

Figure S32. Alignment between Zm00056a020775 and the B73 homolog. Genomic sequences of the A188 gene Zm00056a020775, 800 bp before the gene start, and 800 bp after the gene end were aligned with the B73Ref4 and the aligned region includes the homologous B73 gene Zm00001d042944. A 120 bp MITE insertion (orange) was found at 141 bp upstream of the gene start in B73. The transcripts of both genes were plotted from 5' to 3' and both are in the minus orientation in the reference genomes, which are indicated by "(-)". The identity between aligned is color-coded. UTR: untranslated region; CDS: coding sequence.