A chromosome-level genome assembly 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 9.48% of the genes, respectively. Taken together, these results indicated that our Dan340 genome assembly presented high quality and coverage.

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 high quality Illumina reads mapped to the reference genome was up to 97.48%. The assembly of this genome will not only facilitate our understanding about intraspecific genome diversity in maize, but also provides a novel resource for maize breeding.

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

48 al., 2018; Li et al., 2019). The maize genome exhibits high levels of genetic diversity
49 among different inbred lines (Lu et al., 2015; Yang et al., 2017; Sun et al., 2018).
50 The maize cultivar Dan340 is an excellent backbone inbred line of Luda Red Cob
51 Group with several desirable characters, such as disease resistance, lodging resistance,
52 high combining ability, wide adaptability and so on (Zhang et al., 2018).
53 In the present study, we constructed a high-quality chromosome-level reference
54 genome for Dan340 by combining PacBio long HiFi sequencing, Illumina short reads
55 and chromosomal conformational capture (Hi-C) sequencing reads. The assembly of
56 this genome will not only facilitate our understanding about intraspecific genome
57 diversity in maize, but also provides a novel resource for maize breeding
58 improvement.

59 **Plant material and DNA sequencing**

60 Inbred line Dan340 was selected for genome sequencing and assembly because it is
61 an elite maize cultivar, which plays an important role in maize breeding and genetic
62 research. The plants were grown at 25°C in a greenhouse of Beijing Academy of
63 Agriculture and Forestry Sciences, Beijing, China. Fresh and tender leaves were
64 harvested from the best-growing individual and immediately frozen in liquid nitrogen,
65 followed by preservation at -80 °C in the laboratory prior to DNA extraction.
66 Genomic DNA was extracted from the leaf tissue of a single plant using the
67 DNAsecure Plant Kit (Tiangen Biotech Co., Ltd., Beijing, China). To ensure that
68 DNA extracts were useable for all types of genomic libraries, the quality and quantity

69 were evaluated using a NanoDrop 2000 spectrophotometer (NanoDrop Technologies,
70 Wilmington, DE, USA) and electrophoresis on a 0.8% agarose gel, respectively.
71 PacBio circular consensus sequencing (CCS) libraries were prepared using SMRTbell
72 Express Template Prep Kit 2.0 (Pacific Biosciences Ref. No. 101-685-400), following
73 the manufacturer's protocols, and subsequently sequenced on the PacBio sequel II
74 platform (Pacific Biosciences, Menlo Park, CA, USA). Consensus reads (CCS reads,
75 also termed HiFi reads) were generated using ccs software v.3.0.0
76 (https://github.com/pacificbiosciences/unanimity/) with the parameter '--min-passes 3
77 --min-rq 0.99'. The total CCS yield was 63.54 Gb, with a mean read length of 15.60
78 kb.
79 In addition, one Illumina paired-end sequencing library with an insert size of 350 bp
80 was generated using NEB Next Ultra DNA Library Prep Kit (NEB, USA) following
81 manufacturer's protocol, and subsequently sequenced using an Illumina HiSeq X Ten
82 platform (Illumina, San Diego, CA, USA) in Novogene Bioinformatics Institute,
83 Beijing, China. Approximately 80.66 Gb (~20X) of Illumina sequencing data were
84 obtained.
85 One Hi-C library was constructed using young leaves following the published
86 procedures with certain modifications (Belton et al., 2012). In brief, approximately
87 5-g leaf samples were cut into minute pieces and cross-linked by 4% formaldehyde
88 solution at room temperature in a vacuum for 30 min. Then, the sample was mixed
89 with excess 2.5 M glycine to quench the cross linking reaction for 5 min and then

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

97 **Genome assembly**

98 To obtain a high quality genome assembly of Dan340, we employed both PacBio HiFi
99 sequences and Illumina short-read, with scaffolding informed by chromosomal
100 conformation capture (Hi-C). The assembly was performed in a stepwise fashion.
101 First, de novo assembly of the long CCS reads generated from the PacBio SMRT
102 Sequencing was performed using Hifiasm (Cheng et al., 2021). In brief, we first used
103 all CCS clean reads to do error correction by performing all-vs-all pairwise alignment
104 between them. In this step, haplotype phasing was achieved using heterozygous SNPs
105 and most sequencing errors had been removed. Then, error-corrected reads were used
106 to construct phased assembly graph by determining if one overlap is among the reads
107 coming from different haplotypes. Next, based on the graph topological structures and
108 the phasing relationship among different haplotypes, Hifiasm produced a primary
109 contigs (p-contigs). Then, Pilon (Walker et al., 2014) was used to perform the second
110 round of error correction with short paired-end reads generated from Illumina Hiseq
111 Platforms. Subsequently, the Purge Haplotigs pipeline

112 (https://bitbucket.org/mroachawri/purge_haplotigs/overview) was used to remove
113 redundant sequences formed due to heterozygosity.
114 For Hi-C reads, to avoid reads with artificial bias, we removed the following type of
115 reads: (a) Reads with ≥ 10% unidentified nucleotides (N); (b) Reads with > 10 nt
116 aligned to the adapter, allowing ≤ 10% mismatches; (c) Reads with > 50% bases
117 having phred quality < 5. The filtered Hi-C reads were aligned against the contig
118 assemblies with BWA (version 0.7.8). Reads were excluded from subsequent analysis
119 if they did not align within 500 bp of a restriction site or did not uniquely map, and
120 the number of Hi-C read pairs linking each pair of scaffolds is tabulated. ALLHiC
121 v0.8.12 (Zhang et al., 2019) was used in simple diploid mode to scaffold the genome
122 and optimize the ordering and orientation of each clustered group, producing a
123 chromosomal level assembly. Juicebox Assembly Tools v1.9.8
124 (https://github.com/aidenlab/Juicebox) (Durand et al., 2016) was finally used to
125 visualize and manually correct the large-scale inversions and translocations to obtain
126 the final pseudo-chromosomes. The final assembly of Dan340 genome was 2,348.72
127 Mb, including 2,738 contigs and 2,315 scaffolds with a N50 of 41.49 Mb and 215.35
128 Mb (Table 1), respectively. A total of 91.30% sequences was anchored to 10
129 pseudo-chromosomes.
130 **Evaluation of assembly quality**
131 We assessed the quality of the assembly by two independent methods. First, the short
132 reads obtained from the Illumina sequencing data were aligned to the final assembly
133 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, 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.

Conclusions

1. We assembled the chromosome-level genome of maize elite inbred line Dan340 based on long CCS reads from the third-generation PacBio Sequel II sequencing platform, with scaffolding informed by chromosomal conformation capture (Hi-C).
2. 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.

References

Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. Hi-C: a comprehensive technique to capture the conformation of genomes. Methods. 2012; 58, 268-276.

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. (2021) Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 18, 170-175.

154 Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL.
155 (2016) Juicebox Provides a Visualization System for Hi-C Contact Maps with
156 Unlimited Zoom. Cell Syst. 3, 99-101.
157 Hake S, Ross-Ibarra J. (2015) Genetic, evolutionary and plant breeding insights from
158 the domestication of maize. Elife. 4, e05861.
159 Hubert B, Rosegrant M, Boekel MA, Ortiz R. (2010). The future of food: scenarios
160 for 2050. Crop Sci. 50, 33-50.
161 Hirsch CN, Hirsch CD, Brohammer AB, Bowman MJ, Soifer I, Barad O, Shem-Tov
162 D, Baruch K, Lu F, Hernandez AG, Fields CJ, Wright CL, Koehler K, Springer NM,
163 Buckler E, Buell CR, de Leon N, Kaeppler SM, Childs KL, Mikel MA. (2016) Draft
164 Assembly of Elite Inbred Line PH207 Provides Insights into Genomic and
165 Transcriptome Diversity in Maize. Plant Cell. 28, 2700-2714.
166 Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, Campbell MS, Stein JC, Wei X,
167 Chin CS, Guill K, Regulski M, Kumari S, Olson A, Gent J, Schneider KL,
168 Wolfgruber TK, May MR, Springer NM, Antoniou E, McCombie WR, Presting GG,
169 McMullen M, Ross-Ibarra J, Dawe RK, Hastie A, Rank DR, Ware D. (2017)
170 Improved maize reference genome with single-molecule technologies. Nature. 546,
171 524-527.
172 Li C, Song W, Luo Y, Gao S, Zhang R, Shi Z, Wang X, Wang R, Wang F, Wang J,
173 Zhao Y, Su A, Wang S, Li X, Luo M, Wang S, Zhang Y, Ge J, Tan X, Yuan Y, Bi X,
174 He H, Yan J, Wang Y, Hu S, Zhao J. (2019) The HuangZaoSi Maize Genome

175 Provides Insights into Genomic Variation and Improvement History of Maize. Mol
176 Plant. 12, 402-409.
177 Li H, Durbin R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler
178 transform. Bioinformatics. 26, 589-595.
179 Lu F, Romay MC, Glaubitz JC, Bradbury PJ, Elshire RJ, Wang T, Li Y, Li Y,
180 Semagn K, Zhang X, Hernandez AG, Mikel MA, Soifer I, Barad O, Buckler ES.
181 (2015) High-resolution genetic mapping of maize pan-genome sequence anchors. Nat
182 Commun. 6, 6914.
183 Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J,
184 Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C,
185 Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K,
186 Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, Chen
187 W, Yan L, Higginbotham J, Cardenas M, Waligorski J, Applebaum E, Phelps L,
188 Falcone J, Kanchi K, Thane T, Scimone A, Thane N, Henke J, Wang T, Ruppert J,
189 Shah N, Rotter K, Hodges J, Ingenthron E, Cordes M, Kohlberg S, Sgro J, Delgado B,
190 Mead K, Chinwalla A, Leonard S, Crouse K, Collura K, Kudrna D, Currie J, He R,
191 Angelova A, Rajasekar S, Mueller T, Lomeli R, Scara G, Ko A, Delaney K, Wissotski
192 M, Lopez G, Campos D, Braidotti M, Ashley E, Golser W, Kim H, Lee S, Lin J,
193 Dujmic Z, Kim W, Talag J, Zuccolo A, Fan C, Sebastian A, Kramer M, Spiegel L,
194 Nascimento L, Zutavern T, Miller B, Ambroise C, Muller S, Spooner W, Narechania
195 A, Ren L, Wei S, Kumari S, Faga B, Levy MJ, McMahan L, Van Buren P, Vaughn

196 MW, Ying K, Yeh CT, Emrich SJ, Jia Y, Kalyanaraman A, Hsia AP, Barbazuk WB,
197 Baucom RS, Brutnell TP, Carpita NC, Chaparro C, Chia JM, Deragon JM, Estill JC,
198 Fu Y, Jeddeloh JA, Han Y, Lee H, Li P, Lisch DR, Liu S, Liu Z, Nagel DH, McCann
199 MC, SanMiguel P, Myers AM, Nettleton D, Nguyen J, Penning BW, Ponnala L,
200 Schneider KL, Schwartz DC, Sharma A, Soderlund C, Springer NM, Sun Q, Wang H,
201 Waterman M, Westerman R, Wolfgruber TK, Yang L, Yu Y, Zhang L, Zhou S, Zhu
202 Q, Bennetzen JL, Dawe RK, Jiang J, Jiang N, Presting GG, Wessler SR, Aluru S,
203 Martienssen RA, Clifton SW, McCombie WR, Wing RA, Wilson RK. (2009) The
204 B73 maize genome: complexity, diversity, and dynamics. Science. 326, 1112-1115.
205 Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. (2015)
206 BUSCO: assessing genome assembly and annotation completeness with single-copy
207 orthologs. Bioinformatics. 31, 3210-3212.
208 Sun S, Zhou Y, Chen J, Shi J, Zhao H, Zhao H, Song W, Zhang M, Cui Y, Dong X,
209 Liu H, Ma X, Jiao Y, Wang B, Wei X, Stein JC, Glaubitz JC, Lu F, Yu G, Liang C,
210 Fengler K, Li B, Rafalski A, Schnable PS, Ware DH, Buckler ES, Lai J. (2018)
211 Extensive intraspecific gene order and gene structural variations between Mo17 and
212 other maize genomes. Nat Genet. 50, 1289-1295.
213 AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 34,
214 W435-9.
215 Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA,
216 Zeng Q, Wortman J, Young SK, Earl AM. (2014) Pilon: an integrated tool for

comprehensive microbial variant detection and genome assembly improvement. PLoS One. 9(11):e112963.

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.

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.

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.

Table 1 Results of genome assembly assisted by Hi-C technology.

	Contig length(bp)	Scaffold length(bp)	Contig number	Scaffold number
Total	2,348,678,871	2,348,721,171	2,738	2,315
Max	97,718,682	343,826,844	-	-
Number >=2000	-	-	2,738	2,315
N50	41,489,332	215,353,106	20	5
N60	32,900,726	193,940,028	26	6
N70	22,291,114	184,974,134	35	7
N80	11,599,268	165,218,414	49	9
N90	3,788,017	153,467,472	80	10