


SOFTWARE

Open Access



gcaPDA: a haplotype-resolved diploid assembler

Min Xie^{2†}, Linfeng Yang^{1,2†}, Chenglin Jiang¹, Shenshen Wu¹, Cheng Luo¹, Xin Yang², Lijuan He², Shixuan Chen², Tianquan Deng², Mingzhi Ye², Jianbing Yan^{1,3} and Ning Yang^{1,3*} 

*Correspondence:

ningy@mail.hzau.edu.cn

[†]Min Xie and Linfeng Yang are contributed equally to this work

¹ National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China

Full list of author information is available at the end of the article

Abstract

Background: Generating chromosome-scale haplotype resolved assembly is important for functional studies. However, current *de novo* assemblers are either haploid assemblers that discard allelic information, or diploid assemblers that can only tackle genomes of low complexity.

Results: Here, Using robust programs, we build a diploid genome assembly pipeline called gcaPDA (gamete cells assisted Phased Diploid Assembler), which exploits haploid gamete cells to assist in resolving haplotypes. We demonstrate the effectiveness of gcaPDA based on simulated HiFi reads of maize genome which is highly heterozygous and repetitive, and real data from rice.

Conclusions: With applicability of coping with complex genomes and fewer restrictions on application than most of diploid assemblers, gcaPDA is likely to find broad applications in studies of eukaryotic genomes.

Keywords: Haplotype-resolved *de novo* assembler, Diploid, Highly heterozygous genomes, Gamete cells

Background

Deciphering the genetic blueprint of a species is of fundamental importance for related researches. Nowadays, genome sequence with quality on par with, if not superior to, the human reference genome could be easily generated with long read sequencing [1, 2] and assistant approaches [3, 4]. Most genomes of the plant and animal species are diploid, however, current long read *de novo* assemblers [5–7] are mainly aiming to generate high-contiguity haploid mixed-haplotypes assembly. In haploid mixed-haplotypes assembly, homozygous or low heterozygous regions are collapsed into a single mixed haplotype, whereas highly heterozygous regions are assembled into separate contigs and the allelic contigs (haplotigs) would be removed. Therefore, haploid mixed-haplotypes assembly can't fully represent the complete genetic blueprint of diploid species [8]. However, accurately phased genomes are essential for many population genetics analysis and the understanding sequence-specific variation such as allele-specific expression, methylation effects and so on.



© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

To resolve haplotypes, *de novo* assemblers such as FALCON-unzip [9], hifiasm [10] and SDip [11] try to build diploid assembly by distinguishing long reads of different haplotypes based on heterozygous single nucleotide polymorphisms (SNPs). This strategy requires reads error rate to be much lower than genome heterozygous rate. In addition, heterozygous SNPs are not evenly distributed along chromosomes, with long stretch of homozygous region scattered in the genome. When long reads fail to span adjacent heterozygous SNPs, haplotype phasing across this region can't be correctly inferred and lead haplotype switches. Therefore, supplementary data that can assist in long-range phasing is imperative to achieve chromosome-scale haplotype construction. In this regard, parental whole genome shotgun reads (WGS) data [12, 13], Hi-C data [14], Strand-seq data [15] and gamete cell data [16, 17] have been used to bridge adjacent haplotype blocks.

Trio binning [13] uses WGS data of two parents to partition the progeny's long reads and then assembles paternal and maternal genome respectively. It can provide entirely phased diploid assembly, except for novel mutations that unique to the progeny's genome or heterozygous variants that exist in the trio samples. However, parental samples are not always available, especially in plant and animal studies. Diploid *de novo* assemblers such as DipAsm [14], PGAS (trio-free Phased Diploid genome Assembly using Strand-Seq) [15], Gamete-binning [16] share a common assembly framework (Additional file 1: Figure S1), which involving building an initial assembly, calling and phasing SNPs using sequencing reads, partitioning long reads based on phased SNPs and assembling each haplotype genome separately with partitioned long reads. This framework has gain success in genomes of low heterozygosity. However, it is not well suited for species with highly heterozygous genomes, which are quite prevalent scenarios in nature. Highly heterozygous genomes pose great challenge to haploid assembler and usually result in low quality initial assembly with fragmented contigs, abundant mis-assemblies and haplotigs. It is difficult to remove haplotigs thoroughly, while retaining haplotigs in the initial assembly will mess with SNP calling [18] and phasing process and generate poor haplotype-resolved assembly. Hence, it is particularly urgent to develop a diploid assembler that can cope with highly heterozygous genomes.

In this study, we build a diploid assembly pipeline called gcaPDA (gamete cells assisted Phased Diploid Assembler) based on robust programs, that can generate chromosome-scale phased diploid assemblies for highly heterozygous and repetitive genomes using PacBio HiFi data, Hi-C data and gamete cell WGS data. gcaPDA offers equivalent performance to the trio-dependent method, with 98% phasing accuracy and >99% genome completeness. Both of the reconstructed haplotype assemblies generated using gcaPDA have excellent collinearity with their corresponding reference assemblies. Additionally, structural variations between reference genomes, including inversions and InDels, are well-resolved. Having demonstrated its utility with maize and rice genome, it is plausible that gcaPDA can be easily applied to most diploid eukaryotic species and may find broad application in the coming diploid genome era.

Results

Schematic of the gcaPDA assembler

We have developed a diploid assembler, gcaPDA, to generate chromosome-scale phased genome assembly for diploid species. gcaPDA requires PacBio HiFi data and Hi-C data

of an individual. In addition, short read WGS data of haploid gamete cells from the same individual are required to assist in long-range phasing. As illustrated in Fig. 1 and Additional file 1: Figure S2, gcaPDA consists of 4 major steps: (1) generating an initial assembly; (2) reconstructing of haplotypes; (3) partition and normalization of gamete cell reads and (4) generating chromosome-scale phased diploid assembly.

Generating sequencing data for a maize F₁ hybrid

Maize has a very representative large and complex genome [19], which is suitable for testing the performance of gcaPDA. We selected a maize F₁ hybrid by crossing two inbred lines B73 and SK as test sample. High quality reference genomes are available for both parents (B73 [20] and SK [21]) which could be used to benchmark the phased diploid assemblies. In total, 126 Gb and 129 Gb PacBio HiFi reads, with an average read length of 14 Kb, were simulated based on published genome sequences of B73 [20] and SK [21], respectively (Additional file 1: Table S1). Combining simulated B73 and

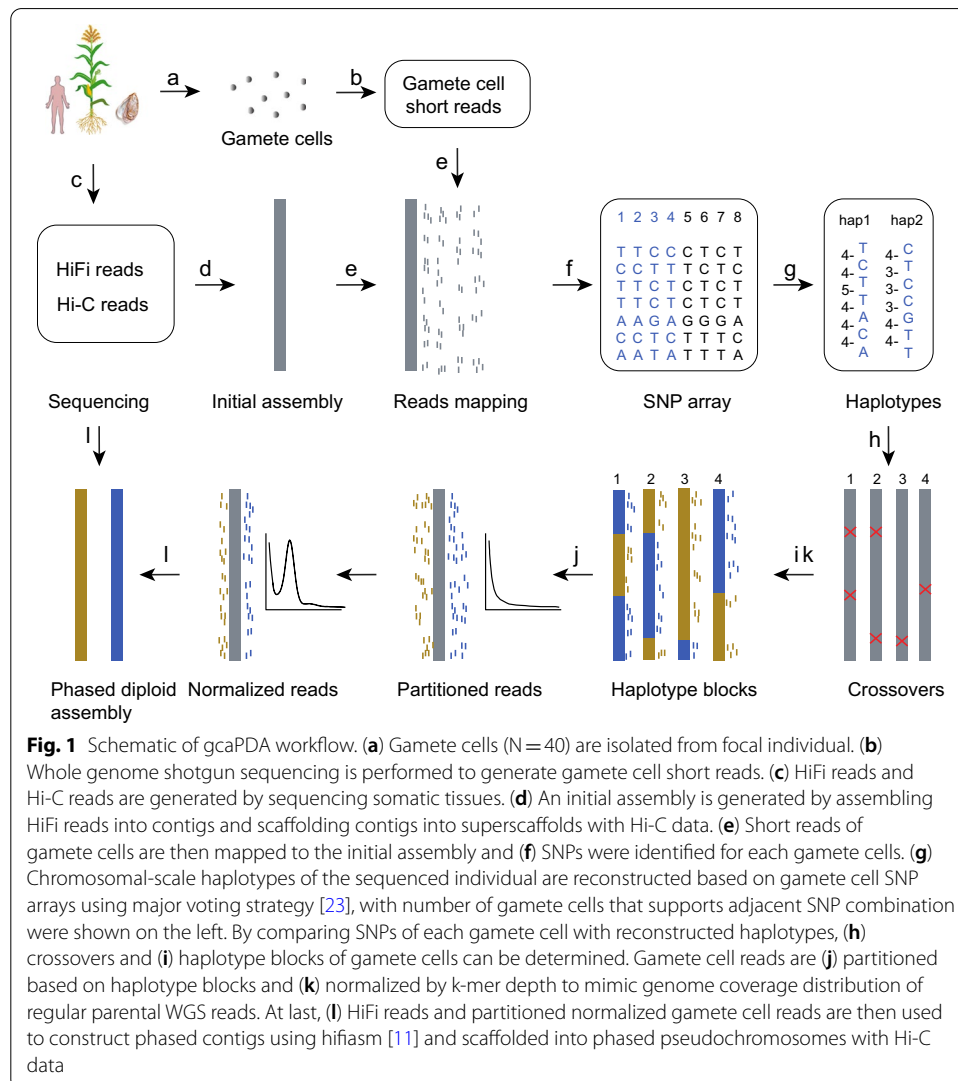


Fig. 1 Schematic of gcaPDA workflow. (a) Gamete cells (N=40) are isolated from focal individual. (b) Whole genome shotgun sequencing is performed to generate gamete cell short reads. (c) HiFi reads and Hi-C reads are generated by sequencing somatic tissues. (d) An initial assembly is generated by assembling HiFi reads into contigs and scaffolding contigs into superscaffolds with Hi-C data. (e) Short reads of gamete cells are then mapped to the initial assembly and (f) SNPs were identified for each gamete cells. (g) Chromosomal-scale haplotypes of the sequenced individual are reconstructed based on gamete cell SNP arrays using major voting strategy [23], with number of gamete cells that supports adjacent SNP combination were shown on the left. By comparing SNPs of each gamete cell with reconstructed haplotypes, (h) crossovers and (i) haplotype blocks of gamete cells can be determined. Gamete cell reads are (j) partitioned based on haplotype blocks and (k) normalized by k-mer depth to mimic genome coverage distribution of regular parental WGS reads. At last, (l) HiFi reads and partitioned normalized gamete cell reads are then used to construct phased contigs using hifiasm [11] and scaffolded into phased pseudochromosomes with Hi-C data

SK PacBio HiFi reads together represents the sequenced reads of the F_1 hybrid. K-mer analysis shows that the heterozygosity, haploid genome size and repeat content of the F_1 hybrid genome are 1.98%, 2.16 Gb and 78% (Additional file 1: Figure S3), respectively, which indicates it's a large, highly heterozygous and repetitive genome.

To assist chromosome-scale assembly, 40 microspores (hereafter referring to as gamete cells) were isolated from 10 tetrads of the F_1 hybrid (Fig. 1a). DNA was extracted from each gamete cell and followed by multiple displacement amplification (MDA) [22] to generate enough DNA for library construction and sequencing [23] (Fig. 1b). Around 45 Gb (~20-fold) high quality short reads were generated for each gamete cell (Additional file 1: Table S2). In addition, 427 Gb Hi-C data were also generated from root tissues of the F_1 hybrid (Fig. 1c and Additional file 1: Table S3).

Generating an initial assembly for the maize F_1 hybrid

The simulated HiFi reads were assembled into contigs using haploid assembler FALCON [9] (Fig. 1d). The FALCON assembly is comprised of 2903 Mb primary contigs and 685 Mb alternative contigs, with a contig N50 of 1.3 Mb. The total length of primary contigs was larger than both the reference genomes of SK (2161 Mb) and B73 (2106 Mb), indicating there are ~742 Mb haplotigs in the primary contigs. Of these, 432 Mb haplotigs can be tagged by `purge_haplotigs` [18], while the remaining 310 Mb haplotigs are undetectable (Additional file 1: Table S4). Since haplotigs in the primary contigs are hard to remove thoroughly, we choose to keep all of them and generated an initial assembly by scaffolding all the primary contigs into super-scaffolds with Hi-C data (Fig. 1d). The longest 10 super-scaffolds (corresponding to 10 chromosomes) covers 97.5% of the initial assembly.

Reconstruction of chromosome-scale haplotypes for the maize F_1 hybrid

Short reads of gamete cells were mapped to the initial assembly and SNPs were identified for each gamete cell (Fig. 1e, f). Nine out of the 40 gamete cells have abnormal SNP heterozygous rate or missing rate (Additional file 1: Table S5 and Figure S4), which could be caused by contamination of adjacent cells or insufficient genome coverage. These gamete cells were considered to be low-quality and excluded from downstream analyses. In general, 1,387,594 to 2,423,031 high confident SNPs were identified in each gamete cell (Additional file 1: Table S5).

Chromosome-scale haplotypes of the sequenced F_1 hybrid were reconstructed based on the SNPs that located on 10 chromosomes using a major voting strategy with R package `Hapi` [24] (Fig. 1g). Totally, we obtained 2,721,839 phased SNPs in the reconstructed chromosome-scale haplotypes, which are evenly distributed across chromosomes (Additional file 1: Figures S5 and S6). Haplotype blocks in each gamete cell could be identified by comparing the genotype at each SNP locus with that of the reconstructed chromosome-scale haplotypes (Fig. 1h, i and Additional file 1: Figure S6).

Partition and normalization of gamete cell short reads

Reads of each gamete cell that belonged to haplotype blocks of B73 or SK were extracted and merged separately (Fig. 1j). Reads of each haplotype were normalized by k-mer depth to mitigate the extremely uneven genome coverage caused by MDA procedure

(Fig. 1k). After normalization, k-mer depth distribution of the haplotype reads is similar to that of parental WGS data (Additional file 1: Figure S7) and ready for use by trio-dependent *de novo* assemblers (HiCanu, hifiasm, etc.).

Generating phased diploid genome assemblies for the maize F₁ hybrid

The phased diploid genome assemblies were generated using hifiasm [10], with simulated HiFi reads and k-mers derived from normalized haplotype reads (Fig. 1l). In total, 2162 Mb contigs were assigned to SK haplotype (HapSK contigs) and 2159 Mb contigs were assigned to B73 haplotype (HapB73 contigs), with a contig N50 of 55.3 Mb and 57.0 Mb, respectively (Table 1). The hapSK contigs and hapB73 contigs were further scaffolded into chromosomes with Hi-C data, respectively. This chromosome-scale phased diploid assembly, including both hapSK assembly and hapB73 assembly, is referred to gcaPDA assembly hereafter.

Testing the generalizability of gcaPDA on real data of rice F₁ hybrid

We selected a rice F₁ hybrid by crossing two inbred lines MH63 and ZS97 as test sample. Gap-free reference genomes are available for both parents [25] which could be used to benchmark the phased diploid assemblies. In total, 23 Gb PacBio HiFi reads, with an average read length of 14.9 Kb of the rice hybrid was generated (Additional file 1: Table S6). K-mer analysis shows that the heterozygosity and haploid genome size of the F₁ hybrid genome are 0.72% and 386 Mb respectively. To assist chromosome-scale assembly, 24 gamete cells were isolated from 6 tetrads of the rice hybrid. Around 10 Gb (~25-fold) high quality short reads were generated for each gamete cell (Additional file 1: Table S7). In addition, 101 Gb Hi-C data were also generated from root tissues of the rice hybrid (Additional file 1: Table S8). Then we perform gcaPDA on rice hybrid same with maize.

Comparing gcaPDA with other methods

To compare the performance of gcaPDA with other methods, we generated a “Hifiasm assembly” with using hifiasm with only HiFi reads, a “Trio assembly” using hifiasm with HiFi reads and simulated parental WGS reads and a “Hifiasm + Hi-C assembly” with using hifiasm with both HiFi and Hi-C reads. Genome assemblies were evaluated in three aspects: contiguity, completeness and accuracy.

Genome contiguity is usually measured with contig N50. All the assembler generated two haplotypic assemblies. The overall contig N50 of gcaPDA is comparable to that of other approaches in with both maize and rice dataset (Table 1, Additional file 1: Table S9).

Genome completeness were evaluated by k-mer, and BUSCOs [26] (Benchmarking Universal Single-Copy Orthologs), and assembled genome size. All assemblies achieved >99% k-mer completeness (Table 1, Additional file 1: Table S9). In addition, all assemblies achieved overall BUSCO completeness comparable to that of the parental genomes. In accordance with the evaluation by k-mer and BUSCOs, assembled genome size of all assemblies are comparable to that of the parental genomes (Table 1, Additional file 1: Table S9). When we looked into the completeness of haplotypic assemblies, we found that duplicated and missing BUSCOs rate are higher

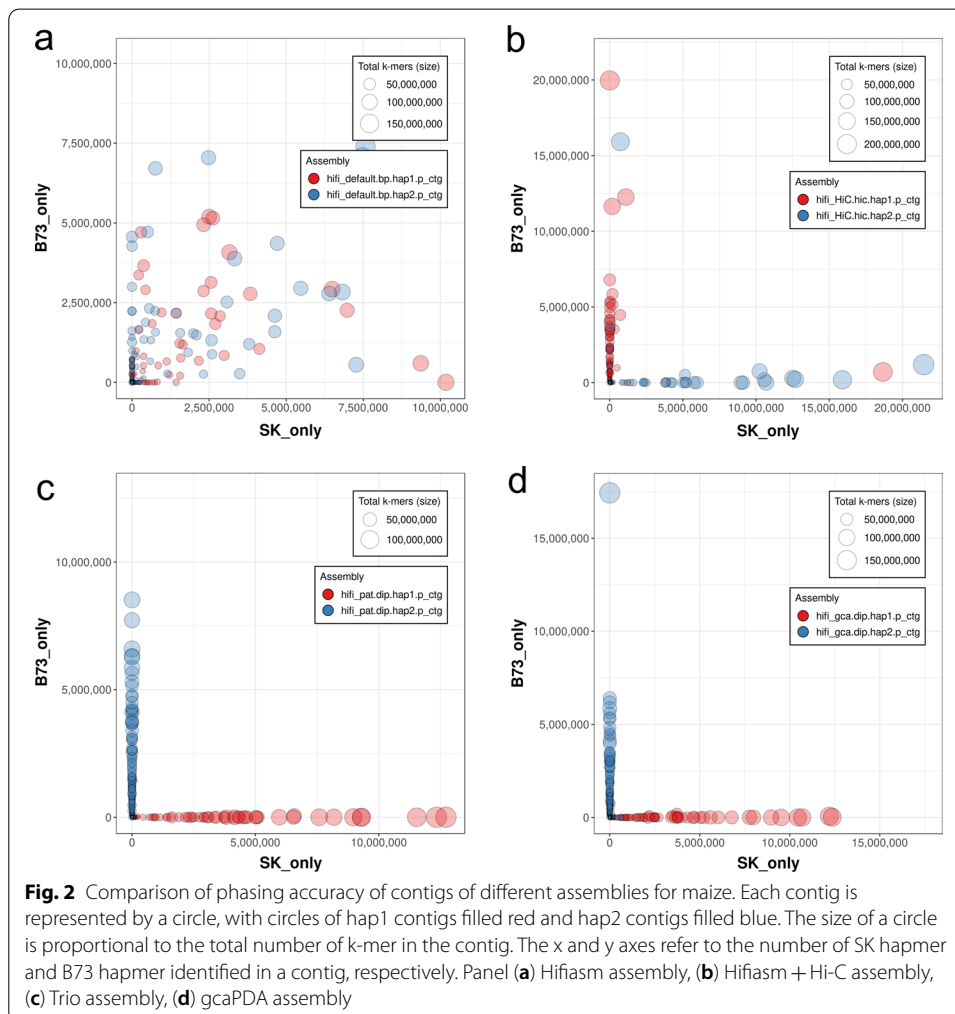
Table 1 Evaluation of Hifiasm assembly, Trio assembly, Hifiasm + Hi-C assembly and gcaPDA assembly

Assembly	Contigs			k-mer completeness (%)				Gene completeness (%)				
	Total (Mb)	No	N50 (Mb)	All	Hap SK	Hap B73	PPV*	Comp	Dup	Frag	Mis	
<i>Reference</i>												
SK	2154	671	223.2	76.15	100.00	0.00	NA	96.1	6.0	1.3	2.6	
B73	2104	267	220.8	75.97	0.00	100.00	NA	95.5	6.1	1.5	3.0	
B73 + SK	4258	938	223.2	100.00	100.00	100.00	NA	96.9	94.8	0.9	2.2	
<i>Hifiasm</i>												
Hap1	2160	800	77.3	75.69	54.91	48.75	53.16	94.4	8.6	1.0	4.6	
Hap2	2160	673	72.4	75.89	48.47	55.35	53.12	95.2	8.3	0.8	4.0	
Hap1 + Hap2	4319	1473	75.5	99.86	99.75	99.67	53.14	97.6	94.2	0.2	2.2	
<i>Hifiasm + Hi-C</i>												
HapSK	2134	671	96.9	75.98	97.53	2.49	97.53	96.4	5.3	0.9	2.7	
HapB73	2124	690	48.0	76.02	2.81	97.76	97.18	96.1	5.8	1.2	2.7	
HapSK + hapB73	4258	1361	69.6	99.96	99.89	99.93	97.36	97.7	94.9	0.2	2.1	
<i>Trio</i>												
HapSK	2160	1118	60.0	76.06	99.62	0.58	99.43	96.1	7.6	0.8	2.7	
HapB73	2118	811	35.6	75.98	0.49	99.85	99.51	96.5	7.2	1.2	2.9	
HapSK + hapB73	4278	1929	44.9	99.90	99.65	99.92	99.47	97.8	94.9	0.4	2.1	
<i>gcaPDA</i>												
HapSK	2162	767	55.3	76.19	98.87	1.78	98.25	95.9	6.5	1.4	2.7	
HapB73	2159	699	57.0	76.41	2.25	99.46	97.77	95.4	6.2	1.6	3.0	
HapSK + hapB73	4320	1466	55.3	99.67	99.06	99.56	98.01	96.9	94.5	1.0	2.2	

*PPV indicates positive predictive value

in maize Hifiasm hap1/hap2 assemblies compared to haplotypic assemblies of other approaches. This might be caused by insufficient phasing and partition of contigs.

Phasing accuracy of genome assemblies were evaluated with hapmer. The overall phasing accuracy (positive predictive value, PPV) of the gcaPDA assembly is 98.01% and 98.5% for maize and rice separately, comparable to that of the Trio assembly and Hifiasm + Hi-C assembly (Table 1, Additional file 1: Table S9) and assemblies reported in recent studies [16, 27, 28]. Furthermore, we investigated chimeric contigs that contain both SK and B73 hapmer. In Hifiasm assembly, there are many contigs contain hampers from both haplotypes (Fig. 2, Additional file 1: Figure S8). In comparison, only few contigs in gcaPAD assembly, Trio assembly and Hifiasm + Hi-C contain hapmer from the other haplotype (Fig. 2, Additional file 1: Figure S8). With hapmer, we also detected haplotype blocks from the other haplotype in the hapSK and hapB73 assembly of gcaPDA (Fig. 3 and Additional file 1: Table S10). It is worth to note that un-anchored sequences are more prone to be mis-assigned to the other haplotype, when comparing with sequences that anchored to chromosomes (Additional file 1: Table S10).



Comparing gcaDPA assembly with parental reference genomes

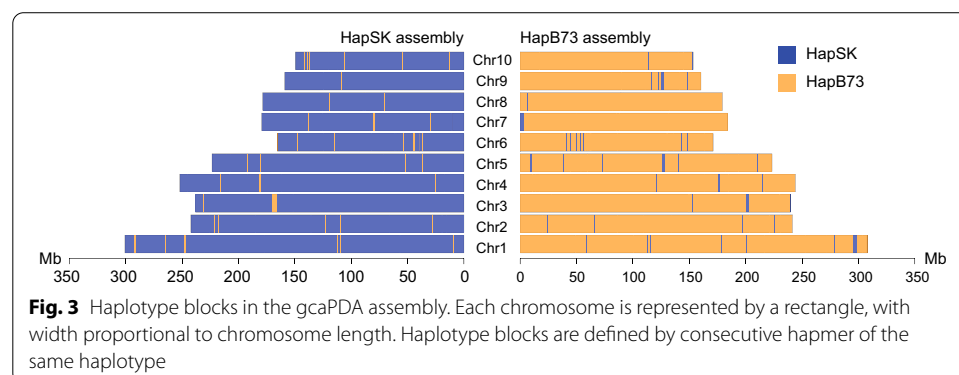
With reference genome sequences for both parents of the F₁ hybrid available, accuracy of gcaPDA assembly was further evaluated by comparing with parental genomes. The HapSK and HapB73 assembly were compared with SK and B73 reference genomes, respectively. The hapSK assembly covers 99.04% of the SK reference genome, with an average alignment identity of 99.99%, while the hapB73 assembly covers 99.72% of the B73 reference genome, with an average alignment identity of 99.99% (Additional file 1: Table S11). In general, near perfect collinearity was observed between haplotype-resolved assemblies and corresponding parental genomes (Additional file 1: Figure S9).

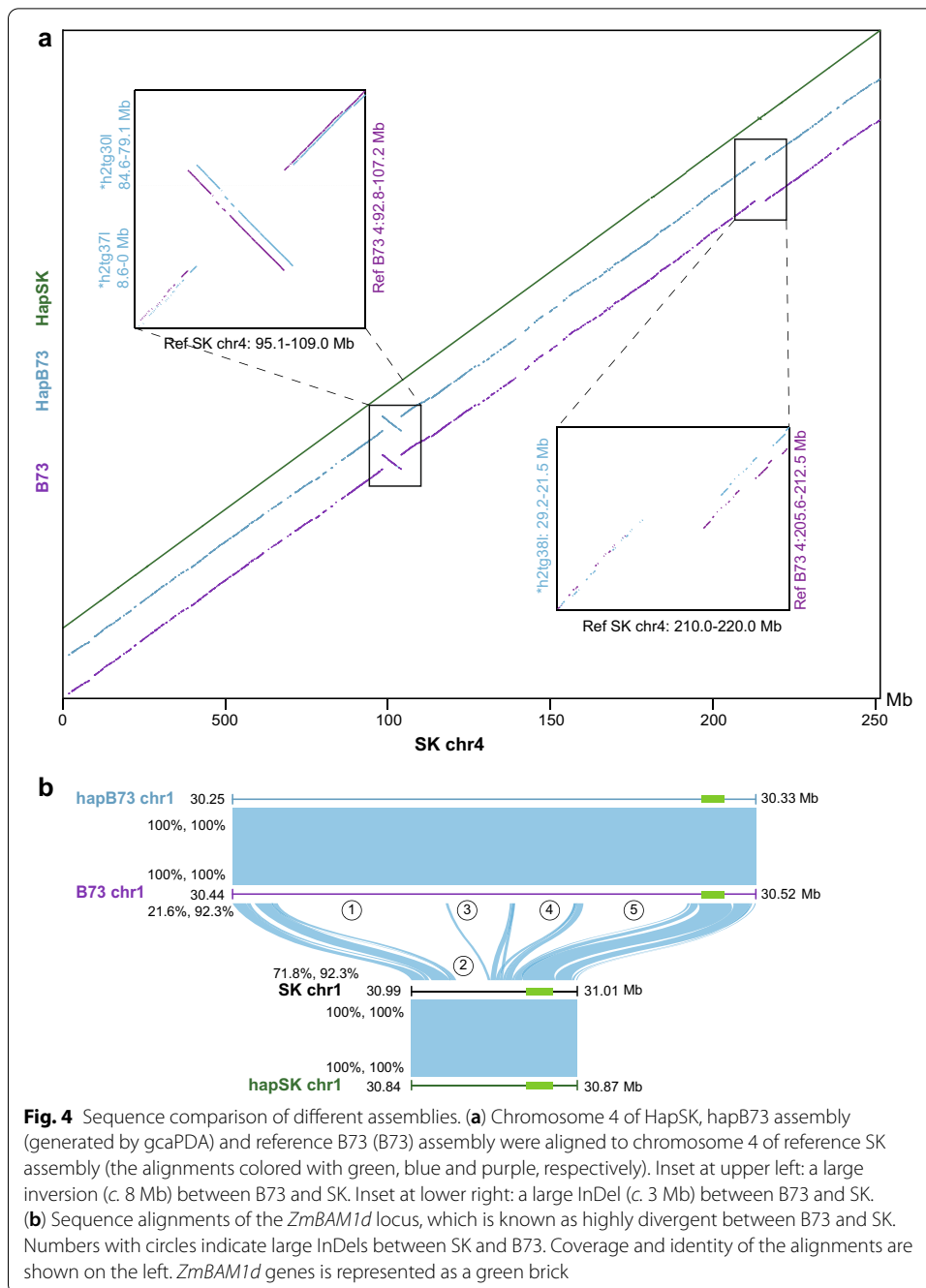
Notably, gcaPDA could phase the structural variations between B73 and SK genome properly. For example, a large inversion (*c.* 8 Mb) and a large InDel (*c.* 3 Mb) between B73 and SK genomes were correctly recovered in the hapSK and hapB73 assembly (Fig. 4a). In addition, we inspected the *ZmBAM1d* locus¹⁹ which is highly divergent between B73 and SK genomes. We found that the *ZmBAM1d* locus were also perfectly phased in the gcaPDA assembly, resulting in haplotype sequences identical to respective reference genome (Fig. 4b).

Discussion

Diploid genomes comprise two sets of homologous chromosomes that are slightly different from each other. Current diploid assemblers (such as DipAsm [14], gamete binning [16], PGAS [15]) are tailored for genomes of low complexity and not suitable for heterozygous genomes. In the present study, we developed gcaPDA, a gamete cell-based method which generates chromosome-scale haplotype-resolved diploid assembly for species with highly heterozygous genomes, thus providing access to the genetic variations present in both sets of homologous chromosomes of diploid cells.

In standard trio binning [13], parental data are used to resolve haplotypes. For a highly heterozygous individual, it is natural to assume that its parents are highly heterozygous, too. Notably, phasing information can be fuzzy at sites where all trio-samples have a heterozygous genotype [10]. And parent information is not always available, which limits the application of trio binning method. In contrast, gcaDPA does not require parental information. The gamete cells used to generate input data for gcaPDA are naturally haploid, and preserves unambiguous chromosome-scale phasing information. Thus, gcaPDA's performance is not affected by the heterozygosity level of the sequenced individual.





In DipAsm [14], gamete binning [16], PGAS [15], and gcaDPA, an initial assembly is generated for SNP calling and phasing (Additional file 1: Figure S1). For highly heterozygous genome, the initial assembly may contain haplotigs, which lead to missing SNP calls (false negative) [18]. In addition, SNP calling at repetitive genomic regions are prone to generate false positive calls [29]. Both false positive and negative SNP calls can complicate SNP-based long reads partition in DipAsm, gamete binning, and PGAS methods. Furthermore, partition long reads before assembly process is not the recommended way of diploid assembly [10]. In contrast, gcaPDA partitions gamete cell reads

based on haplotype blocks instead of individual SNPs, to increase the tolerance to potential false negative and positive SNP calls. In addition, gcaPDA used all the HiFi reads to construct assembly graphs, with haplotype-specific k-mer derived from gamete cell reads to assist in resolving graph, and generate both haplotype assembly simultaneously (Additional file 1: Figures S1 and S2). Furthermore, the input data required by gcaPDA were generated by standard wet-lab and sequencing approaches which are accessible to researchers. In contrast, gamete binning needs to sequence hundreds of thousands of gamete cell genomes by 10× sc-CNV sequencing [16], while this service is no longer supported by 10× Genomics. The Hi-C data used by DipAsm were generated with four restriction enzymes based on a modified protocol with Arima-HiC kit [14] to achieve uniform per-base coverage of the genome and maintain the highest long-range contiguity signal, while it may be difficult to generate Hi-C data of comparable quality for plant species. Generation of Strand-seq [30] data required by PGAS pose great challenge for non-human species, which limits the application of PGAS.

There are several factors that affect the performance of gcaPDA. First, contiguity and accuracy the initial assembly. The genome of F₁ hybrid is highly repetitive and heterozygous, which poses a great challenge for FALCON and results in abundant short contigs. Short contigs that couldn't be scaffolded into chromosomes won't be phased by gcaPDA, while short contigs wrongly placed during Hi-C scaffolding analysis resulting in mis-assemblies (false inversion, translocation) in the initial assembly [31], which in turn lead to chunks of gamete cells reads wrongly assigned to the other haplotype. Improving contiguity of the initial assembly with longer HiFi reads, or incorporating optical mapping data to correct mis-placed sequences [32] might mitigate these issues and improve the performance of gcaPDA. Second, number of gamete cells. In this study, 31 gamete cells (~20× WGS data for each cell) were used by gcaPDA. According to k-mer coverage accumulation curve (Additional file 1: Figure S10), 20 gamete cells shall suffice. However, with more gamete cells sequenced, phasing errors introduced by mis-placed contigs shall be alleviated and higher phasing accuracy can be achieved by gcaPDA. Third, gcaPDA integrates Hi-C and gamete cells reads to assist genome assembly and phasing resulting in higher computational resources (Additional file 1: Tables 12–13).

All in all, taking the assembly of a real large, highly heterozygous and repetitive maize F₁ hybrid genome as the positive control, we proved that gcaPDA could generate high quality haplotype-resolved and chromosome-scale diploid assembly for diploid species. In contrast to other diploid assemblers, gcaPDA does not rely on paternal information, tremendous amount of gamete cells or special sequencing approaches. As a result, gcaPDA is a good alternative option to perform haplotype-resolved genome assembly.

Materials and methods

Hi-C library construction and sequencing for F₁ hybrid

The maize F₁ hybrid (B73 × SK) seeds were from our own lab in Huazhong Agricultural University. The rice F₁ hybrid (MH63 × ZS97) seeds were provided by Dr. Xiangchun Zhou in Huazhong Agricultural University. The maize F₁ hybrids were planted in 2020 Wuhan, China. The rice F₁ hybrids were planted in 2021 Wuhan, China. Young root tissues of F₁ hybrid were harvested. Hi-C proximity libraries were constructed by

the previously described method with restriction enzyme MboI [4]. The libraries were size-selected to retain 350 bp DNA fragments and sequenced on MGISEQ2000 platform (MGI-Tech) to generate 150 bp paired-end reads.

Generating HiFi reads for rice F₁ hybrid

High molecular weight DNA was extracted from young root of the rice hybrid F₁ using modified cetyltrimethylammonium bromide (CTAB) method [33]. PacBio HiFi library was constructed and sequenced with PacBio Sequel II system in BGI-Shenzhen according to manufacturer's guidance.

Simulating reads for maize F₁ hybrid

Reference genome sequences of *Zea mays* var. SK was downloaded from ZEAMAP database (http://www.zeamap.com/ftp/01_Genomics/Genomes/) [34], while reference genome sequences of *Zea mays* var. B73 was downloaded from Ensembl Plant database [35] (release 46). Reference genome sequences of MH63 and ZS97 was downloaded from https://rice.hzau.edu.cn/rice_rs3/. In order to lift the contiguity limit capped by the reference sequences, unambiguous bases ('N's) within sequences were removed to generate gapless sequences. Reads were simulated based on SK or B73 gapless sequences. For short reads simulation, wgsim (parameters: -e 0.01) from samtools package [36] (version 1.9) was used. Pbsim [37] (version 1.0.4) was used to simulate PacBio HiFi reads with read length and quality score randomly sampled from the previously published PacBio HiFi data [38].

Genome survey analysis

Genome survey analysis was performed to profile features of the F₁ hybrid genome. Simulated SK, B73 and real rice HiFi reads were broken into k-mer and then counted by Jellyfish [39] (version 2.2.10, k=21). Genome features such as haploid genome size, heterozygosity, repeat content were estimated using genomescope [40] (version 1) with k-mer frequencies outputted by Jellyfish.

Single cell isolation, DNA extraction, amplification and sequencing

The seeds of F₁ hybrid were planted and then immature tassels were harvested before they had emerged. Gamete cells (microspores) were isolated from tetrads as described in a previous study [23]. DNA was extracted from each gamete cell using QIAGEN REPLI-g Single Cell Kit (Cat No. 150343), followed by multiple displacement amplification (MDA) [22] procedure to generate enough DNA for downstream experiments. A sequencing library was constructed for each gamete cell and then to be sequenced on MGISEQ2000 platform (MGI-Tech) to generate 150 bp paired-end reads.

Generating initial assembly

HiFi reads were assembled into contigs using FALCON [9] (version 1.4.4) from pb-assembly packages (<https://github.com/PacificBiosciences/pb-assembly# citations>), with parameter settings:

```
input_type = preads ovlp_daligner_option = -e.98 -s1000 -h1024 -l2500 -k25  
ovlp_HPCdaligner_option = -v -B4 -M35 -T6
```

```
overlap_filtering_setting=-max-diff 100 -max-cov 150 -min-cov 4 -n-core 24 -bestn
10 -min-len 5000
```

```
length_cutoff_pr=5000
```

This assembly was referred to as FALCON assembly in the main text.

To link contigs to superscaffolds, Hi-C scaffolding analysis was performed. Briefly, Hi-C reads were pre-processed using HiC-Pro program [41] (v2.11.1) and then mapped to contigs using BWA [42] (v0.7). Low quality mapping (MAPQ=0) and duplicates were removed and Hi-C contact matrices were calculated using Juicer [43] (v1.6.2). The Hi-C contact matrices were fed to the 3D-DNA pipeline [31] (v180922, parameters: -m haploid -r 0) to order and orient contigs. Potential mis-joins were corrected manually to generate superscaffolds. The longest 10 superscaffolds were considered as pseudochromosomes of the initial assembly.

Reconstruction of haplotypes and gamete cell reads partition

The raw reads of gamete cell were preprocessed to filter adapter sequences and low-quality reads using bbduk.sh (parameters: ktrim=r k=17 mink=7 hdist=1 tpe tbo qtrim=rl trimq=15 minlength=80) from BBTools package (<https://sourceforge.net/projects/bbmap/>). Reads that passed quality filtering were then mapped to the initial assembly using bowtie2 [44] (version 2.3.5.1), with parameters “-X 800”. Single nucleotide polymorphisms (SNPs) were identified using bcftools [45] ‘mpileup’ (version 1.8, parameters: -d 500 -q 10 -ff SECONDARY), followed by bcftools ‘call’ (parameters: -Ob -cv -p 0.01) implementation. Only bi-allelic SNPs, with quality value ≥ 20 and allele frequency between 0.3 and 0.7 were selected. SNPs that located adjacent to (<5 bp) InDels were also filtered. Furthermore, SNPs in each cell sample with supporting reads depth <5 were replaced with ‘NA’ and treated as missing calls. Heterozygous rate and missing rate were calculated for each cell based on SNP calls. Since we sequenced haploid single cells, most of the SNPs identified in each cell should be homozygous, with lots of missing calls due to insufficient read coverage. Cells with abnormal level of heterozygous rate of SNPs (>5%) or missing call rate (<30% or >70%) indicate contamination or insufficient sequencing coverage, hence were considered as low-quality cells and excluded from downstream analyses.

The paternal and maternal haplotypes were reconstructed based on SNP array of qualified cells using Hapi [24]. By comparing genotypes of each gamete cell with reconstructed parental haplotypes, haplotype blocks can be identified for each cell. Gamete cell reads that mapped to haplotype blocks with the same parental origin were extracted and merged. Due to MDA procedure, the read depth of each cell is unevenly distributed across the genome. To mitigate this issue, merged reads of each haplotype were normalized by k-mer depth using BBnorm.sh (<https://sourceforge.net/projects/bbmap/>), to mimic regular whole genome sequencing (WGS) reads. The normalized short reads were then used as parental WGS data by diploid *de novo* assembler.

De novo genome assembly

The normalized short reads of gamete cells of each haplotype were broken into k-mers using yak (<https://github.com/lh3/yak>), respectively. HiFi reads together with parental k-mers, were *de novo* assembled using hifiasm [10] to generate phased diploid genome

assembly. The haplotype-resolved contigs were linked into superscaffolds using Hi-C reads, with methods described above. The final results of gcaPDA were two haplotype-resolved chromosome-level assemblies: hapSK assembly and hapB73 assembly, hapMH63 and hapZS97 assembly. This final assembly was referred to as gcaPDA assembly in the main text.

In the same time, another haplotype-resolved diploid assembly were generated with HiFi reads and simulated parental WGS short reads using hifiasm. This assembly was referred to as Trio assembly.

A pseudo-haplotype assembly was generated using hifiasm, with only HiFi reads. This assembly was referred to as Hifiasm assembly.

A haplotype-resolved assembly was generated using hifiasm with HiFi and Hi-C reads. This assembly was referred to as Hifiasm + Hi-C assembly.

Evaluation of genome assemblies

We broke the gapless reference genomes of B73 and SK, MH63 and ZS97 into k-mer using meryl utility from Merqury package [46] (release 20200430). Total k-mer set and haplotype-specific k-mer set (hapmers) were computed based on B73 and SK k-mer set and MH63 and SK k-mer set. Genome completeness of each assemblies was evaluated with total k-mer set and phasing accuracy were evaluated with hapmers using Merqury package [46].

Gene completeness of the assemblies was evaluated using BUSCO [26] (version 3.0.2, lineage setting: embryophyta_odb9).

Whole genome sequence comparison between assemblies and reference genomes were performed using nucmer and visualized using mummerplot from MUMMER package [47] (version 4.0). Coverage and identity of the alignments were calculated using dnadiff implementation from MUMMER package. Only alignments span > 10 Kb were counted.

Availability and requirements

Project name: gcaPDA

Project home page: <https://github.com/BGI-shenzhen/gcaPDA>

Operating system(s): e.g. Platform independent

Programming language: Perl, R

Other requirements:

R(v3.4.3): packages such as optparse, HMM, Hapi are required.

FALCON (falcon-kit 1.4.4): <https://github.com/PacificBiosciences/pb-assembly>

juicer: <https://github.com/aidenlab/juicer>

3d-dna: <https://github.com/aidenlab/3d-dna>

bbmap: <https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbmap-guide>

bowtie2 (v2.3.5.1): <http://bowtie-bio.sourceforge.net/bowtie2/index.shtml>

samtools (v1.9): <http://www.htslib.org/download>

bcftools (v1.9): <http://www.htslib.org/download>

Hapi: <https://cran.r-project.org/web/packages/Hapi/>

yak: <https://github.com/lh3/yak>
hifiasm: <https://github.com/chhylp123/hifiasm>

License: No restrictions on non-commercial use.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12859-022-04591-4>.

Additional file 1. Supplementary figures and tables.

Acknowledgements

The computations in this paper were run on the bioinformatics computing platform of the National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University.

Authors' contributions

J-BY, NY and L-FY designed the project. C-LJ, NY, CL, S-SW contributed to sample preparation and wet-lab experiments. MX, L-FY, XY, L-JH, S-XC, T-QD and M-ZY performed all data analysis. MX wrote the first draft of this manuscript. MX, NY, and J-BY wrote the final manuscript. All authors have read and approved the final manuscript.

Funding

This research was supported by the National Natural Science Foundation of China (31900494, 31730064) and Young Elite Scientists Sponsorship Program by CAST (2019QNRC001). The funding bodies did not participate in any roles in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript other than the financial support.

Availability of data and materials

The data reported in this study are available in the CNGB Nucleotide Sequence Archive (CNSA: <https://db.cngb.org/cnsa>; accession number CNP0001702 (maize), CNP0002490 (rice)). The code of gcaPDA are available at <https://github.com/BGI-shenzhen/gcaPDA>.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

M.X., L-F.Y., X.Y., L-J.H., S-X.C., T-Q.D. and M-Z.Y. are employees of BGI-Shenzhen. The gcaPDA methodology are covered in pending patents.

Author details

¹National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China.

²Guangdong Engineering Research Center of Plant and Animal Genomics, BGI Genomics, BGI-Shenzhen, Shenzhen 518083, China. ³Hubei Hongshan Laboratory, Wuhan 430070, China.

Received: 21 December 2021 Accepted: 29 January 2022

Published online: 14 February 2022

References

1. Eid J, et al. Real-time DNA sequencing from single polymerase molecules. *Science*. 2009;323:133–8.
2. Bayley H. Nanopore sequencing: from imagination to reality. *Clin Chem*. 2015;61:25–31.
3. Lam ET, et al. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat Biotechnol*. 2012;30:771–6.
4. Lieberman-Aiden E, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*. 2009;326:289–93.
5. Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. *Nat Methods*. 2020;17:155–8.
6. Xiao CL, et al. MECAT: fast mapping, error correction, and *de novo* assembly for single-molecule sequencing reads. *Nat Methods*. 2017;14:1072–4.
7. Koren S, et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res*. 2017;27:722–36.
8. Zhang X, Wu R, Wang Y, Yu J, Tang H. Unzipping haplotypes in diploid and polyploid genomes. *Comput Struct Biotechnol J*. 2020;18:66–72.
9. Chin CS, et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods*. 2016;13:1050–4.

10. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nat Methods*. 2021;18:170–5.
11. Heller D, Vingron M, Church G, Li H, Garg S. SDip: A novel graph-based approach to haplotype-aware assembly based structural variant calling in targeted segmental duplications sequencing. *bioRxiv*. 2020. <https://doi.org/10.1101/2020.1102.1125.964445>.
12. Garg S, et al. A haplotype-aware *de novo* assembly of related individuals using pedigree sequence graph. *Bioinformatics*. 2020;36:2385–92.
13. Koren S, et al. *De novo* assembly of haplotype-resolved genomes with trio binning. *Nat Biotechnol*. 2018;36:1174–82.
14. Garg S, et al. Chromosome-scale, haplotype-resolved assembly of human genomes. *Nat Biotechnol*. 2021;39:309–12.
15. Ebert P, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*. 2021;372:eabf7117.
16. Campoy JA, et al. Gamete binning: chromosome-level and haplotype-resolved genome assembly enabled by high-throughput single-cell sequencing of gamete genomes. *Genome Biol*. 2020;21:306.
17. Shi D, et al. Single-pollen-cell sequencing for gamete-based phased diploid genome assembly in plants. *Genome Res*. 2019;29:1889–99.
18. Roach MJ, Schmidt SA, Borneman AR. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinform*. 2018;19:460.
19. Sun S, et al. Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. *Nat Genet*. 2018;50:1289–95.
20. Jiao Y, et al. Improved maize reference genome with single-molecule technologies. *Nature*. 2017;546:524–7.
21. Yang N, et al. Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. *Nat Genet*. 2019;51:1052–9.
22. Dean FB, Nelson JR, Giesler TL, Lasken RS. Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res*. 2001;11:1095–9.
23. Li X, Li L, Yan J. Dissecting meiotic recombination based on tetrad analysis by single-microspore sequencing in maize. *Nat Commun*. 2015;6:6648.
24. Li R, et al. Inference of chromosome-length haplotypes using genomic data of three or a few more single gametes. *Mol Biol Evol*. 2020;37:3684–98.
25. Song JM, et al. Two gap-free reference genomes and a global view of the centromere architecture in rice. *Mol Plant*. 2021;14:1757–67.
26. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 2015;31:3210–2.
27. Zhou Q, et al. Haplotype-resolved genome analyses of a heterozygous diploid potato. *Nat Genet*. 2020;52:1018–23.
28. Porubsky D, et al. Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nat Biotechnol*. 2020;39:302–8.
29. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet*. 2011;13:36–46.
30. Sanders AD, Falconer E, Hills M, Spierings DCJ, Lansdorp PM. Single-cell template strand sequencing by Strand-seq enables the characterization of individual homologs. *Nat Protoc*. 2017;12:1151–76.
31. Dudchenko O, et al. *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*. 2017;356:92–5.
32. Udall JA, Dawe RK. Is it ordered correctly? Validating genome assemblies by optical mapping. *Plant Cell*. 2018;30:7–14.
33. Doyle JJ, Doyle JL. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem Bull Bot Soc Am*. 1987;19:11–5.
34. Gui S, et al. ZEAMAP, a comprehensive database adapted to the maize multi-omics era. *iScience*. 2020;23:101241.
35. Bolser DM, Staines DM, Perry E, Kersey PJ. Ensembl plants: integrating tools for visualizing, mining, and analyzing plant genomic data. *Methods Mol Biol*. 2017;1533:1–31.
36. Li H, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;25:2078–9.
37. Ono Y, Asai K, Hamada M. PBSIM: PacBio reads simulator toward accurate genome assembly. *Bioinformatics*. 2013;29:119–21.
38. Wenger AM, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol*. 2019;37:1155–62.
39. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*. 2011;27:764–70.
40. Vurture GW, et al. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*. 2017;33:2202–4.
41. Servant N, et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol*. 2015;16:259.
42. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25:1754–60.
43. Durand NC, et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst*. 2016;3:95–8.
44. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9:357–9.
45. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetic parameter estimation from sequencing data. *Bioinformatics*. 2011;27:2987–93.
46. Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol*. 2020;21:245.
47. Marçais G, et al. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput Biol*. 2018;14:e1005944.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.