A chromosome-level genome assembly and annotation of a maize elite breeding line Dan340

Yikun Zhao[1,5], Yuancong Wang[2,5], De Ma[3,5], Guang Feng[4], Yongxue Huo[1], Zhihao Liu[1], Ling Zhou[2], Yunlong Zhang[1], Liwen Xu[1], Liang Wang[4], Han Zhao[2], Jiuran Zhao[1,*], Fengge Wang[1,*]

F.W., J.Z. and H.Z. conceived the project; Y-K.Z. D.M. and Y.W. wrote, modified the manuscript; L.X. G.F. and L.W. performed the experiments; Y.H. L.Z. Y-L.Z. and Z.L. analyzed data. All authors read and approved the final manuscript.

[1] Maize Research Center, Beijing Academy of Agricultural and Forest Sciences (BAAFS)/Beijing Key Laboratory of Maize DNA Fingerprinting and Molecular Breeding

[2] Provincial Key Laboratory of Agrobiology, Institute of Crop Germplasm and Biotechnology, Jiangsu Academy of Agricultural Sciences

[3] Novogene Bioinformatics Institute

[4] Dandong Academy of Agricultural Sciences

[5] These authors contributed equally to this work.

[*] **Correspondence**: Jiuran Zhao (maizezhao@126.com), Fengge Wang (fenggewangmaize@126.com)

Abstract

Background: Maize is not only one of the most important crops grown worldwide for food, forage, and biofuel, but also an important model organism for fundamental

research in genetics and genomics. Owing to its importance in crop science, genetics and genomics, several reference genomes of common maize inbred line (genetic material) have been released, but some genomes of important genetic germplasm resources in maize breeding research are still lacking. The maize cultivar Dan340 is an excellent backbone inbred line of the Luda Red Cob Group with several desirable characteristics, such as disease resistance, lodging resistance, high combining ability, and wide adaptability. **Findings:** In this study, we constructed a high-quality chromosome-level reference genome for Dan340 by combining PacBio long HiFi sequencing reads, Illumina short reads and chromosomal conformational capture (Hi-C) sequencing reads. The final assembly of the Dan340 genome was 2,348.72 Mb, including 2,738 contigs and 2,315 scaffolds with N50 of 41.49 Mb and 215.35 Mb, respectively. Repeat sequences accounted for 73.40% of the genome size and 39,733 protein-coding genes were annotated. Analysis of genes in the Dan340 genome, together with those from B73, Mo17 and SK, were clustered into 27,654 gene families. There were 1,806 genes from 359 gene families that were specific to Dan340, of which many had functional gene ontology annotations relating to "porphyrin-containing compound metabolic process", "tetrapyrrole biosynthetic process", and "tetrapyrrole metabolic process". **Conclusions:** The completeness and continuity of the genome were comparable to those of other important maize inbred lines. The assembly and annotation of this genome not only facilitates our understanding ~~about~~ of intraspecific genome diversity in maize, but also provides a novel resource for maize breeding improvement.

Research Areas:
Genetics and Genomics**;** Agriculture, Plant Genetics

Data Description

Background

Maize (*Zea mays* ssp. *mays* L., NCBI:txid381124) is one of the most important crops grown worldwide for food, forage, and biofuel, with an annual production of more than 1 billion tons [1]. Owing to rapid human population growth and economic demand, maize has been predicted to account for 45% of total cereal demand by the year 2050 [2]. In addition, it is an important model organism for fundamental research in genetics and genomics [3].

Because of its importance in crop science, genetics and genomics, several reference genomes of common maize inbred lines used in breeding have been released since 2009 [4-8]. However, comparative genomic analyses have found that maize genomes exhibit high levels of genetic diversity among different inbred lines [1, 7, 9]. Meanwhile, accumulating studies have suggested that one or a few reference genomes cannot fully represent the genetic diversity of a species [7, 10, 11].

The maize cultivar Dan340 is an excellent backbone inbred line of the Luda Red Cob Group that has several desirable characteristics, such as disease resistance, lodging resistance, high combining ability, and wide adaptability. More than 50 maize hybrid breeds have been derived from Dan340 since 2000, and their planting area has reached 19 million ha. It is considered that Dan340 originated from a landrace in China and exhibits large genetic differences from other maize germplasms that represent the most important core maize germplasms in China [12]. Therefore, it

71 could serve as a model inbred line for the genetic dissection of desirable agronomic
72 traits, combining ability, heterosis, and breeding history.
73 In the present study, we constructed a high-quality chromosome-level reference
74 genome for Dan340 by combining PacBio long HiFi sequencing reads, Illumina short
75 reads and chromosomal conformational capture (Hi-C) sequencing reads. The
76 completeness and continuity of the genome were comparable with those of other
77 important maize inbred lines, B73 [4], Mo17 [7], SK [13], PH207 [5], and HZS [8].
78 Furthermore, comparative genomic analyses were performed between Dan340 and
79 other maize lines, and genes and gene families that were specific to Dan340 were
80 identified. In addition, large numbers of structural variations between Dan340 and
81 other maize inbred lines were detected. The assembly and annotation of this genome
82 will not only facilitate our understanding of intraspecific genomic diversity in maize,
83 but also provides a novel resource for maize breeding improvement.

84 **Plant materials and DNA sequencing**

85 Inbred line Dan340 (Fig. 1) was selected for genome sequencing and assembly
86 because it is an elite maize cultivar, that plays an important role in maize breeding and
87 genetic research. The plants were grown at 25°C in a greenhouse of the Beijing
88 Academy of Agriculture and Forestry Sciences, Beijing, China. Fresh and tender
89 leaves were harvested from the best-growing individual and immediately frozen in
90 liquid nitrogen, followed by preservation at -80 °C in the laboratory prior to DNA
91 extraction. Genomic DNA was extracted from the leaf tissue of a single plant using

the DNAsecure Plant Kit (Tiangen Biotech Co., Ltd., Beijing, China). To ensure that DNA extracts were useable for all types of genomic libraries, the quality and quantity were evaluated using a NanoDrop 2000 spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA) and electrophoresis on a 0.8% agarose gel, respectively.

In recent years, third-generation DNA sequencing technologies have undergone rapid technological innovation and are now widely used in genome assembly. In this study, PacBio CCS libraries were prepared using the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences, Menlo Park, CA, USA; Ref. No. 101-685-400), following the manufacturer's protocols, and they were subsequently sequenced on the PacBio sequel II platform (Pacific Biosciences, RRID:SCR_017990). As a result, 63.53 Gb (approximately 27× coverage) of HiFi reads was generated and used for the genome assembly.

In addition, one Illumina paired-end sequencing library, with an insert size of 350 bp, was generated using the NEB Next Ultra DNA Library Prep Kit (NEB, Ipswich, MA, USA) following the manufacturer's protocol, and it was subsequently sequenced using an Illumina HiSeq X Ten platform (Illumina, San Diego, CA, USA, RRID:SCR_016385) at the Novogene Bioinformatics Institute, Beijing, China. Approximately 80.66 Gb (~34×) of Illumina sequencing data were obtained.

One Hi-C library was constructed using young leaves following previously published procedures, with slight modifications [14]. In brief, approximately 5-g leaf samples from seedling were cut into minute pieces and cross-linked by 4% formaldehyde

113 solution at room temperature in a vacuum for 30 min. Then,-each sample was mixed
114 with excess 2.5 M glycine to quench the crosslinking reaction for 5 min and then
115 placed on ice for 15 min. The cross-linked DNA was extracted and then digested for
116 12 h with 20 units of *DpnII* restriction enzyme (NEB) at 37 °C, and the resuspended
117 mixture was incubated at 62 °C for 20 min to inactivate the restriction enzyme. The
118 sticky ends of the digested fragments were biotinylated and proximity ligated to form
119 enriched ligation junctions and then ultrasonically sheared to a size of 200 - 600 bp.
120 The biotin-labelled DNA fragments were pulled down and ligated with Illumina
121 paired-end adapters, and then amplified by PCR to produce the Hi-C sequencing
122 library. The library was sequenced using an Illumina HiSeq X Ten platform with 2 ×
123 150 bp paired-end reads (Illumina, San Diego, CA, USA). After removing low-quality
124 sequences and trimming adapter sequences, 304.37 Gb (approximately 130×) of
125 clean data were generated and used for the genome assembly.

126 **Genome assembly**

127 To obtain a high-quality genome assembly of Dan340, we employed both PacBio
128 HiFi reads and Illumina short reads, with scaffolding informed by high-throughput
129 chromosomal conformation capture (Hi-C).

130 The assembly was performed in a stepwise fashion. First, de novo assembly of the
131 long CCS reads generated from PacBio SMRT sequencing was performed using
132 Hifiasm [15] (RRID:SCR_021069)(https://github.com/chhylp123/hifiasm). A total of
133 two SMRT cells produced 4,073,418 subreads, with an average length of 15,598 bp

and a read N50 of 15,715 bp. Generation of HiFi reads and adapter trimming were performed using PacBio SMRTLink v8.0 [16] with default parameters, followed by the deduplication of reads with pbmarkdup v0.2.0 [17] as recommended by PacBio. Next, HiFi reads were aligned to each other and assembled into genomic contigs using Hifiasm [15] with default parameters. The produced primary contigs (p-contigs) were then polished using Quiver [18] by aligning SMRT reads. Then, Pilon [19] (RRID:SCR_014731) was used to perform the second round of error correction with short paired-end reads generated from Illumina Hiseq platforms. Subsequently, the Purge Haplotigs pipeline [20] was used to remove redundant sequences formed as a result of heterozygosity. The draft genome assembly was 2348.68 Mb, which reached a high level of continuity, with a contig N50 length of 45.11 Mb.

For Hi-C reads, to avoid reads having an artificial bias, we removed the following type of reads using HICUP software [21] (RRID:SCR_005569)(http://www.bioinformatics.babraham.ac.uk/projects/hicup/): (a) Reads with ≥ 10% unidentified nucleotides (N); (b) Reads with > 10 nt aligned to the adapter, allowing ≤ 10% mismatches; and (c) Reads with > 50% bases having a phred quality < 5. The filtered Hi-C reads were aligned against the contig assemblies using BWA (version 0.7.8, RRID:SCR_010910)(http://bio-bwa.sourceforge.net/). Reads were excluded from subsequent analyses if they did not align within 500 bp of a restriction site or did not uniquely map, and the number of Hi-C read pairs linking each pair of scaffolds was tabulated. ALLHiC v0.8.12 [22] was used in simple diploid

mode to scaffold the genome and optimize the ordering and orientation of each clustered group, producing a chromosome-level assembly. Juicebox Assembly Tools v1.9.8 [23] (RRID:SCR_021172) was used to visualize and manually correct the large-scale inversions and translocations to obtain the final pseudo-chromosomes (Fig. 2). Finally, a total of 2315 scaffolds (representing 91.30% of the total length) were anchored to 10 chromosomes (Fig. 3).

The final assembly of the Dan340 genome was 2,348.72 Mb, including 2,738 contigs and 2,315 scaffolds, with N50 of 41.49 Mb and 215.35 Mb, respectively (Table 1).

Evaluation of assembly quality

We assessed the quality of the assembly using several independent methods. First, the short reads obtained from the Illumina sequencing data were aligned to the final assembly using BWA version 0.7.8 [24]. Our result showed that the percent of reads mapped to the reference genome was up to 97.48%. Second, a total of 248 conservative genes existing in six eukaryotic model organisms were selected to form the core gene library for the CEGMA (Core Eukaryotic Genes Mapping Approach) [25] (RRID:SCR_015055) evaluation. Our assembled Dan340 genome was aligned to this core gene library using TBLASTN [26] (RRID:SCR_011822), Genewise version 2.2.0 [27] (RRID:SCR_015054), and Geneid v1.4 [28] (RRID:SCR_021639) software tools to evaluate its integrity. The result showed that 238 complete (95.97%) and 243 partial (97.98%) genes were detected in our assembly. Third, the completeness was assessed using the benchmarking universal single-copy orthologs (BUSCO) [29]

(RRID:SCR_015008). The final assembly was tested against the database embryophyta_odb10, which includes 1,614 conserved core genes. The result showed that 98.08% (1,583), 1.11 % (18), and 0.81% (13) of the plant single-copy orthologs were present in the assembled Dan340 genome as complete, fragmented, and missing genes, respectively. Fourth, the long-terminal repeat (LTR) Assembly Index (LAI) metric was used to evaluate assembly continuity in Dan340 and three other maize genomes (B73, Mo17 and SK, Fig. 4). Intact LTR retrotransposons in the four genomes were identified using LTRharvest v1.6.1 [30] (RRID:SCR_018970), LTR_FINDER v1.07 [31] (RRID:SCR_015247), and LTR_retriever v2.9.0 [32] (RRID:SCR_017623). The LAI pipeline was executed using the following parameter settings: -t 20 -intact genome.fasta.pass.list -all genome.ltr.fasta.out. Our Dan340 genome had a LAI score of 25.13, which was relatively high among the four maize genomes compared in this study. B73, Mo17, and SK produced scores of 24.94, 24.45, and 27.12, respectively (Fig. 4 and Table 2). A higher LAI score indicates more complete genome assembly because more intact LTR retrotransposons are identified, as in our Dan340 genome. Furthermore, whole-genome sequence alignments of Dan340 to the genomes of the other three maize inbred lines demonstrated that our assembly has highly collinear relationships with other published maize genomes (Fig. 5). Taken together, the assessment results suggested that the Dan340 genome assembly was of high quality.

Genome annotation

Repeat sequences of the Dan340 genome were annotated using both *ab initio* and homolog-based search methods. For the *ab initio* prediction, RepeatModeler v1.0.8 [33] (RRID:SCR_015027), RepeatScout version 1.0.5 [34] (RRID:SCR_014653), and LTR_FINDER version 1.07 [31] were used to discover transposable elements (TEs) and to build TEs library. An integrated TEs library and a known repeat library (Repbase V15.02, homolog-based)(RRID:SCR_021169) were subjected to RepeatMasker v3.3.0 [35] (RRID:SCR_012954) to predict TEs. For homolog-based predictions, RepeatProteinMask was performed to detect TEs in the genome by comparing it against a TE protein database. Tandem repeats were ascertained in the genome using Tandem Repeats Finder (TRF, version 4.07b) [36] (RRID:SCR_022193). As a result, 1723.99 Mb of repeat sequences were identified, accounting 73.40% of the genome size. Among these repeat sequences, 1,555.57 Mb were predicted ~~as~~ to be long-terminal repeat (LTR) retrotransposons and 44.53 Mb were predicted ~~as~~ to be DNA transposons, accounting for 66.23% and 1.60% of the genome, respectively. Furthermore, among the LTR retrotransposons, the Gypsy and Copia superfamilies comprised 23.81% and 12.75% of the genome, respectively. Thus, retrotransposons accounted for a large proportion of the-Dan340 genome, which was consistent with the genomic characteristics of other maize inbred lines (Table 2).

All repetitive regions except tandem repeats were soft-masked for protein-coding gene annotation. Five *ab initio* gene prediction programs, Augustus v3.0.2 [37-39] (RRID:SCR_008417), Genscan v1.0 [40] (RRID:SCR_013362), Geneid v1.4 [28],

GlimmerHMM v3.0.2 [41] (RRID:SCR_002654) and SNAP v2013-02-16 [42] (RRID:SCR_007936), were used to predict genes. In addition, the protein sequences of five homologous species (*Sorghum bicolor*, *Setaria italica*, *Hordeum vulgare*, *Triticum aestivum*, and *Oryza sativa*) were downloaded from Ensembl and NCBI. Homologous sequences were aligned against the genome using TBLASTN (E-value 1E-05). Genewise version 2.2.0 [27] was employed to predict gene models on the basis of the sequence alignment results.

For RNA-seq prediction, fresh samples of six tissues (stem, endosperm, embryo, bract, silk, and ear tip) were collected. Total RNA was extracted from each sample using an RNAprep Pure Plant Kit (Tiangen Biotech Co., Ltd., Beijing, China). Isolated purified RNA was the template for the construction of a cDNA library, having fragment lengths of approximately 300 bp, using the NEBNext Ultra RNA Library Prep Kit for Illumina (New England Biolabs, Ipswich, MA, USA) in accordance with the manufacturer's instructions. Sequencing was performed on an Illumina HiSeq X Ten platform and 150-bp paired-end reads were generated. Raw reads were trimmed by removing adapter sequences, reads with more than 5% of unknown base calls (N), and low-quality bases (base quality less than 5, $Q \leqslant 5$). Clean paired-end reads were aligned to the genome using Tophat version 2.0.13 [43] (RRID:SCR_013035) to identify exon regions and splice positions. The alignment results were then used as input for cufflinks version 2.1.1 [44] (RRID:SCR_014597) to assemble transcripts to the gene models. In addition, RNA-seq data were assembled using Trinity version

2.1.1 [45] (RRID:SCR_013048), creating several pseudo-ESTs. These pseudo-ESTs were also mapped to the assembled genome using BLAT and gene models were predicted using PASA [46] (RRID:SCR_014656). A weighted and non-redundant gene set was generated using EVidenceModeler (EVM, version 1.1.1) [47] (RRID:SCR_014659), which merged all the gene models predicted by the above three approaches. Finally, PASA was used to adjust the gene models generated by EVM. As a result, a total of 39,733 protein-coding genes were annotated in the final set. To better understand gene functions, we used all 39,733 protein-coding genes as query against public protein databases, including NCBI non-redundant protein sequences (Nr), Swiss-Prot, Protein family (Pfam), Kyoto Encyclopedia of Genes and Genomes (KEGG), InterPro, and Gene Ontology (GO). In total, 39,646 genes (99.8%) could be annotated from these databases and 24,402 genes (61.41%) were supported by RNA-seq data. Furthermore, the gene number, gene length distribution, and exon length distribution were all comparable to those of other maize inbred lines and common crop species (Table 3).

Transfer RNA (tRNA) genes were predicted using tRNAscan-SE software v. 1.4 [48] (RRID:SCR_010835) with the default parameters. Ribosomal RNAs (rRNAs) were annotated on the basis of their homology levels with the rRNAs of several species of higher plants using BLASTN with an E-value of 1e-5. The microRNA (miRNA) and small nuclear RNA (snRNA) fragments were identified by searching the Rfam database v. 11.0 (RRID:SCR_007891) using INFERNAL v. 1.1 software

(RRID:SCR_011809) [49, 50]. Finally, 4,547 miRNAs, 5,963 tRNAs, 63,564 rRNAs, and 1,422 snRNAs were identified, which had average lengths of 126.79, 75.25, 309.47, and 132.10 bp, respectively (Table 4).

Comparative genomic analysis between Dan340 and other maize lines

We applied the OrthoMCL pipeline [51] to identify orthologous gene families among the four maize inbred lines, including Dan340, B73, Mo17, and SK. The longest protein from each gene was selected, and the proteins with a length less than 30 amino acids were removed. Subsequently, pairwise sequence similarities between all input protein sequences were calculated using BLASTP with an E value cut-off of 1×10^{-5}. Markov clustering (MCL) of the resulting similarity matrix was used to define the ortholog cluster structure of the proteins, using an inflation value (-I) of 1.5 (default setting in OrthoMCL). Next, comparative analyses were performed among Dan340, B73, Mo17, and SK (Fig.-6A).

Analysis of genes in the Dan340 genome, together with those from B73, Mo17 and SK, were clustered into 27,654 gene families. Of these, 15,690 families were shared among the four maize inbred lines, representing a core set of genes across these maize genomes. There were 1,806 genes from 359 gene families that were specific to Dan340, of which many had functional gene ontology annotations relating to “protein phosphorylation”, “single-organism catabolic process”, and “pheromone binding” (Fig. 6B). In KEGG functional enrichment, the most enriched pathway of Dan340

specific genes were “antifolate resistance”, “epithelial cell signaling in helicobacter pylori infection”, and “pentose and glucuronate interconversions” (Fig. 6C).

In addition, OrthoMCL was used to identify the core and dispensable gene sets on the basis of gene family. The gene families that were shared among the four inbred lines were defined as core gene families. Furthermore, gene families that were shared among three inbred lines, between two inbred lines, and those that were only present in one inbred line (private gene families) were also displayed in Fig 6D.

Genetic variation analysis

To investigate the genetic variations between Dan340 and other maize inbred lines, we used PBSV version 2.2.2 [52] to detect structural variations. First, PacBio reads of B73 and Mo17 were downloaded from MaizeGDB [53], and PacBio reads of SK were obtained from the National Genomics Data Center [54]. Next, subreads were aligned to the Dan340 reference genome assembled in this study using pbmm2 to generate a bam file. Then, Samtools v1.7 [55] (RRID:SCR_002105) was used to identify and split the bam file on the basis of chromosomes and scaffolds. Afterwards, pbsv discover was used to generate svsig result files for different chromosomes and scaffolds. The svsig files of multiple samples were then used together to perform the SV joint calling with “pbsv call”, and finally, the vcf files were obtained.

The high-quality Dan340 reference genome allowed us to identify large SVs presented in different maize inbred lines. By mapping the PacBio long-reads of B73 to Dan340 genome, we identified a total of 8,289 structural variations (length longer

than 500 bp) between the two representative maize genomes, including 1,653 insertions, 6,537 deletions, 36 inversions, and 63 duplications (Table 5). Furthermore, the structural variations presented in Mo17 and SK were also detected in this study (Table 6 & 7). This dataset provides abundant variation resources for molecular improvement and breeding in maize in the future.

Conclusions

We assembled the chromosome-level genome of the maize elite inbred line Dan340 using long CCS reads from the third-generation PacBio Sequel II sequencing platform, with scaffolding informed by chromosomal conformation capture (Hi-C). The final assembly of the Dan340 genome was 2,348.72 Mb, including 2,738 contigs and 2,315 scaffolds with N50 of 41.49 Mb and 215.35 Mb, respectively. Comparisons of the Dan340 genome with the reference genomes of three other common maize inbred lines identified 1,806 genes from 359 gene families that were specific to Dan340. In addition, we also obtained large numbers of structural variants between Dan340 and other maize inbred lines, and these may be underlying the mechanisms responsible for the phenotypic discrepancies between Dan340 and other maize varieties. Therefore, the assembly and annotation of this genome not only facilitates our understanding of the intraspecific genomic diversity in maize, but they also serve as novel resources for maize breeding improvement.

Data Availability

322 The raw sequence data have been deposited in NCBI under project accession No.

323 PRJNA795201. Data is also available in the *GigaScience* GigaDB repository [56].

324

325 **Abbreviations**

326 BUSCO: Benchmarking Universal Single-Copy Orthologs; Hi-C: chromosomal

327 conformation capture; CCS: circular consensus sequencing; HiFi: long high-fidelity;

328 CEGMA: Core Eukaryotic Genes Mapping Approach; LTR: long-terminal repeat;

329 LAI: long-terminal repeat assembly index; TEs: transposable elements; TRF: Tandem

330 Repeats Finder; EVM: EVidenceModeler; KEGG: Kyoto Encyclopedia of Genes and

331 Genomes; GO: Gene Ontology; MCL: Markov clustering; NCBI: National Center for

332 Biotechnology Information; PacBio: Pacific Biosciences; tRNA: Transfer RNA;

333 rRNAs: Ribosomal RNAs; miRNA: microRNA; snRNA: small nuclear RNA;.

334

335 **Competing Interests**

336 The authors declare that they have no competing interests.

337

338 **Authors' contributions**

339 F.W., J.Z. and H.Z. conceived the project; Y-K.Z. D.M. and Y.W. wrote, modified the

340 manuscript; L.X. G.F. and L.W. performed the data curation; Y.H. L.Z. Y-L.Z. and

341 Z.L. analyzed data. All authors read and approved the final manuscript.

342

Acknowledgements

This research was supported by grants from the special project for the construction of scientific and technological innovation capacity of Beijing Academy of Agriculture and Forestry Sciences (NO. KJCX20200305).

References

1. Yang N, Xu XW, Wang RR, Peng WL, Cai L, Song JM, Li W, Luo X, Niu L, Wang Y, Jin M, Chen L, Luo J, Deng M, Wang L, Pan Q, Liu F, Jackson D, Yang X, Chen LL, Yan J. (2017) Contributions of Zea mays subspecies mexicana haplotypes to modern maize. Nat Commun. 8, 1874.
2. Hubert B, Rosegrant M, Boekel MA, Ortiz R. (2010). The future of food: scenarios for 2050. Crop Sci. 50, 33-50.
3. Hake S, Ross-Ibarra J. (2015) Genetic, evolutionary and plant breeding insights from the domestication of maize. Elife. 4, e05861.
4. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, Chen W, Yan L, Higginbotham J, Cardenas M, Waligorski J, Applebaum E, Phelps L, Falcone J, Kanchi K, Thane T, Scimone A, Thane N, Henke J, Wang T, Ruppert J, Shah N, Rotter K, Hodges J, Ingenthron E, Cordes

M, Kohlberg S, Sgro J, Delgado B, Mead K, Chinwalla A, Leonard S, Crouse K, Collura K, Kudrna D, Currie J, He R, Angelova A, Rajasekar S, Mueller T, Lomeli R, Scara G, Ko A, Delaney K, Wissotski M, Lopez G, Campos D, Braidotti M, Ashley E, Golser W, Kim H, Lee S, Lin J, Dujmic Z, Kim W, Talag J, Zuccolo A, Fan C, Sebastian A, Kramer M, Spiegel L, Nascimento L, Zutavern T, Miller B, Ambroise C, Muller S, Spooner W, Narechania A, Ren L, Wei S, Kumari S, Faga B, Levy MJ, McMahan L, Van Buren P, Vaughn MW, Ying K, Yeh CT, Emrich SJ, Jia Y, Kalyanaraman A, Hsia AP, Barbazuk WB, Baucom RS, Brutnell TP, Carpita NC, Chaparro C, Chia JM, Deragon JM, Estill JC, Fu Y, Jeddeloh JA, Han Y, Lee H, Li P, Lisch DR, Liu S, Liu Z, Nagel DH, McCann MC, SanMiguel P, Myers AM, Nettleton D, Nguyen J, Penning BW, Ponnala L, Schneider KL, Schwartz DC, Sharma A, Soderlund C, Springer NM, Sun Q, Wang H, Waterman M, Westerman R, Wolfgruber TK, Yang L, Yu Y, Zhang L, Zhou S, Zhu Q, Bennetzen JL, Dawe RK, Jiang J, Jiang N, Presting GG, Wessler SR, Aluru S, Martienssen RA, Clifton SW, McCombie WR, Wing RA, Wilson RK. (2009) The B73 maize genome: complexity, diversity, and dynamics. Science. 326, 1112-1115.

5. Hirsch CN, Hirsch CD, Brohammer AB, Bowman MJ, Soifer I, Barad O, Shem-Tov D, Baruch K, Lu F, Hernandez AG, Fields CJ, Wright CL, Koehler K, Springer NM, Buckler E, Buell CR, de Leon N, Kaeppler SM, Childs KL, Mikel

MA. (2016) Draft Assembly of Elite Inbred Line PH207 Provides Insights into Genomic and Transcriptome Diversity in Maize. Plant Cell. 28, 2700-2714.

6. Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, Campbell MS, Stein JC, Wei X, Chin CS, Guill K, Regulski M, Kumari S, Olson A, Gent J, Schneider KL, Wolfgruber TK, May MR, Springer NM, Antoniou E, McCombie WR, Presting GG, McMullen M, Ross-Ibarra J, Dawe RK, Hastie A, Rank DR, Ware D. (2017) Improved maize reference genome with single-molecule technologies. Nature. 546, 524-527.

7. Sun S, Zhou Y, Chen J, Shi J, Zhao H, Zhao H, Song W, Zhang M, Cui Y, Dong X, Liu H, Ma X, Jiao Y, Wang B, Wei X, Stein JC, Glaubitz JC, Lu F, Yu G, Liang C, Fengler K, Li B, Rafalski A, Schnable PS, Ware DH, Buckler ES, Lai J. (2018) Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. Nat Genet. 50, 1289-1295.

8. Li C, Song W, Luo Y, Gao S, Zhang R, Shi Z, Wang X, Wang R, Wang F, Wang J, Zhao Y, Su A, Wang S, Li X, Luo M, Wang S, Zhang Y, Ge J, Tan X, Yuan Y, Bi X, He H, Yan J, Wang Y, Hu S, Zhao J. (2019) The HuangZaoSi Maize Genome Provides Insights into Genomic Variation and Improvement History of Maize. Mol Plant. 12, 402-409.

9. Lu F, Romay MC, Glaubitz JC, Bradbury PJ, Elshire RJ, Wang T, Li Y, Li Y, Semagn K, Zhang X, Hernandez AG, Mikel MA, Soifer I, Barad O, Buckler ES.

(2015) High-resolution genetic mapping of maize pan-genome sequence anchors. Nat Commun. 6, 6914.

10. Haberer G, Kamal N, Bauer E, Gundlach H, Fischer I, Seidel MA, Spannagl M, Marcon C, Ruban A, Urbany C, Nemri A, Hochholdinger F, Ouzunova M, Houben A, Schön CC, Mayer KFX. (2020) European maize genomes highlight intraspecies variation in repeat and gene content. Nat Genet. 52, 950-957.
11. Hufford MB, Seetharam AS, Woodhouse MR, Chougule KM, Ou S, Liu J, Ricci WA, Guo T, Olson A, Qiu Y, Della Coletta R, Tittes S, Hudson AI, Marand AP, Wei S, Lu Z, Wang B, Tello-Ruiz MK, Piri RD, Wang N, Kim DW, Zeng Y, O'Connor CH, Li X, Gilbert AM, Baggs E, Krasileva KV, Portwood JL 2nd, Cannon EKS, Andorf CM, Manchanda N, Snodgrass SJ, Hufnagel DE, Jiang Q, Pedersen S, Syring ML, Kudrna DA, Llaca V, Fengler K, Schmitz RJ, Ross-Ibarra J, Yu J, Gent JI, Hirsch CN, Ware D, Dawe RK. (2021) De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. Science. 373, 655-662.
12. Zhang R, Xu G, Li J, Yan J, Li H, Yang X. (2018) Patterns of genomic variation in Chinese maize inbred lines and implications for genetic improvement. Theor Appl Genet. 131, 1207-1221.
13. Yang N, Liu J, Gao Q, Gui S, Chen L, Yang L, Huang J, Deng T, Luo J, He L, Wang Y, Xu P, Peng Y, Shi Z, Lan L, Ma Z, Yang X, Zhang Q, Bai M, Li S, Li W, Liu L, Jackson D, Yan J. (2019) Genome assembly of a tropical maize inbred

425 line provides insights into structural variation and crop improvement. Nat Genet.
426 51, 1052-1059.
427 14. Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. Hi-C: a
428 comprehensive technique to capture the conformation of genomes. Methods.
429 2012; 58, 268-276.
430 15. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. (2021) Haplotype-resolved de
431 novo assembly using phased assembly graphs with hifiasm. Nat Methods. 18,
432 170-175.
433 16. PacBio SMRTLink. https://www.pacb.com/support/software-downloads/.
434 Accessed 2021 May 16.
435 17. Pbmarkdup. https://github.com/PacificBiosciences/pbmarkdup. Accessed 2021
436 May 16.
437 18. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A,
438 Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J. (2013) Nonhybrid,
439 finished microbial genome assemblies from long-read SMRT sequencing data.
440 Nature Methods. 10, 563-569.
441 19. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA,
442 Zeng Q, Wortman J, Young SK, Earl AM. (2014) Pilon: an integrated tool for
443 comprehensive microbial variant detection and genome assembly improvement.
444 PLoS One. 9(11):e112963.

20. Purge Haplotigs pipeline. https://bitbucket.org/mroachawri/purge_haplotigs/overview. Accessed 2021 Jun 20.

21. Wingett S, Ewels P, Furlan-Magaril M, Nagano T, Schoenfelder S, Fraser P, Andrews S. (2015) HiCUP: pipeline for mapping and processing Hi-C data. F1000Res. 4:1310.

22. Zhang X, Zhang S, Zhao Q, Ming R, Tang H. (2019) Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. Nature Plants. 5, 833-845.

23. Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL. (2016) Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. Cell Syst. 3, 99-101.

24. Li H, Durbin R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 26, 589-595.

25. Parra G, Bradnam K, Korf I. (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 23, 1061-1067.

26. Gertz EM, Yu YK, Agarwala R, Schäffer AA, Altschul SF. (2006) Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. BMC Biol. 4:41.

27. Birney E, Clamp M, Durbin R. (2004) GeneWise and Genomewise. Genome Res. 14, 988-995.

28. Alioto T, Blanco E, Parra G, Guigó R. (2018) Using geneid to Identify Genes. Curr Protoc Bioinformatics. 64, e56.

29. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 31, 3210-3212.

30. Ellinghaus D, Kurtz S, Willhoeft U. (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics. 9:18.

31. Xu Z, Wang H. (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 35, W265-8.

32. Ou S, Chen J, Jiang N. (2018) Assessing genome assembly quality using the LTR Assembly Index (LAI). Nucleic Acids Res. 46(21):e126.

33. RepeatModeler. http://www.repeatmasker.org/RepeatModeler/. Accessed 2021 Aug 6.

34. Price AL, Jones NC, Pevzner PA. (2005) De novo identification of repeat families in large genomes. Bioinformatics. 21 Suppl 1, i351-8.

35. RepeatMasker. http://www.repeatmasker.org/. Accessed 2021 Aug 16.

36. Benson G. (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27, 573-580.

37. Stanke M, Steinkamp R, Waack S, Morgenstern B. (2004) AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res. 32, W309-12.

487 38. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. (2006)
488 AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res.
489 34, W435-9.
490 39. Stanke M, Morgenstern B. (2005) AUGUSTUS: a web server for gene prediction
491 in eukaryotes that allows user-defined constraints. Nucleic Acids Res. 33,
492 W465-7.
493 40. Burge C, Karlin S. (1997) Prediction of complete gene structures in human
494 genomic DNA. J Mol Biol. 268, 78-94.
495 41. Majoros WH, Pertea M, Salzberg SL. (2004) TigrScan and GlimmerHMM: two
496 open source ab initio eukaryotic gene-finders. Bioinformatics. 20, 2878-2879.
497 42. Korf I. (2004) Gene finding in novel genomes. BMC Bioinformatics. 5, 59.
498 43. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. (2013)
499 TopHat2: accurate alignment of transcriptomes in the presence of insertions,
500 deletions and gene fusions. Genome Biol. 14, R36.
501 44. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H,
502 Salzberg SL, Rinn JL, Pachter L. (2012) Differential gene and transcript
503 expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat
504 Protoc. 7, 562-578.
505 45. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis
506 X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A,
507 Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N,

508 Regev A. (2011) Full-length transcriptome assembly from RNA-Seq data without
509 a reference genome. Nat Biotechnol. 29, 644-652.
510 46. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, Maiti
511 R, Ronning CM, Rusch DB, Town CD, Salzberg SL, White O. (2003) Improving
512 the Arabidopsis genome annotation using maximal transcript alignment
513 assemblies. Nucleic Acids Research. 31, 5654-5666.
514 47. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR,
515 Wortman JR. (2008) Automated eukaryotic gene structure annotation using
516 EVidenceModeler and the Program to Assemble Spliced Alignments. Genome
517 Biol. 9, R7.
518 48. Lowe TM, Eddy SR. (1997) tRNAscan-SE: a program for improved detection of
519 transfer RNA genes in genomic sequence. Nucleic Acids Res. 25, 955-964.
520 49. Nawrocki EP, Kolbe DL, Eddy SR. (2009) Infernal 1.0: inference of RNA
521 alignments. Bioinformatics. 25, 1335-137.
522 50. Nawrocki EP, Eddy SR. (2013) Infernal 1.1: 100-fold faster RNA homology
523 searches. Bioinformatics. 29, 2933-2935.
524 51. Li L, Stoeckert CJ Jr, Roos DS. (2003) OrthoMCL: identification of ortholog
525 groups for eukaryotic genomes. Genome Res. 13, 2178-2189.
526 52. PBSV. https://github.com/PacificBiosciences/pbsv. Accessed 2021 Sep 11.
527 53. MaizeGDB. https://maizegdb.org/download. 52. Accessed 2021 Aug 8.

528 54. National Genomics Data Center.
529 https://ngdc.cncb.ac.cn/search/?dbId=gsa&q=CRA001371. Accessed 2021 Aug
530 8.
531 55. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham
532 A, Keane T, McCarthy SA, Davies RM, Li H. Twelve years of SAMtools and
533 BCFtools. Gigascience. 2021 Feb 16;10(2):giab008. doi:
534 10.1093/gigascience/giab008.
535 56. Zhao Y; Wang Y; Ma D; Feng G; Huo Y; Liu Z; Zhou L; Zhang Y; Xu L; Wang
536 L; Zhao H; Zhao J; Wang F (2022): Supporting data for "A chromosome-level
537 genome assembly and annotation of a maize elite breeding line Dan340"
538 GigaScience Database. http://dx.doi.org/10.5524/102221

539

540 Figure legends

541 Figure 1. Ear appearances of the maize inbred lines Dan340, B73, Mo17, and SK.

542 Figure 2: Hi-C contact heat map displaying the inter- and intra-chromosomal
543 interactions in maize inbred line Dan340 genome.

544 Figure 3. Circos plot of genomic features. Outer-to-inner tracks indicate the following:
545 A, Chromosome numbers of Dan340 and B73; B, Repeat density; C, Histogram of
546 gene density distributions along the chromosomes; D, Histogram of GC content
547 distributions along the chromosomes; E, Syntenic relationships of gene pairs between
548 Dan340 and B73 genomes identified using the best-hit method.

549 Figure 4. Genome-wide LTR Assembly Index (LAI) scores for Dan340, B73, Mo17
550 and SK.

551 Figure 5. Pairwise comparison of genome sequences using a dot plot between the
552 Dan340 line and B73 (23,350 gene pairs), Mo17 (21,913 gene pairs), SK (23,016
553 gene pairs). The horizontal axis represents the target species, the vertical axis
554 represents the reference species, C1 – C10 represent the respective chromosomes 1 –
555 10, 0 – 35 k represent the chromosome length scale marks, which mainly reflect the
556 lengths of the chromosomes, and a point represents a pair of common genes.

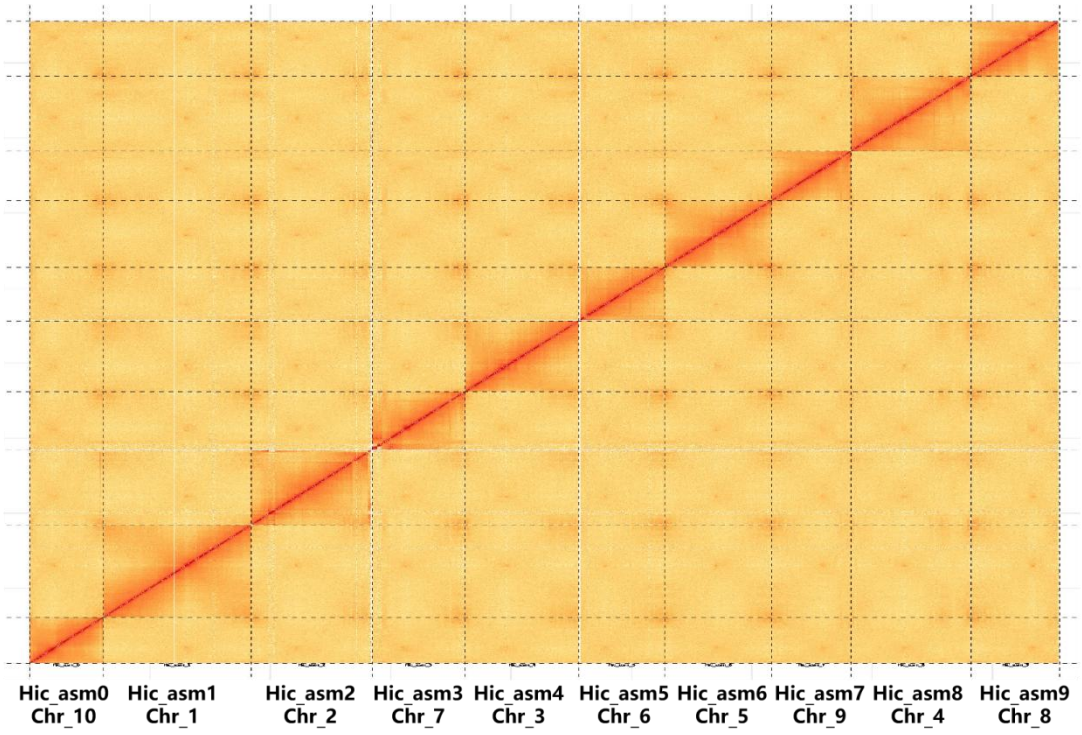
557 Figure 6. Gene family analyses and core- and pan-genomes of maize. A, Comparisons
558 of gene families in Dan340, B73, Mo17, and SK. The Venn diagram illustrates shared
559 and unique gene families among the four maize inbred lines. B, Gene ontology
560 enrichment analysis of Dan340-specific genes. C, Kyoto Encyclopedia of Genes and
561 Genomes analysis of Dan340-specific genes. D, Core- and pan-genome of maize. The
562 histograms show the core-gene clusters (shared in all four genomes), dispensable gene
563 clusters (present in three or two genomes) and specific gene clusters (present only in
564 one genome).

565 Figure 1

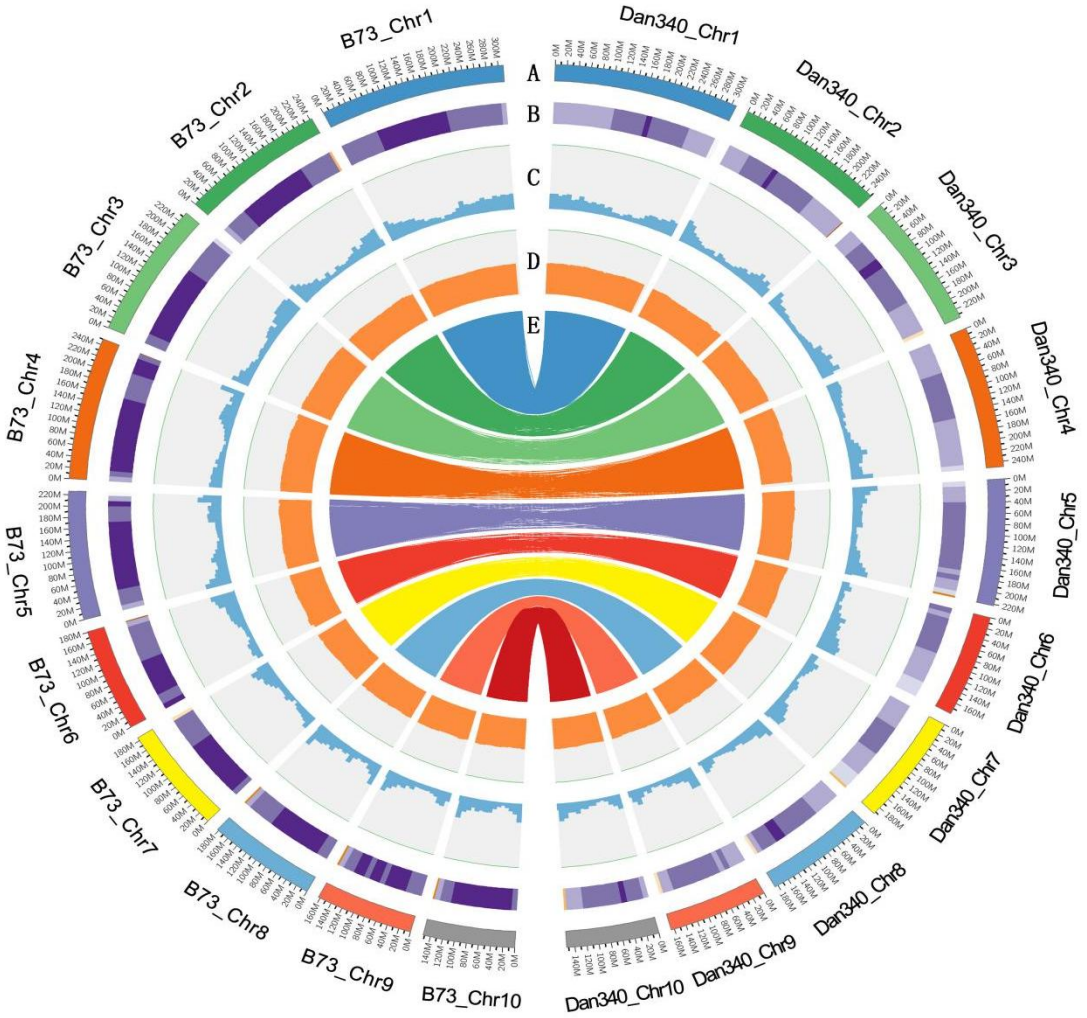
566

567

Figure 2: Hi-C contact heat map displaying the inter- and intra-chromosomal interactions in maize inbred line Dan340 genome.



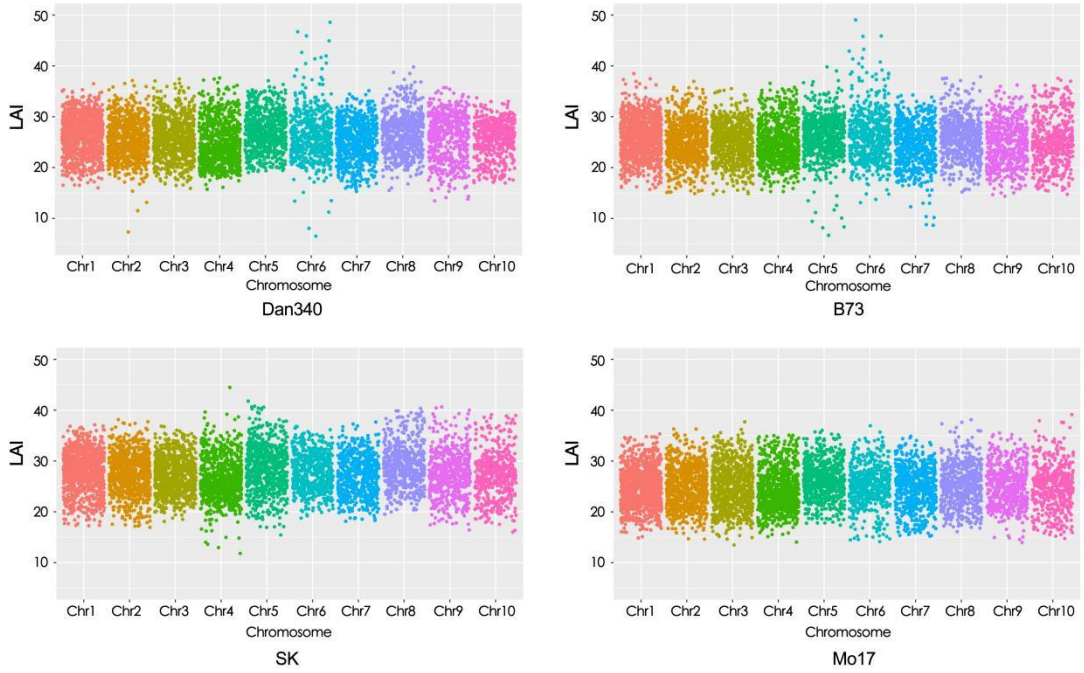
582 Figure 3



583

584

585 Figure 4

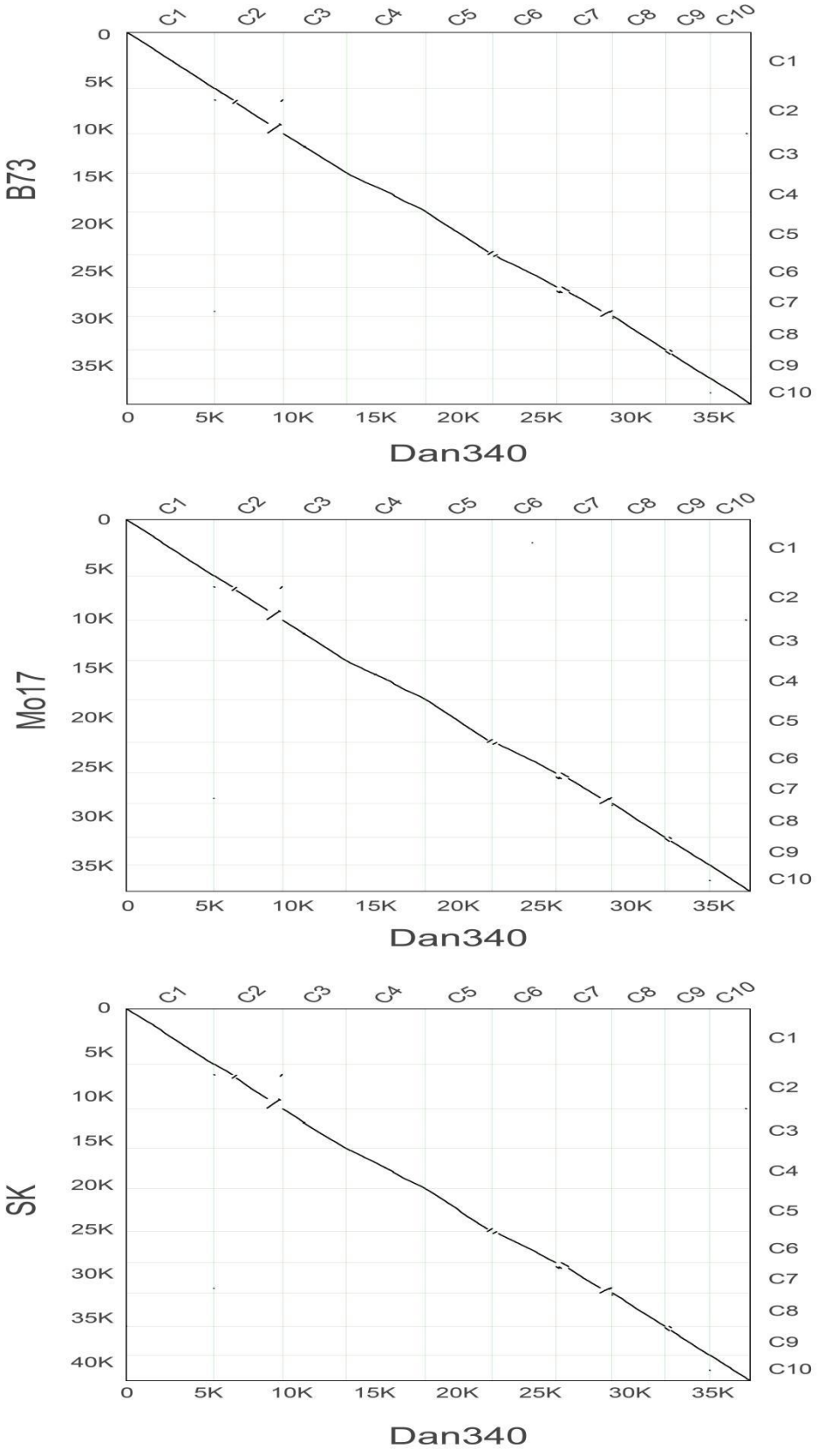


586

587

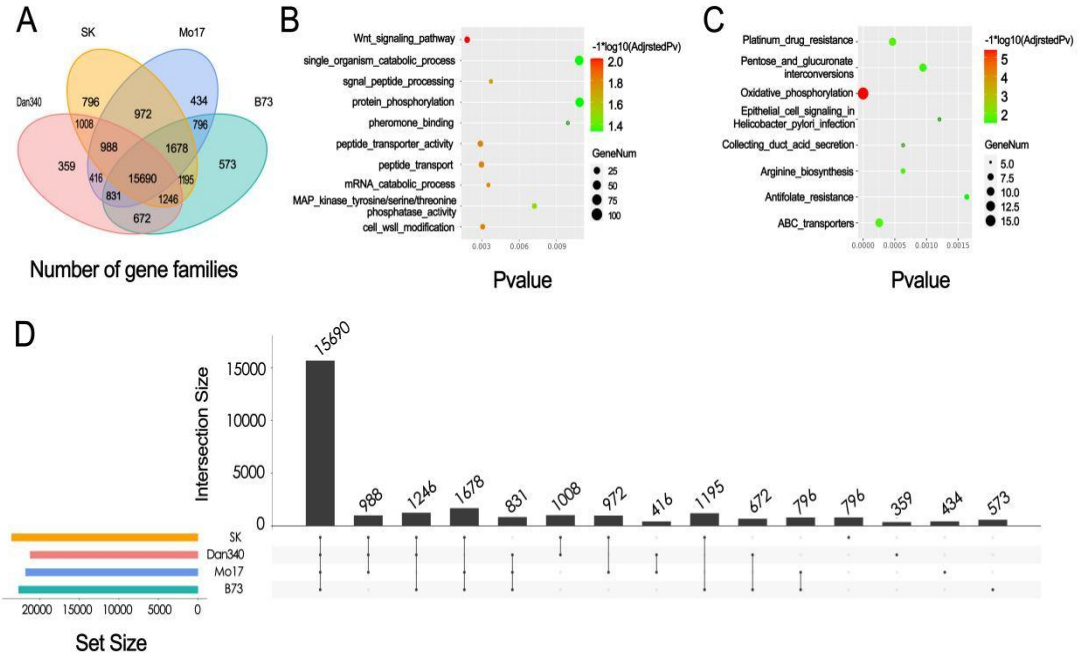
588

589 Figure 5



590

591 Figure 6



592

593

594

595 Table 1 Genome assembly and annotation statistics for the four tested maize inbred

596 lines.

Genomic features	Dan340	B73	Mo17	SK
assembled genome size (bp)	2,348,678,871	2,182,075,994	2,104,465,715	2,161,392,594
Number of scaffolds	2315	687	2203	671
Total length of scaffolds (Mb)	2,348.72	2,182	2,182	2,162
Scaffold N50	222,765,871	226,353,449	220,382,597	73,237,962
Number of contigs	2,738	1,395	9,040	1,090
Total length of contigs (Mb)	2,144,444	2,178,268	2,147,495	2,150,874
Contig N50	45,109,016	47,037,903	1,491,782	15,776,512
Number of genes	39,733	39,756	38,620	43,271

597

598

599 Table 2. LAI scores of the four tested maize inbred lines.

Lines	**~~D340~~ Dan340**	**B73**	**Mo17**	**SK**
LAI	25.13	24.94	24.45	27.12

600

601

602 Table 3 Summary statistics of annotated protein-coding genes in Dan340 and other

603 maize inbred lines and common crop species.

Species	**Number**	**Average transcript length(bp)**	**Average CDS length(bp)**	**Average exons per gene**	**Average exon length(bp)**	**Average intron length(bp)**
Dan340	39,733	3,793.47	1,140.91	4.69	243.47	719.61
B73	39756	3511.78	1102.11	4.58	240.64	673.10
Mo17	38620	3362.68	1140.26	4.69	242.98	601.83
SK	42942	3857.18	1179.17	4.83	243.93	698.48
Hvu	24,286	2,116.13	1,093.77	4.1	267.02	330.19
Osa	35,679	2,165.58	991.55	3.78	262.57	422.87
Sbi	34,008	2,626.44	1,164.14	4.31	270.09	441.76
Sit	27,233	2,982.22	1,336.29	5.14	260.2	397.98
Tae	103,539	3,087.61	1,277.31	4.51	283.23	515.78

604 Abbreviations: Hvu: *Hordeum vulgare*; Osa: *Oryza sativa*; Sbi: *Sorghum bicolor*; Sit:

605 *Setaria italica*; Tae: *Triticum aestivum*.

606 Table 4. Annotation statistics of non-coding RNAs in the Dan340 genome using

607 different databases.

	Type	Copy(w*)	Average length(bp)	Total length(bp)
miRNA		4,547	126.79	576,516	0.024546
tRNA		5,963	75.25	448,705	0.019104
rRNA		63,564	309.47	19,671,118	0.84
rRNA	18S	6,607	1,727.38	11,412,778	0.49
rRNA	28S	25,188	143.61	3,617,315	0.15
rRNA	5.8S	25,181	153.48	3,864,710	0.16
rRNA	5S	6,588	117.84	776,315	0.033053
snRNA		1,422	132.1	187,845	0.007998
snRNA	CD-box	647	103.2	66,768	0.002843
snRNA	HACA-box	123	126.27	15,531	0.000661
snRNA	splicing	651	161.72	105,278	0.004482

608

609

610

611

612

613

614 Table 5 Structural variations between Dan340 and B73.

Chr. Number	Insertion		Deletion		Inversion		Duplication	
	Number	Length (bp)	Number	Length (bp)	Number	Length (bp)	Number	Length (bp)
1	266	432723	998	13194452	7	120136	8	311433
2	183	316530	743	10075332	7	177706	9	261774
3	201	381536	838	11475573	5	24457	7	37754
4	183	303142	698	9809218	2	113807	7	255958
5	181	330690	651	9519227	3	108429	6	121490
6	122	252117	527	7723482	3	55199	8	300610
7	140	226942	512	7146020	2	49011	4	122886
8	140	204697	577	7720113	1	80400	3	42148
9	121	219627	569	7539785	3	17955	9	114595
10	116	209091	424	6256310	3	130724	2	40306

615

616

617

618

619

620

621

622

623

624

625

626

627

628 Table 6 Structural variations between Dan340 and Mo17.

Chr. Number	Insertion		Deletion		Inversion		Duplication	
	Number	Length (bp)	Number	Length (bp)	Number	Length (bp)	Number	Length (bp)
1	171	226204	1049	14591535	10	223256	6	157653
2	102	164536	668	9437198	4	40341	5	149456
3	121	196809	767	11363033	0	0	4	26512
4	96	146230	628	8996741	3	50713	10	169360
5	102	147883	660	9449523	3	190422	7	85046
6	75	101725	567	8423089	3	130693	3	20573
7	84	119614	521	7784957	4	159203	6	275006
8	83	123397	603	8751739	2	30781	5	214099
9	61	83589	529	7194721	1	10328	3	55228
10	75	126350	497	6832836	5	227243	4	159642

629

630

631

632

633

634

635

636

637

638

639

640

Table 7 Structural variations between Dan340 and SK.

Chr. Number	Insertion		Deletion		Inversion		Duplication	
	Number	Length (bp)	Number	Length (bp)	Number	Length (bp)	Number	Length (bp)
1	244	355348	1226	17656467	15	228724	16	404487
2	133	220966	842	12542190	6	244805	5	76672
3	173	286569	894	12961969	4	7657	6	96601
4	211	338458	1012	14231859	3	70762	14	315182
5	160	248233	786	10660145	8	280670	5	206286
6	93	118612	588	8923936	4	55606	2	1760
7	115	175971	622	8865791	1	30856	9	233675
8	143	210185	685	10443437	2	3061	6	152412
9	114	186005	647	8682441	4	9595	7	37664
10	109	159668	559	8726188	4	132337	4	95237