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,*]

[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 article.

[*] **Correspondence**: Han Zhao (zhaohan@jaas.ac.cn), Jiuran Zhao (maizezhao@126.com), Fengge Wang (fenggewangmaize@126.com)

Key findings

1. The final assembly of Dan340 genome was 2,348.72 Mb, including 2,738 contigs and 2,315 scaffolds with a N50 of 41.49 Mb and 215.35 Mb, respectively.
2. The percent of reads mapped to the reference genome was up to 97.48%.
3. The results showed that 96.84% of the plant single-copy orthologues were complete. Complete single-copy and multi copy genes accounted for 87.36% and 9.47% of the genes, respectively. Taken together, these results indicated that our Dan340 genome assembly presented high quality and coverage.

Keywords

Dan340, comparative genomic analysis, tandem duplication, pedigree analysis

Abstract

The maize cultivar Dan340 is an excellent backbone inbred line of Luda Red Cob Group with several desirable characters, such as disease resistance, lodging resistance, high combining ability, wide adaptability and so on. In this study, we constructed a high-quality chromosome-level reference genome for Dan340 by combining PacBio long HiFi sequencing, Illumina short reads and chromosomal conformational capture (Hi-C) sequencing reads. The final assembly of Dan340 genome was 2,348.72 Mb, including 2,738 contigs and 2,315 scaffolds with a N50 of 41.49 Mb and 215.35 Mb, respectively. The percent of reads mapped to the reference genome was up to 97.48%. The completeness and continuity of the genome were comparable to those of other important maize inbred lines. The assembly and annotation of this genome will not only facilitate our understanding about intraspecific genome diversity in maize, but also provides a novel resource for maize breeding improvement.

Background

Maize (*Zea mays* ssp. *mays* L.) is one of the most important crops grown worldwide for food, forage, and biofuel, with an annual production of more than 1 billion tons (Yang et al., 2017). With the rapid growth of human population and economic demand, it has been predicted that maize will account for 45% of total cereal demand by the year 2050 (Hubert et al., 2010). In addition, it is also an important model organism for fundamental research in genetics and genomics (Hake and Ross-Ibarra, 2015).

Several reference genome of common maize inbred line in breeding have been released since 2009 (Schnable et al., 2009; Hirsch et al., 2016; Jiao et al., 2017; Sun et

al., 2018; Li et al., 2019). The maize genome exhibits high levels of genetic diversity among different inbred lines (Lu et al., 2015; Yang et al., 2017; Sun et al., 2018). The maize cultivar Dan340 is an excellent backbone inbred line of Luda Red Cob Group with several desirable characters, such as disease resistance, lodging resistance, high combining ability, wide adaptability and so on (Zhang et al., 2018).

In the present study, we constructed a high-quality chromosome-level reference genome for Dan340 by combining PacBio long HiFi sequencing, Illumina short reads and chromosomal conformational capture (Hi-C) sequencing reads. The completeness and continuity of the genome were comparable to those of other important maize inbred lines, B73 (Schnable et al., 2009), Mo17 (Sun et al., 2018), PH207 (Hirsch et al., 2016), and HZS (Li et al., 2019). The assembly and annotation of this genome will not only facilitate our understanding about intraspecific genome diversity in maize, but also provides a novel resource for maize breeding improvement.

Plant material and DNA sequencing

Inbred line Dan340 was selected for genome sequencing and assembly because it is an elite maize cultivar, which plays an important role in maize breeding and genetic research. The plants were grown at 25°C in a greenhouse of Beijing Academy of Agriculture and Forestry Sciences, Beijing, China. Fresh and tender leaves were harvested from the best-growing individual and immediately frozen in liquid nitrogen, followed by preservation at -80 °C in the laboratory prior to DNA 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. PacBio circular consensus sequencing (CCS) libraries were prepared using SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences Ref. No. 101-685-400), following the manufacturer's protocols, and subsequently sequenced on the PacBio sequel II platform (Pacific Biosciences, Menlo Park, CA, USA). Consensus reads (CCS reads, also termed HiFi reads) were generated using ccs software v.3.0.0 (https://github.com/pacificbiosciences/unanimity/) with the parameter '--min-passes 3 --min-rq 0.99'. The total CCS yield was 63.54 Gb, with a mean read length of 15.60 kb.

In addition, one Illumina paired-end sequencing library with an insert size of 350 bp was generated using NEB Next Ultra DNA Library Prep Kit (NEB, USA) following manufacturer's protocol, and subsequently sequenced using an Illumina HiSeq X Ten platform (Illumina, San Diego, CA, USA) in Novogene Bioinformatics Institute, Beijing, China. Approximately 80.66 Gb (~20X) of Illumina sequencing data were obtained.

One Hi-C library was constructed using young leaves following the published procedures with certain modifications (Belton et al., 2012). In brief, approximately 5-g leaf samples were cut into minute pieces and cross-linked by 4% formaldehyde solution at room temperature in a vacuum for 30 min. Then, the sample was mixed

with excess 2.5 M glycine to quench the crosslinking reaction for 5 min and then placed on ice for 15 min. The cross-linked DNA was extracted and then digested with MboI restriction enzyme. The sticky ends of the digested fragments were biotinylated and proximity ligated to form ligation junctions that were enriched for and then ultrasonically sheared to a size of 200-600 bp. The biotin-labelled DNA fragments were pulled down and ligated with Illumina paired-end adapters and then amplified by PCR to produce the Hi-C sequencing library. The library was sequenced using an Illumina HiSeq X Ten platform (Illumina, San Diego, CA, USA).

Genome assembly

To obtain a high quality genome assembly of Dan340, we employed both PacBio HiFi sequences and Illumina short-read, with scaffolding informed by chromosomal conformation capture (Hi-C). The assembly was performed in a stepwise fashion. First, de novo assembly of the long CCS reads generated from the PacBio SMRT Sequencing was performed using Hifiasm (Cheng et al., 2021). In brief, we first selected the longest 30X of subreads as seeds to do error correction. Then, error-corrected reads were aligned to each other and assembled into genomic contigs using Hifiasm with default parameters.

The produced primary contigs (p-contigs) were then polished using Quiver (Chin et al., 2013) by aligning SMRT reads.

Then, Pilon (Walker et al., 2014) 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

(https://bitbucket.org/mroachawri/purge_haplotigs/overview) was used to remove redundant sequences formed due to heterozygosity.

For Hi-C reads, to avoid reads with artificial bias, we removed the following type of reads: (a) Reads with ≥ 10% unidentified nucleotides (N); (b) Reads with > 10 nt aligned to the adapter, allowing ≤ 10% mismatches; (c) Reads with > 50% bases having phred quality < 5. The filtered Hi-C reads were aligned against the contig assemblies with BWA (version 0.7.8). Reads were excluded from subsequent analysis 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 is tabulated. ALLHiC v0.8.12 (Zhang et al., 2019) was used in simple diploid mode to scaffold the genome and optimize the ordering and orientation of each clustered group, producing a chromosomal level assembly. Juicebox Assembly Tools v1.9.8 (https://github.com/aidenlab/Juicebox) (Durand et al., 2016) was finally used to visualize and manually correct the large-scale inversions and translocations to obtain the final pseudo-chromosomes.

The final assembly of Dan340 genome was 2,348.72 Mb, including 2,738 contigs and 2,315 scaffolds with a N50 of 41.49 Mb and 215.35 Mb, respectively.

Evaluation of assembly quality

We assessed the quality of the assembly by three independent methods. First, the short reads obtained from the Illumina sequencing data were aligned to the final assembly with BWA version 0.7.8 (http://bio-bwa.sourceforge.net/) (Li and Durbin, 2010). Our result showed that the percent of reads mapped to the reference genome

was up to 97.48%. Second, CEGMA version 2.5 (Parra et al., 2007) was used to access the integrity of 248 conserved core genes for eukaryotes. Third, the BUSCO (Benchmarking Universal Single-Copy Otheologs, Simao et al., 2015) database embryophyta_odb10 was used to evaluate the completeness of gene regions, which including 1614 conserved core genes. The results showed that 96.84% of the plant single-copy orthologues were complete. Complete single-copy and multicopy genes accounted for 87.36% and 9.47% of the genes, respectively. Taken together, these results indicated that our Dan340 genome assembly presented high quality and coverage.

Genome annotation

Repeat sequences of the Dan340 genome were annotated using both ab initio and homolog-based search methods. In ab initio prediction, RepeatModeler v1.0.8 (http://www.repeatmasker.org/RepeatModeler/), RepeatScout version 1.0.5 (Price et al., 2005), and LTR_FINDER version 1.07 (Xu and Wang, 2007) were used to discover TEs and to build TEs library. A integrate TEs library and a known repeat library (Repbase V15.02, homolog-based) were subjected to RepeatMasker v3.3.0 (http://www.repeatmasker.org/) to predict TEs. For homolog-based predictions, RepeatProteinMask was performed to detect TEs in the genome by comparing against to the TE protein database. Tandem repeats were ascertained in the genome using Tandem Repeats Finder (TRF, version 4.07b) (Benson, 1999).

All repetitive regions except tandem repeats were soft-masked for protein-coding gene annotation. Five ab initio gene prediction programs were used to predict genes,

158 including Augustus v3.0.2 (Stanke et al., 2004, 2006; Stanke and Morgenstern, 2005),
159 Genscan v1.0 (http://hollywood.mit.edu/GENSCAN.html) (Burge and Karlin S, 1997),
160 Geneid v1.4 (Alioto et al., 2018), GlimmerHMM v3.0.2 (Majoros et al., 2004) and
161 SNAP v2013-02-16 (Korf ,2004). Protein sequences of 5 homologous species
162 (*Sorghum bicolor*, *Setaria italica*, *Hordeum vulgare*, *Triticum aestivum*, *Oryza sativa*)
163 were downloaded from Ensembl and NCBI. Homologous sequences were aligned
164 against to the genome using TBLASTN (E-value 1E-05). Genewise version 2.2.0
165 (Birney et al.,2004) was employed to predict gene models based on the sequences
166 alignment results. For RNA-seq prediction, transcriptomic data from six tissues (stem,
167 endosperm, embryo, bract, silk and ear tip) were aligned to the genome using Tophat
168 version 2.0.13 (http://ccb.jhu.edu/software/tophat/manual.shtml) (Kim et al., 2013) to
169 identify exons region and splice positions. The alignment results were then used as
170 input for cufflinks version 2.1.1 (http://cole-trapnell-lab.github.io/cufflinks/manual/)
171 (Trapnell et al., 2012) to assemble transcripts to the gene models. In addition,
172 RNA-seq data were assembled by Trinity version 2.1.1 (Grabherr et al., 2011),
173 creating several pseudo-ESTs. These pseudo-ESTs were also mapped to the
174 assembled genome by BLAT and gene models were predicted using PASA (Haas et
175 al., 2003). A weighted and non-redundant gene set was generated by
176 EVidenceModeler (EVM, version 1.1.1) (Haas et al., 2008), which merged all genes
177 models predicted by the above three approaches. Finally, PASA was used to adjust

178 the gene models generated by EVM. The gene number, gene length distribution, and
179 exon length distribution were all comparable to those of other maize inbred line.

180 **Retrotransposons and transposable analysis**

181

182 **Comparative genomic analysis between Dan340 and other maize lines**

183 Analyses of genes in the Dan340 genome together with those from B73 (Schnable et
184 al., 2009), Mo17 (Sun et al., 2018), PH207 (Hirsch et al., 2016), and HZS (Li et al.,
185 2019).

186 **Conclusions**

187 1. We assembled the chromosome-level genome of maize elite inbred line Dan340
188 based on long CCS reads from the third-generation PacBio Sequel II sequencing
189 platform, with scaffolding informed by chromosomal conformation capture (Hi-C).
190 2. The final assembly of Dan340 genome was 2,348.72 Mb, including 2,738 contigs
191 and 2,315 scaffolds with a N50 of 41.49 Mb and 215.35 Mb, respectively.
192 3. The assembly and annotation of this genome will not only facilitate our
193 understanding about intraspecific genome diversity in maize, but also serves as a
194 novel resource for maize breeding improvement.

195

196 **References**

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

199 Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. Hi-C: a
200 comprehensive technique to capture the conformation of genomes. Methods. 2012; 58,
201 268-276.
202 Benson G. (1999) Tandem repeats finder: a program to analyze DNA sequences.
203 Nucleic Acids Res. 27, 573-580.
204 Birney E, Clamp M, Durbin R. (2004) GeneWise and Genomewise. Genome Res. 14,
205 988-995.
206 Burge C, Karlin S. (1997) Prediction of complete gene structures in human genomic
207 DNA. J Mol Biol. 268, 78-94.
208 Cheng H, Concepcion GT, Feng X, Zhang H, Li H. (2021) Haplotype-resolved de
209 novo assembly using phased assembly graphs with hifiasm. Nat Methods. 18,
210 170-175.
211 Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A,
212 Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J. (2013) Nonhybrid,
213 finished microbial genome assemblies from long-read SMRT sequencing data. Nature
214 Methods. 10, 563-569.
215 Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL.
216 (2016) Juicebox Provides a Visualization System for Hi-C Contact Maps with
217 Unlimited Zoom. Cell Syst. 3, 99-101.
218 Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X,
219 Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind

220 N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A.
221 (2011) Full-length transcriptome assembly from RNA-Seq data without a reference
222 genome. Nat Biotechnol. 29, 644-652.
223 Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, Maiti R,
224 Ronning CM, Rusch DB, Town CD, Salzberg SL, White O. (2003) Improving the
225 Arabidopsis genome annotation using maximal transcript alignment assemblies.
226 Nucleic Acids Research. 31, 5654-5666.
227 Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR,
228 Wortman JR. (2008) Automated eukaryotic gene structure annotation using
229 EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 9,
230 R7.
231 Hake S, Ross-Ibarra J. (2015) Genetic, evolutionary and plant breeding insights from
232 the domestication of maize. Elife. 4, e05861.
233 Hubert B, Rosegrant M, Boekel MA, Ortiz R. (2010). The future of food: scenarios
234 for 2050. Crop Sci. 50, 33-50.
235 Hirsch CN, Hirsch CD, Brohammer AB, Bowman MJ, Soifer I, Barad O, Shem-Tov
236 D, Baruch K, Lu F, Hernandez AG, Fields CJ, Wright CL, Koehler K, Springer NM,
237 Buckler E, Buell CR, de Leon N, Kaeppler SM, Childs KL, Mikel MA. (2016) Draft
238 Assembly of Elite Inbred Line PH207 Provides Insights into Genomic and
239 Transcriptome Diversity in Maize. Plant Cell. 28, 2700-2714.

240 Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, Campbell MS, Stein JC, Wei X,
241 Chin CS, Guill K, Regulski M, Kumari S, Olson A, Gent J, Schneider KL,
242 Wolfgruber TK, May MR, Springer NM, Antoniou E, McCombie WR, Presting GG,
243 McMullen M, Ross-Ibarra J, Dawe RK, Hastie A, Rank DR, Ware D. (2017)
244 Improved maize reference genome with single-molecule technologies. Nature. 546,
245 524-527.
246 Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. (2013) TopHat2:
247 accurate alignment of transcriptomes in the presence of insertions, deletions and gene
248 fusions. Genome Biol. 14, R36.
249 Korf I. (2004) Gene finding in novel genomes. BMC Bioinformatics. 5, 59.
250 Li C, Song W, Luo Y, Gao S, Zhang R, Shi Z, Wang X, Wang R, Wang F, Wang J,
251 Zhao Y, Su A, Wang S, Li X, Luo M, Wang S, Zhang Y, Ge J, Tan X, Yuan Y, Bi X,
252 He H, Yan J, Wang Y, Hu S, Zhao J. (2019) The HuangZaoSi Maize Genome
253 Provides Insights into Genomic Variation and Improvement History of Maize. Mol
254 Plant. 12, 402-409.
255 Li H, Durbin R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler
256 transform. Bioinformatics. 26, 589-595.
257 Lu F, Romay MC, Glaubitz JC, Bradbury PJ, Elshire RJ, Wang T, Li Y, Li Y,
258 Semagn K, Zhang X, Hernandez AG, Mikel MA, Soifer I, Barad O, Buckler ES.
259 (2015) High-resolution genetic mapping of maize pan-genome sequence anchors. Nat
260 Commun. 6, 6914.

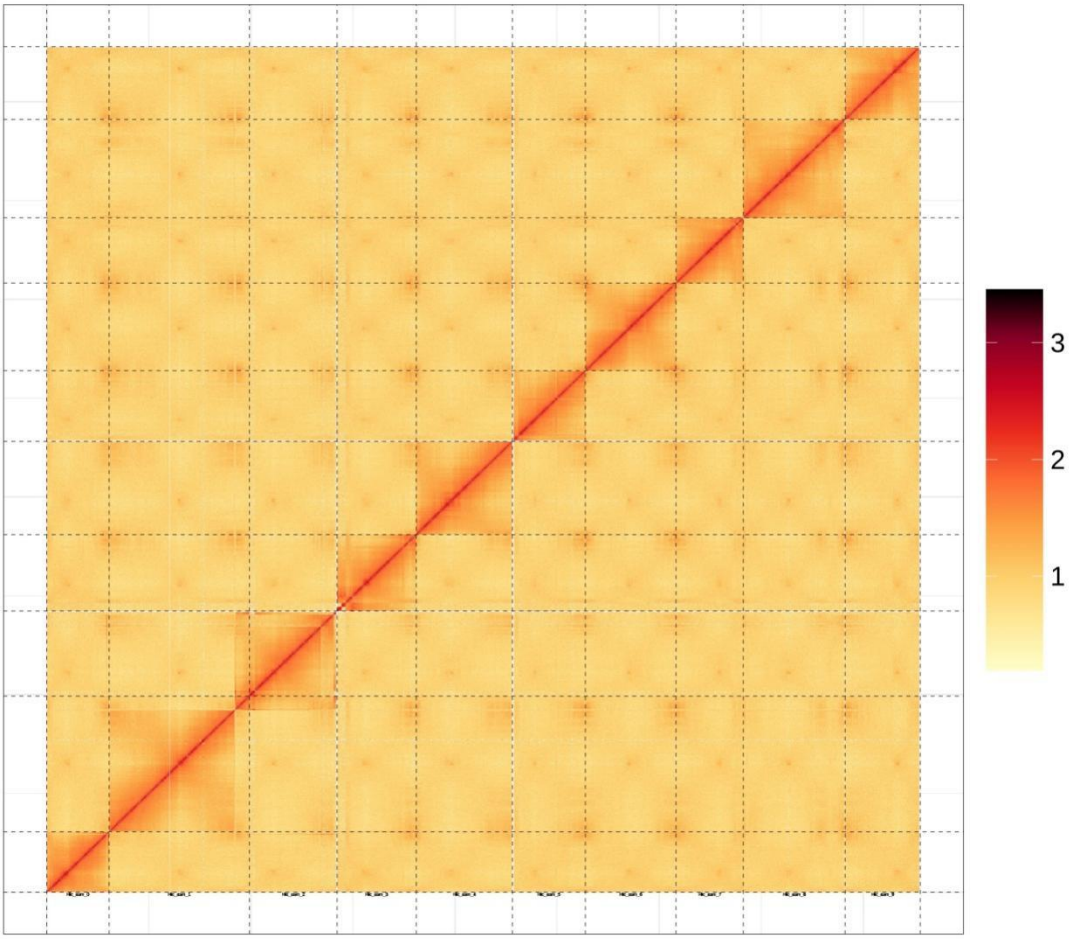
261 Majoros WH, Pertea M, Salzberg SL. (2004) TigrScan and GlimmerHMM: two open
262 source ab initio eukaryotic gene-finders. Bioinformatics. 20, 2878-2879.
263 Parra G, Bradnam K, Korf I. (2007) CEGMA: a pipeline to accurately annotate core
264 genes in eukaryotic genomes. Bioinformatics. 23, 1061-1067.
265 Price AL, Jones NC, Pevzner PA. (2005) De novo identification of repeat families in
266 large genomes. Bioinformatics. 21 Suppl 1, i351-8.
267 Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J,
268 Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C,
269 Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K,
270 Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, Chen
271 W, Yan L, Higginbotham J, Cardenas M, Waligorski J, Applebaum E, Phelps L,
272 Falcone J, Kanchi K, Thane T, Scimone A, Thane N, Henke J, Wang T, Ruppert J,
273 Shah N, Rotter K, Hodges J, Ingenthron E, Cordes M, Kohlberg S, Sgro J, Delgado B,
274 Mead K, Chinwalla A, Leonard S, Crouse K, Collura K, Kudrna D, Currie J, He R,
275 Angelova A, Rajasekar S, Mueller T, Lomeli R, Scara G, Ko A, Delaney K, Wissotski
276 M, Lopez G, Campos D, Braidotti M, Ashley E, Golser W, Kim H, Lee S, Lin J,
277 Dujmic Z, Kim W, Talag J, Zuccolo A, Fan C, Sebastian A, Kramer M, Spiegel L,
278 Nascimento L, Zutavern T, Miller B, Ambroise C, Muller S, Spooner W, Narechania
279 A, Ren L, Wei S, Kumari S, Faga B, Levy MJ, McMahan L, Van Buren P, Vaughn
280 MW, Ying K, Yeh CT, Emrich SJ, Jia Y, Kalyanaraman A, Hsia AP, Barbazuk WB,
281 Baucom RS, Brutnell TP, Carpita NC, Chaparro C, Chia JM, Deragon JM, Estill JC,

282 Fu Y, Jeddeloh JA, Han Y, Lee H, Li P, Lisch DR, Liu S, Liu Z, Nagel DH, McCann
283 MC, SanMiguel P, Myers AM, Nettleton D, Nguyen J, Penning BW, Ponnala L,
284 Schneider KL, Schwartz DC, Sharma A, Soderlund C, Springer NM, Sun Q, Wang H,
285 Waterman M, Westerman R, Wolfgruber TK, Yang L, Yu Y, Zhang L, Zhou S, Zhu
286 Q, Bennetzen JL, Dawe RK, Jiang J, Jiang N, Presting GG, Wessler SR, Aluru S,
287 Martienssen RA, Clifton SW, McCombie WR, Wing RA, Wilson RK. (2009) The
288 B73 maize genome: complexity, diversity, and dynamics. Science. 326, 1112-1115.
289 Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. (2015)
290 BUSCO: assessing genome assembly and annotation completeness with single-copy
291 orthologs. Bioinformatics. 31, 3210-3212.
292 Sun S, Zhou Y, Chen J, Shi J, Zhao H, Zhao H, Song W, Zhang M, Cui Y, Dong X,
293 Liu H, Ma X, Jiao Y, Wang B, Wei X, Stein JC, Glaubitz JC, Lu F, Yu G, Liang C,
294 Fengler K, Li B, Rafalski A, Schnable PS, Ware DH, Buckler ES, Lai J. (2018)
295 Extensive intraspecific gene order and gene structural variations between Mo17 and
296 other maize genomes. Nat Genet. 50, 1289-1295.
297 Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. (2006)
298 AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 34,
299 W435-9.
300 Stanke M, Morgenstern B. (2005) AUGUSTUS: a web server for gene prediction in
301 eukaryotes that allows user-defined constraints. Nucleic Acids Res. 33, W465-7.

302 Stanke M, Steinkamp R, Waack S, Morgenstern B. (2004) AUGUSTUS: a web server
303 for gene finding in eukaryotes. Nucleic Acids Res. 32, W309-12.
304 Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL,
305 Rinn JL, Pachter L. (2012) Differential gene and transcript expression analysis of
306 RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 7, 562-578.
307 Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA,
308 Zeng Q, Wortman J, Young SK, Earl AM. (2014) Pilon: an integrated tool for
309 comprehensive microbial variant detection and genome assembly improvement. PLoS
310 One. 9(11):e112963.
311 Xu Z, Wang H. (2007) LTR_FINDER: an efficient tool for the prediction of
312 full-length LTR retrotransposons. Nucleic Acids Res. 35, W265-8.
313 Yang N, Xu XW, Wang RR, Peng WL, Cai L, Song JM, Li W, Luo X, Niu L, Wang
314 Y, Jin M, Chen L, Luo J, Deng M, Wang L, Pan Q, Liu F, Jackson D, Yang X, Chen
315 LL, Yan J. (2017) Contributions of Zea mays subspecies mexicana haplotypes to
316 modern maize. Nat Commun. 8, 1874.
317 Zhang R, Xu G, Li J, Yan J, Li H, Yang X. (2018) Patterns of genomic variation in
318 Chinese maize inbred lines and implications for genetic improvement. Theor Appl
319 Genet. 131, 1207-1221.
320 Zhang X, Zhang S, Zhao Q, Ming R, Tang H. (2019) Assembly of allele-aware,
321 chromosomal-scale autopolyploid genomes based on Hi-C data. Nature Plants. 5,
322 833-845.

323

324 Figure 1Hi-C interaction heat map for maize inbred line Dan340 genome.



325

326

327 Figure 2 Summary of Dan340 genomic features.

328

329

330 Figure 3 Comparison of the gene families among major maize inbred lines.

331

332

333 Table 1 Sequencing data used for Dan340 genome assembly.

Sequencing type	Platform	Library size (bp)	Clean data (Gb)	Application
Genome short reads	Illumina HiSeq 4000	350	47.42	Genome survey and assessment
Nanopore reads	Nanopore platform	20,000	100.10	Contig assembly
Hi-C reads	Illumina HiSeq 4000	300–700	67.25	Chromosome construction
Transcriptome short reads	Illumina HiSeq 4000	200–500	11.06	Genome annotation and assessment

334

335

336 Table 2 Summary of Dan340 genome assembly and annotation.

	Number	**Size**
Genome assembly		
Total contigs	467	348.38 Mb
Contig N50	15	4.36 Mb
Contig N90	137	265 kb
Total scaffolds	321	348.53 Mb
Scaffold N50	12	13.52 Mb
Scaffold N90	23	10.83 Mb
Pseudochromosomes	23	315.08 Mb
Genome annotation		
Repetitive sequences	52.36%	182.52 Mb
Noncoding RNAs	4235	1.12 Mb
Protein-coding genes	28,321	109.30 Mb
Genes in pseudochromosomes	28,297 (99.92%)	109.24 Mb

337

338

339