Assembly of whole-chromosome pseudomolecules for polyploid plant genomes using outbred mapping populations

Chenxi Zhou1,2, Bode Olukolu3,4, Dorcus C. Gemenet5,6, Shan Wu7, Wolfgang Gruneberg5, Minh Duc Cao1, Zhangjun Fei7, Zhao-Bang Zeng8,9, Andrew W. George10, Awais Khan5,11, G. Craig Yencho3 and Lachlan J. M. Coin1,2,12

Despite advances in sequencing technologies, assembly of complex plant genomes remains elusive due to polyploidy and high repeat content. Here we report PolyGembler for grouping and ordering contigs into pseudomolecules by genetic linkage analysis. Our approach also provides an accurate method with which to detect and fix assembly errors. Using simulated data, we demonstrate that our approach is of high accuracy and outperforms three existing state-of-the-art genetic mapping tools. Particularly, our approach is more robust to the presence of missing genotype data and genotyping errors. We used our method to construct pseudomolecules for allotetraploid lawn grass utilizing PacBio long reads in combination with restriction site-associated DNA sequencing, and for diploid Ipomoea trifida and autotetraploid potato utilizing contigs assembled from Illumina reads in combination with genotype data generated by single-nucleotide polymorphism arrays and genotyping by sequencing, respectively. We resolved 13 assembly errors for a published I. trifida genome assembly and anchored eight unplaced scaffolds in the published potato genome.

High-quality genome assembly plays an essential role in plant genomic and genetic analyses. Even though the recent advances in third-generation long-read sequencing technologies significantly improved the contiguity of genome assemblies; it is so far not possible to construct a complete polyploid plant genome with sequence data alone due to polyploidy and high repeat content. Long-distance linkage information, such as physical maps, genetic maps, optical maps, syntenic maps, chromatin interactions and Hi-C contact maps, is crucial in constructing complete genome assemblies. Genetic maps have been widely adopted as they provide chromosomal-scale linkage information. The construction of many high-quality chromosomal-level genome assemblies for plants involved scaffolding with genetic linkage maps as the last step for building pseudomolecules. The idea is straightforward: a genetic map is constructed for genetic markers associated with scaffolds and the map is then used to anchor scaffolds to construct pseudomolecules. The completeness of the pseudomolecules largely depends on the density of the genetic linkage map. To obtain a full map of chromosomes, genetic markers should cover as much of the genome as possible. However, genome-wide genetic marker discovery for a mapping population with a considerable number of individuals is nontrivial. Several methods have been proposed for high-throughput genetic marker discovery with cost-effective next-generation sequencing (NGS) technology, such as reduced-representation libraries6, restriction site-associated DNA sequencing (RAD-seq)7 and genotyping by sequencing (GBS)8. The genotype data generated with NGS-based high-throughput genetic marker discovery methods, including reduced-representation libraries, RAD-seq and GBS, are usually of large scale with an abundance of missing values and genotyping errors. Genetic linkage analysis for such datasets is challenging. The conventional genetic mapping tools, such as MAPMAKER9 and R/qtl10, have been optimized for relatively small but high-quality marker sets. Moreover, these tools were designed for inbred lines. The development of inbred lines, however, could be difficult, expensive and time consuming, especially for polyploids. Several methods have since been proposed for outbred lines, including OneMap11, JoinMap11 and Lep-MAP2 (ref. 12), but only for diploids. TetraploidMap was the first tool specifically designed for mapping outbred tetraploid species based on dominant and codominant marker information. This tool was later extended to TetraploidSNPMap to make full use of the single-nucleotide polymorphism (SNP) dosage data. polymapR is another genetic mapping tool for outbred polyploids utilizing SNP dosage data. Due to the underlying design, however, these methods are intrinsically sensitive to missing values and genotyping errors, and thus require stringent quality control on the input genetic markers. For whole-chromosome pseudomolecule construction, this could dramatically reduce the proportion of the genome covered by the genetic linkage map.

Here we describe a novel method for the construction of whole-chromosome pseudomolecules for polyploids by genetic mapping. This method relies on the availability of a high-density marker
Fig. 1 | PolyGembler framework. a, Construction of reference contigs and scaffolds and genotyping of mapping populations with high-throughput genome-wide genetic marker discovery technologies. Although we show the genotype data here, PolyGembler can handle allele depth data. b, Construction of haplotypes for each scaffold and for each sample with a HMM. A well-trained HMM for a diploid genotype is depicted in the top panel. The scaffold consists of ten markers illustrated as ellipses. The numbers within the ellipses indicate the probabilities of the parental haplotypes associating with the A and B allele. The width of the line connecting two ellipses is proportional to the RF between parental haplotypes at that position. Recombinations are prohibited between paternal and maternal haplotypes. The model assumes that each F1 progeny inherits one haplotype from each parent’s gamete and maximizes the likelihood of the marker data. This model can identify assembly errors. The RF between the fifth and sixth markers is abnormally large, as indicated by the numeric vaules in the bottom panel, suggesting a misassembly. In the middle panel, we show the positions of markers along the scaffold. c, RF estimation between each pair of scaffolds by counting the mismatches of parental haplotypes. d, The estimated RFs reflect the physical distances. The dot plot and histogram are for scaffold pairs from the same chromosome and different chromosomes, respectively. e, Linkage group construction for the reference scaffolds with a graph partitioning algorithm. f, The order of the scaffolds within each linkage group is determined by the MDS algorithm. The directions of the scaffolds are further determined by solving a modified TSP.

Results

Overview of method. We have developed a new method called PolyGembler (polyploid genetic linkage assembler) for the assembly of polyploid genomes by genetic linkage analysis. Figure 1 provides an overview of PolyGembler. The method assumes the availability of genome-wide data for genotyping, such as GBS, RAD-sequencing and array data, collected on an F1 outbred family, as well as high coverage (that is, greater than 30x) whole-genome sequence data on a reference sample, or alternatively a set of reference contigs or scaffolds. First, the genotyping data of the mapping population are mapped to the reference scaffolds to call variants (Fig. 1a). The resultant genotype data or allele depth data are then used to infer haplotypes for each scaffold. The haplotypes are utilized to detect assembly errors and to calculate recombination fractions (RFs) between each pair of scaffolds (Fig. 1b–d). Next, the linkage information between the scaffold pairs is used for a graph-based clustering algorithm for linkage group construction (Fig. 1e). Ideally, the scaffolds in each linkage group are originated from the same chromosome. The order of the scaffolds in each linkage group is determined by a multidimensional scaling (MDS) algorithm and the directions of the scaffolds are calculated by solving a carefully designed traveling salesman problem (TSP) to minimize the sum of adjacent RFs along the linkage map (Fig. 1f). Finally, the scaffold-based genetic linkage maps are used to construct pseudomolecules.

Application to simulated data. Simulated data were used to evaluate PolyGembler. First, we simulated a diploid genome and a tetraploid genome. Genome-wide Illumina short reads and nanopore long reads were then simulated, respectively, from the diploid and tetraploid genome for genome assembly. To construct pseudomolecules from the genome assemblies, we also simulated outbred mapping populations consisting of two parents and 190 F1 progeny and then further simulated GBS data (Methods).
The longest scaffold was ~625 kilobases (kb) and the N50 statistic simulated data consisted of 83,325 scaffolds of ~493 megabases (Mb).

Pseudomolecules constructed from the diploid (d) and tetraploid (e) mapping populations. Collinear plots between the reference chromosomes and the pseudomolecules constructed from the diploid (e) and tetraploid (f) mapping populations.

**Fig. 2 | Pseudomolecule construction for simulated datasets.** a-c. Dot plots of the physical distances against the estimated RFs for scaffold pairs mapped to the same reference chromosome for diploid (a) and tetraploid (c) mapping populations. b-d. Histograms of the estimated RFs for scaffold pairs mapped to different chromosomes for diploid (b) and tetraploid (d) mapping populations. e-f. Collinear plots between the reference chromosomes and the pseudomolecules constructed from the diploid (e) and tetraploid (f) mapping populations.

**Diploid simulation.** The genome assembly constructed from the simulated data consisted of 83,325 scaffolds of ~493 megabases (Mb). The longest scaffold was ~625 kilobases (kb) and the N50 statistic was ~90 kb. The GBS data were simulated at ~10× around restriction recognition sites (Supplementary Note 1). We identified 24,488 SNPs from the GBS data for pseudomolecule construction (Methods).
total of 3,348 scaffolds of ~338 Mb were covered by these SNPs, constituting ~69% of the genome. Among the 20 assembly errors confirmed with the Quality Assessment Tool for Genome Assemblies (QUAST; Methods), PolyGembler identified three without false positives. To investigate the accuracy of RF estimation, we mapped the scaffolds back to reference chromosomes to determine the physical positions (Methods). For scaffold pairs that located on the same chromosomes, the estimated RFs correlated with the physical distances very well, while for scaffold pairs that mapped to different chromosomes the RFs were consistently large (Fig. 2a,b). The linkage groups constructed by this method perfectly recognized 15 reference chromosomes. Within each linkage group, the order of the mapped scaffolds demonstrated high correlation with the true order on the corresponding reference chromosome (Fig. 2c). The Spearman’s rank correlation coefficients between them ranged from 0.967–1.000.

Tetraploid simulation. A total of ~10.9 million nanopore reads of ~48.5 gigabases (Gb) with an N50 of ~6.7 kb were simulated, representing ~100x depth of coverage of the genome (Methods). The resultant genome assembly consisted of 722 scaffolds of ~484 Mb with an N50 of ~1.1 Mb. The GBS data were simulated at ~40x around restriction recognition sites (Supplementary Note 1). We identified 58,733 SNPs from the GBS data located on 707 scaffolds for pseudomolecule construction. These scaffolds covered nearly the whole genome. Out of the 135 misassembled scaffolds confirmed with QUAST (Methods), PolyGembler identified 86 without false positives. The results for RF estimation were similar to those for diploid simulation (Fig. 2c,d). Fifteen pseudomolecules corresponding to reference chromosomes were constructed without misgrouping. The resultant pseudomolecules were highly collinear with reference chromosomes (Fig. 2f). The Spearman’s rank correlation coefficients between the scaffold order determined by PolyGembler and the true order fell within the range of 0.956–1.000. To investigate the proposed method thoroughly, we also simulated GBS data at ~20x. With the lower-depth data, the number of assembly errors detected by PolyGembler significantly decreased (19 out of 135). However, the results for linkage mapping and pseudomolecule construction were still quite accurate (Extended Data Fig. 1).

Application to *Ipomoea trifida* GBS data. The *Ipomoea trifida* mapping population consists of two parents and 210 F1 progeny. It was derived from the crossing of two diploid *Ipomoea* heterozygous genotypes and is called M9×M19. We ran PolyGembler with the GBS data of this mapping population to construct pseudomolecules from the de novo genome assembly ITR_r1.0 (ref. 13). The ITR_r1.0 genome assembly consists of 77,400 scaffolds. The total size is ~512 Mb and the N50 statistic is ~43 kb. We identified 35,646 SNPs covering 2,469 scaffolds of ~231 Mb, constituting ~45% of the genome. We used another *Ipomoea trifida* genome assembly, NCNSP0306 (ref. 3), as a reference genome to validate the results.

We detected 13 misassembled scaffolds in ITR_r1.0. Figure 3a depicts the RF estimations along the scaffold Itr_sc000015 with 30 independent runs. There was a potential misassembly around 177 kb, where the mean RF estimation was 0.432 (s.d. = 0.012). By mapping it to NCNSP0306 chromosomes, we found that while the first ~166 kb was mapped to chromosome 4, the last ~138 kb was mapped to chromosome 8 (Fig. 3b,c). The mapping results supported the misassembly of scaffold Itr_sc000015, as well as the misassembly position. The other misassemblies detected by PolyGembler were also confirmed by mapping to NCNSP0306 chromosomes (Supplementary Table 1). We split the potentially misassembled ITR_r1.0 scaffolds to generate new scaffolds and excluded the original ones from linkage analysis.

We mapped the scaffolds to the NCNSP0306 chromosomes to investigate the accuracy of RF estimation. When the physical distance was small, it showed a strong correlation with the estimated RFs. As the physical distance increased, the estimated RFs clearly deviated from the expected values calculated from conventional genetic mapping functions (Fig. 4a). One possible explanation is that the calculation of physical distances between scaffolds was not accurate due to structural variations between the two genome varieties or misassemblies in the reference chromosomes. Another possible reason, and probably the major reason, is the absence of genetic recombinations in large pericentromeric heterochromatin blocks in the *Ipomoea trifida* genome, as has been widely reported for the potato genome17. The RFs for scaffolds mapped to different chromosomes were consistently large (Fig. 4b).

Fifteen pseudomolecules in high concordance with the reference chromosomes were constructed (Fig. 4c). The pseudomolecules were not well assembled in several pericentromeric regions. This was mainly because these regions are highly repetitive and the original genome assembly ITR_r1.0 is extremely fragmented in these regions. To complete the genome assembly, we probably need to improve the quality of scaffolds in these regions; for example, with long reads. There was a clear discrepancy between the pseudomolecule and NCNSP0306 chromosome 3 (Fig. 4e). It is not known whether it was a misassembly by our approach, a misassembly in NCNSP0306 chromosome 3 or a structure variation between the two genome varieties. We mapped the pseudomolecule to the genome assembly of *Ipomoea nil*18—another species from the genus *Ipomoea*. The result showed that the pseudomolecule represented a better collinearity with *Ipomoea nil* chromosome 2, suggesting that NCNSP0306 chromosome 3 might be misassembled (Extended Data Fig. 2).
Fig. 4 | Pseudomolecule construction for M9 × M19 and B2721 mapping populations. a–c. Dot plots of the physical distances against the estimated RFs for scaffold pairs mapped to the same reference chromosome for the M9 × M19 (a) and B2721 (c) mapping populations. b, d. Histograms of the estimated RFs for scaffold pairs mapped to different chromosomes for the M9 × M19 (b) and B2721 (d) mapping populations. e. Collinear plots between the NCNSP0306 reference chromosomes and the pseudomolecules constructed from the M9 × M19 mapping population. f. Genetic linkage map constructed from the B2721 mapping population.

Application to Infinium 8303 potato array data. Two outbred mapping populations genotyped with the Infinium 8303 SNP array were used to construct pseudomolecules from Potato Genome Sequencing Consortium (PGSC) doubled-monoploid version 3 scaffolds. The B2721 mapping population was derived from the cross Atlantic × B1829-5, consisting of 156 F1 progeny. A total of 5,736 SNPs were identified covering 721 scaffolds, of which 175 scaffolds with fewer than three SNPs were excluded from the further
The other mapping population was derived from the cross 12601ab1 × Stirling, consisting of 190 F1 progeny. The variant calling ended up with 4,109 SNPs from 679 scaffolds, of which 455 scaffolds with at least three SNPs were retained. The phasing algorithm failed to generate any feasible haplotype calls for 79 and 100 scaffolds, respectively, for the two mapping populations due to a high level of homozygosity (Supplementary Note 2). The pairwise RF estimations showed very similar patterns to those for the *I. trifida* dataset (Fig. 4c,d and Extended Data Fig. 3a,b). For genetic linkage map construction, scaffolds were assigned to 12 linkage groups for both datasets, corresponding to the 12 PGSC version 4.03 chromosomes, without misassignment. The order of scaffolds within each linkage group was consistent with the order on chromosomes despite a few incorrect orderings for closely located scaffolds (Fig. 4f and Extended Data Fig. 3c). The shapes of the genetic linkage maps were nearly identical to those constructed from TetraploidSNPMap (Extended Data Figs. 4 and 5). The total sizes of the anchored scaffolds were ~429 and ~366 Mb, constituting ~59 and ~53% of the genome, respectively. We anchored eight scaffolds of ~2.8 Mb, which were not placed in any PGSC version 4.03 chromosomes, from the two mapping populations (Supplementary Table 2).

**Application to *Z. japonica* RAD-seq data.** *Z. japonica* has an allo-tetraploid genome with 20 chromosomes. A total of ~4.2 million PacBio reads of ~38.6 Gb with an N50 of ~12.2 kb were assembled for the *Z. japonica* cultivar Yaji (Methods). The resultant genome assembly consisted of 1,075 scaffolds of ~329 Mb with an N50 of ~2.6 Mb. The RAD-seq data of a mapping population comprising 60 F1 progeny derived from a cross between Carrizo and El Toro were used for pseudomolecule construction with PolyGembler. We identified 25,080 SNPs covering 297 scaffolds of ~303 Mb, constituting ~92% of the genome. PolyGembler identified six misassembled contigs of approximately 24 Mb, which were split before linkage analysis. The final genetic map consisted of 21 linkage groups, of which a small linkage group comprising two contigs of ~222 kb in total was excluded from the pseudomolecule construction. The resultant

**Fig. 5 | Collinear plots between the *Z. japonica* pseudomolecules constructed from PolyGembler and the *O. sativa* chromosomes.** The *Z. japonica* pseudomolecules labeled with asterisks were converted to reverse complement before performing alignment.
### Table 1 | Results for comparisons of PolyGembler with OneMap, Lep-MAP2 and TetraploidSNPMap on simulated genotype data

| Missing data rate | Genotyping error rate | PolyGembler(2) | OneMap | Lep-MAP2 | PolyGembler(4) | TetraploidSNPMap |
|-------------------|-----------------------|---------------|--------|----------|---------------|-----------------|
|                   |                       | NMI           | R      | T(s)     | NMI           | R               |
| 0                 | 1.000                 | 1.000         | 1.000  | 1.000    | 1.000         | 1.000           |
| 0.05              | 0.998                 | 0.999         | 0.999  | 0.999    | 1.000         | 0.999           |
| 0.10              | 0.996                 | 0.997         | 0.997  | 0.997    | 1.000         | 0.999           |
| 0.20              | 0.991                 | 0.996         | 0.996  | 0.996    | 1.000         | 0.999           |
| 0.30              | 0.990                 | 0.995         | 0.995  | 0.995    | 1.000         | 0.999           |
| 0.50              | 0.989                 | 0.994         | 0.994  | 0.994    | 1.000         | 0.999           |
| 0.60              | 0.988                 | 0.994         | 0.994  | 0.994    | 1.000         | 0.999           |
| 0.80              | 0.985                 | 0.993         | 0.993  | 0.993    | 1.000         | 0.999           |
| 1.00              | 0.982                 | 0.991         | 0.991  | 0.991    | 1.000         | 0.999           |

The statistics for PolyGembler, OneMap and Lep-MAP2 are averaged on 30 independent runs (Intel Xeon Processor E5-2667 v3 @ 3.50 GHz; 32 GB RAM) while for TetraploidSNPMap they are averaged on five independent runs (Intel Core i5-2520M CPU; 2.50 GHz). Column N reports the number of SNPs in the genetic linkage map. N was sometimes <1,000 because SNPs were either filtered out or failed to be assigned to any linkage group. Column T(s) is not reported for PolyGembler, as it was always 1,000. PolyGembler(2) and PolyGembler(4) report the results for diploid and tetraploid simulations, respectively. –, no meaningful results reported.
pseudomolecules can be grouped into ten homoeologous pairs, each with similar length, ranging from approximately 10–23 Mb. We mapped the pseudomolecules to the rice reference genome (Oryza sativa release 7)\(^\text{18}\) for validation (Fig. 5). The pseudomolecules were mapped in pairs and showed clear collinearity with the rice chromosomes, suggesting a high degree of authenticity.

The dataset was also used for genetic linkage map construction in the originally study\(^\text{19}\). The genetic linkage maps reported for the parents Carrizo and El Toro consisted of 2,408 and 1,230 RAD markers, distributing across 21 and 20 linkage groups, respectively. Compared with the genetic linkage maps constructed using our approach, the genetic markers included were significantly fewer. The genetic linkage map for the parent El Toro was further used in another study for genome assembly of Z. japonica accession Nagirizaki for pseudomolecule construction\(^\text{21}\). In that study, total sequences of ~273 Mb were anchored in pseudomolecules (30 Mb fewer than the pseudomolecules constructed in the current study). We performed pairwise comparisons between the pseudomolecules from the two genome assemblies. The results showed that the two genome assemblies were highly consistent (Extended Data Fig. 6).

Comparison with well-known genetic mapping tools. We compared our genetic mapping method with three existing tools designed for outbred mapping populations, including OneMap (version 2.0-4)\(^\text{10}\) and Lep-MAP2 (ref. \(^\text{13}\)) for diploid genomes and TetraploidSNPMap\(^\text{14}\) for tetraploid genomes. The marker sets used for comparisons consisted of 1,000 SNPs sampled from a diploid and a tetraploid simulated outbred mapping population comprising 190 F\(_1\) progeny with known linkage groups and order (Methods).

The missing data rates ranged from 0–0.8, while the genotyping error rates ranged from 0–0.4. For each configuration, we conducted 30 independent experiments for PolyGembler, OneMap and Lep-MAP2 and five experiments for TetraploidSNPMap, for which only a graphical user interface was provided by the author. We compared the accuracy for grouping and ordering, as well as the running time. The accuracy for grouping was measured using the normalized mutual information (NMI) between the constructed linkage groups and true groups\(^\text{15}\). NMI ranges from 0–1. Larger NMI values suggest more similar groupings, and specifically NMI = 1 indicates identical groupings. The accuracy for ordering was measured by the switch error rate (R\(_s\)) of the marker order on the genetic linkage map compared with the true order on chromosomes\(^\text{15}\). We also investigated the phasing accuracy of PolyGembler using the correct phasing rate (R\(_p\)) as an evaluation, which was calculated as the proportion of the correctly phased loci with respect to the true haplotypes (Methods).

The results are reported in Table 1, with details in Supplementary Table 3. Evidently, PolyGembler was never worse and often better than the other methods in accuracy for grouping, and consistently outperformed the other methods in accuracy for ordering. Besides, PolyGembler can handle datasets with much higher missing and error rates. Here we call a linkage map with NMI \(\geq 0.9\) and \(R_s \geq 0.1\) acceptable. For the diploid simulation, we gained acceptable results for datasets constituting up to 70% missing genotypes from PolyGembler and 40% missing genotypes from Lep-MAP2. With OneMap, we obtained accurate groupings for datasets with up to 50% missing data, but the orderings were consistently error prone. The performance on the genotyping errors showed a similar trend. The largest error rate that PolyGembler could handle reached 30%, while for Lep-MAP2 it dropped to 5%. For OneMap, the groupings were perfect for datasets with up to 10% genotyping errors; however, the orderings were inaccurate. PolyGembler also outperformed the other two methods for the data with mixtures of missing data and genotyping errors. PolyGembler produced acceptable results for up to 30% missing data and 10% genotyping errors, while for Lep-MAP2 acceptable results were generated for up to 20% missing data and a 5% error rate. Lep-MAP2 filtered out a significant proportion of SNPs for high missing data or genotyping error rates. Regarding the running time, PolyGembler was the fastest, followed by Lep-MAP2, and both were much faster than OneMap. Lep-MAP2 was occasionally faster than PolyGembler as it discarded many SNPs. For tetraploid datasets, PolyGembler consistently outperformed TetraploidSNPMap in both grouping and ordering. TetraploidSNPMap generated acceptable results for datasets of missing data up to 40% and genotyping errors up to 5%, or a combination of them up to 10% missing data and a 2% error rate. In contrast, PolyGembler can handle datasets of missing data up to 50% or genotyping errors up to 20%, or a combination of 30% missing data and a 10% error rate. For speed, TetraploidSNPMap had an average running time of less than 1 min, which was substantially faster (40–50×) than PolyGembler. A high level of phasing accuracy was observed for PolyGembler. The \(R_p\) was consistently \(>95\)% with diploid datasets of missing data up to 70% or genotyping errors up to 20%, or a combination of them up to 30% missing data and a 10% error rate. For tetraploid datasets, the missing and error rates that could be well handled by PolyGembler slightly decreased. To gain an \(R_p\) of the same level, the datasets needed to be with a missing rate no greater than 40% or an error rate no greater than 10%, or a combination of them no greater than 20% and 5%, respectively. Since we included the SNP loci for introducing missing data and genotyping errors in the calculation of \(R_p\) (Methods), the results indicated that PolyGembler imputed the missing data as well as corrected genotyping errors with high accuracy.

Discussion
We have described a genetic anchoring method that harnesses genotype data from a mapping population and reference genome assembly to construct chromosomal-scale pseudomolecules. Several methods have been proposed under this framework, such as POPSEQ\(^\text{18}\) and RPGC\(^\text{19}\). These methods have mainly focused on integration of established tools for read mapping, variant calling and genetic mapping, to build computational pipelines for organizing contigs or scaffolds. Since they largely depend on conventional genetic mapping algorithms, these methods have several limitations. First, they require high-quality genetic markers and lack robustness to genotyping errors and missing data. However, variant calling for such high-quality genetic markers from genome-wide genotyping sequence data, especially from low-coverage sequence data, is nontrivial. Second, these methods do not scale to large datasets. The conventional genetic mapping tools have been designed to handle datasets with up to several thousand genetic markers but do not scale to hundreds of thousands of genetic markers generated with the current whole-genome genotyping approaches. Lastly, these methods can rarely be used for polyploid genomes. To solve these problems, we used a divide-and-conquer strategy for genetic anchoring. We first performed haplotype phasing using a hidden Markov model (HMM) at the scaffold level. The genetic distances between each pair of scaffolds were then estimated from the haplotypes and further used to perform a scaffold-level linkage analysis. Finally, the genetic linkage maps of scaffolds were converted to whole-chromosome pseudomolecules.

Scaffold-level linkage analysis effectively reduces the computational complexity. Presumably, a well-designed heuristic algorithm is required to order the markers within each linkage group as it is a non-deterministic polynomial hard problem. However, in our method, the size of the problem is dramatically reduced, which enables us to combine the MDS algorithm with the exact TSP solver CONCORDE to order markers efficiently and accurately. The scaffold-level design also enables us to perform haplotype phasing in a highly parallel manner. In the present study, the method was used to handle datasets with up to 3,348 scaffolds. Moreover, the computational time and resources required for haplotype phasing were almost linearly related to the number of genetic markers, and the
method can easily handle scaffolds with up to 10,000 genetic markers (Extended Data Fig. 7). The high scalability makes it possible to apply the method to other more extensive genotyping techniques, such as whole-genome or transcriptome re-sequencing approaches, which could have millions of genetic markers. The generalization of the proposed method to high-ploidy genomes is straightforward. In this study, we focused on diploids and tetraploids. However, this method can construct genetic linkage maps for higher-ploidy genomes. Our biggest challenge when dealing with higher-ploidy species is computational, especially in the haplotype phasing step. For hexaploids, the number of hidden states for HMM increases to 14,400. This is a large number of states but remains computationally tractable. However, for even higher levels of ploidy, computation becomes difficult.

The fundamental idea for haplotype phasing is to model the process of gamete formation of the parents with a Markov chain that starts with the first and ends with the last genetic marker along a chromosome. Similar models were used in PolyHap and TetraOrigin. PolyHap assumed a given number of ancestral haplotypes shared by the target population. The recombinations were allowed between any ancestral haplotypes, leading to an intractably large state space for TetraOrigin and our method are nearly identical except that TetraOrigin also considered double reductions. Double reduction was not considered in our method mainly because the reduction was not considered in our method mainly because the accurate identification of them requires high-quality genotyping data. PolyHap used SNP array data to detect double reductions. However, with low-coverage GBS or RAD-seq data, it is difficult to obtain such high-quality genotypes without filtering out considerable markers, which is undesirable if we want to cover the whole genome. Even though double reduction could introduce some bias in RF estimation, it can be safely ignored if quadrivalent pairing is rare. For organizing the assembly scaffolds, the bias is negligible.

We have demonstrated the ability of our method in the construction of whole-chromosome pseudomolecules for genome assembly. For the real datasets, the genome coverage of the pseudomolecules ranged from 45% for the diploid I. trifida to 92% for the allotetraploid Z. japonica. The huge difference arose because of the quality of the input genome assemblies, especially the contiguity. The scaffold N50 statistics of the input genome assemblies were approximately 43 kb and 2.6 Mb, respectively, for I. trifida and Z. japonica. Even though generating high-quality plant genome assemblies remains challenging, the recent advances in the sequencing technologies and related computational tools facilitate the construction of highly contiguous genome assemblies even for very complex plant genomes. In the context of the rapid development of genome assembly techniques, it is hoped that our method could serve as an option among the others to finish a genome assembly.

**Online content**

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-020-00717-7.

Received: 10 March 2017; Accepted: 15 September 2020; Published online: 30 October 2020

**References**

1. Kyriakidou, M., Tai, H.-H., Anglin, N. L., Ellis, D. & Strömvik, M. V. Current strategies of polyploid plant genome sequence assembly. *Front. Plant Sci.* **9**, 1660 (2018).

2. Bancroft, I. et al. Dissecting the genome of the polyploid crop oilseed rape by transcriptome sequencing. *Nat. Biotechnol.* **29**, 762–766 (2011).

3. Wu, S. et al. Genome sequences of two diploid wild relatives of cultivated sweetpotato reveal targets for genetic improvement. *Nat. Commun.* **9**, 4580 (2018).

4. Fierst, J. L. Using linkage maps to correct and scaffold de novo genome assemblies: methods, challenges, and computational tools. *Front. Genet.* **6**, 220 (2015).

5. Alshuler, D. et al. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**, 513–516 (2000).

6. Baird, N. A. et al. Rapid SNP discovery and genetic mapping using sequenced rad markers. *PLoS ONE* **3**, e3376 (2008).

7. Elshire, R. J. et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* **6**, e19379 (2011).

8. Lander, E. S. & Green, P. Construction of multicopy genetic linkage maps in humans. *Proc. Natl Acad. Sci. USA* **84**, 2363–2367 (1987).

9. Broman, K. W., Wu, H., Sen, S. & Churchill, G. A. R/qtl: QTL mapping in experimental crosses. *Bioinformatics* **19**, 889–890 (2003).

10. Margarido, G., Souza, A. & Garcia, A. OneMap: software for genetic mapping of outcrossing species. *Hereditas* **144**, 78–79 (2007).

11. Van Ooijen, J. Multipoint maximum likelihood mapping in a full-sib family of an outbreeding species. *Genet. Res.* **93**, 343–349 (2011).

12. Rastas, P., Calboli, F. C., Guo, B., Shikano, T. & Merilä, J. Construction of ultradense linkage maps with Lep-MAP2: stickleback F2 recombinant crosses as an example. *Genome Biol. Evol.* **8**, 78–93 (2016).

13. Hackett, C. & Luo, Z. TetraploidMap: construction of a linkage map in autotetraploid species. *J. Hered.* **94**, 358–359 (2003).

14. Hackett, C. A., Boskamp, B., Vogogias, A., Preedy, K. F. & Milne, L. TetraploidSNMap: software for linkage analysis and QTL mapping in autotetraploid populations using SNP dosage data. *J. Hered.* **108**, 438–442 (2017).

15. Bourke, P. M. et al. polymapR—linkage analysis and genetic map construction from F2 populations of outcrossing polyploids. *Bioinformatics* **34**, 3496–3502 (2018).

16. Hirakawa, H. et al. Survey of genome sequences in a wild sweet potato, *Ipomoea trifida* (H. B. K.) G. Don. *DNA Res.* **22**, 171–179 (2015).

17. Consortium, P. G. S. et al. Genome sequence and analysis of the tuber crop potato. *Nature* **475**, 189–195 (2011).

18. Hoshino, A. et al. Genome sequence and analysis of the Japanese morning glory *Ipomoea nil*. *Nat. Commun.* **7**, 13295 (2016).

19. Wang, F. et al. Sequence-tagged high-density genetic maps of *Zayisia japonica* provide insights into genome evolution in Chloridoideae. *Plant J.* **82**, 744–755 (2017).

20. Ouyang, S. et al. The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res.* **35**, D883–D887 (2006).

21. Tanaka, H. et al. Sequencing and comparative analyses of the genomes of zosysiaagrasses. *DNA Res.* **23**, 171–180 (2016).

22. Strehl, A. & Ghosh, J. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.* **3**, 613–637 (2003).

23. Wu, Y., Bhat, P. R., Close, T. J. & Lonardi, S. Efficient and accurate correcting draft genomes using recombinant populations. *Genome Biol. Evol.* **8**, 358–373 (2016).

24. Su, S.-Y., White, J., Balding, D. J. & Coin, L. J. Inference of haplotypic phase and missing genotypes in polyploid organisms and variable copy number genomic regions. *BMC Bioinformatics* **9**, 153 (2008).

25. Zheng, C. et al. Probabilistic multilocus haplotype reconstruction in experimental crosses. *PLoS ONE* **3**, e1000212 (2008).

26. Mascher, M. et al. Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ). *Plant J.* **76**, 718–727 (2013).

27. Hahn, M. W., Zhang, S. V. & Moyle, L. C. Sequencing, assembling, and correcting draft genomes using recombinant populations. *G3 (Bethesda)* **4**, 669–679 (2014).

28. Su, S.-Y., White, J., Balding, D. J. & Coin, L. J. Inference of haplotypic phase and missing genotypes in polyploid organisms and variable copy number genomic regions. *BMC Bioinformatics* **9**, 513 (2008).

29. Zheng, C. et al. Probabilistic multilocus haplotype reconstruction in outcrossing tetraploids. *Genetics* **203**, 119–131 (2016).

30. Jiao, W.-B. & Schneeberger, K. The impact of third generation genomic technologies on plant genome assembly. *Curr. Opin. Plant Biol.* **36**, 64–70 (2017).

31. Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autoploid genomes based on Hi-C data. *Nat. Plants* **5**, 833–845 (2019).

32. Kyriakidou, M., Anglin, N. L., Ellis, D., Tai, H. H. & Strömvik, M. V. Genome assembly of six polyploid potato genomes. *Sci. Data* **7**, 88 (2020).

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2020
Methods

Data simulation. We simulated a reference genome with the sequence data of *I. trifida*. The sequence data were first assembled to generate contigs (Supplementary Table 4). Contigs were then used to generate complex repeats considering the copy number of the contigs inferred from the sequence depth (Supplementary Table 5 and Supplementary Note 1). Finally, we randomly concatenated these repetitive sequences and generated a reference genome comprising 15 chromosomes with a total size of ~482 Mb (Supplementary Table 6).

We simulated the genotypes of a diploid and a tetraploid outbred mapping population both consisting of two parents and 190 F1 progeny using PedSim (version 2.0)10. The genotype data for comparison studies were randomly sampled from the mapping populations (Supplementary Note 1). To simulate GBS data, we mapped the genotypes back to the reference genome to generate individual genomes and simulated the GBS protocol including sequence errors to produce GBS sequences (Supplementary Note 1). To simulate the reference genome assemblies for diploid genomes, we simulated Illumina reads with ART (version 2.5.8)10, followed by DISCOVAR de novo (version 524883)11 assembly, whereas for tetraploid genomes, we simulated nanopore reads with DeepSimulator (version 1.5)14 and used Flye (version 2.5.3)14 for assembly (Supplementary Note 1).

Construction of *Z. japonica* genome assembly. The publicly available whole-genome PacBio sequence data for the *Z. japonica* cultivar Yai were downloaded from the National Center for Biotechnology Information (NCBI) repository under the accession number SRP110561. The sequence data consist of ~4.2 million reads, with a total size of ~38.6 Gb and an N50 read length of ~12.2 kb. Flye (version 2.5.3)14 was used to assemble the sequence data with default parameters.

Variant calling for mapping populations. The variant calling for GBS and RAD-seq data was performed with TASSEL-GBS (version 5.2.43)10 and Stacks (version 2.4.1)10, respectively, using default parameters. Then, we extracted the allele depth of each SNP and performed a new round of genotyping with the R package updog (version 1.1.3)10, taking into account that the mapping population is a recombinant of parental chromosomes inherited from a gamete. Assume that we have count, respectively, for an individual at the $j$th marker, which could be either a parent or F1 progeny. Here we drop the subscript of individuals for convenience. For example, with a diploid genome (that is, $h = 2$), we have $S = \{1,2,3\}$, $\{1,3\}$, $\{1,4\}$, $\{2,3\}$, $\{2,4\}$, where the first two are for parents and the last four are for F1 progeny. Denote by the probability of parental chromosome $i$ being associated with the reference allele at the $j$th marker, and therefore $1 - \theta_i$ the probability of the alternative allele, where $\theta_i \in \{0,1\}$, $i = 1,2...h$ and $j = 1,m$.

Modeling the F1 progeny. An F1 progeny can be decomposed to $F = \sum P_i \hat{Q}_i$, where $P_i = \{f_{1,i},...,f_{h,i}\}$ and $Q_i = \{f_{i+1,j},...,f_{i,m}\}$, representing the genotypes inherited from $P$ and $Q$, respectively. The formation of gametes for two parents are the same. Here we only consider $P$. We further write $f_i \in \{f_{1,i},...,f_{h,i}\}$, where $f_i \in \{1,2\}$, $i = 1,...,2h$ and $j = 1,m$. Now, consider $P_j$ and $P_{j+1}$ (that is, the $j$th and $(j+1)$th markers on the chromosome). We introduce a parameter set $\alpha = \{\alpha_{1,2},...\}$ to model the transitions between the homologous chromosomes during the meiosis process. Precisely, $\alpha_i$ is the probability of a jump occurring between different homologous chromosomes between the $j$th and $(j+1)$th markers.

This can be written as:

$$Pr(f_{ij} = l | f_{i,j+1} = k) = \left\{ \begin{array}{ll} 1 - \alpha_i & \text{if } k = l \\ \frac{\alpha_i}{2} & \text{otherwise} \end{array} \right.$$ (1)

The parameter $\alpha_i$ is positively related to the physical distance between the $j$th and $(j+1)$th markers. When the two markers are tightly linked, $\alpha_i$ should be extremely small to prevent haplotype states from jumping frequently. Actually, $\alpha_i$ could be used to approximate the RF between the two markers. The transition probability between $P_j$ and $P_{j+1}$ is then written as:

$$Pr(P_j = \{f_1,...,f_{2h}\} | P_{j+1} = \{f_1',...,f_{2h}'\}) = \sum_{P_j \in \Pi(P_j)} \prod_{i = 1,\ldots,2h} Pr(f_{ij} = l | f_{i,j+1} = k_{ij})$$ (2)

For unphased $P_j$ and $P_{j+1}$, the transition probability is written as:

$$Pr(F_{j,i} = \{f_1,...,f_{2h}\} | P_j = \{f_1',...,f_{2h}'\}) = \prod_{P_j \in \Pi(P_j)} \prod_{i = 1,\ldots,2h} Pr(f_{ij} = l | f_{i,j+1} = k_{ij})$$ (3)

Since the inheritance of chromosomes from two parents is independent, the transition probability between $F_i$ and $P_j$ in the unphased model is then written as:

$$Pr(F_{j,i} \mid F_{j+1,i}) = Pr(F_{j,i} | P_j) Pr(Q_{j+1} | Q_j)$$ (4)

Likelihood of allele depth data. For an individual, the probability of observing the reference allele at a marker with parental chromosome subset $s \subseteq S$ is calculated as:

$$\theta_s = \frac{1}{h} \sum_{i \in s} \theta_i$$ (5)

We drop the subscript for $s$, and $\theta_s$ is calculated for each $s$. The count of the reference allele at the $j$th marker, $y_j$, follows a binomial model of probability $\theta_j$ with $x_j$ trials; that is:

$$y_j \sim \text{Binomial}(x_j, \theta_j)$$ (6)

Likelihood of genotype data. With genotype data, $x_j$ and $y_j$ represent the total allele count and reference allele dosage, respectively. Apparently, we have $x_j = h$, irrespective of $j$, and $y_j = 0...h$. For a chromosome set $s \subseteq S$, $y_j$ follows a Poisson binomial distribution (that is, the sum of $h$ independent Bernoulli trials with success probabilities associated with $s$). The probability of $y_j = g$, where $g = 0...h$, is calculated as:

$$Pr(y_j = g) = \sum_{A \subseteq s} \prod_{a \in A} \theta_a \prod_{b \in A^c} (1 - \theta_b)$$ (7)
where \( g \) is the collection of all subsets of \( g \) elements that can be selected from \( s \). \( A' \) is the complement of \( A \) with respect to \( g \) (that is, \( A' = s \setminus A \)). For example, if \( g = \{1,2,3,6\} \) and \( g = \{3,5,6\} \), then \( A' = \{1,2,4\} \).

More computational details. The model described above can be summarized as a HMM. Each marker is a time step; the hidden states are the underlying parental chromosomes \( S \) and the observations are the allele depth or the genotypic data collected on the mapping population. The transition probabilities are calculated with equation (4) and the emission probabilities are calculated with equation (6) or (7). The parameters for the model include \( \theta = \{ \theta_1, \ldots, \theta_m \} \) and \( \{ \theta_i \} \), where \( i = 1, \ldots, 2h \). The Baum–Welch algorithm was used to estimate these parameters. We used Dirichlet priors on all parameters. Let \( \Theta_i \sim Beta(\mu_i, (1-\beta_i) \mu_i) \), where \( \beta_i \) is the physical distance between the \( j \)th and \( (j+1) \)th markers along the chromosome, and \( \mu_i \sim 10^{-\gamma} \) for base pair, reflecting the background recombination rate. Let \( \Theta_i \sim Beta(\mu_i, (1-\beta_i) \mu_i) \), where \( \beta_i \) is the allele frequency estimated from the genotype data for the \( j \)th marker. We used \( \mu_i = 10^{-1} \) and \( \mu_i = 10 \) for initialization of the expectation–maximization algorithm and \( \mu_i = 0.1 \) for the maximization step.

RF estimation. The RF between two markers was estimated by the proportion of the number of recombinants to the total number of haplotypes in the F1 progeny. Assume a genome of ploidy \( f \) and a mapping population of \( f \) F1 progeny, and \( r \) recombinants between the two markers in the \( F1 \) haplotypes. The RF was then calculated as:

\[
\tau = \frac{r}{nh} \tag{8}
\]

Two kinds of RFs were estimated in the proposed method: (1) to detect assembly errors, we calculated RFs between adjacent markers along the scaffold; and (2) to estimate genetic distance between two scaffolds, we calculated RFs between the outermost markers. The RF calculations within a scaffold were straightforward. We counted the number of jumps from one parental haplotype to another in \( F1 \) haplotypes. However, the calculation of RF between two scaffolds needed to consider all four possible directions as the orientations of the scaffolds on the chromosome were unknown. Besides, as we ran the haplotype phasing algorithm for each scaffold independently, we considered all possible correspondences between parental haplotypes. See Supplementary Note 3 for more computational details and an example.

Assembly error detection. To detect misassemblies, RFs between the adjacent markers along a scaffold were calculated. If the RF between two adjacent markers was larger than the predefined threshold 0.1, we considered it a misassembly at that position. Apparently, the physical distance between two markers should be taken into consideration. However, it is difficult to derive a universal relationship between the physical distances and RFs, especially when the physical distance becomes large. Therefore, we did not consider positions where the physical distance between the adjacent markers was larger than 1 Mb. The incorrect scaffolds detected by the algorithm were split at the misassembled positions to generate new scaffolds. We ran haplotype phasing analysis on the new scaffolds and recalculated the pairwise RFs. For the genome assemblies generated from the simulated data, QUAST (version 5.0.2) was employed to detect the genuine misassemblies by comparing the scaffolds with the reference chromosomes with default parameters. The genuine misassemblies identified by QUAST were used to validate the misassemblies detected by the proposed method.

Superscaffold construction and multipoint analysis. We introduced multipoint haplotype phasing analysis to improve the accuracy of RF estimations. The multipoint analysis ran the haplotype phasing algorithm on superscaffolds, which were constructed by nearest neighbor joining. Precisely, we chose the scaffold pair that represented the smallest RF. The scaffolds were combined to form a superscaffold if the RF was smaller than the predefined threshold 0.27 (approximately 30 cM with the Kosambi mapping function). The gap between the two scaffolds was measured by the RF. The RFs between the superscaffold and the other scaffolds were then updated. This process was repeated until there was no such closely linked scaffold pair. Multipoint analysis is crucial for short scaffolds as the number of markers that mapped to them could be small, which makes haplotype phasing solely based on them inaccurate.

Linkage group construction. We constructed a graph for linkage group construction with scaffolds as vertices and genetic linkages as edges. To simplify this, we excluded edges with RFs greater than the predefined threshold 0.38 (approximately 50 cM with the Kosambi mapping function). The retained edges were weighted based on the RFs. For an edge with RF \( \tau \), the weight was calculated as \( \tau(1-\tau) \cdot k(p) \), where \( k(p) \) is a function that converts the RFs to genetic distances with the Kosambi mapping function. For graph clustering, an information-theoretic approach, as implemented in the R package igraph, was employed, which detects the modularity structure of a graph by minimizing the expected description length of the trajectory of a random walk.

Scaffold ordering within linkage groups and pseudomolecule construction. The scaffold ordering within each linkage group was performed with the MDS algorithm, as implemented in the R package MDSMap (version 1.1). Once the order of the scaffolds was determined, we further performed a TSP to find the optimum directions of the scaffolds to minimize the sum of adjacent RFs. The start and end of a scaffold were treated as two independent points in the TSP. The TSP was carefully designed so that in the optimum solution: (1) the Hamiltonian circuit can be equivalently converted to a Hamiltonian path, as is required for scaffold ordering; (2) the order of scaffolds determined by the MDS algorithm remains unchanged; and (3) the start and end of a scaffold are adjacent (Supplementary Note 3). The exact TSP solver CONCORDE (release 03.12.19) was employed in this research to find exact solutions of TSPs to guarantee feasible scaffold orderings. The genetic linkage maps of scaffolds were converted to pseudomolecules. The adjacent scaffolds were joined with 10-kb gaps.

True linkage groups, scaffold orders and distances, and collinear blocks. We mapped scaffolds to the reference chromosomes using Minimap2 (version 2.1.7). A scaffold was considered as being mapped to a reference chromosome if at least 50% of the base pairs were collinear with that, and only that, chromosome. The true linkage groups of scaffolds were determined by the reference chromosomes they mapped to. The order of the scaffolds within a linkage group and the pairwise distances were determined by the positions on the chromosome. We used MUMmer (version 3.2.3) for mapping pseudomolecules to reference chromosomes to detect collinear blocks.

Calculation of the correct phasing rate. The correct phasing rate is calculated as the proportion of the correctly phased loci with respect to the true haplotypes. Denote by \( H = (H_1, \ldots, H_h) \) and \( H^* = (H_1^*, \ldots, H_h^*) \) the haplotypes constructed by PolyGembler and the true haplotypes, respectively. Denote by \( \hat{H} \) a one-to-one mapping between the haplotypes in \( H \) and \( H^* \). The number of correctly phased loci is then calculated as \( M(\hat{H}) = \sum_{i=1}^{h} 1 \) \( \{ \hat{H}(j) = H_i^* \} \), where \( m \) is the number of genetic markers and \( \hat{H} \) is the indicator function. The correct phasing rate \( R \) is defined as:

\[
R = \frac{\max \{ M(\hat{H}) \}}{mh} \tag{9}
\]

For a dataset of multiple individuals and multiple scaffolds, we calculated \( R \) as the total number of matched loci divided by the total number of loci. To evaluate the power of inferring missing data and the robustness in handling genotyping errors for PolyGembler, the haplotypes before introducing missing data and genotyping errors were used as true haplotypes.

Statistical testing. For SNP filtering, we performed exact multinomial tests with the xmulti function from the R package XNomial (version 1.0.4) to check whether the segregation of a SNP followed the law of segregation. The log-likelihood ratio was used as the test statistic to calculate \( \hat{P} \). The original sample sizes for the MS9 x M19, B2721, 12601ab1xStirling, Carrizzo x EL Toro and simulated mapping populations were 210, 156, 190, 60 and 190, respectively. We required a minimum sample size of 50 for testing after removing samples with no genotypes or unexpected genotypes with respect to the parental genotypes. We rejected the null hypothesis if the \( \hat{P} \) value was <0.001. To compare PolyGembler with OneMap and Lep-MAP2, we performed two-tailed paired-sample \( t \)-tests to compare the accuracy for grouping and ordering and the central processing unit (CPU) time (Supplementary Table 3). The sample size was 30.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The dataset for the simulation studies, including comparisons with other methods and studies of MS9 x M19 I. trifida and the B2721 potato, are available from http://data.genomicsresearch.org/Projects/polyGembler. Data for the 12601ab1xStirling potato mapping population were provided by C. Hackett. Data for the Z. japonica mapping population Carrizzo x EL Toro are available from the NCBI repository under the accession number SRP055007. The whole-genome PacBio sequence data for the Z. japonica cultivar Yai are available from the NCBI repository under the accession number SRP110561. Data related to the PGSC version 4.03 pseudomolecules are available from http://solanaeae.plantbiology.msu.edu. The I. trifida de novo genome assembly ITR_r1.0 is available from http://sweetpotato-garden.kazusa.or.jp. The I. trifida de novo genome assembly NCNSP0306 is available from http://sweetpotato-plantbiology.msu.edu. Release 7 of the O. sativa reference genome is available from http://phytomize.jgi.doe.gov. The genome assembly of the Z. japonica accession Nagirizaki is available from http://zoyssa.kazusa.or.jp. Source data are provided with this paper.

Code availability

The software PolyGembler, presented in this article, and its documentation are publicly available at GitHub (https://github.com/c-zhou/polyGembler).
References

31. Voorrips, R. E. & Maliepaard, C. A. The simulation of meiosis in diploid and tetraploid organisms using various genetic models. BMC Bioinformatics 13, 248 (2012).

32. Huang, W., Li, L., Myers, J. R. & Marth, G. T. Art: a next-generation sequencing read simulator. Bioinformatics 28, 593–594 (2012).

33. Love, R. R., Weisenfeld, N. I., Jaffe, D. B., Besansky, N. J. & Neafsey, D. E. Evaluation of DISCOVAR de novo using a mosquito sample for cost-effective short-read genome assembly. BMC Genomics 17, 187 (2016).

34. Li, Y. et al. DeepSimulator: a deep simulator for nanopore sequencing. Bioinformatics 34, 2899–2908 (2018).

35. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. Nat. Biotechnol. 37, 540–546 (2019).

36. Glaubitz, J. C. et al. TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS ONE 9, e90346 (2014).

37. Rochette, N. C., Rivera-Colón, A. G. & Catchen, J. M. Stacks 2: analytical methods for paired-end sequencing improve RADseq-based population genomics. Mol. Ecol. 28, 4737–4754 (2019).

38. Gerard, D., Ferrão, L. F. V., Garcia, A. A. F. & Stephens, M. Genotyping polyploids from messy sequencing data. Genetics 210, 789–807 (2018).

39. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29, 1072–1075 (2013).

40. Csardi, G. & Nepusz, T. The igraph software package for complex network research. Int. J. Complex Syst. 1695, 1–9 (2006).

41. Rosvall, M. & Bergstrom, C. Maps of information flow reveal community structure in complex networks. Proc. Natl Acad. Sci. USA 105, 1118–1123 (2008).

42. Preedy, K. & Hackett, C. A rapid marker ordering approach for high-density genetic linkage maps in experimental autotetraploid populations using multidimensional scaling. Theor. Appl. Genet. 129, 2117–2132 (2016).

43. Li, H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094–3100 (2018).

44. Kurtz, S. et al. Versatile and open software for comparing large genomes. Genome Biol. 5, R12 (2004).

45. Xie, M., Wu, Q., Wang, J. & Jiang, T. H-PoP and H-PoPG: heuristic partitioning algorithms for single individual haplotyping of polyploids. Bioinformatics 32, 3735–3744 (2016).

Acknowledgements

We thank F. Diaz for developing the M9 × M19 I. trifida mapping population and M. David for extracting and quantifying DNA from the M9 × M19 cross. The 12601ah1 × Stirling Infinium 8303 potato array data were provided by C. A. Hackett. This research was supported by grants from the Bill & Melinda Gates Foundation (OPP1052983) and Australian Research Council (DP170102626 awarded to L.J.M.C.). The work at the International Potato Center (CIP) was carried out as part of the Consultative Group for International Agricultural Research (CGIAR) Research Program on Roots, Tubers and Bananas, which is supported by CGIAR Fund Donors (http://www.cigiar.org/about-us/our-funders/). This research was also supported by use of the NeCTAR Research Cloud, by QCIF and by the University of Queensland's Research Computing Centre. The NeCTAR Research Cloud is a collaborative Australian research platform supported by the National Collaborative Research Infrastructure Strategy.

Author contributions

C.Z. and L.J.M.C. designed the study and wrote the software. A.K., D.C.G. and W.G. developed and provided the I. trifida mapping population materials. B.O., D.C.G., S.W. and W.G. generated data for the M9 × M19 I. trifida mapping population. C.Z. performed the analysis. C.Z. and L.J.M.C. wrote the manuscript. L.J.M.C., G.C.Y., A.K., M.D.C., A.W.G., Z.-B.Z. and Z.F. supervised the project. All authors contributed to editing the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41588-020-00717-7.

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-020-00717-7.

Correspondence and requests for materials should be addressed to L.J.M.C.

Reprints and permissions information is available at www.nature.com/reprints.
Extended Data Fig. 1 | Pseudomolecule construction for 20× tetraploid simulated GBS data. A total of 42,715 SNPs located on 678 scaffolds were used for linkage analysis. These scaffolds of ~482 Mb covered approximately 99.6% of the genome. 

**a.** Dot plot for the RF estimations for scaffold pairs mapped to the same reference chromosome. The x- and y-axis represent the physical distances and the estimated RFs, respectively. 

**b.** Histogram of the RF estimations for scaffold pairs mapped to different reference chromosomes. 

**c.** Collinear plots of pseudomolecules mapped to reference chromosomes. The x- and y-axis represent physical positions (Mb) on the reference chromosomes and pseudomolecules, respectively. Each line represents a collinear block between the reference chromosome and the pseudomolecule. The diagonal line in each plot indicates a high correlation between the reference chromosome and the pseudomolecule constructed from scaffolds.
Extended Data Fig. 2 | Collinear plots between the *Ipomoea nil* reference chromosomes and pseudomolecules constructed from the *Ipomoea trifida* genotype data. The x- and y-axis represents the physical positions (Mb) on the reference chromosomes and pseudomolecules, respectively. Each line represents a collinear block between the *Ipomoea nil* reference chromosome and the pseudomolecules.
Extended Data Fig. 3 | Genetic linkage map construction from the Infinium 8303 SNP array data of the Stirling×12601ab1 mapping population.

**a**, Dot plot for RF estimations between scaffold pairs mapped to the same PGSC v4.03 chromosomes. The x- and y-axis represents the physical distances and the estimated RFs, respectively. **b**, Histogram of the RF estimations for scaffold pairs mapped to different PGSC v4.03 pseudomolecules. **c**, Comparison between the genetic linkage map constructed by the proposed method and the PGSC v4.03 pseudomolecules. Twelve genetic linkage groups corresponding to 12 pseudomolecules were constructed. In each plot, the x-axis represents the positions (Mb) on the PGSC v4.03 pseudomolecules, and the y-axis represents the positions (cM) on the genetic linkage map.
Extended Data Fig. 4 | Genetic linkage map constructed from the Infinium 8303 SNP array data of the B2721 mapping population with TetraploidSNPMap. Each dot represents a SNP. The x-axis represents the positions (Mb) on the PGSC v4.03 pseudomolecules, and the y-axis represents the positions (cM) on the genetic linkage map. The genetic linkage map comprises a total of 4,745 SNPs including 56 SNPs located on the unplaced PGSC v4.03 scaffolds (red) and 76 SNPs placed in incorrect PGSC v4.03 pseudomolecules (blue). Since the physical positions of the red and blue dots cannot be determined, they were set to zero in the plots.
Extended Data Fig. 5 | Genetic linkage map constructed from the Infinium 8303 SNP array data of the Stirling×12601ab1 mapping population with TetraploidSNPMap. Each dot represents a SNP. The x-axis represents the positions (Mb) on the PGSC v4.03 pseudomolecules, and the y-axis represents the positions (cM) on the genetic linkage map. The genetic linkage map comprises a total of 3,593 SNPs including 54 SNPs located on the unplaced PGSC v4.03 scaffolds (red) and 35 SNPs placed in incorrect PGSC v4.03 pseudomolecules (blue). Since the physical positions of the red and blue dots cannot be determined, they were set to zero in the plots.
Extended Data Fig. 6 | Collinear plots between the pseudomolecules of Zoysia japonica accession Yaji and Nagirizaki. The x- and y-axis represent the positions (Mb) on the pseudomolecules. Each line represents a collinear block between the pseudomolecules.
Extended Data Fig. 7 | Relationship between the number of genetic markers and computational resources required for the haplotype phasing algorithm. The x- and y-axis (in logarithm scale) represents the number of genetic markers and the consumption of resources, respectively. **a**, CPU time and **b**, Memory. Each point in the plot was averaged over 30 independent experiments (Intel® Xeon® Processor E5-2667 v3 CPU, 3.20GHz). The error bar for one standard deviation was included at each point.
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
- Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

1. ART (Version 2.5.8) was used to simulate Illumina short reads.
2. DeepSimulator (Version 1.5) was used to simulate Nanopore long reads.
3. PedgreeSim (Version 2.0) was used to simulate F1 mapping populations.
4. The custom software PolyGembler was used to simulate GBS data. PolyGembler together with its documentation have been made publicly available at GitHub https://github.com/c-zhou/polyGembler.

Data analysis

1. SOAPdenovo2 (Version 2.0.4) was used for short reads assembly.
2. DISCOVAR de novo (Version 52488) was used for short reads assembly.
3. Flye (Version 2.5) was used for long reads assembly.
4. SGA (Version 0.10.15) was used for scaffolding in genome assembly.
5. QUAST (Version 5.0.2) was used for assembly error detection.
6. TASSEL-GBS (Version 5.2.43) was used for variant calling from GBS data.
7. Stacks (Version 2.41) was used for variant calling from RAD-seq data.
8. R (Version 3.5.2) was used for running custom code and necessary R packages.
9. R package RPoly (Version 3.0.0) was used for genotyping from the Infinium i8003 potato array data.
10. R package updag (Version 1.1.3) was used for genotyping from allele depth data for F1 outbred mapping populations.
11. R package ggplot2 (Version 2.2.1) was used for generating the dot plot and histogram.
12. R package MDSMap (Version 1.1) was used for running multidimensional scaling algorithm.
13. R package Onemap (Version 2.0-4) was used for genetic mapping.
14. R package XNomial (Version 1.0.4) was used for performing multinomial goodness of fit tests.
15. TetraploidSNPMap (no version control) was used for genetic mapping.
16. Lep-MAP2 (no version control) was used for genetic mapping.
17. MUMmer (Version 3.23) was used for genome comparison.
18. Minimap2 (Version 2.17) was used for alignment between pseudomolecules and reference chromosomes.
20. The custom software PolyGembler was used to perform genetic linkage mapping and pseudomolecule construction. PolyGembler together with its documentation have been made publicly available at GitHub https://github.com/c-zhou/polyGembler.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

**Data**

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

1. Data for the Zoysia japonica mapping population 'Carrizo'x'El Toro' is publicly available from the NCBI repository under the accession number SRP055007.
2. The whole genome PacBio sequence data for the Zoysia japonica cultivar Yaji is publicly available from the NCBI repository under the accession number SRP110561.
3. Data related to the PGSC v4.03 pseudomolecules are publicly available at the website http://solanaceae.plantbiology.msu.edu.
4. The Ipomoea trifida de novo genome assembly ITR_r1.0 is publicly available at the website http://sweetpotato-garden.kazusa.or.jp.
5. The Ipomoea trifida de novo genome assembly NCN00306 is publicly available at the website http://sweetpotato.plantbiology.msu.edu.
6. The Release 7 of the Oryza sativa reference genome is publicly available at the website http://phytozome.jgi.doe.gov.
7. The Zoysia japonica accession Nagirizaki genome assembly is publicly available at the website http://zoysia.kazusa.or.jp.
8. Data for the 'M9'x'M19' Ipomoea trifida and B2721 potato datasets are available at http://data.genomicsresearch.org/Projects/polyGembler.
9. Data for the 12601ab1xStirling potato mapping population was provided by Dr. Christine Hackett upon request.
10. Data related to the simulation studies including the comparisons to other methods are available at the website http://data.genomicsresearch.org/Projects/polyGembler.

**Field-specific reporting**

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences
- Behavioural & social sciences
- Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

**Life sciences study design**

All studies must disclose on these points even when the disclosure is negative.

| Sample size | 'M9'x'M19' Ipomoea trifida mapping population: two parents + 210 F1 progeny |
|-------------|--------------------------------------------------------------------------------|
|             | B2721 potato mapping population: two parents + 156 F1 progeny                  |
|             | 12601ab1xStirling potato mapping population: two parents + 190 F1 progeny     |
|             | 'Carrizo'x'El Toro' Zoysia japonica mapping population: two parents + 60 F1 progeny |
|             | Simulated mapping populations: two parents + 190 F1 progeny                    |

**Data exclusions**

No samples were excluded in the analysis. For SNP filtering, we performed multinomial goodness of fit tests to check if the segregation of a SNP follows Mendel's law of segregation. Depending on whether the parental genotypes were known or not, one or all possible segregation ratios were tested. A SNP was filtered out if the p-value<0.001 for all possible segregation ratios. PolyGembler requires a scaffold/contig to contain at least three SNPs, and thus scaffolds/contigs covered by less than three SNPs were excluded from genetic linkage analysis.

**Replication**

N/A

**Randomization**

N/A

**Blinding**

N/A

**Reporting for specific materials, systems and methods**

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.
## Materials & experimental systems

| Involved in the study |
|-----------------------|
| Antibodies           |
| Eukaryotic cell lines |
| Palaeontology and archaeology |
| Animals and other organisms |
| Human research participants |
| Clinical data |
| Dual use research of concern |

## Methods

| Involved in the study |
|-----------------------|
| ChIP-seq |
| Flow cytometry |
| MRI-based neuroimaging |

## Dual use research of concern

Policy information about [dual use research of concern](#)

### Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

| No | Yes |
|----|-----|
| Public health |   |
| National security |   |
| Crops and/or livestock |   |
| Ecosystems |   |
| Any other significant area |   |

### Experiments of concern

Does the work involve any of these experiments of concern:

| No | Yes |
|----|-----|
| Demonstrate how to render a vaccine ineffective |   |
| Confer resistance to therapeutically useful antibiotics or antiviral agents |   |
| Enhance the virulence of a pathogen or render a nonpathogen virulent |   |
| Increase transmissibility of a pathogen |   |
| Alter the host range of a pathogen |   |
| Enable evasion of diagnostic/detection modalities |   |
| Enable the weaponization of a biological agent or toxin |   |
| Any other potentially harmful combination of experiments and agents |   |