Supplementary Information for

"Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm"

1 Software commands

1.1 Hifiasm

To produce primary assemblies of homozygous samples (*M. musculus*, *Z. mays* and CHM13), hifiasm (version 0.12) was run with the following command which does not purge haplotig duplications:

```
hifiasm -o <outputPrefix> -t <nThreads> -l0 <HiFi-reads.fasta>
```

For heterozygous samples, hifiasm was run with the following command:

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

We added '`-D10`' for the octoploid *F.* × *ananassa* because the default k-mer cutoff seems too low:

```
hifiasm -o <outputPrefix> -t <nThreads> -D10 <HiFi-reads.fasta>
```

For trio-binning assembly, we first built the paternal trio index and the maternal trio index by yak (version r55) 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 Falcon-Unzip

Falcon-kit (version 1.8.1) was run with the following HiFi-specific options:

```
length_cutoff_pr = 8000
ovlp_daligner_option = -k24 -h1024 -e.98 -l1500 -s100
ovlp_HPCdaligner_option = -v -B128 -M24
ovlp_DBsplit_option = -s400
overlap_filtering_setting = --max-diff 200 --max-cov 200 --min-cov 2 --n-core 24 --min-idt
98 --ignore-indels
```

Falcon-unzip-kit (version 1.3.7) was run with default options.

1.3 HiCanu

For primary assembly, HiCanu (version 2.1) 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. For trio-binning assembly, we ran HiCanu in two steps as recommended. We partitioned the HiFi reads by parental short reads with the following command:

```
canu -haplotype -p asm -d <outDir> genomeSize=<GSize> useGrid=false \
   maxThreads=<nThreads> -haplotypePat <pat-reads.fq> -haplotypeMat <mat-reads.fq> \
   -pacbio-raw <HiFi-reads.fasta>
```

Note that '`-pacbio-raw`' was used to partition HiFi reads followed the document of HiCanu. We then perform HiCanu assemblies on partitioned reads.

1.4 Peregrine

For primary assemblies of all samples except maize, Peregrine (version 0.1.6.1) was run with the following command, where 48 is the number of threads in use:

```
docker run -it -v <workDir>:/wd --user $(id -u):$(id -g) cschin/peregrine:0.1.6.1 asm \
   /wd/Input.fnfo 48 48 48 48 48 48 48 48 48 --with-consensus --with-alt --shimmer-r 3 \
   --best_n_ovlp 8 --output <outDir>
```

For the highly repetitive maize genome, we employed the following options to improve contig contiguity:

```
docker run -it -v <workDir>:/wd --user $(id -u):$(id -g) cschin/peregrine:0.1.6.1 asm \
   /wd/Input.fnfo 48 48 48 48 48 48 48 48 48 --with-consensus --with-alt --shimmer-r 3 \
   --best_n_ovlp 255 --mc_upper 8192 --shimmer-w 32 --output <outDir>
```

For trio-binning assembly, we first used HiCanu to partition HiFi reads by parental short reads, and then assembled each haplotype individually by Peregrine.

1.5 Purge_dups

Purge_dups (version 1.2.3) was used to postprocess the output primary assemblies of HiCanu for all heterozygous samples except *R. muscosa*. The commands are as follows:

```
minimap2 -I6G -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 -I6G -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>
```

Since running purge_dups in default cannot produce primary assembly of HiCanu with right size for HG002, we manually adjusted the cutoff thresholds of purge_dups as follows "`5 7 11 30 22 42`".

1.6 Running asmgene

We aligned the cDNAs to the reference genome (Zm-B73-REFERENCE-NAM-5.0 for maize, GRCm38 for mouse, CHM13 reference generated by the T2T consortium for CHM13 and GRCh38 human genome for HG00733 and HG002) and contigs by minimap2 r974 and evaluated the gene completeness with paftools.js from the minimap2 package:

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

We set the sequence identity threshold to be 97% with '`-i.97`' to tolerate low per-base accuracy of ONT assemblies. For trio binning assemblies, we added option '`-a`' to evaluate genes mapped to the autosomes only. When evaluating multi-copy genes retained in an assembly, we replaced '`-i.97`' to '`-i.99`' to increase the resolution.

1.7 Computing NGA50

We used minigraph (version 0.10-dirty-r361) and paftools (version 2.17-r974-dirty) to calculate the NGA50 of HG00733 and HG002 assemblies:

```
minigraph -xasm -K1.9g --show-unmap=yes -t <nThreads> <ref.fa> <asm.fa> > <asm.paf>
paftools.js asmstat <ref.fa.fai> <asm.paf>
```

where the '`ref.fa`' is the GRCh38 version of human reference genome. In comparison to minimap2, minigraph tends to generate longer alignments and is more robust to highly variable regions. For CHM13, the CHM13 reference (v0.9) generated by the T2T consortium was used as '`ref.fa`'. The reason is that all CHM13 assemblies and the CHM13 reference were produced using the data from the exactly same cell line. So we ran minigraph with more stringent parameters as follows:

```
minigraph -xasm -g10k -r10k --show-unmap=yes -t <nThreads> <ref.fa> <asm.fa> > <asm.paf>
```

1.8 BUSCO

BUSCO (version 3.0.2) was used with the following command:

```
python3 run_BUSCO.py -i <asm.fa> -m genome -o <outDir> -c <nThreads> -l <lineage_dataset>
```

where '`lineage_dataset`' was set to *tetrapoda* for *R. muscosa*, set to *embryophyta* for *Z. mays*, *F.* × *ananassa* and *S. sempervirens* and set to *mammalia* for *M. musculus* and human genomes.

1.9 Determining resolved BACs

The resolution of BAC for different assemblies was evaluated using the pipeline at: https://github.com/skoren/bacValidation, except that we added option '`-I6g`' to minimap2. There are 341 BACs for CHM13 but 11 of them cannot be resolved by CHM13 T2T reference, indicating these 11 BACs may have cloning artifacts. These problematic BACs were excluded when evaluating CHM13 assemblies. For the fully phased diploid assembly, we ran this pipeline on each haplotype. If a BAC is correctly reconstructed in any haplotype, we think it is resolved in the diploid assembly. The BAC libraries for CHM13 and HG00733 can be found at https://www.ncbi.nlm.nih.gov/nuccore/?term=VMRC59+and+complete and https://www.ncbi.nlm.nih.gov/nuccore/?term=VMRC62+and+complete, respectively.

1.10 QV evaluation

We used yak (version r55) to measure the per-base consensus accuracy (QV). To this end, we first built the index for the short reads coming from the same sample:

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

1.11 Dipcall

For the male sample HG002, we ran dipcall (version 0.1) as follows:

```
dipcall.kit/run-dipcall -x dipcall.kit/hs37d5.PAR.bed <prefix> hs37d5.fa \
   <pat-asm.fa> <mat-asm.fa> > <prefix.mak>
make -j2 -f <prefix.mak>
```

For the female sample HG00733, we removed option '`-x`'. We used the GRCh37 variant of '`hs37d5.fa`' here because GIAB works best with hs37d5.

1.12 Evaluating collapsed misassemblies for inbred samples

We used scripts at: https://github.com/lh3/CHM-eval/blob/master/misc/clustreg.js, and https://github.com/lh3/CHM-eval/blob/master/misc/select-collapse-het.js. The commands are as follows:

```
minimap2 -axasm20 -t <asm.fa> <HiFi-reads.fasta> | samtools sort -o <aln.bam> -
htsbox pileup -vcf <asm.fa> -q20 -Q20 -l5000 -S5000 -s5 <aln.bam> > <var.vcf>
./select-collapse-het.js -c <readCoverage> <var.vcf> | ./clustreg.js -n10
```

where '`-l`' and '`-S`' filter out alignments shorter than 5kb.

1.13 QUAST

QUAST (version 5.0.2) was used with the following command for all CHM13 assemblies:

```
quast.py -t <nThreads> --large --skip-unaligned-mis-contigs -r <ref.fa> -o <outDir> <asm.fa>
```

The CHM13 reference (v0.9) generated by the T2T consortium (see Data availability) was employed as the '`ref.fa`'.

1.14 Running CHM13 centromere evaluation

To identify potential centromeric regions on CHM13 reference (v0.9) by the T2T consortium, dna-nn (version r60) was used with the following command:

```
dna-brnn -Ai models/attcc-alpha.knm -t <nThreads> <ref.fa> > <ref.bed>
```

Short repetitive regions which are less than 100 kbp reported by dna-nn were filtered to avoid false positives. We then merged all remaining regions of each chromosome to get the potential centromeric regions. After that, all CHM13 assemblies were aligned to CHM13 reference with minimap2:

```
minimap2 -cx asm5 -z10000,200 -t <nThreads> <ref.fa> <asm.fa>
```

and winnowmap (version 1.1):

```
winnowmap -W <repeat.list> -cx asm5 -z10000,200 -t <nThreads> <ref.fa> <asm.fa>
```

where '`-z10000,200`' was used to generate longer alignments. The alignments around potential centromeric regions were then inspected manually.

1.15 HG002 Major Histocompatibility Complex (MHC) evaluation

We aligned the HG002 MHC reference sequences (https://github.com/NCBI-Hackathons/TheHumanPangenome/tree/master/MHC/assembly/MHCv1.1) to each HG002 assembly with the following command:

```
minimap2 -cx asm20 -t <nThreads> <asm.fa> <MHC_ref.fa>
```

The alignment was then inspected manually.

1.16 Merqury

To evaluate QV with Merqury, the version 1.1 was used with the following command:

```
meryl count threads=<nThreads> k=21 output <index> <short-reads.fastq>
merqury.sh <index> <asm.fa> <output>
```

Supplementary Table 1: Misassembly statistics of CHM13 assemblies

Dataset	Assembler	NG50 (Mb)	# QUAST misassemblies		Diff per 1Mbp (paftools)	
			misassemblies	local misassemblies	mismatches	indels
CHM13 (primary)	hifiasm	88.9	186	290	1.9	5.4
	HiCanu	69.7	111	238	1.2	3.6
	Peregrine	37.8	1,100	967	78.8	19.4
	Falcon	27.1	1,035	518	42.3	21.4
	Canu (ONT)	80.0	1,103	2,142	60.5	645.2
	Flye (ONT)	37.5	883	1,364	65.8	595.3
	Shasta (ONT)	41.3	343	1,770	67.5	1,192.0

Supplementary Table 2: Completeness of HG002 fully phased assemblies and HG00733 fully phased assemblies

Dataset	Assembler	Haplotype	Multi-copy genes retained (%)	Gene completeness (asmgene)	
				Complete (%)	Duplicated (%)
HG00733	hifiasm (trio)	paternal	85.24	99.49	0.36
		maternal	82.68	99.35	0.39
	HiCanu (trio)	paternal	83.80	99.41	0.41
		maternal	84.84	99.25	0.47
	Peregrine (trio)	paternal	38.07	99.00	0.30
		maternal	37.11	98.94	0.28
	Peregrine (Hi-C)	haplotype 1	33.76	99.03	0.39
		haplotype 2	32.72	99.02	0.40
	Peregrine (Strand-seq)	haplotype 1	34.00	99.14	0.16
		haplotype 2	32.08	99.16	0.14
HG002	hifiasm (trio)	paternal	79.97	99.33	0.47
		maternal	81.32	99.25	0.34
	HiCanu (trio)	paternal	79.33	99.07	0.43
		maternal	81.56	99.10	0.45
	Peregrine (trio)	paternal	38.23	98.77	0.36
		maternal	39.27	98.85	0.34

Supplementary Table 3: Potential centromeric regions on CHM13 reference by the T2T consortium

Chromosome	Centromere start	Centromere end	Length
chr1	121,677,850	142,241,850	20,564,000
chr2	90,991,658	94,695,197	3,703,539
chr3	90,712,050	96,424,150	5,712,100
chr4	49,705,150	55,303,250	5,598,100
chr5	46,850,351	50,962,200	4,111,849
chr6	58,285,850	61,068,750	2,782,900
chr7	58,348,190	63,744,850	5,396,660
chr8	44,238,850	46,331,250	2,092,400
chr9	44,938,545	76,593,802	31,655,257
chr10	38,527,450	42,982,850	4,455,400
chr11	51,001,050	54,490,050	3,489,000
chr12	34,580,550	37,221,350	2,640,800
chr16	35,884,750	52,219,457	16,334,707
chr17	21,915,450	27,577,348	5,661,898
chr18	15,627,250	21,120,600	5,493,350
chr19	24,565,150	29,770,450	5,205,300
chr20	26,377,644	32,969,579	6,591,935
chrX	57,801,750	60,937,950	3,136,200

Short repetitive regions which are less than 100 kbp reported by dna-nn were filtered to avoid false positives (see Supplementary Section 1.14). Acrocentric chromosomes chr13, chr14, chr15, chr21 and chr22 are not shown.

Supplementary Table 4: Centromere alignment to the CHM13 reference by the T2T consortium

Assembler	Chromosome	Reference centromere length	Aligner	Length diff (reference - assembly)	Alignment type	Is misassembly
hifiasm	chr2	3,703,539	minimap2	0	go through	no
			winnowmap	0	go through	
	chr8	2,092,400	minimap2	23,124	broken	yes
			winnowmap	24,317	broken	
	chr11	3,489,000	minimap2	73,920	broken	yes
			winnowmap	75,629	broken	
	chr12	2,640,800	minimap2	0	go through	no
			winnowmap	0	go through	
HiCanu	chr8	2,092,400	minimap2	143	broken	no
			winnowmap	0	go through	
	chr20	6,591,935	minimap2	-7,135	broken	unknown
			winnowmap	-7,136	broken	
Falcon	chr8	2,092,400	minimap2	2,107,885	broken	yes
			winnowmap	2,107,885	broken	

For “Alignment type”, “go through” indicates there is an alignment covering the whole centromere, and “broken” indicates alignments are broken at the centromeric region. If the assembly is shorter than the CHM13 reference, the assembly is likely to be misassembled. Peregrine assembly and ONT assemblies break into large numbers of pieces around centromeres.

Supplementary Table 5: BUSCO scores of human primary assemblies

Dataset	Assembler	Complete BUSCOs		Fragmented BUSCOs (%)	Missing BUSCOs (%)
		Single-copy (%)	Duplicated (%)		
human reference	GRCh38	94.03	0.80	2.49	2.68
CHM13	T2T consortium	94.03	0.78	2.51	2.68
	hifiasm	94.09	0.95	2.31	2.66
	HiCanu	93.74	1.05	2.53	2.68
	Peregrine	93.93	0.83	2.49	2.75
	Falcon	94.05	0.88	2.36	2.70
	Canu (ONT)	91.98	0.93	4.02	3.07
	Flye (ONT)	92.20	0.93	3.83	3.05
	Shasta (ONT)	89.52	0.71	5.60	4.17
HG00733	hifiasm (purge)	94.01	1.05	2.31	2.63
	HiCanu (purge)	93.88	0.95	2.53	2.63
	Peregrine	94.05	1.00	2.39	2.56
	Falcon	92.71	0.90	2.27	4.12
	Canu (ONT)	90.98	0.90	4.78	3.34
	Flye (ONT)	91.45	0.63	4.90	3.02
	Shasta (ONT)	91.72	0.58	4.68	3.02
HG002	hifiasm (purge)	94.40	0.80	2.07	2.73
	HiCanu (purge)	94.23	0.85	2.31	2.61
	Peregrine	93.91	0.93	2.46	2.70
	Falcon	94.20	0.88	2.29	2.63
	Shasta (ONT)	90.86	0.56	5.04	3.53

Supplementary Table 6: BUSCO scores of haplotype-resolved human assemblies

Dataset	Assembler	Haplotype	BUSCOs (%)			
			Complete		Fragmented	Missing
			Single-copy	Duplicated		
HG00733	hifiasm (trio)	paternal	93.01	0.85	2.36	3.78
		maternal	93.86	0.97	2.41	2.75
	HiCanu (trio)	paternal	93.71	0.97	2.56	2.75
		maternal	93.88	1.05	2.39	2.68
	Peregrine (trio)	paternal	94.35	0.97	2.22	2.46
		maternal	93.76	0.88	2.66	2.70
	Peregrine (Hi-C)	haplotype 1	93.96	1.00	2.39	2.66
		haplotype 2	94.15	0.90	2.36	2.58
	Peregrine (Strand-seq)	haplotype 1	94.01	0.93	2.53	2.53
		haplotype 2	94.05	1.05	2.29	2.61
HG002	hifiasm (trio)	paternal	93.03	0.80	2.36	3.80
		maternal	94.23	0.90	2.34	2.53
	HiCanu (trio)	paternal	92.79	1.10	2.78	3.34
		maternal	94.05	0.95	2.46	2.53
	Peregrine (trio)	paternal	92.86	0.93	2.92	3.29
		maternal	94.10	0.95	2.34	2.61

Supplementary Table 7: BUSCO scores of non-human assemblies

Dataset	Assembler	Complete BUSCOs		Fragmented BUSCOs (%)	Missing BUSCOs (%)
		Single-copy (%)	Duplicated (%)		
M. musculus	hifiasm	94.13	1.27	2.24	2.36
	HiCanu	94.15	1.29	2.07	2.49
	Peregrine	94.10	1.29	2.27	2.34
	Falcon	94.20	1.29	2.07	2.44
Z. mays	hifiasm	87.48	8.43	1.47	2.60
	HiCanu	88.41	8.12	0.99	2.48
	Peregrine	87.92	8.24	1.30	2.54
	Falcon	88.04	8.12	1.24	2.60
F. × ananassa	hifiasm (purge)	5.02	93.43	0.06	1.49
	HiCanu	5.14	92.94	0.06	1.86
	HiCanu (purge)	41.70	55.08	0.25	2.97
	Peregrine	6.63	91.70	0.06	1.61
	Falcon	5.45	92.81	0.06	1.67
R. muscosa	hifiasm (purge)	64.91	1.70	13.29	20.10
	HiCanu	61.62	3.92	13.59	20.86
	Peregrine	65.11	1.72	12.94	20.23
S. sempervirens	hifiasm (purge)	21.89	39.42	8.87	29.82
	Peregrine	27.27	35.93	9.02	27.78

Supplementary Table 8: Merqury QV of human assemblies

Assembler	CHM13 (primary)	HG00733 (primary)	HG002 (primary)	HG00733 (fully phased)		HG002 (fully phased)	
				pat / hap1	mat / hap2	pat / hap1	mat / hap2
hifiasm	54.55	52.47	52.87				
hifiasm (trio)				52.22	52.73	52.68	54.08
HiCanu	54.53	53.38	55.43				
HiCanu (trio)				51.84	51.59	48.88	50.40
Peregrine	45.14	42.44	42.21				
Peregrine (trio)				44.52	44.00	43.36	44.25
Peregrine (Hi-C)				43.60	43.73		
Peregrine (Strand-seq)				47.60	48.20		
Falcon	50.76	48.44	48.26				
Canu (ONT)	34.06	30.46					
Flye (ONT)	34.62	31.10					
Shasta (ONT)	31.90	31.49					

Supplementary Table 9: Statistics of haplotype-resolved HG002 MHC assemblies

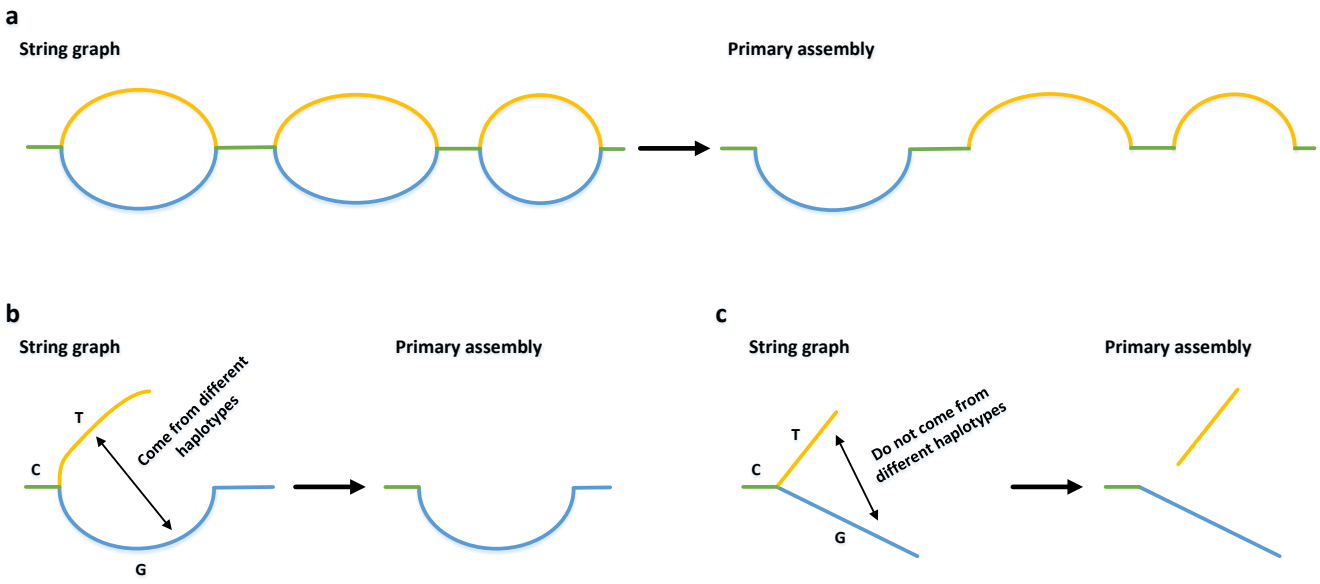
Assembler	Haplotype	Reference MHC length	Alignment breakpoints	Alignment uncover length	Length diff (reference - assembly)	# diff bases
hifiasm (trio)	paternal	5,077,518	0	7	0	569
	maternal	4,979,891	0	7	0	125
HiCanu (trio)	paternal	5,077,518	1	627	0	340
	maternal	4,979,891	0	7	0	559
Peregrine (trio)	paternal	5,077,518	1	7	32,852	826
	maternal	4,979,891	1	7	32,856	1,310

"Length diff (reference - assembly)" represents how many bases that are collapsed on assembly. "# diff bases" was collected from "NM:i" field of PAF format.

Supplementary Table 10: Run time and peak memory usage of different assemblers

Dataset	Metric	hifiasm	HiCanu	HiCanu + purge_dups	Peregrine	Falcon
M. musculus (primary)	Elapsed time (h)	3.9	25.2		3.0	
	CPU time (h)	147.1	493.6			2,478.9
	Peak Memory (Gb)	77.5	40.1		464.2	
Z. mays (primary)	Elapsed time (h)	3.7	46.1		35.3	
	CPU time (h)	146.4	1,789.4			3,016.0
	Peak Memory (Gb)	71.1	65.5		319.0	
F. × ananassa (primary)	Elapsed time (h)	3.9	19.0	19.5	2.7	
	CPU time (h)	159.6	641.0	656.7		979.8
	Peak Memory (Gb)	92.2	15.8	15.8	206.3	
R. muscosa (primary)	Elapsed time (h)	69.0				
	CPU time (h)	2,856.5				
	Peak Memory (Gb)	484.1				
S. sempervirens (primary)	Elapsed time (h)	72.9			375.2	
	CPU time (h)	4,492.0				
	Peak Memory (Gb)	719.7			>3,000	
CHM13 (primary)	Elapsed time (h)	7.6	49.2		3.9	
	CPU time (h)	292.7	1,057.2			1,541.8
	Peak Memory (Gb)	124.7	91.6		684.5	
HG00733 (primary)	Elapsed time (h)	7.7	47.4	51.4	4.7	
	CPU time (h)	281.9	1,133.4	1,256.8		2,585.6
	Peak Memory (Gb)	134.9	70.8	70.8	707.2	
HG002 (primary)	Elapsed time (h)	9.4	75.5	78.7	4.9	
	CPU time (h)	357.3	1,174.3	1236.4		3,349.1
	Peak Memory (Gb)	142.9	83.7	83.7	749.8	
HG00733 (trio)	Elapsed time (h)	9.5	43.1		10.2	
	CPU time (h)	331.0	925.5			
	Peak Memory (Gb)	134.9	42.5		445.1	
HG002 (trio)	Elapsed time (h)	10.8	45.4		9.7	
	CPU time (h)	403.2	986.1			
	Peak Memory (Gb)	142.9	41.1		467.8	

The majority of hifiasm, HiCanu and Peregrine assemblies were generated using a same machine with 48 CPU threads. The hifiasm *S. sempervirens* assembly was produced on another machine with 80 CPU threads. All Falcon assemblies, the Peregrine *R. muscosa* and *S. sempervirens* assemblies and the HiCanu *R. muscosa* assembly were generated using a cluster. Peregrine was run using docker. It is difficult to measure all the metrics of these assemblies without the assemblers logging the information.



Supplementary Fig. 1: Constructing primary assembly from string graph. **(a)** Each bubble in the graph is reduced into a single path using bubble popping. **(b)** Given a tip unitig T in yellow that is broken in one end but connected to the green unitig C in another end, hifiasm checks if there is another blue unitig G, which is also connected to C, coming from the different haplotypes of T. In this case, T should be discarded from the primary assembly. **(c)** If G does not come from the different haplotypes of T, hifiasm uses the "best overlap graph" strategy to break this branch but still keeps T in the primary assembly.