nature portfolio

https://doi.org/10.1038/s41587-022-01261-x

Supplementary information

Haplotype-resolved assembly of diploid genomes without parental data

In the format provided by the authors and unedited

Supplementary Information for

"Haplotype-resolved assembly of diploid genomes without parental data"

1 Software commands

1.1 Hifiasm

To produce primary assemblies and alternate assemblies, hifiasm (version 0.16.1-r375) was run with the following command:

```
hifiasm -o <outputPrefix> -t <nThreads> --primary <HiFi-reads.fasta>
```

For partial phased assemblies only with HiFi, hifiasm was run with the following command:

```
hifiasm -o <outputPrefix> -t <nThreads> <HiFi-reads.fasta>
```

For fully phased assemblies with Hi-C, hifiasm was run with the following command:

```
hifiasm -o <outputPrefix> --h1 <HiC-reads-R1.fasta> --h2 <HiC-reads-R2.fasta> -t <nThreads> <HiFi-reads.fasta>
```

For trio-binning assembly, we first built the paternal trio index and the maternal trio index by yak (version 0.1-r62-dirty) with the following commands:

```
yak count -b37 -t <nThreads> -o <pat.yak> <paternal-short-reads.fastq>
yak count -b37 -t <nThreads> -o <mat.yak> <maternal-short-reads.fastq>
```

and then we produced the paternal assembly and the maternal assembly with the following command:

```
hifiasm -o <outputPrefix> -t <nThreads> -1 <pat.yak> -2 <mat.yak> <HiFi-reads.fasta>
```

1.2 HiCanu

For primary assembly, HiCanu (version 2.1.1) was run with the following command line:

```
canu -p asm -d <outDir> genomeSize=<GSize> useGrid=false maxThreads=<nThreads> \
   -pacbio-hifi <HiFi-reads.fasta>
```

The contigs labeled by 'suggestedBubbles=yes' were removed from the primary assembly. We ran purge_dups (version v1.2.5) to postprocess the HG002, HG00733 and Sterlet assemblies. The HiCanu assemblies of European badger and Black Rhinoceros were not purged as purge_dups resulted in significantly worse assemblies.

1.3 Purge_dups

Purge_dups (version 1.2.5) was used to postprocess the output primary assemblies of HiCanu for HG002, HG00733 and Sterlet. The commands are as follows:

```
minimap2 -I10G -xmap-pb <asm.fa> <HiFi-reads.fasta> -t <nThreads> > <read-aln.paf>
bin/pbcstat <read-aln.paf>
bin/calcuts PB.stat > cutoffs
bin/split_fa <asm.fa> > <split.fa>
minimap2 -I10G -xasm5 -DP <split.fa> <split.fa> -t <nThreads> > <ctg-aln.paf>
bin/purge_dups -2 -T cutoffs -c PB.base.cov <ctg-aln.paf> > <dups.bed>
bin/get_seqs <dups.bed> <asm.fa>
```

We then manually adjusted the cutoff thresholds of purge_dups as "5 7 11 30 22 42", "5 7 11 30 22 42" and "5 24 24 25 25 92" for HG002, HG00733 and Sterlet, respectively.

1.4 Running asmgene

For HG00733 and HG002, we aligned the cDNAs to the GRCh38 human reference genome and contigs by minimap2 (version 2.20-r1061), and evaluated the gene completeness with paftools.js from the minimap2 package:

```
minimap2 -cxsplice:hq -t <nThreads> <ref.fa> <cDNAs.fa> > <ref.paf>
minimap2 -cxsplice:hq -t <nThreads> <asm_contig.fa> <cDNAs.fa> > <asm.paf>
paftools.js asmgene -a -i.97 <ref.paf> <asm.paf>
```

When evaluating multi-copy genes missed in an assembly, we replaced '`-i.97`' to '`-i.99`' to increase the resolution.

1.5 BUSCO

BUSCO (version 5.1.3) was used with the following command:

```
busco -i <asm.fa> -m genome -o <outDir> -c <nThreads> -l <lineage_dataset>
```

where '`lineage_dataset`' was set to *mammalia_odb10* for European badger and Black Rhinoceros, set to *actinopterygii_odb10* for Sterlet and set to *aves_odb10* for South Island takahe.

1.6 QV evaluation

We used yak (version 0.1-r62-dirty) to measure the per-base consensus accuracy (QV). To this end, we built the index for the short reads coming from the same sample and did evaluation:

```
yak count -b37 -t <nThreads> -o <sr.yak> <short-reads.fastq>
yak qv -t <nThreads> <sr.yak> <asm.fa>
```

1.7 Phasing accuracy evaluation

We used yak (version 0.1-r62-dirty) to measure the hamming error rate and the switch error rate:

```
yak trioeval -t <nThreads> <paternal.yak> <maternal.yak> <asm_contig.fa>
```

1.8 IGV visualization

We used IGV (version 2.10.0) to visualize alignment around complex genes on the GRCh38 human reference genome. HiFi reads were aligned by minimap2 (version 2.20-r1061) as follows:

```
minimap2 -ax map-hifi -t <nThreads> <ref.fa> <HiFi-reads.fasta>
```

Assemblies were aligned with the following command:

```
minimap2 -ax asm5 -t <nThreads> <ref.fa> <asm.fa>
```

1.9 Generating chromosome-level phasing plots

For HG00733 and HG002, the hifiasm (Hi-C) phased contigs were aligned by minimap2 (version 2.23-r1114-dirty) to the reference combing the T2T CHM13 assembly (v1.1) and the Y chromosome of GRCh38:

```
minimap2 -cxasm20 -r1k -t <nThreads> <ref.fa> <asm_contig.fa> > <ref.paf>
```

The phasing density of each contig was determined by the number of paternal- and maternal-specific k-mers on this contig, which were obtained by yak (version 0.1-r62-dirty):

```
yak trioeval -t <nThreads> <paternal.yak> <maternal.yak> <asm_contig.fa>
```

Supplementary Table 1: Statistics of different assemblies

Dataset	Assembler	QV	Switch error (%)	Completeness (asmgene or BUSCO)			
				Single (%)	Dup (%)	Frag (%)	Miss (%)
HG002 (HiFi + trio/Hi-C)	hifiasm (Hi-C)	52.3/52.7	1.10/0.72	98.97/98.75	0.32/0.32	0.27/0.28	0.45/0.64
	FALCON-Phase (Hi-C)	43.4/43.6	1.52/1.48	96.15/96.13	3.14/3.13	0.23/0.30	0.45/0.44
	hifiasm (trio)	52.7/52.3	0.82/0.98	98.88/98.91	0.29/0.33	0.29/0.27	0.54/0.49
HG002 (HiFi only)	hifiasm (dual)	52.0/52.3	1.18/0.89	98.76/98.73	0.35/0.31	0.32/0.32	0.57/0.64
	hifiasm (primary/alt)	52.5/51.8	0.89/092	99.09/85.43	0.34/2.67	0.24/7.46	0.33/4.43
	HiCanu (primary/alt)	52.2/38.1	1.20/0.70	98.69/80.48	0.19/5.15	0.39/9.33	0.73/5.04
HG00733 (HiFi + trio/Hi-C /Strand-seq)	hifiasm (Hi-C)	50.9/50.8	1.16/0.98	99.14/99.16	0.31/0.35	0.24/0.23	0.32/0.26
	DipAsm (Hi-C)	41.6/41.7	2.30/2.15	98.64/98.64	0.39/0.40	0.48/0.47	0.49/0.49
	PGAS (Strand-seq)	45.5/46.1	1.44/1.63	98.99/99.03	0.16/0.15	0.40/0.44	0.45/0.38
	hifiasm (trio)	50.8/51.1	0.97/0.97	99.08/98.96	0.42/0.32	0.24/0.25	0.26/0.46
HG00733 (HiFi only)	hifiasm (dual)	50.8/51.1	1.34/1.32	99.02/98.76	0.34/0.42	0.25/0.29	0.38/0.52
	hifiasm (primary/alt)	51.0/49.5	1.29/0.90	99.06/82.06	0.51/2.89	0.21/9.42	0.21/5.63
	HiCanu (primary/alt)	50.8/36.6	1.38/0.99	98.75/76.49	0.14/6.29	0.35/10.90	0.76/6.32
European badger (HiFi + trio/Hi-C)	hifiasm (Hi-C)	50.3/51.7	1.18/1.73	95.09/92.71	1.68/1.63	0.67/0.73	2.56/4.94
	hifiasm (trio)	51.6/50.2	0.82/2.70	92.74/93.43	1.70/1.68	0.70/0.73	4.86/4.16
European badger (HiFi only)	hifiasm (dual)	50.6/50.8	1.52/1.37	93.67/94.49	1.65/1.65	0.70/0.67	3.98/3.19
	hifiasm (primary/alt)	50.2/50.1	1.15/1.31	95.15/50.24	1.67/1.35	0.65/3.27	2.53/45.13
	HiCanu (primary/alt)	49.0/35.3	1.11/1.34	94.79/35.74	1.96/2.57	0.61/3.53	2.64/58.16
Sterlet (HiFi + trio/Hi-C)	hifiasm (Hi-C)	47.5/47.3	0.84/0.84	35.22/34.81	57.83/58.35	2.25/2.39	4.70/4.45
	hifiasm (trio)	47.3/47.3	1.07/0.59	34.15/35.36	59.15/57.91	2.20/2.28	4.51/4.45
Sterlet (HiFi only)	hifiasm (dual)	47.2/47.4	0.85/0.84	36.48/34.01	56.92/58.79	2.25/2.39	4.34/4.81
	hifiasm (primary/alt)	47.2/47.5	0.83/0.88	34.42/36.98	59.01/55.66	2.31/2.28	4.26/5.08
	HiCanu (primary/alt)	47.8/41.5	0.82/0.93	49.01/30.27	42.47/59.97	2.12/2.06	6.40/7.69
South Island takahe (HiFi + trio/Hi-C)	hifiasm (Hi-C)		0.43/0.41	96.47/89.82	0.54/0.46	0.58/0.68	2.41/9.04
	hifiasm (trio)		1.65/0.18	90.59/91.68	0.64/0.61	0.65/0.62	8.13/7.09
South Island takahe (HiFi only)	hifiasm (dual)		0.47/0.42	92.07/93.93	0.49/0.52	0.66/0.70	6.78/4.86
	hifiasm (primary/alt)		0.38/0.40	96.52/44.60	0.59/0.73	0.59/1.91	2.30/52.76
Black Rhinoceros (HiFi + trio/Hi-C)	hifiasm (Hi-C)		0.66/0.68	95.66/96.04	0.82/0.78	0.76/0.69	2.75/2.48
	hifiasm (trio)		0.82/0.45	95.24/95.91	0.89/0.90	0.69/0.67	3.18/2.51
Black Rhinoceros (HiFi only)	hifiasm (dual)		0.77/0.74	93.69/95.12	0.80/0.87	0.72/0.72	4.79/3.30
	hifiasm (primary/alt)		0.73/0.61	95.99/83.75	0.80/1.01	0.69/2.53	2.51/12.71
	HiCanu (primary/alt)		0.87/0.58	95.26/68.73	1.53/1.38	0.72/3.74	2.49/26.15

Each assembly consists of two sets of contigs. The two sets represent paternal/maternal with trio binning, haplotype 1/haplotype 2 with haplotype-resolved assembly or hifiasm dual assembly, or represent primary/alternate contigs. The two numbers in each cell give the metrics for the two sets of contigs, respectively. QV is the Phred-scaled contig base error rate measured by comparing 31-mers in contigs to 31-mers in short reads from the same sample. QV was not reported for the assemblies of south island takahe and black rhinoceros as there are no short reads from the same sample. The phasing switch error rate is the percentage of adjacent parental-specific 31-mer pairs that come from different parental haplotypes. It was calculated with yak.

Supplementary Table 2: Run time and peak memory usage of different assemblers

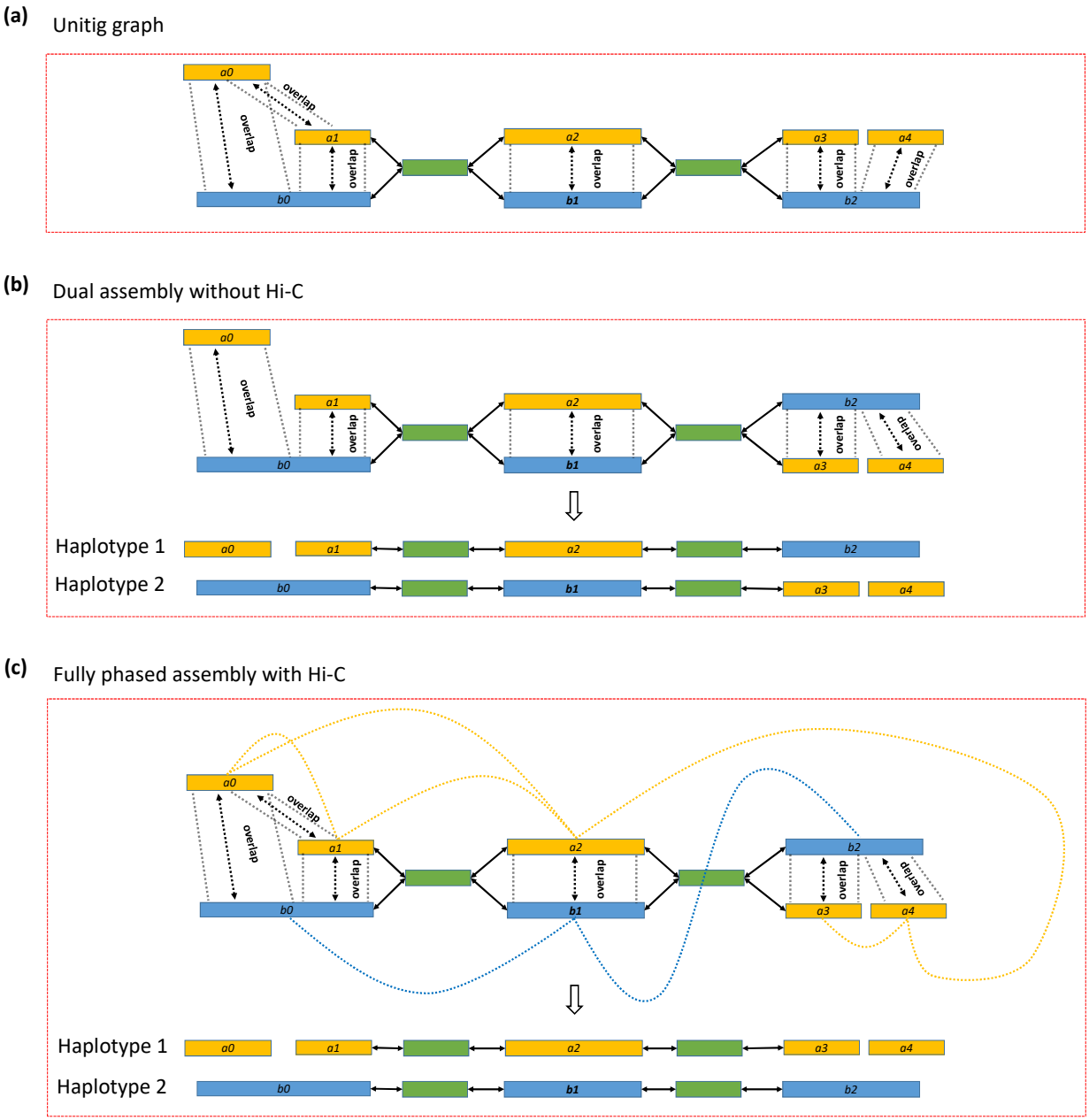
Dataset	Metric	hifiasm (Hi-C)	hifiasm (trio)	hifiasm (dual)	hifiasm (primary/alt)	HiCanu (primary/alt)
HG002	Elapsed time (h)	10.8	10.1	9.5	9.4	48.5
	CPU time (h)	366.9	368.3	358.1	357.2	1,369.3
	Peak Memory (Gb)	149.5	142.9	142.9	142.9	83.7
HG00733	Elapsed time (h)	9.1	8.6	7.7	7.7	43.9
	CPU time (h)	292.2	293.7	282.0	281.9	1,202.2
	Peak Memory (Gb)	147.1	134.9	134.9	134.9	70.8
European badger	Elapsed time (h)	13.2	11.7	10.9	10.9	65.3
	CPU time (h)	466.3	452.5	440.2	440.1	2,185.1
	Peak Memory (Gb)	178.8	178.8	178.8	178.8	131.2
Sterlet	Elapsed time (h)	14.5	13.6	13.1	13.0	186.6
	CPU time (h)	565.4	570.1	551.8	551.7	7,701.7
	Peak Memory (Gb)	184.1	184.1	184.1	184.1	41.2
South Island takahe	Elapsed time (h)	4.3	3.9	3.7	3.6	
	CPU time (h)	149.1	147.3	143.2	143.0	
	Peak Memory (Gb)	70.0	66.1	66.1	66.1	
Black Rhinoceros	Elapsed time (h)	13.7	12.2	11.2	11.2	41.9
	CPU time (h)	461.5	451.7	436.0	435.7	1,358.5
	Peak Memory (Gb)	194.5	194.5	194.5	194.5	42.6

All assemblies were generated using the same machine with 48 CPU threads. For HiCanu (primary/alt), we ran purge_dups to postprocess the HG002, HG00733 and Sterlet assemblies. The HiCanu assemblies of European badger and Black Rhinoceros were not purged as purge_dups resulted in significantly worse assemblies. For South Island takahe, HiCanu could not produce assembly in 3 weeks. According to the publications, the phasing step of Falcon-Phase (excluding contig assembly) took 46 wall-clock hours and 579 CPU hours over 64 threads for a human HiFi dataset. The complete PGAS pipeline took 2,000 CPU hours for human HiFi data at 30-fold coverage. DipAsm took a day over 64 threads to assemble a 30-fold human dataset with the commercial HiRise scaffolder from DoveTail Genomics or took two weeks with the open source 3D-DNA pipeline.

Supplementary Table 3: Hifiasm (Hi-C) assemblies of HG002 with different times of flipping (option --n-perturb)

Strategy	Assembly	# phase flips					
		5,000		10,000		15,000	
		Switch error (%)	Hamming error (%)	Switch error (%)	Hamming error (%)	Switch error (%)	Hamming error (%)
Without optimization	haplotype 1	0.90	20.09	0.91	19.87	0.89	20.02
	haplotype 2	1.16	25.82	1.15	25.98	1.18	27.79
With optimization	haplotype 1	0.72	0.82	1.10	1.42	0.79	0.90
	haplotype 2	1.11	1.43	0.72	0.82	1.02	1.14

Hifiasm has an optimization strategy that may flip a pair of unitig sets at the same time if the two sets of unitigs are inferred to come from opposite phases and to be homologous to each other. Supplementary Table 3 shows the phasing accuracy of HG002 assemblies with and without this optimization strategy.



Supplementary Fig. 1: Outline of the hifiasm phasing algorithm. **(a)** Rectangles in orange and blue represent heterozygous unitigs (haplotigs) coming from haplotype 1 and haplotype 2, respectively. Rectangles in green are homozygous unitigs. Solid lines between unitig ends indicate that the corresponding unitigs are linked in unitig graph. By utilizing all-to-all overlap calculation with *trans* read overlaps, hifiasm identifies 6 overlaps between heterozygous unitigs: $a0 \leftrightarrow a1$, $a0 \leftrightarrow b0$, $a1 \leftrightarrow b0$, $a2 \leftrightarrow b1$, $a3 \leftrightarrow b2$ and $a4 \leftrightarrow b2$. All overlaps to homozygous unitigs are filtered out as homozygous unitigs do not need to be phased. **(b)** Without Hi-C, the optimization model for dual assembly assigns $a0$ and $a1$ to the same haplotype based on all overlaps between heterozygous unitigs. As a result, it identifies $a0 \leftrightarrow a1$ does not represent the homologous relationship between different haplotypes. Hifiasm phasing algorithm bins homologous unitigs to different haplotypes to produce two non-redundant primary assemblies. **(c)** Dashed lines represent Hi-C contacts between unitigs. The Hi-C contacts to homozygous unitigs are discarded as homozygous unitigs are not informative for phasing. Two fully phased assemblies are generated by integrating Hi-C data.