Haplotype-resolved assembly of a tetraploid potato genome using long reads and low-depth offspring data

Rebecca Serra Mari, Sven Schrinner, Richard Finkers, Freya Maria Rosemarie Ziegler, Paul Arens, Maximilian H.-W. Schmidt, Björn Usadel, Gunnar W. Klau & Tobias Marschall

METHOD

Open Access

Haplotype-resolved assembly of a tetraploid potato genome using long reads and low-depth offspring data

Rebecca Serra Mari[1,2], Sven Schrinner[3,2], Richard Finkers[4,5], Freya Maria Rosemarie Ziegler[6,7,8,9], Paul Arens[5], Maximilian H.-W. Schmidt[6,7], Björn Usadel[6,7,8,9*], Gunnar W. Klau[3,6*] and Tobias Marschall[1,2*]

*Correspondence:
b.usadel@fz-juelich.de;
gunnar.klau@hhu.de; tobias.marschall@hhu.de

[1] Institute for Medical Biometry and Bioinformatics, Medical Faculty and University Hospital Düsseldorf, Heinrich Heine University Düsseldorf, Düsseldorf, Germany
[2] Center for Digital Medicine, Heinrich Heine University Düsseldorf, Düsseldorf, Germany
[3] Algorithmic Bioinformatics, Faculty of Mathematics and Natural Sciences, Heinrich Heine University Düsseldorf, Düsseldorf, Germany
[4] Gennovation B.V., Agro Business Park 10, 6708, PW, Wageningen, The Netherlands
[5] Plant Breeding, Wageningen University & Research, Wageningen, The Netherlands
[6] Cluster of Excellence on Plant Sciences (CEPLAS), Heinrich Heine University Düsseldorf, Düsseldorf, Germany
[7] Forschungszentrum Jülich, Institute of Bio and Geosciences, Bioinformatics (IBG-4), Jülich, Germany
[8] Bioeconomy Science Center, c/o Forschungszentrum Jülich, Jülich, Germany
[9] Biological Data Science, Faculty of Mathematics and Natural Sciences, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

Abstract

Potato is one of the world's major staple crops, and like many important crop plants, it has a polyploid genome. Polyploid haplotype assembly poses a major computational challenge. We introduce a novel strategy for the assembly of polyploid genomes and present an assembly of the autotetraploid potato cultivar Altus. Our method uses low-depth sequencing data from an offspring population to achieve chromosomal clustering and haplotype phasing on the assembly graph. Our approach generates high-quality assemblies of individual chromosomes with haplotype-specific sequence resolution of whole chromosome arms and can be applied in common breeding scenarios where collections of offspring are available.

Background

Polyploidy is common in plant genomes and two forms are recognized. Allopolyploids arise from interspecific or intergeneric hybridization events, and the difference between subgenomes is usually sufficient to assemble them like diploids. This has been demonstrated for rapeseed, wheat, and strawberry, among others [1]. In contrast, autopolyploids arise from genome duplications, and the presence of multiple sets of the same homologous chromosomes means that haplotype-resolved sequence assemblies are much more challenging. One example is potato (*Solanum tuberosum*), most cultivars of which are autotetraploid [2]. Potato is a vital food crop in many developing countries [3], and the global production volume exceeds 300 million tons per year [4]. Because of this agronomic value, efforts to assemble potato genomes are of crucial importance.

The haplotype-resolved assembly of diploid genomes has been progressively refined, and accurate results are now possible as we have shown previously [5, 6]. In contrast, computational methods for polyploid haplotype assembly rarely lead to satisfying results, particularly for autotetraploids. Reference-based approaches for haplotype phasing in polyploid species align reads to an existing reference sequence but are often inaccurate [7]. Especially in the presence of structural variation, reference-based approaches in general have severe limitations [6]. For potato haplotype phasing, two reference genomes are currently used: the synthetic double monoploid potato clone DM1–3 516 R44 [8] and Solyntus, which is based on a diploid potato cultivar [9]. Reference-based algorithms for polyploid haplotype phasing include HapTree [10] and H-PoP [11]. Other methods target selected genomic regions to resolve haplotypes locally, for example using integer linear programming [12]. We previously developed WhatsHap polyphase, which was an improvement over contemporaneous methods but still relied on a reference genome [13].

The de novo assembly of polyploid genomes without a reference is still an emerging strategy. Recently proposed workflows involve the building of a "squashed" assembly with no or limited haplotype resolution at first, and using this as the basis for haplotype phasing. Even long sequencing reads are generally insufficient for long-range phasing, and auxiliary data types are required. One example is single pollen cell sequencing [14], which was recently used for comprehensive haplotype reconstruction in autotetraploid potato [15]. Another example is the recent publication of a potato assembly [16] where a selfing population of 1034 samples was used.

Here, we propose an alternative method in which PacBio HiFi reads of the potato cultivar Altus are combined with cost-effective low-coverage short-read sequences from multiple offspring samples. Accordingly, we generated PacBio HiFi reads (96× coverage) and created an initial assembly using hifiasm [17]. We assembled the individual haplotypes from the resulting assembly graph using sequencing data from 193 offspring of two potato cultivars (Altus and Colomba) at low coverage (~1.5× per haplotype) combined with a novel approach based on *k*-mers to identify the four haplotypes. Our assembly mapped well to the latest version of the monoploid DM1–3 516 R44 reference (DMv6.1) and yielded haplotype-resolved assemblies of individual chromosomes with phased haplotype block lengths of up to 34 Mb, phased contig N50 values of up to 12 Mb, and a genome-wide phased contig N50 value of 7.5 Mb. Our approach also allows the detection and correction of assembly errors in the assembly graph as well as in previously published references.

Results

Overall assembly strategy

A high-level overview of our workflow is shown in Fig. 1. Starting with PacBio HiFi reads derived from the Altus genome (Fig. 1a), we built an assembly graph using hifiasm, resulting in a partially haplotype-resolved graph with bubble-like structures representing the different haplotypes (Fig. 1b). For each so-called unitig in the assembly graph, we detected unique *k*-mers (Fig. 1b). We then estimated the dosage of each unitig, defined as the number of haplotypes to which each unitig contributes (Fig. 1c). In the next step, we counted the formerly detected unique *k*-mers in the Illumina reads for each of the 193 offspring samples (Fig. 1d). Each unitig is thus represented by a *k*-mer count pattern consisting of 193 values. Nodes with similar count patterns, implying the inheritance of a node by the same subset of offspring samples, are therefore likely to be part of the same haplotype. We then made use of the *k*-mer count patterns to perform an initial

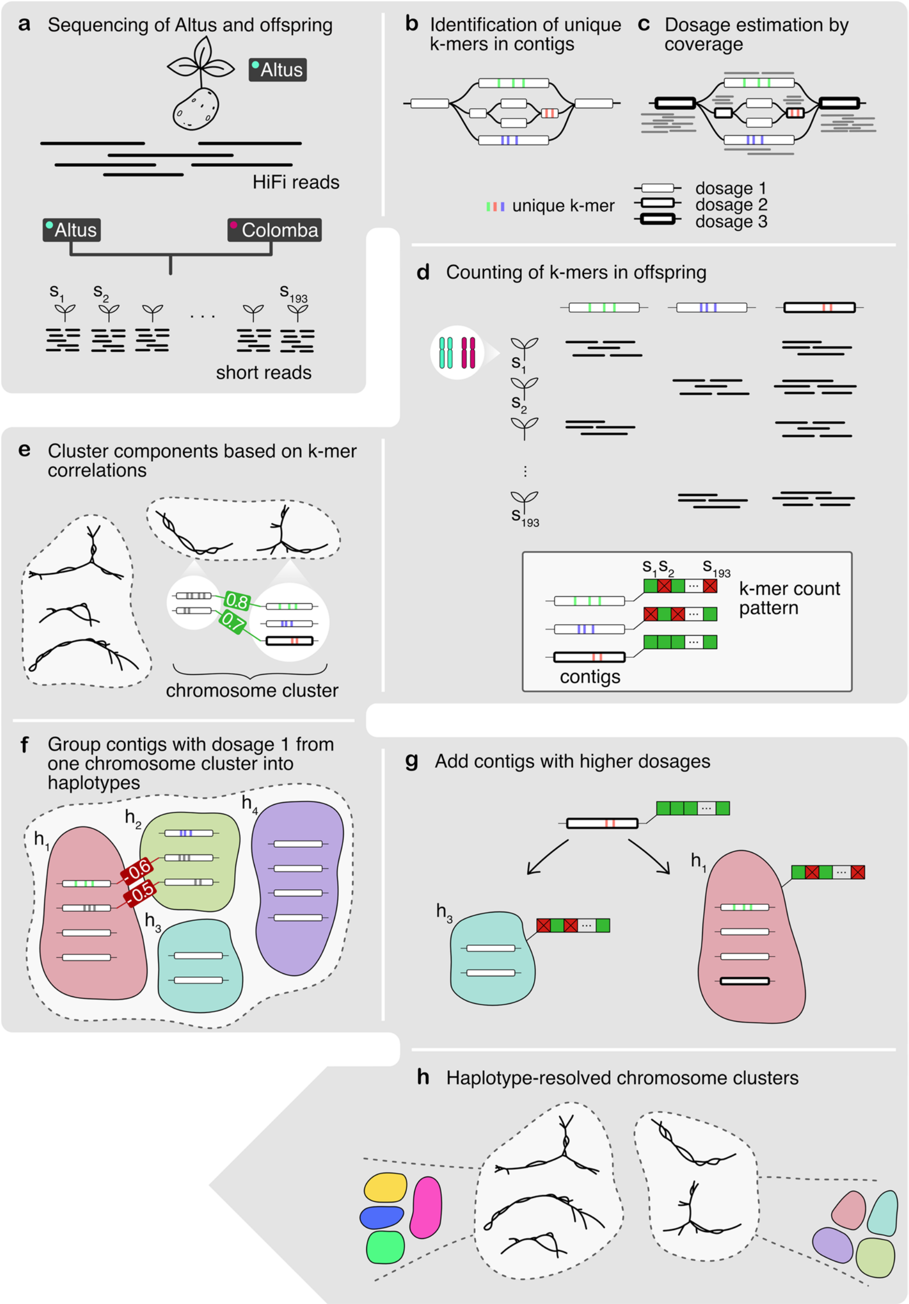
a Sequencing of Altus and offspring
Altus
HiFi reads
Altus
Colomba
s1
s2
s193
short reads
b Identification of unique k-mers in contigs
unique k-mer
c Dosage estimation by coverage
dosage 1
dosage 2
dosage 3
d Counting of k-mers in offspring
s1
s2
s193
s1 s2
s193
k-mer count pattern
contigs
e Cluster components based on k-mer correlations
0.8
0.7
chromosome cluster
f Group contigs with dosage 1 from one chromosome cluster into haplotypes
h1
h2
h3
h4
0.6
0.5
g Add contigs with higher dosages
h3
h1
h Haplotype-resolved chromosome clusters

Fig. 1 Overview of the workflow. **a** The Altus genome was sequenced using PacBio HiFi technology, whereas the 193 genomes of the cross Altus × Colomba were sequenced on the Illumina platform. **b** We used hifiasm to assemble the Altus HiFi reads into an assembly graph. For each contig in the graph, unique *k*-mers were detected (denoted by the colored bars). **c** The HiFi reads were aligned to the contigs and the mapping depth was used to estimate dosages (1 to 4) for each contig. The different dosages are denoted by the thickness of the contig line (thicker outlines mean higher dosage). **d** The unique *k*-mers were counted in the short reads of the offspring samples in order to compose a count pattern for each contig. **e** For all nodes from the assembly graph components, the pairwise correlation of *k*-mer count patterns was computed and components were clustered to represent chromosomes. **f** In each chromosome cluster, the nodes with estimated dosage 1 were first clustered into the four haplotypes, again based on pairwise correlations. **g** The contigs with dosages > 1 were added to the clusters that contain the most matching nodes in terms of *k*-mer count pattern correlations. **h** This process resulted in chromosome clusters that contain subclusters for each haplotype

clustering of the nodes into chromosomes (Fig. 1e). The clustering procedure was followed by a step to determine the four haplotypes among nodes with dosage 1 (Fig. 1f), and another step to add nodes with higher dosages (Fig. 1g). Ultimately, this yielded a set of four haplotype clusters for each chromosome (Fig. 1). We completed the assembly by finding graph traversals through the clustered assembly graph and thereby assembling haplotype blocks (haplotigs).

Initial assembly

We first sequenced the Altus genome using PacBio HiFi technology to produce highly accurate long reads with an average coverage of 24× per haplotype (73.7 Gb in total). We also acquired Illumina short-read sequencing data representing 193 offspring of the cross between Altus and another cultivar (Colomba). The data consisted of 2 × 150 bp paired-end reads with an average coverage of 1.5× per haplotype.

We assembled the HiFi reads using hifiasm v0.13, which outputs an assembly graph that contains all the assembled, unprocessed (raw) unitigs, which partially resolve the four haplotypes. Variation is represented by bubble structures in the graph, where a unitig branches into two or more other unitigs.

The initial graph consisted of 20,216 nodes (unitigs) and 26,566 edges and contained 2798 Mb of sequence data. The N50 value of the unitigs was 1.34 Mb. The nodes of the unitig graph (Additional file 1: Fig. S1) within the 10 largest connected components covered 91–190 Mb each (1.27 Gb in total), 11 further components covered 45–66 Mb each (555.2 Mb in total), and a set of smaller components covered 20–32 Mb each (249.1 Mb in total). Additionally, 699 unitigs were not connected to any other node. In summary, the initial raw unitig graph provided a certain degree of haplotype resolution, indicated by the total amount of sequence data (3.8× the size of the DMv6.1 reference genome), but did not provide longer-range phasing at many loci, indicated by the substantial number of nodes shorter than 50 kb (Fig. 2a).

Dosage analysis

For each unitig, we estimated the dosage (number of haplotypes represented), which for a tetraploid genome can be any value from the set {1, 2, 3, 4}. This was achieved by analyzing the coverage of reads aligned to the unitigs. First, we aligned all Altus HiFi reads to the graph unitigs using minimap2 [18] and filtered out all alignments with mapping quality below 60. Using the remaining alignments, we computed the sequencing depth at each base position. Given that hifiasm graphs usually contain overlaps, we computed the intervals of non-overlapping sequences per node (the region of each node that is not part of any overlap with its neighboring nodes) and only computed the depth in these unique regions, leading to an average depth per node. Nodes with a non-overlapping sequence of 100 kb or longer (Fig. 2b) covered ~80% of the total sequence in the graph. Three peaks were observed, representing approximate coverage values of 23, 46, and 69, consistent with dosages of 1, 2, and 3. A fourth peak (~92) was missing for the long contigs (Fig. 2b) and barely visible for all contigs (Additional file 1: Fig. S2). This may indicate the existence of only a few homozygous regions and the complete absence of long homozygous stretches exceeding 100 kb.

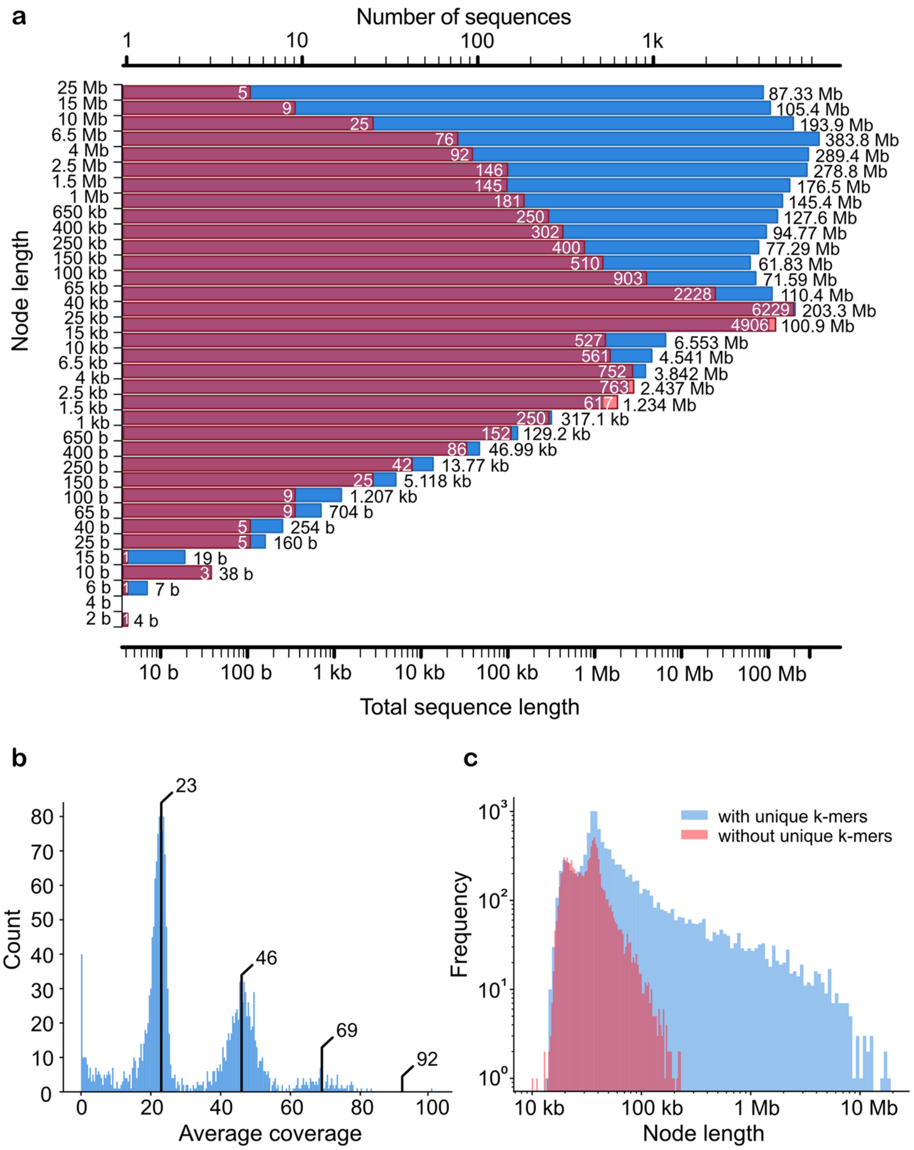

Fig. 2 Initial assembly. **a** Distribution of node lengths of the initial assembly graph. Red represents the count of each binned contig length (the peak is 25–40 kb). Blue represents the aggregate length of a contig bin, measured in bases. The two visible peaks show that the total sequence of contigs between 25 and 40 kb is on par with the sequence taken up by those between 4.0 and 6.5 Mb. **b** Dosage distribution of contigs, excluding those with a unique sequence < 100 kb. The proportion of sequence that is covered by contigs at least 100 kb in length is 80%. The dosage peaks are marked by black bars (approximate coverage values of 23, 46, and 69). The peak for dosage 4 would be ~92. **c** Length distribution of contigs with unique *k*-mers compared to contigs without unique *k*-mers

For 6212 contigs, the sequence consisted solely of overlaps to both neighboring nodes. Given the absence of a unique region, we therefore omitted these contigs from the computation of coverage. These reads accounted for 0.148 Gb in total, where the longest node was 42,105 bp.

Of the 8290 nodes with a depth value above zero, 72.77% were labeled as dosage one, 15.01% as dosage two, 7.95% as dosage three, and 2.97% as dosage four. The remaining 1.3% of the contigs exceeded dosage four and are presumed to represent repetitive regions. In total, these nodes that could be assigned a dosage estimate consisted of 2.396 Gb.

These dosage estimates are a key step in our assembly process and are, due to repetitive sequence, difficult to validate using short-read technologies. We therefore produced 162 Gb of long-read sequencing data from the Oxford Nanopore Technologies (ONT) platform (see the "Methods" section). We again estimated the dosage as described above. Comparing dosage estimates based on alignments of ONT reads to those based on HiFi reads, we found that for 6982 nodes (84.2%), both estimates were equal. These nodes had a total length of 2.233 Gb, which is 93.22% of the formerly described 8290 nodes. The joint distribution of ONT- and HiFi-based coverage estimates can be found in Additional file 1: Fig. S3, confirming the robustness of the dosage estimates, especially for nodes with sufficient amounts of unique sequence.

Analysis of *k*-mers

In the next step, we counted all possible *k*-mers (fragments of length *k*, in our case *k* = 71) within the unitigs. We then identified those appearing exactly once in the entire graph. Additionally, we ensured the *k*-mers are unique for the Altus genome by disregarding those also found in the Colomba genome. We reported the resulting set as unique *k*-mers. Approaches based on unique *k*-mers have facilitated the haplotype-resolved assembly of diploid parent-offspring trios [19] and challenging regions of human chromosome 8, such as the centromere [20]. In the latter example, the authors created a library of singly unique nucleotide *k*-mers (SUNKs) to barcode long reads and assemble them into scaffolds especially in complex regions. Here, we have developed a novel approach to phase the assembly graph of a parent genome from a polyploid offspring panel. For each unitig, we used the corresponding set of unique *k*-mers as an identifier for the node. The *k*-mer counting process is based on the Jellyfish API [21].

The resulting set of unique *k*-mers was counted in the 193 offspring samples. Given that each of the tetraploid offspring inherits two haplotypes from Altus and two from Colomba, we inferred the number of inherited copies of a unitig by assessing the abundance of unique *k*-mers for that unitig. Based on the unitig dosage in the parental Altus genome, there are different possible dosages in the offspring. For dosage 1 in the parent, the *k*-mer representing the unitig can be absent or present in an offspring genome, whereas for parental dosage 2, the *k*-mer can be absent, present once, or present twice in the offspring. For parental dosage 3, dosages of 1 and 2 are possible in the offspring, and for parental dosage 4, both inherited haplotypes must arise from this unitig.

Based on the above, we can denote unitigs with unique *k*-mers as *phase informative* and those without unique *k*-mers as *phase uninformative*. The analysis of node lengths for the sets of phase informative and uninformative nodes is shown in Fig. 2c. As anticipated, the uninformative unitigs were generally the shorter ones. Among the complete set of 20,216 nodes, we found that 10,784 (53.34%) were phase informative. The different node sets are visualized in Additional file 1: Fig. S4, showing the overlaps between nodes that are, for instance, phase informative, those for which the dosage could be confidently estimated, and others. Recall that 6212 contigs did not have a unique region due to overlaps, so that unique *k*-mers cannot be present in these nodes. The length of the sequence covered by informative nodes in relation to the sequence covered by all nodes was 88.15% (2.466 of 2.798 Gb), showing that phase uninformative nodes tend to be shorter than phase informative nodes. Specifically, the average node length in the set of phase informative nodes was 228.7 kb (N50 = 1.89 Mb) whereas the average for uninformative nodes was 35.1 kb (N50 = 37 kb). The longest unitig without a unique *k*-mer was 237 kb, compared to 19.11 Mb for the longest informative unitig. Thus, despite the relatively high number of phase uninformative nodes, most of the sequence (88.15%) was generally amenable to offspring-based phasing using our technique.

Correlation analysis

For our correlation-based approach, we exploit that unitigs from the same haplotype have similar *k*-mer count patterns because the corresponding haplotype context is transmitted to the same subset of offspring samples. The unitigs used for this procedure are (a) phase informative and (b) have the same dosage estimates based on both the HiFi and the ONT data. This ensures that the correlation clustering process starts with the most reliable dosage estimates. We computed the Spearman correlation coefficients (ρ) between the *k*-mer count patterns of the candidate unitigs and analyzed the distribution of correlations throughout the assembly graph. Genomic loci that are at close distance (and thus, tightly linked) are likely to be transmitted together to progeny, and accordingly, the offspring-based haplotype signal gets weaker with increasing distance due to recombination. In line with this expectation, we observed a strong correlation for contigs at distances < 10 Mb and a decreasing correlation for greater distances (Fig. 3a).

Based on high correlation values, we were able to reconstruct areas from the graph that were unconnected in the initial assembly, such as broken bubble structures and unconnected fragments. A representative reconstruction of chromosome 3 is shown in Fig. 3b. In the initial assembly, this chromosome consisted of three connected components and two longer unconnected contigs. By connecting contig pairs with very high positive correlation coefficients ($\rho \geq 0.8$), we were able to reconstruct the phased structure of the chromosome and to order the components and contigs accordingly.

Graph traversal and final haplotype assembly

We assigned each unitig to a haplotype based on our novel clustering procedure, which is described in more detail in "Methods" section. We extended the clustering beyond the graph components and thereby matched the components belonging to the same chromosome. This resulted in 12 pseudo-chromosomes, each consisting of four clusters of haplotagged unitigs. Additionally, we connected the resulting clustered contigs as far as possible by finding graph traversals within the assembly graph, yielding blocks corresponding to the four haplotypes, which we describe as haplotigs. The longest haplotig per chromosome ranged from 11.50 Mb on chromosome 10 to 34.09 Mb on chromosome 7. The haplotig N50 value was $\leq$ 12 Mb and the total N50 value was 7.54. The full dataset is presented in Table 1.

To investigate whether the same clustering performance could be achieved with fewer offspring samples, we repeated the clustering after downsampling offspring data to 50, 100, 120, and 150 samples, respectively. For the clustering procedure to work as expected, we found that at least 150 samples are necessary. With 50 and 100 samples, no significant clusters were created. With 120 samples, the clustering outputs 5 clusters

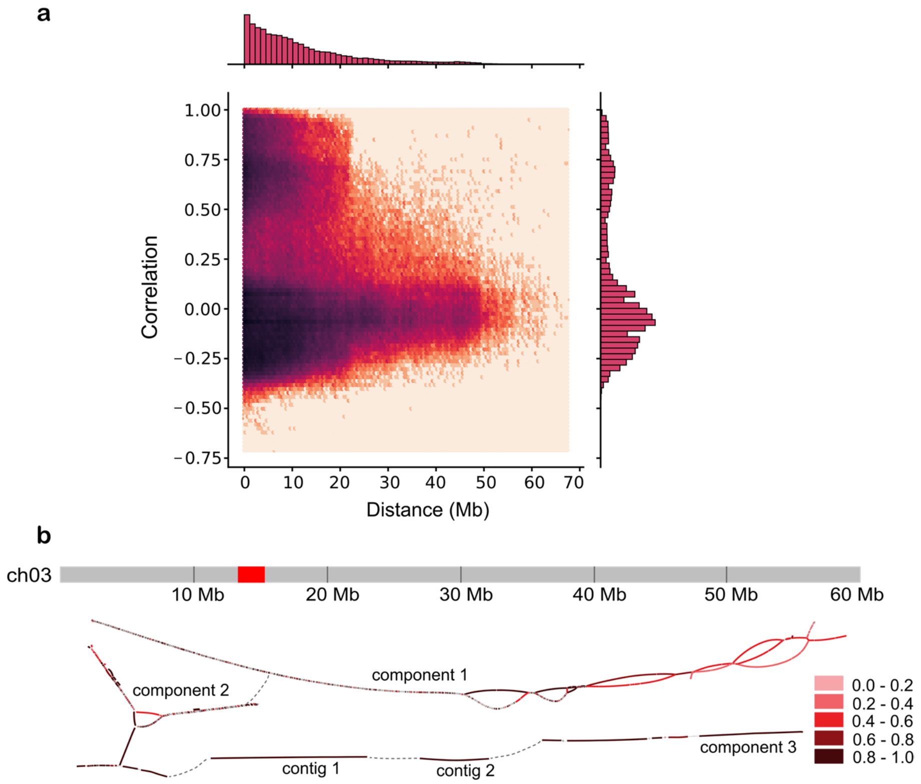

Fig. 3 Correlation analysis. **a** The correlation of all node pairs (nodes with dosage 1) in the 20 largest connected components as a function of the distance between nodes (in megabases). The 4830 dosage-1 nodes of the largest components account for 947.78 Mb. After removing pairs which had no valid correlation (NaN), 701,582 pairs remained in the dataset for plotting. **b** Reconstruction of the structure of chromosome 3 based on high correlation coefficients between nodes. Chromosome 3 is shown above, with the red block labeling the centromere as reported in the DMv6.1 annotation. The initial assembly consisted of three connected components and two additional contigs, which were manually placed at their approximate genomic location along the *x*-axis as determined by mapping the unitigs to DMv6.1 (the darker the color of a contig, the higher the maximum correlation to any other contig beyond its component). Contig pairs with the highest correlation (here denoted by the darkest color, representing a correlation coefficient of ≥ 0.8) could then be connected, revealing a more complete structure of the haplotype-resolved chromosome. The connected node pairs are marked by the dotted gray line

instead of the expected 12 (one for each chromosome). Note that in this step, 12 is the expected number of clusters since the first step of our clustering method determines the chromosome clusters, of which we expect 12. It is only in the second step that these clusters are further refined to specific haplotype clusters. Since in our experiments with offspring sample numbers lower than 150, we noticed that even the 12 chromosomal clusters were not distinguished properly, we refrained from any further phasing steps. Only with an increased sample size of 150 offspring and more, the *k*-mer counts get differentiated enough to distinguish between 12 chromosome clusters and output the 12 pseudo-chromosomes. The following analyses are based on using the full set of 193 offspring samples.

We compared the assembled pseudo-chromosomes to the latest version of the monoploid reference, DMv6.1 [8]. Its length of 731.3 Mbp is consistent with a *k*-mer-based haploid genome size estimate from our HiFi data (730.6 Mb, Additional file 1: Fig. S5). To compute the N50 measures, we estimated the tetraploid genome size by using fourfold

Table 1 Comparative length of the reference genome and the phased assembly after clustering into haplotypes and constructing the final haplotigs. The phased length is defined as the sum of the contig lengths contained in the four haplotypes for each chromosome. The N50 value is computed with four times the reference length as the underlying genome size

Chromosome	Length of DMv6.1 (Mb)	N50 (Mb)	Longest haplotig (Mb)	Sum of haplotigs (Mb)
01	88.59	6.01	19.45	429.25
02	46.10	8.66	18.97	234.51
03	60.71	9.36	25.36	204.66
04	69.24	4.56	19.04	257.50
05	55.60	12.33	23.33	224.21
06	59.09	11.94	25.95	238.74
07	57.64	11.31	34.09	204.34
08	59.23	4.35	13.97	169.03
09	67.60	7.17	17.59	234.41
10	61.04	3.83	11.50	185.79
11	46.78	11.93	31.99	171.43
12	59.67	9.93	23.09	227.18
Total	731.29	7.54	34.01	2781.06

the length of DMv6.1 (2.925 Gb). The cumulative size comparison of each chromosome based on our assembled pseudo-chromosomes and DMv6.1 is provided in Table 1. The size of the individual phased chromosome was 3.5–4 times as large as the reference, and the total phased length was ~3.8 times as large, consistent with structural variation and sequence loss on some of the haplotypes as previously observed for other cultivars [15].

Sizes of the resulting haplotype clusters alongside the size of the DMv6.1 chromosome are visualized in Fig. 4b. Some chromosomes, for instance 7 and 11, are rather complete, while others (e.g., chromosome 10) exhibit shorter lengths on all four clusters. For the latter, more nodes had to be excluded as they did not contain enough valid sequence to be reliably assigned. The light gray bars above each cluster size bar indicate how much sequence was contained in the first clustering step to form the chromosomal cluster, but had to be excluded from the second step. This value is not haplotype-specific, as the correct amount of sequence that each haplotype cluster is lacking cannot be specified further, so we assigned equal portions of the unphased sequence from the chromosomal cluster to each haplotype for the purpose of the length estimate in Fig 4b (light gray part split equally between haplotypes).

For chromosome 2, one haplotype appears larger than expected. To test whether this is due to an assembly misjoin to a different chromosome, we mapped the four haplotype clusters to the reference (Additional file 1: Fig. S6). It becomes visible that the nodes from this haplotype map specifically to chromosome 2 in the reference, so it is unlikely to be a misjoin and more likely to be a higher amount of duplicated sequence.

For a general comparison of the Altus assembly and DMv6.1, we mapped the resulting clusters to the reference using minimap2. The corresponding mapping intervals (Fig. 4a) indicated that all chromosomes in the assembly were nearly complete, and no large gaps were detected. In all chromosomes, the assembly consisted entirely of contigs from one single cluster, supporting the robustness of our chromosome clustering

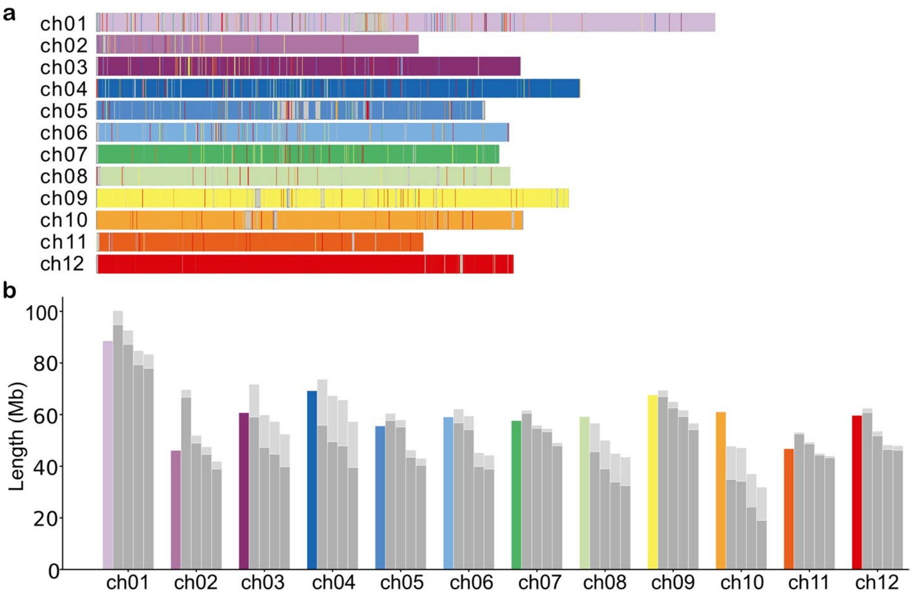

Fig. 4 Clustering results. **a** The contigs of each chromosome cluster are mapped to the reference sequence DMv6.1, and the mapped interval is colored accordingly. A different color is used for each cluster. Ideally, one chromosome contains a single color. **b** Length comparison of the four haplotypes (gray bars) compared to the reference (colored bars). The length is computed as the sum of the contig lengths for all contigs in a haplotype cluster. We have added the information on how much node sequence was included in the first clustering step to form the chromosomal cluster, but had to be excluded from the second step of phased clustering due to not being phase informative (= lacking the unique *k*-mer information), marked by the light gray bars atop of each cluster size bar. This visualization distributes the unphased sequence uniformly to each haplotype cluster. Note that the true assignment to haplotypes is, by definition, unknown for the unphased sequence

process. For a more detailed visualization showing mappings of all 48 chromosomal haplotype clusters, see Additional file 1: Fig. S7. The findings coincide with the results from Table 1, where some chromosomes exhibit a high mapping completeness on all four haplotypes, such as chromosomes 7 and 11, while others lack parts of their haplotypes, such as chromosome 10.

Comparison of earlier reference assemblies to reveal structural differences

The correlation signal underlying our chromosome clustering approach was used to detect structural differences between our assembly graphs and previous reference assemblies. Such differences can indicate assembly errors in either of the two assemblies, as well as structural differences in all or some haplotypes. When comparing the initial assembly graph (the hifiasm output) to the DMv6.1 reference, we detected two sets of nodes present on the same component of the graph that mapped to different chromosomes in DMv6.1 (Additional file 1: Fig. S8). For two contig sets on separate chromosomes, we would expect to see little to no correlation between node pairs from the two sets. Indeed, for the two sets in question, the correlation distribution was very similar in shape to the correlation between one of the sets and a comparison set from a different chromosome (Additional file 1: Fig. S8a). This probably indicates a false join in the hifiasm graph, which we corrected by manual curation. In this way, correlation analysis provides an opportunity to detect and correct residual assembly errors.

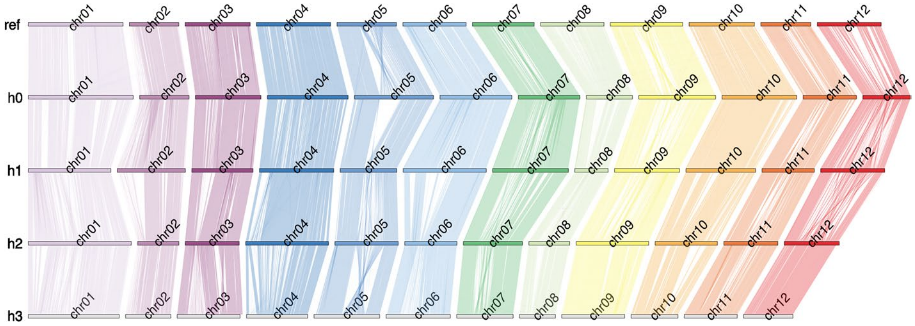

Fig. 5 Synteny analysis. High-level overview of structural differences between chromosomal haplotypes. Top row: DMv6.1, rows 1–4: Haplotypes 0, 1, 2, and 3

We then compared our assembly graphs to the diploid reference Solyntus [9] and found a number of larger structural differences (Additional file 1: Fig. S9). One example can be found in chromosome 8, where two regions are assembled from contigs that belong to the same clusters as chromosome 7 and chromosome 1, respectively. To investigate whether this was a clustering artifact, an error in the Solyntus assembly, or a true structural difference, we mapped the connected components from the graph representing chromosome 8 individually to the Solyntus reference and identified one component that contained a large fragment of chromosome 1 but also the inserted region on chromosome 8 (Additional file 1: Fig. S10). We again compared the *k*-mer count correlations of all node pairs within the component, distinguishing between the sets of contigs mapping to chromosomes 1 and 8. The former contained 563 nodes, covering 110.32 Mb, of which 315 featured unique *k*-mers and were thus suitable for the correlation computations (covered sequence = 102.48 Mb), whereas the latter contained 527 nodes, covering 74.6 Mb, of which 297 featured unique *k*-mers (covered sequence = 67.05 Mb). Again, we expect to see little or no correlation if two node sets originate from separate chromosomes. In this case, however, the distribution of correlations was consistent with the connections suggested by the assembly graph—contradicting the structure of the Solyntus reference (Additional file 1: Fig. S10a). These results suggest there is either a large rearrangement that distinguishes between the Altus and Solyntus genomes or a structural error in the Solyntus reference genome.

Structural analysis

To analyze structural distinctions among the chromosomes in our assembly, we employed two distinct approaches for synteny analysis: SyntenyPlotteR [22] and SyRI [23].

Initially, we generated scaffolds from the assembled haplotigs using RagTag [24] and the DMv6.1 reference sequence, as the aforementioned tools require a singular chromosome-wide sequence as input for the synteny analysis. Subsequently, pairwise alignments were computed between the scaffolded haplotypes (h1 to h0, h2 to h1, h3 to h2, and h0 to DMv6.1) using minimap2. The resulting synteny data were then input into SyntenyPlotteR.

The resulting synteny plot (Fig. 5) provides an overview of the structural differences between the chromosomal haplotypes, including the reference. Overall, high synteny is observed among the haplotypes and also with the reference. Notably, there are no translocations of sequence fragments between different chromosomes, underscoring the robustness of the clustering process.

For further analysis, we applied SyRI and plotsr [25] to the previously generated haplotype scaffolds in order to analyze structural differences for each individual chromosome. Synteny was computed between the scaffolded haplotypes and the reference DMv6.1 (Additional File 1: Fig. S11-16) as well as among the four haplotypes (Additional File 1: Fig. S17-28).

These synteny patterns reveal extensive rearrangements, for instance on chr02, similar to those reported for other cultivars [15]. However, we caution that the outcomes of such analyses are dependent on alignment parameters, and a comprehensive investigation of rearrangement histories should be based on a larger number of different haplotypes.

Quality evaluation and comparison to other assemblies

To assess the quality of our assembly, we performed further qualitative and quantitative analyses and compared the results to additional potato assemblies. To estimate the completeness of gene content, we have computed BUSCO scores with BUSCO v5.4.7 [26] using the solanales_odb10 database and obtained a completeness score of 96.8%. In comparison, this is similar to the completeness score listed for the DMv6.1 reference (97.9%), the assembly of C88 [16] (96.28%), and the assembly of Otava [15] (97.3%).

Furthermore, we have annotated repeats using RepeatModeler2 [27] and RepeatMasker (http://www.repeatmasker.org) and found 66.81% of repetitive content, the largest part of which was made up of LTRs. This is also in line with what was reported by Sun et al. in the assembly of Otava (66% of repetitive content, LTRs being the most abundant group) and by Pham et al. for the DMv6.1 assembly (66.8%).

For the qualitative assessment, we have applied Merqury [28], a reference-free assembly evaluation method comparing *k*-mer databases of a read set and the assembly. Using Illumina data of Altus, we have counted k-mers ($k = 21$) using Meryl v1.4 (https://github.com/marbl/meryl) and used the resulting *k*-mer database as input for Merqury, resulting in a QV value of 45.6677. This is similar to the QV value of C88 (46.6), while Sun et al. report a slightly better QV value of 51.7 for Otava. We also used Merqury to assess assembly completeness at different stages of our pipeline: For the phased result, i.e., taking into account only the fully phase-resolved parts of the haplotigs, completeness was estimated as 89.48%, while the result after the clustering—i.e., contigs that have been clustered into the chromosome, but might lack accurate phasing information due to insufficient *k*-mer information—is 97.93%. This value is similar to previously reported completeness scores of 97.3% for Otava and 99.05% for C88. The BUSCO scores per haplotype are 83.1%, 84%, 81.7%, and 81.7%, respectively.

To assess the accuracy of phasing, we generated ~77Gb of Pore-C data for Altus. Subsequently, we aligned all Pore-C reads to the assembly graph nodes using minimap2 and filtered for a mapping quality of 50. We then identified node pairs connected by at least one shared Pore-C read, considering their "coverage"—the number of Pore-C reads covering the node pair. Our evaluation of phasing results involves two criteria: Node pairs clustered to the same haplotype (cis pairs) ideally should be supported by Pore-C reads (true positives), while node pairs from different haplotypes (trans pairs) should lack Pore-C support (true negatives).

Parameters that influence the ability of Pore-C data to assess a node pair in this manner include the distance between the two nodes in the graph, node length, and Pore-C "coverage."

To quantify this, we conducted an analysis focusing on pairs of nodes longer than 1 Mb, applying various distance cutoffs (3 Mb, 5 Mb, 7 Mb, and 10 Mb) and Pore-C coverage cutoffs ranging from 0 to 70 (in steps of 1) and a number of higher cutoffs (100 to 1200). The corresponding true positive and false positive rates were computed, and the ROC curves are presented in Additional file 1: Fig. S29.

The area under the curve (AUC) for the different curves are 0.96, 0.94, 0.93, and 0.93 for distance cutoffs of 3 Mb, 5 Mb, 7 Mb, and 10 Mb, respectively. These results suggest that Pore-C data can effectively support the correctness of phasing.

For a parameter setting with a distance cutoff of 5 Mb and a Pore-C coverage cutoff of 30, for instance, the achieved true positive rate is 95.22%, and the true negative rate is 92.56%. A total of 251 cis and 309 trans node pairs meet this condition.

Even for node pairs further apart (up to 10 Mb), encompassing a larger number of pairs (502 cis, 706 trans), we maintain sensitivity and specificity > 0.9, even though the Pore-C signal expectedly declines with increasing distance. In summary, the Pore-C data robustly validates the haplotype concordance of our assemblies.

Discussion

We have developed a de novo assembly approach that uses accurate long reads and low-depth sequencing data from offspring samples to produce a phased assembly with haplotig lengths up to the length of chromosome arms. To achieve this, our method features multiple innovations. In particular, we designed a complete pipeline that uses haplotype-unique *k*-mers to chromosome sort and phase an assembly graph representing an autopolyploid genome. Importantly, this avoids intermediate steps that flatten the assemblies into contigs, instead resolving the haplotypes directly in the context of the graph topology, which might allow the unified integration of additional data types in the future.

The pseudo-chromosomes resulting from our assembly mapped well to the current monoploid reference genome, but we obtained ~3.8 times as much sequence data, which indicates comprehensive haplotype resolution. By using low-pass offspring sequencing, our approach is immediately accessible in breeding and research settings where a population of offspring and standard sequencing facilities are available. It avoids the need for single-cell pollen sequencing technology, which is an alternative route to assemblies of comparable quality [15]. In parallel to our developments, Bao et al. [16] have published a similar phasing strategy to the one presented by us, but with the distinction that it uses a selfing population as offspring. We therefore view the two methods as complementary, as they cater to different application settings.

Conclusions

Current limitations lie in the completeness of our assembly, which could still use improvement for some chromosomes. This is somewhat expected for two reasons: First, our method utilizes unique *k*-mers, which poses a challenge in genomes as repeat-rich as the potato genome. Second, we use sequencing data from an offspring of the cross Altus × Colomba, so those *k*-mers which are also present in the Colomba genome have to be discarded in order to make sure that only the signal from Altus is taken into account. However, our method presents a solution for practical data encountered in breeding research where offspring genotyping data is typically determined at low coverage or reduced representation for cost constraints. There, the practice of crossbreeding distinct cultivars is the norm and extensive data from self-crosses is typically not available without additional experimental work. Here we show that our method yields a reliable genome assembly tailored to such settings.

We believe our presented assembly will be a valuable contribution to the potato genomics community. As the number of published assemblies of different potato cultivars slowly increases [29, 30], it becomes evident that graph-based pangenomics studies, like recently demonstrated for human genomes [31], will become possible for tetraploid potatoes in the near future.

Despite the rapid advances in phased plant genome assembly, haplotype-resolved chromosome-level assemblies remain challenging for complex autopolyploid genomes. The complete resolution of a haploid human genome foreshadows this development and highlights the methodological advantage of working directly on assembly graphs [32]. To resolve the most recalcitrant genomic loci, ultra-long Oxford Nanopore Technologies (ONT) reads have been aligned to assembly graphs constructed from PacBio HiFi reads [33]. We envision that our approach will be combined with such additional data types in future studies. This is currently hampered by difficulties in the preparation of ultra-long sequencing reads (> 100 kb) for plant genomes and the read length N50s of ONT reads produced in our study are currently smaller (N50 ≤ 33.43 kb). But we anticipate the technical challenges will be overcome in the next few years. In our present HiFi-based graphs, shorter contigs tended to lack unique *k*-mers and 12% of the genome was part of such contigs. Mapping additional sequencing data such as ultra-long ONT reads to the graphs could help to bridge the remaining gaps, allowing the inclusion of further graph nodes in the haplotype sequences.

Methods

Data production

Plants were grown under greenhouse conditions at Heinrich Heine University Düsseldorf. After 3 weeks, young leaves from a single plant were harvested and immediately frozen in liquid nitrogen. Next, DNA was extracted from 1 g of frozen leaf material as previously described [34]. The DNA was size-selected using the Circulomics Short-Read Eliminator XL Kit (Circulomics Cat# SKU SS-100-111-01). DNA quality was assayed on a 1% agarose gel and using a NanoDrop Spectrophotometer (Thermo Fisher Scientific, USA). Sequencing libraries were prepared using the Oxford Nanopore Technologies (ONT) library preparation and sequencing kit SQK-LSK114, following standard protocols suggested by the manufacturer (Oxford Nanopore, UK). In brief, genomic DNA fragments were repaired and 3′-adenylated using the NEBNext FFPE DNA Repair Mix and the NEBNext Ultra II End Repair/ A-Tailing Module (New England Biolabs, USA). Sequencing adapters provided by ONT were then ligated using NEBNext Quick Ligation Module (NEB). After purifying the product with AMPure XP beads (Beckmann Coulter, CA, USA), libraries were loaded onto primed 10.4.1 Spot-On Flow Cells and sequenced using a PromethION sequencer (Oxford Nanopore Technologies, Oxford, UK) for 72 h. Basecalling was performed using Oxford Nanopore guppy software (v6.3.8) with "super" accuracy models resulting in 162 Gb of ONT reads passing the quality filter.

Restriction enzyme Pore-C libraries (RE-Pore-C) were prepared following the ONT Plant RE-Pore-C protocol (Oxford Nanopore Technologies), employing DpnII as the restriction enzyme [35]. After an overnight incubation, the enzyme was heat-denatured to facilitate the ligation of adjacent DNA clusters. Subsequent processes included protein degradation and decrosslinking, liberating chimeric Pore-C double-stranded DNA polymers. DNA quality and concentration were monitored using 1% agarose gel electrophoresis, a NanoDrop spectrophotometer (Thermo Fisher Scientific), and the Qubit DNA Assay Kit with a Qubit fluorimeter (Thermo Fisher Scientific). Genomic DNA fragments underwent repair, end repair, and A-tailing via NEBNext FFPE DNA Repair Mix (New England BioLabs Inc) and the NEBNext Ultra II End Repair/A-Tailing Module (New England BioLabs Inc). Afterwards, adapters were ligated and cleanup was performed (ONT ligation sequencing DNA V14 (SQK-LSK114) protocol). Resulting libraries were sequenced on R10.4.1 PromethION flow cells, with a runtime set to 100 h in accurate speed mode (260 base pairs per second). Flowcells were flushed and reloaded after 24, 48, and 72 h. In total, seven Runs were performed and basecalling was done using guppy v6.4.8 (ONT).

Dosage estimation in unitigs

The first step of our method is to compute the average coverage for each contig of the initial hifiasm assembly graph. This is done by aligning the HiFi reads with the contigs using minimap2. We only considered positions covering the unique sequence of the contig, meaning that overlaps to both neighboring contigs (if they exist) were not considered. The average coverage c_i of a contig i was then computed as the average read depth over all positions. We also computed the total average coverage m. Finally, we estimated the dosage d_i of contig i as follows:

$$d_i = d \text{ for } (d - 0.5)\mathrm{m} < \mathrm{c_i} \le (d + 0.5)\mathrm{m}, d \in \{1, 2, 3, 4\}.$$

For $c_i > 4.5m$, we assigned $d_i = 5$ to denote a repetitive contig. The same procedure is carried out using the ONT reads.

Connection of graph components

The clustering of unitigs into haplotype-resolved chromosome clusters involved two steps. First, we attempted to resolve the genome at the chromosome level. Chromosomes may feature several connected components plus additional singletons, so it was necessary to determine which components from the graph belong together. Second, we divided each chromosomal cluster into four distinct clusters, one for each haplotype.

We made use of the previously computed *k*-mer counts in the progeny to cluster unitigs with a similar *k*-mer count pattern. Our clustering procedure followed the idea that we can assign unitigs showing highly similar patterns to the same haplotype, whereas unitigs with opposing patterns are likely to be from the same chromosome but a different haplotype, and unitigs with seemingly unrelated count patterns are probably from different chromosomes.

The similarity between the *k*-mer count patterns of two nodes was assessed by computing the Spearman correlation coefficient (ρ). Two nodes with highly positively correlated *k*-mer count patterns should therefore reflect the same haplotype, whereas highly negative correlations would indicate that the nodes lie on distinct haplotypes. Only nodes from the same chromosome should be highly correlated (positively or negatively), whereas for nodes lying on two separate chromosomes, the *k*-mer counts should be unrelated and any similarities would occur by chance, resulting in low correlation coefficients.

We initially clustered the components and single unitigs into chromosome clusters by grouping all nodes showing the highest pairwise correlation coefficients ($\rho > 0.5$). These initial clusters were merged when the contigs therein were found to stem from the same graph components. This first clustering step yielded 12 large clusters that were defined as the corresponding chromosomal clusters or pseudo-chromosomes.

Clustering of unitigs based on similar k-mer patterns

To determine the individual haplotypes for each chromosome, we used the previously computed dosage estimation and started by clustering unitigs with dosage 1, because those nodes can only be assigned to a single cluster. We followed an agglomerative method that starts by building seed clusters with the highest correlations and then merges them into larger clusters as well as adding more nodes.

For each node n with dosage 1, we initially created one cluster for n containing only those unitigs with a high correlation to n ($\rho > 0.5$), producing a set of seed clusters. We then merged these clusters according to the number of common nodes they contain. To distinguish the different linkage groups, we made use of the high negative correlation between two nodes representing different haplotypes.

We created a negative edge between clusters c_i and c_j if there was one node pair (n_i, n_j), where $n_i \in c_i$ and $n_j \in c_j$, with a high negative correlation ($\rho_{i,j} < -0.3$). Conversely, we created a positive edge if at least one node pair (n_i, n_j) existed with a high positive correlation ($\rho_{i,j} > 0.5$). Two clusters c_i and c_j that were connected by a positive edge could be merged if no contradicting edge existed, such as a positive edge from c_i to another cluster c_k connected negatively to c_j.

After the merging steps, all nodes with the highest correlation to other nodes were assigned to clusters. Given that the subset of nodes not highly correlated to any other node ($\rho < 0.5$) was left out during this procedure, we included these remaining nodes by assigning them to a cluster c_i if the three best hits (the nodes n_a, n_b, and n_c with the highest correlation to n) all belonged to c_i. If this was not the case, we were unable to assign n unambiguously to a single cluster and it was left unclustered.

Finally, we assigned unitigs with higher dosages to the previously computed haplotype clusters. To cluster a node n with dosage x ($x \in \{2, 3, 4\}$), we computed the pairwise correlations between n and all nodes of all clusters c_1, c_2, c_3, and c_4 and added the node to the x clusters with the highest ratio of nodes that correlated positively with n.

Because of this rather conservative nature of our clustering procedure to avoid misassemblies, it may happen that a set of contigs is not added to a haplotype cluster and remains unphased. Therefore, we employed a post-processing step in which we reviewed the remaining unphased nodes and assigned them to the cluster with the highest correlation.

Assembly of clustered unitigs

Starting with the cluster of contigs for each chromosome, we reconstructed the ordering of contigs throughout the chromosomes to find all possible connections between them in order to create haplotypes with the greatest contiguity. First, we implemented the obvious extensions. If a phased node had only one neighbor in either direction, that neighbor was also considered to be phased. For simple bubble structures (four nodes, including source, sink, and two branching nodes) where both the source and the sink node were phased, one of the two branching nodes was assumed to be on the phasing path. If both branches lacked phase, no information was available to pick the correct one, so the node was chosen arbitrarily and the corresponding sequence was filled with placeholder characters instead of the node sequence to indicate the absence of correct haplotype sequence information.

We then considered the set of all phased nodes isolated from the rest of the graph. These formed a set of linear block structures, for each of which we were able to identify the two end nodes and recreate the node path (and therefore the sequence) through the block. To also reconstruct the order of these haplotype blocks, we then searched for paths between the end nodes of different blocks that solely contained unphased nodes. For blocks that could be connected uniquely to one additional block, we concatenated the two block sequences and again used placeholder characters for the length of the intervening unphased fragment. Finally, we resolved any remaining overlaps between the extended node paths, resulting in the final output sequences.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-023-03160-z.

Additional file 1: Fig. S1. Bandage visualization of the hifiasm raw unitig graph. **Fig. S2.** Dosage distribution of unitigs. **Fig S3.** Joint coverage of ONT reads and HiFi reads, mapped to the hifiasm assembly graph of Altus. **Fig S4.** UpSet plot of the various node sets used throughout this study. **Fig S5.** Genome size estimation using GenomeScope. **Fig S6.** Mapping of the contigs from all four haplotype clusters of chromosome 2 to the DMv6.1 reference. **Fig S7.** Haplotype-resolved version of (main) Fig. 4a. **Fig. S8.** Analysis of assembly errors in the hifiasm graph between chromosomes 10 and 12 of DMv6.1. **Fig. S9.** Mapping of the clusters to the Solyntus v1.1 reference sequence. **Fig. S10.** Analysis of assembly inserts between chromosomes 1 and 8 in the Solyntus v1.1 reference sequence. **Figs. S11-S16.** Synteny analysis between assembled haplotypes and DMv6.1. **Figs. S17-S28.** Synteny analysis between all four haplotypes per chromosome. **Fig. S29.** Pore-C data analysis for phasing validation: ROC curves for four different distance cutoffs.

Additional file 2: Review history.

Acknowledgements
We acknowledge the computational support provided by the Centre for Information and Media Technology (ZIM) at the Heinrich Heine University Düsseldorf. We thank Dr. Richard M Twyman for manuscript editing.

Review history
The review history is available as Additional file 2.

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.

Authors' contributions
R.S.M. and T.M. designed the assembly method, with input from S.S., R.F., B.U., and G.W.K. R.S.M. implemented the assembly method, performed the assemblies, and created the figures. R.S.M., S.S., R.F., G.W.K., B.U., and T.M. discussed and interpreted results. R.S.M., R.F., B.U., and T.M. wrote the manuscript, with input from S.S. and G.W.K. P.A. provided offspring sequencing data. M.HW.S. produced HiFi data and ran an initial assembly. F.M.R.Z. produced ONT data and wrote text describing data production. All authors read and approved the final manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL. We acknowledge funding provided by the German Research Foundation (DFG)—395192176 (to G.W.K. and T.M.)—as well as under Germany's Excellence Strategy—EXC 2048/1—390686111 (to G.W.K. and B.U.), by the German Ministry of Education and Research (BMBF) 031A536C (to B.U.), by the National Institutes of Health (NIH) R01HG011274 (to T.M.), and by the Dutch TKI top-sector projects "Genetics Assisted Assembly of Complex Genomes" (project number BO-68-001-033-WPR) and "LWV20.112 Application of sequence-based multi-allelic markers in genetics and breeding of polyploids" (project number BO-68-001-042-WPR).

Availability of data and materials
The offspring short reads are available via the NCBI BioProject under accession number PRJEB48582 [36]. Similarly, the HiFi reads for Altus are available under accession number PRJNA778192 [37]. The Oxford Nanopore data for Altus is available under the accession number PRJNA1049180 [38]. The Illumina reads of Altus are available via SRA under the accession numbers SRR14993639 and SRR14993640 [39].
The implementation of the workflow described herein is available as open-source code under the MIT license [40]. The version at the time of manuscript creation is archived under [41].

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 31 August 2022 Accepted: 27 December 2023
Published online: 19 January 2024

References

1. Kyriakidou M, Tai HH, Anglin NL, Ellis D, Strömvik MV. Current strategies of polyploid plant genome sequence assembly. Front Plant Sci. 2018;9:1660.
2. Petek M, Zagorščak M, Ramšak Ž, Sanders S, Tomaž Š, Tseng E, et al. Cultivar-specific transcriptome and pan-transcriptome reconstruction of tetraploid potato. Sci Data. 2020;7:249.
3. Devaux A, Kromann P, Ortiz O. Potatoes for sustainable global food security. Potato Res. 2014;57:185–99.
4. Birch PRJ, Bryan G, Fenton B, Gilroy EM, Hein I, Jones JT, et al. Crops that feed the world 8: Potato: are the trends of increased global production sustainable? Food Secur. 2012;4:477–508.
5. Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science. 2021:372. https://doi.org/10.1126/science.abf7117.
6. Porubsky D, Ebert P, Audano PA, Vollger MR, Harvey WT, Marijon P, et al. Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. Nat Biotechnol. 2021;39:302–8.
7. Motazedi E, Finkers R, Maliepaard C, de Ridder D. Exploiting next-generation sequencing to solve the haplotyping puzzle in polyploids: a simulation study. Brief Bioinform. 2018;19:387–403.
8. Pham GM, Hamilton JP, Wood JC, Burke JT, Zhao H, Vaillancourt B, et al. Construction of a chromosome-scale long-read reference genome assembly for potato. Gigascience. 2020:9. https://doi.org/10.1093/gigascience/giaa100.
9. van Lieshout N, van der Burgt A, de Vries ME, Ter Maat M, Eickholt D, Esselink D, et al. Solyntus, the new highly contiguous reference genome for potato (Solanum tuberosum). G3: Genes, Genom, Genet. 2020;G3(10):3489–95.
10. Berger E, Yorukoglu D, Peng J, Berger B. HapTree: a novel Bayesian framework for single individual polyplotyping using NGS data. PLoS Comput Biol. 2014;10:e1003502.
11. Xie M, Wu Q, Wang J, Jiang T. H-PoP and H-PoPG: heuristic partitioning algorithms for single individual haplotyping of polyploids. Bioinformatics. 2016;32:3735–44.

12. Siragusa E, Haiminen N, Finkers R, Visser R, Parida L. Haplotype assembly of autotetraploid potato using integer linear programing. Bioinformatics. 2019;35:4534.
13. Schrinner SD, Mari RS, Ebler J, Rautiainen M, Seillier L, Reimer JJ, et al. Haplotype threading: accurate polyploid phasing from long reads. Genome Biol. 2020;21:252.
14. Zhang W, Luo C, Scossa F, Zhang Q, Usadel B, Fernie AR, et al. A phased genome based on single sperm sequencing reveals crossover pattern and complex relatedness in tea plants. Plant J. 2021;105:197–208.
15. 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; https://doi.org/10.1038/s41588-022-01015-0.
16. 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.
17. 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.
18. Li H. Minimap2: pairwise alignment for nucleotide sequences [Internet]. Bioinformatics. 2018:3094–100. https://doi.org/10.1093/bioinformatics/bty191.
19. Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, et al. De novo assembly of haplotype-resolved genomes with trio binning. Nat Biotechnol. 2018; https://doi.org/10.1038/nbt.4277.
20. Logsdon GA, Vollger MR, Hsieh P, Mao Y, Liskovykh MA, Koren S, et al. The structure, function and evolution of a complete human chromosome 8. Nature. 2021;593:101–7.
21. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27:764–70.
22. Quigley S, Damas J, Larkin DM, Farré M. syntenyPlotteR: a user-friendly R package to visualize genome synteny, ideal for both experienced and novice bioinformaticians. Bioinform Adv. 3(1):vbad161. https://doi.org/10.1093/bioadv/vbad161.
23. Goel M, Sun H, Jiao W-B, Schneeberger K. SyRI: identification of syntenic and rearranged regions from whole-genome assemblies. bioRxiv. 2019;1 Available from: https://scholar.archive.org/work/bxysy2kf2na6rnyshmkfygzply/access/wayback/https://www.biorxiv.org/content/biorxiv/early/2019/02/11/546622.full.pdf
24. Alonge M, Lebeigle L, Kirsche M, Jenike K, Ou S, Aganezov S, et al. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. Genome Biol. 2022;23:258.
25. Goel M, Schneeberger K. plotsr: visualizing structural similarities and rearrangements between multiple genomes. Bioinformatics. 2022;38:2922–6.
26. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol. 2021;38:4647–54.
27. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A. 2020;117:9451–7.
28. Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 2020;21:245.
29. Tang D, Jia Y, Zhang J, Li H, Cheng L, Wang P, et al. Genome evolution and diversity of wild and cultivated potatoes. Nature. 2022;606:535–41.
30. Hoopes G, Meng X, Hamilton JP, Achakkagari SR, de Alves Freitas Guesdes F, Bolger ME, et al. Phased, chromosome-scale genome assemblies of tetraploid potato reveal a complex genome, transcriptome, and predicted proteome landscape underpinning genetic diversity. Mol Plant. 2022;15:520–36.
31. Liao W-W, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, et al. A draft human pangenome reference. Nature. 2023;617:312–24.
32. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. bioRxiv. 2021;376(6588):44–53. Available from: https://www.biorxiv.org/content/10.1101/2021.05.26.445798v1.full?utm_source=Nature+Briefing&utm_campaign=daa292b748-briefing-dy-20210601&utm_medium=email&utm_term=0_c9dfd39373-daa292b748-44906129
33. Rautiainen M, Marschall T. GraphAligner: rapid and versatile sequence-to-graph alignment. Genome Biol. 2020;21:253.
34. Pucker B. Plant DNA extraction and preparation for ONT sequencing v1. 2020. https://doi.org/10.17504/protocols.io.bcvyiw7w.
35. Deshpande AS, Ulahannan N, Pendleton M, Dai X, Ly L, Behr JM, et al. Identifying synergistic high-order 3D chromatin conformations from genome-scale nanopore concatemer sequencing. Nat Biotechnol. 2022;40:1488–99.
36. Kodde L, Arens P, Finkers R. Sequence data solanum tuberosum "Altus" x "Columba." Datasets. Sequence Read Archive. Available from: https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJEB48582
37. Usadel B. Phased sequencing of the altus potato cultivar. Datasets. Sequence Read Archive. Available from: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA778192/
38. Ziegler FMR, Usadel B. Oxford Nanopore data of the Altus potato cultivar. Datasets. Sequence Read Archive. Available from: https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA1049180
39. Potato PanGenome Consortium. WGS of Solanum tuberosum: Altus Paired-End Reads 470bp. Datasets. Sequence Read Archive. Available from: https://www.ncbi.nlm.nih.gov/sra/SRR14993639
40. Serra Mari R. Polyploid potato assembly [Internet]. Github. Available from: https://github.com/rebeccaserramari/polyploid-potato-assembly
41. Serra Mari R. Polyploid potato assembly [Internet]. zenodo. Available from: https://doi.org/10.5281/zenodo.10160515

Publisher's Note