Genome Biology

RESEARCH

Open Access

Haplotype-resolved and near telomere-to-telomere assembly of the autotetraploid potato genome

Pei-Xuan Xiao[1,2,3], Lei Tan[1,2,3], Jianke Dong[1,2,4], Jing Huang[1,2,3], Yuhong Huang[2,5], Jia-Bao He[1,2,3], Handong Su[2,5], Botao Song[1,2,4] and Wen-Biao Jiao[1,2,3,4*]

*Correspondence:
jiao@mail.hzau.edu.cn

[1] National Key Laboratory for Germplasm Innovation and Utilization of Horticultural Crops, Huazhong Agricultural University, Wuhan 430070, China
[2] Hubei Hongshan Laboratory, Wuhan 430070, China
[3] Hubei Key Laboratory of Agricultural Bioinformatics, College of Informatics, Huazhong Agricultural University, Wuhan 430070, China
[4] Key Laboratory of Potato Biology and Biotechnology, Ministry of Agriculture and Rural Affairs, Huazhong Agricultural University, Wuhan 430070, China
[5] National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China

Abstract

Background: Potato *(Solanum tuberosum)* breeding is severely hindered by its highly heterozygous autotetraploid genome, where complex allelic interactions impede precise trait selection. Reconstructing complete haplotype-resolved assemblies is crucial for genome-assisted breeding. However, current assembly methods for autopolyploids often generate fragmented sequences, haplotype-switch errors, and gaps in complex regions such as centromeres.

Results: To address these challenges, we develop PHap, a haplotype assembly pipeline tailored for autopolyploids, using only standard sequencing data, including long-reads and Hi-C. Applying PHap to the autotetraploid potato cultivar HuaShu4, we generate a haplotype-resolved, near telomere-to-telomere assembly of 3.12 Gb with an N50 of 32.7 Mb and 99.7% haplotype accuracy. Comparisons with alternative methods and existing assemblies highlight PHap's advantages in assembly quality and cost-effectiveness. Integration of transcriptomic and epigenomic data demonstrates that the genomic and methylation divergence across haplotypes drives substantial allelic expression differentiation. Time-course RNA-seq further reveals, for the first time, that 55% of genes exhibit divergent allelic expression, with dynamic shifts in dominant or suppressed alleles during tuber development. Additionally, our assembly resolves high-resolution haplotype-specific structures in centromeres and subtelomeres, as well as haplotype divergence of structural rearrangements. It also shows neocentromere formation via the expansion of megabase-scale satellite arrays.

Conclusions: These findings provide insights into the architecture of autopolyploid genomes and establish a foundation for genomics-assisted breeding of polyploid potatoes.

Keywords: Haplotype assembly, Telomere-to-telomere assembly, Potato, Allelic specific expression, Tuber development, Polyploid

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Background

Potato (*Solanum tuberosum* L.) ranks as the third largest food crop globally. It is cultivated in approximately 140 countries with an annual production of about 360 million tons [1]. Despite its significance, potato breeding has lagged behind other crops due to the complexity of highly heterozygous autotetraploid cultivar genomes [2, 3]. Although these genomes are complex, autotetraploid potatoes offer polyploid advantages including autopolyploid progressive heterosis (APH) [4, 5]. Reconstructing all haplotype sequences of autotetraploid genomes is crucial for deepening our understanding of their genome architecture [6, 7], thereby promoting genomics-assisted breeding to fully utilize these advantages [8, 9].

Recent advancements have enabled haplotype-resolved telomere-to-telomere (T2T) assembly for diploid genomes using common sequencing data such as PacBio HiFi, Oxford Nanopore Ultra-Long (ONT-UL), and chromatin conformation capture sequencing (Hi-C) data [10–12]. However, this remains challenging for autopolyploid species like potato, where a considerable amount of highly similar sequences (identity-by-descent (IBD) blocks) are shared among haplotypes due to repeated crossing of the same parents during breeding [13].

Previous methods for haplotype-resolved assembly of autotetraploid potato genomes typically begin by grouping long sequencing reads from the same haplotype to enable more accurate haplotype-level assembly. However, this grouping process relies on single-cell sequencing of gamete genomes (e.g., Otava's gamete binning) or sequencing of genetic populations (e.g., C88 (herein designated as C88(Bao)) and Altus) [13–15], both of which increase costs and time. A new version (herein designated as C88(Cheng)) of long read de novo assembler Hifiasm showed much improvements for the assembly of C88 with contig N50 37.13 Mb, however, still relying on the genetic map [16]. Other methods such as ALLHiC and HapHiC directly utilize Hi-C data to distinguish intra-haplotype and inter-haplotype interactions [17, 18], allowing haplotype-aware scaffolding of primary assembly contigs. However, they struggles with assembling cultivar potato genomes due to frequent IBD blocks [16, 17]. A latest study on European potato pangenomes released a new workflow for assembling tetraploid potatoes only using HiFi and Hi-C data while the resulting contigs had N50 values only around 5 Mb [19]. So far, there is no assembly of haplotype-resolved autotetraploid potato reaching the telomere-to-telomere level, which hinders our deep understanding the complex regions such as centromeres.

In this study, we introduce PHap, a novel tool that enables the assembly of each haplotype to near T2T resolution using only common sequencing data. Using PHap, we successfully accomplished a haplotype-resolved and near T2T genome assembly for the autotetraploid potato cultivar HuaShu4, covering nearly all difficult-to-access genomic regions. This high-quality assembly further empowers the comprehensive investigations into genome structures in both euchromatin and heterochromatin regions, as well as their impacts on gene expression and tuber development.

Results

The haplotype-resolved and near T2T genome assembly of an autotetraploid potato

To obtain a haplotype-resolved and T2T genome assembly for autotetraploid potato, we developed a new workflow, PHap, using sequencing reads of PacBio HiFi, ONT-UL, and Hi-C. Briefly, the PHap pipeline consists of seven steps (Fig. 1). First, PHap

employs the tool Hifiasm [16] to generate a primary contig assembly (*p_ctg*) and a phased unitig assembly (*p_utg*) based on HiFi and ONT-UL reads (Fig. 1A). Second, PHap performs dosage analysis to identify collapsed assembly regions due to high similarity between haplotypes by remapping sequencing reads to the *p_utg* assembly. These *p_utg* unitigs are categorized into non-collapsed (haplotigs) and collapsed unitigs (diplotigs, triplotigs, or tetraplotigs for an autotetraploid genome). Specifically, haplotigs represent only one haplotype, while others represent two, three, and four haplotypes, respectively, indicating assembly collapse. Simultaneously, PHap produces a mosaic T2T (mT2T) assembly through all-vs-all self-alignment of the highly continuous contigs *p_ctg* (Fig. 1B).

Third, PHap aligns *p_utg* to the mT2T reference to identify allelic unitigs (Fig. 1C, "Methods"). Fourth, PHap clusters allelic unitigs from the sample haplotype, guided by the strength of Hi-C signals (Fig. 1D). Unitigs within the same window are mutually assigned to different groups. In particular, haplotigs are assigned to the group with the strongest signal, diplotigs and triplotigs are assigned to the top two and three groups, respectively, and tetraplotigs are assigned to all four groups. Fifth, unitigs that failed to align to the mT2T reference or showed weak Hi-C signals during the initial clustering are re-clustered. This re-clustering is based on the intensity of Hi-C interaction with each haplotype group (Fig. 1E). Sixth, raw reads are remapped to clustered unitigs, assigned to haplotype groups (read phasing), and then independently de novo assembled using Hifiasm for each haplotype (Fig. 1F). Finally, chromosome-level

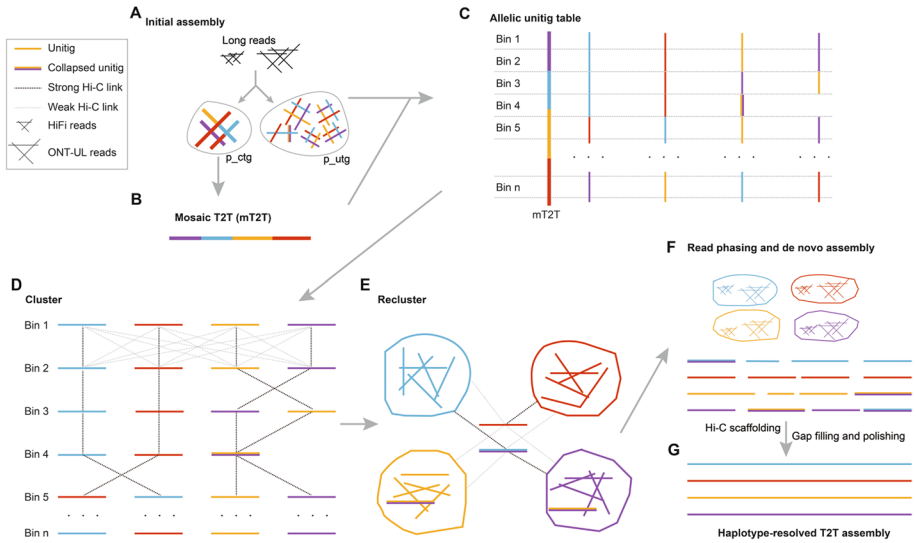

Fig. 1 Overview of the PHap pipeline. **A** Initial assembly. A primary contig assembly (*p_ctg*) and a phased unitig assembly (*p_utg*) are assembled using hifiasm with the long-read sequencing data including PacBio HiFi and ONT-UL reads. **B** mT2T assembly. A mosaic T2T (mT2T) reference is assembled based on the all-vs-all alignments from the *p_ctg* assembly contigs. **C** Allelic unitig table construction. All unitigs of the *p_utg* assembly are aligned to the mT2T reference to build an allelic table. **D** Unitig clustering. Unitigs in the allelic table are clustered according to the strength of Hi-C signals and guided by alignments against mT2T. **E** Re-clustering for remaining unitigs based on their relative intensity of Hi-C interaction with each group. **F** Read phasing and de novo assembly. Long reads are mapped to the *p_utg* unitigs, assigned into haplotypes, and de novo assembled independently for each haplotype. **G** Chromosome-scale assembly. Contigs of each haplotype assembly are scaffolded with Hi-C data, followed by gap filling and polishing with long reads

scaffolding is achieved using Hi-C data to anchor contigs, followed by gap filling and assembly polishing with long reads (Fig. 1G).

Using the PHap pipeline, we assembled the autotetraploid potato genome of cultivar 'HuaShu4' with 93.4 Gb PacBio HiFi reads, 117.2 Gb ONT-UL reads, and 257.2 Gb Hi-C reads (Additional file 2: Table S1 and Additional file 1: Fig. S1). By de novo assembling all long reads with Hifiasm, we initially obtained 742 (1.72 Gb) primary assembly contigs (*p_ctg*) with an N50 of 47.40 Mb, and 4,565 (2.58 Gb) phased assembly unitigs (*p_utg*) with an N50 of 4.46 Mb (Additional file 2: Table S2). The size of the *p_utg* accounted for only 79.66% of the estimated genome size, due to assembly collapse in highly similar allelic regions. Based on the mapping coverage of HiFi reads against the *p_utg* assembly, PHap classified the *p_utg* unitigs into haplotigs and collapsed unitigs. The collapsed unitigs included diplotigs (15.50%, 400.06 Mb), triplotigs (3.29%, 84.92 Mb), and tetraplotigs (0.22%, 5.68 Mb), corresponding to their origins from two, three, and four haplotypes, respectively (Additional file 2: Table S3). Through all-vs-all alignments of *p_ctg* contigs, a 750.32 Mb mT2T assembly was generated, encompassing 18 telomeres of 12 chromosomes (Additional file 2: Table S2). Compared to the T2T assembly of the haploid potato reference genome DM8, this mT2T assembly exhibited comparable assembly completeness (mT2T: 99.66%, DM8: 99.48%) and duplication rate (mT2T: 1.29%, DM8: 1.33%) as assessed by BUSCO analysis (Additional file 1: Fig. S2). Nearly all *p_utg* unitigs (99.01%) had homologous regions in the mT2T assembly. This mT2T assembly provided around 33 Mb more sequence space for identifying allelic unitigs among *p_utg* in the following step, compared to the reference DM8.

Subsequently, by leveraging Hi-C signals and alignments against the mT2T assembly, 2,102 unitigs were clustered into 48 groups with a total length of 2.88 Gb. Further re-clustering of 1,284 unitigs that were either small or had low Hi-C signals resulted in updated 48 groups corresponding to 48 haplotypes, with a total length of 3.05 Gb (Additional file 2: Table S2). Next, HiFi, ONT-UL, and Hi-C reads were mapped to these grouped unitigs and assigned to their respective originating haplotype(s). The assigned reads amounted to 85.7 Gb, 82.8 Gb, and 239.4 Gb, respectively (Additional file 3: Table S1 and Additional file 1: Fig. S1). After de novo assembly of these assigned long reads from each haplotype, followed by Hi-C scaffolding, gap filling, and polishing, we obtained a haplotype-resolved and chromosome-level assembly with a total length of 3.12 Gb (Fig. 2A-B and Additional file 2: Table S2). This final assembly featured a contig N50 of 32.68 Mb, which was approximately 7.69-fold, 1.74-fold, 5.31-fold, and 11.88-fold that of Otava [13], C88(Bao) [14], Eigenheimer (the one with the highest contig N50 length among the ten assemblies from ref. [19]), and Altus [15], respectively. It was also close to the assembly C88(Cheng), which has a N50 of 37.13 Mb [16]. However, it should be noted that despite its high N50 value, the assembly of C88(Cheng) incorporated additional genetic maps. It was excluded for further chromosome-level comparisons because its release version remains at contig level and lacks gene annotation. Moreover, the HuaShu4 assembly reached the T2T level with six gap-free chromosomes and a total of 174 gaps across the entire genome. In comparison, the Otava, C88, Eigenheimer, and Altus assemblies contained 5,657, 947, 3,542, and 2,909 gaps, respectively (Fig. 2A and Additional file 3: Table S2).

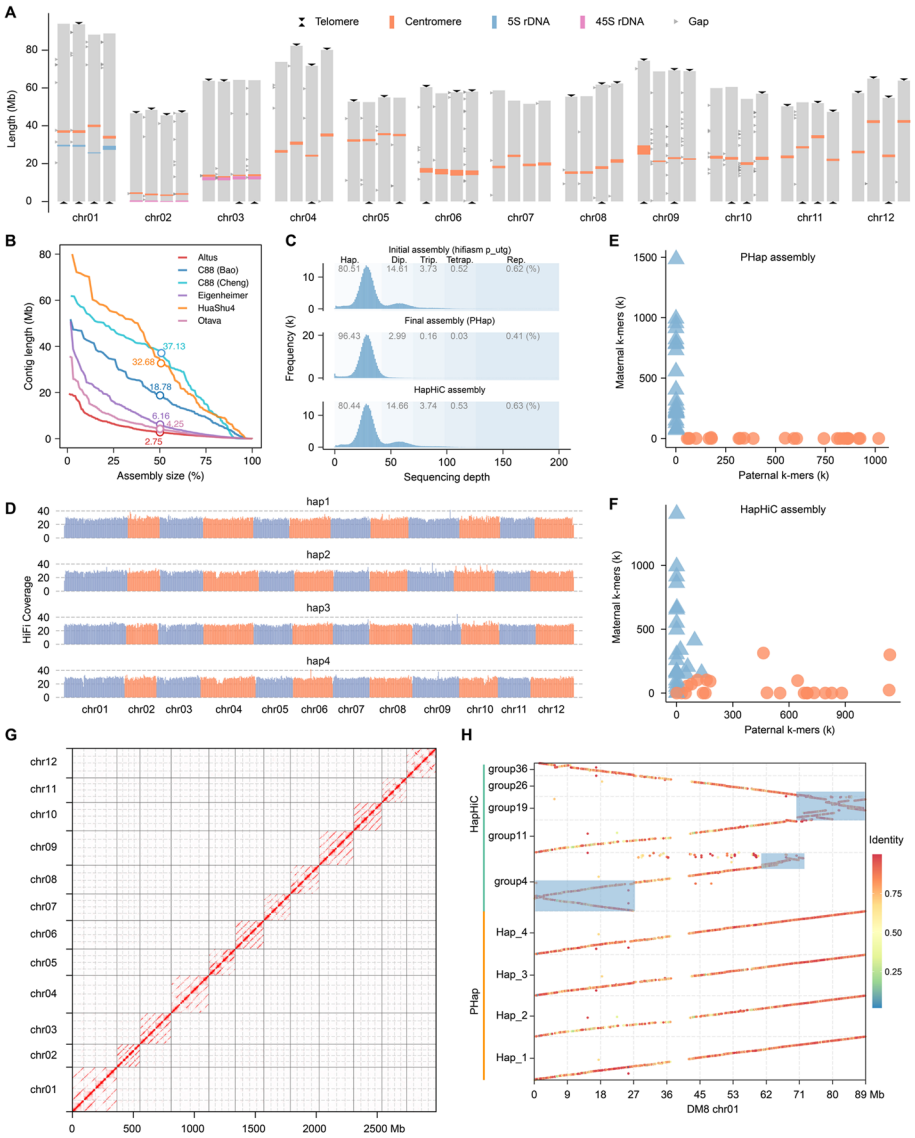

Fig. 2 Haplotype-resolved and near telomere-to-telomere assembly of the autotetraploid potato genome HuaShu4. **A** Karyotype of the autotetraploid potato genome HuaShu4. **B** Assembly contiguity comparisons of the HuaShu4, Altus [15], C88(Bao) [14], C88(Cheng) [16], Eigenheimer (the one with the highest contig N50 length among the ten assemblies from ref. [19]), and Otava [13] genomes. The x-axis represents the cumulative assembly size, calculated as the percentage of the total assembly length, with contigs arranged in descending order of length. The y-axis represents the length of individual contigs. Contig N50 (Mb) values are highlighted with open circles. **C** Dosage analysis of the initial assembly (*p_utg*) using hifiasm, the PHap assembly, and the HapHiC assembly for the HuaShu4 genome, based on the remapping of PacBio HiFi reads against the assembly. **D** Mapping coverage of four haplotypes in the HuaShu4 genome assembly by remapping PacBio HiFi reads to the final assembly. **E**, **F** *k*-mer based evaluation for the haplotype accuracy of the PHap based assembly (**E**) and the HapHiC based assembly (**F**). Each symbol represents an individual haplotype assembly of one chromosome. The x-axis and y-axis indicate the number of *k*-mers in the haplotype sequence that are specific in either of the parental genomes of Redsen or Barbara. **G** The heatmap showing Hi-C based chromatin interaction intensity. **H** Dot plot showing sequence alignments of chromosome 1 between the potato reference assembly DM8 and the PHap assembly, as well as the HapHiC assembly. Shaded areas indicate mis-assembled regions in the HapHiC assembly

We assessed the assembly quality of the HuaShu4 genome using multiple approaches. First, raw sequencing reads, including HiFi, ONT-UL, and Illumina reads (214.8 Gb), were aligned to the final assembly, achieving mapping rates of 99.91%, 99.21%, and 99.13%, respectively. Notably, dosage analysis based on this remapping indicated negligible assembly collapse across homologous haplotypes (Fig. 2C). Additionally, the remapping coverage displayed an almost uniform distribution along each chromosome, further supporting the high-quality of the HuaShu4 assembly (Fig. 2D). Besides, BUSCO analysis revealed completeness scores of 99.01%, 98.62%, 98.54%, and 98.07% for the four haplotypes, and 99.74% for the whole assembly (Additional file 1: Fig. S2). LTR Assembly Index (LAI) analysis indicated that our assembly had higher assembly completeness of LTR retrotransposons (LAI values) compared to other autotetraploid assemblies such as C88 and Otava (Additional file 2: Table S4).

Moreover, we sequenced the parental genomes of HuaShu4, Barbara and Redsen (Additional file 1: Fig. S3), generating 94.3 Gb (29.5X per haplotype) and 89.2 Gb (27.9X per haplotype) of Illumina paired-end reads, respectively, to evaluate haplotype assembly accuracy. Analysis of maternal and paternal specific *k*-mers derived from short reads revealed a haplotype accuracy rate of 99.74% (Fig. 2E, F), slightly surpassing that of the Otava genome assembly (99.6%), which required additional sequencing of single gamete genomes. Furthermore, the haplotype accuracy was corroborated by the distribution of Hi-C interaction signals (Fig. 2G). Collectively, these evaluations confirm that we have achieved a haplotype-resolved, near-complete autotetraploid potato genome assembly exhibiting higher contiguity, completeness, and accuracy.

We also compared PHap with the current state-of-the-art autopolyploid haplotype-aware scaffolding tool, HapHiC, using the sequencing data of HuaShu4. The HapHiC assembly exhibited a haplotype accuracy of only 89.91% (Fig. 2F), suggesting that HapHiC is more prone to errors in scaffolding inter-haplotype contigs. Such issues were further evidenced by the abnormal chromosome lengths resulting from inaccurate grouping of unitigs (Additional file 1: Figs. S4-S5). For example, when aligning HuaShu4 assemblies generated by HapHiC and PHap to the reference DM8, we observed that HapHiC erroneously clustered sequences from different haplotypes together (Fig. 2H). Similar errors have been previously reported in the HapHiC-based assembly of the C88 genome [17]. In contrast, PHap produced a more accurate assembly with no noticeable errors (Fig. 2H; Additional file 1: Fig. S5).

Allelic sequence and gene divergence between haplotypes

As indicated by substantial collapsed regions in the primary assembly, the HuaShu4 genome likely harbors significant inbreeding footprints. Our analysis confirmed that HuaShu4 displayed a notable level of inbreeding footprints, with 34.82% (1.04 Gb) of the genome being identical-by-descent (IBD) between haplotypes (Additional file 2: Table S5 and Additional file 1: Figs. S6-S8). This observation was supported by a bimodal distribution of SNP density between haplotypes (Additional file 1: Fig. S9), where the first peak near zero corresponded to the SNP density within IBD blocks. Compared to other tetraploid potato genomes, HuaShu4's IBD level was similar to Otava but substantially higher than that of C88 and Altus (Fig. 3A). Remarkably, 64.05 Mb of IBD blocks were shared among all four cultivars, predominantly located on chromosome 10 (Fig. 3B

and Additional file 1: Fig. S10). These shared IBD blocks are also consistent with the recurrent use of common parental germplasm such as Aquila for breeding cultivars Altus, Otava, and HuaShu4 (Additional file 1: Figs. S3 and S11). However, we cannot rule out that common IBD blocks can alternatively arise from ancient haplotype sharing

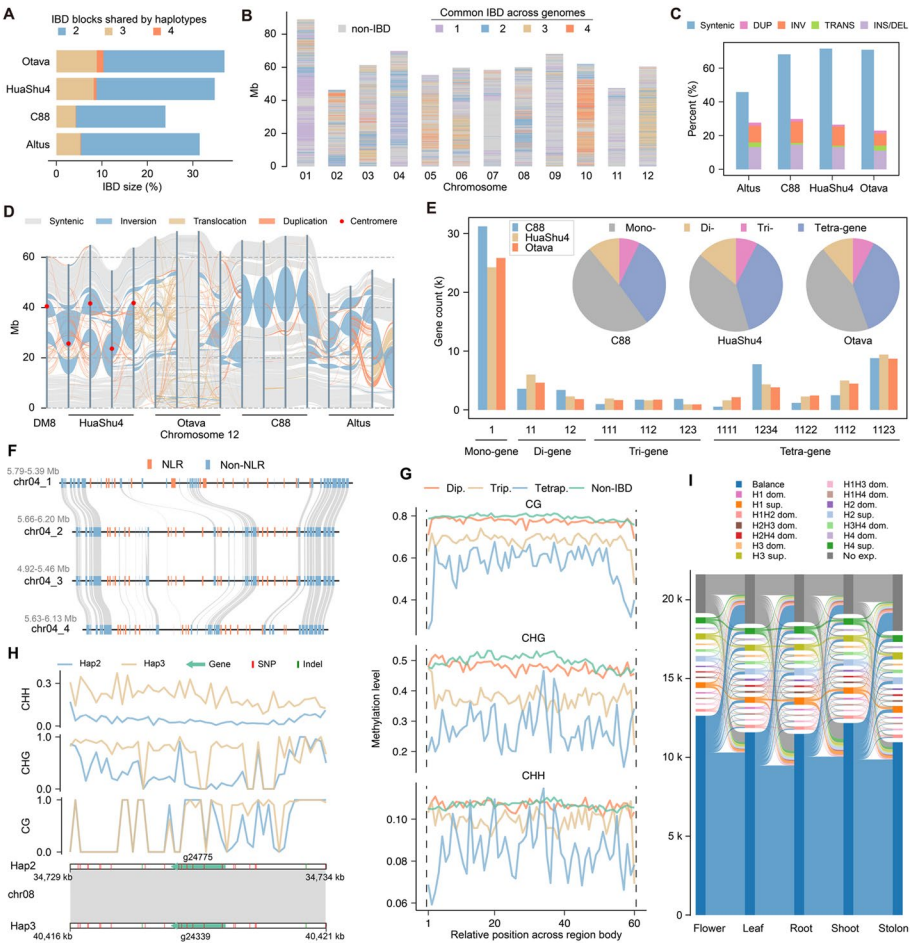

Fig. 3 Haplotype divergence of autotetraploid potato genomes at genomic, transcriptomic, and epigenomic levels. **A** The percent of different level of identity-by-descent (IBD) blocks among four autotetraploid potato genomes. Dip./Trip./Tetra. indicates IBD blocks shared by two/three/four haplotypes. **B** IBD blocks shared by different numbers of autotetraploid potato genomes, including HuaShu4, Otava, C88, and Altus. **C** The average percent of syntenic and rearranged regions in all pairwise haplotype sequence alignments for four autotetraploid potato genomes. INS: insertion, DEL: deletion, DUP: duplication, INV: inversion, TRANS: translocation. **D** Syntenic alignments among homologous haplotypes of chromosome 12 of the reference genome DM and four potato cultivar genomes. **E** Distribution of different allelic copies and genotypes of genes among autotetraploid potato genomes. The pie charts show the proportion of genes with different allelic copies. The bar charts indicate the number of gene loci with different genotypes. **F** An example showing copy number variation (CNV) of potential R genes (NLR, nucleotide-binding and leucine-rich repeat) on chromosome 4 of HuaShu4. **G** The distribution of DNA methylation level around IBD blocks shared by different number (Dip./Trip./Tetra. indicates 2/3/4) of haplotypes and non-IBD blocks. Methylation level values range from 0.0 to 1.0, where 1.0 indicates fully methylated cytosines and 0.0 indicates completely unmethylated cytosines. **H** An example for one region of chromosome 8 showing different methylation patterns at highly similar regions between two haplotypes. The top panel displays CHH, CHG, and CG methylation levels for haplotypes 2 and 3 across the region in 50 equally sized bins. The bottom panel illustrates the syntenic region (grey) between these two haplotypes. Green arrows represent annotated genes, red vertical lines indicate SNPs, and green vertical lines indicate indels (< 50 bp). **I** Sankey diagram showing dynamic switches among different types of allelic differential expression for tetra-allelic genes across five tissues

caused by historical genetic bottlenecks and reduced nucleotide diversity, as previously reported in cultivated potatoes [19, 20].

Pairwise comparisons of haplotype sequences revealed that, on average, 71.0% (529.26 Mb) of the regions were syntenic between any two homologous haplotypes, and 65.74% (491.22 Mb) of the regions showed synteny across all four haplotypes (Additional file 2: Table S6). The average synteny (defined based on Synteny Diversity [21]) among HuaShu4 haplotypes was comparable to that in Otava (68.31%) and C88 (66.71%), but significantly higher than in Altus (58.11%) (Fig. 3C). In non-IBD blocks, the average rates of allelic SNPs and indels in pairwise haplotype alignments were 17.3 and 2.8 per kb in HuaShu4, showing only minor differences compared to the other three genomes (Additional file 2: Table S7). Additionally, we identified a total of 73,033 structural variations (>50 bp), including 56,996 insertions or deletions, 777 inversions, 9,587 duplications, and 5,673 translocations between the haplotypes of HuaShu4 (Fig. 3C and Additional file 2: Table S8). Among these, 367 SVs including 261 inversions were larger than 100 kb. The largest pericentric inversion, measuring 20.06 Mb on chromosome 12, was also found in other potato genomes, either as an isolated event or in conjunction with additional rearrangements (Fig. 3D; Additional file 1: Fig. S12). Compared to diploid potato genomes, HuaShu4 had far more potentially deleterious variations, a trend also observed in other tetraploid cultivars (Additional file 3: Table S3). The vast majority (91.51%) of these deleterious variations were heterozygous in tetraploid genomes, which is consistent with previous findings that autopolyploid species exhibit increased tolerance to mutations [14, 22].

By integrating 31.09 Gb of RNA-seq data from five different tissues (Additional file 2: Table S9), we annotated 37,213, 37,385, 36,526, and 36,446 protein-coding genes for the four haplotype sets of HuaShu4, respectively (Additional file 2: Table S10). These genes formed a consensus allelic pan-gene set of 60,209 with varying allele copy numbers (Fig. 3E and Additional file 2: Table S11). Specifically, 38.23% (23,018) were present in all four haplotypes (tetra-genes), 40.28% (24,254) were haplotype-specific (mono-genes), and the remaining were present in two (di-genes) or three (tri-genes) haplotypes. Among di-gene, tri-gene, and tetra-gene loci, the proportions of genes with identical coding sequences were 72.45%, 43.00%, and 7.17%, respectively. In HuaShu4, the average copy number per gene was 3.18, while the average number of distinct alleles per gene was only 1.85, both similar to those in other tetraploid potato genomes (Additional file 2: Table S12). Apart from gene presence-absence variants (PAV) between haplotypes, we identified 374 genes present in all four haplotypes that had copy number variation (CNV) in their alleles. These genes were significantly enriched for GO terms related to defense response and the biosynthesis of secondary metabolites such as terpenoids (Additional file 1: Figs. S13-S14). Notably, 67.82% of potential resistance (R) genes showed allelic CNVs (Additional file 2: Table S13), as exemplified in Fig. 3F.

Allelic DNA methylation and gene expression differences between haplotypes

We further investigated the inter-haplotype variations at the DNA methylation level using 78.54 Gb of Illumina whole-genome BS-seq data, complemented by PacBio HiFi and ONT raw sequencing data. All types of IBD blocks showed relatively lower methylation levels compared to non-IBD blocks across CG, CHG, and CHH contexts (Fig. 3G).

In particular, IBD blocks shared by more haplotypes displayed lower methylation levels in all contexts, particularly in CG and CHG contexts within euchromatin regions (Additional file 1: Fig. S15). For non-IBD blocks, we identified allelic distinct methylation level in 1,930 blocks (865.14 Mb), with 73% of these blocks containing SVs. Remarkably, based on the methylation signals in long reads, we detected 171 Mb of regions (> 10 kb) that showed highly differential (10%) methylation levels across haplotypes, despite having highly similar (> 98% identity) allelic sequences (an example is presented in Fig. 3H).

To explore whether distinct allelic sequences and/or methylation patterns lead to allelic differential expression (ADE), we conducted expression analysis across allelic copies. As previously defined [14, 23], for tetra-genes, ADE can be further categorized into 14 types of dominant and suppressed expression patterns, depending on the relative expression levels of the four alleles (Fig. 3I and Additional file 1: Fig. S16). Among 21,020 tetra-genes with sequence variation, 9.34% (1,964) genes showed ADE in at least three different tissues. By integrating DNA methylation data, we found that the expression of 634 (32.3%) of these genes was significantly correlated to the methylation status within the gene body or in the regions 2 kb upstream or downstream of the gene. Additionally, 26.8% of genes with ADE contained indels or SVs in the promoter regions or the first introns. Since many SVs in these regions have been reported to be associated with the differential expression [24, 25], they could potentially introduce different transcriptional regulatory elements, thereby resulting in allelic expression divergence.

Allelic differential expression during tuber development

To gain deeper insights into the role of allelic divergence in important potato traits, we explored allelic gene expression throughout tuber development. We generated 159.71 Gb of RNA-seq data from 24 tuber samples collected at eight consecutive time points (S1-S8) during development (Fig. 4A and Additional file 2: Table S14). Correlation analysis of expressed genes across replicates revealed high Pearson correlation coefficient (r) values at each time point (mean $r = 0.92$, $p < 1e\text{-}10$), indicating the high-quality of the RNA-seq data (Additional file 1: Fig. S17). Principal component analysis (PCA) further supported the high consistency of biological replicates and the transcriptomic differences at different time points (Additional file 1: Fig. S18). Additionally, pseudotime analysis demonstrated that these samples could be classified into five distinct stages, corresponding to phases of tuber development: induction (S1-S2), initiation (S3), setting (S4-S5), bulking (S6-S7), and maturation (S8) (Additional file 1: Fig. S19).

Across these time points, we identified a total of 16,070 differentially expressed genes (DEGs). These DEGs were grouped into ten clusters, each representing a distinct expression pattern, and showed enrichment in specific transcription factor families and gene functions (Additional file 1: Figs. S20-S21). Notably, 42.6% of these DEGs exhibited ADE. Additionally, we identified 13,577 DEGs associated with transitions between the five stages of tuber development (Additional file 1: Figs. S22-S23), among which 5,663 (35.2%) also displayed ADE. These findings suggest that allelic genes likely display divergent expression patterns during tuber development.

This discovery motivated us to conduct a clustering analysis based on the expression of all alleles (showing differential expression across time points) across all

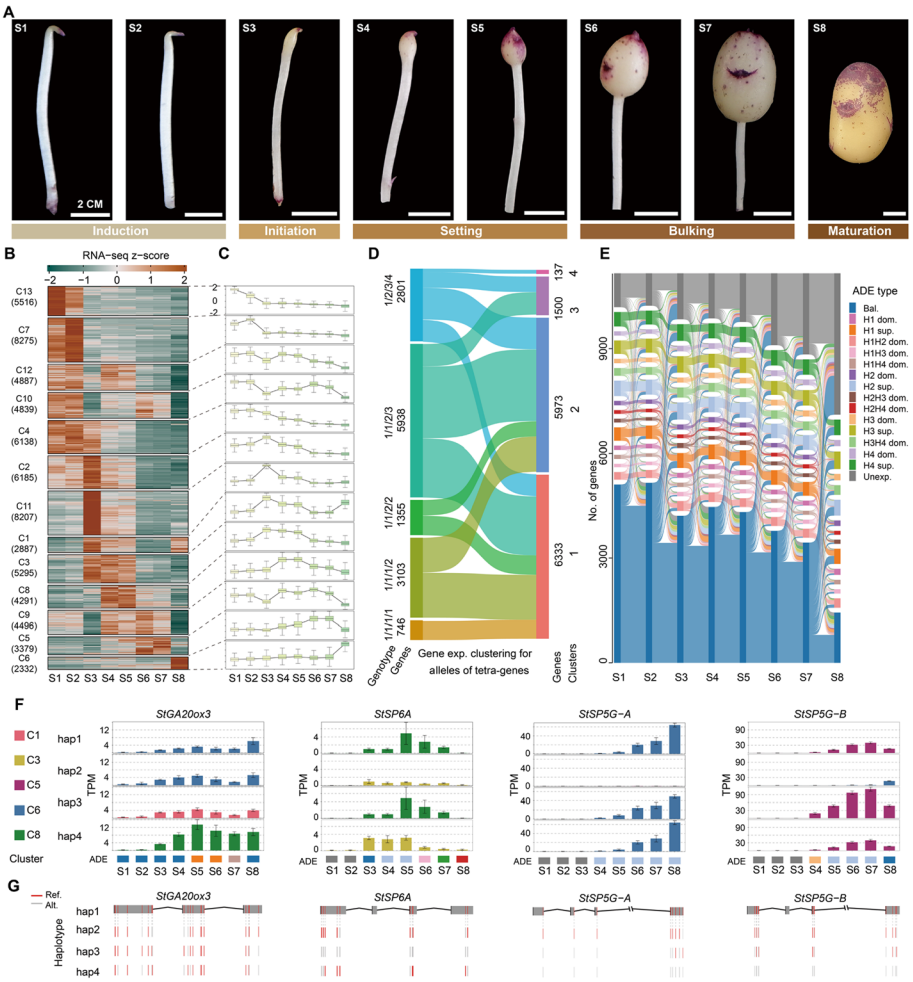

Fig. 4 Transcriptome dynamics and allelic differential expression during tuber development. **A** Morphology during various stages of potato tuber development after the transfer of plants from long-day to short-day conditions. Specifically, S1-S2 correspond to the induction stage, S3 to the initiation stage, S4-S5 to the setting stage, S6-S7 to the bulking stage, and S8 to the maturation stage. **B** K-means clustering analysis based on expression levels of all alleles showing differential expression across time points. The expression level was measured as Transcripts per Kilobase per Million mapped reads (TPM) and standardized into z-scores ranging from −2 (low expression) to 2 (high expression). The numbers in parentheses represent the number of alleles within each cluster. **C** The boxplot illustrates the expression trends of allelic genes within each cluster across the eight time points. Each box represents the median and interquartile range (IQR), with the whiskers extending to the maximum and minimum values, but not beyond Q ± 1.5 * IQR. **D** The number of clusters (corresponding to **B**) for all four alleles of 13,943 tetra-genes with different genotypes. **E** Sankey diagram showing dynamic changes among different types of allelic differential expression for 9,047 tetra-genes across eight tuber developmental stages. Note: tetra-genes that were not expressed or remained balanced across all stages were not included. **F** Examples showing the expression levels of each allele of tetra-genes during tuber development stages (S1-S8). The expression levels are shown as the mean ± standard error of three biological replicates. Colors in histograms indicate the corresponding clusters of alleles in the k-means clustering analysis of allele expression as shown in (**B**). Colors below histograms indicate the ADE types as shown in (**E**). **G** Haplotypes of allelic copies for the genes shown in (**F**). The red rectangles indicate variant positions between haplotypes

haplotypes during tuber development (Fig. 4B-C, and Additional file 1: Fig. S24). Our analysis revealed that for 7,610 (54.58%) of the 13,943 tetra-genes, not all four alleles were grouped into the same cluster. For 1,637 genes, their four alleles were distributed into three or four clusters, showing a much higher degree of expression divergence

(Fig. 4D). This further provides strong evidence that allelic genes exhibit divergent expression patterns during tuber development. Specifically, some genes showed relatively balanced expression levels across their four alleles in one stage, but dominant or suppressed alleles appeared in other stages (Fig. 4E).

For instance, the four distinct alleles of the gene *StGA20ox3* exhibited distinct expression patterns throughout tuber development, corresponding to clusters C6, C6, C1, and C8, respectively (Fig. 4D-F). Among these clusters, C6 genes were highly expressed at the maturation stage, C1 genes showed high expression at the initiation and maturation stages, and C8 genes were highly expressed at the setting stage. Additionally, this tetragene showed balanced allele expression at stages S1-S4 and S8, but ADE was observed at stages of S5-S7. Such divergent expression patterns among homologous alleles were also found in 31 other genes related to tuber development, such as *StSP5G-A*, *StSP5G-B*, and *StSP6A* (Fig. 4F and Additional file 3: Table S4). These findings together revealed that the frequent switches between allelic balanced, dominant, and suppressed expression considerably increase the transcriptomic complexity and dynamics during potato tuber development.

Haplotype divergence of centromeres

Previous cytogenetic experiments and the C88 assembly have shown that centromeres between homologous chromosomes in tetraploid potato genomes even exhibit divergence in sequence and structure [14, 26]. However, due to the constraints of the cytogenetic experiments and assembly gaps in centromeric regions, our understanding of centromere divergences remains limited. We annotated the centromeric regions based on 13.78 Gb of ChIP-seq (Chromatin Immunoprecipitation and Sequencing) data from two replicates, generated using an anti-CENH3 (centromeric histone H3) antibody. The HuaShu4 assembly successfully spanned the entirety of 41 out of the 48 centromeres without gaps. The centromeres ranged in size from 0.25–4.5 Mb, with an average size of 1.1 Mb (Fig. 5A and Additional file 3: Table S5). The relative positions of centromeres on chromosomes, defined by the length ratio of the long arm and the short arm (L/S ratio), varied substantially, ranging from 1.1 to 13.5 (Fig. 5B, and Additional file 2: Table S15). Based on the L/S ratios, most (40 out of 48) chromosomes were classified as metacentric or submetacentric. The sequence similarity between centromeres of different chromosomes was much lower than that between homologous chromosomes (Additional file 1: Fig. S25), implying the existence of chromosome-specific centromeric repeat libraries. We identified eight centromeric satellite monomers including homologs of three previously characterized repeat units (Additional file 1: Fig. S26). Two contrasting patterns of centromeric structure were found in this genome where 15 centromeres contained megabase-sized satellite arrays but not for the rest, consistent with the previous research on the potato DM genome [27]. Particularly, chromosomes 5, 6, 8, 10, and 12 lacked detectable centromeric satellite arrays in any haplotype (Additional file 3: Table S5).

Comparative analysis across haplotypes revealed remarkable variation in the centromeres of homologous chromosomes (Fig. 5C; Additional file 1: Figs. S27-31). First, centromeres differed in size across haplotypes, even when they shared the same type of satellite monomers (Fig. 5A). For example, all centromeres of chromosome 3 contained satellite arrays of HS_03_5390, yet their sizes ranged from 155 to 662 kb (Additional

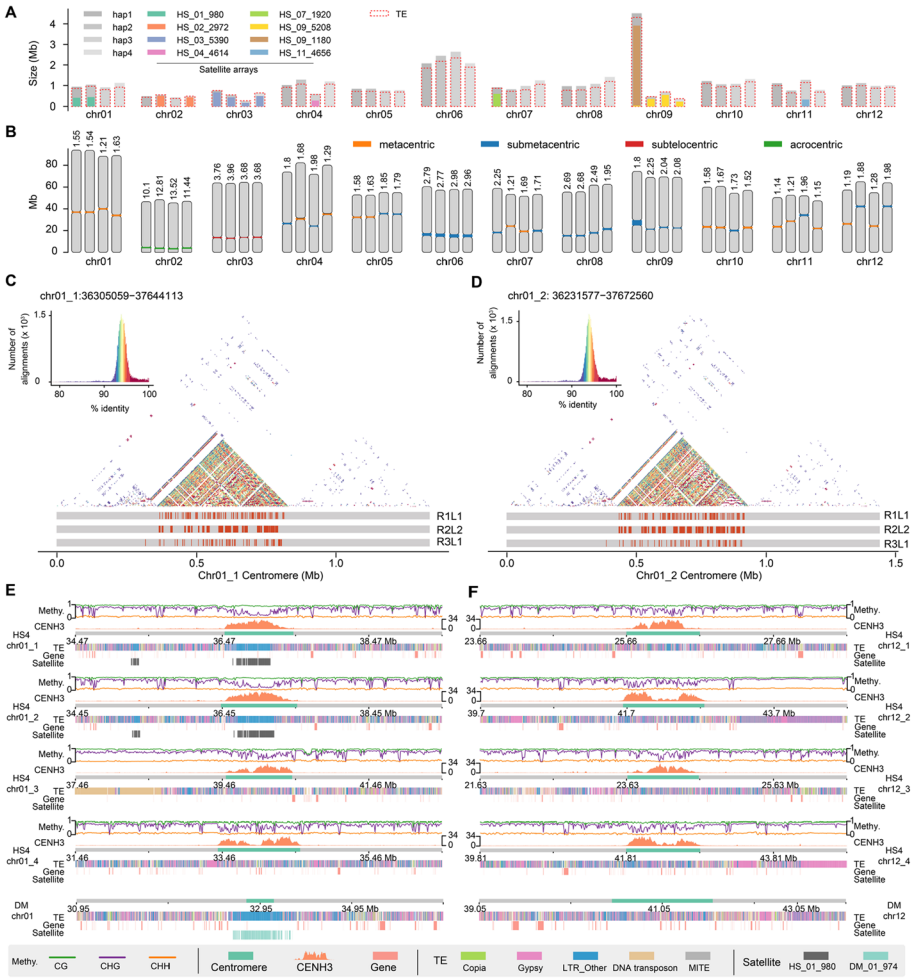

Fig. 5 Haplotype divergence of potato centromeres. **A** The size and repeat content of all 48 centromeres in the HuaShu4 genome assembly. Colored rectangles represent the content of satellite arrays with different monomer sequences. The red dashed rectangles indicate the content of TEs in centromeres. **B** Centromere positions in the assembly of the HuaShu4 genome. The number on each chromosome indicates the ratio between the long arm length and the short arm length. **C**, **D** Sequence identity and higher-order repeats (HORs) structure of centromeres, including adjacent upstream and downstream 200 kb regions, for chromosome 1 haplotype 1 (**C**) and haplotype 2 (**D**). **E**, **F** DNA methylation and sequence composition in centromeric regions of the chromosome 1 (**E**) and 12 (**F**) of HuaShu4 (four haplotypes) genome and the reference DM genome. The DNA methylation level, CENH3 occupancy, centromere position, gene density, density of different TE families, density of different monomer types of satellite arrays are indicated by different colored curves or rectangles

file 1: Fig. S28). The relative positions of centromeres could also vary between haplotypes (Fig. 5B), as exemplified in chromosome 12, which underwent a 20.06-Mb pericentric inversion (Fig. 3E). Interestingly, such centromeric inversions frequently occur in potato cultivar genomes (Additional file 1: Fig. S12).

Moreover, the structure of satellite arrays could vary between allelic centromeres, although phylogenetic clustering of satellite monomers indicated some degree of sequence exchange between haplotypes (Additional file 1: Fig. S32). For chromosomes 1, 2, 4, 7, and 11 (Fig. 5A and C), only some of the four haplotypes contained satellite arrays, as supported by previous FISH experiments [26]. The centromere of chromosome

9 haplotype-1 (Cen9-h1) harbored two types of satellite arrays, HS4_09_1920 and HS4_09_5208, whereas the other haplotypes only had HS4_09_5208 (Additional file 1: Fig. S30). In particular, nearly all (85.8%) of the Cen9-h1 sequences were composed of the HS4_09_1920 satellites. Conversely, satellite-enriched centromeres across haplotypes displayed similar patterns at level of higher-order repeat (HOR) structures (Fig. 5C, D and Additional file 1: Figs. S33-S38).

Furthermore, both homologous centromeres with and without satellite arrays displayed divergence in TE composition (Fig. 5A, Additional file 1: Fig. S39, and Additional file 3: Table S5). The proportions of the two most enriched TE subfamilies, Tekay and Athila, varied across haplotypes. For example, Tekay was the most abundant in three haplotypes of Cent1, while Athila dominated the remaining one. Additionally, the centromeric-specific TE subfamily CRM (Additional file 1: Fig. S40) showed significant variation among homologous centromeres. In chromosome 2, CRM occupied approximately 50% of TEs in one haplotype but less than 25% in the other three haplotypes. Across the 48 centromeric regions, we predicted 407 protein-coding genes, 148 of which had homologs in the reference DM genome (Additional file 1: Fig. S41). Notably, 82.3% of these genes displayed PAV across haplotypes, suggesting that the gene content in centromeric regions might be highly dynamic and subject to frequent changes.

Similar to other plant genomes such as *A. thaliana* [28], CHG methylation was relatively lower in regions with higher CENH3 occupancy (Fig. 5E, F; Additional file 1: Figs. S28-31). In addition, distinct CENH3 patterns were observed across haplotypes. For example, all Cent1 and Cent12 regions were approximately 1.0 Mb, but they exhibited distinct CENH3 peak patterns, with one or two major peaks (Fig. 5F). Taken together, our haplotype-resolved near T2T assembly provides comprehensive insights into the highly divergent haplotypes within centromeric regions of an autopolyploid genome.

Haplotype divergence of subtelomeres

Our HuaShu4 assembly covered 49 telomeres, with lengths ranging from 0.5 to 52.4 kb and an average length of 8.4 kb (Additional file 3: Table S6). A total of 87 chromosome terminals featured typical subtelomeric satellite repeat units (Fig. 6A and Additional file 1: Figs. S42-S47). The majority (63.04%) of these subtelomeres were situated in close proximity to the telomeres, with distances less than 500 bp. The length of subtelomeres varied significantly, ranging from 4.84 to 1,544.57 kb, with an average length of 391.86 kb (Additional file 3: Table S7). We identified three subtypes of repeat monomers that were homologous to the previously known potato subtelomeric satellites CL14 (subtel_rep1) and CL34 (subtel_rep2) [29] (Fig. 6B). In addition, considerable sequence variations were observed among the monomers of the satellite repeat arrays (Fig. 6C). In HuaShu4, 51 chromosome ends contained both CL14 and CL34 homologs, whereas 36 chromosome ends harbored only one type of these repeats. The subtelomeric regions displayed relatively stable and lower GC content compared to other genomic regions. Notably, the CG, CHG, and CHH methylation levels within the subtelomeric regions were all considerably higher than those in other regions (Fig. 6D).

Notably, the allelic subtelomeres in HuaShu4 demonstrated extensive variations in sequence composition and structure (Fig. 6E, Additional file 3: Table S7, and Additional

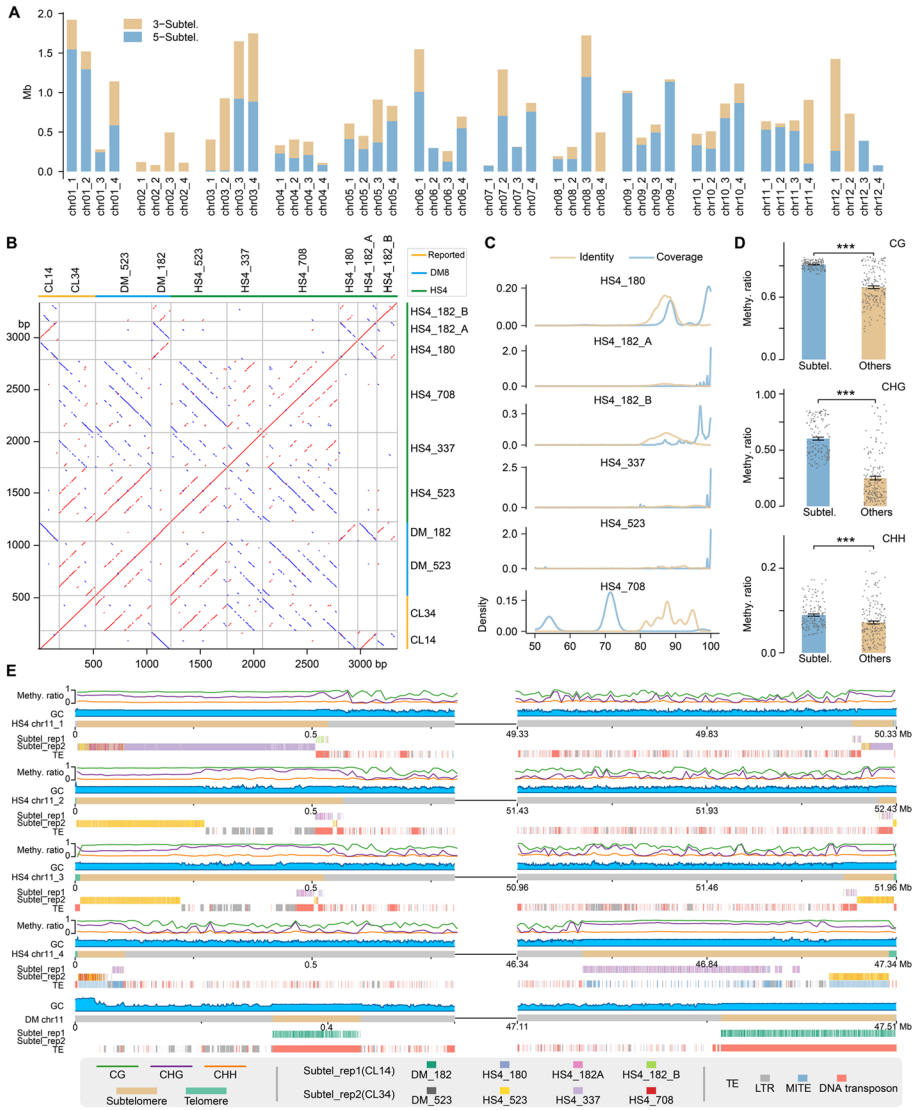

Fig. 6 Haplotype divergence of potato subtelomeres. **A** Length of subtelomeres across 48 haplotypes of the HuaShu4 genome. **B** Dot plot showing sequencing alignment of subtelomeric satellite repeat monomers from different potato genomes. Red: forward alignment, blue: reverse alignment, search window size: 10 bp. **C** The alignment identity and coverage of different satellite repeat monomers in the HuaShu4 genome. **D** Comparison of DNA methylation levels between subtelomeric regions and other genomic regions. *** indicates $p < 0.001$. **E** DNA methylation and sequence composition in subtelomeric regions of chromosome 11 of the HuaShu4 genome and the reference DM genome. The DNA methylation level, GC content, subtelomeres, telomeres, density of different TE families, density of different monomer types of satellite arrays (CL14 and CL34) are indicated by different colored curves or rectangles

file 1: Figs. S42-S47). For example, the homologous subtelomeres of chromosome 11 displayed a wide range of sizes (45.84–807.71 kb). Nearly all subtelomeres in chromosome 11 were enriched with relatively more satellites of CL34 homologs, with the exception of the 3' chromosome ends of haplotypes 2 and 4, where CL14 homologs were more predominant (Fig. 6E). TE and nonrepetitive sequences were interspersed between the CL14 and CL34 satellite arrays, as exemplified by the 5' chromosome ends of haplotypes 2 and 3 of chromosome 11. Moreover, the TE percentages (7.18%−33.56%) and their

compositions differed significantly across the subtelomeres of chromosome 11. Long terminal repeats (LTRs) were uniquely present in haplotypes 2 and 3, while miniature inverted-repeat transposable elements (MITEs) were exclusive to haplotype 4 (Fig. 6E). Additionally, the distance between subtelomeres and telomeres varied greatly among different haplotypes, as seen in chromosome 12 (Additional file 1: Fig. S47). Overall, our assembly demonstrated the noteworthy structural diversity and complexity of the subtelomeric regions, even among homologous chromosomes.

Discussion

Reconstructing complete T2T sequences is crucial for advancing genomics-assisted breeding efforts [30]. Owing to remarkable advancements in sequencing technologies [31], T2T assemblies have been reported for many crops, such as rice, maize, and diploid potato [11, 12, 32, 33]. However, fully assembling each haplotype in autopolyploid organisms, like autotetraploid potato, remains a formidable challenge. In this study, we developed PHap, a novel pipeline tailored for achieving haplotype-resolved T2T assembly of autopolyploid potato genomes. PHap effectively integrates the complementary strengths of commonly available sequencing data types, thereby avoiding the need for additional technically challenging or labor-intensive sequencing procedures. Specifically, PHap integrates three key modules: the construction of a mosaic telomere-to-telomere (mT2T) reference, Hi-C based clustering allelic unitigs following dosage analysis, and the phasing and re-assembly of haplotype-specific reads. The use of an unphased assembly (mT2T) as a reference for phasing has demonstrated advantages in haplotype assembly of diploid genomes [34, 35]. Compared to other reference genomes such as DM8, our mT2T reference can provide more sequence space for identifying allelic unitigs. Additionally, dosage analysis can detect collapsed unitigs and recover their original copies, thereby improving the efficiency of Hi-C based allelic unitig clustering.

Notably, assemblies generated by PHap achieved high contiguity, completeness, and accuracy. With a haplotype accuracy of 99.7%, it outperforms, albeit marginally, the accuracy achieved via gamete binning [13]. This high-quality assembly also provides a valuable resource for benchmarking new assembly methods and phasing populations with only short sequencing reads [19]. Since no specific features of potato genomes have been utilized in the PHap pipeline, this pipeline should also be applicable to other autotetraploid genomes. Theoretically, with adjustments to allelic unitig identification, clustering of unitigs from the same chromosome, and potentially integration of long-read sequencing-based chromatin interaction data (e.g., Pore-C) [36], our approach should be extendable to other more complex polyploid genomes as well.

Compared to diploid potato genomes, tetraploid ones feature significantly higher gene content, gene redundancy, and allelic diversity [37]. Collectively, these characteristics may provide a robust genomic foundation for progressive heterosis in autopolyploids [4]. The HuaShu4 assembly enabled us to comprehensively dissect these features at the genome, methylome, and transcriptome levels. Intriguingly, we detected substantial allelic differential expression across various tissues and during tuber development. This dynamic allelic expression frequently varied across different developmental stages, augmenting transcriptomic complexity and plasticity. Such variability may endow polyploids with advantages, enhancing their adaptability and functional diversity [23].

Nevertheless, the substantial presence of heterozygous deleterious variants and SVs, combined with tetrasomic inheritance and inbreeding depression, poses additional obstacles to obtaining homozygous tetraploid lines through selfing for breeding purposes [22, 38]. In contrast, historic breeding practices, which involved frequent crosses with common ancestors, have introduced significant inbreeding footprints in the genomes of cultivated potatoes. Although these practices may have effectively stabilized desirable traits, it remains unclear to what extent they reduce the benefits of heterosis or improve the dosage of dominant alleles.

Beyond euchromatin regions, we conducted the comprehensive exploration of genomic 'dark matter' regions in an autopolyploid genome. Our assembly uncovered new types or subtypes of satellite repeats in centromeres and subtelomeres. Remarkably, sequences in these regions demonstrated substantial allelic variations in size, sequence composition, and structure. Such diversity may indicate the inherent plasticity and potentially dynamic evolution of potato centromeres and subtelomeres [26]. Notably, this could also partly stem from extensive introgressions from wild species during domestication and modern breeding. Additionally, it remains uncertain whether these centromeric dynamics between homologous chromosomes impact chromosome pairing during meiosis and long-term genomic stability. In the future, more phased T2T assemblies from other potato cultivars, wild potato species, and their phylogenetic relatives should provide deeper insights into the evolutionary mechanisms of potato centromeres.

Although genomics-assisted breeding has proven effective in accelerating the improvement of many crops including the diploid potato breeding [39–41], polyploid breeding strategies offer unique advantages [4, 5]. Further implementation of our approach will facilitate population-level diversity analysis of polyploid potato and other complex polyploid crops. Ultimately, this will contribute to the genome design of optimal haplotypes for polyploid crops [9], enhancing their agronomic potential and resilience.

Conclusions

We have developed PHap, a haplotype assembly pipeline tailored for autopolyploid genomes using only standard long-read and Hi-C sequencing data. Applying it to the autotetraploid potato cultivar HuaShu4 yielded a high-quality, haplotype-resolved near-T2T assembly, outperforming alternative methods and existing assemblies in integrity and cost-effectiveness. Integrative analyses based on this assembly revealed extensive haplotype divergence in genome structure, DNA methylation, and allelic expression. Notably, time-course RNA-seq first uncovered divergent allelic expression with dynamic shifts in allelic dominance during tuber development. Furthermore, it enabled high-resolution characterization of centromeres, subtelomeres, and neocentromere formation driven by satellite array expansion. Collectively, this work establishes PHap as a robust tool for autopolyploid assembly, offers insights into autopolyploid genome architecture, and lays a foundation for polyploid potato genomics-assisted breeding.

Methods

Plant materials and growth conditions

The tetraploid cultivar HuaShu4 (*Solanum tuberosum*) was obtained from the Germplasm Resource Bank of the Key Laboratory of Potato Biology and Biotechnology, Ministry of Agriculture and Rural Affairs, Huazhong Agricultural University (Wuhan, China). Plants were grown according to the method described by Jing et al. [42]. For in vitro propagation, single-node cuttings were cultured on Murashige and Skoog (MS) medium supplemented with 4% w/v sucrose at 20 ℃ under long-day conditions (16 h light/8 h dark). After three weeks, the in vitro grown plants were transplanted into 10 cm × 10 cm pots (one plant per pot) and grown for an additional four weeks under the same long-day condition (16 h light/8 h dark) in a controlled growth chamber maintained at 20 ℃. Each experiment included 12 pots, with three biological replicates.

Whole genome DNA sequencing and Hi-C sequencing

Leaf samples from four-week-old in vitro grown HuaShu4 plants were collected and immediately frozen in liquid nitrogen. DNA libraries were prepared and sequenced at GrandOmics company (Wuhan, China). Specifically, for PacBio HiFi sequencing, high-quality genomic DNA was extracted using the QIAGEN Genomic-tip kit. The extracted DNA was randomly sheared using g-TUBEs (Covaris, USA). Library preparation was performed using the SMRTbell Prep Kit 3.0. The resulting libraries were combined with enzyme complexes and loaded into the PacBio Revio platform for sequencing. For ONT Ultra-Long sequencing, high-quality genomic DNA was extracted using the BAC-long Genomic DNA Extraction Kit (catalog number: XJZZ-BDE-003). Library preparation was conducted using the SQK-LSK114 kit. The libraries were then loaded into sequencing chips and sequenced on the Nanopore PromethION P48 platform.

For whole genome short read sequencings, libraries were constructed using MGIEasy Universal DNA Library Prep Kit V1.0 (CAT#1,000,005,250, MGI) according to the standard protocol. Briefly, 1μg genomic DNA was randomly fragmented using Covaris. Fragmented DNA was size-selected to an average size of 200–400 bp using MGIEasy DNA Clean beads (CAT#1,000,005,279, MGI). The selected fragments underwent end-repair, 3'adenylation, and adapter ligation. The DNA samples were amplified by PCR, and the products were purified using MGIEasy DNA Clean beads. The double-stranded PCR products were heat-denatured and circularized using the splint oligo sequence provided in MGIEasy Circularization Module (CAT#1,000,005,260, MGI). The single-strand circle DNA (ssCir DNA) formed the final library, which was qualified by QC. Qualified libraries were sequenced on DNBSEQ-T7RS platform.

For Hi-C sequencing, we constructed the Hi-C library and performed sequencing using the MGI-2000 platform. About 2 g of plant material was cut into 1 to 2 mm strips. The strips were fixed with 2% final concentration fresh formaldehyde in NIB buffer (20 mM HEPES, pH 8.0, 250 mM sucrose, 1 mM MgCl2,5mM KCl, 40% (v/v) glycerol, 0.25% (v/v) Triton X-100, 0.1 mM PMSF, and 0.1% (v/v) β-mercaptoethanol) at 4 °C for 45 min under vacuum. Formaldehyde fixation was quenched by adding glycine to a final concentration of 0.375 M under vacuum infiltration for an additional 5 min. The samples were washed twice in ice-cold water, then frozen in liquid nitrogen and ground to a powder. The powdered samples were resuspended in the NIB buffer and filtered through one

layer of Miracloth. The isolated nuclei were lysed with 0.1% (w/v) final concentration SDS at 65 °C for 10 min. SDS molecules were added using Triton X-100 at a 1% (v/v) final concentration. The DNA in the nuclei was digested by adding 200U DpnII (NEB) and incubating the samples at 37 °C for 2 h. Restriction fragment ends were labeled with biotinylated cytosine nucleotides using Biotin-14-dATP (TriLINK). Blunt-end ligation was performed at 16 °C overnight in the presence of 50 Weiss units of T4 DNA ligase. Cross-linking was reversed by 200 μg/mL proteinase K (Thermo) at 65 °C overnight. Purified DNA was extracted using the QIAamp DNA Mini Kit (Qiagen) according to the manufacturer's instructions. The purified DNA was sheared to a length of ~ 400 bp. Finally, the Hi-C libraries were quantified and sequenced using the MGI-2000 platform.

Transcriptome sequencing

For RNA-seq, samples were collected from five different tissues of four-week-old plants, including roots, shoots, leaves, followers, and stolons. Additionally, to study the transcriptome during tuber development, four-week-old plants grown under long-day conditions (16 h light/8 h dark, 20 ℃) were transferred to a short-day growth chamber (8 h light/16 h dark, 20 ℃). Samples were collected at eight continuous developmental timepoints (S1-S8) of tuber formation, defined from the hook stage of stolon apices to mature tuber development. Each developmental stage includes three biological replicates.

The mRNA-seq library was constructed as follow. Firstly, mRNA was enriched using magnetic beads with Oligo(dT). The enriched mRNA was randomly fragmented using Fragmentation Buffer. Secondly, first-strand cDNA was synthesized using the fragmented mRNAs as templates with random hexamer primers. Second-strand cDNA was synthesized by adding buffer, dNTPs, RNase H, and DNA polymerase I. The resulting double-stranded cDNA was purified using AMPure XP beads. Next, the purified double-stranded cDNA underwent end repair, A-tailing, and adapter ligation, followed by size selection with AMPure XP beads. Then, the cDNA library was enriched by PCR amplification. The concentration and insert size of the library were assessed using Qubit 2.0 and Agilent 2100, respectively. Effective library concentration was precisely quantified using Q-PCR to ensure library quality. Finally, qualified libraries were sequenced on Illumina NovaSeq X Plus platform with paired-end 150 bp reads.

Haplotype assembly of *S. tuberosum* cultivar HuaShu4

Initial assembly

An initial assembly was performed using hifiasm (v0.19.8) under the hybrid assembly model with all HiFi and ONT-UL reads, resulting in two assemblies: *p_ctg* and *p_utg*. To remove plastid genome sequences, we aligned both assemblies against publicly available chloroplast and mitochondrial sequences of potato (NCBI accession numbers: MH021593.1, MN104801.1, MN104802.1, and MN104801.1) using minimap2 [43] (v2.24) with the option of -ax asm20. A total of 122 sequences with an alignment coverage greater than 50% were removed using SeqKit2 [44] (v2.6.1). To further eliminate bacterial contaminant sequences, we aligned the remaining sequences against a subset of the NCBI NT nucleotide database containing bacterial sequences, using blastn with an e-value cutoff of 1e-5. Sequences with an alignment coverage over 50% and identity exceeding 80% were removed using SeqKit2. In total, 410 sequences were

filtered out through this process. After removing both plastid and bacterial contaminant sequences, the HuaShu4 assembly *p_utg* consisted of 4,564 unitigs with a total length of 2,580,736,761 bp and an N50 of 4,457,220 bp.

Dosage analysis

The dosage analysis followed the previous work [13]. All HiFi reads were aligned to the *p_utg* assembly using minimap2 with default parameter settings. Primary alignments were retained by removing supplementary alignments using SAMtools [45] with the option 'view -F 3840'. The average mapping depth for each *p_utg* unitig was calculated using PanDepth [46] (v2.25) with a 10 kb window size. Given the average HiFi mapping depth of 29X for the entire *p_utg* assembly, each window of *p_utg* unitigs was categorized into haplotigs ([0, 45X]), diplotigs ([45X, 74X]), triplotigs ([74X, 103X]), tetraplotigs ([103X, 132X]), and replotigs ($\geq$ 132X) based on the mapping depth.

Mosaic telomere-to-telomere genome assembly

The *p_ctg* assembly was utilized to generate a telomere-to-telomere assembly. Pairwise sequence distances between *p_ctg* contigs were computed using Mash [47] (v2.3) with a similarity threshold set at < 0.15 to identify similar contig pairs. Subsequently, these similar contig pairs were aligned using minimap2, producing a self-alignment of the *p_ctg* assembly. An overlapping graph was constructed according to the alignments. A greedy algorithm was employed to iteratively connect contigs by selecting the largest overlapping neighbor at each step, resulting in the assembly of longer sequences. Finally, we removed redundant sequences based on overlapping regions of adjacent contigs and obtained a mosaic telomere-to-telomere (mT2T) reference genome.

Build allelic unitig table

An allelic *p_utg* unitig table was constructed based on the alignments between *p_utg* unitigs and the mT2T assembly. This table included information on which *p_utg* unitigs were allelic within each 100 kb (default size) window of mT2T (serving as the reference). All *p_utg* unitigs were aligned to the mT2T assembly using minimap2 with the option '-cx asm5'.

Unitig clustering and reclustering

PHap employed a haplotype-aware unitig clustering strategy modified from the HapHiC pipeline. First, Hi-C reads were aligned to the *p_utg* assembly using BWA [48]. Alignments were filtered using SAMtools with the options 'view -F 3340' to remove PCR duplicates, secondary alignments, and supplementary alignments. Only paired reads that both were mapped with a mapping quality greater than 0 and an edit distance smaller than 3 were retained. Valid Hi-C links were defined as paired-end reads mapped to different unitigs. For each unitig pair, the Hi-C link density was calculated by dividing the number of valid Hi-C links by their restriction site counts. This density was defined as the Hi-C signal.

For the autotetraploid ($2n = 4x$) genome, the initial number of groups was set to 4x ($x = 12$ for potato). Each chromosome was processed sequentially based on the order defined in the allelic unitig table. For each unitig to be assigned, Hi-C signals between

the unitig and previously processed groups were calculated. Unitigs were then assigned to groups based on the strongest Hi-C signals. Specifically, haplotigs were assigned to the group with the strongest Hi-C signal, while diplotigs were assigned to the top two groups, triplotigs to the top three, and tetraplotigs to all four groups. Unitigs not included in the allelic unitig table were clustered based on Hi-C signal strength and assigned to groups using the same rules described above.

Finally, unaligned unitigs (those that could not be aligned to the mT2T assembly) and unitigs with weak Hi-C signals (with Hi-C link density less than 5) were rescued. Hi-C signals between these unitigs and all chromosome haplotypes were calculated, and these unitigs were assigned to groups following the same assignment rules as described above. In addition, unitigs assigned to a chromosome but not to a specific haplotype were denoted as "HapUn" and included into the final assembly, whereas those unassigned to both chromosome and haplotype were denoted as "ChrUn".

Read phasing and assembly

All raw sequencing data, including HiFi, ONT reads (> 100 kb), and Hi-C reads, were aligned to the *p_utg* unitigs using minimap2 for long reads and BWA for Hi-C reads. Based on the unitig clustering results, reads were assigned to haplotype-specific groups. Specifically, for reads aligned to diplotigs, triplotigs, and tetraplotigs, they were randomly assigned to one of the two, three and four haplotype groups, respectively. Reads in each haplotype group were then de novo assembled using hifiasm, followed by scaffolding with HapHiC [17] (v1.0.5). Finally, Juicebox [49] was employed to correct scaffolding errors based on the Hi-C interaction signal map.

Gap filling and genome polishing

For each chromosome, TGS-GapCloser [50] was used to fill gaps using phased ONT reads resulting from the previous step. This was followed by assembly refinement using the T2T-Automated-Polishing strategy [51]. Specifically, phased HiFi reads were aligned to the gap-filled genome using Winnowmap2 [52] with the options '-k 15 -ax map-pb -MD'. The alignments were then filtered using 'falconc bam-filter-clipped tool' with the options '-t -F 0×104' to remove secondary alignments and excessively clipped reads. Next, the assembly was polished using Racon [53] with default parameters. Finally, nucleotide-level errors were identified using Merfin [54], and the final consensus sequences were generated using BCFtools [45].

Genome assembly quality assessment

Multiple complementary approaches were utilized to evaluate the assembly quality in terms of completeness and accuracy. First, the haplotype accuracy was assessed using paired-end Illumina reads from the two parental lines of HS4. *k*-mer frequencies for the parental lines (Redsen and Barbara) were calculated using yak [55] with options 'yak count -b37 -t32'. Parental-specific *k*-mers in all unitigs of each haplotype sequence were identified using yak trioeval. The haplotype accuracy was then calculated based on the number of specific *k*-mers. Second, the LTR assembly index (LAI) was computed using LTR_retriever [56] to evaluate the completeness of repetitive sequences. Additionally, genome completeness was assessed using compleasm [57] with the options '-l

eudicots_odb10 -m busco' by searching the eudicots_odb10 database of 2,326 genes. Lastly, raw genome sequencing data, including HiFi, ONT, Illumina, and RNA-seq reads, were remapped to the genome assembly to calculate mapping rate using minimap2, Bowtie2 [58], and Hisat2 [59]. Mapping depths were calculated using PanDepth [46].

Repeat and gene annotation

A potato repeat library was constructed using EDTA [60] and then used to annotate repeat sequences of potato genomes using RepeatMasker (http://repeatmasker.org). For protein-coding gene annotation, we utilized a workflow BRAKER [61] by integrating evidence from ab initio prediction, RNA-seq based transcripts, and homologous protein alignments. Initially, AUGUSTUS [62] (v3.4.0) was used for ab initio gene prediction. Next, GeneMark-ETP [63] was employed for homology-based gene predictions using protein sequences from representative plant genomes, including *Arabidopsis thaliana*, *Oryza sativa*, *Solanum melongena*, *Solanum lycopersicum*, *Capsicum annuum*, and potato DM8. Additionally, we integrated RNA-Seq data and these protein sequences to further optimize and validate the gene models. Specifically, RNA-Seq reads were processed using fastp [64] (v0.23.4) for quality control and removal of low-quality reads. The filtered reads were aligned to the assembly using Hisat2 [59] (v2.2.1), and those with mapping quality below 30 were filtered out using SAMtools. The remaining reads were assembled into transcripts using StringTie [65] (v2.2.1). Finally, we used the BRAKER3 [61] (v3.0.8) pipeline to integrate evidence from previous steps to generate consensus gene models. Genes with coding regions overlapping more than 30% with TE regions were marked as TE-related genes.

In addition, we annotated tRNA using tRNAscan-SE [66] (v2.0.12) with default options and rRNA using RNAmmer [67] (v1.2). For lncRNA annotation, we first used FEELnc [68] (v0.2.1) to filter out transcripts shorter than 200 bp and those overlapping with mRNAs. Two deep learning-based tools, LncFinder-plant and CPAT-plant were also employed [69]. Predicted lncRNA transcripts were further aligned to the Uniprot protein database using DIAMOND [70] (v2.0.15) to remove potential mRNAs. The results from these three methods were cross-validated to obtain the final lncRNA annotation. Additionally, Infernal [71] (v1.1.5) was used to annotate other types of non-coding RNAs, such as snRNA, snoRNA, and miRNA.

Annotation of the NLR disease resistance gene family

Firstly, InterProScan [72] (v5.72–103.0) was used to search for typical domains of NLR (Nucleotide-binding Leucine-rich Repeat) resistance genes including the Toll/Interleukin-1 receptor (TIR) domain, the coiled-coil (CC) domain, and the RPW8 domain. Consequently, these three NLR subfamilies were named TIR-NLR (TNL), CC-NLR (CNL), and RPW8-NLR (RNL) [73]. Besides, NLR-Annotator [74] (v2.0) was employed to annotate candidate NLR-type R genes. Since NLR-Annotator was unable to identify the RPW8 domain, candidate R genes resulting from InterProScan that contained both RPW8 and NB-ARC domains were classified as RNL type. These candidates were then combined with the annotation of NLR-Annotator.

Identification of IBD blocks and analysis of haplotype divergence

IBD blocks on the final assembly were determined based on the alignments between *p_utg* unitigs and the final assembly using minimap2 with the parameter '-cx asm5'. The types of *p_utg* unitigs were determined through dosage analysis during the PHap assembling process. Specifically, regions aligned by diplotigs, triplotigs, and tetraplotigs were annotated as IBD blocks, representing genomic segments shared among two, three, or four homologous haplotypes. Regions aligned by haplotigs were annotated as non-IBD blocks, representing unique haplotype sequences. The IBD blocks shared by different tetraploid potato cultivars were identified by aligning their sequences to the reference genome DM8 using minimap2 with the parameter '-cx asm 5', followed by checking overlapping regions with BEDTools [75].

Pairwise alignment of haplotype sequences was performed using minimap2 with options '-ax asm5 –eqx'. Syntenic regions and sequence variations, including SNPs, indels, and SV (>=50 bp), were identified using SyRI [76] with options '-F B -k'. SnpEff [77] pipeline was used to identify deleterious mutations in both tetraploid cultivars and natural diploid potato accessions [40, 78]. The synteny diversity of an autopolyploid genome was defined as the average ratio of non-syntenic sites found within all pairwise alignments of haplotypes within the genome according to our previous study [21]. This synteny diversity was calculated using SynDiv [79].

Allelic genes were identified based on haplotype collinearity and gene homology. First, MCscan [80] was used to obtain collinear gene pairs located within the synteny blocks with the parameter 'jcvi.compara.catalog ortholog –csore=0.99'. Next, Orthofinder [81] (v2.5.4) was used to identify homologous genes across four haplotypes. Subsequently, collinear gene pairs were excluded if they belonged to different ortholog groups. For genes with multiple pairing candidates, all possible pairs were retained, and the pair with the highest collinearity score was designated as the representative. Genes with alleles present in four/three/two/one haplotypes are classified as tetra-/tri-/bi-/mono-genes.

Identification of telomeres and subtelomeres

Telomeric repeats in potato genomes were annotated using TIDK [82] (v0.2.31) with plant typical telomeric repeat unit (TTTAGGG/CCCTAAA). Potential subtelomeric satellites were identified in both ends of 1 Mb chromosomes using SRF [83] with parameter setting '-k 71 -ci 5 -cs 10,000'. Re-dot-able (https://www.bioinformatics.babraham.ac.uk/projects/redotable/) was used to perform sequence alignments of satellite repeat units. The chromosome terminal regions were visualized using karyotypeR [84].

ChIP-seq and analysis of centromeres

Approximately 1–3 g of leaf tissue was harvested and immediately frozen in liquid nitrogen. The frozen tissue was then finely ground using pre-chilled mortars and pestles. The native Chromatin Immunoprecipitation (ChIP) protocol was adapted from a previously described method that involves micrococcal nuclease digestion of the DNA [85]. Anti-rabbit CNEH3 polyclonal antibodies were raised against peptides corresponding to the shared N-terminus peptide sequence (ARTKHLAKRSRTKPSV-C) of *S. tuberosum* CENH3 proteins. The polyclonal antibodies were generated using the method previously described [86] and were provided by GL Biochem (Shanghai) Ltd. The chromatin

sample that did not undergo CENH3 enrichment was used as the Input control. Both ChIP-enriched and Input DNA samples were then used for library preparation with the Illumina ThruPLEX® DNA-Seq Kit (Takara, #R400674) to generate 150-bp paired-end reads, which were sequenced by Kindstar Sequenon (Wuhan, China).

ChIP-seq reads were aligned to the assembly using Bowtie2 [58] (v2.4.5). Duplicate reads were removed using Sambamba [87] (v1.0.1), followed by peak calling with MACS2 [88] (v2.2.6). Centromeric regions were initially identified through manual inspection based on the peak. StainedGlass [89] (v0.6) was used to assess the sequence consistency of centromeric repeats and generate heatmaps with a window size set to 2 kb. The monomers of satellite arrays within centromeric regions were identified using SRF. Pairwise alignments of monomers were performed using Re-dot-able (https://www.bioinformatics.babraham.ac.uk/projects/redotable/). The high-order repeat (HOR) structures of centromeric regions were identified using HiCAT [90] (v1.1.0). The TE composition in centromeres were analyzed based on the TE annotation using BEDTools [75]. Subclade composition of LTRs were annotated using TEsorter [91] (v1.3).

Whole genome bisulfite sequencing and DNA methylation analysis

After samples DNA testing, positive control DNAs were added. The DNA was then fragmented into 200–400 bp using Covaris S220 (Covaris, USA). Next, ssDNA fragments were bisulfite-treated. After bisulfite treatment, unmethylated cytosine were converted into uracil, while methylated cytosine remained unchanged. Subsequently, methylation sequencing adapters were ligated, followed by size selection and PCR amplification steps to generate DNA fragments using the Accel-NGS Methyl-Seq DNA Library Kit (Swift, USA, Catalog #:30,096). Library quality was assessed on the Agilent 5400 system (Agilent, USA) and quantified by qPCR (1.5 nM). Pair-end sequencing of the library was performed on Illumina platform.

Raw reads were quality-controlled using fastp with default parameters. Clean reads were then aligned to the *p_utg* assembly using Bismark [92] with options of '–hisat2 –score_min L,0,−0.6'. Based on the alignments and the haplotype that unitig were clustered into during the PHap assembling process, reads were assigned to the corresponding haplotypes. Specifically, reads aligned to collapsed unitigs (diplotig, triplotig, and tetraplotig), were randomly assigned to one of the haplotypes. This process generated 48 haplotype-specific datasets. Next, these reads were aligned to their respective haplotype sequences using Bismark with '–hisat2 –score_min L,0,−0.6'. Duplicates were removed using deduplicate_bismark, and methylation information for CG, CHG, and CHH contexts was extracted using bimark_methylation_extractor (-p –comprehensive –CX –bedGraph –parallel 10 –cytosine_report). ViewBS [93] MethGeno (–win 500,000 –step 500,000) was used to compute methylation levels in non-overlapping 500 kb windows, and ViewBS GlobalMethLev was used to calculate the global methylation level for each haplotype.

Long read-based DNA methylation analysis

For ONT-UL reads, Dorado basecaller (v0.9.0) (https://github.com/nanoporetech/dorado) was first used to perform base calling on the phased ONT reads generated during the PHap assembly process. Subsequently, Minimap2 was used to build an index

for the 48 haplotypes of the HuaShu4 assembly, and the Dorado aligner was applied to align the reads to the haplotype sequences. Methylated cytosines were extracted using Modkit (v0.4.2) (https://github.com/nanoporetech/modkit) with options '–motif CG 0 –combine-strands –ignore h, –motif CHG 0 –ignore h, and –motif CHH 0 –ignore h'. Additionally, phased HiFi reads were aligned to the HuaShu4 assembly using pbmm2 (v1.16.99) (https://github.com/PacificBiosciences/pbmm2) with options '–preset CCS –sort'. The probabilities of CG methylation site were determined using the aligned_bam_to_cpg_scores tool from pb-CpG-tools software package (v2.3.2) (https://github.com/PacificBiosciences/pb-CpG-tools), with the trained model 'pileup_calling_model.v1.tflite'. For comparisons of methylation levels in highly similar (> 98% identity) regions, a difference was considered meaningful if the absolute methylation difference exceeded 10%.

RNA-seq analysis of tuber

Raw RNA-seq data were preprocessed using fastp to remove adapters and low-quality reads. Clean reads were aligned to the HuaShu4 genome assembly via Hisat2, with gene-level read counts quantified by featureCounts [94] (v1.6.2). Normalized counts matrices were used for principal component analysis (PCA) using the *PlotPCA* function in the DESeq2 package.

Differential expression analysis during tuber development was determined through likelihood ratio testing (LRT) implemented in DESeq2 [95] (v1.44.0), with thresholds (FDR < 0.01 and |log2FoldChange| > 1). Transcripts per million (TPM)-normalized expression values were Z-score standardized. Resultant differentially expressed genes (DEGs) were clustered via k-means in R (seed = 1998, iter.max = 50). Transcription factor (TF) families were annotated using PlantTFdb [96] (https://planttfdb.gao-lab.org/). The enrichment of TF families in different gene clusters were assessed via enricher() in the R package clusterProfiler [97] (v4.12.6).

Pseudotime analysis followed established methods [98]. Genes with TPM > 1 in any development stage (S1-S8) were selected for PCA-based sample stratification. Adjacent time point distances were calculated using Euclidean metrics and linearly scaled to a 0–10 pseudotime axis. Z-scaled expression profiles were interpolated at 500 pseudotime points via locally weighted regression, followed by trajectory ordering through Atan2-transformed PCA coordinates. Visualization was executed with ComplexHeatmap [99] (v2.20.0).

Allelic differential expression analysis: Allele-specific TPM values were quantified by kallisto [100] (v0.51.1) and normalized to total gene-level TPM. Following polyploid allelic expression classification frameworks [14, 23], allelic expression patterns were categorized into 15 types (H1_dominant H1H2_dominant, H1H3_dominant, H1H4_dominant, H1_suppressed, H2_dominant, H2H3_dominant, H2H4_dominant, H2_suppressed, H3_dominant, H3H4_dominant, H3_suppressed, H4_dominant, H4_suppressed, and Balanced.) through Euclidean distance minimization between observed and theoretical expression vectors.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-026-03980-9.

Additional file 1: Supplementary Figures. Supplementary Figures S1-S47.

Additional file 2: Supplementary Tables. Supplementary Tables S1-S15.

Additional file 3: Supplementary Tables. Supplementary Tables S1-S7.

Peer review information
Wenjing She was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions
W.B.J and P.X.X conceived and designed the project. W.B.J supervised the project. P.X.X developed the genome assembly pipeline, performed data analysis, and drafted the manuscript. J.D prepared the plant materials. L.T, J.H, and J.B.H, performed data analysis. Y.H performed ChIP experiments. W.B.J revised the manuscript with contributions from B.S and H.S. All authors read and approved the final manuscript.

Funding
This project was supported by grants from the National Natural Science Foundation of China (no. 32270685), the National Natural Science Fund for Excellent Young Scientists Fund Program (Overseas), and the "Young Scientist Fostering Funds for the National Key Laboratory for Germplasm Innovation & Utilization of Horticultural Crops".

Data availability
All raw sequencing data and assembly sequences have been deposited at National Genomics Data Center with BioProject number PRJCA036064 [101]. Genome assembly and annotation files can also be accessed in Figshare [102]. Custom scripts and codes used in this study are available at GitHub (https://github.com/JiaoLab2021/PHap) under the MIT License [103] and Zenodo [104].

Declarations

Ethics approval and consent to participate
Ethics approval is not applicable.

Competing interests
The authors declare no competing interests.

Received: 28 July 2025 Accepted: 28 January 2026
Published online: 03 February 2026

References

1. Qu L, Huang X, Su X, Zhu G, Zheng L, Lin J, et al. Potato: from functional genomics to genetic improvement. Mol Hortic. 2024;4:34.
2. Jansky SH, Charkowski AO, Douches DS, Gusmini G, Richael C, Bethke PC, et al. Reinventing potato as a diploid inbred line–based crop. Crop Sci. 2016;56:1412–22.
3. Douches D, Maas D, Jastrzebski K, Chase R. Assessment of potato breeding progress in the USA over the last century. Crop Sci. 1996;36:1544–52.
4. Washburn JD, Birchler JA. Polyploids as a "model system" for the study of heterosis. Plant Reprod. 2014;27:1–5.
5. Comai L. The advantages and disadvantages of being polyploid. Nat Rev Genet. 2005;6:836–46.
6. Kong W, Wang Y, Zhang S, Yu J, Zhang X. Recent advances in assembly of complex plant genomes. Genomics Proteomics Bioinf. 2023;21:427–39.
7. Zhang Q, Qi Y, Pan H, Tang H, Wang G, Hua X, et al. Genomic insights into the recent chromosome reduction of autopolyploid sugarcane *Saccharum spontaneum*. Nat Genet. 2022;54:885–96.
8. Yuan Y, Scheben A, Edwards D, Chan T-F. Toward haplotype studies in polyploid plants to assist breeding. Mol Plant. 2021;14:1969–72.
9. Wang Y, Fuentes RR, van Rengs WM, Effgen S, Zaidan MWAM, Franzen R, et al. Harnessing clonal gametes in hybrid crops to engineer polyploid genomes. Nat Genet. 2024;56:1075–9.
10. Li H, Durbin R. Genome assembly in the telomere-to-telomere era. Nat Rev Genet. 2024;25:658–70.
11. Feng Y, Zhou J, Li D, Wang Z, Peng C, Zhu G. The haplotype-resolved T2T genome assembly of the wild potato species *Solanum commersonii* provides molecular insights into its freezing tolerance. Plant Commun. 2024;5:100980.
12. Yang X, Zhang L, Guo X, Xu J, Zhang K, Yang Y, et al. The gap-free potato genome assembly reveals large tandem gene clusters of agronomical importance in highly repeated genomic regions. Mol Plant. 2023;16:314–7.
13. Sun H, Jiao W-B, Krause K, Campoy JA, Goel M, Folz-Donahue K, et al. Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. Nat Genet. 2022;54:342–8.

14. Bao Z, Li C, Li G, Wang P, Peng Z, Cheng L, et al. Genome architecture and tetrasomic inheritance of autotetraploid potato. Mol Plant. 2022;15:1211–26.
15. Serra Mari R, Schrinner S, Finkers R, Ziegler FMR, Arens P, Schmidt MH-W, et al. Haplotype-resolved assembly of a tetraploid potato genome using long reads and low-depth offspring data. Genome Biol. 2024;25:26.
16. 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. 2024;21:967–70.
17. Zeng X, Yi Z, Zhang X, Du Y, Li Y, Zhou Z, et al. Chromosome-level scaffolding of haplotype-resolved assemblies using Hi-C data without reference genomes. Nat Plants. 2024;10:1184–200.
18. 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. 2019;5:833–45.
19. Sun H, Tusso S, Dent CI, Goel M, Wijfjes RY, Baus LC, et al. The phased pan-genome of tetraploid European potato. Nature. 2025;642:389–97.
20. Uitdewilligen JG, Wolters A-MA, D'hoop BB, Borm TJ, Visser RG, Van Eck HJ. A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. PLoS One. 2013;8:e62355.
21. Jiao W-B, Schneeberger K. Chromosome-level assemblies of multiple Arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics. Nat Commun. 2020;11:989.
22. Zhang C, Wang P, Tang D, Yang Z, Lu F, Qi J, et al. The genetic basis of inbreeding depression in potato. Nat Genet. 2019;51:374–8.
23. Ramírez-González R, Borrill P, Lang D, Harrington S, Brinton J, Venturini L, et al. The transcriptional landscape of polyploid wheat. Science. 2018;361:eaar6089.
24. Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, et al. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. Cell. 2020;182:145-161.e123.
25. Li X, Wang Y, Cai C, Ji J, Han F, Zhang L, et al. Large-scale gene expression alterations introduced by structural variation drive morphotype diversification in Brassica oleracea. Nat Genet. 2024;56:517–29.
26. Wang L, Zeng Z, Zhang W, Jiang J. Three potato centromeres are associated with distinct haplotypes with or without megabase-sized satellite repeat arrays. Genetics. 2014;196:397–401.
27. Gong Z, Wu Y, Koblížková A, Torres GA, Wang K, Iovene M, et al. Repeatless and repeat-based centromeres in potato: implications for centromere evolution. Plant Cell. 2012;24:3559–74.
28. Naish M, Alonge M, Wlodzimierz P, Tock AJ, Abramson BW, Schmücker A, et al. The genetic and epigenetic landscape of the Arabidopsis centromeres. Science. 2021;374:eabi7489.
29. Torres GA, Gong Z, Lovene M, Hirsch CD, Buell CR, Bryan GJ, et al. Organization and evolution of subtelomeric satellite repeats in the potato genome. G3 Genes Genomes Genetics. 2011;1:85–92.
30. Bhat JA, Yu D, Bohra A, Ganie SA, Varshney RK. Features and applications of haplotypes in crop breeding. Commun Biol. 2021;4:1266.
31. Jiao W-B, Schneeberger K. The impact of third generation genomic technologies on plant genome assembly. Curr Opin Plant Biol. 2017;36:64–70.
32. Song J-M, Xie W-Z, Wang S, Guo Y-X, Koo D-H, Kudrna D, et al. Two gap-free reference genomes and a global view of the centromere architecture in rice. Mol Plant. 2021;14:1757–67.
33. Chen J, Wang Z, Tan K, Huang W, Shi J, Li T, et al. A complete telomere-to-telomere assembly of the maize genome. Nat Genet. 2023;55:1221–31.
34. Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al. Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods. 2016;13:1050–4.
35. Garg S, Fungtammasan A, Carroll A, Chou M, Schmitt A, Zhou X, et al. Chromosome-scale, haplotype-resolved assembly of human genomes. Nat Biotechnol. 2021;39:309–12.
36. Zhang Z, Yang T, Liu Y, Wu S, Sun H, Wu J, et al. Haplotype-resolved genome assembly and resequencing provide insights into the origin and breeding of modern rose. Nat Plants. 2024;10:1659–71.
37. Hardigan MA, Laimbeer FPE, Newton L, Crisovan E, Hamilton JP, Vaillancourt B, et al. Genome diversity of tuber-bearing *Solanum* uncovers complex evolutionary history and targets of domestication in the cultivated potato. Proc Natl Acad Sci U S A. 2017;114:E9999-10008.
38. Bradshaw JE. Breeding diploid F1 hybrid potatoes for propagation from botanical seed (TPS): comparisons with theory and other crops. Plants. 2022;11:1121.
39. Zhang C, Yang Z, Tang D, Zhu Y, Wang P, Li D, et al. Genome design of hybrid potato. Cell. 2021;184:3873-3883.e3812.
40. Cheng L, Wang N, Bao Z, Zhou Q, Guarracino A, Yang Y, et al. Leveraging a phased pangenome for haplotype design of hybrid potato. Nature. 2025;640:408–17.
41. Mascher M, Jayakodi M, Shim H, Stein N. Promises and challenges of crop translational genomics. Nature. 2024;636:585–93.
42. Jing S, Sun X, Yu L, Wang E, Cheng Z, Liu H, et al. Transcription factor stabi5-like 1 binding to the FLOWERING LOCUS T homologs promotes early maturity in potato. Plant Physiol. 2022;189:1677–93.
43. Li H, Birol I. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.
44. Shen W, Sipos B, Zhao L. SeqKit2: a swiss army knife for sequence and alignment processing. Imeta. 2024;3:e191.
45. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10:giab008.
46. Yu H, Shi C, He W, Li F, Ouyang B. Pandepth, an ultrafast and efficient genomic tool for coverage calculation. Brief Bioinform. 2024;25:bbae197.
47. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016;17:1–14.
48. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013;arXiv:1303.3997.
49. Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst. 2016;3:99–101.

50. Xu M, Guo L, Gu S, Wang O, Zhang R, Peters BA, et al. TGS-gapcloser: a fast and accurate gap closer for large genomes with low coverage of error-prone long reads. Gigascience. 2020;9:giaa094.
51. Mc Cartney AM, Shafin K, Alonge M, Bzikadze AV, Formenti G, Fungtammasan A, et al. Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. Nat Methods. 2022;19:687–95.
52. Jain C, Rhie A, Hansen NF, Koren S, Phillippy AM. Long-read mapping to repetitive reference sequences using Winnowmap2. Nat Methods. 2022;19:705–10.
53. Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 2017;27:737–46.
54. Formenti G, Rhie A, Walenz BP, Thibaud-Nissen F, Shafin K, Koren S, et al. Merfin: improved variant filtering, assembly evaluation and polishing via k-mer validation. Nat Methods. 2022;19:696–704.
55. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021;18:170–5.
56. Ou S, Jiang N. Ltr_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. Plant Physiol. 2018;176:1410–22.
57. Huang N, Li H. Compleasm: a faster and more accurate reimplementation of BUSCO. Bioinformatics. 2023;39:btad595.
58. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. Nat Methods. 2012;9:357–9.
59. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37:907–15.
60. Ou S, Su W, Liao Y, Chougule K, Agda JR, Hellinga AJ, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 2019;20:1–18.
61. Gabriel L, Brůna T, Hoff KJ, Ebel M, Lomsadze A, Borodovsky M, et al. BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. Genome Res. 2024;34:769–77.
62. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 2008;24:637–44.
63. Brůna T, Lomsadze A, Borodovsky M. GeneMark-etp significantly improves the accuracy of automatic annotation of large eukaryotic genomes. Genome Res. 2024;34:757–68.
64. Chen S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. Imeta. 2023;2:e107.
65. Shumate A, Wong B, Pertea G, Pertea M. Improved transcriptome assembly using a hybrid of long and short reads with StringTie. PLoS Comput Biol. 2022;18:e1009730.
66. Chan PP, Lin BY, Mak AJ, Lowe TM. TRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. Nucleic Acids Res. 2021;49:9077–96.
67. Lagesen K, Hallin P, Rødland EA, Stærfeldt H-H, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 2007;35:3100–8.
68. Wucher V, Legeai F, Hédan B, Rizk G, Lagoutte L, Leeb T, et al. FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. Nucleic Acids Res. 2017;45:e57–e57.
69. Tian X-C, Chen Z-Y, Nie S, Shi T-L, Yan X-M, Bao Y-T, et al. Plant-LncPipe: a computational pipeline providing significant improvement in plant lncRNA identification. Hortic Res. 2024;11:uhae041.
70. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods. 2021;18:366–8.
71. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013;29:2933–5.
72. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30:1236–40.
73. Li Q, Jiang X-M, Shao Z-Q. Genome-wide analysis of NLR disease resistance genes in an updated reference genome of barley. Front Genet. 2021;12:694682.
74. Steuernagel B, Witek K, Krattinger SG, Ramirez-Gonzalez RH, Schoonbeek H-j, Yu G, et al. The NLR-annotator tool enables annotation of the intracellular immune receptor repertoire. Plant Physiol. 2020;183:468–82.
75. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
76. Goel M, Sun H, Jiao W-B, Schneeberger K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome Biol. 2019;20:1–13.
77. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;6:80–92.
78. Zhang Z, Zhang P, Ding Y, Wang Z, Ma Z, Gagnon E, et al. Ancient hybridization underlies tuberization and radiation of the potato lineage. Cell. 2025;188:5249-5265.e5215.
79. Du Z-Z, He J-B, Jiao W-B. Syndiv: an efficient tool for chromosome collinearity-based population genomics analyses. Plant Commun. 2024;5:101071.
80. Tang H, Krishnakumar V, Zeng X, Xu Z, Taranto A, Lomas JS, et al. JCVI: a versatile toolkit for comparative genomics analysis. Imeta. 2024;3:e211.
81. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20:1–14.
82. Brown MR, Gonzalez de La Rosa P, Blaxter M. tidk: a toolkit to rapidly identify telomeric repeats from genomic datasets. Bioinformatics. 2025:btaf049.
83. Zhang Y, Chu J, Cheng H, Li H. De novo reconstruction of satellite repeat units from sequence data. Genome Res. 2023;33:1994–2001.
84. Gel B, Serra E. KaryoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. Bioinformatics. 2017;33:3088–90.

85. Nagaki K, Talbert PB, Zhong CX, Dawe RK, Henikoff S, Jiang J. Chromatin immunoprecipitation reveals that the 180-bp satellite repeat is the key functional DNA element of *Arabidopsis thaliana* centromeres. Genetics. 2003;163:1221–5.
86. Su H, Liu Y, Dong Q, Feng C, Zhang J, Liu Y, et al. Dynamic location changes of Bub1-phosphorylated-H2AThr133 with CENH3 nucleosome in maize centromeric regions. New Phytol. 2017;214:682–94.
87. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. Bioinformatics. 2015;31:2032–4.
88. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9:1–9.
89. Vollger MR, Kerpedjiev P, Phillippy AM, Eichler EE. Stainedglass: interactive visualization of massive tandem repeat structures with identity heatmaps. Bioinformatics. 2022;38:2049–51.
90. Gao S, Yang X, Guo H, Zhao X, Wang B, Ye K. HiCAT: a tool for automatic annotation of centromere structure. Genome Biol. 2023;24:58.
91. Zhang R-G, Li G-Y, Wang X-L, Dainat J, Wang Z-X, Ou S, et al. Tesorter: an accurate and fast method to classify LTR-retrotransposons in plant genomes. Hortic Res. 2022;9:uhac017.
92. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics. 2011;27:1571–2.
93. Huang X, Zhang S, Li K, Thimmapuram J, Xie S. Viewbs: a powerful toolkit for visualization of high-throughput bisulfite sequencing data. Bioinformatics. 2018;34:708–9.
94. Liao Y, Smyth GK, Shi W. Featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30:923–30.
95. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):1–21.
96. Tian F, Yang D-C, Meng Y-Q, Jin J, Gao G. Plantregmap: charting functional regulatory maps in plants. Nucleic Acids Res. 2020;48:D1104-13.
97. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. Clusterprofiler 4.0: a universal enrichment tool for interpreting omics data. Innovation. 2021;2:100141.
98. Liu X, Bie XM, Lin X, Li M, Wang H, Zhang X, et al. Uncovering the transcriptional regulatory network involved in boosting wheat regeneration and transformation. Nat Plants. 2023;9:908–25.
99. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016;32:2847–9.
100. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34:525–7.
101. Xiao P-X, Tan L, Dong J, Huang J, Huang Y, He J-B, Su H, Song B, Jiao W-B. Haplotype-resolved and near telomere-to-telomere assembly of the autotetraploid potato genome. Datasets. National Genomics Data Center. 2025. https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA036064.
102. Xiao P-X, Tan L, Dong J, Huang J, Huang Y, He J-B, Su H, Song B, Jiao W-B. Haplotype-resolved and near telomere-to-telomere assembly of the autotetraploid potato genome. Datasets. 2025. https://figshare.com/s/df110f76382070bc3cc3.
103. Xiao P-X, Tan L, Dong J, Huang J, Huang Y, He J-B, Su H, Song B, Jiao W-B. PHap: A haplotype-resolved and telomere-to-telomere genome assembly pipeline tailored for autopolyploids. GitHub. 2025. https://github.com/JiaoLab2021/PHap.
104. Xiao P-X, Tan L, Dong J, Huang J, Huang Y, He J-B, et al. PHap: A haplotype-resolved and telomere-to-telomere genome assembly pipeline tailored for autopolyploids. 2026. Zenodo. https://doi.org/10.5281/zenodo.18223394.

Publisher's Note

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