Bioinformatics, 2021, 1–8
doi: 10.1093/bioinformatics/btab068
Advance Access Publication Date: 4 February 2021
Original Paper

OXFORD

Genome analysis

Accurate haplotype-resolved assembly reveals the origin of structural variants for human trios

Mengyang Xu [1,2,†], Lidong Guo[1,3,†], Xiao Du[1,2], Lei Li[1,4], Brock A. Peters[2,5], Li Deng[1], Ou Wang[2], Fang Chen[6], Jun Wang[1], Zhesheng Jiang[1], Jinglin Han[1], Ming Ni[1,2], Huanming Yang[2], Xun Xu[2], Xin Liu[1,2,7,*], Jie Huang [8,*] and Guangyi Fan[1,2,7,*]

[1]BGI-QingDao, Qingdao 266555, China, [2]BGI-Shenzhen, Shenzhen 518083, China, [3]BGI Education Center, University of Chinese Academy of Sciences, Shenzhen 518083, China, [4]School of Future Technology, University of Chinese Academy of Sciences, Beijing 101408, China, [5]Complete Genomics Inc., 2904 Orchard Pkwy, San Jose, CA 95134, USA, [6]MGI, BGI-Shenzhen, Shenzhen 518083, China, [7]State Key Laboratory of Agricultural Genomics, BGI-Shenzhen, Shenzhen 518083, China and [8]National Institutes for food and drug Control (NIFDC), No. 2 Tiantan Xili, Dongcheng District, Beijing 10050, China

*To whom correspondence should be addressed.
[†]The authors wish it to be known that these authors contributed equally.
Associate Editor: Janet Kelso

Received on August 27, 2020; revised on December 7, 2020; editorial decision on January 25, 2021; accepted on January 28, 2021

Abstract

Motivation: Achieving a near complete understanding of how the genome of an individual affects the phenotypes of that individual requires deciphering the order of variations along homologous chromosomes in species with diploid genomes. However, true diploid assembly of long-range haplotypes remains challenging.

Results: To address this, we have developed Haplotype-resolved Assembly for Synthetic long reads using a Trio-binning strategy, or HAST, which uses parental information to classify reads into maternal or paternal. Once sorted, these reads are used to independently *de novo* assemble the parent-specific haplotypes. We applied HAST to cobarcoded second-generation sequencing data from an Asian individual, resulting in a haplotype assembly covering 94.7% of the reference genome with a scaffold N50 longer than 11 Mb. The high haplotyping precision (~99.7%) and recall (~95.9%) represents a substantial improvement over the commonly used tool for assembling cobarcoded reads (Supernova), and is comparable to a trio-binning-based third generation long-read-based assembly method (TrioCanu) but with a significantly higher single-base accuracy [up to 99.99997% (Q65)]. This makes HAST a superior tool for accurate haplotyping and future haplotype-based studies.

Availability and implementation: The code of the analysis is available at https://github.com/BGI-Qingdao/HAST

Contact: liuxin@genomics.cn or jhuang5522@126.com or fanguangyi@genomics.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Determining the combination of allelic variants along each homologous chromosome, haplotyping, provides more information than nonhaplotype-resolved genetic variations for diploid or polyploid species. Population-variation-based haplotype inference has been used for many years (Hill, 1974; O'Connell *et al.*, 2016), but the direct determination of an individual's haplotype has been challenging. More recently, whole genome sequence data have been applied to reconstructing haplotypes. By aligning the sequencing data to the reference genome, reads harboring two or more heterozygous variants can be used to capture allelic variations (alignment-based haplotyping) (Snyder *et al.*, 2015). Despite the high efficiency of this approach, especially when using long reads (Edge *et al.*, 2017) or cobarcoded (Peters *et al.*, 2014) long fragment reads (Peters *et al.*, 2012) (LFR), alignment-based haplotyping relies on the reference genome and variation calling. This makes it difficult to phase haplotypes that include structural variations (SVs) or sequences not found in the reference genome. One solution to this problem would be to *de novo* assemble the complete sequence of each haplotype (Myers, 2005). For example, during graph-based genome assembly, handling the variation-induced 'bubble' structures can facilitate the

construction of haplotypes (Chin *et al.*, 2016; Weisenfeld *et al.*, 2017). Although assembly-based haplotyping performs better in regions with SV (Zhang *et al.*, 2020), the overall efficiency and precision underperforms that of alignment-based haplotyping.

To improve upon this, trio-binning-based strategies have been developed (Eberle *et al.*, 2017; Marchini *et al.*, 2006). These strategies use heterozygous variants not shared between the two parents to separate the child's sequencing data into maternal and paternal groups. With the read data classified, genome assembly is more efficient and results in improved haplotype contiguity and precision (Koren *et al.*, 2018; Low *et al.*, 2020). However, it is difficult for haplotype-specific *k*-mers to capture heterozygous alleles in all three individuals in the trio without local phasing information (Garg *et al.*, 2018).

Recently, cobarcoding-based strategies that retain long DNA information along with cost-effective, highly accurate short reads, such as MGI's single-tube long fragment read (Wang *et al.*, 2019) (stLFR) and 10× Genomics' linked-read (Zheng *et al.*, 2016) technologies, have been successfully used in alignment-based haplotyping (Wang *et al.*, 2019), SV detection (Bishara *et al.*, 2015; Zhou et al., 2019) and *de novo* genome assembly (Kuleshov *et al.*, 2016).

In this study, we describe Haplotype-resolved Assembly for Synthetic long reads using a Trio-binning strategy (HAST, https://github.com/BGI-Qingdao/HAST), which is the first trio-binning-assembly-based haplotyping tool for cobarcoded reads. The principle of this pipeline is to reconstruct two haplotype-resolved assemblies with high precision and continuity based on the combination of global haplotyping information from trios and local phasing information from barcodes. HAST first identifies haplotype-specific *k*-mers using sequencing data from the parents, then employs these markers to partition the offspring cobarcoded sequencing data into haplotypes, and finally assembles each haplotype independently. Using HAST to assemble haplotypes of an Asian individual, we generated results comparable to trio-binning-assembly-based haplotyping using long third generation reads (TrioCanu) (Koren *et al.*, 2018).

2 Materials and methods

2.1 Sample preparation and sequencing of the HJ trios

We sequenced a Han Chinese volunteer (Research ethics ID: XHEC-C-2019-086, HJ) using the MGIEasy stLFR Library Prep Kit on the MGISEQ-2000 platform (DNBSEQ-G400, RRID: SCR_017980) and generated a total of 632 Gb of read data. We applied SOAPfilter (Luo *et al.*, 2012) (version 2.2) to remove possible adapter contamination, low quality reads, duplicated reads and PE reads with a short insert size (<300 bp). This reduced the amount of data to 355 Gb. Both of HJ's parents were sequenced to ∼30× coverage using massively parallel sequencing (MPS) short reads (Rogers and Venter, 2005). From this data, haplotype *k*-mers were extracted. In addition, to construct haplotypes for comparison with HAST results, PacBio SMRT libraries were sequenced on a Sequel instrument (PacBio Sequel System, RRID: SCR_017989). In total, we generated 138 Gb PacBio high-fidelity circular consensus sequence (CCS) long reads.

2.2 Generation of parental unique markers

The classification of parent-specific *k*-mers is similar to the method described in TrioCanu (Koren *et al.*, 2018). Due to its high speed and low requirement of memory, Jellyfish (Marcais and Kingsford, 2011) (version 0.6.1; Jellyfish, RRID: SCR_005491) was chosen to generate, count and output distinct *k*-mers in the parental genomes. The haplotype-specific *k*-mers were identified by mixing 1 copy of paternal *k*-mers with 2 copies of maternal and then counting the total frequencies. The *k*-mers that occurred exactly once in the mixture were identified to belong to the paternal group, while those that occurred twice were maternal-specific.

The majority of unshared *k*-mers originated from the sequencing errors and as such were present in low copy numbers. In addition, there were a number of *k*-mers with excessively high coverage due to repetitive regions of the genome. To reduce the computational load from these unuseful *k*-mers, we limited the parental *k*-mer library size based on the coverage distribution. For the plot of *k*-mer frequency f against coverage c, the low coverage range C_{low} was determined based on the first lowest point of the profile C_1 and the highest point of the main peak C_2 (Supplementary Fig. S1). The profile can be fitted by a mixture model of negative binomial model terms with a long tail in the high-coverage region (Vurture *et al.*, 2017). We simplified the variances in the tail region and defined the high-coverage range C_{high} as twice of C_{low}. It can also be calculated by C_1 and C_2

$$C_{high} = 3 \times C_2 - 2 \times C_1$$

Only those *k*-mers within the thresholds were exported to identify long fragments in the next step. Note that the numbers of unshared *k*-mers for two haplotypes are not necessarily equal.

The performance of the classification also substantially depends on the *k* value. Theoretically, a read has a greater possibility of matching heterozygous markers by using a smaller *k*-mer size. However, too small of a *k* value can cause random *k*-mer collisions, especially for large genomes. According to the formula of *k*-mer collision, for a rate r given a random distribution of *k*-mers in the genome G (Fofanov *et al.*, 2004) ,

$$1/r = 1 + (4^k/G)$$

a *k* of 19 provides a collision rate of 1% for the human genome, while $k = 21$ and 31 reduce r below 0.1% and $1e^{-9}$, respectively.

2.3 Partition of stLFR long fragments

The concept of read binning prior to assembly has been used in long reads (Koren *et al.*, 2018) and metagenomics (Wu and Ye, 2011). Ideally, cobarcoded reads can be categorized if they possess individual *k*-mers from one haplotype and none from others. Assuming that heterozygous sites and sequencing errors are randomly distributed in the genome and there are no overlaps between long DNA fragments or between reads sharing the same barcode, the expected number of unshared *k*-mers, p in one long DNA fragment can be derived as

$$p = 2 \cdot h \cdot k \cdot N\ (l\ -\ k\ +\ 1)(1\ -\ e)^k$$

where h is the genome heterozygosity, N is the number of read pairs in one long DNA fragment, l is the read length and e is the sequencing error rate. The read clustering based on barcodes overcomes the limit of short-read length, especially for genomes with low heterozygosity. The high base-calling accuracy also improves the sensitivity and precision of *k*-mer mapping. Given $h = 0.1\%$, $N = 20$, $l = 100$, $k = 21$, $e = 0.1\%$ for a typical stLFR library of a human sample, the expected number of unshared *k*-mers is larger than 65 for each barcode.

In the case of trios, long DNA fragments are assigned to the paternal group if they only own paternal-specific *k*-mers and vice versa. If both types of parental *k*-mers match different portions of the same DNA fragment, then the proper partitioning will be determined by which parental genome has a higher probability to be mapped by the fragment. The probability is first normalized to the parent-specific *k*-mer libraries with different sizes, and then multiplied by a correction factor of sex chromosome size variance. The factor attempts to neutralize the inherent discrepancy of parents due to the constitution of sex chromosome (Bachtrog and Charlesworth, 2001). The remaining DNA fragments with equal *k*-mer counts are discarded as errors, while those with no *k*-mers are regarded as homozygous. Note that each DNA fragment is identified by its barcode, which corresponds to all fragments in the same cobarcoding compartment (Wang *et al.*, 2019). To accelerate millions of *k*-mer queries in the partitioning, we binarized the *k*-mer characters, hashed the *k*-mer database and parallelized the classifying procedure. Note that HAST (version 1.0.0; HAST, RRID: SCR_018247) also accepts 10× Genomics reads, PacBio and ONT long reads, and shows significant improvements in speed and memory efficiency

relative to the trio-binning scripts in the published TrioCanu (Koren *et al.*, 2018) (Supplementary Table S19). We notice that the latest version of Canu (Koren *et al.*, 2017) has also optimized its thread and memory consumption.

2.4 Hast assembly pipeline

HAST utilizes parental variation information to facilitate a completely haplotype-resolved diploid assembly in the following three steps (Fig. 1): (a) generation of parental unshared *k*-mers from MPS reads and sequencing the child with cobarcoded reads, (b) determination of the parent of origin of each of the child's long fragments (the set of cobarcoded reads sharing the same barcode and representing a single long DNA fragment) based on the parental unshared *k*-mer sets, and finally (c) haplotype-resolved assemblies of both parental-inherited chromosomes.

The global haplotyping information from the trio straightforwardly partitioned reads with haplotype-specific *k*-mers. Meanwhile, reads without global markers could also be grouped if they share the same barcode with adjacent alleles that are trio-resolved due to the long-range information. All DNA fragments in paternal or maternal groups along with homozygous fragments were transformed into 10× Genomics data format, and passed to Supernova (Weisenfeld *et al.*, 2017) (version 2.1.1; Supernova assembler, RRID: SCR_016756) to assemble. The assembly graph was simplified by barcode information to resolve repeat-induced junctions.

Note that the direct Supernova output for both groups still retains phasing errors. For errors due to inaccurate base calling, we recovered the bubble structures in the assembly graph, and used the haplotyped-specific *k*-mers again to determine which arm truly represented the allelic variants. For phase variants which are heterozygous in all three individuals in the trio, their corresponding reads were identified as homozygous and used for both assemblies if they have no adjacent marked alleles sharing the same barcode. We also distinguished two arms of each bubble to determine if they belong to heterozygous variants in all three individuals, and kept both allelic sequences in the final output.

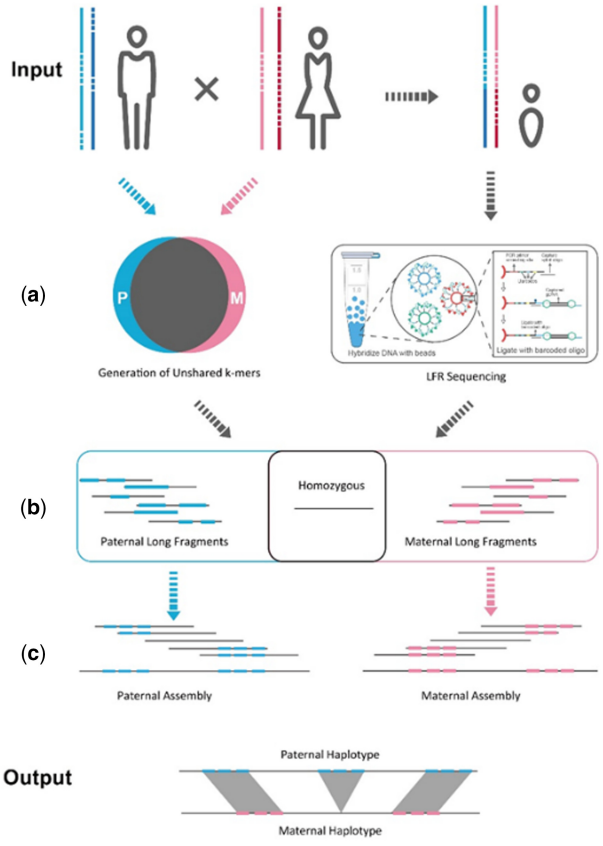

Fig. 1. A schematic of HAST workflow. (**a**) Generation of parental unshared *k*-mers from MPS reads and sequencing the child with cobarcoded reads, (**b**) determination of the parent of origin of each of the child's long fragments based on the parental unshared *k*-mer sets, and finally (**c**) haplotype-resolved assemblies of both parental-inherited chromosomes. The output includes two complete haplotypes. The inserted illustration of cobarcoded technique in (a) is from the reference (Wang *et al.*, 2019)

2.5 Haplotyped-resolved assembly for CCS data

We used the TrioCanu module implemented in Canu (version 2.0; Canu, RRID: SCR_015880) to classify the PacBio CCS long reads with an average read length of 10.2 kb. The identified ratio was 34.3% and 34.8% for paternal and maternal, respectively (Supplementary Table S5). To compare with HAST, the haplotypes were individually assembled by Canu using corresponding haplotype-specific long reads and unassigned long reads without any markers.

2.6 Validation of haplotyping effects

Th hg19 assembly (Church *et al.*, 2011) was used as the reference for *Homo sapiens*. The QUAST (Gurevich *et al.*, 2013) (version 5.0.2; QUAST, RRID: SCR_001228) was used with default parameters to report the assembly statistics including total length, scaffold N50, contig N50, as well as scaffold NGA50, contig NGA50, genome fraction, misassemblies and local misassemblies based on the mapping relation with the reference. Note that QUAST splits scaffolds as contigs when there is a continuous stretch of N's of length ≥10. In addition, Merqury (Rhie *et al.*, 2020) (version 1.0) was applied for the evaluation of haplotyping precision and recall according to the reliable specific *k*-mers from the child's and parental reads. In total, we generated 18.6 Mb paternal hap-mers and 16.3 Mb maternal hap-mers to evaluate. To quantify the effect of HAST on the downstream bioinformatics analysis, we ran BUSCO (Simao *et al.*, 2015) (version 3.0.2; BUSCO, RRID: SCR_015008) analysis for all the human assemblies against the vertebrata_odb9.

3 Implementation

3.1 Assignment of long fragments into haplotype groups

The effectiveness of grouping the raw sequencing data into their specific haplotypes is important for an assembly-based haplotyping method. The MPS datasets of ∼30× coverage per parent were used to generate unique markers (unshared *k*-mers) to partition the child's cobarcoded reads to separate haplotypes. The accurate selection of markers depends on the sequencing error rate and influences the accuracy of partitioning. In this study, we first applied 21-mers to partitioning those unassigned reads, and then used 19- and 31-mers to enhance the efficiency. Based on the *k*-mer distribution, 21-mers that occurred <9 times (sequencing errors) or >58 (high-frequency duplications) were removed from the dataset (Supplementary Fig. S1) resulting in ∼40 M unique markers for maternal and ∼60 M for paternal (Supplementary Table S1). Due to the relatively short-read length (100 bp), only ∼1.4% of reads had at least one unshared *k*-mer (Supplementary Table S2). Nonetheless, stLFR barcodes enabled the clustering of short reads and extended the haplotype information to the entire long fragment. There were total of 155 196 643 barcodes for two stLFR libraries in the child's clean reads, of which 23.8% (correspondingly 58.3% read pairs) were uniquely assigned into haplotypes.

To investigate the grouping effectiveness, we calculated the assigned ratio for four groups: paternal, ambiguous, maternal and homozygous. We observed that all the ratios showed a clear dependence of read-pair number sharing the same barcode. The ratio of cobarcoded read groups with no haplotype-specific *k*-mers were exponentially decreasing from 93% to 0% with the increasing number of read pairs per barcode, while the assigned barcode numbers were evenly growing (Fig. 2a). This suggested that the distinguishable

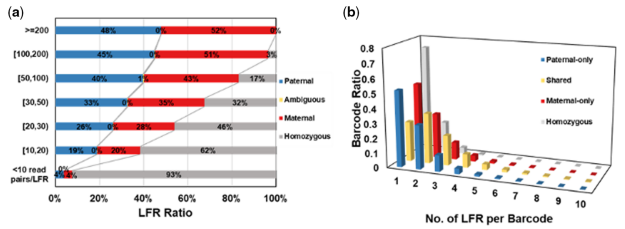

Fig. 2. Statistics for the clustering of cobarcoded reads. (**a**) The assigned ratio of barcodes for different numbers of read pairs per barcode. The assigned ratio is plotted on the *x*-axis and consists of four groups: paternal or maternal (contain more paternal or maternal-specific *k*-mers than the other), ambiguous (have equal numbers of both types of haplotype-specific *k*-mers) or homozygous (have no haplotype-specific *k*-mers). Barcodes with a smaller number of reads (<10 read pairs) are likely to be from shorter DNA fragments and result in mostly homozygous assignments. As the number of cobarcoded read pairs per barcode increases (likely due to longer DNA fragments being captured on the stLFR beads), the majority of these barcodes are assigned to a haplotype-specific group. The ambiguous barcodes are likely to be caused by multiple DNA fragments being captured on a single stLFR bead, but they are relatively infrequent in our data. (**b**) The stLFR collision rate for four groups: paternal-only or maternal-only (only contain at least one paternal or maternal-specific *k*-mer), shared (have both types of haplotype-specific *k*-mers) or homozygous (have no haplotype-specific *k*-mers). The *x*-axis represents the number of DNA fragments per barcode, and *y*-axis represents the ratio of barcodes for each group to the total barcode number. Not surprisingly, the ambiguous barcodes have a peak at 2, indicating a higher possibility of collision

barcodes had more read pairs and the unassigned barcodes contained DNA fragments that were of a shorter overall length and lacked heterozygous variants. When we filtered out the short fragments (i.e. <20 read pairs per barcode), there were total of 31.0% and 32.7% of barcodes (42.0% and 44.1% of read pairs) classified as inherited from the child's father and mother, respectively. This demonstrates the feasibility and high efficiency of our trio-binning strategy using stLFR data.

3.2 Collision rates of cobarcoded read groups

The haplotyping precision of cobarcoded reads is dependent on how often an stLFR barcode is associated with long DNA fragments from disparate parts of the genome (the stLFR collision rate). To better understand the stLFR collision effect, all the cobarcoded reads were mapped against the human reference genome (hg19). For the child's stLFR library the overall collision rate was 1.47, similar to what as previously been reported (Wang *et al.*, 2019). The barcode assignment was used to further differentiate the collision rate. The group of paternal-only, maternal-only, shared and homozygous had a collision rate of 1.58, 1.54, 2.12 and 1.19, respectively (Supplementary Table S4).

The distribution of long fragment counts per barcode shows that the majority of paternal-only, maternal-only and homozygous barcodes have only one long fragment, while the shared barcode group exhibits a peak at two (Fig. 2b). The shared barcodes can also be classified based on the relative marker ratio to tolerate mis-mapped *k*-mers by possible sequencing errors. Although this data tends to introduce defective long-range information, the overall amount is insignificant thanks to the high base-calling accuracy and low LFR collision rate of short-read-based cobarcoding techniques. Moreover, this spurious information can be avoided during the assembly process by considering the fact that the stLFR collision rate randomly occurs among the long fragments and thus two fragments from disparate parts of the genome are practically never found repeated on another barcode.

3.3 Complete reconstruction of individual haplotypes

Using HAST, we individually assembled the classified stLFR data from an Asian male (HJ) to reconstruct haplotypes, and compared these to direct pseudo-haplotyping assembly that did not use trio-binning information. The HAST generated assembly was entirely in phase and the origin of each sequence was unambiguous, while the pseudo-haplotype assembly arbitrarily combined one arm of each 'bubble' structure in the graph resulting in many incorrectly linked haplotypes. Although it was complicated to maintain the long phasing information due to the relatively sparse heterozygous variant sites in humans (~0.1%), HAST was able to cluster most of the fragments based on the long-range cobarcode information and successfully assemble haploid chromosomes. For the genomes with higher heterozygosity, this complexity would decrease.

HAST generated a 6.0 Gb assembly of HJ with a scaffold N50 of 18.3 and 11.4 Mb for the maternal and paternal portions, respectively (Supplementary Table S7). The longer scaffold length of the maternal assembly is likely due to the presence of large *X* chromosome. The contig N50 is about 60 kb. The alignment to the human reference genome resulted in 94.95% and 94.40% coverages with a scaffold NGA50 > 1 Mb. This suggests the assembly is highly accurate and complete. In contrast, the pseudo-haplotype assembly without trio binning produced a comparable scaffold N50 (16.7 Mb) but a shorter contig N50 (38 kb). The base-level accuracy (quality value, QV) and the reappearance capacity of reliable *k*-mers from input reads (completeness) were estimated for each assembly. The HAST assembly showed a slightly higher average QV value (Q63versus Q61) with similar completeness (~97%) to the pseudo-haplotype assembly (Table 1).

The contiguity metrics were reported by QUAST. The assembly quality and haplotyping were evaluated by Merqury, including *k*-mer-based QV, assembly completeness and haplotyping precision and recall. The QV was calculated using the following equation: $QV = -10\log_{10}E$, where E refers to the single-base-level assembly errors. The haplotyping estimates were given by the reliable haplotype-specific *k*-mers. BUSCO v3 was run with vertebrata_odb9 database, where Comp., Complete; Dup., Duplicated; Frag., Fragmented and Mis., Missing.

We also examined HAST effect using other two pilot human genomes HG001/NA12878 and HG002/NA24385. Compared with HJ, the diploid assemblies showed average scaffold N50s of 4.3 and 5.7 Mb due to relatively shorter DNA fragment length but larger average contig N50s of 123 and 93 kb for two datasets, respectively (Supplementary Table S11). The single-base accuracy reached Q66 and Q62 with a similar completeness (~99%) (Supplementary Table S9).

3.4 Assembly haplotype validation

As a sanity check, we applied Merqury (Rhie *et al.*, 2020), a reference-free phasing assessment tool based on the *k*-mer analysis, to haplotype evaluation. We chose this method because the direct alignment to the human reference does not provide phasing information and currently there are no other assemblies available for HJ or HJ's parents. By counting the number of expected parent-specific *k*-mers present in the child's diploid assembly, it demonstrated that each haplotype recovered >95% of the parental heterozygous sites (Table 1). This was significantly higher than the pseudo-haplotypes without trio binning, which recovered only 64.91% and 54.05% of the paternal and maternal alleles, respectively (Table 1).

The haplotyping error is related to the presence of unexpected haplotype-specific *k*-mers in the assembly. The haplotype precisions of the HAST paternal and maternal assemblies were 99.42% and 99.94%, respectively. In contrast, it was only 64.09% or 44.05% for Supernova assembled pseudo-haplotypes. The haplotype recall and precision rates achieved 95.75% and 99.89% for HG001 and 91.81% and 99.81% for HG002 (Supplementary Table S9 and Fig. S2). Moreover, the HAST haplotypes contained fewer fragmented or missing BUSCO genes compared to the pseudo-haplotypes and achieved 94.8% complete genes. This suggests that there are more allelic variations improperly assembled in the pseudo-haplotypes.

In the *k*-mer multiplicity copy-number plots (Fig. 3a), the first small 1-copy peak (red) represents *k*-mers unique to each parental haplotype, the second large 2-copy peak (green) corresponds to sequences shared between both haplotypes or haplotype-specific duplications and higher copies (blue, purple, orange) are repetitive regions. Those *k*-mers that only occur in the sequence reads (grey) possibly came from sequencing errors or indicated missing genomic

Table 1. Haplotype-resolved assembly quality statistics for HJ

Assembly	Contiguity				Quality		Haplotyping			BUSCO (%)			
	Scaffold Max (Mb)	Scaffold N50 (Mb)	Contig Max (Mb)	Contig N50 (Mb)	QV (Phred)	Completeness (%)	Precision (%)	Recall (%)	F1-score (%)	Comp.	Dup.	Frag.	Mis.
HAST													
Paternal	59.7	11.4	0.61	0.06	61.83	96.61	99.42	95.83	97.60	91.2	1.4	5.2	3.6
Maternal	70.9	18.3	0.51	0.06	65.13	96.83	99.94	96.02	97.94	91.7	1.3	4.9	3.4
Combined	70.9	14.9	0.61	0.06	63.18	99.13	99.67	95.92	97.76	94.8	94.4	2.5	2.7
Supernova													
Pseudohap1	105.1	16.7	0.34	0.04	60.62	97.10	61.09	64.91	62.94	87.2	1.4	8.3	4.5
Pseudohap2	105.1	16.7	0.34	0.04	60.46	97.10	44.05	54.05	48.54	87.0	1.4	8.2	4.8
Combined	105.1	16.7	0.34	0.04	60.54	98.23	52.51	59.83	55.93	87.6	87.1	8.1	4.6
TrioCanu													
Paternal	—	—	15.0	3.06	45.47	95.49	99.94	97.95	98.94	93.6	3.6	3.2	3.2
Maternal	—	—	20.8	4.56	47.17	97.49	99.97	97.41	98.67	94.6	4.0	2.9	2.5
Combined	—	—	20.8	3.81	46.24	99.46	99.95	97.70	98.81	95.8	95.6	2.0	2.2

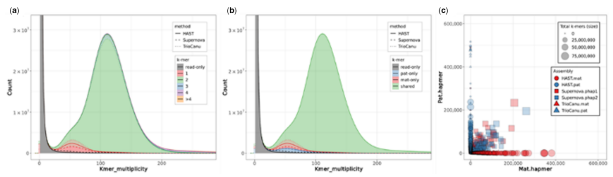

Fig. 3. Haplotype *k*-mer spectra for the diploid HJ genome. (**a**) The stacked copy-number spectra of *k*-mer multiplicity for evaluating assembly errors and heterozygous recall. The error bars next to the zero multiplicity indicate mis-assembled *k*-mers absent from sequencing reads. (**b**) The stacked haplotyping assembly spectra of *k*-mer multiplicity for evaluating haplotyping recall and balance. (**c**) Hap-mer plots for evaluating haplotyping contiguity and precision, where the 'hap-mer' refers to the haplotype-specific *k*-mer of the genome (Rhie *et al.*, 2020). The blob size is proportional to the scaffold/contig size, and the position is plotted based on the number of contained paternal (*y*-axis) and maternal (*x*-axis) hap-mers. Solid lines and rounds are generated by HAST-assembled stLFR data, dashed lines and squares by Supernova-assembled stLFR data and dotted lines and triangles by TrioCanu-assembled PacBio CCS data

regions that could not be assembled. HAST and Supernova had comparable assembly completeness and quality. The >Q60 single-base-level QV was evident in the *k*-mer analysis as a relatively small bar at the zero multiplicity which corresponded to *k*-mers found in the assembly but absent from the reads. We also analyzed the *k*-mer spectra of each haplotype individually. In the HAST assembly, the paternal (blue) and maternal (red) haplotypes generated approximately identical numbers of *k*-mers in the heterozygous peak (Fig. 3b). This was also true for the Supernova assembly, where the arbitrary combination of 'heterozygous sites' resulted in equal numbers of *k*-mers for each pseudo-haplotype, but the total number of classified allelic variations was lower by 37%. The fraction of shared *k*-mers (green) is expected to be larger, given that the human heterozygosity is only ~0.1%, of which the stacked portion at heterozygous peak corresponds to the heterozygous variants shared by both parents. The hap-mer blob plots (Fig. 3c) describe the contiguity and precision of phased scaffolds. The HAST assembly generated near perfect separation of haplotype-specific *k*-mers across all scaffold lengths (round). On the contrary, numerous scaffolds in each pseudo-haplotype assembled by Supernova shared both paternal and maternal-specific *k*-mers (square). Only short sequences in the Supernova assembly had properly separated *k*-mers and suggested that pseudo-haplotypes could not reconstruct the true homologous chromosomes (Table 1).

The long-range information from the cobarcoded read data improves the contiguity and precision of phase blocks in the assemblies. Incorrect grouping of haplotype-specific variants introduces switch errors, which split the scaffold or contig and reduce the block size. Using the advantage of parental information, the phase block length N50s were up to 11.6 and 18.3 Mb for HAST paternal and maternal haplotypes, respectively, with the longest block reaching the chromosome level (69.2 Mb). Importantly, these haplotype N50 values were close to the scaffold N50s (Tables 1 and 2). By contrast, there was a 10-fold decrease in the contiguity from scaffolds to phase blocks for the Supernova pseudo-haplotypes. In addition, the HAST haplotypes covered almost the entire genome and did so with switch error rate roughly 100-fold less than the pseudo-haplotypes. The HAST assemblies of HG001 and HG002 also showed large phase block sizes close to scaffolds with low switch error rates (Supplementary Table S10).

The accurate reconstruction of haplotypes is not limited to stLFR format data. We also applied HAST to a 10× Genomics format dataset for HG001. HAST partitioned 10× Genomics barcodes and read pairs into haplotype-specific and homozygous groups for assembling. However, there were up to 95.09% read pairs (correspondingly 34.93% barcodes) have both types of parental unique markers (Supplementary Table S12). This is because each barcode corresponds to more DNA long fragments relative to stLFR cobarcoded reads, and the higher LFR collision ratio makes the clustering more complex and error-prone. Two assemblies showed scaffold N50s of 6.84 and 13.7 Mb, contig N50s of 97 and 123 kb with single-base accuracy of Q46 and Q48 for the paternal and maternal haplotypes, respectively (Supplementary Table S13). Compared to the stLFR data, the 10× Genomics linked reads also obtained a similar haplotyping precision rate of 96.85% and a recall rate of 93.66% on average, beyond those of Supernova pseudo-haplotypes (Supplementary Table S13 and Fig. S3). However, the relatively higher LFR collision rate leads to more switch errors, reducing the contiguity of phase blocks (Supplementary Table S14).

A phase block refers to the region in the sequence where at least two paternal or maternal-specific *k*-mers occur. Two consecutive conflicting haplotype-specific *k*-mers within a certain range are marked as a switch error and split the sequence into two phase blocks. We allowed at most 100 switches within 20 kb range to calculate the phase block.

3.5 Investigation of phased SVs

As an additional validation, we generated ~46× coverage PacBio CCS long reads from the same diploid sample and examined the phased SVs by HAST and Supernova assemblies. Currently there are no reference assemblies for the parents so the relatively long and accurate CCS reads were ideal to inspect the haplotypes and the associated phased SVs. The CCS reads were partitioned using TrioCanu (Koren *et al.*, 2018) and individually mapped against each haplotype or pseudo-haplotype. We investigated regions corresponding to all 52 high-confidence SVs that were discovered by various long read

Table 2. Statistics for phase blocks and switch error rates

Phasing blocks	No. of blocks	Genome covered (bp)	Block size Max	Block size N50	Switch (%)
HAST					
Paternal	17 501	2 740 909 157	38 467 450	11 618 507	0.05
Maternal	14 667	2 831 811 967	69 220 504	18 343 560	0.03
Supernova					
Pseudohap1	53 472	2 632 497 377	5 995 447	962 576	4.54
Pseudohap2	53 736	2 635 501 290	5 447 421	996 724	4.31
TrioCanu					
Paternal	5164	2 650 387 668	14 920 882	3 239 941	0.05
Maternal	3595	2 819 463 735	20 744 553	4 742 849	0.02

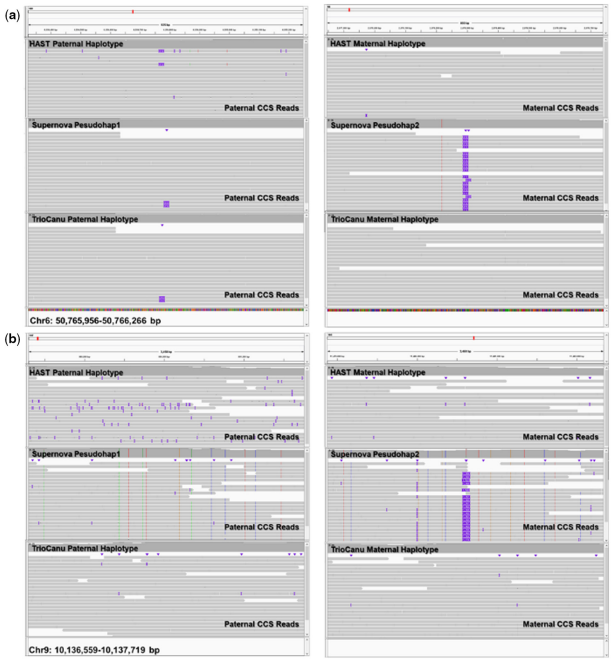

Fig. 4. Structural variants visualized in IGV for HAST's haplotypes, Supernova's pseudo-haplotypes and TrioCanu's haplotypes. CCS reads were mapped against each assembly and displayed in the Integrative Genomics Viewer (IGV) (Robinson *et al.*, 2011). The pseudo-haplotypes are missing two phased SVs: paternal-specific deletions in (**a**) hg19 Chr6: 50 765956–50 766 266 bp and (**b**) Chr9: 10 136 559–10 137 719 bp. The CCS reads were first partitioned by TrioCanu, and then individually mapped to the corresponding assemblies with minimap2 (Li, 2018). All assemblies were aligned with the reference hg19 to obtain proper coordinates

sequencing platforms and validated by PCR and Sanger sequencing in the SV benchmark for HJ sample (Du *et al.*, 2020). Those SVs were identified as heterozygous according to the supporting CCS read number, which is approximately half of that for homozygous SVs. About 90.6% of phased SVs longer than 150 bp were detected by the HAST haplotype-resolved assemblies, of which 75.9% appeared only in one assembly (Supplementary Table S16). By contrast, the pseudo-haplotypes by Supernova recovered 75.0% phased SVs, but only 20.0% of them supported the heterozygosity.

Moreover, we mapped the partitioned CCS reads against the SV regions in assemblies to visualize the haplotyping. Each group of haplotype-specific CCS reads, for instance, was well aligned with the corresponding HAST haplotype in Figure 4, indicating a good concordance of phased sequences. In contrast, there were overlaps of ~310 bp (Fig. 4a) and 1160 bp (Fig. 4b) in the mapped CCS reads for one of the pseudo-haplotypes by Supernova, while the other pseudo-haplotype showed good alignment. As the CCS reads most likely represent the true chromosomal configuration, the 'overlaps' represent two phased deletions in the paternal-inherited chromosome 6 and 9, respectively. Both SVs were accurately assembled by HAST but were missing in the Supernova assembly.

3.6 Comparison to PacBio CCS haplotypes

TrioCanu (Koren *et al.*, 2018) also provided a haplotype-resolved CCS assembly using the trio-binning strategy, with contig N50s of 3.06 and 4.56 Mb for paternal and maternal portions, respectively (Supplementary Table S7). Mapping against the reference genome resulted in contig NGA50s and genome coverages from 1.39 to 1.70 Mb and 95.20% to 97.33%. The variations between two haplotypes indicate the presence of the *X* chromosome in maternal, consistent with results from HAST. With the benefit of local phasing information from long read lengths, the CCS-based assembly exhibited similar haplotyping precision (99.94% and 99.97%) and recall rates (97.95% and 97.41%) relative to HAST, as well as high completeness in the BUSCO assessment (93.6% and 94.6%, Table 1). Almost all the haplotype-resolved contigs were aligned along either the *x*-axis or *y*-axis in the blob plot, which demonstrated the excellent haplotyping precision and efficiency resulting from trio binning (Fig. 3c). In addition, we also investigated the phased SVs recovered by TrioCanu assemblies and compared with those by HAST. Totally 93.8% of 52 phased SVs were detected by TrioCanu, among which the haplotyping consistency with HAST is 92.9% and 100.0% for paternal- and maternal-specific groups, respectively (Supplementary Table S16). In general, the reconstruction of two haplotypes by TrioCanu was equivalent to that of HAST and superior to Supernova.

The average *k*-mer-level accuracy in the CCS assemblies was Q46 (Table 1), substantially lower than stLFR assemblies by HAST or Supernova (>Q60). The base-calling accuracy has been improved by the consensus procedure during the CCS read generation. The residual errors in the input reads introduced a remarkable number of mis-assembled *k*-mers in the final result (Fig. 3a and b), with an error bar higher than the stLFR assemblies. Nevertheless, the haplotype-resolved CCS assembly produced a 60-fold longer contig N50 compared to that of HAST (Table 1), but with a 4-fold shorter phase block N50 (Table 2). It indicates the absence of long-range scaffolding information as CCS reads are typically around 10 kb but stLFR long fragment length can reach almost 300 kb.

3.7 Assembled MHC/KIR regions with accurate phasing

To better understand the structural accuracy of phased assemblies, we investigated the highly repetitive and polymorphic regions of biological importance such as Major Histocompatibility Complex (MHC) and Killer-cell Immunoglobulin-like Receptor (KIR). The MHC region contains human leukocyte antigen (HLA) genes, important to cancer and autoimmunity studies (Brandt *et al.*, 2015). A single long scaffold of HAST maternal haplotype covered the entire MHC region, while the paternal haplotype assembled two scaffolds to cover (>99%) (Fig. 5). There were obvious distinctions between two haplotype-specific assemblies. Previous studies showed that HLA type phasing was consistent with a trio structure (Chin *et al.*, 2020). We observed that 23 and 22 out of 24 HLA class I and II genes (Horton *et al.*, 2004) were recovered by HAST maternal and paternal haplotypes, respectively (Supplementary Table S20).

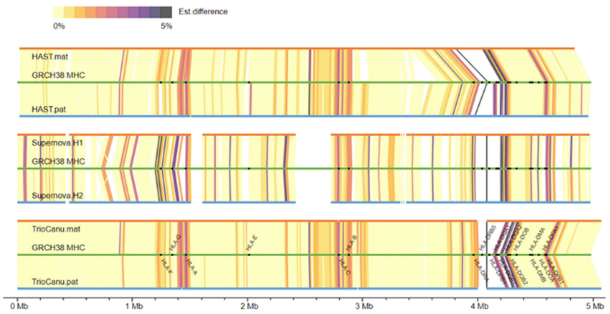

Fig. 5. Alignments of two haplotypes to the GRCh38 MHC region. The haplotype-resolved scaffolds or contigs were mapped to the reference using minimap2 (Li, 2018). The scaffolds or contigs were locally anchored to the reference using a greedy strategy. The edit distance between 10 kb bins and reference was individually computed using a modified O(ND) dynamics programming alignment algorithm (Myers, 1986) (penalty score: 1 for an indel, 1 for a substitution, 0 for an N). The white strips corresponding to the regions with >30% N's were rescaled based on the aligned positions of adjacent regions, and their differences were not computed

By comparison, Supernova assemblies were more fragmented in this region covering ~86% bases, and erroneously showed overall homozygous structures. As a result, Supernova failed to assemble 4 HLA genes for each pseudo-haplotype, including *HLA-DRB1*, *HLA-DQB1*, *MICA* and *MICB*. TrioCanu also correctly reconstructed most of the MHC region (>99%) for two haplotypes with several contigs, and only paternal haplotype missed 1 HLA genes.

Similarly, all three methods successfully assembled the KIR region with one scaffold or contig for each haplotype (Supplementary Fig. S5). However, only Supernova failed to represent the structural heterogeneity between two haplotypes.

4 Results

In this study, we provided a direct way to individually assemble the haplotypes of an individual using parental sequencing data. Although the DNA fragment length and read coverage of each fragment vary for different cobarcoded datasets, HAST can cluster reads sharing the same barcodes and retain the long-range phased sequence information. The application of trio binning by HAST simplifies the assembly problem and achieves significant improvements in the contiguity, precision and recall of haplotyping. The assembly of sex chromosomes could be further enhanced by the identification of male-specific *k*-mers (Wang et al., 2020). This concept could also be applied to complex polyploid genomes if parents or related species are available, which is a potentially important application for animal and plant genetics and breeding programs.

The long-range Hi-C conformation data and long reads can also provide accurate phasing information ranging from single-base level to chromosome level (Garg *et al.*, 2020). Following the HAST assembly, two sets of haplotype-resolved scaffolds were possible to be improved in collaboration with other sequencing platforms. With additional haplotype-resolved Hi-C data, the chromosome-level genomes were obtained with scaffold N50s of 145.0 and 153.2 Mb (Supplementary Table S17). Additionally, the PacBio CCS reads could extend contigs with a >10-fold increase in contig N50 (Xu *et al.*, 2020) (Supplementary Table S18). The improved contiguity, haplotype precision and recall of the final assembly will provide access to the reference-level genomes for different individuals with haplotype-specific SVs, which is essential for the studies of genomic diseases and evolutionary relationships.

Acknowledgements

The data that support the findings of this study have been deposited into CNGB Nucleotide Sequence Archive (CNSA) (Guo *et al.*, 2020) of China National GeneBank DataBase (CNGBdb) (Chen Fengzhen, 2020) with accession number CNP0001199.

Funding

This work was supported by the Qingdao Applied Basic Research Projects (Grant No. 19-6-2-33-cg) and the National Key Research and Development Program of China (Grant No. 2018YFD0900301-05).

Conflict of Interest: Some of the authors are employees of BGI Group. The authors otherwise declare that they have no competing interests.

Data availability

The sequencing data for the sample HJ (including stLFR and PacBio CCS) has been deposited in the CNGB under accession number CNP0000091. The stLFR data for HG001/NA12878 and HG002/NA24385 is available in the CNGB under accession number CNP0000066 (PRJEB27414). We downloaded the 10X Genomics linked reads of HG001/NA12878 from GIAB (ftp://ftp.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/10Xgenomics_ChromiumGenome_LongRanger2.0_06202016/NA12878.fastqs/). All the evaluated assemblies generated by us can be obtained in the CNGB under accession number CNP0001199. All codes, scripts and manuals to build haplotype-specific k-mer sets, and classify stLFR reads are available at https://github.com/BGI-Qingdao/HAST.

References

Bachtrog,D. and Charlesworth,B. (2001) Towards a complete sequence of the human Y chromosome. *Genome Biol.*, **2**, reviews1016.1–reviews1016.5.

Bishara,A. *et al.* (2015) Read clouds uncover variation in complex regions of the human genome. *Genome Res.*, **25**, 1570–1580.

Brandt,D.Y. *et al.* (2015) Mapping bias overestimates reference allele frequencies at the HLA genes in the 1000 genomes project phase I data. *G3 (Bethesda)*, **5**, 931–941.

Chen Fengzhen,Y.L. *et al.* (2020) CNGBdb: China National GeneBank DataBase. *Yi Chuan*, **42**, 799–809.

Chin,C.S. *et al.* (2016) Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods*, **13**, 1050–1054.

Chin,C.-S. *et al.* (2020) A diploid assembly-based benchmark for variants in the major histocompatibility complex. *Nat. Commun.*, **11**, 4794.

Church,D.M. *et al.* (2011) Modernizing reference genome assemblies. *PLoS Biol.*, **9**, e1001091.

Du,X. *et al.* (2021) Robust benchmark structural variant calls of an asian using the state-of-art long fragment sequencing technologies. *Genomics Proteomics Bioinform.*

Eberle,M.A. *et al.* (2017) A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Res.*, **27**, 157–164.

Edge,P. *et al.* (2017) HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res.*, **27**, 801–812.

Fofanov,Y. *et al.* (2004) How independent are the appearances of n-mers in different genomes? *Bioinformatics*, **20**, 2421–2428.

Garg,S. *et al.* (2018) A graph-based approach to diploid genome assembly. *Bioinformatics*, **34**, i105–i114.

Garg,S. *et al.* (2020) Chromosome-scale, haplotype-resolved assembly of human genomes. *Nat. Biotechnol.*

Guo,X. *et al.* (2020) CNSA: a data repository for archiving omics data. *Database*, **2020**.

Gurevich,A. *et al.* (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, **29**, 1072–1075.

Hill,W.G. (1974) Estimation of linkage disequilibrium in randomly mating populations. *Heredity*, **33**, 229–239.

Horton,R. *et al.* (2004) Gene map of the extended human MHC. *Nat. Rev. Genet.*, **5**, 889–899.

Koren,S. *et al.* (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.*, **27**, 722–736.

Koren,S. *et al.* (2018) De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.*, **36**, 1174–1182.

Kuleshov,V. *et al.* (2016) Genome assembly from synthetic long read clouds. *Bioinformatics*, **32**, i216–i224.

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

Low,W.Y. *et al.* (2020) Haplotype-resolved genomes provide insights into structural variation and gene content in Angus and Brahman cattle. *Nat. Commun.*, **11**, 2071.
Luo,R. *et al.* (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*, **1**, 18.
Marcais,G. and Kingsford,C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, **27**, 764–770.
Marchini,J. *et al.* (2006) A comparison of phasing algorithms for trios and unrelated individuals. *Am. J. Hum. Genet.*, **78**, 437–450.
Myers,E.W. (1986) AnO(ND) difference algorithm and its variations. *Algorithmica*, **1**, 251–266.
Myers,E.W. (2005) The fragment assembly string graph. *Bioinformatics*, **21**, ii79–ii85.
O'Connell,J. *et al.* (2016) Haplotype estimation for biobank-scale data sets. *Nat. Genet.*, **48**, 817–820.
Peters,B.A. *et al.* (2012) Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature*, **487**, 190–195.
Peters,B.A. *et al.* (2014) Co-barcoded sequence reads from long DNA fragments: a cost-effective solution for "perfect genome" sequencing. *Front. Genet.*, **5**, 466.
Rhie,A. *et al.* (2020) Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.*, **21**, 245.
Robinson,J.T. *et al.* (2011) Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24–26.
Rogers,Y.-H. and Venter,J.C. (2005) Massively parallel sequencing. *Nature*, **437**, 326–327.
Simao,F.A. *et al.* (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.
Snyder,M.W. *et al.* (2015) Haplotype-resolved genome sequencing: experimental methods and applications. *Nat. Rev. Genet.*, **16**, 344–358.
Vurture,G.W. *et al.* (2017) GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*, **33**, 2202–2204.
Wang,O. *et al.* (2019) Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. *Genome Res.*, **29**, 798–808.
Wang,X. *et al.* (2020) SRY: an effective method for sorting long reads of sex-limited chromosome. *bioRxiv 2020.05.25.115592*.
Weisenfeld,N.I. *et al.* (2017) Direct determination of diploid genome sequences. *Genome Res.*, **27**, 757–767.
Wu,Y.W. and Ye,Y. (2011) A novel abundance-based algorithm for binning metagenomic sequences using l-tuples. *J. Comput. Biol.*, **18**, 523–534.
Xu,M. *et al.* (2020) TGS-GapCloser: a fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *GigaScience*, **9**, **9**.
Zhang,X. *et al.* (2020) Unzipping haplotypes in diploid and polyploid genomes. *Comput. Struct. Biotechnol. J.*, **18**, 66–72.
Zheng,G.X. *et al.* (2016) Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol.*, **34**, 303–311.
Zhou,X. *et al.* (2019) Aquila stLFR: assembly based variant calling package for stLFR and hybrid assembly for linked-reads. *bioRxiv 2019:742239*.