






In the format provided by the authors and unedited.

Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement

Ning Yang ^{1,4}, Jie Liu ^{1,4}, Qiang Gao^{2,4}, Songtao Gui¹, Lu Chen¹, Linfeng Yang², Juan Huang¹, Tianquan Deng², Jingyun Luo¹, Lijuan He², Yuebin Wang¹, Pengwei Xu ², Yong Peng¹, Zhuoxing Shi², Liu Lan¹, Zhiyun Ma², Xin Yang², Qianqian Zhang², Mingzhou Bai², San Li², Wenqiang Li¹, Lei Liu^{1,3}, David Jackson ^{1,3} and Jianbing Yan ^{1*}

¹National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan, China. ²BGI Genomics, BGI Shenzhen, Shenzhen, China.

³Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA. ⁴These authors contributed equally: Ning Yang, Jie Liu, Qiang Gao.

*e-mail: yjianbing@mail.hzau.edu.cn

Supporting Online Material for

Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement

Ning Yang^{1#}, Jie Liu^{1#}, Qiang Gao^{2#}, Songtao Gui¹, Lu Chen¹, Linfeng Yang², Juan Huang¹, Tianquan Deng², Jingyun Luo¹, Lijuan He², Yuebin Wang¹, Pengwei Xu², Yong Peng¹, Zhuoxing Shi², Liu Lan¹, Zhiyun Ma², Xin Yang², Qianqian Zhang², Mingzhou Bai², San Li², Wenqiang Li¹, Lei Liu^{1,3}, David Jackson^{1,3}, Jianbing Yan^{1*}

*To whom correspondence should be addressed: yjianbing@mail.hzau.edu.cn

This PDF file includes:

Materials and Methods

Supplementary Figs. 1 to 21

Supplementary Tables 1, 3 to 5, 7 to 10, 12 to 14, 16

References (62-82)

Plant material, DNA and RNA extraction for sequencing

The teosinte (*Zea mays* ssp. *parviglumis*) shown in figure 1 was from CIMMYT and its accession number is 27479. And planted with maize in the same conditions in a tropical environment of Hainan experimental farm in 2016.

Maize inbred line SK (a landrace from Peru) was sown in 2017 in Wuhan, Hubei Province, China. Young leaves were collected and frozen at -80 °C for DNA extraction. High molecular weight DNA extraction and purification was performed using a DNeasy Plant Maxi Kit (Qiagen, Germany). 0.8g of young leaves was ground to a fine powder in liquid nitrogen using a mortar and pestle and then transferred to a 15 mL centrifuge tube. The supplied 5 mL Buffer AP1 and 10 µL RNase A were added to the tube and mixed vigorously until there were no visible tissue clumps. The tube was then incubated at 65 °C for 60 minutes and gently inverted every 10 minutes during incubation. Buffers P3, AW1, and AW2 were added, followed by centrifugation, according to the kit's protocol. DNAase-free water was used for elution. DNA concentration was measured using Nanodrop (Thermo Fischer, Schwerte, Germany) and Qubit 2.0 (Invitrogen, Karlsruhe Germany).

For the preparations of RNA samples, roots and leaves of seedlings, immature ears, immature tassels, internodes of elongation stage, pollen, silk, kernel of 15 DAP, and SAM were collected. Tissues were immediately frozen in liquid N₂. For each tissue, at least 10 plants were pooled and ~30 ears and tassels were pooled for each of three biological replicates. Total RNA was prepared by grinding tissue in TRIzol reagent (Invitrogen 15596026) on dry ice and processed following the protocol provided by the manufacturer. To remove DNA, an aliquot of total RNA was treated with RQ1 DNase (Promega M6101), followed by phenol/chloroform/isoamyl alcohol extraction, chloroform/isoamyl alcohol extraction using Phase Lock Gel Light tubes (5 PRIME 2302800) and ethanol precipitation. Precipitated RNA was stored at -20 °C.

PacBio data generation

The genomic DNA was purified by PacBio official Magnetic beads and sheared to a size range of 20–50 kb using G-tubes (Covaris®). Unpaired DNA fragments were eliminated by digestion with exonucleases (ExoVII). Third, damaged DNA fragments were repaired and hairpin adapters ligated. Fourth, the remaining damaged DNA fragments and those without adapters at both ends were eliminated by digestion with exonuclease (Exo III). Finally, the resulting SMRTbell templates were size-selected by Blue Pippin electrophoresis (Sage Sciences) and templates ranging from 20 to 50 kb were sequenced on a PacBio Sequel instrument using kit2.0 sequencing chemistry. To acquire long reads, all data were collected as 10-hour or more sequencing movies. Finally, 43 cells of PacBio data were generated, with 19,700,810 reads, with a total length of 199 Gb. 16,295,345 subreads used for assembly with a total length 196 Gb were obtained after filtering reads which were shorter than 2 kb.

Illumina data generation and filtering

Paired-end (PE) libraries with insert sizes of 410 and 670 bp, as well as mate-pair (MP) libraries with insert sizes of 2, 5, 10, and 20 kb were constructed, following a standard protocol provided by Illumina (San Diego, CA, USA). Low-quality reads meeting any of the following criteria were filtered out: (i) containing >2% ambiguous “N” bases; (ii) >30% low-quality bases (quality value less than 20); (iii) containing >30% of adapter sequence; and (iv) remove duplication and index. After that, sequenced data totaled 531.59 Gb (Supplementary Table 1), about 229.13 X coverage.

BioNano data generation.

Based on standard BioNano protocols³⁹, nicking, labelling, repair, and staining processes were performed. Specifically, DNA was digested by the single-stranded nicking endonuclease Nt.BspQI. A total of 7,782,681 BioNano molecules were obtained with total length 671.14 Gb, about 289.28 X coverage of the SK genome.

10X genomics data generation.

The Chromium Genome Reagent Kit⁴¹ (10×Genomics) was used to prepared sample indexing and partitioned barcoded libraries. Sequencing was conducted with Illumina HiSeq X ten to generate linked reads, generating 384.40 Gb of usable sequence, about 165.69 X coverage of the genome.

Genome size estimation by k-mer analyses

Genome size was estimated using the sequence reads of 410bp library (PE250), using the program Jellyfish⁶². The genome size was estimated to be 2.32 Gbp. 0.16% heterozygous rate was estimated by *k*-mer distribution of heterozygous sequences.

***De novo* assembly using PacBio sequence data and assembly result correction**

The PacBio data was assembled using the FALCON assembler¹⁸, and polished with the Arrow program (<https://github.com/PacificBiosciences/SMRT-Link>), to improve the accuracy of the assembly. FALCON implements a hierarchical assembly approach; the initial step is to error-correct long reads by aligning all reads to a subset of the longest reads. PacBio data shorter than 2 kb were filtered before correction. Finally, we used PacBio raw data with an average length of 12 kb, and total bases of 196 Gb. For the full raw data set, only reads longer than 8 kb were corrected, and generated 107 Gb of error-corrected reads bases, with an N50 size of 13 kb. To identify overlaps between raw sequences, we used “-t 20 -e 0.75 -l 2000 -k 18 -h 300 -w 8 -s 1000” for Daligner (<https://github.com/thegenemyers/DALIGNER>). Using these parameters, only overlaps longer than 2,000 bp were considered for error correction with seed matches >300 bp. In the next step, we used “-M32 -k25 -h1050 -e.96 -l3000 -s1000” for Daligner to get overlap relationships, which was used in the process of constructing the string graph. Error correction and assembly of the corrected reads was completed using ~114,000 CPU hours, which took ~3 weeks. For the polishing step, we improved PacBio’s *de novo* assembly pipeline from the SMRT Analysis package (<https://github.com/PacificBiosciences/SMRT-Link>). First, we aligned the raw reads of Pacbio to the original contigs with BLASR⁶³ using the parameters “--minMatch 12 --bestn 10 --minPctSimilarity 70.0”. Then we combined the mapping

result in one file. In the next step, we split our contigs into 200 segments, and generated consensus of the segments independently, the consensus param was “--minConfidence 40 --minCoverage 5 --algorithm best”. Finally, we combined all of the consensus results. The polishing processes took 25,000 CPU hours, a total of less than 5 days.

We used Illumina data to improve the assembly result by Pilon³⁸, an integrated tool for comprehensive variant detection and genome assembly improvement, and the parameter we used was “--changes --vcf --vcfqe --tracks --diploid --fix snps,indels --flank 50 --K 101”. These steps produced an assembly, containing 2.14 Gb in 1,316 contigs with an N50 size of 5.93Mb. The NG50 of contigs calculated based on genome size of 2.32 Gb was 5.13 Mb.

Construction of optical genome maps using the Irys system

Optical maps were assembled with BioNano IrysView⁴⁰ analysis software, only single molecules with a minimum length of 100kb and six labels per molecule were used. The p-values for the initial assembly and extension of the assembly were set to 1e-09 and 1e-10, respectively. The assembly optical maps contained 2,313.74Mb in 1,890 contigs with a N50 of 1.22Mb.

In our study, 164 conflicts were found between the optical maps and Pacbio assembly contigs using Next-Generation Mapping (NGM) Tools from BioNano Irys System (-B2 -N2). For the optical maps of conflicts, 5 conflicts were eliminated, 9 conflicts were confirmed to not exist because both the optical maps and Pacbio assembly were mapped well to the optical mapping reads. The remaining 150 conflicts in contigs were validated and the break points were found manually according to the Pacbio long reads and Illumina short reads alignments. Alignments between molecules and contigs demonstrated that there were 144 overlaps at the tail of contigs, which were merged by Minus2⁶⁴ if the overlaps between two contigs were longer than 500 bp, and the identities were over 98%. In total, we merged 139 contigs into 66 larger contigs. Furthermore, we used the paired-end information from all MP libraries to check whether the 66 larger contigs were merged accurately. First,

we aligned the MP reads to the larger contigs, and then drew the pair-ends relationships, then checked the map to verify the merge points visually. Finally, it was found that all of the 66 larger contigs could be merged without problems.

After resolving the conflicts and merging the overlaps, the Next-Generation Mapping (NGM) Hybrid Scaffold Tool from BioNano Irys System was used to scaffold the Pacbio assembly contigs and optical maps, with p-values for the conflicting alignments flagging and merging set to $1e-11$, minimum alignment length and maximum end outlier was set to 80. A new run of scaffolding was performed with a higher p-value ($1e-06$) using Next-Generation Mapping (NGM) Tools from BioNano Irys System, to insert more contigs into the hybrid assembly. A hybrid assembly was created with 2154.95Mb in 870 scaffolds with a N50 of 25.65Mb.

PacBio sequence gap filling and correction of the result

PBjelly¹⁹ was used to close gaps in the Bionano assembly with PacBio sequence (v.15.2.20, <https://sourceforge.net/projects/pb-jelly/files/>), using default parameters. The sequence was further polished with Pilon after PBjelly. The major parameters were the same as previously described⁵, but we used “--bestn 10 --hitPolicy randoibest --minReadLength 1000 --minSubreadLength 1000 --minAlnLength 500 --minPctSimilarity 70” for consensus in the polishing step. We also added the gap filling options during the Pilon process. The final assembly produced a genome length of 2.16 Gb, the scaffold N50 was 25.66 Mb and the contig N50 was 8.68 Mb, and the NG50 of scaffolds, which was calculated based on genome size of 2.32 Gb, was 22.08 Mb.

Scaffold construction using 10X genomics data

Scaffolding was performed using 10x Chromium technology linked reads based the “Assembly Roundup by Chromium Scaffolding” (ARCS) pipeline. Linked reads with barcodes that did not match the company’s barcode whitelist were filtered out. ARCS was run with sensitive parameters specified in the previous study²⁰, which utilized the BWA to align the linked reads to the draft assembly and then used the

ARCS to statistically validate the alignment information for scaffolding, LINKS was used to construct scaffolds. Default parameters were used for the BWA step. Because of the relatively high raw-read coverage (100X), it was completed using ~2,600 CPU hours. The linked reads alignment information for sequence (scaffold) output from ARCS required further conversion to a tab-separated value file listing all possible oriented sequence pairs. In this step, we use the ARCS parameter `-c 5 -r 0.05`. Given that the reference scaffolds were very long (N50=25.6M), we set the parameter `-l 60000`. These set parameters meant that there were three thresholds: (a) we just considered unique alignment pair reads that aligned to the 5' and 3' end (`-e=60,000bp`) bases of each sequence, (b) the same barcode pair reads must have sufficient numbers (`-c=5`), (c) the binomial distribution test was used to distinguish whether the distribution of the same barcode pair reads aligning to the 5' or 3' end of a sequence was substantially different from the uniform distribution (`-r=0.05`). ARCS cost ~12 CPU hours. As for as LINKS parameters, we used the recommended values^{20,65} (`-a 0.5` and `-l 5`), so sequences were joined only if the number of links connecting a sequence pair was not less than the minimum (`-l=5`), and ambiguous links were resolved when the ratio of barcode links of the second-most to top-most supported edge was equal or below a threshold (`-a=0.5`). The other parameters were default. The CPU time consumed by LINKS was less than 0.5 hour.

In order to validate the linked scaffolds, we used a consensus approach which combined evidence from three different sources: (a) Irys optical maps, (b) PacBio long reads alignments to the scaffolds, (c) Illumina HiSeq reads alignments to the scaffolds. We found that linking of 110 paired scaffolds was supported by Irys, while 62 pairs of scaffolds did not align with the Irys optical map. There were three possible reasons for the conflicts in Irys: an incorrect result from the ARCS pipeline, a big gap between the scaffolds or very short scaffolds (<200kb) which did not have enough labels to align to the Irys optical maps. All of the conflicting links were removed. We used the 599 "N" to manually fill the gaps generated by the ARCS pipeline and some small gaps were manually revised. We got assemble result with genome length of

2.16G in 708 scaffolds with a N50 is 73.28 Mb and the contig N50 is 9.17 Mb. The NG50 of scaffolds which is calculated based on genome size of 2.32 Gb is 60.43 Mb.

Anchoring of the assembled scaffolds

In order to anchor the scaffolds, a high-density genetic linkage map was developed using the RIL population with 263 recombination inbred lines derived from a SK-Zheng58 cross and genotyped with 56 K SNP array⁶⁶. The genetic map spanned 1,858.9 cM and contained 2,796 bins derived from 13,883 high-quality SNPs. The sequences of probes from the Illumina MaizeSNP50 array⁶⁶ were mapped to the 10X genomics assembly using BLAT⁴³. We used “minIdentity=95” for alignment, all of the other parameters were default. We screened the alignment results with the following conditions: mapped sequence length greater than or equal to 48bp, mismatch less than or equal to 2, and the sequence was the unique alignment. ~2.094 Gb (47 scaffolds) could be anchored to 10 chromosomes by genetic linkage mapping, which made up 96.90% of the 10X genomics assembly. To further filling gaps, we allocated the corrected Pacbio long reads to 10 chromosomes by mapping them onto the 10 pseudo chromosomes and then reassembled them respectively. We aligned the contigs resulted from reassembly onto the 10 pseudo chromosomes and filling gaps manually. BioNano map assisted gap filling. The BioNano de novo assembly and the BioNano molecules were used to estimate the gap length. Then we filled gaps using corrected Pacbio long reads with PBjelly¹⁹. Finally, the filled regions were polished with Plion³⁸. Irys optical maps and Illumina HiSeq reads were used to examine these areas over again, and got the final assemble result. The final assembly result had a genome length of 2.16 Gb in 708 scaffolds with a N50 73.24 Mb and the contig N50 of 15.78 Mb. Genotype by sequencing (GBS) probes from the high-resolution genetic mapping of the maize pan-genome⁴⁴ were also mapped to the 10x genomics assembly using blat software. We used “minIdentity=95” for alignment, all of the other parameters were default. We screened the alignment results with the following conditions: (a) mapped sequence length greater than or equal to 48bp and the mismatch less than or equal to 2, (b) the sequence was the unique alignment (c) the scaffold of the 10X genomics assembly result must have 5 or more sequences with unique alignment

support. At least 80% of these sequences belonged to a single chromosome. Finally, 151 scaffolds could be assigned to a chromosome, but they could not be ordered and positioned on the chromosome. The size of these 151 scaffolds was 26.25 MB. The SK assembly had 238 gaps, of which 48.3% (n = 115) had an estimated median gap length of 23.3 kb. The gaps introduced by 10X genomics were labeled by 599 “N” and the gaps introduced by genetic map were labeled by 999 “N”. The gaps shorter than 399bp were uniformly labeled by 399 “N” in the assembly. We also evaluated the contigs using Benchmarking Universal Single-Copy Orthologs (BUSCO)²³, the Complete BUSCOs is 96.4%.

ChIA-PET sequencing and bioinformatic analysis

SK seeds were germinated in the greenhouse (Huazhong Agricultural University). Seedlings were harvested at 16 days after planting. Seedlings were harvested by cutting just above the soil line and collecting rapidly in PBS (1x). Formaldehyde was added to 1% (v/v) to fix the seedlings at room temperature, vacuum infiltrated for 5 min, then the gas was released, and held for 25 min. Formaldehyde crosslinking was quenched by adding glycine to a final concentration of 0.2 M. Seedlings were rinsed three times with deionized water. The supernatant was discarded. Tissue was transferred into 1.5 mM EGS (Dissolve 0.06 g EGS in 600 ul DMSO at 37 °C for 5 min, and then mix EGS/DMSO with 89.4 ml 1×PBS; the EGS mixture was placed at 37 °C before use). The solution was vacuumed for 5 min, then gas was released, and held for 40 min. Seedlings were rinsed three times with deionized water. Surface water was removed from tissues using paper towels. Tissues were frozen in liquid nitrogen and kept in -80°C freezer for the ChIA-PET experiment. The Illumina sequencing libraries were constructed on beads as previously described⁶⁷. ChIA-PET data for RNAPII were processed using ChIA-PET v2 (long-read) sequence data processing. The reference genome was the maize genome (SK).

Transposable element annotation

Transposable elements in the SK genome were identified using an integration of

independent *de novo* predictions and homology searching from RepeatMasker using P-MITE⁴⁹ and Repbase databases⁵⁰ as repeat libraries.

LTR retrotransposons: LTR retrotransposons were annotated through a combination of homology searching and *de novo* prediction. The Repbase library was searched against the SK genome using RepeatMasker. For *de novo* predictions, we used RepeatModeler to build a consensus sequence library and this library was searched against the SK genome using RepeatMasker. For *de novo* predictions, we also used LTRharvest and LTRdigest to annotate the intact LTRs in the SK genome as described in Jiao et al. 2017.

TIR: TIR transposons were annotated through homology searching. We used the P-MITE database as a library, which is a plant MITE database with manual correction. This P-MITE library was searched against the SK genome using RepeatMasker.

Helitron: Helitron transposons were annotated through *de novo* prediction using HelitronScanner. We used the same pipeline as used in the B73 annotation⁵.

SINE and LINE: SINE elements were annotated through *de novo* prediction using SINE-finder⁴⁷. LINE elements were annotated through a combination of homology searching and *de novo* prediction. This Repbase library was searched against the SK genome using RepeatMasker. For *de novo* predictions, we used RepeatModeler to build a consensus sequence library and this library was also searched against the SK genome using RepeatMasker.

Gene annotation

The pipeline for gene prediction included *de novo* prediction of the repeat-masked genome and evidence-based predictions using MAKER⁵¹ and PASA⁵² (Supplementary Fig. 7). For homolog evidence, 744,030 annotated protein sequences of six species (*Arabidopsis thaliana*, *Brachypodium distachyon*, *Oryza sativa*, *Setaria italica*, *Sorghum bicolor*, *Zea mays*) were aligned to the genome using exonerate⁶⁸, and then clustered and filtered to result in the final homolog gene set. Transcript evidence included 327,904 high-quality full-length transcripts for nine tissues (male

spikelet, female spikelet, internode, seedling root, seedling leaf, mature pollen, unpollinated silks, 15DAP kernels, and vegetative meristem) from Iso-seq and 1,795,841 Trinity-assembled transcripts from the RNA-seq of the same nine tissues. The transcripts from RNA-seq and Iso-Seq were further validated by PASA⁵². For *de novo* gene prediction, we used Augustus and FGENESH (<http://www.softberry.com/berry.phtml>) trained on 2,000 homolog genes which were supported by Iso-Seq full-length transcripts and monocots transcripts, respectively. All the evidence was submitted to MAKER⁵¹ resulting in 40,936 gene models and 48,224 transcripts. The output of MAKER⁵¹ was refined again by PASA⁵² only retaining the validated transcripts. This resulted in 43,271 genes and 95,938 transcripts. Functional annotation of protein-coding genes was achieved using BLASTP (E-value 1×10^{-5}) against two integrated protein sequence databases: SwissProt (http://web.expasy.org/docs/swiss-prot_guideline.html, version: release-2017_09) and NR (database version: 20170924, 02GO version: 20171220). Protein domains were annotated by searching against the InterPro using InterProScan (version: interproscan-5.16-55.0). The Gene Ontology (GO, <http://www.geneontology.org/page/go-database>) terms for each gene were obtained from the corresponding InterPro. The pathways in which the genes might be involved were assigned by BLASTP against the KEGG database (<http://www.kegg.jp/kegg/kegg1.html>, version 84), with an E-value cut-off of 1×10^{-5} .

Assembly and annotation validation with BUSCO

Genome assembly and annotation completeness was assessed using the plants database of 1,440 plant conserved plant genes using BUSCO (embryophyta_odb9, creation date: 2016-11-01)

SV genotyping and filtering

With the known sequences and positions of 386,014 SVs, SVs were genotyped in the 521 lines of the association panel with DNA deep re-sequencing data (~20x) using BayesTyper v1.3.110. The WGS reads of each line were trimmed using Trimmomatic

and the reads k-mers of each line were counted using KMC v3.1.0⁶⁹ (with parameters ‘-k55 -cil’). A read k-mer bloom filter for each sample were built using ‘bayesTyperTools makeBloom’ module in BayesTyper v1.3.1 with default parameters. The trimmed reads of each line were mapped to B73 and SK reference genome using Bowtie2 version 2.3.4.1 (with the parameters ‘--very-fast --end-to-end’). According to the mapping results, each line of the association panel was genotyped using ‘bayesTyper genotype’ module in BayesTyper v1.3.1 with default parameters. To accurately estimate the noise parameters, the genotyping procedure included all the SVs and 1 million SNVs, and all the unplaced contigs and organelle contigs in the reference genome were treated as decoy. The genotyping outputs were merged using bcftools v1.8 with the parameters ‘--force-samples --filter-logic x --info-rules ACP:max’. After genotyping SVs in each inbred line, we merged the individual results into one hapmap file. In this hapmap file, 382,254 SV sites had been genotyped in at least one line (because the power of genotyping software BayesTyper is limited, other SV sites had failed to be genotyped by BayesTyper in any inbred lines, so we filtered them). To try our best to ensure the accuracy of pSVs, we set some strict filter conditions. 1) The SV must be genotyped in at least one line of B73, Mo17 and SK using DNA resequencing data; 2) the alleles of B73, Mo17 or SK genotyped by DNA resequencing data should be cross-validated by the results based on contig alignments; 3) the genotyped SVs using DNA resequencing data should be polymorphic across the population. 1) and 2) mean we only kept the cross-validated SVs between resequencing results and contig alignment results in at least one line of B73, Mo17 and SK. 3) means two alleles found by contig alignments should be genotyped by DNA resequencing data across our panel. After applying these filters, a SVs hapmap including 80,614 polymorphic SVs (pSVs) were obtained. However, it is hard to say all of the filtered SVs are false positives, because the identification SVs with short reads is still a challenging question and the power of BayesTyper is not 100% for all SVs. For GWAS and eQTL analysis, each site of pSVs should be genotyped in at least 50 individuals and with $MAF \geq 0.05$ (Supplementary Fig. 9).

LD analysis of pSV

We refer to the method in a previous study about TE variants⁷⁰. For each common pSV (MAF>5%), the nearest 150 upstream and 150 downstream common SNPs (MAF>0.05%) were selected. Pairwise genotype LD (r^2 values) for all complete cases were obtained for SNP-SNP and SNP-pSV variant states. r^2 values were then ordered by decreasing rank and a median SNP-SNP rank value was calculated. For each of the 300 ranked surrounding positions, the number of times the pSV rank was greater than the SNP-SNP median rank was calculated as a relative LD metric of pSV to SNP. pSV with less than 100 ranks over the SNP-SNP median were classified as low-LD insertions. pSV with ranks between 100 and 200 were classified as mid-LD, while pSV with greater than 200 ranks above their respective SNP-SNP median value were classified as variants in high LD with flanking SNPs.

Phylogenetic analysis of genes encoding both LRR and protein kinase domains in maize, rice, and Arabidopsis

Zm00001d028317 encoded both LRR and protein kinase domains. We used hmmsearch in HMMER⁷¹ to search LRR domain (PF00560.28) and protein kinase domain (PF00069.20) against the proteomes of rice, maize, and Arabidopsis. Next, we checked if the identified proteins had signal peptide with SignalP 4.1⁷² (<http://www.cbs.dtu.dk/services/SignalP/>). Thus, we identified 677 genes in total. The protein sequences were aligned with Clustal X⁷³ and the phylogenetic tree of LRR kinases was constructed using MEGA⁷⁴ software with the neighbor-joining method and 1000 times bootstraps.

Quantification of DNA methylation level in NILs

The DNAs of NIL^{SK} and NIL^{ZHENG58} were extracted from the leaves and then treated with the DNA Bisulfite Conversion Kit (TIANGEN BIOTECH Co., LTD). The purified DNAs were used as templates to run PCR reactions. PCR products were then inserted into the pMD18-T vector for sequencing. We randomly selected 30 positive clones for sequencing per sample and we performed three biological replicates for

each genotype. The methylation levels were analyzed with Kismeth⁷⁵ (<http://katahdin.mssm.edu/kismeth/revpage.pl>). Primers used were listed in Supplementary Table 16.

The detection of present variations. We took B73 v4 genome as the initial reference genome and added one another genome into reference sequences in each step. The uniquely present sequences in one genome compared with reference genomes consisted of the following steps: (1) query scaffolds were aligned to the reference genomes with NUCmer⁷⁶ (-c 90 -l 40 -maxmatch); (2) Unaligned sequences in the query genome were extracted and gap regions containing undetermined N bases were filtered out; (3) The filtered query genome sequences (> 100 bp) were aligned to the reference genomes with BLASTN⁷⁷ (-evalue 1e-5 -perc_identity 90); (4) Unaligned sequences were kept as variations uniquely present in the query genome.

References:

62. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764-770 (2011). doi: 10.1093/bioinformatics/btr011; pmid: 21217122
63. Chaisson, M.J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics*. **13**, 238 (2012). doi: 10.1186/1471-2105-13-238; pmid: 22988817
64. Sommer, D.D. *et al.* Minimus: a fast, lightweight genome assembler. *BMC Bioinformatics*. **8**, 64, (2007). doi: 10.1186/1471-2105-8-64; pmid: 17324286
65. Ren é L. *et al.* LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads. *GigaScience*. **4**, 35 (2015). doi: 10.1186/s13742-015-0076-3; pmid: 26244089
66. Ganai, M.W. *et al.* A large maize (*Zea mays* L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. *PLoS One*. **6**, e28334 (2011). doi:

10.1371/journal.pone.0028334; pmid: 22174790

67. Li, X. *et al.* Long-read ChIA-PET for base-pair-resolution mapping of haplotype-specific chromatin interactions. *Nat Protoc* 12:899-915 (2017). doi: 10.1038/nprot.2017.012; pmid: 28358394
68. Slater, G.S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*. 6:31 (2005). doi: 10.1186/1471-2105-6-31; pmid:15713233
69. Kokot, M., Długosz, M. & Deorowicz, S. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics*. **33**, 2759–2761 (2017). doi: 10.1093/bioinformatics/btx304; pmid: 28472236
70. Stuart, T. *et al.* Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. *Elife*. **5**, e20777 (2016). doi: 10.7554/eLife.20777; pmid: 27911260
71. Mistry, J., Finn, R.D., Eddy, S.R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res*. **41**, e121 (2013). doi: 10.1093/nar/gkt263; pmid: 23598997
72. Nielsen, H. Predicting secretory proteins with SignalP. *Methods Mol Biol*. 1611, 59-73 (2017). doi: 10.1007/978-1-4939-7015-5_6; pmid: 28451972
73. Larkin, M.A. *et al.*, Clustal W and Clustal X version 2.0. *Bioinformatics*. **23**, 2947-2948 (2007). doi: 10.1093/bioinformatics/btm404; pmid: 17846036
74. Tamura, K., Dudley, J., Nei, M. & Kumar, S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol*. **24**, 1596-1599 (2007). doi: 10.1093/molbev/msm092; pmid: 17488738
75. Gruntman, E. *et al.* Kismeth: analyzer of plant methylation states through bisulfite sequencing. *BMC Bioinformatics*. **9**, 371. (2008). doi: 10.1186/1471-2105-9-371; pmid: 18786255
76. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol*. **5**, R12 (2004). doi: 10.1186/gb-2004-5-2-r12; pmid: 14759262

77. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402 (1997). doi: 10.1093/nar/25.17.3389; pmid: 9254694
78. Kim, D., Langmead, B. & Salzberg, S.L. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods.* **12**, 357-360 (2015). doi: 10.1038/nmeth.3317; pmid: 25751142
79. Wu, T.D. & Watanabe, C.K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics.* **21**, 1859-1875 (2005). doi: 10.1093/bioinformatics/bti310; pmid: 15728110
80. Li, C. *et al.* The HuangZaoSi maize genome provides insights into genomic variation and improvement history of maize. *Mol Plant.* (2019). doi: 10.1016/j.molp.2019.02.009
81. Sandra1, U. *et al.* European Flint reference sequences complement the maize pan-genome. *bioRxiv* (2017). doi: 10.1101/103747
82. Hirsch, C.N. *et al.* Draft Assembly of Elite Inbred Line PH207 Provides Insights into Genomic and Transcriptome Diversity in Maize. *Plant Cell.* **28**, 2700-2714 (2016). doi: 10.1105/tpc.16.00353; pmid: 27803309

Supplementary Table 1. Summary of library construction and sequencing output.

Type	Insert size	Read length (bp)	Total cleaned data (Gp)	Cleaned data Sequence depth (fold) *	Raw data Sequence (Gp)	Sequencing platform
PE	410bp	250	133.36	57.48	184.68	HiSeq 2500
	670bp	250	84.06	36.23	181.76	HiSeq 2500
MP	2kb	150	84.80	36.55	160.16	HiSeq 4000
	5kb	150	65.11	28.06	158.97	HiSeq 4000
	10kb	150	102.28	44.09	262.88	HiSeq 4000
	20kb	150	61.98	26.72	325.62	HiSeq 4000
Total	-	-	531.59	229.13	1274.07	-

*Sequencing coverage was estimated assuming the genome size of maize as 2.32 Gb.

Supplementary Table 2. Gap estimation in the SK genome. (Excel)

Supplementary Table 3. The results of the evaluation of SK genome by BUSCOs

	Genome		Gene	
	Number	Percent (%)	Number	Percent (%)
Complete single-copy BUSCOs*	1,304	90.6	1,281	89.6
Complete duplicated BUSCOs	81	5.8	91	6.5
Fragmented BUSCOs	20	1.3	31	1.9
Missing BUSCOs	35	2.3	37	2.0

*The database was embryophyta_odb9 (Creation date: 2016-11-01, number of species: 30, number of BUSCOs: 1440)

Supplementary Table 4. The megabase size structural variations between B73 v4 and SK.

1	35509696	36893158	35263462	36993880	*		1730418
2	99884568	100856889	99350165	100975444	*		1625279
2	238663116	240275904	240606483	242457364	*		1850881
2	240317707	241805059	242603313	244413647	*		1810334
3	86812024	88305196	85193610	86988142	*		1794532
3	204679156	205876918	202066110	204443130	*		2377020
4	98978587	104753846	97027719	105001372	*		7973653
4	153113709	154844569	151132134	152973827	*		1841693
4	168311820	169101183	165073095	166920504	*		1847409
5	173995955	175168825	173054589	174861535	*		1806946
7	163126609	164560724	167270729	168382147	*		1111418
7	173068552	175145983	177046706	179918095	*		2871389
9	77505775	79638904	74003318	76969980	*		2966662
3	197104635	198822228	197023482	197186764		*	1717593
4	213998858	217188618	212216120	212383030		*	3189760
5	199291845	200807132		200541812		*	1515287
9	120197490	121844530		117598200		*	1647040
10	56714249	58295691		58852672		*	1581442
1	168276529	168593751	170378169	171785312		*	1407143
6		22903202	22499642	25500062		*	3000420
7	55736396	56270424	56678673	58835904		*	2157231
7		57431556	59637378	61487243		*	1849865

Supplementary Table 5. Transposable elements in the SK genome.

Classification	No.TE in SK/B73	SK /B73 (%)	Total SK/B73 (Mb)
Class I: Retroelement	980,375/944,784	77.72/80.34	1,788,781,507/1,715,116,367
LTR Retroelement	958,144/924,038	76.28/79.46	1,769,697,988/1,696,032,848
Copia	368,852/356,766	26.99/28.29	626,231,425/603,868,246
Gypsy	544,325/522,052	45.43/47.11	1,053,942,249/1,005,532,736
other	44,967/45,220	3.86/4.06	89,524,314/86,631,866
non-LTR retotransposon	22,231/20,746	0.82/0.88	19,083,519/188,812,335
SINE	638/608	0.009/0.009	197,562/191,608
LINE1	18,056/16,630	0.71/0.77	16,630,897/16,455,166
RTE	3,448/3,481	0.10/0.10	2,248,570/2,163,627
Class II: DNA transposon	305,371/227,661	6.71/6.76	155,584,493/144,324,064
hAT	164,643/78,332	1.33/0.78	30,946,024/16,742,504
CACTA	29,962/30,858	0.31/0.34	7,153,193/7,360,447
Tc1/Mariner	4,434/4,896	0.033/0.04	756,570/861,003
Mutator	3,325/4,216	0.025/0.032	569,358/686,607
PIF/Harbinger	89,204/95,368	0.59/0.70	13,729,603/14,966,480
Helitron	13,803/13,991	4.42/4.86	102,429,745/103,707,023

Supplementary Table 6. The summary of ISO-seq and RNA-seq data. (Excel)

Supplementary Table 7. Summary of predicted gene models of SK genome.

I*	Number	Total size (bp)	Mean length(bp)
Genes	43,271	202,451,390	4,678.69
Transcripts	95,938	592,216,966	6172.91
Exons	664,177	220,956,291	332.68
CDSs	592,524	133,285,332	1389.29
3'UTRs	115,910	53,126,433	458.34
5'UTRs	117,402	34,544,526	294.24
Introns	568,239	371,260,675	653.35

II	<i>ab initio</i> support	Protein support	Isoseq support
Genes level	38,574 (89.15%)	33,386 (77.16%)	26,383 (60.97%)
Transcript level	87,127 (90.82%)	81,048 (84.48%)	76,690 (79.94%)

	<i>ab</i>		
	RNA-seq support	Protein/Isoseq/RNA-seq	<i>initio</i>/Protein/Isoseq/RNA-seq
Gene level	27,753 (64.14%)	38,916 (89.94%)	42883 (99.10%)
Transcript level	78,586 (81.91%)	91,301 (95.17%)	95371 (99.41%)

* The numbers are based on the longest transcript.

Supplementary Table 8. Summary of the annotation of predicted protein-coding genes in the SK genome assembly.

	Number	Percentage
Total	95,938	100.00%
Nr	93,819	97.79%
Swissprot	74,919	78.09%
KEGG	71,270	74.29%
COG	38,857	40.50%
TrEMBL	93,607	97.57%
Interpro	90,289	94.11%
GO	57,318	59.74%
Annotated	94,361	98.36%
Unannotated	1,577	1.64%

Supplementary Table 9. The significant SVs associated with oil traits. at the Bonferroni-corrected threshold $-\log_{10}(P) > 8.36$ ($\alpha = 0.05$) even using 11,496,863 SNPs.

Traits	Name of SVs	$-\log_{10}(P)$	LD type with nearby SNPs
OIL	chr4_55114800_55115455_deletion_656	12.12	low
OIL	chr4_57343067_57343076_deletion_10	11.77	mid
OIL	chr4_55298680_55298679_insertion_90	11.71	mid
OIL	chr4_55308766_55309104_deletion_339	11.06	low
OIL	chr4_55257554_55257553_insertion_542	10.38	low
OIL	chr4_55116283_55116282_insertion_97	10.19	low
OIL	chr4_55128244_55128243_insertion_285	9.70	low
OIL	chr4_55093533_55093652_deletion_120	9.27	low
OIL	chr4_55147776_55148642_deletion_867	8.89	low
C18_1	chr4_55114800_55115455_deletion_656	10.43	low
C18_1	chr4_55298680_55298679_insertion_90	10.11	mid
C18_1	chr4_55257554_55257553_insertion_542	9.64	low
C18_1	chr4_57343067_57343076_deletion_10	9.56	mid
C18_1	chr4_55116283_55116282_insertion_97	9.43	low
C18_1	chr4_55128244_55128243_insertion_285	8.54	low
C18_1	chr4_55308766_55309104_deletion_339	8.40	low
C18_2	chr4_57343067_57343076_deletion_10	9.78	mid
C18_2	chr4_55114800_55115455_deletion_656	9.54	low
C18_2	chr4_55298680_55298679_insertion_90	9.33	mid
C18_2	chr4_55308766_55309104_deletion_339	9.27	low
C20_1	chr4_57343067_57343076_deletion_10	9.90	mid
C20_1	chr4_55302716_55302715_insertion_624	9.86	mid
C20_1	chr4_55114800_55115455_deletion_656	8.48	low

Supplementary Table 10. The annotation of pSVs identified between SK and B73.

Annotation of variants	Num.
intergenic_variant	19662
upstream_gene_variant	8061
downstream_gene_variant	5329
intron_variant	2843
5_prime_UTR_variant	548
3_prime_UTR_variant	898
stop_gained	278
splice_region_variant	62
frameshift_variant	171
splice_donor_variant	6
inframe_insertion	18
stop_lost	1
start_lost	1

Supplementary Table 11. The enrichment analysis of lead SV eQTL (excel file).

Supplementary Table 12. Multiple replicates confirmed the effect of *qHKW1* on HKW

Year	Location	Number of individuals (NIL ^{SK} / NIL ^{ZHENG58})	NIL ^{SK}	NIL ^{ZHENG58}	P-Value
2013	Sanya	134/139	16.15 ± 2.02	18.69 ± 2.33	6.50 × 10 ⁻¹⁹
2014	Sanya	85/49	16.24 ± 2.43	22.57 ± 3.10	1.38 × 10 ⁻²⁵
2014	Sanya	61/62	17.63 ± 2.39	20.52 ± 2.18	1.19 × 10 ⁻¹⁰
2015	Sanya	94/97	15.78 ± 1.40	17.15 ± 1.33	5.73 × 10 ⁻¹¹
2015	Sanya	203/210	14.10 ± 1.00	14.84 ± 1.10	3.18 × 10 ⁻¹²
2017	Sanya	89/82	20.40 ± 1.54	21.33 ± 1.80	3.62 × 10 ⁻⁰⁴

Supplementary Table 13. The phenotypic data for five agricultural traits between over-expression positive, CRISPR-Cas9 edited transgenic lines and wild-type.

Trait	OE (+)		Cas9 edits		WT
	Mean \pm SD	P-value	Mean \pm SD	P-value	Mean \pm SD
Plant height (cm)	139.17 \pm 5.75	0.72	136.64 \pm 6.58	0.11	139.93 \pm 6.02
Ear height (cm)	51.33 \pm 6.09	0.76	49.24 \pm 4.76	0.29	50.77 \pm 3.47
Tassel branch number	6.20 \pm 0.76	0.90	6.49 \pm 1.08	0.34	6.17 \pm 0.83
Number of leaves above the primary ear	4.92 \pm 0.47	0.80	4.73 \pm 0.51	0.18	4.89 \pm 0.31
Number of leaves below the primary ear	5.00 \pm 0.55	0.71	4.82 \pm 0.49	0.30	4.95 \pm 0.40

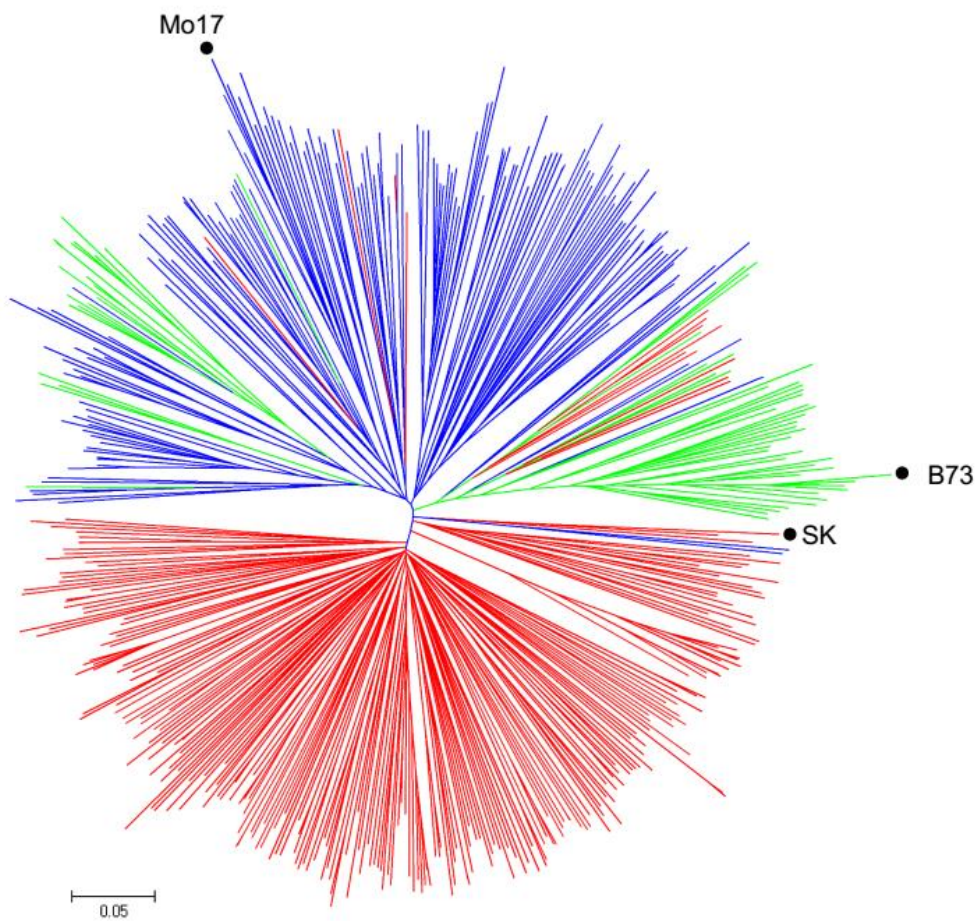
Supplementary Table 14. The phenotypic data for 13 agricultural traits between NIL^{SK} and NIL^{ZHENG58}.

Trait	NIL ^{SK}		NIL ^{ZHENG58}		P-value
	Mean \pm SD	N	Mean \pm SD	N	
Plant height (cm)	129.5 \pm 3.11	22	130.0 \pm 2.90	11	0.66
Heading date (days)	65.15 \pm 0.74	34	65.00 \pm 1.29	13	0.63
Days to anthesis	68.56 \pm 0.66	34	68.62 \pm 0.65	13	0.79
Days to silking	69.50 \pm 0.56	34	69.62 \pm 0.65	13	0.55
Tassel branch number	2.76 \pm 1.24	29	2.63 \pm 0.94	48	0.59
Leaf number above ear	4.40 \pm 0.58	45	4.54 \pm 0.51	28	0.31
Leaf number below ear	4.76 \pm 0.48	45	4.68 \pm 0.55	28	0.53
Ear length (cm)	11.13 \pm 1.00	23	10.77 \pm 1.02	28	0.22
Ear diameter (mm)	34.72 \pm 0.81	22	34.92 \pm 1.21	26	0.51
Ear weight (g)	58.44 \pm 8.88	22	55.54 \pm 5.79	25	0.19
Cob weight (g)	8.94 \pm 1.02	19	8.44 \pm 0.84	25	0.08
Kernel length (mm)	9.28 \pm 0.26	19	9.36 \pm 0.29	22	0.34
Kernel width (mm)	6.71 \pm 0.19	18	6.89 \pm 0.17	25	0.001

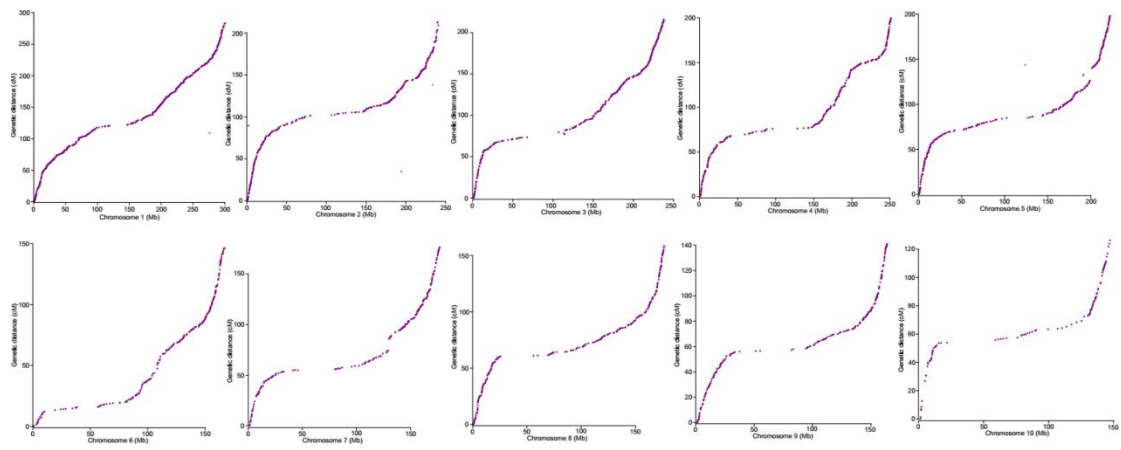
Supplementary Table 15. SV validation Sanger sequencing (excel file).

Supplementary Table 16. Primers used for fine mapping of *qHKW1* and transgenic experiments.

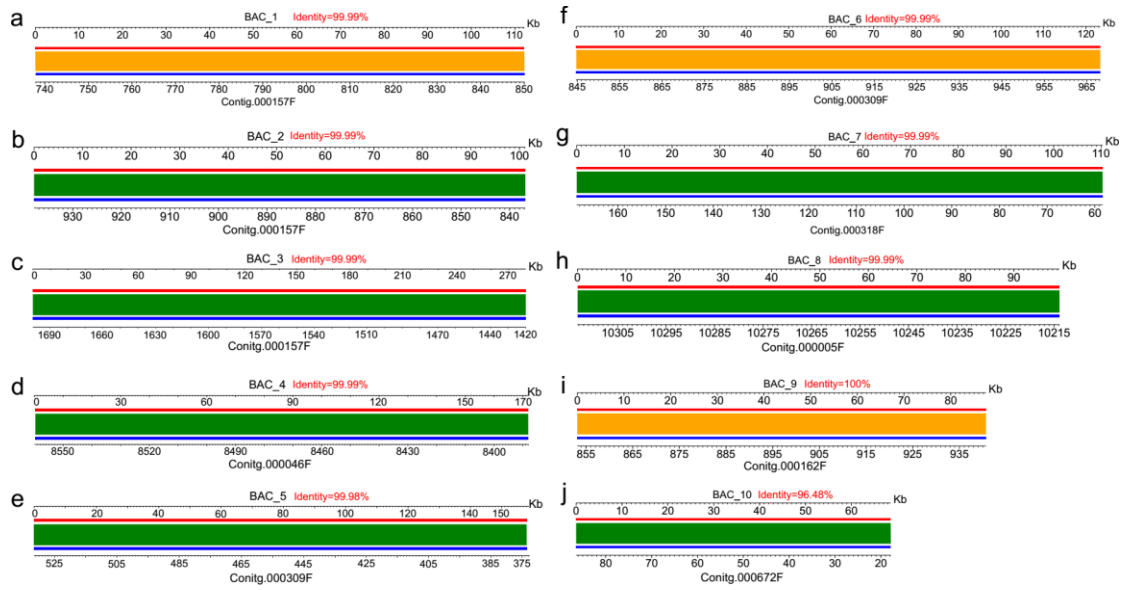
No.	ID	Sequence	Purpose
1	P02	5'-CTGCGGCAACTCTTTTGTAGTA-3'/5'-CAACCATTCCGGGAACATGC-3'	Markers for fine mapping
2	P13	5'-TCGATCGAACGAGGACCAAC-3'/5'-TCCAAGTAATCCAAACATGCCC-3'	Markers for fine mapping
3	ID02	5'-CCGGCTGCCTGCCTTATCAT-3'/5'-CGTGGATGGAAATGGAACACG-3'	Markers for fine mapping
4	M49	5'-CTAGCACACTGGCACACGAC-3'/5'-CATGGATCGGTACCACTAGAC-3'	Markers for fine mapping
5	SSR107	5'-TTGTCATCGAGCCGAAGGAG-3'/5'-ACTAAGAAGATCCTGCTGCGG-3'	Markers for fine mapping
6	M40	5'-TCCGGGATTGCTGTTTGGAG-3'/5'-AGATGACGTACGAAACCGGC-3'	Markers for fine mapping
7	SSR70	5'-CGATGCGGTGCAAACACTACAG-3'/5'-TAAAACCACGCACAACCTGGT-3'	Markers for fine mapping
8	1483	5'-TGCATTAGAACCACACTGTCCAC-3'/5'-ACCATGGGTTCCAATGGCTT-3'	Markers for fine mapping
9	028317&YFP	5'-gacaaacgcactagtagtcccgggATGGCGATGGCGCCCTCC-3/ 5'-tcaccatggcgccttcccgggATACTAAGAAGATCCTG-3'	Infusion Zm00001d028317-YFP
10	YFP-V	5'-actagatcccgggaaggcgcgccATGGTGAGCAAGGGCGAGG-3/ 5'-tgaacgataagcttagcgcgccCTACTTGTACAGCTCGTCC-3'	YFP into vector
11	CRISPR-F	5'-TGCACTGCACAAGCTGCTGTTTTTGTAGCCCCATCG-3'	CRISPR-Cas9 Construct
12	pU6F2	5'-TGCTTTTTTTAAGCTGCTGTTTTTGTAGCCCCATCG-3'	CRISPR-Cas9 Construct
13	CRISPR-R	5'-GGCCAGTGCCAAGCTTAAAAAAGCACCGACTCG-3'	CRISPR-Cas9 Construct
14	CRISPR-g1F	5'-GGCGGGATACCCACGAACCTGTTTTAGAGCTAGAAATAGCAAGTT-3'	CRISPR-Cas9 Construct
15	CRISPR-g2F	5'-GTGGAGCCGCATCCGGTACCGTTTTAGAGCTAGAAATAGCAAGTT-3'	CRISPR-Cas9 Construct
16	CRISPR-g1R	5'-AGGTTTCGTGGGTATCCCGCCAATTCGGTGCTTGC GGCTC-3'	CRISPR-Cas9 Construct
17	CRISPR-g2R	5'-GGTACCGGATGCGGCTCCACAATTCGGTGCTTGC GGCTC-3'	CRISPR-Cas9 Construct
18	CRISPR-C	5'-TCACGCCCTTTTAAATATCCGA-3'/5'-AGATAAACTGCACCTCAAACAAGT-3'	Cas9 detection
19	CRISPR-PD	5'-GCAGATCGTGGTATGTGCC-3'/5'-CCTTCCTGGTGGAAGAGGATA-3'	Cas9 detection
20	CRISPR-D	5'-TCCACTCACCACACCATCCA-3'/5'-AGAAGACACCTCCCTCCGT-3'	Sequencing editing target
21	q028317&Y	5'-GCCGCAGCAGGATCTTCTTA-3'/5'-GTCAGCTTGCCGTAGGTGG-3'	qRT-PCR for exogenous Zm00001d028317-YFP
22	q028317	5'-GCCGCAGCAGGATCTTCTTA-3'/5'-AATCACCGCCTGCATACACA-3'	qRT-PCR for endogenous Zm00001d028317
23	ZmUBI	5'-GGCCGCACCTTAGCAGACTA-3'/5'-ATGGAGAGGGCACCAGACGA-3'	Control for qRT-PCR
24	028317-P	5'-GGAGTAGAGAAGCATTTGAGGC-3'/5'-ATGGTGTGGTGTGAGTGGAGTG-3'	Amplify promoter of Zm00001d028317
25	Methy-S	5'-TTTTAATAATTTGTTTTTGGGAG-3'/5'-ACATTCCTAAAACCTATATTTTAAAAA-3'	Methylation level quantification of NIL(SK)
26	Methy-Z	5'-AATGTAGGGATGTATATTTTTTATTTT-3'/5'-ATAAAACCCACTTCRACATATA-3'	Methylation level quantification of NIL(ZHENG58)



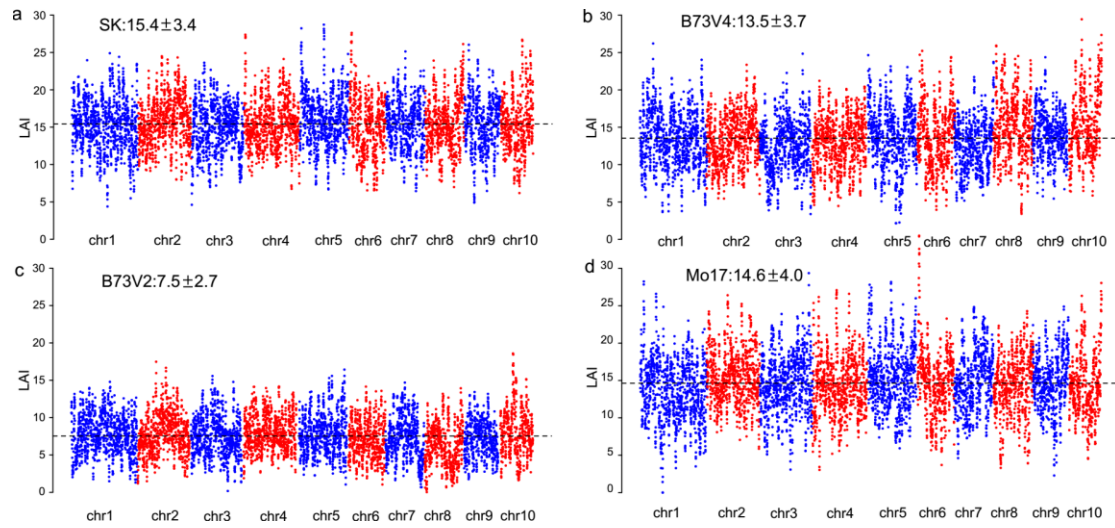
Supplementary Fig. 1 | Phylogenetic tree of association panel²⁷. SK does not belong to the two North American heterotic groups: Stiff Stalk (SS, green lines) and Non-Stiff Stalk (NSS, blue lines). SK is derived from a landrace with tropical background (TST, red lines).



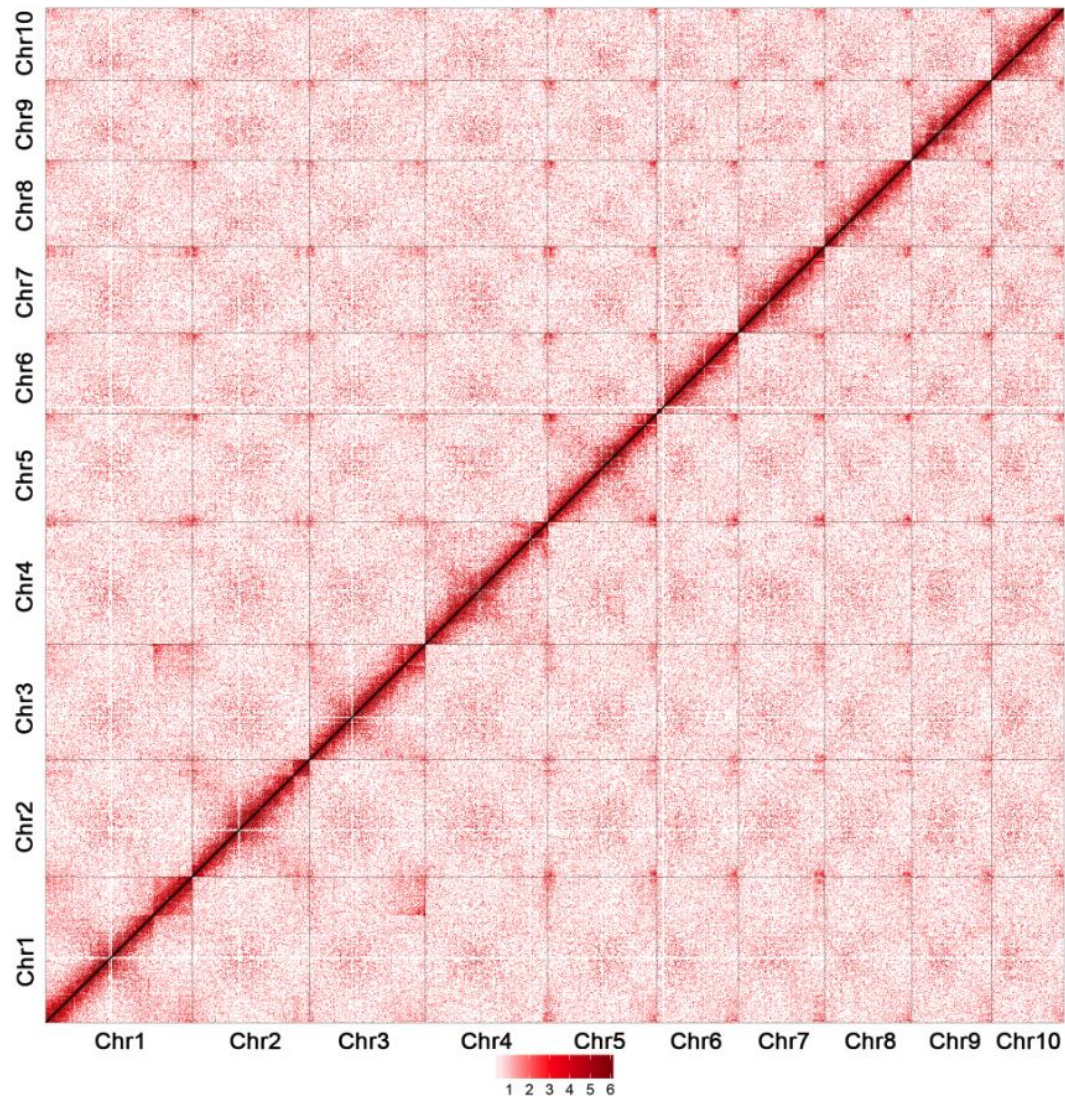
Supplementary Fig. 2 | Graphical representation of the location of SNP markers on the physical map (x-axis), as compared to their position on the integrated genetic map (y-axis) of the SK genome.



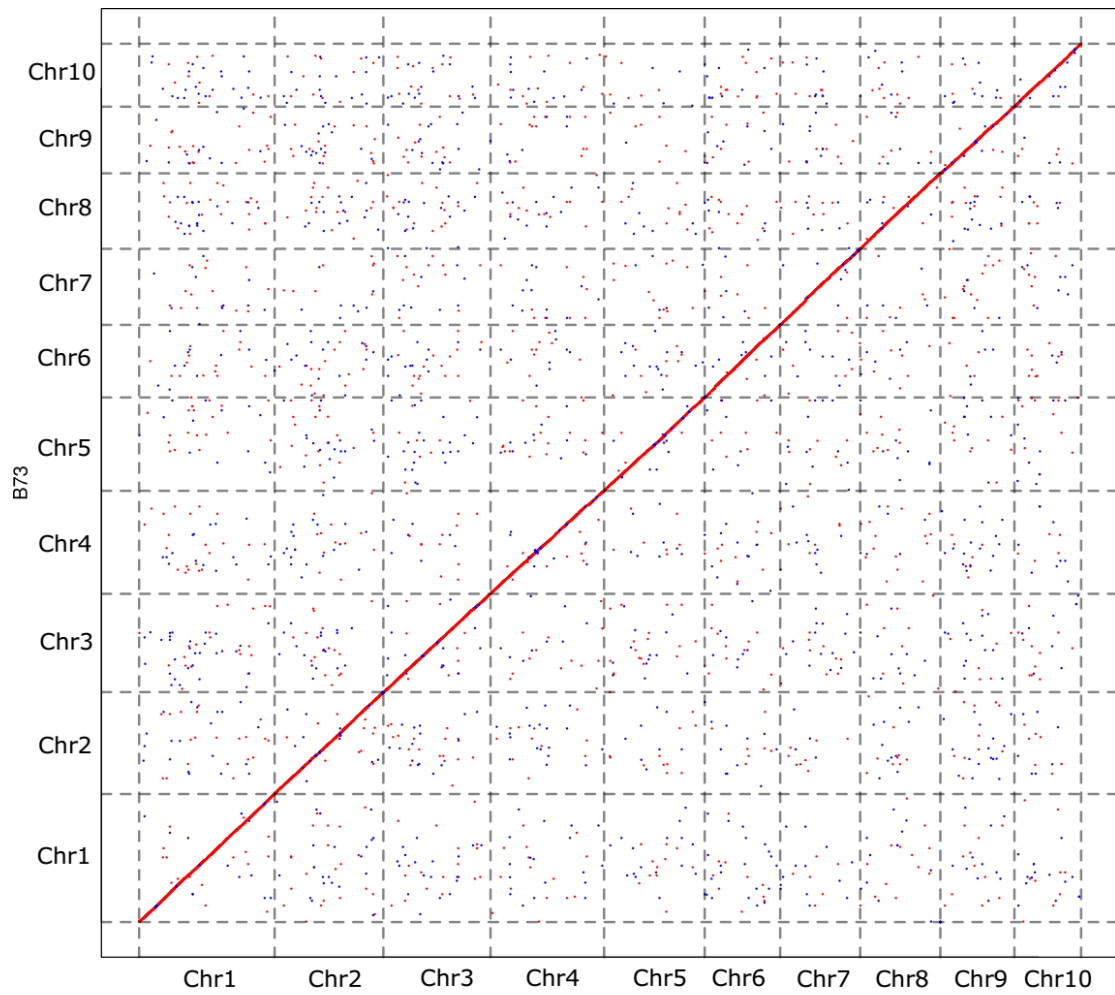
Supplementary Fig. 3 | Assessment of the SK genome assembly by ten fully sequenced SK BACs. The red line indicates each BAC sequence and the blue line represents the scaffold. The orange blocks and the green blocks show aligned regions between the BAC sequences and the scaffolds. Blue = plus strand, orange = minus strand.



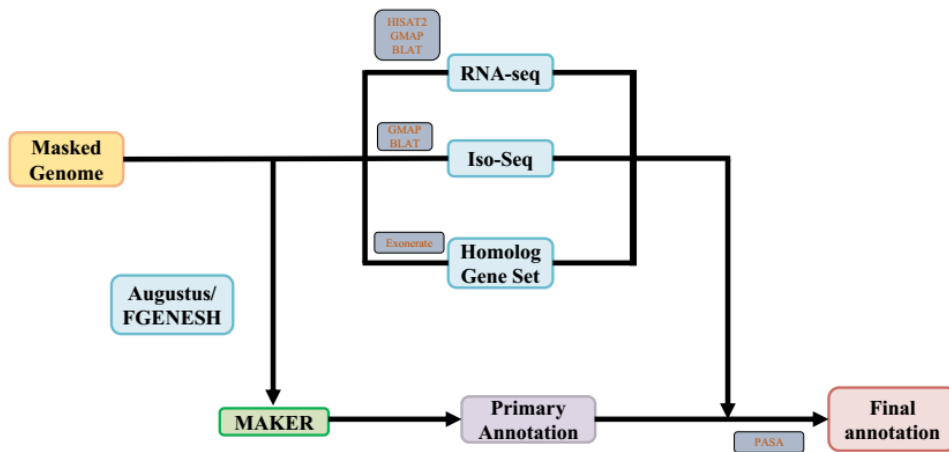
Supplementary Fig. 4 | Assessment of genome assemblies by LTR Assembly Index (LAI)²⁴. The LAI score for SK (A), B73 v4 (B), B73V2 (C) and Mo17 (D). These results indicate SK had the best assembly continuity.



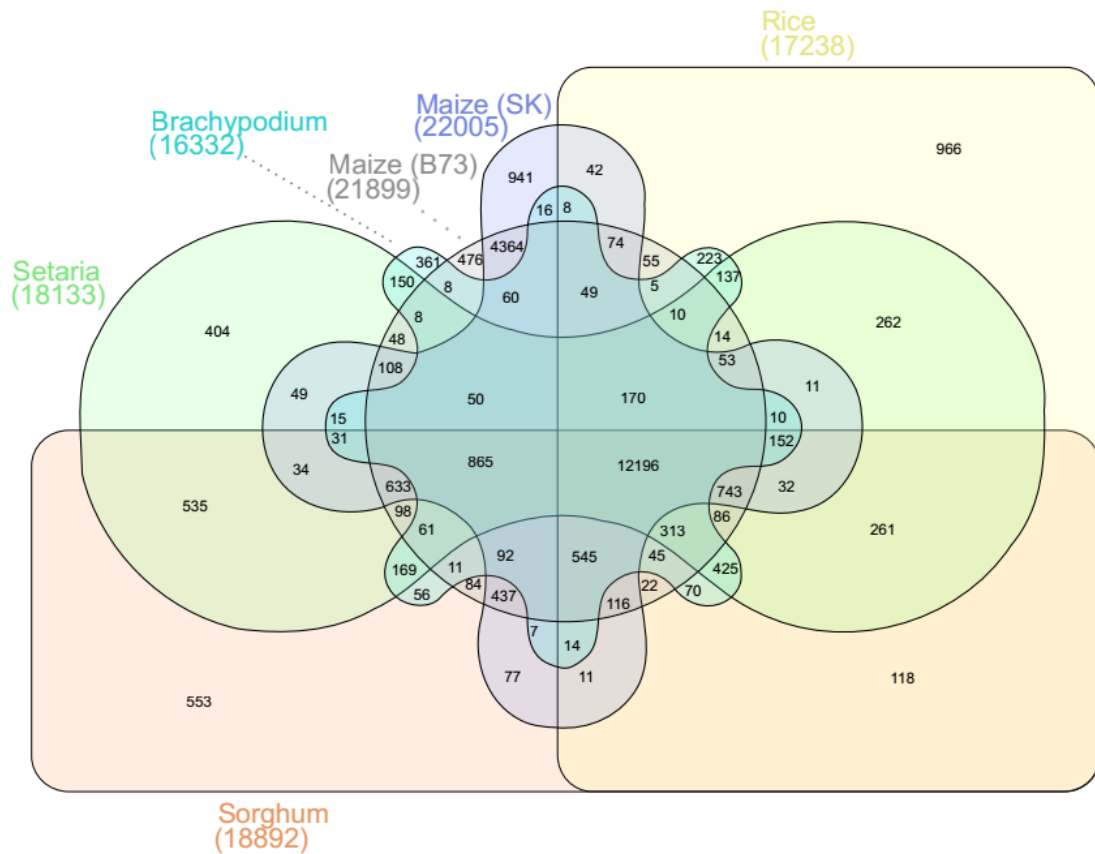
Supplementary Fig. 5 | Heat map of whole-genome chromatin contact matrices generated by aligning a ChIA-PET data set of RNAPII binding sites to SK assembly. The frequency of interactions was calculated using 1-Mb bin size.



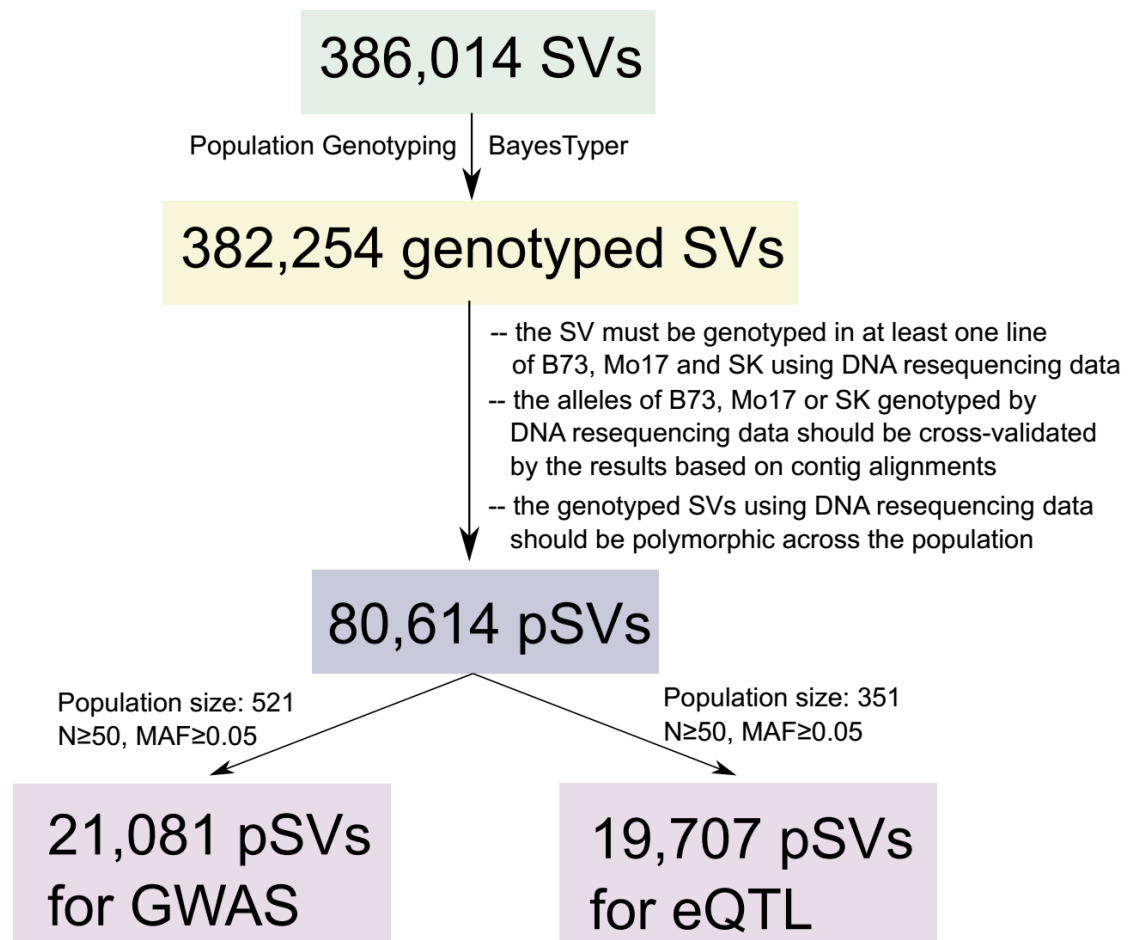
Supplementary Fig. 6 | The colinearity between SK and B73 genomes. Each dot indicates an aligned region whose length is at least 5 kb.



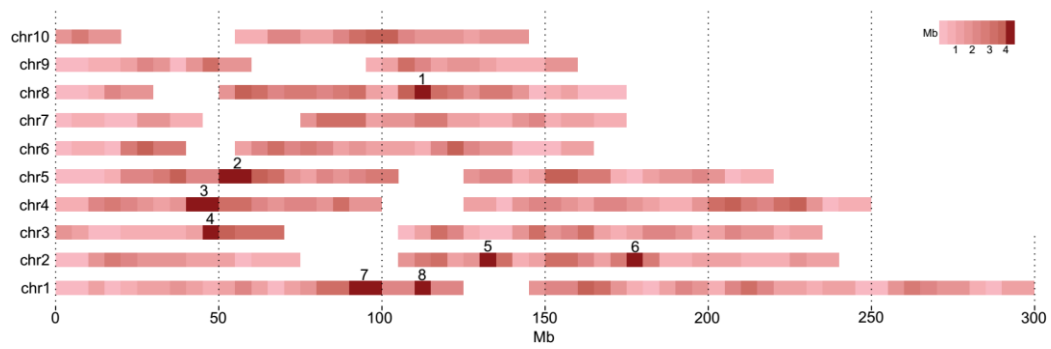
Supplementary Fig. 7 | Pipeline used for gene annotation. HISAT2⁷⁸, GAMP⁷⁹, BLAT⁴³, Exonerate⁶⁸, PASA⁵³.



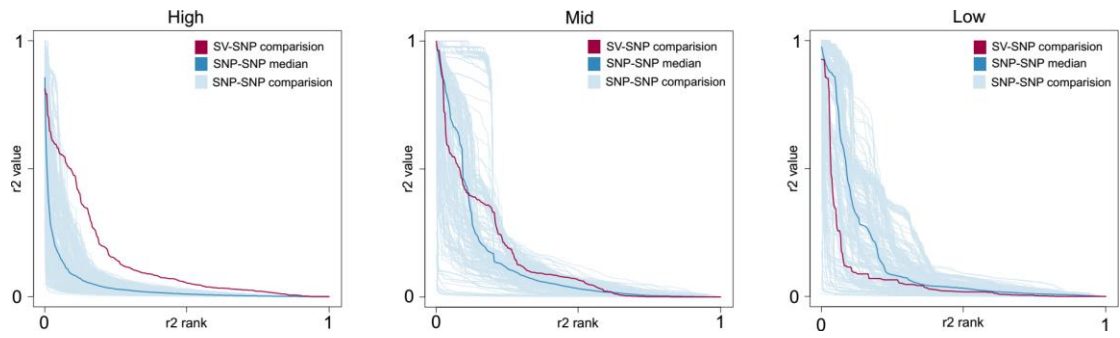
Supplementary Fig. 8 | Comparison of gene families among six fully sequenced grass genomes. Unique and shared gene families between *Zea mays* (SK), *Zea mays* (B73, v4), *Setaria italica* (Setaria, JGI-V2.0), *Sorghum bicolor* (sorghum, V2), *Oryza sativa* (rice, IRGSP-1.0) and *Brachypodium distachyon* (Brachypodium, V1.0) are depicted in the 6-way Venn diagram.



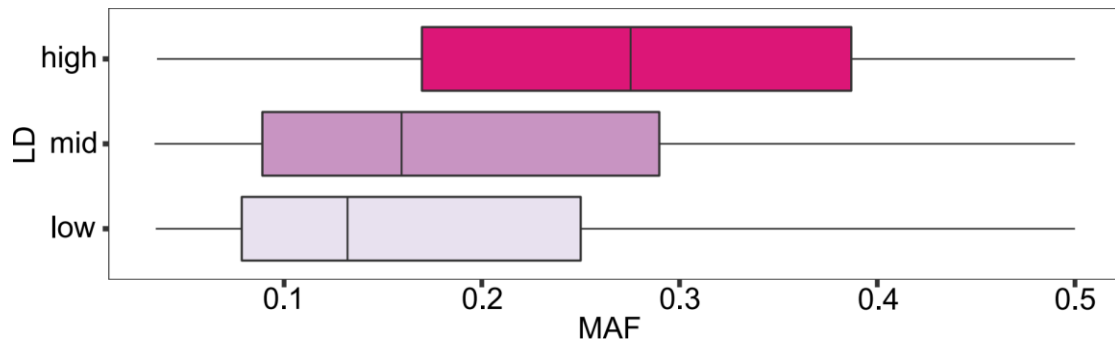
Supplementary Fig. 9 | The follow chart of filter conditions for genotyped SVs. Because only 368 lines had RNA-seq data²⁸, the population size was 351 after taking the intersection for eQTL analysis.



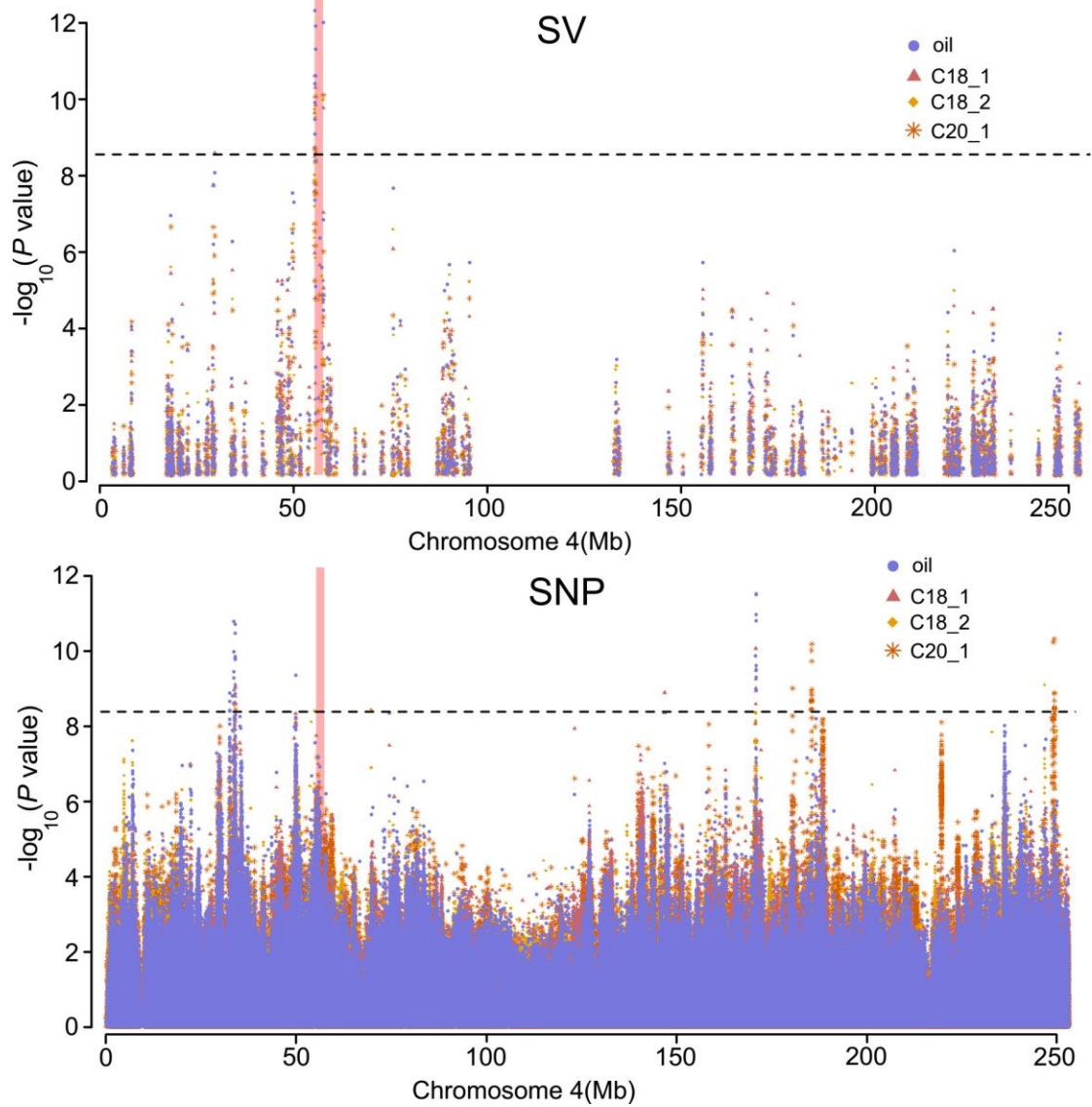
Supplementary Fig. 10 | A map of pSVs. The color denotes the number of pSVs bases (Mb), within a 10-Mbp sliding window (5-Mbp step).



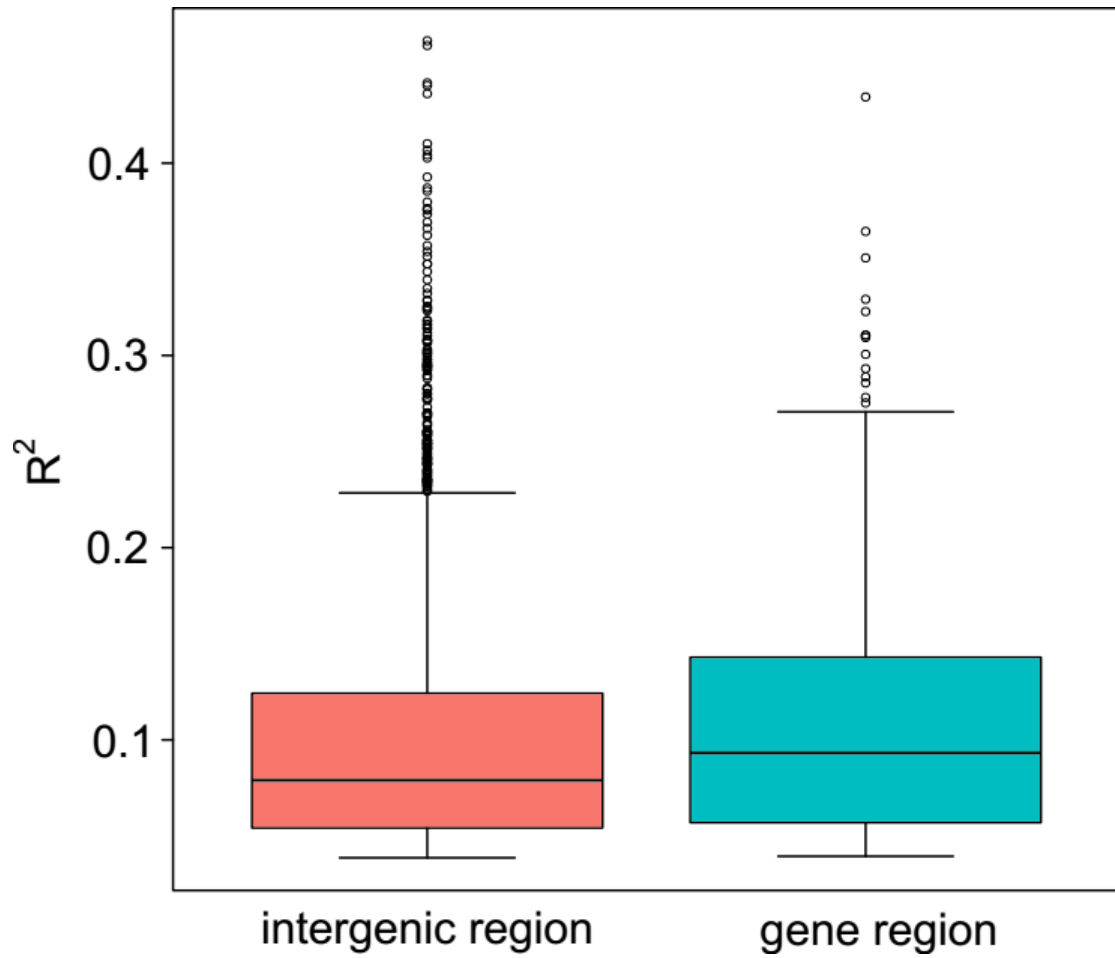
Supplementary Fig. 11 | Rank order plots for individual representative high, middle and low-LD SVs. Blue line indicates the median r^2 value for each rank across SNP-based values. Red line indicates r^2 values for SV-SNP comparisons. Light blue lines indicate all individual SNP-SNP comparisons.



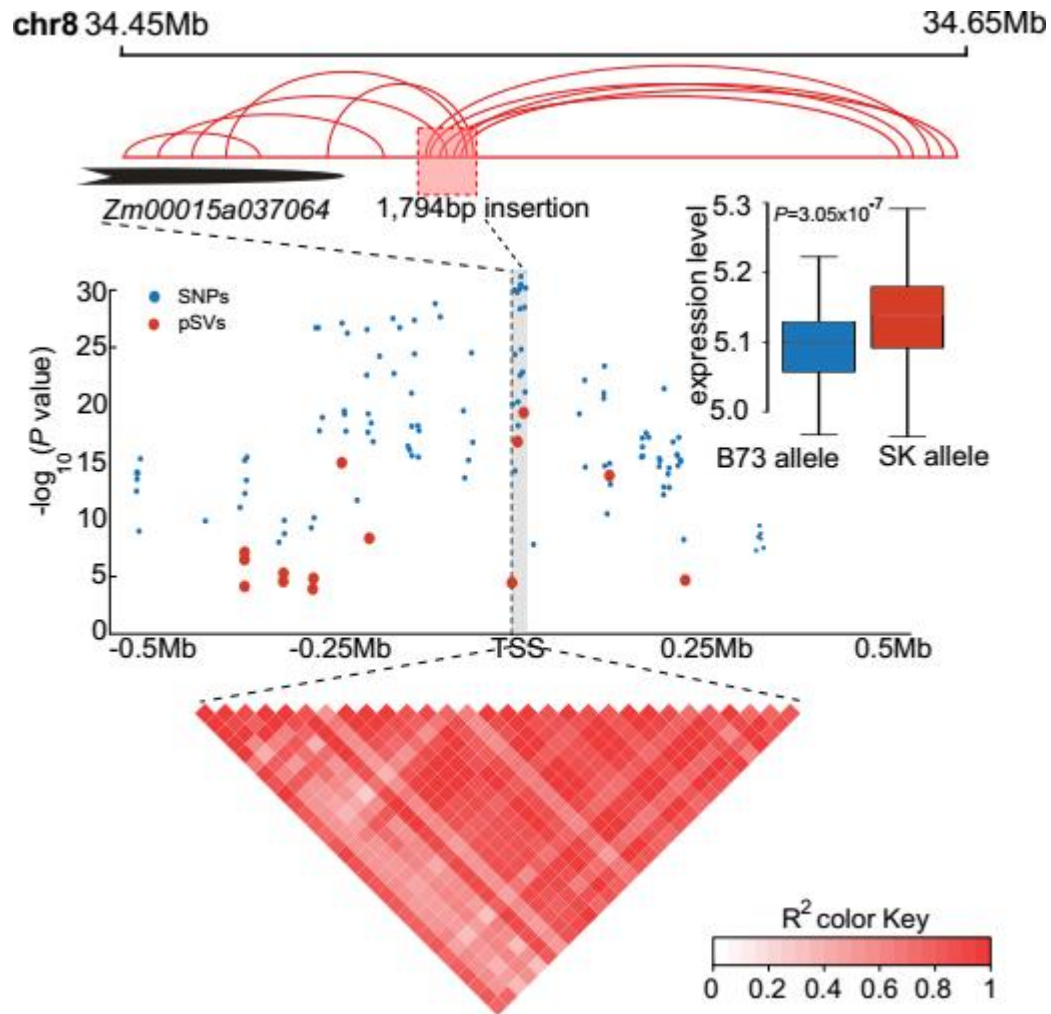
Supplementary Fig. 12 | Boxplots showing distribution of minor allele frequencies for each LD category.



Supplementary Fig. 13 | Manhattan plot of SV-GWAS and SNP-GWAS for oil concentration, C18_1, C18_2 and C20_0. There was no significant SNPs in the candidate region detected by SV-GWAS. The red rectangle indicates the candidate region: chr4: 55.09 Mb-55.34Mb.

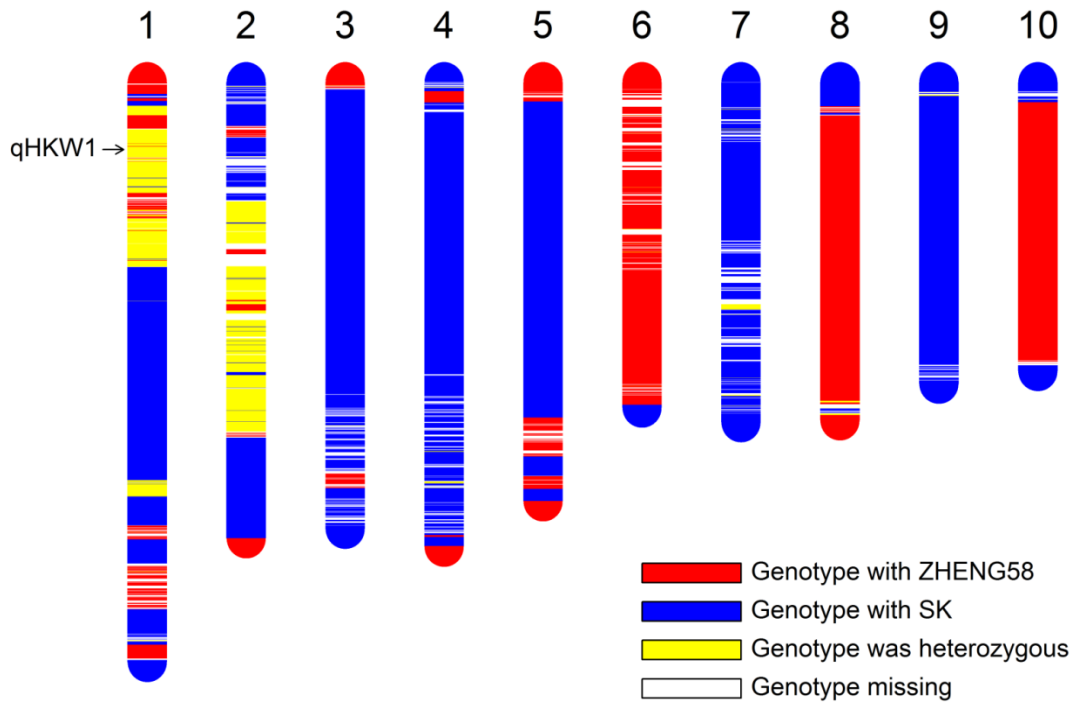


Supplementary Fig. 14 | The effect of lead SV eQTLs and LD-linked SV eQTLs in different region ($P=4.4 \times 10^{-4}$).

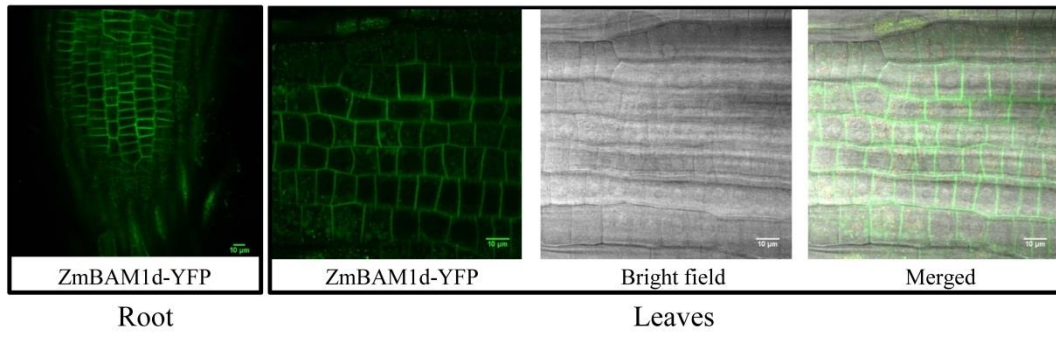


Supplementary Fig. 15 | A 1,794 bp SV was a *cis* expression quantitative trait locus of *Zm00015a037064* and could affect gene expression by affecting chromatin interactions.

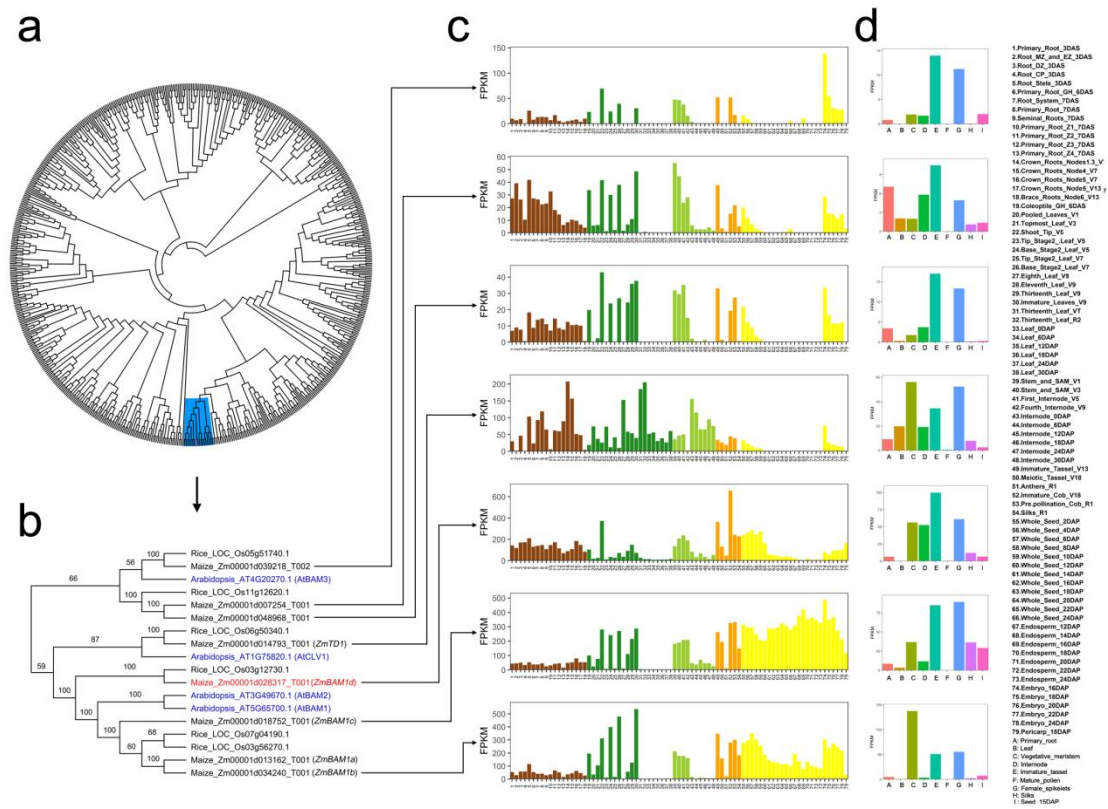
HIF_Founder HZAU1348



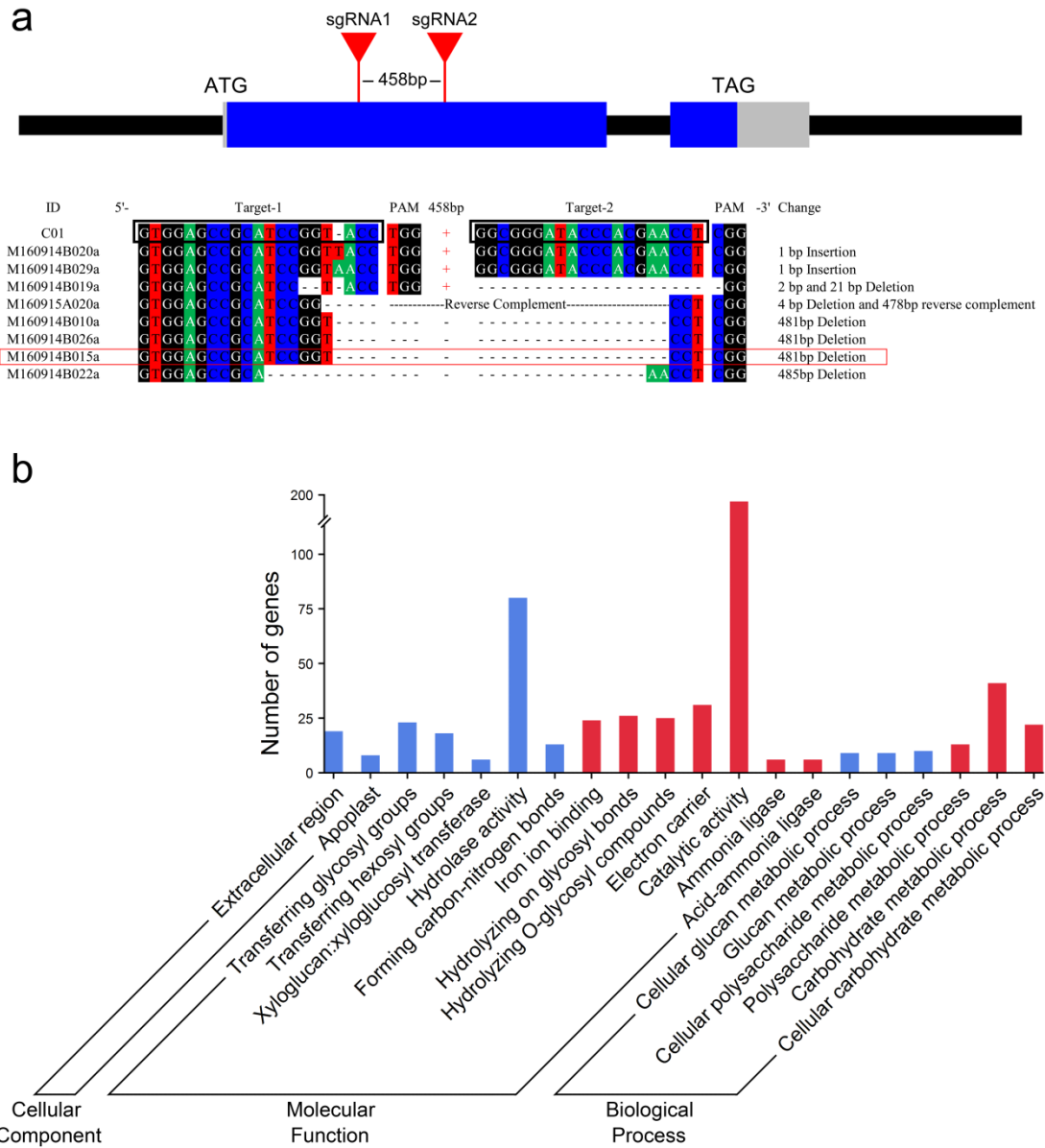
Supplementary Fig. 16 | The genetic background of HIF founder for fine-mapping of *qHKW1*.



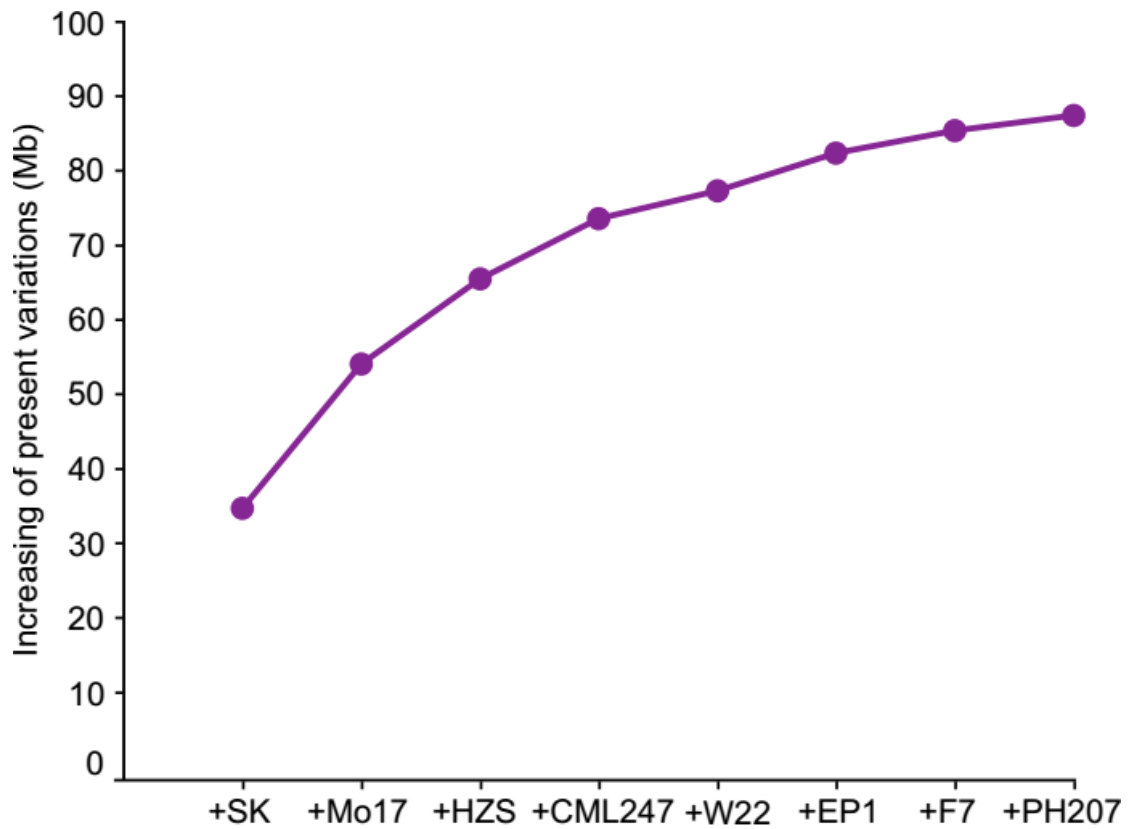
Supplementary Fig. 17 | YFP tagged ZmBAM1d protein is localized on the plasma membrane in ZmBAM1d-YFP transgenic lines.



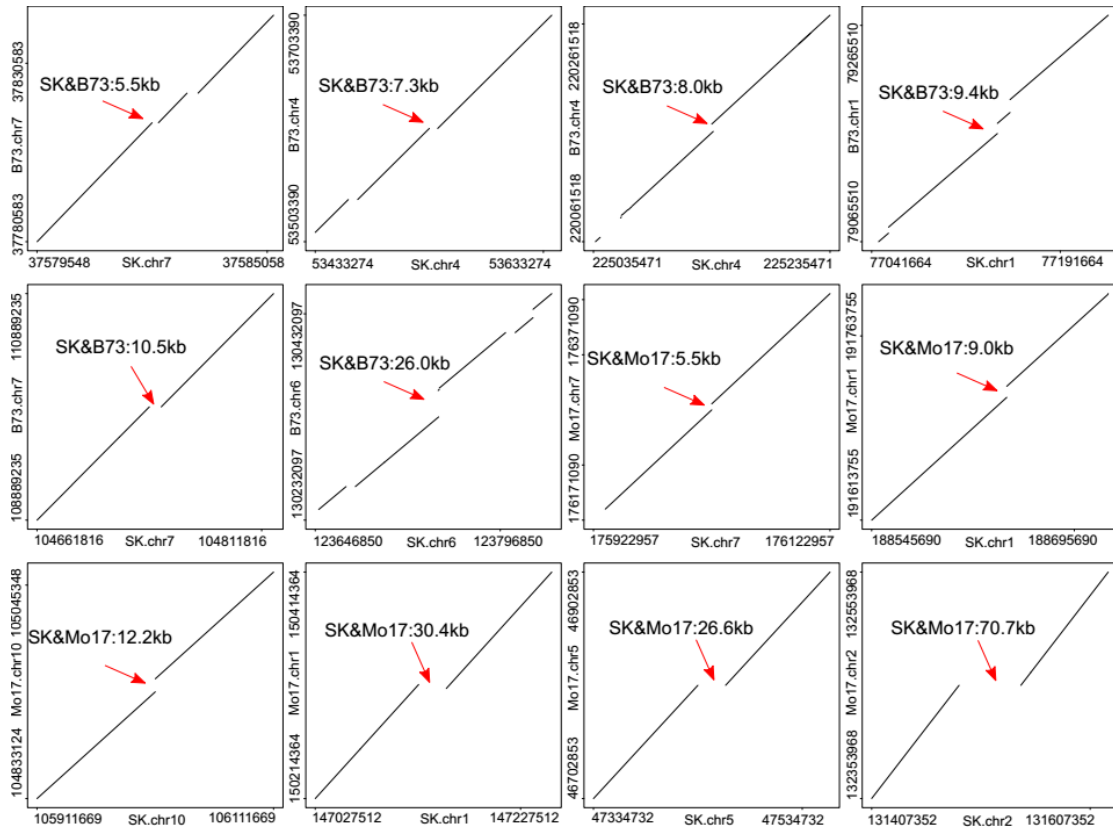
Supplementary Fig. 18 | (a) Phylogenetic tree analysis of genes encoding both LRR and protein kinase domain in maize, rice, and Arabidopsis. The group which *ZmBAM1d* locate in is indicated with blue background. (b) Red indicates *ZmBAM1d* and blue indicate *BAM* genes and *AtCLV1* in Arabidopsis. (c) The expression patterns of *ZmBAM1d* and its homologous in 79 tissues of B73 (data source: MaizeGDB (<https://www.maizegdb.org/>)). (d) The expression patterns of *ZmBAM1d* and its homologous in 9 tissues of SK.



Supplementary Fig. 19 | The 481bp deletion CRISPR edited event which was used for RNA-Seq and the significantly enriched GO terms in differential expressed genes between over-expression transgenic positive and negative lines. (a) CRISPR edited events of *ZmBAM1d* and the event indicated with red box was used for RNA-Seq. (b) The red GO terms are also significantly enriched in differential expressed genes detected between CRISPR edited and control plants.



Supplementary Fig. 20 | The increasing size of pan-genome of *Zea mays* predicted by present variations using public maize genomes, including Mo17¹¹, HZS⁸⁰, CML247⁴⁴, W22¹², EP1⁸¹, F7⁸¹ and PH207⁸².



Supplementary Fig. 21 | The visual inspection for 12 randomly selected SVs (from 5kb to 70kb).