The Crop Journal xxx (xxxx) xxx

Contents lists available at ScienceDirect

The Crop Journal

journal homepage: www.keaipublishing.com/en/journals/the-crop-journal/

Genome assembly of the maize inbred line A188 provides a new reference genome for functional genomics

Fei Ge [a,1], Jingtao Qu [a,1], Peng Liu [a], Lang Pan [a], Chaoying Zou [a], Guangsheng Yuan [a], Cong Yang [a], Guangtang Pan [a], Jianwei Huang [b], Langlang Ma [a,*], Yaou Shen [a,*]

[a] Key Laboratory of Biology and Genetic Improvement of Maize in Southwest Region, Maize Research Institute, Sichuan Agricultural University, Chengdu 611130, Sichuan, China
[b] Berry Genomics Corporation, Beijing 100015, China

ARTICLE INFO

Article history:
Received 30 April 2021
Revised 30 July 2021
Accepted 20 August 2021
Available online xxxx

Keywords:
Maize
Embryonic callus
A188
Genome assembly
Single-molecule sequencing

ABSTRACT

The current assembled maize genomes cannot represent the broad genetic diversity of maize germplasms. Acquiring more genome sequences is critical for constructing a pan-genome and elucidating the linkage between genotype and phenotype in maize. Here we describe the genome sequence and annotation of A188, a maize inbred line with high phenotypic variation relative to other lines, acquired by single-molecule sequencing and optical genome mapping. We assembled a 2210-Mb genome with a scaffold N50 size of 11.61 million bases (Mb), compared to 9.73 Mb for B73 and 10.2 Mb for Mo17. Based on the B73_RefGen_V4 genome, 295 scaffolds (2084.35 Mb, 94.30% of the final genome assembly) were anchored and oriented on ten chromosomes. Comparative analysis revealed that ~30% of the predicted A188 genes showed large structural divergence from B73, Mo17, and W22 genomes, which causes high protein divergence and may lead to phenotypic variation among the four inbred lines. As a line with high embryonic callus (EC) induction capacity, A188 provides a convenient tool for elucidating the molecular mechanism underlying the formation of EC in maize. Combining our new A188 genome with previously reported QTL and RNA sequencing data revealed eight genes with large structural variation and two differentially expressed genes playing potential roles in maize EC induction.

© 2021 Crop Science Society of China and Institute of Crop Science, CAAS. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Maize (*Zea mays* L.) is used for animal feed and human consumption worldwide. Maize shows great phenotypic polymorphism and genetic diversity [1–5]. More than 80% of the maize genome consists of repetitive element sequences [6–9]. However, next-generation sequencing reads are too short to cover repetitive sequences, resulting in many gaps in assembled maize genomes [6]. Only a few quantitative trait loci (QTL) for agronomic traits have been cloned, owing to a lack of high-quality reference maize genomes. This problem can be overcome by the recently established single-molecule sequencing platform [6], which is able to generate long sequencing reads. By use of this platform, the new B73 genome has been improved with a 52-fold increase in contig length [9].

The temperate line A188 [10] shows great phenotypic difference from other inbred lines, such as in plant height [11], tassel branch number, ear number, days to tassel [11], days to pollination, days to silk [11], oil concentration [12], protein concentration [12], starch concentration [12], and embryonic callus (EC) induction capability. This variation facilitates cloning genes controlling these traits.

Genetic transformation has been an effective approach for elucidating gene function in plants. However, maize genetic transformation is highly reliant on the use of EC induced from immature embryos. Only a few lines possess the ability to efficiently form embryonic callus, including inbred lines such as A188, B104, H99, C01 and the combination Hi-II (A × B) [13–16]. Since plant regeneration from maize tissue culture was first reported in 1975 [17], little is known about how maize EC is induced from the immature embryos despite the efforts of generations of researchers [13–15,18]. A few QTL have been identified as controlling callus induction or plant regeneration [19,20]. A B73 near-isogenic line, WCIC2 (with donor parent A188) with a high frequency of EC initiation was used to genetically fine-map QTL for EC response, and a QTL

* Corresponding authors.
E-mail addresses: sxyljxml@163.com (L. Ma), shenyaou@sicau.edu.cn (Y. Shen).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.cj.2021.08.002
2214-5141/© 2021 Crop Science Society of China and Institute of Crop Science, CAAS. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Please cite this article as: F. Ge, J. Qu, P. Liu et al., Genome assembly of the maize inbred line A188 provides a new reference genome for functional genomics, The Crop Journal, https://doi.org/10.1016/j.cj.2021.08.002

located in a 3.06-Mb region on chromosome 3 was identified as controlling EC formation and regeneration [13]. Owing to the difference between the A188 genome and the B73 reference genome, QTL for embryo culture response have not been cloned to date, limiting their application to improving EC formation capability.

Assembly of a high-quality A188 reference genome would be helpful for identifying the molecular mechanisms underlying EC induction and other agronomic traits. We combined single-molecule sequencing and BioNano optical-mapping technologies to produce a *de novo* assembly of the A188 genome.

2. Materials and methods

2.1. Phenotypic evaluation of maize inbred lines

The maize inbred lines A188, B73, Mo17, and W22, provided by the Maize Research Institute of Sichuan Agricultural University, were grown in Chengdu (Sichuan province, China, N30°67′, E104°06′) in 2018. They were planted in a randomized complete block design with three replicates and two rows per line. Each 3-m row contained 14 plants, with a row spacing of 0.75 m. At 10 days after pollination (DAP), plant height and tassel branch numbers (TBN) were measured as described previously [21]. The duration from seeding to half of the plants tasseling, pollinated, and silking was recorded as days to tassel, days to pollination and days to silk, respectively. Ear number was recorded at 30 DAP. Three mature seeds of each line were crushed and subjected to measurement of protein concentration using the RAPID N exceed (Elementar, Langenselbold, Germany).

To evaluate the EC induction ratio, A188, B73, Mo17 and W22 were planted in a greenhouse (14 h/10 h light/dark, at 28 °C and 70% relative humidity). At 12 DAP, 108 immature embryos (1.2–1.5 mm in length) from each line were collected and evenly distributed among three Petri dishes containing modified N6 medium [22] to induce EC with three repetitions. After aseptic incubation for 21 days in darkness at 28 °C, the EC induction ratio was recorded as (number of immature embryos successfully inducing EC/number of inoculated immature embryos) × 100%.

2.2. Genome assembly and annotation

Libraries for genome sequencing were constructed as previously described [23]. High-quality genomic DNA was sheared to ~20 kb with Agilent 2100 (Agilent Technologies Inc., CA, USA). The resulting PCR-free SMRTbell libraries were sequenced on the PacBio Sequel platform (Pacific Biosciences, Menlo Park CA, USA). A total of 63 SMT cells were run on the PacBio Sequel instrument, generating 27.25 million reads with a total length of 224.03 Gb. Reads longer than 10 kb were used for contig assembly with Falcon [23] and polished with the Arrow program (https://www.pacb.com/support/software-downloads/).

2.3. Optical library construction and sequencing

Nicking, labeling, repair, and staining were performed as standard BioNano protocols (BioNano Genomics, San Diego, CA, USA). The High Molecular Weight genome was specifically recognized by the *BspQ* I enzyme to identify the site to be labeled. Optical maps were assembled with the BioNano Irys system (BioNano Genomics, San Diego, CA, USA). Raw BNX files were filtered and assembled into genome maps using the BioNano Solve pipeline (https://bionanogenomics.com/support-page/bionano-solve/).

2.4. Hybrid assembly of PacBio contigs and BioNano optical maps

The PacBio-assembled contigs and BioNano-assembled genome maps were subjected to hybrid assembly using the HybridScaffold module of IrysSolve as described previously [8]. Briefly, the PacBio genome maps were aligned to an *in silico* BspQI-digested cmap. The BioNano genome maps were then aligned to the PacBio genome maps with RefAligner, followed by identifying and resolving the conflict points. The BioNano and PacBio genome maps were then merged to generate a hybrid scaffold. The PacBio genome maps were mapped to hybrid scaffolds again to identify overlaps. If the overlap between PacBio contigs and hybrid scaffold was longer than 1 kb and identity was ≥95%, the two sequences were merged. Based on the alignment information, super-scaffolds were built.

2.5. Construction of pseudomolecules

The reference genome of B73_RefGen_v4 [9] was used to anchor A188 scaffolds to chromosomes. Contigs were ordered using Bwa (version bwa-0.7.15, http://bio-bwa.sourceforge.net/bwa.shtml) and placed on chromosomes based on synteny between A188 and B73.

2.6. Assembly evaluation

BUSCO (Benchmarking Universal Single-Copy Orthologs: http://busco.ezlab.org/) combined with tblastn, augustus (http://bioinf.uni-greifswald.de/augustus/), and hmmer (http://hmmer.org/) software was used to evaluate genome-assembly completeness. 'Embryophyta_odb9' containing 1440 single-copy orthologous genes, was used as a search dataset to assess the completeness of the A188 genome assembly.

2.7. Repetitive element prediction

Transposable elements were identified by a combination of homolog-based and *de novo* approaches. TRF v4.07b (http://tandem.bu.edu/trf/trf407b.linux64.download.html) was used to predict tandem repeats. LTR Finder [24], RepeatScout (v1.0.5, http://www.repeatmasker.org), and PILER (v1.0, http://www.drive5.com/piler) were used to predict LTR elements, LINEs, SINEs, and transposable DNA, respectively.

2.8. Isoform-sequencing

Total RNA was extracted from five tissues (12-d seedlings, tassels, silks, pericarp, and 20-DAP seeds) with TRIzol reagent (Invitrogen, Carlsbad, CA) according to the manufacturer's instructions. For each sample, three independent biological replicates were generated. Equal amounts of RNA (1 μg) for each replicate of each tissue were pooled. One microgram of enriched poly-A RNA was reverse-transcribed into cDNA using the Clontech SMARTER cDNA synthesis kit (Clontech Laboratories, Inc. A Takara Bio Company, Mountain View, CA, USA), and the cDNA was subjected to size selection using the BluePippin system. Size fractions eluted from the run were re-amplified to generate two libraries (0–1 and 1–10 kb). Then 2 μg cDNA of each library was subjected to Iso-Seq SMRTBell library construction as described (https://pacbio.secure.force.com/SamplePrep). The SMRTBell libraries were then subjected to single-molecule sequencing on the PacBio Sequel platform.

2.9. Gene annotation

MAKER2 (http://www.yandell-lab.org/software/maker.html) [25] was used to annotate genes in the A188 genome by the strat-

egy described previously [8]. First, for protein-homology-based prediction, the proteins of the B73, Mo17, and W22 reference genomes were retrieved from Gramene (http://gramene.org/) [26] as input to MAKER2. The A188 transcripts assembled from five different tissues based on single-molecule long-read sequencing, B73 full-length transcripts from Iso-seq [27], and Mo17 transcripts [8] were used for gene transcript prediction. Second, the generated gene models were submitted to Augustus [28], SNAP (http://snap.stanford.edu/snap/download.html), GeneMark-ESSuite (version 4.32 http://topaz.gatech.edu/GeneMark/license_download.cgi), and Glimmerhmm (http://ccb.jhu.edu/software/glimmerhmm/) *ab initio* prediction software to further *de novo* predict gene models. The preliminary prediction gene set was then filtered according to the AED scores generated in MAKER software and high-confidence gene models were generated.

2.10. Identification of PAV sequences

To identify presence/absence-variation (PAV, length >500 bp) sequences, a sliding-window method was used as reported previously [8]. To identify A188-specific PAV sequences by comparison with B73, the A188 genome was divided into 500-bp windows with a step size of 100 bp. Then all of the 500-bp windows were aligned to the B73 genome with BWA mem [29] (http://bio- bwa.-sourceforge.net/) with options "-w 500 -M". A188-specific PAV sequences were sequences that could not be aligned to the B73 genome or whose primary alignment coverage was <25% [8]. Overlapping PAV windows were merged. The same method was used to identify A188-specific PAV sequences relative to Mo17 and to W22, B73-specific PAV sequences relative to A188, Mo17-specific PAV sequences relative to A188, and W22-specific PAV sequences relative to A188. PAV sequences within 100 kb of the physical coordinates were further merged. A merged region with more than 10% PAV sequences was defined as a PAV cluster. Finally, all of the PAVs were anchored back to the corresponding genome. The same method [8] was used to identify A188-specific genes relative to B73, Mo17, and W22. Genes more than 75% of whose coding sequence (CDS) fell in PAV sequences were defined as PAV genes.

2.11. Comparative genomic analysis among A188, B73, Mo17 and W22

Mummer software [30] was used to perform comparative genomic analysis of the A188 and B73 genomes. Each A188 pseudochromosome sequence was mapped to the corresponding B73 chromosome using mummer with the parameters "nucmer -g 1000 -c 90 -l 40". The mapping results were submitted to delta-filter to perform noise filtration with parameters "-r -q". Show-coords was used for conversion of aligned physical coordinates with parameters "-rcloTH". SNPs and small (<100 bp) InDels were identified using show-snps with "-ClrTH". Show-diff was employed with default parameters to obtain inversions. Finally, alignments with aligned physical positions in one genome that were located more than 10 Mb away in another genome were removed. Comparative genomic analyses of A188 with the Mo17, A188, and W22 genomes were performed by the same method.

2.12. Quantitative real-time PCR

Immature maize embryos of the inbred lines A188, C01, Mo17, and B73 were cultured to induce callus as described in section 2.2. After culture for 0 h and 72 h, the embryos or calli were collected for isolating total RNA using Trizol Reagent (Life Technologies, Carlsbad, CA, USA), with three biological repetitions. Reverse transcription was performed following the protocol of PrimeScript RT Reagent Kit with gDNA Eraser (Takara Biotechnology Co., Ltd., Dalian, China). Quantitative real-time PCR was performed on an ABI 7500 real-time PCR System (Applied Biosystems, CA, USA), with *Actin 1* used as the reference gene. The primers used for quantitative real-time PCR are described in Table S12. The $2^{-\triangle\triangle CT}$ method was used to calculate the relative expression levels of candidate genes in all lines.

3. Results

3.1. Phenotypic differences between A188 and the other assembled lines

As a public inbred line, A188 shows an excellent response to tissue culture, with approximately 100% efficiency in forming EC from immature embryos [31,32]. However, in previous studies, B73 and Mo17 both showed a very low frequency of inducing EC under standard conditions [13,33]. Comparison of EC formation ratio among A188, B73, Mo17, and W22 confirmed the high ratio of A188 and the low ratios of the other three lines (Table 1). Other agronomic traits also showed marked differences between A188 and the other lines (Table 1; Fig. 1). Our findings were in agreement with previous studies [11,12] of the phenotypic performance of A188, indicating that A188/B73, A188/Mo17, and A188/W22 are ideal pairs of maize lines for genetic and molecular studies of these traits.

3.2. Genome sequencing and de novo assembly provided a high-quality reference genome of A188

In combination with optical genome mapping with the BioNano Genomics Irys System, PacBio Sequel platform was used to sequence and *de novo* assemble the A188 genome. A >104-fold coverage of sequence (224.03 Gb in total) generated from PacBio Sequel was used for an initial 2127.72 Mb assembly with a contig N50 size of 1.06 Mb and longest contig of 4.97 Mb (Tables 2, S1, and S2). A 631.48-Gb BioNano molecule (287-fold-coverage BioNano optical map) used to scaffold the assembled contigs generated a final assembly of 4469 scaffolds with a scaffold N50 size of 11.61 Mb and longest scaffold of 47.84 Mb (Tables 2, S1, and S2). The total size of the final assembly was 2207.74 Mb, similar to those of the B73 (2106 Mb) [9], Mo17 (2183 Mb) [8] and SK (2094 Mb) [7] genomes (Tables 2, S2). Finally, 295 scaffolds were anchored and oriented onto ten chromosomes (2084.35 Mb, 94.30% of the final genome assembly) and 3704 scaffolds failed to be mapped to chromosomes (5.70% of the final genome assembly) (Table S13). The final A188 assembly had 2480 gaps (89.56 Mb in length), compared with 2522 gaps in the B73 and 238 gaps in the SK genome [7]. Finally, 95.3% of complete single-copy BUSCOs [34] could be aligned to the A188 final assembly, similar to those for the B73 [9], Mo17 [8], W22 [35] and SK [7] genomes, indicating the near completeness of our assembly (Table S4).

3.3. Genome annotation showed a complex genome composition of A188

A total of 80.70% of the A188 genome sequence was identified as transposable-element sequences, including retrotransposons (71.93%), DNA transposons (5.91%), and unclassified elements (2.49%) (Table S5), which was lower than those in B73 [9], Mo17 [8], W22 [35], SK [7] and K0326Y [6] genomes. For retrotransposons, the families of *Copia* and *Gypsy* represented respectively 24.01% and 46.92% of the A188 genome (Table S5). For DNA transposons, the representation of the *hAT* family was much lower than those in the B73 [9] and Mo17 [8] genomes.

In total, 44,653 high-confidence protein-coding gene models with 66,359 transcripts were predicted (Table S6). Among them,

Table 1
Agronomic trait phenotypes in four inbred lines.

Line	ECIR[#] (%)	Plant height[*]	TBN[*]	Ear number[*]	PC[#]	DtT	DtP	DtS
A188	91.53 ± 5.55 a	126.00 ± 16.83 d	12.36 ± 2.71 a	1.38 ± 0.62 a	11.98 ± 0.04 b	51	59	61
W22	1.67 ± 1.85 c	159.50 ± 17.03 c	11.48 ± 2.48 a	1.29 ± 0.60 ab	10.48 ± 0.07 c	67	70	70
Mo17	6.94 ± 1.39 b	184.29 ± 17.99 b	6.21 ± 1.05 b	1.07 ± 0.68 bc	12.59 ± 0.11 a	70	71	72
B73	0 ± 0 c	203.36 ± 13.92 a	6.64 ± 1.23 b	0.92 ± 0.51 c	9.34 ± 0.05 d	71	72	73

Values are means ± SD (*, $n = 42$; #, $n = 3$); Letters a, b, c, d indicate significant differences among lines at $P < 0.05$. ECIR, embryonic callus induction ratio; TBN, tassel branch numbers; PC, protein concentration; DtT, days to tassel; DtP, days to pollination; DtS, days to silk.

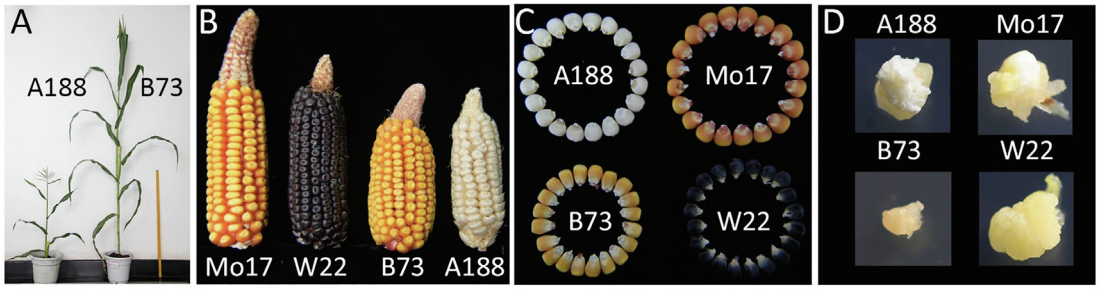

Fig. 1. Trait differences among A188, B73, Mo17, and W22 inbred lines, including plant height (A), ear size (B), kernel size (C) and embryonic callus (D).

Table 2
Global statistics for the A188 genome assembly.

	PacBio assembly	PacBio + BioNano hybrid assembly	Pseudomolecule
Total length of assembly (Mb)	2127.72	2210.33	2084.35
N50 size (Mb)	1.06	11.61	–
Greatest length (Mb)	4.97	47.84	–
Number of sequences	6385	4469	10

10,965 (24.56%) and 16,243 (36.38%) genes were supported by ISO-seq with CDS coverage >90% and >50%, respectively (Table S6). In total, 41,715 (93.42%) of the predicted A188 genes were mapped to 10 pseudochromosomes (Table S3). Of these genes, 62,058 (93.52%) were functionally annotated and deposited in public databases (Fig. S1).

3.4. Genomic polymorphisms among A188, B73, Mo17, and W22

To identify genome differences, we individually aligned the pseudochromosomes of B73, Mo17, and W22 to those of A188. Respectively 62.50% (1316.38 Mb), 63.10% (1327.82 Mb), and 62.59% (1327.48 Mb) of the B73, Mo17 and W22 genome sequences matched in one-to-one syntenic blocks with 63.16% (1316.45 Mb), 63.71% (1328.03 Mb), and 63.69% (1327.43 Mb) of the A188 genome sequence (Figs. 2, S2; Table S7).

Similar numbers of SNPs, insertions, and deletions were identified between A188 and other 3 inbred lines, except more insertions were found between A188 and Mo17 (Figs. 2, S3; Table S7). <2.5% of these variations in A188 were found in CDS regions, with the remainder annotated as intergenic variations (Tables 3, S8). InDels of 3 N ± 1 bp in the CDS region were more abundant than those of 3 N bp in gene coding regions (Table 3), between A188 and any of B73, Mo17 and W22. Comparison of the A188 and B73 genomes revealed 27,240 A188-specific genomic segments (16.92 Mb) and 28,558 B73-specific genomic segments (17.76 Mb). Most of these PAV segments were shorter than 3 kb, with only 1 and 2 longer PAV segments in A188 and B73, respectively (Fig. S4). Similarly, comparison of the A188 and Mo17 genomes revealed 26,983 A188-specific genomic segments (16.76 Mb), and 28,030 Mo17-specific genomic segments (17.44 Mb). Most of the PAV segments were shorter than 3 kb, with only 1 and 3 longer PAV segments in A188 and Mo17, respectively (Fig. S4). Comparison of the A188 and W22 genomes revealed 31,536 A188-specific genomic segments (19.42 Mb) and 29,192 W22-specific genomic segments (17.98 Mb), with 1 and 4 PAV segments longer than 3 kb in A188 and W22, respectively (Fig. S4). According to the criterion that a gene with $\geq$ 75% of coding sequences covered by a PAV sequence can be assigned as a PAV gene [8], we identified 100 A188-specific and 104 B73-specific PAV genes by comparison of the A188 and B73 genomes. Similarly, 116 A188-specific and 146 Mo17-specific PAV genes were found by comparison of A188 and Mo17, and 112 A188-specific and 116 W22-specific PAV genes were identified between A188 and W22 (Tables 3, S9). Thus, the A188 genome showed large variation with respect to the B73, Mo17, and W22 genomes. However, only 9 A188-specific PAV genes were simultaneously identified in comparison with the other three inbred lines, showing that most of the A188-specific PAV genes were already present in other lines (Table S9).

3.5. Gene structural variation among A188, B73, Mo17, and W22

Totals of 20,557, 21,007 and 20,713 genes displayed no variation in the CDS regions between B73 and A188, Mo17 and A188, and W22 and A188, respectively (Table 3). Respectively 17,168, 17,634 and 17,382 A188 genes showed no variations in CDS and intron regions as compared with B73, Mo17 and W22 (Table 3). In particular, as compared with B73, Mo17 and W22, respectively 8647, 9054 and 8854 genes were highly conserved without any genetic variation (including 2 kb upstream and downstream) (Table 3). Respectively 23,989, 24,424 and 20,860 A188 genes showed synonymous variations in CDS compared to B73, Mo17, and W22 (Table 3). Compared with B73, 22,958, 21,975 and 7210 genes in A188 resulted in amino acid changes, missense mutation and non-frameshift InDels, respectively (Table 3). Mapped to Mo17, 23,313, 21,601 and 7257 genes in A188 contained amino acid changes, missense mutations in CDS, or non-frameshift InDels, respectively (Table 3). Aligned to W22, 23,070, 21,869, and 7295 A188 genes showed amino acid changes, missense mutations in CDS, or non-frameshift InDels, respectively (Table 3). All of these

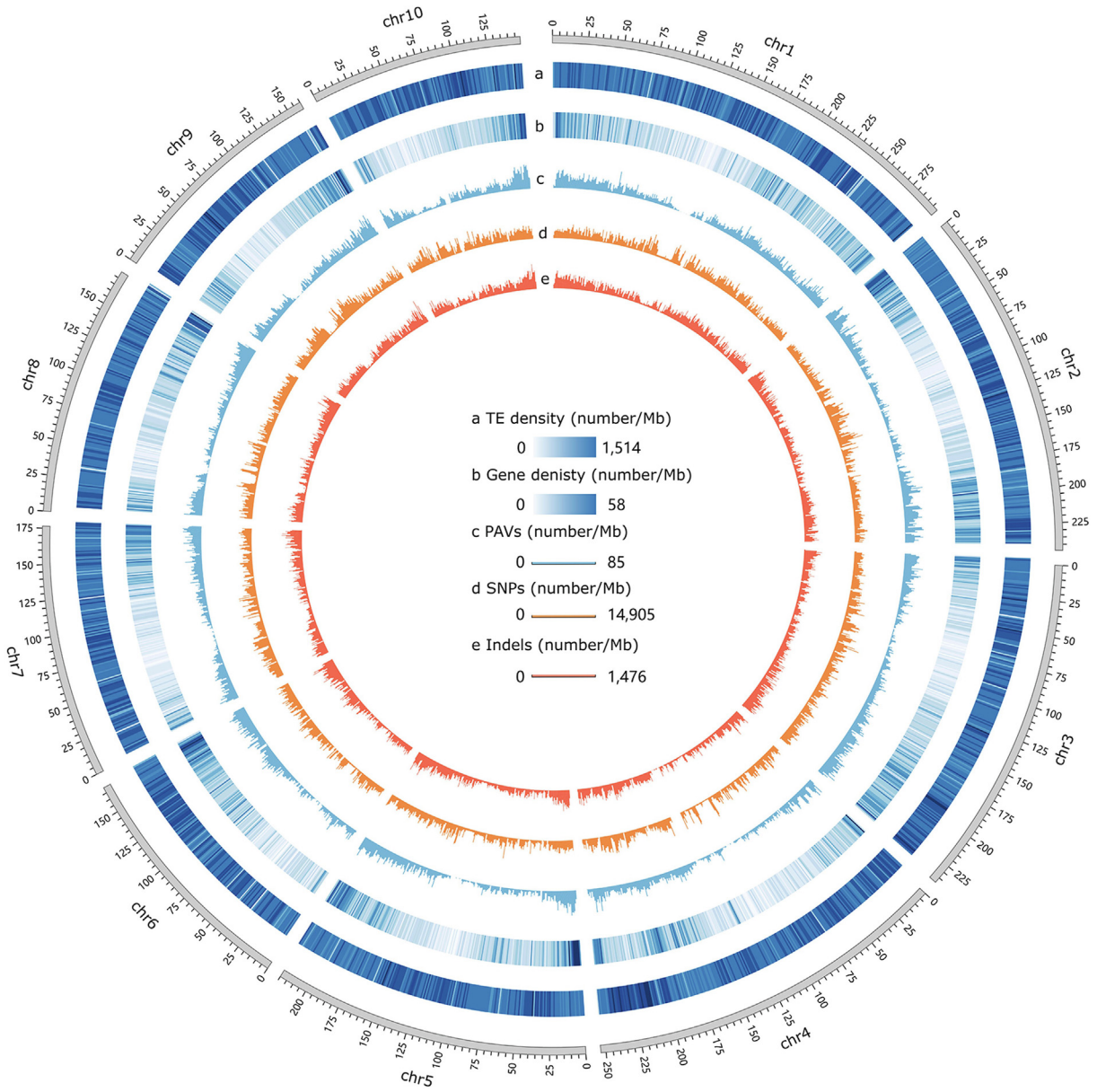

Fig. 2. Features of the A188 genome. (a) Transposable-element density; (b) Gene density; (c, d and e) numbers of PAVs (c), SNPs (d) and InDels (e) compared with B73 genome. The sliding window is 1 Mb for all tracks.

genes were classified as structurally conserved between A188 and the other lines and accounted for ≥68.61% of the annotated A188 genes.

Comparison of the B73 and A188 genomes revealed 737, 506, 841, 1671, 10,834, and 2362 genes in A188 that generated start-codon mutations, stop-codon mutations, splice-donor mutations, splice-acceptor mutations, frameshift InDels in CDS, and premature termination codon mutations, respectively (Table 3). Respectively 747, 504, 801, 1559, 10,982, and 2355 genes in A188 led to start-codon mutations, stop-codon mutations, splice-donor mutations, splice-acceptor mutations, frameshift InDels in CDS, and premature termination codon mutations, as compared with Mo17 (Table 3). Respectively 742, 506, 811, 1486, 10,889, and 2389 genes in A188 showed start-codon mutations, stop-codon mutations, splice-donor mutations, splice-acceptor mutations, frameshift InDels in CDS, and premature termination codon mutations, as compared with W22 (Table 3). Respectively 204, 262 and 228 PAV genes were identified between A188 and the B73, Mo17, and W22 genomes, (Table 3). Respectively 13,224 (29.62%), 13,306 (29.80%) and 13,167 (29.49%) A188 genes had large structural variations (start- or stop-codon mutations, splice-donor or splice-acceptor mutations, frameshift mutations, premature termination codon mutations, or PAV variations) in comparison with the B73, Mo17, and W22 genomes.

3.6. A188 genome-based genetic dissection revealed candidate genes for tissue culture response

Recently [13], using an $F_{3:4}$ population derived from B73 and WCIC2 (a near-isogenic line of B73 containing several A188 segments), a locus associated with embryogenic and regenerative capabilities of immature embryo was fine-mapped to within a 3.06 M region (chr3:178772856–181826658) based on the B73 reference genome, suggesting that the genes harbored by the A188 segment caused the high callus formation ratio. To identify candidate genes for EC induction, we aligned the 3.06 M B73 segment to

Table 3
Variation within genes among the A188, B73, Mo17, and W22 genomes.

Variation type	A188 to B73	A188 to Mo17	A188 to W22
Structurally conserved genes	30,635	31,095	30,764
No DNA variation in CDS	20,557	21,007	20,713
No DNA variation in CDS and intron	17,168	17,634	17,382
No DNA variation in genic region*	8647	9054	8854
Without amino acid substitutions	23,989	24,424	20,860
With amino acid changes	22,958	23,313	23,070
With missense mutation in CDS	21,975	21,601	21,869
With 3 N InDel in CDS	7210	7257	7295
Genes with large structural mutations	13,020	13,044	12,939
Start codon mutation	737	747	742
Stop codon mutation	506	504	506
Splice donor mutation	841	801	811
Splice acceptor mutation	1671	1559	1486
With 3 N ± 1 InDel in CDS	10,834	10,982	10,889
Premature termination codon	2362	2355	2389
PAV genes	204	262	228
A188-present PAV genes	100	116	112
A188-absent PAV genes	104	146	116
Total of genes with large structural variations	13,224	13,306	13,167

*Genic regions include the 2 kb upstream and downstream of the gene body.

the A188 genome and identified a 3.89 M syntenic segment (Fig. 3) in A188. Within the 3.89 M segment, respectively 51, 57, and 57 A188 genes were identified syntenic to B73, Mo17, and W22 syntenic segments (Table S10). Among them, respectively 6, 14, and 6 genes showed large structural variation (LSV: premature termination codon, stop-codon loss, frameshift deletion, or frameshift insertion) relative to B73, Mo17, and W22 (Table S11), and 4 LSV genes were simultaneously identified in A188 by comparison with the other three lines (Table 4).

Respectively 48, 42, and 42 A188 genes in the QTL interval were nonsyntenic in comparison with the B73, Mo17 and W22 genomes (Table S10). To determine whether the nonsyntenic genes have homologs in other sites of the 3 inbred lines, we mapped these nonsyntenic genes to the B73, Mo17, and W22 genomes. Respectively 28, 11, and 24 A188 nonsyntenic genes showed LSV relative to their homologs in B73, Mo17, and W22 (Table S11), and 4 LSV genes in A188 were simultaneously identified in comparison with the other three inbred lines (Table 4).

Changes in gene expression can be induced during somatic embryogenesis to respond to tissue culture [13,36–38]. Based on the reported transcriptome data of A188 [36], 4 of the 99 A188 genes within the mapped QTL region were up- or down-regulated by more than 8 folds in different stages of immature embryo culture, relative to a control. We performed qRT-PCR of the four genes in two lines with high EC induction capacity (A188 and C01), and two lines with low EC induction capacity (Mo17 and B73). During callus induction, the expression of ZmY09GFa039032 was up-regulated in A188 and C01, but down-regulated in Mo17 and B73 (Fig. S5A). The expression levels of ZmY09GFa039032 were higher in A188 and C01 at 72 h of incubation than in Mo17 and B73. The

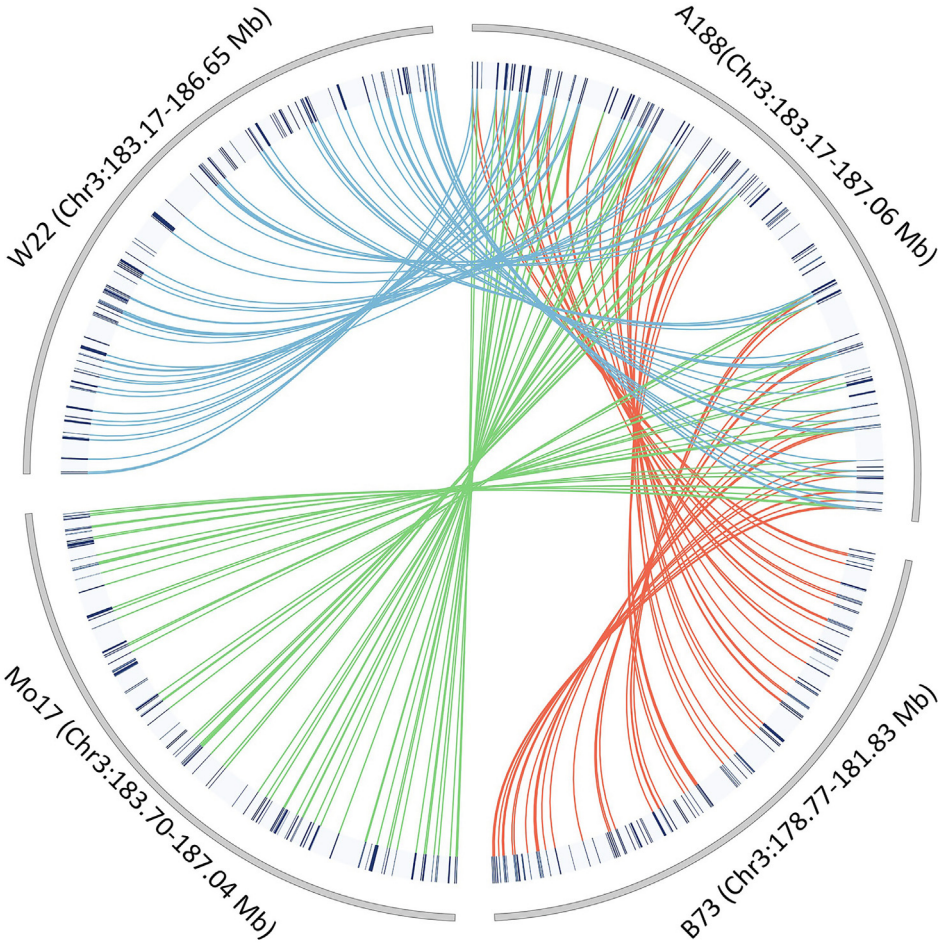

Fig. 3. Tissue culture response candidate loci. The 3.89 M A188 segment (QTL for maize tissue culture response) aligned to syntenic segments in B73, Mo17, and W22 genomes. Red, green and blue lines connect aligned A188 genes in the 3.89 M segment to those in B73, Mo17, and W22, respectively.

Table 4
Tissue culture response candidate genes.

A188 Gene ID	B73 homologous	Mutation type to B73	Mo17 homologous	Mutation type to Mo17	W22 homologous	Mutation type to W22	Homologous type	Annotation
ZmY09GFa037173	GRMZM2G123977	Stop gain	Zm00014a019537	Stop gain	Zm00004b018533	Stop gain	Syntenic gene	Ankyrin repeat-containing protein; signal transduction
ZmY09GFa037487	GRMZM2G359234	Frameshift deletion	Zm00014a019543	Stoploss and frameshift deletion	Zm00004b018529	Frameshift deletion	Syntenic gene	UDP-glucuronic acid decarboxylase
ZmY09GFa038636	GRMZM2G337905	Stop gain	Zm00014a019529	Stop gain	Zm00004b018516	Stop gain	Syntenic gene	Helicase-like protein; DNA repair
ZmY09GFa039738	GRMZM5G856598	Stop gain	Zm00014a039033	Stop gain	Zm00004b018443	Stop gain	Syntenic gene	Probable anion transporter
ZmY09GFa035987	GRMZM2G341918	Frameshift insertion and stop gain	Zm00014a013928	Stop gain	Zm00004b000208	Frameshift, insertion and stop gain	Nonsyntenic homologous	Zinc finger MYM-type protein 1-like
ZmY09GFa037580	GRMZM2G156296	Frameshift insertion and stop gain	Zm00014a010023	Frameshift insertion	Zm00004b030624	Frameshift, insertion and stop gain	Nonsyntenic homologous	Uncharacterized protein loc103635851
ZmY09GFa038110	GRMZM2G084717	Frameshift deletion	Zm00014a020349	Frameshift deletion	Zm00004b017917	Frameshift, deletion	Nonsyntenic homologous	Hypothetical protein
ZmY09GFa038645	GRMZM2G078468	Stop gain	Zm00014a004443	Stop gain	Zm00004b021555	Frameshift, deletion and stop gain	Nonsyntenic homologous	Hypothetical protein
ZmY09GFa038775	GRMZM2G084779	Synonymous SNV	Zm00014a020354	Synonymous SNV	Zm00004b017916	–	DE gene, nonsyntenic homologous	Potasium ion uptake permease
ZmY09GFa039032	GRMZM2G065557	–	Zm00014a036794	–	Zm00004b018459	–	DE gene, syntenic gene	Hypothetical protein

DE gene, differentially expressed genes during tissue culture response.

expression of ZmY09GFa038775 was increased by 247–911 folds among the four lines at 72 h relative to 0 h (Fig. S5C). The transcript abundance of ZmY09GFa038775 was much higher in the two high EC induction rate lines than in the remaining lines at 72 h (Fig. S5C). However, for ZmY09GFa036902 and ZmY09GFa036216, no difference of expression level was observed between the two groups with contrasting EC induction frequencies (Fig. S5B and D). Collectively, the four syntenic genes with LSV, the four nonsyntenic genes with LSV, and the differentially expressed genes ZmY09GFa039032 and ZmY09GFa038775, were assigned as candidate genes responsible for tissue culture capability of A188 immature embryo (Table 4).

4. Discussion

Although A188's application in breeding programs is limited by its poor agronomic traits, A188 shows high phenotypic variation relative to B73, Mo17, and W22, in particular in EC induction ratio. Phenotypic performance is determined by the combination of genotype and environment. To investigate the mechanisms underlying the phenotypic difference between A188, B73, Mo17, and W22, we sequenced and *de novo* assembled the A188 genome into 2207.74 Mb with a scaffold N50 size of 11.61 Mb. As expected, A188 showed large genomic variations as compared with B73, Mo17 and W22. Our new A188 genome provides a resource for mapping causal genes controlling these various traits. We also identified A188 genes presenting structure variation relative to other three inbred lines, such as genes with start- or stop-codon mutations, splice-donor or -acceptor mutations, and frameshift InDels, providing a novel resource for gene function and evolutionary analysis.

We demonstrated the utility of this new genome by using it to dissect the genetic control of EC induction. EC induction from maize immature embryo is highly dependent on genotype, so that only a few functional genes have been identified. Combining our new A188 genome, previously reported QTL, and RNA sequencing data, we identified 10 candidate genes responsible for maize tissue culture response (Table 4). These candidate genes suggest the molecular mechanisms of maize tissue culture response and represent new gene resources for improving maize EC induction and maize genetic transformation. In particular, ZmY09GFa037173 showed a premature termination mutation in A188, which was annotated as an Ankyrin repeat-containing protein and involved in signal transduction. The finding [39] that the *Arabidopsis* homolog *Itn1* regulated ROS accumulation under salt stress via regulation of ABA signaling pathways suggests that ZmY09GFa037173 has potential to induce maize callus formation by regulating ROS levels.

Data availability

All datasets reported in this study have been deposited in GenBank (National Coalition Building Institute, NCBI) with the following accession IDs: Genome assembly, JADZIA000000000; Raw data for genome assembly and gene annotation, PRJNA678284.

CRediT authorship contribution statement

Fei Ge: Conceptualization, Data curation, Formal analysis, Funding acquisition, Project administration, Investigation, Supervision, Validation, Visualization, Writing – original draft, Writing - review & editing. **Jingtao Qu:** Data curation, Formal analysis, Software,

Writing – original draft, Writing - review & editing. **Peng Liu:** Data curation, Formal analysis, Software, Validation, Writing - review & editing. **Lang Pan:** Formal analysis, Investigation, Writing - review & editing. **Chaoying Zou:** Formal analysis, Writing - review & editing. **Guangsheng Yuan:** Software, Writing - review & editing. **Cong Yang:** Software, Writing - review & editing. **Guangtang Pan:** Funding acquisition, Writing - review & editing. **Jianwei Huang:** Data curation, Software, Writing - review & editing. **Langlang Ma:** Conceptualization, Data curation, Project administration, Software, Supervision, Writing - review & editing. **Yaou Shen:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Project administration, Resources, Supervision, Writing – original draft, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (31871637, 32072073, and 32001500), and the Project of Transgenic New Variety Cultivation (2016ZX08003003).

Appendix A. Supplementary data

Supplementary data for this article can be found online at https://doi.org/10.1016/j.cj.2021.08.002.

References

[1] E.S. Buckler, B.S. Gaut, M.D. McMullen, Molecular and functional diversity of maize, Curr. Opin. Plant Biol. 9 (2006) 172–176.
[2] P.S. Schnable, D. Ware, R.S. Fulton, J.C. Stein, F. Wei, S. Pasternak, C. Liang, J. Zhang, L. Fulton, T.A. Graves, The B73 maize genome: complexity, diversity, and dynamics, Science 326 (2009) 1112–1115.
[3] N.M. Springer, K. Ying, Y. Fu, T. Ji, C.T. Yeh, Y. Jia, W. Wu, T. Richmond, J. Kitzman, H. Rosenbaum, Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content, PLoS Genet. 5 (2009) e1000734.
[4] J. Lai, R. Li, X. Xu, W. Jin, M. Xu, H. Zhao, Z. Xiang, W. Song, K. Ying, M. Zhang, Genome-wide patterns of genetic variation among elite maize inbred lines, Nat. Genet. 42 (2010) 1027.
[5] J. Yan, M. Warburton, J. Crouch, Association mapping for enhancing maize (*Zea mays* L.) genetic improvement, Crop Sci. 51 (2011) 433–449.
[6] C. Li, X. Xiang, Y. Huang, Y. Zhou, D. An, J. Dong, C. Zhao, H. Liu, Y. Li, Q. Wang, C. Du, J. Messing, B.A. Larkins, Y. Wu, W. Wang, Long-read sequencing reveals genomic structural variations that underlie creation of quality protein maize, Nat. Commun. 11 (2020) 17.
[7] N. Yang, J. Liu, Q. Gao, S. Gui, L. Chen, L. Yang, J. Huang, T. Deng, J. Luo, L. He, Y. Wang, P. Xu, Y. Peng, Z. Shi, L. Lan, Z. Ma, X. Yang, Q. Zhang, M. Bai, S. Li, W. Li, L. Liu, D. Jackson, J. Yan, Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement, Nat. Genet. 51 (2019) 1052–1059.
[8] S. Sun, Y. Zhou, J. Chen, J. Shi, H. Zhao, H. Zhao, W. Song, M. Zhang, Y. Cui, X. Dong, H. Liu, X. Ma, Y. Jiao, B. Wang, X. Wei, J.C. Stein, J.C. Glaubitz, F. Lu, G. Yu, C. Liang, K. Fengler, B. Li, A. Rafalski, P.S. Schnable, D.H. Ware, E.S. Buckler, J. Lai, Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes, Nat. Genet. 50 (2018) 1289–1295.
[9] Y. Jiao, P. Peluso, J. Shi, T. Liang, M.C. Stitzer, B. Wang, M.S. Campbell, J.C. Stein, X. Wei, C.S. Chin, K. Guill, M. Regulski, S. Kumari, A. Olson, J. Gent, K.L. Schneider, T.K. Wolfgruber, M.R. May, N.M. Springer, E. Antoniou, W.R. McCombie, G.G. Presting, M. McMullen, J. Ross-Ibarra, R.K. Dawe, A. Hastie, D.R. Rank, D. Ware, Improved maize reference genome with single-molecule technologies, Nature 546 (2017) 524.
[10] P. Gacheri, J. Machuka, O. Ombori, B. Bukachi, Agrobacterium mediated transformation of selected maize inbred lines with pPZP200 towards enhancment of lysine and methionine content, J. Biol. Agric. Healthcare 5 (2015) 1–18.
[11] J.A. Peiffer, M.C. Romay, M.A. Gore, S.A. Flint-Garcia, Z. Zhang, M.J. Millard, C.A. Gardner, M.D. McMullen, J.B. Holland, P.J. Bradbury, The genetic architecture of maize height, Genetics 196 (2014) 1337–1356.
[12] J.P. Cook, M.D. McMullen, J.B. Holland, F. Tian, P. Bradbury, J. Ross-Ibarra, E.S. Buckler, S.A. Flint-Garcia, Genetic architecture of maize kernel composition in the nested association mapping and inbred association panels, Plant Physiol. (2011).
[13] S. Salvo, J. Cook, A.R. Carlson, C.N. Hirsch, S.M. Kaeppler, H.F. Kaeppler, Genetic fine-mapping of a quantitative trait locus (QTL) associated with embryogenic tissue culture response and plant regeneration ability in maize (*Zea mays* L.), Plant, Genome (2018) 170111.
[14] C.L. Armstrong, J. Romero-Severson, T.K. Hodges, Improved tissue culture response of an elite maize inbred through backcross breeding, and identification of chromosomal regions important for regeneration by RFLP analysis, Theor. Appl. Genet. 84 (1992) 755–762.
[15] M. Krakowsky, M. Lee, L. Garay, W. Woodman-Clikeman, M. Long, N. Sharopova, B. Frame, K. Wang, Quantitative trait loci for callus initiation and totipotency in maize (Zea mays L.), Theor. Appl. Genet. 113 (2006) 821–830.
[16] F. Bronsema, W. Van Oostveen, A. Van Lammeren, Comparative analysis of callus formation and regeneration on cultured immature maize embryos of the inbred lines A188 and A632, Plant Cell Tissue Organ Cult. 50 (1997) 57–65.
[17] C. Green, R. Phillips, Plant regeneration from tissue cultures of maize, Crop Sci. 15 (1975) 417–421.
[18] G. Pan, Z. Zhang, X. Wei, Y. Song, M. Zhao, Y. Xia, T. Rong, QTL analysis of maize (*Zea mays* L.) embryo culturing capacity, Zuo wu xue bao 32 (2006) 7–13.
[19] B.A. Lowe, M.M. Way, J.M. Kumpf, J. Rout, D. Warner, R. Johnson, C.L. Armstrong, M.T. Spencer, P.S. Chomet, Marker assisted breeding for transformability in maize, Mol. Breed. 18 (3) (2006) 229–239.
[20] K. Lowe, E. Wu, N. Wang, G. Hoerster, C. Hastings, M.-J. Cho, C. Scelonge, B. Lenderts, M. Chamberlin, J. Cushatt, L. Wang, L. Ryan, T. Khan, J. Chow-Yiu, W. Hua, M. Yu, J. Banh, Z. Bao, K. Brink, E. Igo, B. Rudrappa, P.M. Shamseer, W. Bruce, L. Newman, B. Shen, P. Zheng, D. Bidney, C. Falco, J. Register, Z.-Y. Zhao, D. Xu, T. Jones, W. Gordon-Kamm, Morphogenic regulators baby boom and wuschel improve monocot transformation, Plant Cell 28 (2016) 1998–2015.
[21] P.J. Brown, N. Upadyayula, G.S. Mahone, F. Tian, P.J. Bradbury, S. Myles, J.B. Holland, S. Flint-Garcia, M.D. McMullen, E.S. Buckler, T.R. Rocheford, J. Flint, Distinct genetic architectures for male and female inflorescence traits of maize, PLoS Genet. 7 (2011) e1002383.
[22] B.R. Frame, H. Shou, R.K. Chikwamba, Z. Zhang, C. Xiang, T.M. Fonger, S.E.K. Pegg, B. Li, D.S. Nettleton, D. Pei, K. Wang, Agrobacterium tumefaciens-mediated transformation of maize embryos using a standard binary vector system, Plant Physiol. 129 (2002) 13–22.
[23] M. Pendleton, R. Sebra, A.W.C. Pang, A. Ummat, O. Franzen, T. Rausch, A.M. Stütz, W. Stedman, T. Anantharaman, A. Hastie, H. Dai, M.H.Y. Fritz, H. Cao, A. Cohain, G. Deikus, R.E. Durrett, S.C. Blanchard, R. Altman, C.S. Chin, Y. Guo, E.E. Paxinos, J.O. Korbel, R.B. Darnell, W.R. McCombie, P.Y. Kwok, C.E. Mason, E.E. Schadt, A. Bashir, Assembly and diploid architecture of an individual human genome via single-molecule technologies, Nat. Methods 12 (2015) 780.
[24] L. Xu, Y. Zhang, Y. Su, L. Liu, J. Yang, Y. Zhu, C. Li, Structure and evolution of full-length LTR retrotransposons in rice genome, Plant Syst. And Evol. 287 (2010) 19–28.
[25] B.L. Cantarel, I. Korf, S.M. Robb, G. Parra, E. Ross, B. Moore, C. Holt, A.S. Alvarado, M. Yandell, MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes, Genome Res. 18 (2008) 188–196.
[26] M.K. Tello-Ruiz, J. Stein, S. Wei, J. Preece, A. Olson, S. Naithani, V. Amarasinghe, P. Dharmawardhana, Y. Jiao, J. Mulvaney, Gramene 2016: comparative plant genomics and pathway resources, Nucleic Acids Res. 44 (2016) D1133–D1140.
[27] B. Wang, E. Tseng, M. Regulski, T.A. Clark, T. Hon, Y. Jiao, Z. Lu, A. Olson, J.C. Stein, D. Ware, Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing, Nat. Commun. 7 (2016) 11708.
[28] O. Keller, M. Kollmar, M. Stanke, S. Waack, A novel hybrid gene prediction method employing protein multiple sequence alignments, Bioinformatics 27 (2011) 757–763.
[29] H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, arXiv (2013) 1303.3997.
[30] S. Kurtz, A. Phillippy, A.L. Delcher, M. Smoot, M. Shumway, C. Antonescu, S.L. Salzberg, Versatile and open software for comparing large genomes, Genome Biol. 5 (2004) R12.
[31] Y. Ishida, H. Saito, S. Ohta, Y. Hiei, T. Komari, T. Kumashiro, High efficiency transformation of maize (*Zea mays* L.) mediated by Agrobacterium tumefaciens, Nat. Biotechnol. 14 (1996) 745–750.
[32] T. Hodges, K. Kamo, C. Imbrie, M. Becwar, Genotype specificity of somatic embryogenesis and regeneration in maize, Bio-Technology 4 (1986) 219–223.
[33] B.R. Frame, J.M. McMurray, T.M. Fonger, M.L. Main, K.W. Taylor, F.J. Torney, M. M. Paz, K. Wang, Improved Agrobacterium-mediated transformation of three maize inbred lines using MS salts, Plant Cell Rep. 25 (2006) 1024–1034.
[34] F.A. Simão, R.M. Waterhouse, P. Ioannidis, E.V. Kriventseva, E.M. Zdobnov, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, Bioinformatics 31 (2015) 3210–3212.
[35] N.M. Springer, S.N. Anderson, C.M. Andorf, K.R. Ahern, F. Bai, O. Barad, W.B. Barbazuk, H.W. Bass, K. Baruch, G. Ben-Zvi, E.S. Buckler, R. Bukowski, M.S. Campbell, E.K.S. Cannon, P. Chomet, R.K. Dawe, R. Davenport, H.K. Dooner, L.H. Du, C. Du, K.A. Easterling, C. Gault, J.-C. Guan, C.T. Hunter, G. Jander, Y. Jiao, K.E. Koch, G. Kol, T.G. Köllner, T. Kudo, Q. Li, F. Lu, D. Mayfield-Jones, W. Mei, D.R. McCarty, J.M. Noshay, J.L. Portwood, G. Ronen, A.M. Settles, D. Shem-Tov, J. Shi, I. Soifer, J.C. Stein, M.C. Stitzer, M. Suzuki, D.L. Vera, E. Vollbrecht, J.T. Vrebalov, D. Ware, S. Wei, K. Wimalanathan, M.R. Woodhouse, W. Xiong, T.P. Brutnell,

The maize W22 genome provides a foundation for functional genomics and transposon biology, Nat. Genet. 50 (2018) 1282–1288.

[36] X. Zhang, S.A.G.D. Salvo, C.N. Hirsch, C.R. Buell, S.M. Kaeppler, H.F. Kaeppler, Whole transcriptome profiling of maize during early somatic embryogenesis reveals altered expression of stress factors and embryogenesis-related genes, PLoS ONE 9 (2014) e111407.

[37] Y. Shen, Z. Jiang, X. Yao, Z. Zhang, H. Lin, M. Zhao, H. Liu, H. Peng, S. Li, G. Pan, J. Zhang, Genome expression profile analysis of the immature maize embryo during dedifferentiation, PLoS ONE 7 (2012) e32237.

[38] F. Ge, H. Hu, X. Huang, Y. Zhang, Y. Wang, Z. Li, C. Zou, H. Peng, L. Li, S. Gao, Metabolomic and Proteomic Analysis of Maize Embryonic Callus induced from immature embryo, Sci. Rep. 7 (2017) 1004.

[39] H. Sakamoto, O. Matsuda, K. Iba, ITN1, a novel gene encoding an ankyrin-repeat protein that affects the ABA–mediated production of reactive oxygen species and is involved in salt-stress tolerance in Arabidopsis thaliana, Plant J. 56 (2008) 411–422.