Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Enhanced Pore-C with C-Phasing Enables Chromosomal-Scale, Haplotype-Resolved Assembly of Ultra-Complex Genomes

Xingtan Zhang

zhangxingtan@caas.cn

Agricultural Genomics Institute at Shenzhen https://orcid.org/0000-0002-5207-0882

Yibin Wang

Fujian Agriculture and Forestry University

Ping Zhao

State Key Laboratory for Tropical Crop Breeding, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences

Xiaofei Zeng

Southern University of Science and Technology https://orcid.org/0000-0001-8127-3825

Jiaxin Yu

Chinese Academy of Agricultural Sciences

Aoqian Dong

Center for Genomics and Biotechnology, Fujian Agriculture and Forestry University, Fuzhou, Fujian, 350002, China

Yi Liu

State Key Laboratory for Tropical Crop Breeding, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences

Mengwei Jiang

State Key Laboratory for Tropical Crop Breeding, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences

Fang Wang

State Key Laboratory for Tropical Crop Breeding, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences

Xiao Chen

State Key Laboratory for Tropical Crop Breeding, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences

Shengcheng Zhang

Agricultural Genomics Institute at Shenzhen https://orcid.org/0000-0002-3949-4269

Shuai Chen

FAFU and UIUC-SIB Joint Center for Genomics and Biotechnology, National Sugarcane Engineering Technology Research Center, Fujian Provincial Key Laboratory of Haixia Applied Plant Systems Biology, F
https://orcid.org/0000-0002-6861-2682

Yuqing Gong
Center for Genomics and Biotechnology, Fujian Agriculture and Forestry University, Fuzhou, Fujian, 350002, China

Yixing Zhang
Center for Genomics and Biotechnology, Fujian Agriculture and Forestry University, Fuzhou, Fujian, 350002, China

Ruicai Long
Institute of Animal Science, Chinese Academy of Agricultural Sciences, Beijing, China

Maojun Wang
Huazhong Agricultural University https://orcid.org/0000-0002-4791-3742

Haibao Tang
Fujian Agriculture and Forestry University https://orcid.org/0000-0002-3460-8570

Article

Keywords: polyploid, ultra-complex genome, genome assembly, haplotype-resolved, scaffolding

Posted Date: November 6th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7343323/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Enhanced Pore-C with C-Phasing Enables Chromosomal-Scale, Haplotype-Resolved Assembly of Ultra-Complex Genomes

Yibin Wang [1,2,#], Ping Zhao[2,#], Xiaofei Zeng[2], Jiaxin Yu[2], Aoqian Dong[2,3], Yi Liu[2], Mengwei Jiang[2], Fang Wang[2], Xiao Chen[2], Shengcheng Zhang[1,2], Shuai Chen[2], Yuqing Gong[2,3], Yixing Zhang[3], Ruicai Long[4], Maojun Wang[1], Haibao Tang[3,*] and Xingtan Zhang [2,*]

[1] College of Informatics, Huazhong Agricultural University, Wuhan, China

[2] State Key Laboratory of Tropical Crop Breeding, Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China

[3] Fujian Agriculture and Forestry University, Fuzhou, China.

[4] Institute of Animal Sciences, Chinese Academy of Agricultural Sciences, Beijing 100193, China.

[#]These authors contribute equally

[*]Corresponding authors: Xingtan Zhang, zhangxingtan@caas.cn; Haibao Tang, tanghaibao@gmail.com

Abstract

Ultra-complex, repeat-rich polyploids and hybrid genomes with extensive identical-by-descent regions pose a formidable challenge for chromosome-scale, haplotype-resolved assembly, since short-read chromatin conformation capture (3C) technologies such as Hi-C struggle to disambiguate their highly similar sequences. Long-read 3C technologies (e.g., Pore-C) provide potential to overcome this by generating multi-kilobase fragments, high-order concatemers and native 5-methylcytosine (5mC) signals. Here, we first optimized the Pore-C to ePore-C (enhanced Pore-C) protocol for plant tissues, which markedly boosts sequencing yield and reduces the cost per gigabase. We then present C-Phasing, a comprehensive assembly pipeline that integrates ePore-C long reads, high-order concatemers and native 5mC methylation profiles to resolve and scaffold even the most challenging regions. By leveraging methylation-based anchoring and improved statistical modeling, C-Phasing effectively rescues collapsed contigs and corrects chimeras, and increases both completeness and continuity of chromosome-scale assemblies. Fully compatible with both Pore-C and Hi-C data, C-Phasing outperforms existing tools across diverse polyploid genomes, including a modern cultivated sugarcane genome (aneuploid, 2n=9-12x=114), where it markedly enhances assembly quality and completeness over a previous reference.

Keywords: polyploid, ultra-complex genome, genome assembly, haplotype-resolved, scaffolding

Introduction

Polyploidy, the presence of more than two sets of chromosomes in an organism, is prevalent across various species, including plants, animals, and fungi[1,2]. Many polyploid species, such as wheat[3], sweet potato[4], sugarcane[5], and many fishes[6,7] are of significant importance to humans. Understanding the evolutionary history and improving breeding practices necessitates high-quality, haplotype-resolved and chromosome-scale genome assembly[8,9]. However, assembling polyploid genomes is inherently challenging due to their complex genomic structures[10], which often include highly similar or identical homologous sequences that can lead to frequent misassembly and fragmentation.

Recent advancements in sequencing technologies, particularly long-read sequencing methods like Pacific Biosciences (PacBio) circular consensus sequencing (CCS) or Oxford Nanopore Technologies (ONT) ultra-long[11–13], along with long-range linkage techniques, such as Hi-C[14,15], have become indispensable tools for assembling high-quality polyploid genomes. Several polyploid genomes, including those of wild sugarcane (*Saccharum spontaneum*) [16,17], cultivated alfalfa[18,19], cultivated potato[20–22], sweet potato[23], and modern cultivated sugarcane[24–26], have been successfully resolved by incorporating these innovative approaches. Our previous work demonstrated that Hi-C data could effectively phase and scaffold the genomes of wild sugarcane and cultivated alfalfa through the ALLHiC[27] and HapHiC[28] method. However, the genomic architecture of cultivated autopolyploid and allo-autopolyploid crops (such as potato, sweet potato, and sugarcane) has become ultra-complex due to recent inbreeding and introgression, characterized by a high proportion of similar or near-identical genomic regions due to identical-by-descent (IBD)[20,25]. While Hi-C reads, generated on short-read sequencing platforms, offer high base accuracy, they often struggle with mappability in regions of high sequence similarity, complicating haplotype phasing and chromosome-scale scaffolding in ultra-complex polyploid genomes.

To tackle these challenges, it is essential to incorporate additional genomic signals and employ more sophisticated assembly strategies. For instance, assembly of three tetraploid potato genomes and a hexaploid sweet potato genome required the sequencing of recombinant offspring populations or single-cell sequencing of pollen genomes[20–23].

However, the reliance on genetic maps, pollen genomes, trio data, or single flow-sorted chromosome libraries is not always feasible. Ideally, polyploid assemblies should utilize widely available data types, such as HiFi reads, ultra-long reads, and Hi-C data[29]. While contig-level assembly using HiFi and ultra-long reads has been effectively addressed by tools like hifiasm[30,31], challenges in haplotype phasing persist when relying solely on Hi-C data for ultra-complex polyploid assemblies. This is primarily due to the ambiguous mapping associated with the short 150-bp reads. The chance of ambiguous mapping tends to increase drastically as the repeat content, size, and ploidy of the target genome increase.

A promising alternative to Hi-C is the new 3C technology, Pore-C, which functions similarly to Hi-C but relies on sequencing with the ONT platform. Pore-C generates remarkably longer reads, allowing for the detection of high-order chromatin interactions across broader genomic regions[32]. Additionally, Pore-C retains methylation information, offering structural and epigenetic insights crucial for genome assembly. However, existing tools have yet to fully harness the advantages of Pore-C's long-read concatemers (a single molecule linking spatially proximal restriction fragments) and methylation features for haplotype-resolved assembly of ultra-complex polyploid genomes.

The standard Pore-C workflow suffers from low sequencing yield for plants[22,32,33], so we first addressed this by developing an optimized protocol (enhanced Pore-C, ePore-C) that dramatically increases throughput. We then introduce C-Phasing, a novel pipeline that leverages the features of Pore-C reads with several novel algorithmic improvements. First, C-Phasing introduces a function of *Hyperpartition*, a hypergraph-based clustering algorithm that employs high-order concatemers for accurate contig grouping. Second, we develop *Methalign*, a methylation-aware alignment method that utilizes 5mC sites in Pore-C reads to refine ambiguous mapping based on the methylation profiles derived from long-read sequencing. Third, a collapsed contig rescue module identifies and recovers collapsed sequences by duplicating them into their respective groups, substantially improving the completeness of chromosome-scale assemblies. Using C-Phasing, we generated chromosomal-scale and haplotype-resolved genome assemblies for key crops with ultra-complex genomes, such as hexaploid sweet potato and modern cultivated sugarcane. Fully compatible with both Pore-C and Hi-C data, our approach addresses assembly challenges

presented by these ultra-complex genomes and will accelerate the genetics research and breeding efforts for these vital crops.

Results

Problem statement of Hi-C scaffolding in ultra-complex polyploid genomes

Hi-C scaffolding has demonstrated its utility in constructing chromosome-level assemblies for various organisms[8]. However, a significant challenge arises when applying Hi-C scaffolding to ultra-complex polyploid genomes, as the short reads often lack sufficient sequence context to distinguish between homologous sequences efficiently (**Supplementary Fig. 1a-d**). We examined the Hi-C reads mapping results against four plant species with differing ploidies (**Supplementary Table 1**), including *Sorghum bicolor* (2n=2x=20)[34], *S. spontaneum* (1n=4x=32)[17], *S. officinarum* (2n=8x=80), and *S. hybrid* cultivar POJ2878 (2n=10-12x=118). For the diploid sorghum genome, the ratio of uniquely mapped reads is 49.54%; however, this value sharply decreases to 8.36% in the octoploid *S. officinarum* and 7.92% in the ultra-complex hybrid sugarcane genome (**Supplementary Fig. 2a**). The 49.54% uniquely mapped Hi-C reads in *S. bicolor* can span 40.72% (2,012/4,941) of assembled contigs, covering a vast majority (91.66%; 640.3 Mb/698.5 Mb) of the assembled genome size (**Supplementary Table 1**). In contrast, the extremely low level (<10%) of uniquely mapped Hi-C reads covers only approximately 5% of contigs and less than 70% of the assembled genome size in the high-ploidy sugarcane genome (**Supplementary Fig. 2b**), emphasizing the limited power of Hi-C reads in scaffolding complex genomes.

Another challenge in haplotype phasing is that the contig-level assembly may contain a considerable number of artifacts (i.e., chimeric and collapsed errors), leading to incorrect placement and phasing of sequences that compound over the size and complexity of the target genome (**Supplementary Fig. 1e,f**)[8]. Chimeric assembly errors occur when two or more regions of the genome that are not contiguous in reality are incorrectly joined together during the assembly process. The collapsed assembly errors indicate that two or more identical haplotypic regions are merged into a single sequence during the assembly process. We analyzed both types of assembly errors based on contiguous alignment and read-depth of

ONT ultra-long reads for three polyploid contig-level genome assemblies, including alfalfa (*Medicago sativa* L., 2n=4x=32), hexaploid sweet potato (*Ipomoea batatas* L., 2n=6x=90) and *S.* hybrid cv. POJ2878. These genomes were assembled using different contig assemblers (hifiasm[30,31], verkko[35] and HiCanu[36]), and the results reveal that all the assembly algorithms contain varying degrees of assembly errors, with 0.11%-2.37% of chimeric assembly errors and 2.01%-16.83% of collapsed regions (**Supplementary Fig. 2c and Supplementary Table 2**). Notably, the hifiasm primary unitig (p_utg) assemblies show the lowest proportion of chimeric errors and collapsed regions, making this algorithm widely accepted for haplotype phasing in polyploid genomes.

Enhanced Pore-C increased yield and reduced cost through optimized nucleus extraction

While powerful, the application of Pore-C for genome assembly and chromatin structure analysis, particularly in large genomes, is often hindered by limited data yield and high cost. Previous Pore-C protocols typically generated only 25-60 Gb of data per ONT sequencing cell (PromenthION), resulting in costs exceeding $22 per Gb (**Supplementary Table 3**)[22,33,37].

To address these limitations, we developed an enhanced Pore-C (ePore-C) protocol (**Methods**). This protocol incorporates a series of optimizations, most notably in the nucleus extraction step. By employing Percoll density gradient centrifugation, we efficiently removed plant-specific secondary metabolites such as polysaccharides and phenols (**Fig. 1a-b**). This refined extraction process yielded libraries that consistently met the highest quality standards (**Fig. 1c**).

ePore-C maintains the capacity to capture chromatin contacts comparable to the original Pore-C protocol while demonstrating an improved ability to obtain higher-order contacts. For instance, in *Arabidopsis thaliana*, ePore-C achieves a 71.61% rate for contacts of high order (contact order ≥ 3, contact order refers to the number of restriction fragments linked in a single concatemer), compared to 69.0% with the original protocol[37] (**Fig. 1d**). This improvement in high-order contact capture is coupled with a substantial increase in data yield, with an average of ~100 Gb per ONT sequencing cell (**Fig. 1e**). This increased

efficiency greatly reduces the cost of Pore-C sequencing to approximately $10 per Gb (**Fig. 1f)**.

Overview of the C-Phasing framework

To utilize the ePore-C data for haplotype-resolved genome assembly, we developed the C-Phasing algorithm package, which consists of five core components: three innovative algorithms - *Hitig*, *Align/Methalign*, *Hyperpartition*, and two modified functions, *Kprune* and *Scaffolding*, from our original ALLHiC and HapHiC programs (**Fig. 2 and see Methods**).

Hitig. Contig-level assembly errors—particularly chimeric and collapsed errors—are key factors impacting haplotype phasing assembly (**Supplementary Fig. 2**) [27]. Unlike earlier Hi-C scaffolding approaches such as LACHESIS[38] and ALLHiC[27], which infer contig adjacency from global contact strength, our strategy emphasizes the exclusion of erroneous regions prior to contig clustering. To achieve this, the *Hitig* algorithm performs split-read alignment of ONT ultra-long reads to detect structural inconsistencies (**Fig. 2a**). By computing the Longest Increasing Subsequence (LIS) of high-quality alignments, *Hitig* mitigates alignment noise from ONT sequencing errors and robustly identifies chimeric breakpoints based on LIS discontinuities. Collapsed and abnormal coverage regions are detected by modeling deviations in local sequencing depth. Regions exhibiting continuous mapping and normal coverage are retained as high-confidence regions (HCRs) and used for downstream contact filtering. To increase generalizability, *Hitig* also supports not only ultra-long reads but also Hi-C and Pore-C, providing users with flexible input options for diverse genome projects (**Supplementary Notes**).

Align/MethAlign. Pore-C provides longer reads than Hi-C[32], improving resolution of allelic regions with 99.0%-99.99% sequence identity (**Supplementary Fig. 3**). However, mapping ambiguity persists in regions with near-identical (>99.99%), limiting accurate haplotype phasing (**Supplementary Fig. 3**). To address this, the *Align/MethAlign* module combines sequence- and methylation-based alignment strategies (**Fig. 2b**). The *Align* module performs conventional alignment of Pore-C reads to the unitig assembly, followed by stringent MAPQ-based filtering for downstream partitioning in typical genomes. For ultra-complex genomes, *Methalign* leverages 5mC profiles—treated as a fifth base—penalizing

methylation discordance between reads and unitigs (**Fig. 2b**). Using 5mC profiles of unitigs derived from HiFi or ONT reads, we refined both Pore-C's primary and secondary alignments by re-soring based on methylation concordance, reassigning the highest-scoring alignment as primary (**Methods and Supplementary Notes**). This refinement increases the number of valid alignments and enhances haplotype partitioning accuracy in downstream modules.

The resulting alignments are subsequently used to construct a raw hypergraph that encodes high-order chromatin contacts in an $n \times m$ incidence matrix (n, unitig; m, concatemers). Each concatemer is also decomposed into virtual pairwise contacts (VPCs) for downstream scaffolding and contact matrix visualization (**Fig. 2b**).

Kprune: Kprune is a crucial component of the C-Phasing pipeline, building on the principles established by ALLHiC[27] for haplotype phasing. The effectiveness of ALLHiC in this context relies on its ability to identify allelic contigs while eliminating the inter-homologous-chromosome signals (h-*trans* errors), including inter-allelic signals (same locus on different homologous chromosomes) and cross-allelic signals (different loci on homologous chromosomes)[39] (**Supplementary Fig. 4**). This process is facilitated by the ALLHiC-Prune algorithm, which prevents the misjoining of highly similar homologous chromosomes, or "switch" errors. *Kprune* shares this core principle but offers improvements in speed and memory usage. While ALLHiC-Prune relies on a well-annotated reference genome to label allelic contigs, *Kprune employs* a fully reference-free approach, both simplifying the process and boosting accuracy. *Kprune* first scores pairwise contig alignments to quantify sequence similarity and builds a contig graph. Instead of ALLHiC-Prune's greedy pruning based on the labels, *Kprune* applies a maximum matching algorithm to isolate cross-allelic signal by exploiting the fact that *trans* (i.e., inter-chromosomal) contacts are much lower than *cis* contacts (i.e., intra-chromosomal)[38]. Once these inter- and cross-allelic signals are identified, they can be effectively removed from Hi-C or Pore-C data (**Fig. 2c**), ensuring that only true *cis* interactions remain for subsequent haplotype phasing.

Hyperpartition. We observed that h-*trans* signals—artifactual contacts between homologous chromosomes[40]—were significantly stronger than those between

nonhomologous chromosomes, especially in Pore-C data (i.e., nh-*trans*; Wilcoxon rank-sum test, $P<0.001$; **Supplementary Fig. 5**). These signals, often arising from ambiguous read mapping (**Supplementary Fig. 4**), motivated the design of a two-layer Louvain-based partitioning strategy in *Hyperpartition* (**Fig. 2d; Methods**).

In the first layer, contigs are clustered into homologous chromosome groups (HCGs) using a hypergraph constructed from high-order Pore-C interactions at MAPQ ≥ 1, which retains informative h-*trans* signals (**Supplementary Fig. 5**).

In the second layer, we raise the MAPQ threshold to ≥ 2 to remove most h-*trans* and suppress residual h-*trans* noise by reweighting edges with *Kprune*-derived allelic information. This reweighted graph enables accurate haplotype-level partitioning of contigs into chromosomes via community detection (**Fig. 2d**).

Scaffolding. Scaffolding in C-Phasing integrates optimized modules from ALLHiC and HapHiC (*fastsort*, *optimize*, and *build*), enabling efficient contig ordering and orientation within each group. ALLHiC and HapHiC calculate contig ordering and orientation using interactions within a fixed window (< 10 Mb), which reduces accuracy for large contigs (> 5 Mb). In contrast, C-Phasing uses global interaction scores across all contig pairs, improving performance for large contigs. And, parallelized processing substantially accelerates runtime, particularly for large and complex genomes. Unlike conventional methods[15,27,28] that discard low-quality alignments, C-Phasing retains all contacts, including those with MAPQ = 0, to recover connectivity in highly similar regions where signal dropouts are common. Additionally, chromosome-wide orientation is standardized across haplotype using inter-allelic strand information from *Kprune*, facilitating consistent haplotype alignment and clearer heatmap interpretation (**Fig. 2e**).

Benchmarking of the C-Phasing pipeline with simulation data

'Synthetic' polyploid genome. To rigorously benchmark C-Phasing, we generated 'synthetic' polyploid genomes of *Arabidopsis thaliana* with ploidy levels ranging from 2x to 12x. These were constructed using the TAIR10 reference genome[41] as a backbone and introducing homozygous SNPs and InDels from 12 previously published *A. thaliana* accessions (**Supplementary Table 8 and Supplementary Notes**)[42] to model allelic

variation. MASH distance analysis[43] revealed a broad spectrum of genomic similarity among accessions (**Fig. 3a and Supplementary Table 9**), with some (e.g., Altenb-2 and TAL-07; distance = 0.0024) highly similar, and others (e.g., Cvi-0; distance > 0.0054) more divergent.

Simulated chromosome-level assemblies were randomly fragmented to mimic contig-level assemblies with contig N50 values ranging from 50 kb to 2 Mb (**Supplementary Table 10**). To generate realistic contact maps, we incorporated ~50× real Hi-C and Pore-C data from Col-0[37,44] accession and substituted the corresponding variants from each accession. The resulting Pore-C dataset retained a high proportion of multi-way contacts (61.73%) and moderate read accuracy (mean 95.43%, **Fig. 3b,c**). The 'synthetic' dodecaploid genome, characterized by its high complexity and well-organized contact map (**Fig. 3d**), serves as a robust ground truth for assessing the efficacy of our algorithm (**Supplementary Fig. 8**).

Evaluation of Hitig. To assess *Hitig's ability* to identify and correct chimeric assembly errors, we compared it against CRAQ[45], GAEP[46], and tigmint-long[47], using simulated dodecaploid assemblies (2x to 12x ploidy; N50 = 500 kb) and ultra-long reads. To rigorously assess performance, synthetic assemblies were spiked with 2% intra-chromosomal and 18% inter-chromosomal misassemblies.

Hitig demonstrated precision comparable to CRAQ and GAEP in detecting chimeric errors, achieving near-perfect accuracy. However, it outperformed all other tools in recall, identifying a higher proportion (>87.50%) of true positive chimeric errors. This enhanced recall resulted in substantially higher F1 scores (>0.90)—a metric combining precision and recall—across all simulated genomes (**Fig. 3e**). Notably, *Hitig* also exhibited markedly reduced computational demands, featuring faster runtimes and lower memory usage compared to the other programs, thus proving to be both accurate and efficient.

Methalign evaluation. To assess the effectiveness of *Methalign* in improving the mapping accuracy of shorter ONT reads (e.g., Pore-C reads), we evaluated its performance using ultra-long reads (>100 kb) from sweet potato as a ground truth (**Methods**). These long reads' alignments were artificially fragmented to simulate shorter reads ranging from 250 bp to 5 kb, and then re-aligned to the reference genome (**Supplementary Fig. 9**). A fragment

was considered accurately mapped if it aligned to a genomic location with at least 80% overlap with its original position.

Our analysis revealed a direct correlation between read length and mapping accuracy, with longer fragments achieving higher accuracy (**Supplementary Fig. 10**). *Methalign* performed best with a penalty setting of 2, consistently improving alignment outcomes. For instance, without *Methalign*, 74.1% of 700 bp reads and 82.4% of 2 kb reads were accurately mapped (MAPQ ≥ 1). With *Methalign*, these proportions increased to 77.4% and 85.2%, respectively (**Fig. 3f**). Notably, for 700-bp simulated reads, *Methalign* successfully corrected 50.33% of previously misaligned reads (9.1% of total mapped reads) by reassigning them to the correct homologous chromosome, thereby providing the ability of downstream haplotype phasing when using Pore-C data (**Supplementary Table 12**).

Evaluation of Kprune. We benchmarked *Kprune* against HapHiC and ALLHiC using synthetic polyploid genomes with Hi-C data. *Kprune* demonstrated higher accuracy in detecting inter-allelic contig pairs than HapHiC in higher ploidy or greater contiguity, with recall advantages particularly pronounced in high-ploidy or low-contiguity assemblies (**Fig. 3g**). This performance gap arises from HapHiC's reliance on enriched h-*trans* signals between bins with high sequence similar[28]. In our synthetic polyploid genomes, increased haplotype similarity and number lead to a substantial reduction in the proportion of valid contacts, weakening the signal required for accurate inter-allelic identification by HapHiC. Although ALLHiC-Prune showed higher precision in diploid or short-fragment assemblies 0.91 for *Kprune* vs. 0.99 for ALLHiC-Prune at N50 50 kb, its performance is contingent on the availability of a reference genome and is further favored by the allele-consistent structure of our simulated polyploid genomes. In contrast, *Kprune* operates reference-free and achieves higher recall in low-contiguity assemblies (0.95 for *Kprune* vs. 0.82 for ALLHiC-Prune at an N50 of 50 kb).

A crucial step in haplotype phasing is the removal of h-*trans* contacts, including inter-allelic and cross-allelic contig pairs, while preserving true *cis* contacts. *Kprune achieved a superior balance, retaining over 94% of true cis contacts while removing the majority of inter-/cross-allelic artifacts across ploidy levels and assembly qualities* (**Fig. 3g and Supplementary Fig. 11**). In contrast, HapHiC exhibited poor removal efficiency (as low as

9.4% in dodecaploid assemblies), while ALLHiC-Prune suffered from reduced *cis* contact retention. These results highlight *Kprune*'s robustness in resolving inter- and cross-allelic noise, enabling more accurate haplotype phasing and chromosome partitioning in highly complex genomes.

Evaluation of Hyperpartition. We evaluated *Hyperpartition using simulated Hi-C and Pore-C data from the 'synthetic' dodecaploid genome*, comparing its (Hi-C-only) performance with HapHiC and ALLHiC at contig N50 values of 500 kb and 2 Mb. C-Phasing substantially outperformed both methods, achieving 98.75 and 99.87% correct assignment rates, respectively (**Fig. 3h).** In contrast, HapHiC and ALLHiC struggled to resolve highly similar haplotypes E and F, resulting in high misassignments and unanchored rates. For example, HapHiC exhibited overall misassignment and unanchored rates of 8.93% and 17.98% at 500 kb N50, and 4.94% and 4.60% at 2 Mb. Notably, haplotypes E and F were the primary contributors to these errors, with much higher rates: 36.70% and 44.75% at 500 kb N50, 29.67% and 22.95% at 2 Mb N50, respectively (**Fig. 3i and Supplementary Table 11)**. Unlike existing Hi-C-based tools, *Hyperpartition* benefits from Pore-C's enhanced contact resolution. Incorporating Pore-C data further reduced misassignments from 0.09% to 0.18% and the unanchored rates from 0.09% to 1.06% at 500 kb N50, highlighting its robustness for phasing in ultra-complex polyploid genomes.

Evaluation of scaffolding. The scaffolding performance of C-Phasing on a simulated polyploid genome is consistently comparable or better than HapHiC and ALLHiC (Supplementary Fig. 12 and Supplementary Notes). However, consider the relatively short chromosome lengths in the simulated datasets. To further benchmark, we evaluated its performance against state-of-the-art Hi-C-based tools—HapHiC[28], ALLHiC[27], YAHS[48], and 3D-DNA[15]—using haplotype-resolved human assemblies with varying contig N50 values. C-Phasing achieved comparable performance to at least one of these tools across all tested conditions **(Supplementary Fig. 13**).

Using either Hi-C or Pore-C input, C-Phasing produced scaffolds for the maternal haplotype that aligned almost perfectly to the reference. It achieved a Spearman's correlation coefficient of 1.00 at contig N50 values of 2 Mb and 10 Mb, demonstrating exceptional accuracy across a range of contig contiguities. At an N50 of 500 kb, the correlation slightly

declined to 0.92, consistent with other tools, likely due to misordering in repetitive centromeric regions. However, as contig contiguity improved, C-Phasing's performance increased, underscoring the importance of high-quality contig assemblies for optimal scaffolding accuracy.

Evaluation of C-Phasing with real-world polyploid genomes

To assess the robustness of C-Phasing in phasing and scaffolding real-world polyploid genomes, we applied it to three species of increasing ploidy: cultivar tetraploid alfalfa Zhongmu-4 (*M. sativa* L., 2n=4x=32), cultivar hexaploid sweet potato 1393 (*I. batatas* L., 2n=6x=90), and wild decaploid sugarcane 82-114 (*S. spontaneum*, 2n=10x=80) (**Supplementary Table 13**). For each genome, we generated two contig-level assemblies using hifiasm: one using HiFi reads alone (HiFi-only) and another integrating HiFi reads with ONT ultra-long reads (HiFi-UL), reflecting current *de novo* assembly strategies. We observed that HiFi-only assemblies exceeded the estimated genome sizes across all three species (**Fig. 4a and Supplementary Table 14**), indicating potential false sequence duplications during contig assembly in the absence of additional information. In contrast, HiFi-UL assemblies were more continuous, with higher contig NG50 values (e.g., 4.37 Mb vs. 1.54 Mb in alfalfa, 5.52 Mb vs. 1.62 Mb in sweet potato, and 25.75 Mb vs. 5.52 Mb in wild sugarcane; **Fig. 4b and Supplementary Table 14**).

Using Hi-C data, C-Phasing anchored more contigs with fewer misjoins than HapHiC. In alfalfa HiFi-only assembly, the adjusted anchor rate reached 88.86% for C-Phasing, compared to 88.55% for HapHiC, with a lower misjoin rate (0.53% vs. 0.72%). Similar improvements were observed in sweet potato (85.14% vs. 83.40% adjusted anchor rate; 0.98% vs. 2.25% misjoin rate) (**Fig. 4c-d and Supplementary Table 14**). Haplotype completeness and scaffold contiguity, measured by the allele completeness index (ACI) and synteny correlation index (SCI), C-Phasing showed slightly higher ACIs (**Fig. 4e**) but significantly higher SCIs compared to HapHiC ($P<0.01$; **Fig. 4f**), reflecting more complete and accurate scaffolding than HapHiC.

Despite high-depth Hi-C input (~91× for alfalfa, ~162× for sweet potato), the adjusted chromosomal anchor rate—normalized to estimated genome size—remained limited in fragmented or ultra-complex assemblies (e.g., 88.86% for alfalfa HiFi-only, 85.14% for sweet potato HiFi-only; **Supplementary Table 14**). Simulation benchmarks showed that C-Phasing could benefit from Pore-C data to improve anchoring completeness. As expected, using Pore-C data (processed with *Align*) increased anchor length across all assemblies. In alfalfa HiFi-only, Pore-C increased chromosomal anchoring by 3.00% over Hi-C, and in sweet potato, the gains were more substantial, rising to 4.99% and 4.21% for HiFi-only and HiFi-UL, respectively (**Fig. 4c**). Although Pore-C led to slightly higher misjoin rates in complex regions (**Fig. 4d**), it significantly improved ACI in sweet potato (P<0.05; **Fig. 4e**), with more modest gains in alfalfa and wild sugarcane.

These improvements stem from the higher contact yield of Pore-C: it produced more valid contacts (MAPQ ≥ 2) per gigabase, covered more genome regions, and rescued contigs with weak Hi-C connectivity, particularly in sweet potato (**Fig. 4g-i**). Moreover, C-Phasing consistently achieved the highest SCIs with both Pore-C and Hi-C data by fully leveraging all available contacts (including MAPQ = 0) (**Fig. 4f**).

In summary, C-Phasing effectively assembles haplotype-resolved, chromosome-level genomes across diverse ploidy levels using either Pore-C or Hi-C data (**Supplementary Fig. 14-19**). With Hi-C input, C-Phasing performs comparably to or better than HapHiC, and when leveraging Pore-C data, it achieves greater contiguity and completeness, especially in ultra-complex genomes like sweet potato.

Enhancing assembly accuracy and completeness by resolving chimeric and collapsed contigs

To evaluate C-Phasing's capacity to resolve structural errors in polyploid assemblies, we examined both chimeric and collapsed contigs in the alfalfa HiFi-UL assembly. Five chimeric contigs were manually curated based on discordant Pore-C contact patterns and subsequently used to benchmark the *Hitig* algorithm. With ONT ultra-long reads, *Hitig correctly* identified

four of the five; when incorporating Pore-C or Hi-C data, all five were successfully detected, demonstrating the utility of *Hitig* in misassembly detection (**Fig. 5a-b**).

To address collapsed contigs—regions where distinct haplotypes are erroneously merged—we integrated a rescue module into C-Phasing. This module identifies high-copy-number contigs (2 ≤ copy number < ploidy level) using read depth and chromatin interaction patterns (**Fig. 5c**). Contigs exhibiting strong interactions with multiple chromosomal groups in chromatin conformation data were duplicated and re-assigned to their correct haplotypes. Application of this strategy rescued 53 collapsed contigs (total length: 53.88 Mb), reducing the proportion of collapsed regions from 2.59% to 1.09% (**Fig. 5d-f**). Post-rescue analyses confirmed restored synteny (dot plots) and chromatin interaction coherence (Pore-C heatmaps), resolving coverage abnormalities and improving ACI across the 10 reconstructed chromosomes (**Fig. 5g-i**).

To validate the biological relevance of rescued contigs, we analyzed heterozygous single-nucleotide polymorphisms (SNPs) from HiFi read alignments. Of the 53 rescued contigs, 37 lacked heterozygous SNPs, consistent with homozygous regions that were collapsed during initial assembly. The remaining 16 contigs exhibited low heterozygosity (1-2 SNPs per contig, ~1 SNP/Mb), suggesting that these regions represent distinct haplotypes or segmental duplication that were not properly resolved by the assembler due to limited sequence divergence (**Fig. 5j**). These results underscore the feasibility and effectiveness of duplicating collapsed regions to reconstruct complete, haplotype-resolved polyploid genomes.

Enhancing ultra-complex region resolution via 5mC-guided alignment

Despite the high Pore-C sequencing depth (~125×), the total length of sweet potato chromosomes remained below the estimated genome size of 2.78 Gb. Anchoring efficiencies reached 90.13% and 94.27% for HiFi-only and HiFi-UL assemblies, respectively (**Fig. 4c and Supplementary Table 14**). This gap was largely attributed to fragmented assembly and reduced chromatin interaction signals in unanchored contigs, which were characterized by shorter lengths and lower contact density compared to anchored regions (**Fig. 6a-b**). Scaffold comparisons across homologous groups revealed undersized scaffolds, such as Chr01g5,

Chr01g6, and Chr02g6, indicating uneven scaffold completeness across haplotypes (**Supplementary Fig. 20**). Syntenic analysis of Chr01 (**Fig. 6c**) showed structural discontinuities within these size-discrepant scaffolds, spatially aligning with unresolved contig regions. Coverage analysis (**Fig. 6d**) ruled out contig collapse as the primary cause, as most windows showed coverage below 72×, corresponding to a copy number of two. Instead, missing sequences appeared concentrated in fragmented, unanchored contigs lacking chromatin interaction signals, which were often filtered out due to insufficient connectivity.

To overcome this limitation, we employed the *Methalign* module, which leverages allele-specific methylation profiles to refine ambiguous alignments and improve resolution in near-identical regions. Simulated data demonstrated that treating 5mC as a fifth base increases the number of unique alignments. To demonstrate its practical utility, we downsampled sweet potato Pore-C data to ~50× and compared *Methalign* against the conventional sequence alignment module (i.e., *Align*). *Methalign* utilizes two types of datasets—HiFi (~36×; namely HiFi-Methalign) and ONT reads (~41×; namely ONT-Methalign)—to identify the 5mC profiles of the unitig assembly.

We assessed *Methalign*'s impact on alignment quality and chromosomal anchoring using HiFi-only and HiFi-UL assemblies. Compared to the *Align* module, *Methalign* increased the heterozygosity between allelic contigs and between Pore-C reads and contigs (**Supplementary Fig. 21**). Then, *Methalign* substantially increased the proportion of valid alignments (MAPQ ≥ 2) and recovered additional reads previously discarded as singletons. HiFi-Methalign achieved 75.01% and 76.50% valid alignments—improvements of ~9.5% over *Align*—while ONT-Methalign showed similar trends (**Fig. 6e and Supplementary Table 15).** As a result, valid contact counts increased substantially, reaching 46.14 M and 45.53 M for HiFi-Methalign and 48.08 M and 47.65 M for ONT-Methalign, compared to 37.77 M and 39.30 M with *Align*. These additional contacts enhanced connectivity across previously weakly connected contigs, improving scaffold completeness (**Fig. 6e,f**). The total length of anchored contigs increased from 2,393.93 Mb (*Align*, HiFi-only) and 2,535.75 Mb (*Align*, HiFi-UL) to 2,857.71 Mb and 2,725.55 Mb for HiFi-Methalign, and to 2,842.17 Mb

and 2,724.41 Mb for ONT-Methalign, respectiely (**Fig. 6g**). These improvements translated into longer and more complete scaffolds, especially for homologous chromosomes such as Chr01g5/g6, Chr06g6, and Chr10g5/g6 (**Fig. 6h-i, Supplementary Fig. 22 and Fig. 23**). Despite some limitations in nearly homozygous regions with sparse allelic-specific methylation, *Methalign* effectively bridged many previously fragmented homologous regions, underscoring its utility for improving chromosome-scale assembly in ultra-complex genomic regions.

To further demonstrate *Methalign*'s effectiveness under different sequencing depths, we subsampled Pore-C data from 10 M to 170 M contacts. Across all depths, *Methalign* consistently outperformed *Align* in terms of total anchored length (**Fig. 6j**). Notably, just 20 M contacts with *Methalign achieved similar anchoring performance as Align's* maximum (**Fig. 6j**), highlighting that methylation-aware alignment offers greater efficiency than simply increasing sequencing depth to resolve ultra-complex genomic regions. Leveraging *Methalign*, we achieved chromosome-level assemblies of 2.84 Gb for HiFi-only and 2.73 Gb for HiFi-UL, using 70 M contacts, equivalent to approximately 50× Pore-C coverage.

Collectively, these results demonstrate that integrating allelic-specific methylation profiles via *Methalign* enhances alignment accuracy, improves valid contact recovery, and enables more complete chromosome-level assemblies, particularly in complex and highly similar genomic regions.

Improving chromosome-scale assembly of the modern cultivated sugarcane genome

As another case study, we evaluate C-Phasing's capacity to resolve the chromosome-scale architecture of sugarcane cultivar (ZZ01), a genetically complex crop with 9-12 homologous chromosome copies and pervasive IBD sequences from intensive inbreeding, repeated backcrossing and introgression[24–26]. Our prior assembly[24] combined HiFi reads and hifiasm to generate 11.22 Gb of contigs (NG50 = 527 kb), with Hi-C (~93× coverage) and trio sequencing data anchoring 9.06 Gb to 114 chromosomes. However, reliance on trio sequencing (often unavailable in commercial cultivars) and Hi-C short-read feature limited resolution in repetitive regions resulted in 2.16 Gb of the genome remaining unanchored (**Fig.**

7a). This was primarily attributable to the poor Hi-C interconnectivity among 3.68 Gb of contigs and limited Hi-C contacts across 3.76 Gb of genomic regions (**Fig. 7b**). Hence, C-Phasing with Hi-C anchored only 7.86 Gb of the assembled sequences (**Fig. 7b**), but it significantly improved scaffold completeness and continuity compared to the previous assembly (P<0.01; **Fig. 7c-d**).

To address the limitation of Hi-C, we applied ePore-C sequencing, generating 817 Gb (82× coverage) of Pore-C data. Although Hi-C yielded more contacts (281.6 M vs. 155.4 M), Pore-C doubled genomic coverage efficiency, reducing contact-depleted regions to 1.54 Gb and weakly connected contigs to 1.95 Gb (**Fig. 7a**). Integrating C-Phasing with Pore-C anchored 9.35 Gb (**Fig. 7b**)—a 23% improvement over Hi-C-based phasing (P=1.067×10^{-4})—while achieving statistically superior completeness (P=5.000×10^{-24}) and scaffold continuity (P=1.378×10^{-31}) versus prior methods (**Fig. 7c-d**). Synteny analysis against the *Erianthus rufipilus* reference genome[49] (**Fig. 7e**) and chromatin contact heatmaps (**Supplementary Fig. 24**) further supports the substantial improvement of this ultra-complex polyploid genome assembly using C-Phasing combined with Hi-C or Pore-C data.

Discussion

Ultra-complex polyploid genomes remain notoriously difficult to assemble at chromosomal-scale with haplotype-resolution, one reason is that conventional Hi-C-based scaffolding approaches are often limited in ultra-complex polyploids by poor mappability in near-identical haplotypes. To address this obstacle, we present C-Phasing, a modular and end-to-end pipeline that incorporates Pore-C with several novel algorithmic modules to enable high-quality, haplotype-resolved and chromosome-scale assemblies across a wide ploidy range and genome complexities.

The original Pore-C[32,33,37] was limited in yield in plants; we optimized it into ePoreC, introducing a refined nuclei-purification procedure that lifted per-flow-cell yield from 25-65 Gb to ~100 Gb and reduced costs from ~$22/Gb to ~$10/Gb, making Pore-C technology more feasible for large plant genomes.

C-Phasing consistently delivers improvements in completeness, continuity, and phasing accuracy across diverse genomes and ploidy levels, including real-world examples such as tetraploid alfalfa, hexaploid sweet potato, and decaploid wild sugarcane. Notably, in modern cultivated sugarcane—a genome characterized by high ploidy, massive content of IBD—C-Phasing anchored 9.35 Gb of sequences using Pore-C data—a 23% improvement over Hi-C-based phasing—while also significantly improving scaffold continuity.

Several algorithmic advances contribute to this performance. A two-layer hypergraph partitioning module, *Hyperpartition*, integrates h-*trans* contig pairs information derived from *Kprune* to deliver accurate haplotype clustering. It achieved over 98.5% assignment accuracy even in simulated dodecaploid genomes with contig N50s of 500 kb and 2 Mb, establishing Hyperpartition as a key component for accurate contig clustering in polyploid assemblies. The *Hitig efficiently corrects* chimeric and collapsed contigs, establishing high-confidence contig sets and regions for downstream scaffolding. In alfalfa, *Hitig* successfully corrected all five manually identified chimeric contigs. When combined with a collapsed rescue module, it recovered 53.88 Mb of previously collapsed regions, enhancing genome completeness and highlighting robust capability to resolve common misassemblies in complex polyploid genomes. The *Methalign* module introduces a 5mC-aware alignment refinement strategy to increase the yield of valid contacts in challenging regions. In sweet potato, it achieved comparable performance to non-5mC-aware alignment using only 20 million contacts, indicating that incorporating methylation information is more efficient for generating anchor signals than increasing sequencing depth for resolving complex regions. At higher coverage (~50×), *Methalign* effectively bridged fragmented and weakly supported regions, contributing to improved genome completeness.

While tailored to Pore-C, C-Phasing performs competitively or superiorly to existing tools using Hi-C data, yielding higher anchoring rates and lower misjoin errors than HapHiC. It is readily compatible with emerging chromatin conformation capture technologies, such as CiFi[50] (which sequences reads with PacBio platforms yield higher accuracy HiFi reads).

Nonetheless, several limitations remain. The current pipeline still inherits ONT base-error noise, which explains the slightly higher misjoin rate observed with Pore-C vs. Hi-C. The recent development of ONT duplex sequencing and improvements in basecalling

accuracy (e.g., Dorado model v5+), as well as HiFi-based methods (CiFi), offer promising paths forward. Another limitation is the handling of scaffold gaps. Currently, C-Phasing assembles contigs by joining them using placeholder gap sequences. Future extensions of the pipeline will incorporate string graph structures and ultra-long reads to enable more accurate gap filling and sequence recovery.

In conclusion, C-Phasing represents a significant advancement in ultra-complex polyploid genome assembly, addressing key limitations through innovative algorithms and Pore-C's unique features. Its successful application to challenging genomes, such as sweet potato and hybrid sugarcane, demonstrates practical utility for advancing our understanding of complex plant genomes, facilitating the identification of accurate genetic variations, high-resolution genetic map development, and providing a foundation for further studies on the genetics and evolution of polyploid species.

Methods

Experiments in this study

Plant materials. The sugarcane *S. spontaneum* 82-114 and *S.* hybrid cv. Zhong Zhe No. 1 (ZZ01) were cultivated in a greenhouse at 30°C with 16 hours of light per day, while sweet potato 1393 was cultivated at 23°C with 16 hours of light per day at the Chinese Academy of Agricultural Sciences (CAAS) Shenzhen Institute of Genomics. After 15 days of cultivation, young leaves or leaf sheaths were harvested for subsequent experiments.

ePore-C library construction and sequencing. We optimized the nucleus extraction procedures to improve the yield and quality of the Pore-C[32,33,37] in plants. The complete ePore-C protocol is described below:

Optimization of nucleus extraction and cross-linking procedures. About 2 g of young plant tissue was ground to a fine powder in liquid nitrogen and transferred to 15 mL pre-cold nucleus extraction Buffer I (0.4 mol/L sucrose, 10 mmol/L Tris-HCl, pH 8.0, 10 mmol/L KCl, 0.5% Triton X-100, 10 mmol/L EDTA, pH 8.0, 1% PVP40, 0.25% β-Mercaptoethanol, and 0.1 mmol/L PMSF). The homogenate was crosslinked by adding 471 μL of 37% formaldehyde (final concentration: 1%) and incubating for 25 minutes with rotation in a

vacuum desiccator at room temperature, releasing the vacuum every 5 minutes. Crosslinking was terminated by adding 963 μL of 2 mol/L glycine (final concentration: 0.125 mol/L) and incubating for 5 minutes. The mixture was filtered through a single layer of Miracloth (Merck, 475855-1R) followed by a 40 μm filter (Falcon, 32340). The filtrate was layered on 30 mL of dense sucrose buffer (1.7 mol/L sucrose, 20 mmol/L Tris-HCl, pH 8.0, 2 mmol/L KCl, 0.2% Triton X-100, 2 mmol/L EDTA, pH 8.0, 0.25% β-Mercaptoethanol and 0.1 mM PMSF) in a 50 mL Falcon tube and centrifuged at 2200 × g for 20 minutes at 4 °C. The supernatant was discarded, and the nuclear pellet was washed with 10 mL of pre-chilled 1× PBS and centrifuged under the same conditions for 10 minutes. The supernatant was discarded, and the nuclei were resuspended in the nucleus extraction buffer. For gradient centrifugation, 2 mL of 2.5 mol/L sucrose was added to a 15 mL Falcon tube, followed by the slow addition of 5 mL of 50% Percoll (Solarbio, P8370) and the nuclear suspension along the tube wall. The sample was centrifuged at 1800 × g for 30 minutes at 4 °C using a horizontal centrifuge with acceleration and deceleration rates set below 3. The Percoll layer solution was transferred to a new 15 mL Falcon tube, mixed with an equal volume of 0.25 mol/L sucrose solution, and centrifuged at 1800 × g for 20 minutes at 4 °C. The supernatant was discarded, and the pellet was resuspended and washed with pre-chilled 1× PBS. The sample was centrifuged at 1200 × g for 5 minutes at 4 °C, and the washing step was repeated twice. The supernatant was discarded after each wash.

Restriction enzyme digest. Next, 100 μL of 0.5% SDS was added to the nuclei, and the mixture was incubated at 62 °C for exactly 6 minutes. The tube was then allowed to return to room temperature. To quench the SDS, 250 μL of nuclease-free water and 50 μL of 10% Triton X-100 were added. The sample was gently mixed by pipetting up and down and incubated at 37 °C for 15 minutes. Then 30 μL (600 U) of Hind III (NEB, #R0104S), 50 μL of NEBuffer™ r2.1 (NEB, B6002V), and 20 μL of nuclease-free water were added. The reaction was incubated at 37 °C for 18 hours with periodic mixing in a thermomixer at 900 rpm (10 seconds every 30 minutes) to prevent condensation inside the tube lid.

Proximity ligation and reverse cross-linking. To inactivate the restriction enzyme, 55 μL of 1% SDS was added and incubated in a thermomixer at 65 °C for 20 minutes. The sample was allowed to cool to room temperature, and 50 μL was withdrawn, transferred to a

new 1.5 mL centrifuge tube, labeled as the digestion control sample, and stored temporarily at 4 °C. To the remaining nuclear suspension, a ligation reaction mixture was added to a final volume of 1000 μL. This mixture included 100 μL of 10× ligation buffer, 100 μL of 10% Triton X-100 (Sigma-Aldrich, 93443), 5 μL of 20 μg/μL Recombinant Albumin (NEB, B9200S), 40 μL of 400 U/μL T4 DNA ligase (NEB, M0202L), and 250 μL of nuclease-free water. The sample was gently mixed by pipetting up and down with a wide-bore pipette tip. The nuclear suspension was incubated at 16 °C in a thermomixer for 12 hours with periodic mixing at 900 rpm (10 seconds every 30 minutes). To reverse cross-links, the samples were treated with 280 μL of nuclease-free water, 100 μL of 20 mg/mL Proteinase K (NEB, P8107S), 100 μL of 10% SDS (Thermo Fisher Scientific,15553027), 500 μL of 20% Tween-20 (Sigma-Aldrich, P9416), and 20 μL of 100 mg/mL RNase A (Vazyme, DE111-01). The samples were incubated in a thermomixer at 56 °C for 18 hours with periodic mixing at 900 rpm (10 seconds every 30 minutes).

DNA purification. The 2 mL cross-linked sample was divided equally into two tubes (1 mL per tube). An equal volume of pre-chilled DNA extraction solution (phenol: chloroform: isoamyl alcohol = 25:24:1) (G-CLONE, EX0128) was added to each tube. The tubes were placed on a HulaMixer and mixed at 10 rpm for 5 minutes, followed by centrifugation at 15°C and 15,000 × g for 15 minutes. 850 μL of the upper aqueous phase was carefully transferred from each tube into new 2 mL centrifuge tubes. The extraction was repeated by adding an equal volume of pre-chilled DNA extraction solution (phenol: chloroform: isoamyl alcohol = 25:24:1) to each sample. The samples were mixed on the HulaMixer at 10 rpm for 5 minutes, centrifuged at 15°C and 15,000 × g for 15 minutes, and 800 μL of the upper aqueous phase was transferred from each tube into a new 15 mL centrifuge tube and combined. To the combined phase, 0.02 volumes of 5 mol/L NaCl (Thermo Fisher Scientific, AM9760G), 0.1 volumes of 3 mol/L sodium acetate (Solarbio, IS9096), and 3 volumes of absolute ethanol were added. The mixture was gently inverted to mix and precipitated at -80 °C for at least 1 hour or overnight at -20 °C. The sample was centrifuged at 4 °C and 15,000 × g for 30 minutes. The supernatant was discarded, and the pellet was washed with 4 mL of 80% ethanol, followed by centrifugation for 5 minutes. The pellet was then washed again with 2 mL of 70% ethanol, centrifuged for 5 minutes, and air-dried for 5 minutes. The

dried pellet was resuspended in 100 μL of nuclease-free water and incubated at room temperature for 5 minutes. DNA fragments larger than 1 kb were selected using 0.5× DNA Clean Beads (Vazyme, N411). DNA concentration was quantified using the Qubit 1× dsDNA HS Assay Kit (Thermo Fisher Scientific, Q33231), and 200 ng of DNA was loaded onto a 0.7% agarose gel to assess the efficiency of digestion and ligation.

Nanopore sequencing library construction. DNA was prepared for Nanopore sequencing using the Nanopore Ligation Sequencing Kit V14 (ONT, SQK-LSK114) according to the manufacturer's instructions, with the following modifications: The initial input was adjusted to 2.5 μg, yielding approximately 1 μg of DNA library, which is sufficient for sequencing 3-5 cells and helps reduce library preparation and sequencing costs. For DNA repair and end-prep, the samples were incubated in a thermal cycler at 20 °C for 20 minutes and 65 °C for 20 minutes to improve end-repair efficiency. During the adapter ligation and clean-up steps, the reaction was incubated at 30°C for 20 minutes to improve ligation efficiency. Beads were washed with LFB to ensure the retention of fragments larger than 3 kb, and DNA was eluted in 25 μL of EB. Approximately 150 ng of DNA was loaded onto FLO-PRO114M (R10.4.1) flow cells and sequenced on the PromethION for 96 hours. Data collection was managed using MinKNOW (version 24.02.19) software with default parameters.

The C-Phasing framework

This study presents an integrated computational framework, C-Phasing, comprising five core modules and auxiliary utility tools to achieve polyploid phasing and scaffolding. The principal workflow components include: (1) *Hitig*, high-quality contigs or regions identification, (2) *Align/Methalign*, Pore-C data processing without or with rescuing ambiguous alignments using methylation information, (3) *Kprune*, inter-allelic and cross-allelic contig pairs identification, (4) *Hyperpartition*, robust contigs clustering algorithm, and (5) *Scaffolding*, ordering and orienting contigs in each group. Supplementary modules, such as *Plot* (for fast visualization of assembly results via interaction signal heatmaps) and *HiCMapper* (a wrapper pipeline for Hi-C data processing), offer extended analytical functionalities. These components are systematically integrated into a cohesive C-Phasing

pipeline architecture, enabling one-command execution through standardized configuration parameters. The subsequent sections provide detailed descriptions of each module's implementation and its algorithmic foundations.

***Hitig*: high-quality contigs or region identification.** High-quality contig assembly constitutes the prerequisite for accurate phasing and scaffolding. Nonetheless, current *de novo* assemblers (e.g., HiCanu[36], hifiasm[30]) are prone to generate assemblies containing chimeric and collapsed sequences, particularly in complex polyploid genomes. Such assembly artifacts may propagate erroneous clustering results during chromosome partitioning and induce misassemblies, including misorderings or misorientations, in downstream scaffolding (**Supplementary Fig. 1**). To address these limitations, we developed *Hitig*, an ONT ultra-long read-based algorithm for detecting chimeric breakpoints and identifying high-confidence regions. *Hitig* also supports both Pore-C and Hi-C data; implementation details are provided in the Supplementary Notes.

Step A. Ultra-long reads processing. Due to high sequence similarity between homologous regions, error-prone ONT reads are prone to misalignment at homologous switch points, as the aligner tends to extend the longest seed chain, which impedes proper split mapping and leads to missed detection of chimeric assemblies. To improve breakpoints identification, we fragmented ONT ultra-long reads (>50 kb) into 5 kb sliding windows and aligned them using minimap2 v2.24-r1122[51] with parameters of "--qstrand --cs -cx map-ont -p.3". The relaxed threshold ("-p 0.3") preserves more secondary alignments, mitigating alignment truncation artifacts in homologous regions.

Step B. False alignments refinement. In cmplex polyploid genomes with high and extensive repetitive sequences, fragmented reads ofen fail to span adjacent SNP loci, resulting in misalignments. Additionally, split reads are prone to false alignments at contig boundaries, as longer inter-allelic contigs may yield higher alignment scores (AS), misleading the aligner. To address these issues in homologous regions and contig termini, we propose a windowed read correction algorithm based on the Longest Increasing Subsequence (LIS) approach[52]. This algorithm is composed of four key steps:

(1) Building LIS. To refine alignments, we first selected one alignment per read window based on the highest alignment score (AS), labeled as "P", and those with mapping quality ≥ 2, labeled as "M". These candidate alignments were then traversed in read order. During traversal, if the genomic coordinates between adjacent alignment pairs were within 1.25 times the window length, the LIS was extended. The traversal produced a raw LIS for each read, representing a coherent mapping path across homologous regions.

(2) Filling gaps. In regions of high sequence similarity, the optimal alignments reported by minimap2 may be inaccurate due to sequencing errors or misassemblies, causing correct alignments to be marked as secondary and resulting in gaps within the LIS. To address this, we filled gaps using selected secondary alignments. First, the LIS was sorted by genomic coordinates to identify gaps between adjacent windows; gaps exceeding twice the window length were excluded. A secondary alignment was used to fill a gap only if its AS exceeded 75% of the corresponding primary alignment, and was labeled as “S”.

(3) Extending LIS. At contig ends, correct alignments may yield lower AS due to limited length relative to their inter-allelic counterparts. To address this, we introduced a normalized metric, *Pscore* = (AS / Alignment length), to better evaluate alignment quality at contig termini. For any LIS window located at a contig end, we compute the *Pscore* for both primary and secondary alignments. The alignment with the highest *Pscore* is selected as the refined alignment and labeled as “E”.

(4) De-duplicating LIS. Duplicated sequences at contig ends and homologous exchanges can lead to multiple LISs overlapping at identical genomic coordinates. To eliminate redundancy, we first quantify each LIS by its total length, number of high-quality mappings ("M" tags), and overlapping windows. Fully overlapping LISs are removed iteratively by discarding the shortest, retaining those with more “M” tags in case of equal lengths. For partially overlapping LISs we resolved overlaps by traversing sorted pairs: if one contains an "E" tag (suggesting a correct alignment near the contig boundary), the overlapping region was assigned to it. Otherwise, overlaps were assigned to the preceding LIS to maintain structural consistency at contig junctions.

Step C. Detection of split reads and chimeric breakpoints

(1) Identification of split alignments. Reads with a single LIS indicate continuous mapping within a single locus. In contrast, reads containing multiple LISs on distinct contigs may reflect either (a) genuine contig junctions due to fragmented assemblies or (b) structural misassemblies, including inter-chromosomal chimeras (e.g., homologous switch misassemblies). To differentiate these scenarios, LISs are ordered based on their initial read window. For each adjacent LIS pair, we calculate their distances from the contig ends; if the sum is less than 1.2 × window length or both distances are below a defined threshold, the split is likely due to a natural contig boundary. Otherwise, it may signal a misjoin. To reduce noise from homozygous regions, we apply a 5 kb sliding window (1 kb step) to aggregate adjacent split alignments, merging windows supported by at least three consecutive split reads.

(2) Breakpoint identification. For each 5 kb window (1 kb step), we compute coverage depth based on corrected windowed alignments, defined as the total aligned base length divided by window length. To control for coverage fluctuations, we normalize the count of split alignments in each window by its average read depth. Regions with a normalized split-read ratio > 0.5 (i.e., >50% of supporting reads are fragmented) are flagged as candidate breakpoints. These loci are interpreted as potential chimeric juctions and are used to split the contigs for subsequent phasing and scaffolding.

Step D. Identification of High-confidence regions. High-confidence regions (HCRs) are defined as genomic intervals that are free of assembly errors, exhibit sufficient sequence specificity, and maintain consistent read coverage. We identify HCRs through a two-tiered filtering strategy: (a) Reads containing a single LIS with at least 60% of windows labeled as high-quality ("M" tag) are retained. Their mapped genomic coordinates are extracted as candidate high-confidence intervals. (b) Using all such reads, we compute the coverage depth per 5 kb genomic window (1 kb step). Windows with abnormally low coverage (suggesting poor-quality or junk sequences) or excessively high coverage (indicative of colapsed repeats or surious alignments) are excluded. The intersection of LIS-supported and coverage-consistent intervals defines the final set of HCRs. These regions are used to filter noisy or ambigous edges in the hypergraph of the Hyperpartition module, thereby improving contig

clustering accuracy. Additionally, collapsed contig detection within these regions facilitates targeted rescue and proper assignment in downstream C-Phasing processes.

Align/Methalign: Pore-C data processing. Raw Nanopore signal data were basecalled using dorado (v0.3.2+) (https://github.com/nanoporetech/dorado) with the modified base model dna_r10.4.1_e8.2_400bps_sup@v4.2_5mCG_5hmCG to call canonical and modified bases. The "—min-qscore 10" parameter was used to retain only high-quality reads. In this study, we processed Pore-C reads using two distinct strategies, as detailed below:

Step A. Align: Mapping Pore-C reads without methylation information. Pore-C reads were aligned to the assembly using minimap2[51] (2.24-r1122) with the '-cx map-ont –secondary=no' parameters. The resulting alignments were saved in PAF (Pairwise Mapping Format, https://github.com/lh3/miniasm/blob/master/PAF.md) files for downstream filtering.

***Step B. Methalign:* 5mC-guided mapping of Pore-C reads.** To improve mapping accuracy by leveraging allele-specific 5mC signals, we developed *Methalign*, a correction module that adjusts alignments based on methylation consistency. First, 5mC sites were identified from HiFi or ONT reads aligned to the contig assembly. Then, Pore-C reads containing 5mC modifications were aligned using the Dorado aligner with the "—mm2-opts '—secondary=yes'" parameter to retain secondary alignments. For each secondary alignment, its corresponding primary or supplementary alignment was located based on read position. The number of inconsistent 5mC sites between each Pore-C read and the contig was then used to compute an adjusted alignment score:

$$AS_{adj} = AS_{raw} - N \times P$$

where AS_{raw} is the original alignment score, N is the number of inconsistent 5mC modified sites, and P is the mismatch penalty (default = 2). If a secondary alignment's adjusted scores exceed those of its primary or supplementary counterpart, their alignment types are swapped, and the MAPQ is increased to prevent filtering in downstream analysis.

Finally, after completing alignment (Step A or Step B), we removed secondary alignments and filtered out low-quality primary/supplementary alignments with sequence identity < 0.8, alignment length < 150 bp, or singleton mappings. The remaining high-confidence alignments were formatted into a tabular file for subsequent hypergraph construction. To generate chromatin contact information, Pore-C concatemers were

decomposed into virtual pairwise contacts (VPCs) and stored in the 4DN (4D Nucleome) Pairs format[53]. These VPCs were further binned into multi-resolution contact matrices and converted to the 4DN cool format[54] for visualization and analysis.

Kprune: Inter-allelic and cross-allelic contig pairs identification. Contig-level assemblies in polyploid genomes often exhibit two types of h-*trans* artifacts: inter-allelic and cross-allelic (**Supplementary Fig. 3**). Inter-allelic contigs are highly similar homologous sequences that generate spurious contact signals, typically visible as diagonal lines in Hi-C/Pore-C heatmaps. Cross-allelic contigs, in contrast, originate from different homologous chromosomes and exhibit moderate contact frequencies—higher than those between unrelated contigs but lower than true cis contacts—potentially confounding partitioning algorithms.

To identify inter-allelic contig pairs, we used partig (https://github.com/lh3/partig) with parameters '-k 19 -w 19 -m 0.5 -d 0.1', requiring ≥ 10% sequence overlap based on shared k-mers. For cross-allelic pairs, we introduced two key improvements over our previous ALLHiC-Prune method.

First, we normalize the contact frequency between contigs i and j:

$$N_{ij} = \frac{C_{ij}}{\sqrt{C_i C_j}}$$

where C_{ij} is the raw contact count between contig i and j, and C_i, C_j are the total contact counts for contigs i and j.

Second, we applied the Kuhn-Munkres (Hungarian) algorithm to identify optimal one-to-one matches between inter-allelic contigs across haplotypes. For each candidate pair (contigs i_1 and j_1), we construct a bipartite graph between their inter-allelic partners (i_2, i_3, i_4, … and j_2, j_3, j_4, …), weighted by normalized contact scores (N_{ij}). If i_1 and j^1 form the strongest matched pair, they are assigned to the same haplotype; otherwise, they are classified as cross-allelic.

Finally, we constructed a reweighting matrix B for input into the *Hyperpartition* module, incorporating these relationships:

$$B_{ij} = \begin{cases} -1, & if\ i\ and\ j\ are\ interallelic \\ 0, & if\ \ i\ and\ j\ are\ \ crossallelic \\ 1, & otherwise \end{cases}$$

Hyperpartition: partitioning contigs into homologous groups or chromosomes. To group contigs into chromosomes, we developed Hyperpartition, a hypergraph-based clustering framework tailored for both diploid and complex polyploid genomes. In this framework, each contig is treated as a node in a hypergraph $H = (V, E)$, where nodes $V = \{V_1, V_2, V_3, \ldots, V_n\}$ represent contigs, and hyperedges E derive from Pore-C concatemers (monomers from the same contig are collapsed into one node; **Supplementary Table 4**). Low-confidence contigs (degree < 25) and abnormally large concatemers (> 50 monomers) are excluded to reduce noise.

To enable efficient clustering, we apply clique expansion to convert the hypergraph into a weighted adjacency matrix: Using the resulting hypergraph *H,* we apply the to cluster contigs into putative chromosomes (**Supplementary Table 5**). Before clustering, the hypergraph is transformed using clique expansion:

$$A = H(D_e - I)^{-1}H^{\mathrm{T}}$$

where H is the incidence matrix, D_e is the degree matrix of H, and I is the identity matrix. We remove self-loops and perform community detection using the Louvain algorithm[55] (default resolution = 1.0) to obtain initial chromosome-scale clusters. If the number of resulting groups n differs from the expected chromosome count k, we dynamically adjust the resolution: increasing it (by 0.2 increments) if $n<k$, or merging groups by modularity if $n>k$, thus enforcing exact k clustering. This "basal mode" shows strong performance for monoploid and allopolyploid genomes (**Supplementary Notes and Supplementary Tables 6-7**).

However, for highly similar homologs in polyploid genomes, basal clustering is often confounded by h-*trans* errors. To address this, we developed a two-layer phasing mode:

First layer: contigs are grouped into homologous chromosome groups (HCGs) using relaxed contact filtering (MAPQ ≥ 1).

Second layer: we increase the stringency of the MAPQ threshold to ≥ 2, then reweight the adjacency matrix to downweight erroneous inter-homolog links using the inter-allelic and cross-allelic mask matrix B:

$$A_r = A \odot B$$

The refined matrix is again clustered using Louvain, with an optional merge step if $n>k$. To prevent false merges across homologs, we apply an allele-aware merging constraint: groups with > 30% inter-allelic overlap are not merged, regardless of modularity. This ensures high-fidelity haplotype partitioning.

Scaffolding: contig ordering and orientation. To order and orient contigs within each group, C-Phasing builds on our previous ALLHiC[27] strategy by introducing key enhancements for efficiency and accuracy. In contrast to the version of "allhic optimize" used in ALLHiC and HapHiC, which compute ordering and orientation scores based on chromatin interaction within a fixed genomic window (typically < 10 Mb), we revised it to utilize global interaction scores across all contig pairs, which improved its performance of ordering and orientation for large contigs (> 5 Mb). And, rather than loading full contact matrices into memory and initializing from random populations—as done in "allhic optimize"—C-Phasing extracts group-specific link information directly from CLM files and parallelizes optimization across groups. To accelerate convergence, we integrate the results of "HapHiC fastsort" as an initialization step, significantly improving runtime[28].

Unlike prior tools that discard MAPQ=0 alignments, C-Phasing retains them to preserve linkage in highly similar regions, critical for polyploid genomes. Additionally, leveraging known inter-allelic contig pairs, we developed a *HaplotypeAlign* submodule that ensures consistent chromosome orientation across haplotypes. It calculates an orientation score (SO_{kl}) between k and l based on shared minimizers and relative contig directions:

$$SO_{kl} = \sum_{i,j} O_i O_j S_{ij} O_{ij}$$

where S_{ij} is the count of shared minimizers, O_{ij} is the relative orientation, and O_i and O_j represent within chromosome orientations. The first scaffold of HCG serves as the reference; homologs with $SO_{kl} < 0$ are flipped, ensuring uniform orientation and improving assembly readability.

Plot: contact matrix visualization. To facilitate rapid visualization of haplotype-resolved, chromosome-scale assemblies, we implemented an integrated plotting module. This module maps contig-level contact matrices (default resolution: 10 kb) to chromosome coordinates,

reorganizes them accordingly, and coarsens the matrix to a user-defined resolution (typically 500 kb) for heatmap generation.

HiCMapper: Hi-C data processing for C-Phasing. C-Phasing is also compatible with Hi-C data for phasing and scaffolding. Paired-end Hi-C reads were mapped to contig-level assemblies using a modified Chromap (v0.0.24)[56] with the parameters "--remove-pcr-duplicates -q 0 –preset hic". We adapted Chromap to output mapping quality into the 4DN Pairs format. To accommodate large genomes and prevent memory overflow, the k-mer and window sizes were adjusted: "-k 17 -w 7" for genomes < 8 Gb, and "-k 27 -w 14" for larger genomes. Hi-C contacts were converted into an incidence matrix using the same pipeline as for Pore-C, but instead of forming a hypergraph, a conventional graph was constructed, where nodes represent contigs and edges denote paired-end Hi-C contacts.

Collapsed contig identification and rescue. Collapsed contigs were identified based on sequencing depth derived from HiFi, ONT, or Pore-C reads (**Fig. 3c**). For HiFi and ONT, we adopted the sliding-window strategy from Hitig (5 kb window, 1 kb step) to calculate the alignment depth. For Pore-C, all alignments with MAPQ ≥ 0 (including singletons, which are reads that only contain a single fragment without proximal ligation) were processed through the Align module, and depth was similarly computed across sliding windows. Collapsed contigs were defined as those with an average sequencing depth exceeding 1.5× but below the expected ploidy. These contigs often exhibited strong inter-group contact signals beyond their assigned cluster. To address this, we duplicated such contigs into additional homologous groups (excluding those already containing a copy), up to the expected copy number. The rescued contigs were subsequently ordered and oriented using the scaffolding module.

Performance benchmarking methodology of simulation data

Simulation of the '*Synthetic*' polyploids. To benchmark C-Phasing, we simulated complex '***Synthetic'*** polyploid genomes using multiple *A. thaliana* accessions (**Supplementary Table 8**), including two highly similar accessions, Altenb-2 and TAL-07 (Mash distance = 0.0024, **Supplementary Table 9**). Variants data were downloaded from the 1001 Genomes Project[42], which used TAIR10[41] as a reference. Pseudo-haploid genomes for each accession were constructed using BCFtools v1.9[57] consensus.

To simulate Pore-C data for these pseudo-haploid genomes, we use real Pore-C data from the Col-0 accession[37], substituting alleles based on each accession's variant set. Pseudo-haploid genomes and Pore-C reads were then merged to create '*Synthetic*' polyploid genomes of varying ploidy levels. Additionally, to emulate ONT-specific sequencing errors (~5%), we employed PBSIM3[58] with the ERRHMM-ONT model. Finally, we simulated assemblies with varying contig N50 sizes (50 kb to 2 Mb), reflecting different levels of contiguity and assembly difficulty (**Supplementary Table 10**). Hi-C and Ultra-long reads were also simulated, with detailed procedures described in the **Supplementary Notes**.

Benchmark of chimeric contig identification using Ultra-long reads. To evaluate the accuracy of chimeric contig identification, we introduced simulated chimeras into 'syntenic' contig-level polyploid assemblies (N50 = 500 kb), modifying ~20% of contigs to contain artificial chimeric junctions: 10% intra-chromosomal, 72% inter-homologous, and 18% inter-nonhomologous events. Using simulated UL reads, we compared Hitig against existing tools including CRAQ v1.0.9-alpha [45], GAEP v1.2.3[46] and tigmint-long v1.2.10[59].

A breakpoint was considered correctly identified if it fell within ±5 kb of the true simulated location. Precision and recall were calculated accordingly to assess detection performance.

Benchmark of Methalign. To validate the accuracy of Methalign, we employed ONT ultra-long reads as the ground truth, given that longer reads typically have a higher likelihood of achieving correct alignment than shorter ones. We aligned the genomic ultra-long reads of sweet potato to the contigs and retained alignments exceeding 100 bp in length. Based on the length distribution of Pore-C read alignments obtained using different restriction enzymes in Human HG002 datasets, we split these ultra-long alignments into fragments ranging from 250 bp to 5 kb to simulate short ONT reads. These reads were then realigned to the sweet potato assembly using the dorado aligner. These simulated reads will return to their original chromosomes and loci if mapped correctly. Following this principle, we can categorize each mapped read as 'mapped to a homologous chromosome', 'mapped to a nonhomologous chromosome', 'mapped to the same chromosome but incorrect locus', or 'mapped to the same chromosome and correct locus'. This classification and statistical analysis were

conducted on both chromosome 1 and the entire genome using a Python script, `evaluate_alignments.py`.

Benchmark of phasing and scaffolding. To eliminate variability introduced by differing Hi-C mapping strategies, we applied a unified alignment set generated by *HiCMapper* across all tools. As *HiCMapper* outputs contact data in 4DN pairs format, we converted these into BAM format (containing only the information required by other tools) using "cphasing pairs2bam". This enabled consistent downstream analysis using 3D-DNA[15], ALLHiC[27], HapHiC[28], YAHS[48]. For phasing evaluations in polyploid genomes, we used HapHiC v1.0.2 and ALLHiC (commit: 80ce6a), while YAHS v1.2a.1 and 3D-DNA v180922 were included to assess contig ordering and orientation accuracy.

We first evaluated the accuracy of allelic contig identification, a critical step in contig partitioning for polyploid genomes. Ground truth of inter-allelic pairs were defined by overlapping coordinates between homologous chromosomes. Based on this, we computed the precision and recall of inter-allelic contigs identified by C-Phasing, HapHiC and ALLHiC. To further evaluate the accuracy of inter- and cross-allelic contigs identification, we measured the proportion of retained *cis* and h-*trans* Hi-C contacts. Here *cis* contacts refer to links between contigs from the same chromosome, while h-*trans* contacts connect homologous chromosomes. An effective method should preserve most *cis* contacts while removing the majority of h-*trans* contacts.

To assess clustering performance, including haplotype phasing, we computed the proportions of correctly assigned, misassigned and unanchored contigs. For each cluster, the source chromosome was defined as the one contributing the majority of contigs; contigs from other chromosomes within the same cluster were counted as misassignments. Contigs not assigned to any cluster were considered unanchored. We further compared the phasing performance of C-Phasing using Pore-C and Hi-C data.

To further evaluate scaffolding accuracy, we used the haplotype-resolved human HG002 genome[60] alongside corresponding Hi-C[60] and Pore-C[32] data. Chromosomes were fragmented into contigs of varying N50 sizes (500 kb, 2 Mb, and 10 Mb). Hi-C and Pore-C reads were aligned to these assemblies using our *HiCMapper* and *Align* pipelines, retaining all alignments (including MAPQ = 0). Each scaffolding method was assessed via dot plots,

where 10 kb windows of contigs were plotted as data points. Scaffolding accuracy was quantified by calculating the Spearman correlation coefficient (ρ) between the results and reference chromosomal coordinates.

Performance benchmarking methodology of real datasets

Strategy for contig-level assembly of real datasets. To benchmark our pipeline on real-world polyploid genomes, we generated contig-level assemblies for three polyploids using hifiasm v 0.19.1-r559[30,31] with two strategies: (1) HiFi-only, assembled from HiFi reads using default parameters, and (2) HiFi-UL, integrating HiFi and ONT ultra-long reads with "--ul" parameter. The resulting primary unitigs (p_utg) from hifiasm were used as input for all downstream benchmarking analysis.

Processing of Hi-C data on real datasets. To benchmark the performance of our pipeline against HapHiC, Hi-C data were processed following the guidelines provided in each tool's official documentation. For C-Phasing, we employed Chromap to process Hi-C data, following the procedures described in the *HiCMapper* Methods. For HapHiC, we followed the official tutorial (https://github.com/zengxiaofei/HapHiC?tab=readme-ov-file#align), mapping Hi-C data to the contig-level assembly using BWA-MEM v0.7.17[61] with "-5SP" parameter. PCR duplicates were removed using samblaster v0.1.26[62], and unmapped, secondary, and supplementary alignments were excluded using samtools[57] view -F 3340 (v1.15.1). Additional filtering was performed using HapHiC's "filter_bam" utility, removing alignments with mapping quality < 1 and edit distance ≥ 3.

Evaluation of misjoined contigs using ultra-long ONT reads. To assess structural accuracy across assemblies, ultra-long ONT reads were aligned to the final chromosomal-level scaffolds using minimap2 with "-x map-ont –secondary=no" parameters. Alignments with MAPQ < 2 were discarded. Misjoins were detected using "paftools.js misjoin" (https://github.com/lh3/minimap2/blob/master/misc/paftools.js) with "-l 20k" parameter to identify split alignments indicative of structural errors, where a single read maps to two distant regions. We focused exclusively on contig-level misjoins, excluding those attributable

to upstream de novo assembly errors. Candidate misjoins were retained only if (i) both breakpoints overlapped contig ends and (ii) were supported by $\geqslant$ independent reads.

Evaluation of assembly by synteny. In this study, we utilized closely related diploid species of the polyploids as references to guide the renaming of scaffolds using "cphasing rename" (**Supplementary Notes**) and perform coding sequences annotation through the GMAP v20131028[63]. To assess chromosomal assembly completeness, we defined an allele completeness index (ACI) based on gene content. First, allelic sequences were identified from the above annotations following the ALLHiC methodology. Complete allelic sets were clusters containing exactly *n* homologous loci, where *n* corresponds to the experimentally determined ploidy level (e.g., tetraploid *n*=4). The ACI for each chromosomal scaffold was then calculated as:

$$ACI = \frac{N}{N_t}$$

where N is the number of complete allelic sets in the scaffold and N_t is the total number of genes in the reference chromosome, and we only retain genes corresponding to the complete allelic sets.

Furthermore, we constructed synteny dot plots using the JCVI toolkit v1.4.18[64] to systematically evaluate scaffolding quality based on chromosomal collinearity. To objectively quantify assembly continuity, we defined a synteny correlation index (SCI) as the product of the maximum length of the longest increasing or the longest decreasing subsequence (LIS or LDS) of synteny anchors and the total number of anchors on each scaffold.

Evaluating the h-*trans* error between homologous chromosomes

To evaluate the h-*trans* error between homologous chromosomes, we normalized the contacts by following the formula:

$$N_{ij} = \frac{C_{ij}}{\sqrt{C_i C_j}}$$

where C_i and C_j denote the *cis* contacts in chromosome and C_{ij} represents the *trans* contacts between chromosomes i and j.

Whole genome coverage and variant calling of collapsed contig

HiFi reads were aligned to the scaffolds using minimap2 with the parameters "-ax map-hifi –secondary=no". Coverage depth was computed in 500 kb windows using Mosdepth v0.3.3[65] (https://github.com/brentp/mosdepth). To assess heterozygosity between collapsed contigs and their missing homologous copies, SNPs were called from aligned HiFi reads using longcallD v 0.0.3 (https://github.com/yangao07/longcallD) with "—hifi" parameter. Low-confidence variants, those supported by fewer than half the average HiFi read depth (e.g., <20 × in alfalfa), were excluded from downstream analysis.

Data availability

The sequencing data generated in this study have been deposited in the China National Center for Bioinformation (CNCB) GSA database (https://ngdc.cncb.ac.cn/gsa/) under BioProject **No. xxxxxxx.** The *A. thaliana* Col-0 Pore-C data were obtained from GSA in the CNCB under CRA005105, and the Hi-C data were downloaded from the National Center for Biotechnology (NCBI) (SRR2626163). Genetic variants from diverse *A. thaliana* ecotypes were retrieved from the 1001 Genomes Project (https://1001genomes.org/data/GMI-MPI/releases/v3.1/intersection_snp_short_indel_vcf/). The HG002 haplotype-resolved genome v1.1 was downloaded from the Telomere-to-telomere Consortium HG002 "Q100" project https://github.com/marbl/HG002. The HG002 Pore-C data were collected from NCBI SRX8156773 and SRR27664048. The Hi-C data of HG002 were taken from the Human Pangenome Reference Consortium: https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=working/HPRC_PLUS/HG002/raw_data/hic/downsampled/. Hi-C and HiFi data for *M. sativa* Zhongmu-4 were obtained from the CNCB under accession PRJCA004062, and ONT UL data and Pore-C data were retrieved from CNCB under accessions PRJCA030790 and PRJCA041059, respectively. The monoploid genome of alfalfa was derived from the haplotype-resolved assembly of Xinjiangdaye haplotype 1 (https://figshare.com/projects/whole_genome_sequencing_and_assembly_of_Medicago_sativa/66380). The sweet potato *I. batatas* Beauregard consensus genome and annotation were

taken from https://sweetpotato.uga.edu/sweetgains_Beauregard_con_v1_asm_anno.shtml. The genome and annotation of *Erianthus rufipilus* were downloaded from the Saccharum Genome database[66] (https://sugarcane.gxu.edu.cn/scdb/download).

Code availability

The source code of C-Phasing program can be found at GitHub (https://github.com/wangyibin/CPhasing). Detailed documentation is available at https://wangyibin.github.io/CPhasing/latest/.

References

1. Comai, L. The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.* **6**, 836–846 (2005).
2. Van de Peer, Y., Mizrachi, E. & Marchal, K. The evolutionary significance of polyploidy. *Nat. Rev. Genet.* **18**, 411–424 (2017).
3. Brenchley, R. *et al.* Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* **491**, 705–710 (2012).
4. Yang, J. *et al.* Haplotype-resolved sweet potato genome traces back its hexaploidization history. *Nat. Plants* **3**, 696–703 (2017).
5. Garsmeur, O. *et al.* A mosaic monoploid reference sequence for the highly complex genome of sugarcane. *Nat. Commun.* **9**, 2638 (2018).
6. Xu, M.-R.-X. *et al.* Maternal dominance contributes to subgenome differentiation in allopolyploid fishes. *Nat. Commun.* **14**, 8357 (2023).

7. Wang, B. *et al.* Whole-genome sequencing reveals autooctoploidy in Chinese sturgeon and Its evolutionary trajectories. *Genomics Proteomics Bioinformatics* **22**, qzad002 (2024).
8. Kong, W., Wang, Y., Zhang, S., Yu, J. & Zhang, X. Recent advances in assembly of plant complex genomes. *Genomics Proteomics Bioinformatics* (2023) doi:10.1016/j.gpb.2023.04.004.
9. Turner-Hissong, S. D., Mabry, M. E., Beissinger, T. M., Ross-Ibarra, J. & Pires, J. C. Evolutionary insights into plant breeding. *Curr. Opin. Plant Biol.* **54**, 93–100 (2020).
10. Garg, V. *et al.* Unlocking plant genetics with telomere-to-telomere genome assemblies. *Nat. Genet.* **56**, 1788–1799 (2024).
11. Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018).
12. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
13. Sereika, M. *et al.* Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nat. Methods* **19**, 823–826 (2022).
14. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
15. Dudchenko, O. *et al.* De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).

16. Zhang, Q. *et al.* Genomic insights into the recent chromosome reduction of autopolyploid sugarcane Saccharum spontaneum. *Nat. Genet.* **54**, 885–896 (2022).
17. Zhang, J. *et al.* Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L. *Nat. Genet.* **50**, 1565–1573 (2018).
18. Chen, H. *et al.* Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa. *Nat. Commun.* **11**, 2494 (2020).
19. Long, R. *et al.* Genome assembly of alfalfa cultivar Zhongmu-4 and identification of SNPs associated with agronomic traits. *Genomics Proteomics Bioinformatics* **20**, 14–28 (2022).
20. Sun, H. *et al.* Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. *Nat. Genet.* **54**, 342–348 (2022).
21. Bao, Z. *et al.* Genome architecture and tetrasomic inheritance of autotetraploid potato. *Mol. Plant* **15**, 1211–1226 (2022).
22. Serra Mari, R. *et al.* Haplotype-resolved assembly of a tetraploid potato genome using long reads and low-depth offspring data. *Genome Biol.* **25**, 1–19 (2024).
23. Wu, S. *et al.* Phased chromosome-level genome assembly provides insight into the origin of hexaploid sweetpotato. *bioRxiv* 2024.08.17.608395 (2024) doi:10.1101/2024.08.17.608395.
24. Bao, Y. *et al.* A chromosomal-scale genome assembly of modern cultivated hybrid sugarcane provides insights into origination and evolution. *Nat. Commun.* **15**, 3041 (2024).

25. Healey, A. L. *et al.* The complex polyploid genome architecture of sugarcane. *Nature* 1–7 (2024) doi:10.1038/s41586-024-07231-4.
26. Zhang, J. *et al.* The highly allo-autopolyploid modern sugarcane genome and very recent allopolyploidization in *Saccharum*. *Nat. Genet.* 1–12 (2025) doi:10.1038/s41588-024-02033-w.
27. Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat. Plants* **5**, 833–845 (2019).
28. Zeng, X. *et al.* Chromosome-level scaffolding of haplotype-resolved assemblies using Hi-C data without reference genomes. *Nat. Plants* 1–17 (2024) doi:10.1038/s41477-024-01755-3.
29. Li, H. & Durbin, R. Genome assembly in the telomere-to-telomere era. *Nat. Rev. Genet.* 1–13 (2024) doi:10.1038/s41576-024-00718-w.
30. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
31. Cheng, H., Asri, M., Lucas, J., Koren, S. & Li, H. Scalable telomere-to-telomere assembly for diploid and polyploid genomes with double graph. *Nat. Methods* **21**, 967–970 (2024).
32. Deshpande, A. S. *et al.* Identifying synergistic high-order 3D chromatin conformations from genome-scale nanopore concatemer sequencing. *Nat. Biotechnol.* 1–12 (2022) doi:10.1038/s41587-022-01289-z.

33. Huang, X. *et al.* Multi-omics mapping of chromatin interaction resolves the fine hierarchy of 3D genome in allotetraploid cotton. *Plant Biotechnol. J.* **20**, 1639–1641 (2022).
34. Paterson, A. H. *et al.* The Sorghum bicolor genome and the diversification of grasses. *Nature* **457**, 551–556 (2009).
35. Rautiainen, M. *et al.* Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat. Biotechnol.* **41**, 1474–1482 (2023).
36. Nurk, S. *et al.* HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* **30**, 1291–1305 (2020).
37. Li, Z. *et al.* Pore-C simultaneously captures genome-wide multi-way chromatin interaction and associated DNA methylation status in Arabidopsis. *Plant Biotechnol. J.* **20**, 1009–1011 (2022).
38. Burton, J. N. *et al.* Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013).
39. Zhang, X. *et al.* Haplotype-resolved genome assembly provides insights into evolutionary history of the tea plant *Camellia sinensis*. *Nat. Genet.* 1–10 (2021) doi:10.1038/s41588-021-00895-y.
40. Edge, P., Bafna, V. & Bansal, V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res.* **27**, 801–812 (2017).
41. Lamesch, P. *et al.* The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* **40**, D1202–D1210 (2012).

42. Alonso-Blanco, C. *et al.* 1,135 Genomes Reveal the Global Pattern of Polymorphism in *Arabidopsis thaliana*. *Cell* **166**, 481–491 (2016).
43. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).
44. Wang, C. *et al.* Genome-wide analysis of local chromatin packing in Arabidopsis thaliana. *Genome Res.* **25**, 246–256 (2015).
45. Li, K., Xu, P., Wang, J., Yi, X. & Jiao, Y. Identification of errors in draft genome assemblies at single-nucleotide resolution for quality assessment and improvement. *Nat. Commun.* **14**, 6556 (2023).
46. Zhang, Y., Lu, H.-W. & Ruan, J. GAEP: a comprehensive genome assembly evaluating pipeline. *J. Genet. Genomics* (2023) doi:10.1016/j.jgg.2023.05.009.
47. Jackman, S. D. *et al.* Tigmint: correcting assembly errors using linked reads from large molecules. *BMC Bioinformatics* **19**, 393 (2018).
48. Zhou, C., McCarthy, S. A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* **39**, btac808 (2023).
49. Wang, T. *et al.* A complete gap-free diploid genome in *Saccharum* complex and the genomic footprints of evolution in the highly polyploid *Saccharum* genus. *Nat. Plants* **9**, 554–571 (2023).
50. McGinty, S. P. *et al.* CiFi: Accurate long-read chromatin conformation capture with low-input requirements. *bioRxiv* 2025.01.31.635566 (2025) doi:10.1101/2025.01.31.635566.
51. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

52. Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
53. Open2C *et al.* Pairtools: From sequencing data to chromosome contacts. *PLOS Comput. Biol.* **20**, e1012164 (2024).
54. Abdennur, N. & Mirny, L. A. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics* **36**, 311–316 (2020).
55. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, P10008 (2008).
56. Zhang, H. *et al.* Fast alignment and preprocessing of chromatin profiles with Chromap. *Nat. Commun.* **12**, 6566 (2021).
57. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, (2021).
58. Ono, Y., Hamada, M. & Asai, K. PBSIM3: a simulator for all types of PacBio and ONT long reads.
59. Jackman, S. D. *et al.* Tigmint: correcting assembly errors using linked reads from large molecules. *BMC Bioinformatics* **19**, 393 (2018).
60. Jarvis, E. D. *et al.* Semi-automated assembly of high-quality diploid human reference genomes. *Nature* 1–13 (2022) doi:10.1038/s41586-022-05325-5.
61. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv Prepr. ArXiv13033997* (2013).
62. Faust, G. G. & Hall, I. M. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**, 2503–2505 (2014).

63. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
64. Tang, H. *et al.* JCVI: A versatile toolkit for comparative genomics analysis. *iMeta* **3**, e211 (2024).
65. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867–868 (2018).
66. Chen, S. *et al.* ScDB: A comprehensive database dedicated to Saccharum, facilitating functional genomics and molecular biology studies in sugarcane. *Plant Biotechnol. J.* **22**, 3386–3388 (2024).

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2021YFF1000900 to X.Z., 2024YFF1000800 to H.T.), National Natural Science Foundation of China grant (No. 32222019 and No. 32370714 to X.Z.) and Basic Research Center for Agricultural Frontiers and Interdisciplinary Sciences (BRC-AFIS), Innovation Program of Chinese Academy of Agricultural Sciences (CAAS-BRC-AFIS-2025-02 to X.Z.).

Author contributions

X. Zhang conceived the project and coordinated research activities. Y.W. designed and developed the C-Phasing framework and wrote the manuscript. P.Z. developed the ePore-C and performed all experiments. X. Zeng developed the Methalign module. J.Y. developed the Hitig module. A.D., Y.L., and M.J. assisted in the benchmarking of C-Phasing. X.C., F.W., S.Z., and S.C. assisted in collecting data and evaluating different assemblers. Y.G. assisted in the construction of ePore-C libraries. X. Zeng and Y.Z. provided suggestions for the algorithms. M.W. and R.L. supplied the materials and part of the data for cotton and alfalfa,

respectively. X. Zhang and H.T. supervised the entire project. All authors read and approved the final manuscript.

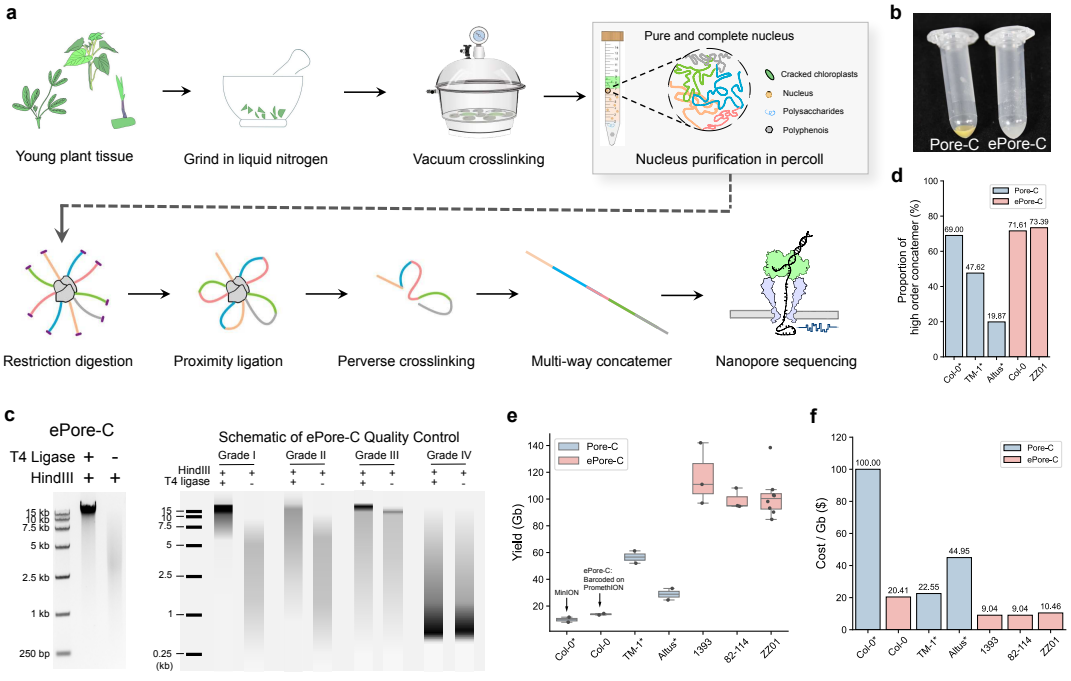


Fig. 1 | Enhanced Pore-C improves sequencing yield. **a**, Overview of the enhanced Pore-C (ePore-C) workflow, highlighting the nucleus purification step for plant tissues. **b**, Nucleus purity comparison between standard Pore-C (left) and ePore-C (right), the nucleus pellet at the bottom of the centrifuge tube is shown; a yellowish-green tint indicates residual impurities. **c**, Quality assessment of ePore-C libraries. Left: agarose-gel electrophoresis of alfalfa ePore-C libraries (Grade I indicates high-quality); Ladder (M) denotes 15 kb. Right: Grading schematic illustrating four ePore-C library quality tiers (Grades I-IV), with Grade I representing fully digested and ligated libraries. **d**, Proportion of high-order concatemers (order ≥ 3) for DpnII-digested samples using Pore-C versus ePore-C. **e,f**, Sequencing yield (**e**) and cost per gigabase (**f**) for Pore-C and ePore-C. Yield comparisons include published Pore-C datasets (Col-0, Li, Z. et al., 2022; TM-1, Huang, X. et al. 2022; Altus, Serra Mari, R. et al. 2024) and ePore-C datasets generated here (Col-0, 1393, 82-114, ZZ01). Cost comparisons span published and newly generated datasets (Col-0, TM-1, Altus; 1393, 82-114, ZZ01). These samples represent different species: Col-0 (*Arabidopsis thaliana*), TM-1 (upland cotton), Altus (cultivar potato), 1393 (sweet potato), 82-114 (wild sugarcane, *Saccharum spontaneum*), and ZZ01 (modern cultivated sugarcane). Star markers (*) in **d-f** indicate referenced studies.

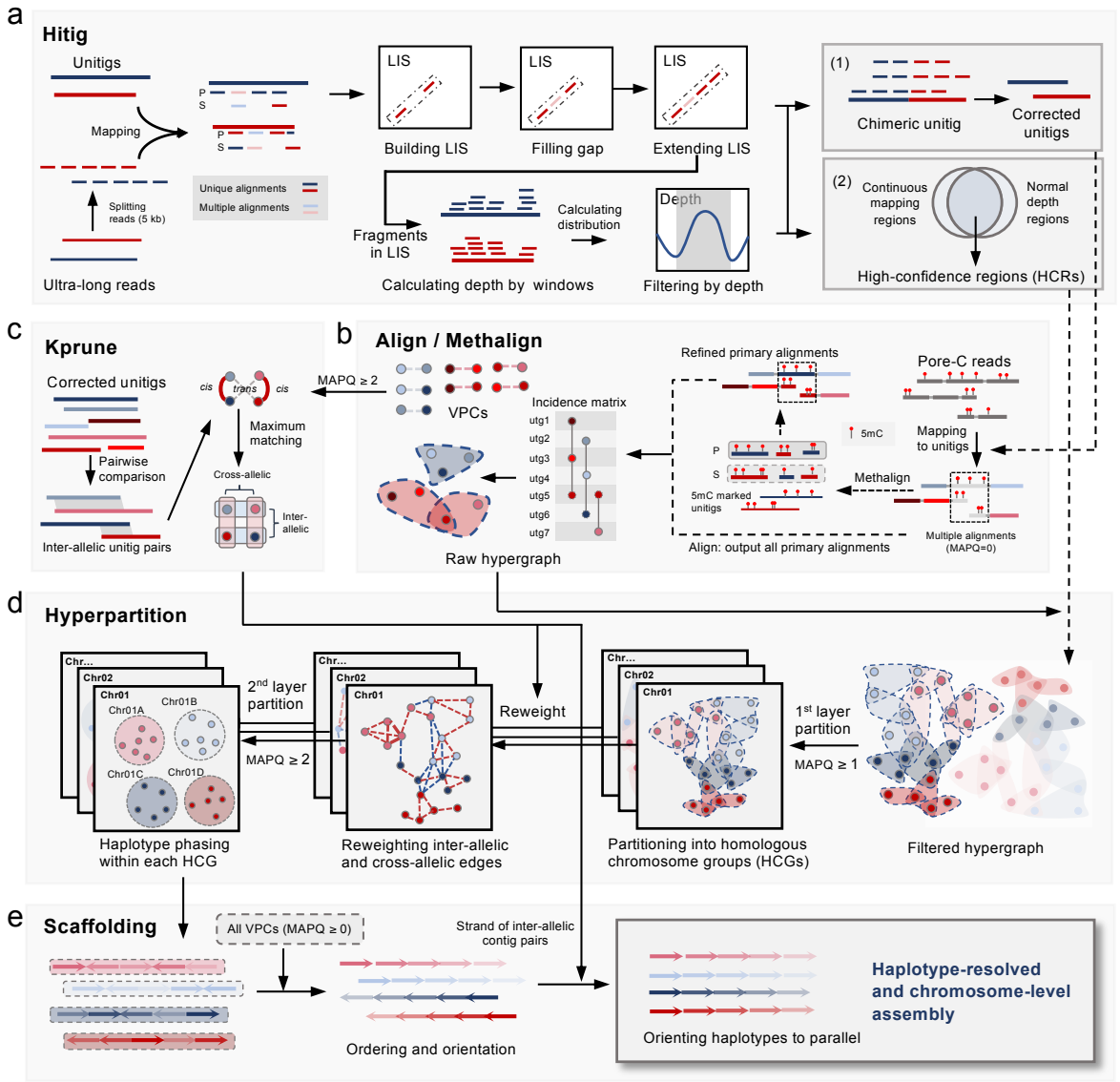


Fig. 2 | Overview of the C-Phasing framework. **a**, In Hitig, ultra-long (UL) reads are split and aligned to unitigs to compute the longest increasing subsequence (LIS) of high-confidence mappings. Gaps and edge alignments are refined, enabling the identification of chimeric unitigs and high-confidence regions (HCRs). **b**, Pore-C reads are mapped to corrected unitigs using either a baseline alignment (Align) or a 5mC-aware correction (Methalign), which refines mappings based on 5mC site concordance. Primary alignments construct a raw hypergraph; virtual pairwise contacts (VPCs) are also extracted. **c**, Kprune detects inter-allelic contig pairs based on sequence similarity, and applies maximum matching to distinguish cross-allelic pairs, guiding accurate phasing. **d**, Hyperpartition constructs a hypergraph filtered by HCRs, followed by a two-layer Louvain-based community detection. In the first layer, unitigs are grouped into homologous chromosome groups (HCGs) based on h-*trans* and *cis* contact signals. In the second layer, within each HCG, contigs are further partitioned into individual chromosomes using a graph reweighted by inter- and cross-allelic information. **e**, Unitigs within each group are ordered and oriented using all VPCs. Inter-allelic orientation consistency is enforced to align all homologous chromosomes to a common haplotype direction.

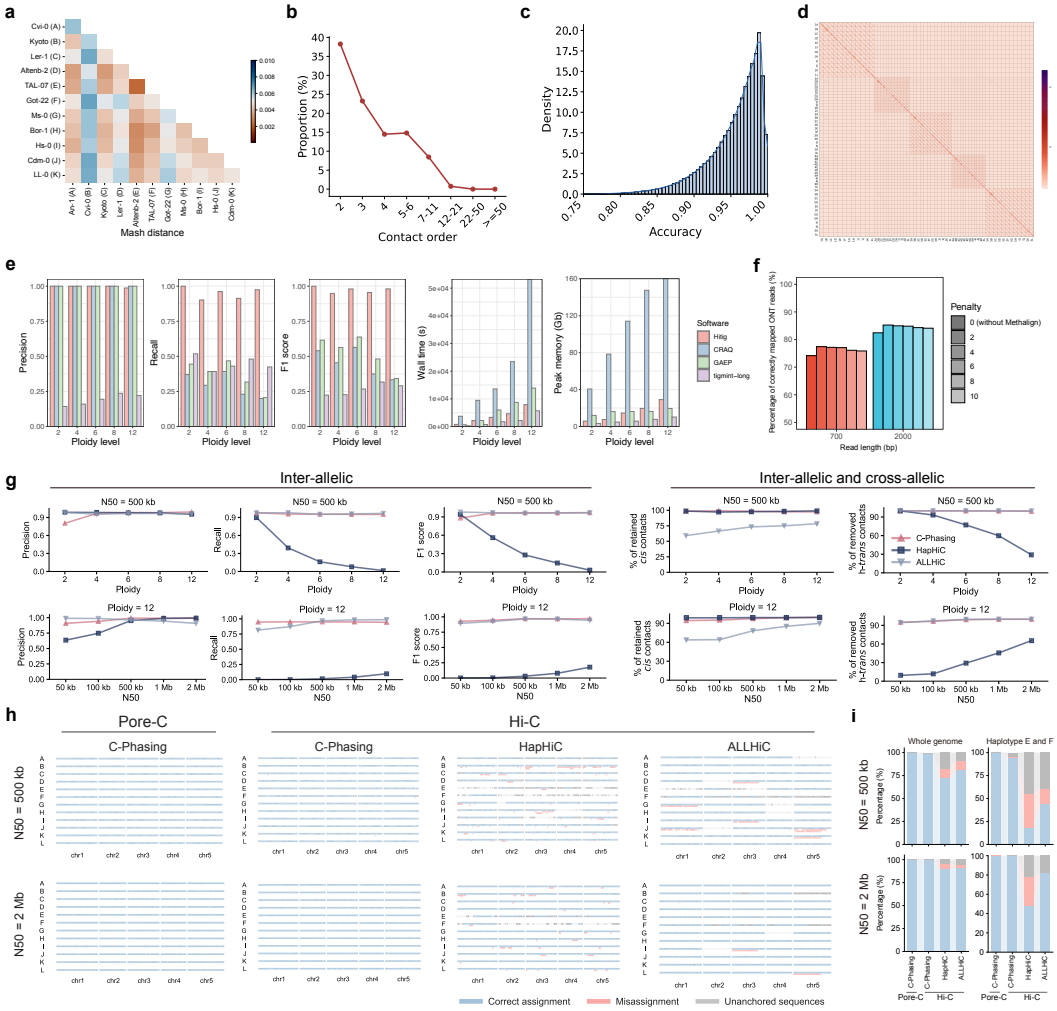


Fig. 3 | Performance evaluation on 'synthetic' datasets. a-d, Characterization of the simulated dodecaploid genome and corresponding Pore-C data. **a**, Pairwise Mash distances among twelve ecotypes (An-1 to LL-0); lower values indicate higher sequence similarity. **b**, Distribution of concatemer orders in simulated Pore-C reads. **c**, Alignment accuracy of Pore-C reads in the simulated dodecaploid genome. **d**, Contact heatmap (100-kb resolution) showing interactions in the simulated dodecaploid genome. **e**, Comparison of Hitig with CRAQ, GAEP, and tigmint-long across five ploidy levels (2x-12x) at N50 = 500 kb, evaluated by precision, recall, F_1 score, wall time, and peak memory usage. **f**, Performance of Methalign in correcting ONT read alignments, showing the proportion of correctly mapped reads under penalty values from 0 to 10. **g**, Evaluation of Kprune for inter-allelic pair identification and h-*trans* removal across varying ploidy and contiguity levels. Top: fixed contig N50 = 500 kb, ploidy = 2x-12x; bottom: fixed ploidy = 12x, contig N50 = 50 kb-2 Mb. Shown are precision, recall, and F_1 scores for inter-allelic detection, retention of *cis* contacts, and removal of h-*trans* contacts. Symbols: red triangles (C-Phasing Kprune), dark-blue rectangles (HapHiC-Remove), light-blue triangles (ALLHiC-Prune). **h, i**, Performance of Hyperpartition on dodecaploid assemblies with contig N50 values of 500 kb and 2 Mb. **h**, Per-chromosome distribution of correctly assigned (blue), misassigned (red), and unanchored (gray) contigs for C-Phasing (Pore-C), C-Phasing (Hi-C), HapHiC (Hi-C), and ALLHiC (Hi-C). **i**, Overall genome-wide proportions of correctly assigned, misassigned, and unanchored contigs (left), and results for haplotypes E and F (right).

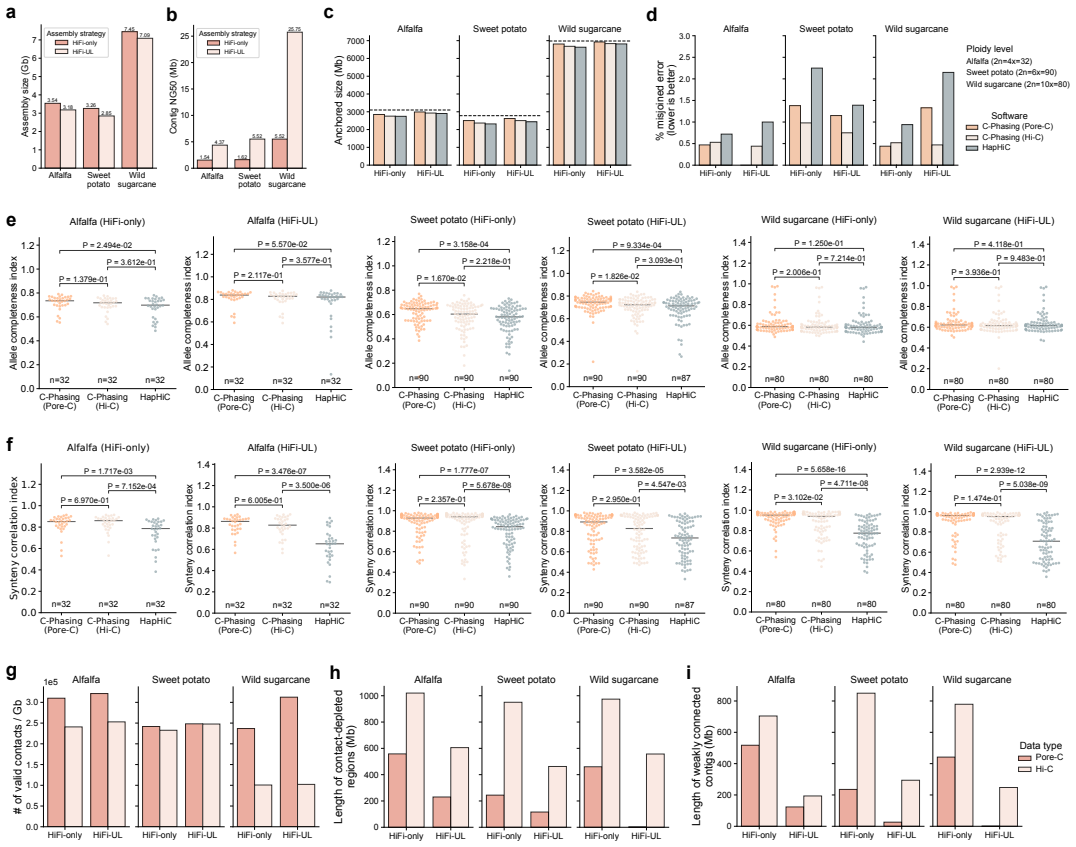


Fig. 4 | Evaluation of phasing and scaffolding performance on real-world datasets. a,b, Assembly contiguity metrics: total assembled length (**a**) and contig NG50 (**b**) across different strategies. **c**, Total length of anchored contigs (i.e., contigs incorporated into chromosome-scale scaffolds) for each scaffolding method. Dashed line indicates the estimated genome size. **d**, Misjoin error rates in the resulting scaffolds; lower values indicate higher phasing and scaffolding accuracy. **e**, Chromosome-level phasing completeness, quantified using the Allele Completeness Index (ACI), where higher values reflect more complete and accurate haplotype resolution. **f**, Scaffold contiguity assessed by the Synteny Correlation Index (SCI), with higher values reflecting improved contig ordering and orientation. In **e** and **f**, each point represents a scaffold; black lines denote median values. Statistical significance was assessed using two-sided Wilcoxon rank-sum tests. **g-i**, Comparison of contact data quality between Pore-C (dark salmon) and Hi-C (light pink). **g**, Number of valid contacts (MAPQ ≥ 2) per gigabase (Gb) of input data. **h**, Cumulative length of contact-depleted regions (10-kb bins lacking any valid contacts). **i**, Total length of weakly connected contigs (defined as having fewer than 25 contacts).

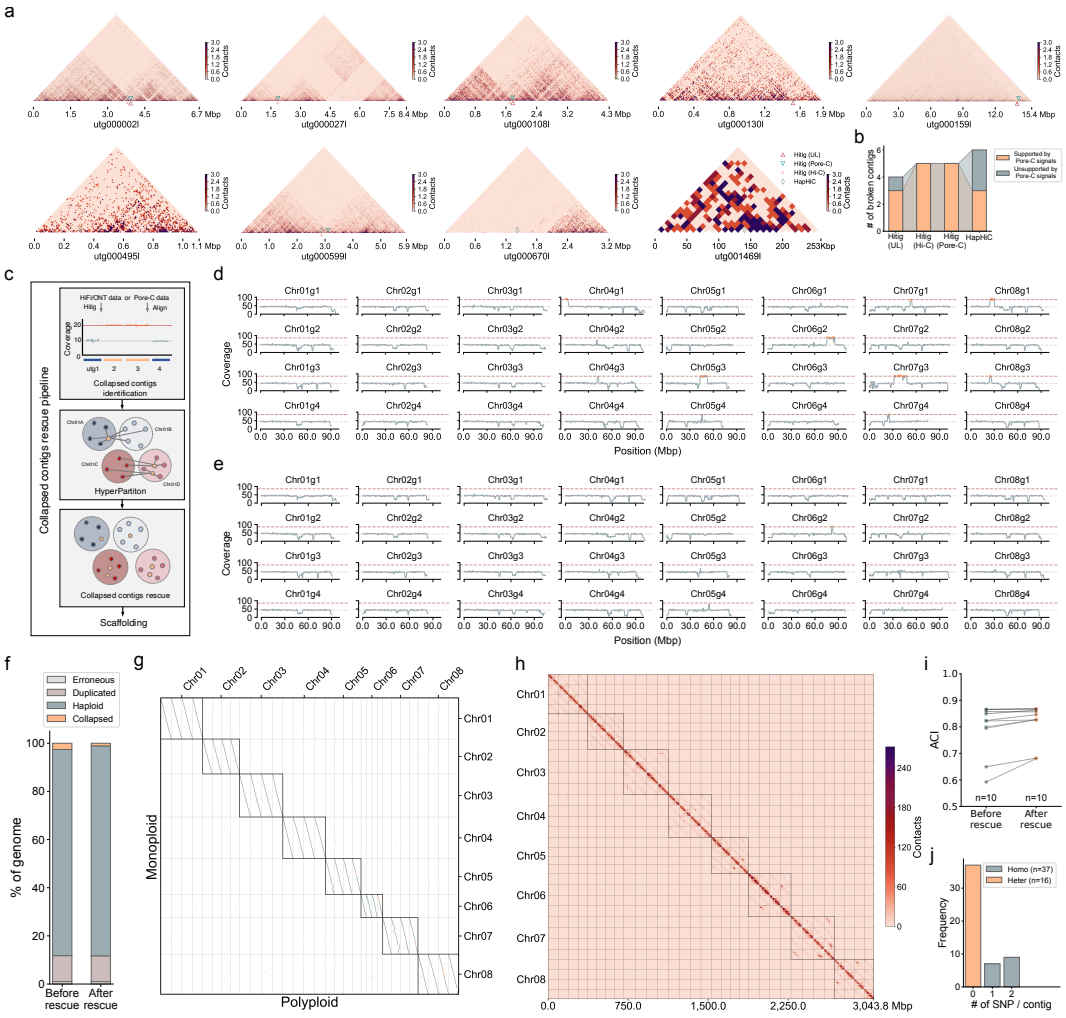


Fig. 5 | Resolving chimeric and collapsed contigs improves assembly accuracy and completeness. a, Contact heatmap (10-kb resolution) of Pore-C data highlights nine contigs in the alfalfa HiFi-UL assembly. Five contigs are chimeras supported by Pore-C signals (utg000002l, utg000027l, utg000108l, utg000159l and utg000599l). Predicted breakpoints are marked by different shapes: dark red upward triangles (Hitig, UL), dark green downward triangles (Hitig, Pore-C), light red Y-shapes (Hitig, Hi-C), and light green diamonds (HapHiC). **b**, Summary of the chimeric contigs in **a**, categorized by whether their breakpoints are supported by Pore-C contact signals. **c**, Overview of the collapsed contig rescue strategy. Collapsed contigs are first identified based on abnormal sequencing coverage, then duplicated and redistributed into other homologous groups according to their contact patterns. The updated groups are then scaffolded into pseudomolecules. **d,e**, Comparison of sequencing coverage profiles before (**d**) and after (**e**) collapsed contig rescue. Coverage was calculated from HiFi alignments in 500-kb windows. Dashed lines indicate average coverage levels: grey (normal, 43×) and red (collapsed, 86×). Windows exceeding the threshold (64.5×) are highlighted in orange. **f**, Classification of genomic bins in **d** and **e**, based on their coverage characteristics. **g,h**, Genome-wide dotplot (**g**) and Pore-C contact heatmap (**h**, 500-kb resolution) of the assembly after collapsed contig resolution, showing improved collinearity and contact density. **i**, Comparison of the Allele Completeness Index (ACI) before and after collapsed contig rescue. **j**, Distribution of SNP counts per rescued contig, reflecting their allelic divergence.

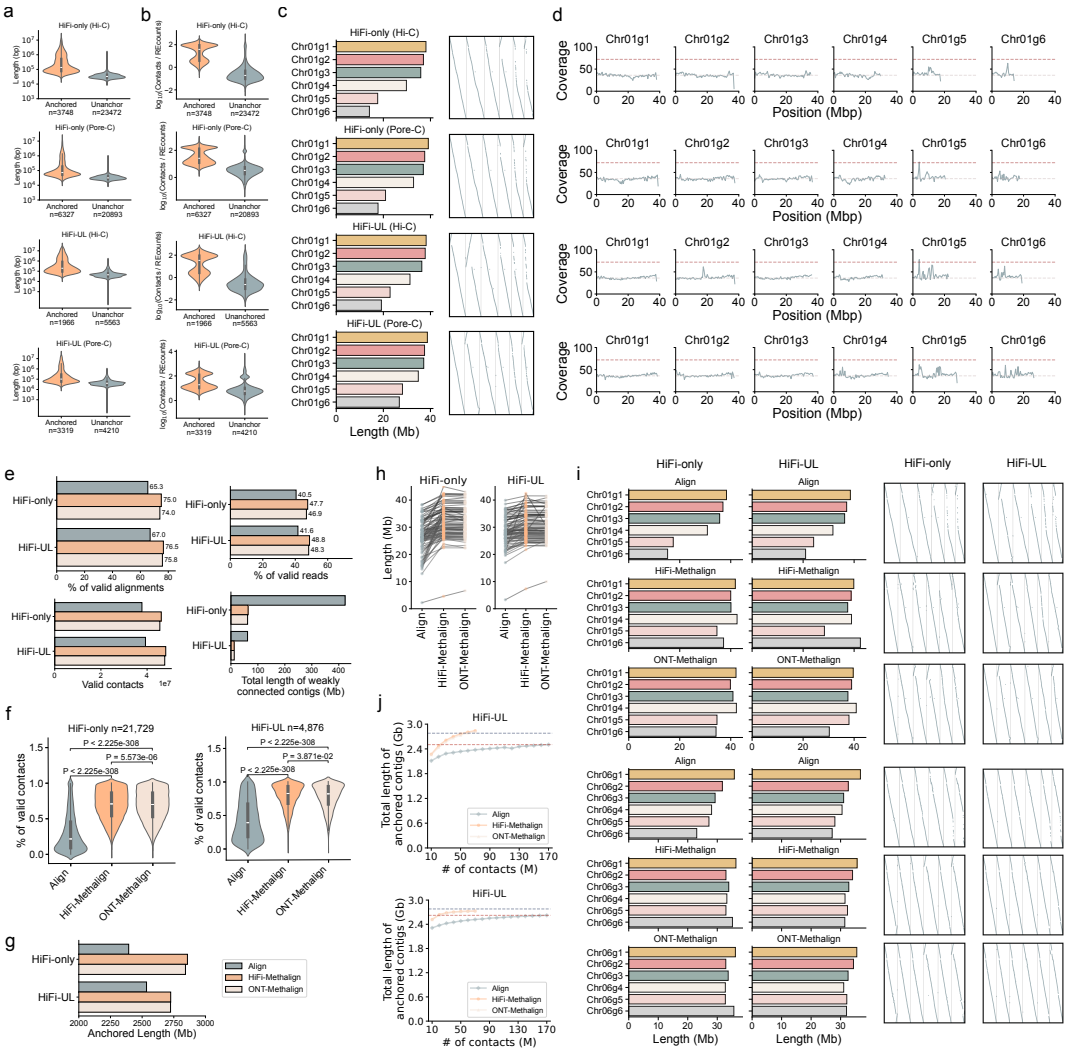


Fig. 6 | Methalign improves phasing in ultra-complex polyploid genomes. **a**, Length distribution of contigs categorized as anchored or unanchored. **b**, Violin plots of normalized contact density between anchored and unanchored contigs. **c**, Left: Scaffold length distribution per chromosome (Chr01g1-Chr01g6). Right: Dot plots showing scaffold alignments against the reference genome, highlighting structural accuracy and potential misassemblies. **d**, Sequencing coverage distribution across 500-kb windows from HiFi read alignments. **e-j**, Evaluation of Align versus Methalign modes, highlighting the impact of methylation-aware Pore-C alignment on phasing and scaffolding. 'Align' indicates direct mapping without methylation consideration. 'Methalign' incorporates 5mC data from assemblies annotated by HiFi (HiFi-Methalign) or ONT UL (ONT-Methalign) reads. **e**, Summary metrics: (top left) percentage of valid alignments (MAPQ $\geq$ 2); (top right) percentage of valid reads ($\geq$ 2 valid alignments); (bottom left) total number of valid contacts; (bottom right) total length of weakly connected contigs. **f**, Proportion of valid contacts (MAPQ $\geq$ 2) among all detected contacts; statistical significance determined by two-sided Wilcoxon rank-sum test. **g**, Total length of anchored contigs across methods. **h**, Scaffold length distribution across strategies. **i**, Left: Scaffold length distribution for homologous groups Chr01 and Chr06. Right: Corresponding dot plots showing alignments to the reference. **j**, Anchored contig length as a function of contact count. Red dashed line: maximum value from Align; blue dashed line: estimated sweet potato genome size (2.78 Gb).

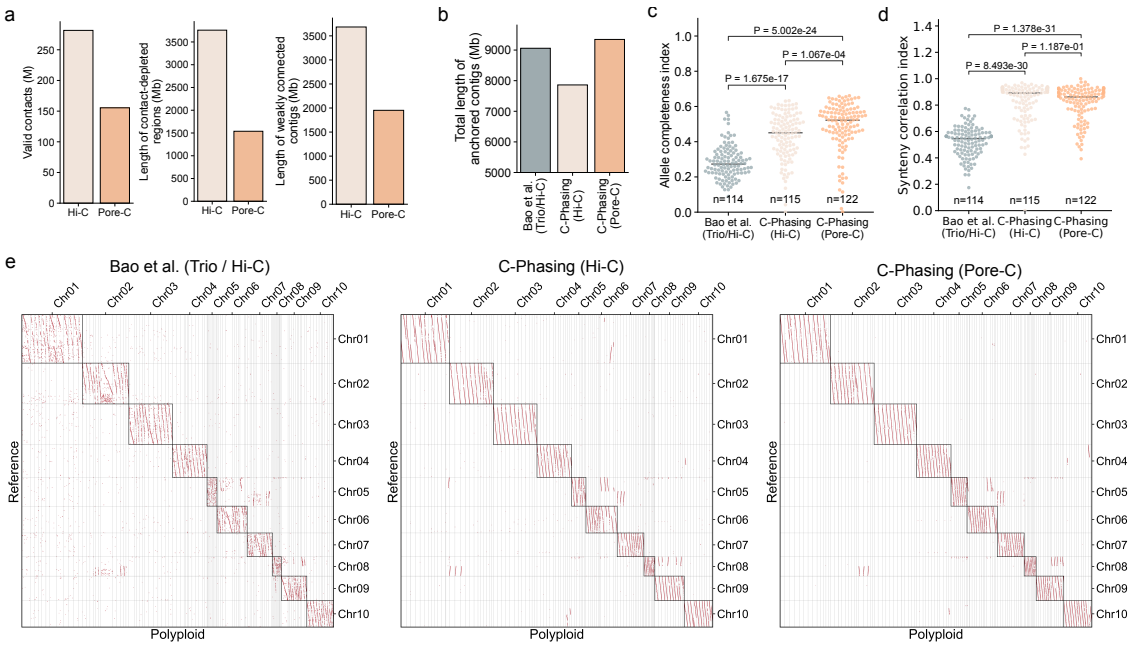


Fig. 7 | Improving chromosome-scale assembly of the modern cultivated sugarcane genome using C-Phasing. **a**, Comparative analysis of valid contacts (left), total length of contact-completed contigs (middle), and total length of weakly connected contigs (right) between Hi-C and Pore-C for the modern cultivated sugarcane genome (ZZ01). **b**, Total length of anchored contigs in the previous assembly, the C-Phasing assembly using Hi-C and the C-Phasing assembly using Pore-C. **c**, **d**, Assessment of allele completeness (**c**) and synteny correlation index (**d**); each dot represents a scaffold. The P-values were derived from two-sided Wilcoxon rank-sum tests. **e**, Whole genome synteny dot plot comparison, using the *Erianthus rufipilus* as the reference

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterials.pdf
- FigS15.pdf
- FigS17.pdf
- FigS19.pdf
- FigS24.pdf