nature portfolio

https://doi.org/10.1038/s41587-022-01435-7

Supplementary information

DeepConsensus improves the accuracy of sequences with a gap-aware sequence transformer

In the format provided by the authors and unedited

Supplementary Notes

Software commands

Generating DeepConsensus input files with pbmm2 and samtools

We started with the subreads BAM and CCS reads BAM files output by pbccs. Samtools v1.1 was used to convert the CCS reads BAM to a FASTA:

```
samtools fasta "ccs.bam" > "ccs.fasta"
```

We used actc v0.0.2 and pbmm2 v1.4.0 for mapping subreads to CCS reads using the following commands:

```
actc -j "$(nproc)" "subreads.bam" "ccs.bam" "aligned.subreads.bam"

pbmm2 align --sample "sample" --preset SUBREAD --sort "ccs.fasta"
"subreads.bam" "aligned.subreads.bam"
```

Only subread alignments to the correct molecule were retained. We used samtools and awk to filter incorrect alignments using the below command:

```
samtools view -h "aligned.subreads.bam" | \
  awk '{ if($1 ~ /^@/) { print; } else { split($1,A,"/"); \
  split($3,B,"/"); if(A[2]==B[2]) { split(A[3],C,"_"); \
  print $0 "\tqs:i:" C[1]; } } }'
```

Running inference with DeepConsensus

DeepConsensus v0.1 was run with the following command:

```
python3 -m deepconsensus.scripts.run_deepconsensus \
  --input_subreads_aligned=subreads.aligned.bam \
  --input_subreads_unaligned=subreads.bam \
  --input_ccs_fasta=ccs.fasta \
  --output_directory=deepconsensus_output \
  --checkpoint=model-50
```

The outputs are sharded FASTQs, which are concatenated together with cat to produce the final FASTQ.

Aligning DeepConsensus predictions with pbmm2

DeepConsensus predictions and CCS reads from pbccs were mapped to the HG002 diploid assembly or reference genome using the below command. The `--unmapped` flag was added to the below command for assembly and variant calling analysis pipelines:

```
pbmm2 align --sample "sample" --preset HIFI --sort -c 0 -y 70
"ref.fa" "query.reads.fasta" "aligned.reads.bam"
```

Generating phased diploid assemblies with hifiasm

HiFiasm (version 0.15.3-r339) was run with the following command to generate the diploid human genome assemblies:

```
hifiasm -o "outputPrefix" -t "nThreads" "reads.fastq"
```

We ran the following commands to convert the output GFA files to fasta files.
For the primary assembly we used:

```
awk '/^S/{print ">"$2;print $3}' "outputPrefix.bp.p_ctg.gfa" >
"outputPrefix.bp.p_ctg.fa"
```

For the haplotype-resolved GFA files for the diploid sample, we used the following commands to get haplotype-resolved assembly:

```
awk '/^S/{print ">"$2;print $3}' "outputPrefix.bp.hap1.p_ctg.gfa" >
"outputPrefix.bp.hap1.p_ctg.fa"

awk '/^S/{print ">"$2;print $3}' "outputPrefix.bp.hap2.p_ctg.gfa" >
"outputPrefix.bp.hap2.p_ctg.fa"
```

Reference free assembly quality estimation with YAK

We used yak (version 0.1-r62-dirty) to derive estimated consensus accuracy (Q) of the assemblies.

To evaluate the assembly of a sample we first built the k-mer database from Illumina paired-end short reads of the same sample using this command:

```
yak count -b37 -t32 -o "KMER_DB" <(zcat sr*.fq.gz) <(zcat sr*.fq.gz)
```

Then we estimated the quality of the phased assembly using the following command:

```
yak qv -t "nThreads" -p -K 3.2g -l 100k "KMER_DB"
"outputPrefix.bp.hap1.p_ctg.fa" > "outputPrefix.hap1.yak.qv.txt"

yak qv -t "nThreads" -p -K 3.2g -l 100k "KMER_DB"
"outputPrefix.bp.hap2.p_ctg.fa" > "outputPrefix.hap2.yak.qv.txt"
```

From the outputs `"outputPrefix.hap1.yak.qv.txt"` and `"outputPrefix.hap2.yak.qv.txt"` we reported the balanced error as the quality of the assembly we evaluated.

Assembly-based small variant calling assessment using dipcall

We used dipcall (version 0.3) to derive small variants from the phased assemblies.

For male samples such as HG002, HG003, HG006, we used the following command:

```
dipcall.kit/run-dipcall -x dipcall.kit/hs38.PAR.bed
"outputPrefix.dipcall_output" "reference"
"outputPrefix.bp.hap1.p_ctg.fa" "outputPrefix.bp.hap2.p_ctg.fa" >
dipcall_instructions.mak

make -j2 -f dipcall_instructions.mak
```

For female samples such as HG004, HG007 we used the following command:

```
dipcall.kit/run-dipcall "outputPrefix.dipcall_output" "reference"
"outputPrefix.bp.hap1.p_ctg.fa" "outputPrefix.bp.hap2.p_ctg.fa" >
dipcall_instructions.mak

make -j2 -f dipcall_instructions.mak
```

From the output we compared `outputPrefix.dipcall_output.dip.vcf.gz` variant call set against GIAB truth set using hap.py small variant evaluation program. For all samples, we used GIAB v4.2.1 benchmarking set and GRCh38 as the reference. The command for hap.py is provided in the associated section.

Gene completeness assessment with asmgene

We assessed the gene completeness of the phased assemblies with asmgene (version v2.21). We first aligned the Ensembl cDNA sequences to the GRCh38 reference genome using minimap2 (v2.21) using the following command:

```
minimap2 -cxsplice:hq -t32 "GRCh38.fa" "homo_sapiens_cdna.fa" >
"asmgene_output_prefix.ref.cdna.paf"
```

Then we align the cDNA sequences to each haplotype of the phased assemblies.

```
minimap2 -cxsplice:hq -t32 "outputPrefix.bp.hap1.p_ctg.fa"
"homo_sapiens_cdna.fa" > "asmgene_output_prefix.HAP1.cdna.paf"

minimap2 -cxsplice:hq -t32 "outputPrefix.bp.hap2.p_ctg.fa"
"homo_sapiens_cdna.fa" > "asmgene_output_prefix.HAP2.cdna.paf"
```

Then we use asmgene tool available from paftools.js to derive the gene completeness metrics for each haplotype.

```
paftools.js asmgene -a "asmgene_output_prefix.ref.cdna.paf"
"asmgene_output_prefix.HAP1.cdna.paf" >
"asmgene_output_prefix.hap1.genestats"

paftools.js asmgene -a "asmgene_output_prefix.ref.cdna.paf"
"asmgene_output_prefix.HAP2.cdna.paf" >
"asmgene_output_prefix.hap2.genestats"
```

Assembly statistics with QUAST

We used QUAST (version v5.0.2) to derive the assembly size, N50 and NG50 metrics of the assembly against the GRCh38 reference sequence. For the assembly size, N50 and NG50 values, we used the primary assembly we get from hifiasm assembler. We run the following command to derive the assembly metrics:

```
python /root/tools/quast/quast-5.0.2/quast-lg.py -t "nThreads" -o
"QUAST_output_directory" -r "GRCh38.fa" --large
outputPrefix.bp.p_ctg.fa
```

Within the output directory the `report.txt` file contains the assembly size, N50 and NG50 numbers that we report in our evaluation.

Variant calling with DeepVariant

We ran DeepVariant with the provided convenience script to call variants. Here is the command used:

```
/opt/deepvariant/bin/run_deepvariant --model_type="PACBIO"
--use_hp_information --ref="ref.fa" --reads="reads.bam"
--output_vcf="output.vcf" --num_shards="$(nproc)"
```

For the pbccs baseline, we used the latest DeepVariant model for PacBio data, v1.2. For DeepConsensus, we trained a custom model which was specified using the `--customized_model` in the above command.

Variant calls were evaluated using v0.3.12 of hap.py from the jmcdani20/hap.py Docker image. The command used was:

```
/opt/hap.py/bin/hap.py "truth.vcf" "output.vcf" -f "truth.bed" -r
"ref.fa" -o "output/happy.output"
--stratification="v2.0-GRCh38-stratifications.tsv" --engine=vcfeval
--pass-only
```

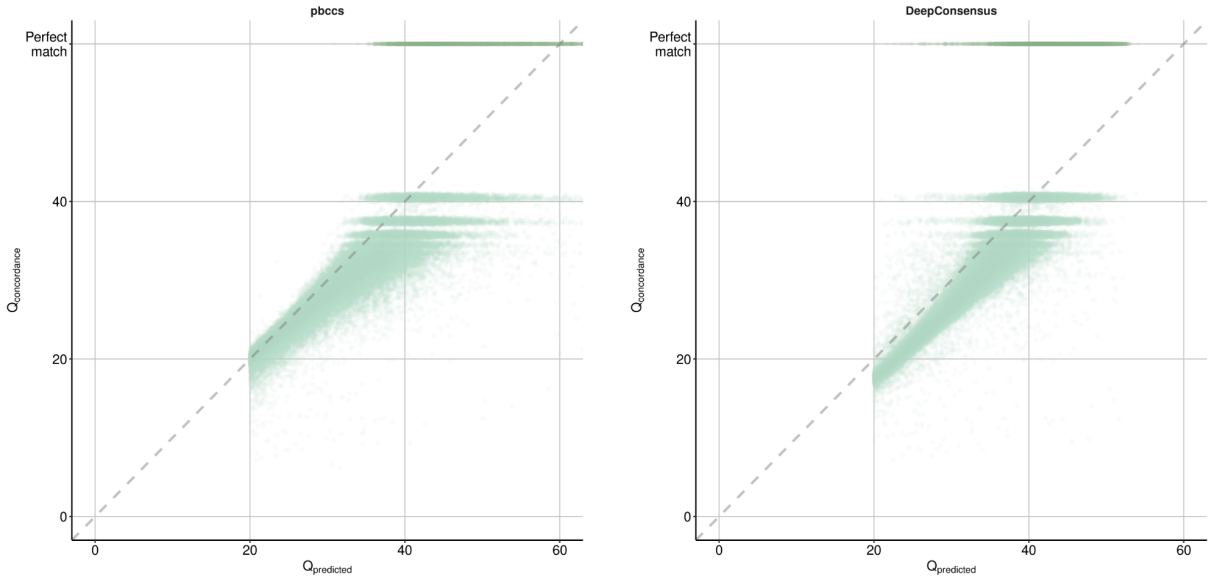
Truth VCF and BED files come from the Genome in a Bottle (GIAB) truth sets. The v4.2.1 truth set was used for all samples.

Comparison with *Lal et al.*

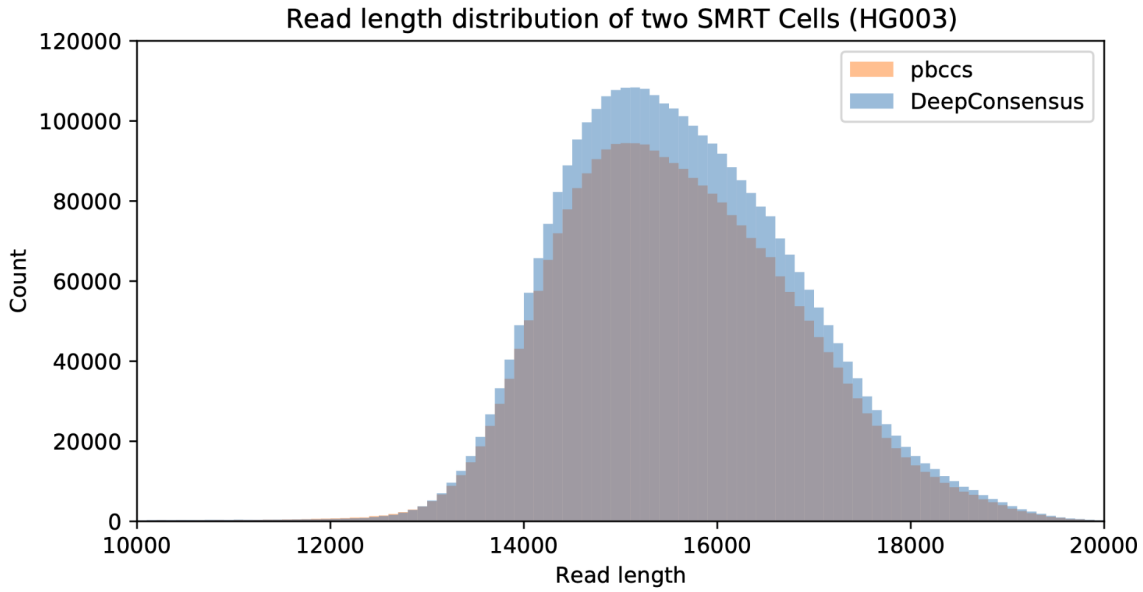
Sequence data from *Lal et al* was retrieved from https://nv-pacbio.s3.us-east-2.amazonaws.com/paper_results.zip. The file `sequelii_after.fastq` consists of reads polished using the approach presented in *Lal et al.* We extracted 760 of 3,586 reads that aligned to chromosome 20, which was present in both our test dataset and the *Lal et al.* test dataset. We then ran bamConcordance to compare the accuracy of both approaches. We calculate base pair errors (bp-errors) as the sum of the lengths of mismatch and indel errors.

Supplementary results

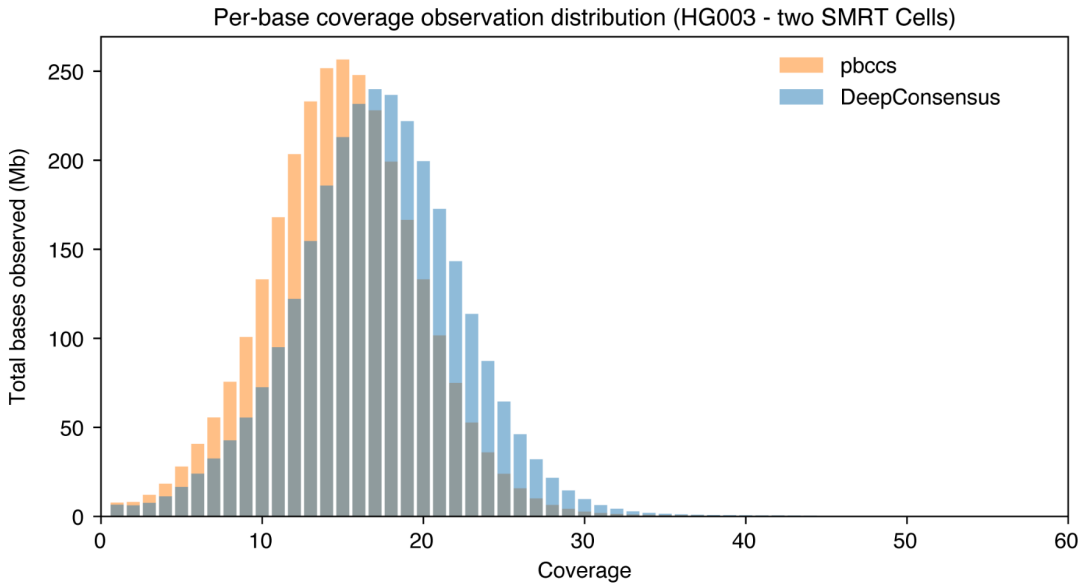
Supplementary figures



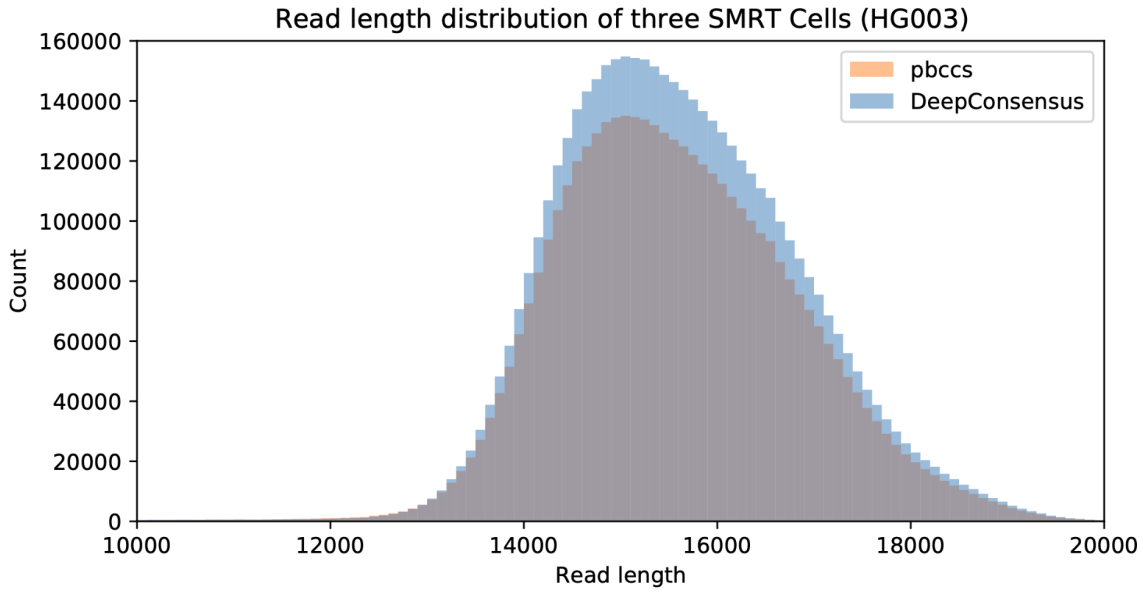
Supplementary Figure 1: The observed read quality ($Q_{concordance}$) and predicted read quality ($Q_{predicted}$) are plotted against one another for both pbccs and DeepConsensus. Reads are from HG002 chr20 and have a length of 11kb.



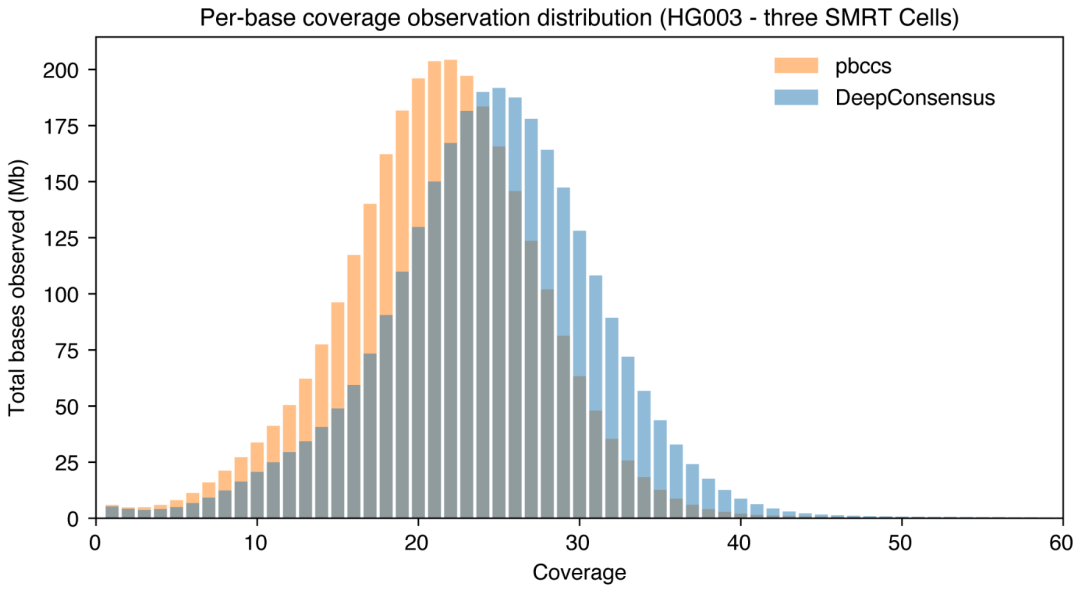
Supplementary Figure 2: Read length distribution of HG003 two SMRT Cells reads from pbccs and DeepConsensus.



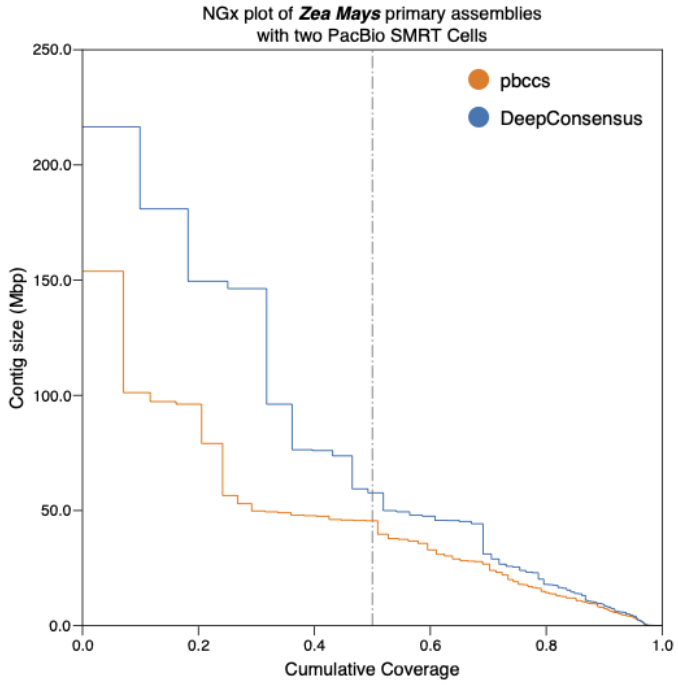
Supplementary Figure 3: Distribution of observed coverage of HG003 two SMRT Cells reads from pbccs and DeepConsensus.



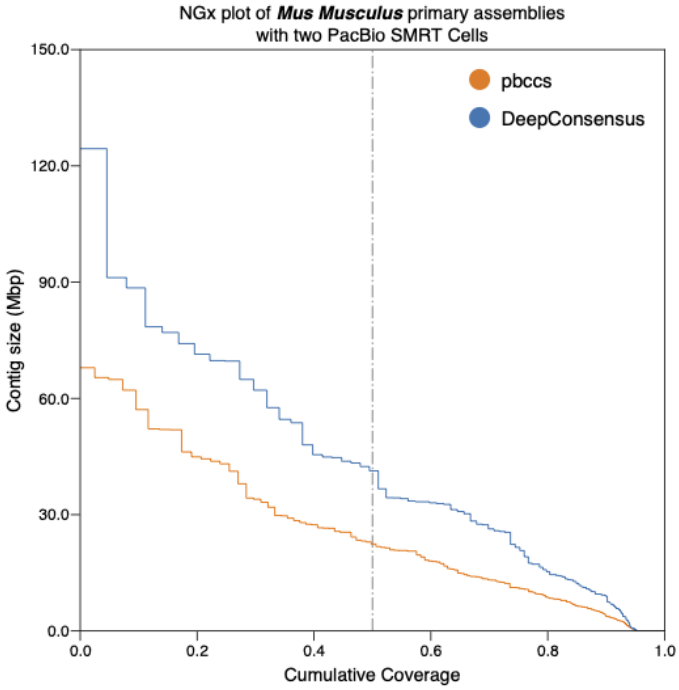
Supplementary Figure 4: Read length distribution of HG003 three SMRT Cells reads from pbccs and DeepConsensus.



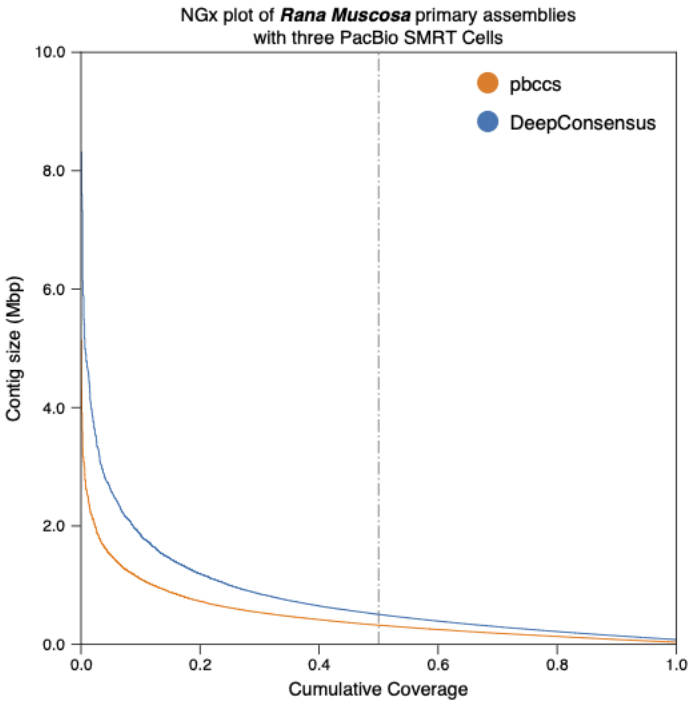
Supplementary Figure 5: Distribution of observed coverage of HG003 two SMRT Cells reads from pbccs and DeepConsensus.



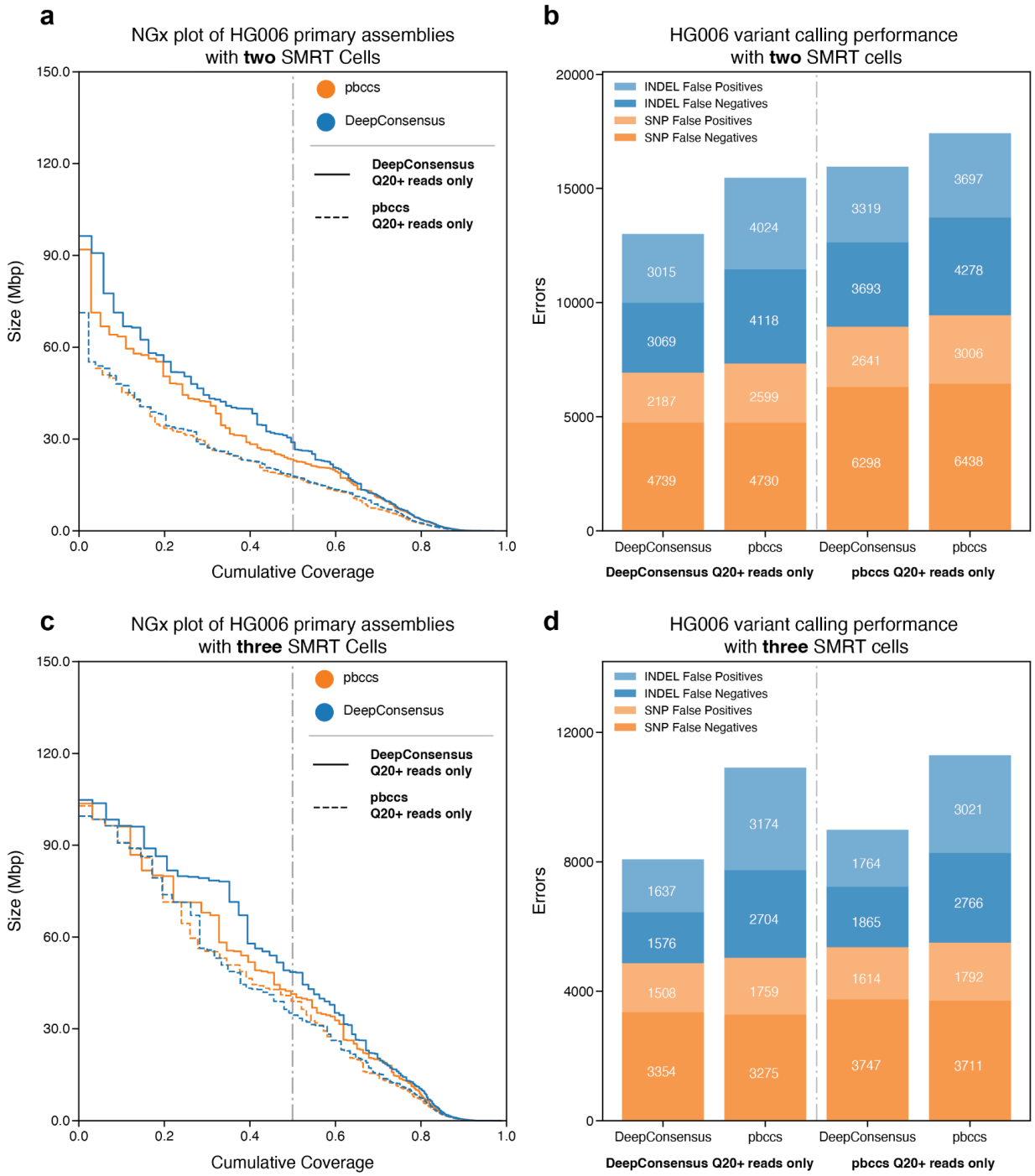
Supplementary Figure 6: Contiguity of the hifiasm assemblies of Maize B73 sample with reads from pbccs and DeepConsensus from two PacBio SMRT Cells.



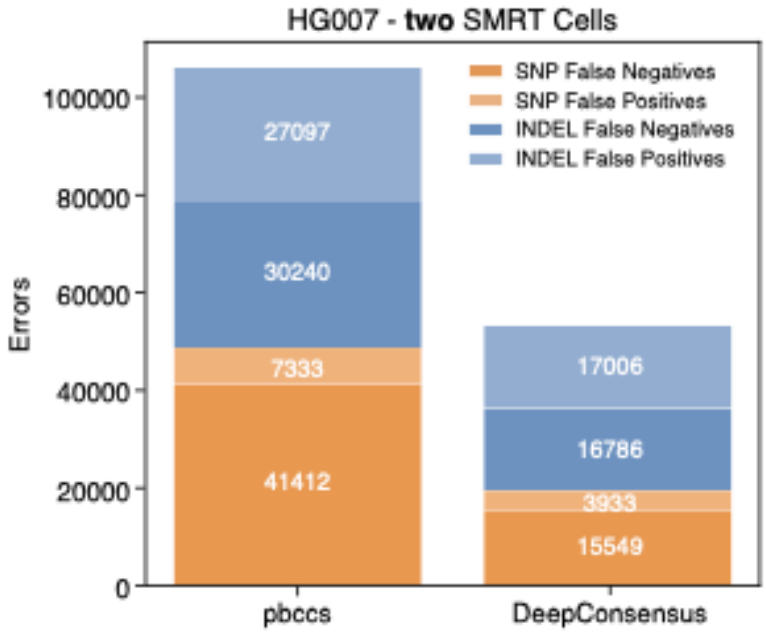
Supplementary Figure 7: Contiguity of the hifiasm assemblies of Mus Musculus sample with reads from pbccs and DeepConsensus from two PacBio SMRT Cells.



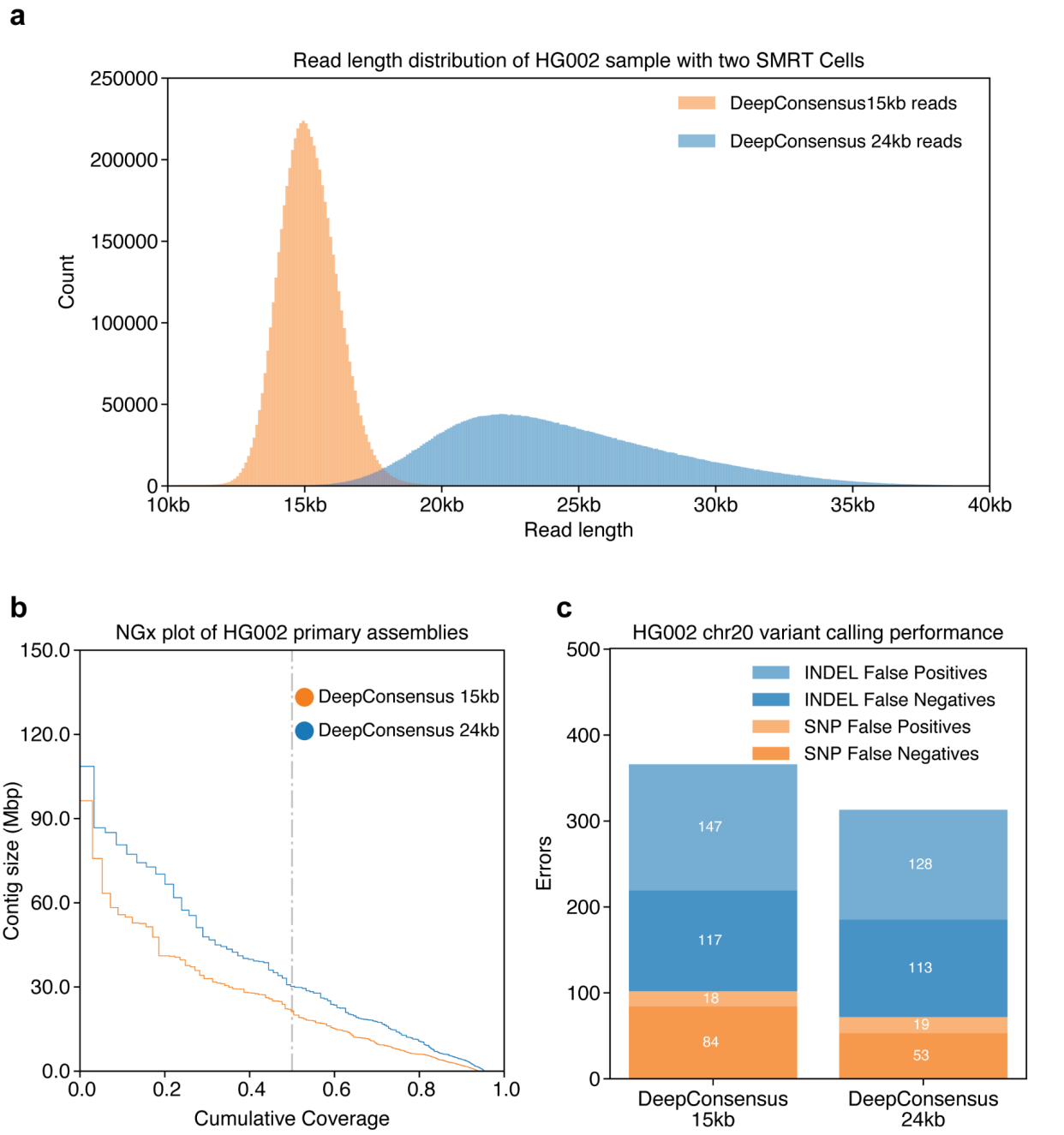
Supplementary Figure 8: Contiguity of the hifiasm assemblies of Rana muscosa sample with reads from pbccs and DeepConsensus from three PacBio SMRT Cells.



Supplementary Figure 9: To isolate the impact of accuracy improvements from coverage gains, analyses were run with the identical set of reads processed with DeepConsensus and pbccs. (a) Contiguity of the hifiasm assemblies using Q20+ reads from each method for two HG006 PacBio SMRT Cells. (b) HG006 variant calling performance using Q20+ reads from each method for two PacBio SMRT Cells. (c) Contiguity of the hifiasm assemblies using Q20+ reads from each method for three HG006 PacBio SMRT Cells. (d) HG006 variant calling performance using Q20+ reads from each method for three PacBio SMRT Cells.



Supplementary Figure 10: Variant calling performance with pbccs and DeepConsensus reads on HG007 sample.



Supplementary Figure 11: DeepConsensus with longer reads improves genome assembly contiguity. (a) HG002 read length distribution for 15kb and 24kb DeepConsensus reads from two SMRT Cells. (b) Contiguity of the HG002 hifiasm assembly with 15kb and 24kb DeepConsensus reads from two SMRT Cells. (c) HG002 variant calling performance for 15kb and 24kb reads from DeepConsensus for two SMRT Cells.

Supplementary tables

Supplementary Table 1: Error reduction on the test split for DeepConsensus models using subsets of input signals and either the custom alignment loss or average cross-entropy loss computed over each predicted position. The released DeepConsensus model uses alignment loss + bases + pulse width (PW) + interpulse duration (IP) + CCS sequence from pbccs + strand for each subread + signal-to noise-ratios (SN). Error reduction here is shown for chr19 and chr20 and includes reads <Q20, which is why the error reduction of the published model is slightly different from other comparisons.

DeepConsensus experiment	Error reduction compared to pbccs
Alignment loss + bases	11.11%
Alignment loss + bases + PW	20.96%
Alignment loss + bases + PW + IP	30.40%
Alignment loss + bases + PW + IP + CCS	31.81%
Alignment loss + bases + PW + IP + CCS + strand	34.53%
Alignment loss + bases + PW + IP + CCS + strand + SN (published model)	37.57%
Cross-entropy loss + bases + PW + IP + CCS + strand + SN	9.31%

Supplementary Table 2: Q-concordance comparison on 760 reads from chr20, which is the only chromosome available across all test datasets.

Dataset	$Q_{concordance}$
pbccs	27.5987
Lal et al.	29.1858
DeepConsensus	29.9320

Supplementary Table 3: Comparison of average number of errors per 10kbp for each error class on 760 reads from chr20, which is the only chromosome available across all test datasets.

Dataset	All Errors (per 10kbp)	Mismatch (per 10kbp)	Non-Homopolymer Insertion (per 10kbp)	Non-Homopolymer Deletion (per 10kbp)	Homopolymer Insertion (per 10kbp)	Homopolymer Deletion (per 10kbp)
pbccs	17.29	1.31	1.13	0.43	7.54	6.88
Lal et al.	12.08	1.08	0.36	0.41	4.69	5.54
DeepConsensus	10.09	0.66	0.18	0.51	3.66	5.08

Supplementary Table 4: DeepConsensus and pbccs yield for reads at Q20 (predicted 99% accuracy) or higher.

Sample	Read Length	SMRT Cells	Yield	
			pbccs	DeepConsensus
HG003	15 kb	2	46.15 Gb	53.02 Gb
		3	65.62 Gb	75.41 Gb
HG004	15 kb	2	45.55 Gb	52.28 Gb
		3	64.70 Gb	74.26 Gb
HG006	15kb	2	61.10 Gb	69.24 Gb
		3	83.17 Gb	96.56 Gb
HG007	15kb	2	39.56 Gb	49.54 Gb

Supplementary Table 5: QUAST reported assembly statistics of hifiasm assemblies with pbccs and DeepConsensus reads.

Sample	SMRT Cells	Method	N50 (Mb)	NG50 (Mb)	Assembly size (Gb)	Assembly completeness against GRCh38 (%)
HG003 15kb	2	pbccs	4.74	4.91	3.16	**97.67**
		DeepConsensus	**17.23**	**17.23**	3.1	97.62
	3	pbccs	33.64	33.64	3.1	97.88
		DeepConsensus	**55.54**	**55.54**	3.1	**97.94**
HG004 15kb	2	pbccs	3.43	3.72	3.23	97.31
		DeepConsensus	**12.76**	**12.37**	3.07	**97.34**
	3	pbccs	36.24	36.24	3.07	97.44
		DeepConsensus	**41.07**	**41.07**	3.06	**97.47**
HG006 15kb	2	pbccs	18.89	18.55	3.01	97.37
		DeepConsensus	**31.68**	**31.54**	3.02	**97.5**
	3	pbccs	42.87	41.19	3.04	97.76
		DeepConsensus	**51.68**	**51.68**	3.04	**97.88**
HG007 15kb	2	pbccs	1.89	1.94	3.15	97.1
		DeepConsensus	**8.52**	**8.48**	3.07	**97.36**

Supplementary Table 6: YAK estimated phred-scale Q value of hifiasm assemblies with PBCCS and DeepConsensus reads.

Sample	SMRT Cells	Method	HAP1 Q	HAP2 Q
HG003 15kb	2	pbccs	43.391	43.066
		DeepConsensus	**45.107**	**45.181**
	3	pbccs	48.68	48.569
		DeepConsensus	**50.281**	**50.425**
HG004 15kb	2	pbccs	43.446	43.096
		DeepConsensus	**44.916**	**44.925**
	3	pbccs	48.881	48.547
		DeepConsensus	**50.333**	**50.404**
HG006 15kb	2	pbccs	45.778	45.806
		DeepConsensus	**46.637**	**46.503**
	3	pbccs	46.972	46.868
		DeepConsensus	**47.421**	**47.483**
HG007 15kb	2	pbccs	39.219	39.067
		DeepConsensus	**42.052**	**41.935**

Supplementary Table 7: Assembly-based small variant calling SNP evaluation of hifiasm assemblies with pbccs and DeepConsensus reads.

Sample	SMRT Cells	Method	SNPs					
			True positives	False negatives	False positives	Precision	Recall	F1-score
HG003 15kb	2	pbccs	3011035	316460	31051	0.9898	0.9049	0.9454
		DeepConsensus	**3141265**	**186230**	**28454**	**0.991**	**0.944**	**0.9670**
	3	pbccs	3235037	92458	9978	0.9969	0.9722	0.9844
		DeepConsensus	**3262865**	**64630**	**8834**	**0.9973**	**0.9806**	**0.9889**
HG004 15kb	2	pbccs	2902441	444169	34944	0.9881	0.8673	0.9238
		DeepConsensus	**3110169**	**236441**	**31955**	**0.9898**	**0.9293**	**0.9586**
	3	pbccs	3220904	125706	**12159**	0.9962	0.9624	0.9790
		DeepConsensus	**3268274**	**78336**	12173	**0.9963**	**0.9766**	**0.9863**
HG006 15kb	2	pbccs	3093947	175913	13177	0.9958	0.9462	0.9703
		DeepConsensus	**3174650**	**95210**	**11817**	**0.9963**	**0.9709**	**0.9834**
	3	pbccs	3198263	71597	**10844**	**0.9966**	0.9781	0.9873
		DeepConsensus	**3214917**	**54943**	11265	0.9965	**0.9832**	**0.9898**
HG007 15kb	2	pbccs	2620694	663768	52719	0.9803	0.7979	0.8797
		DeepConsensus	**2919344**	**365118**	**46821**	**0.9842**	**0.8888**	**0.9341**

Supplementary Table 8: Assembly-based small variant calling INDEL evaluation of hifiasm assemblies with pbccs and DeepConsensus reads.

Sample	SMRT Cells	Method	INDELs					
			True positives	False negatives	False positives	Precision	Recall	F1-score
HG003 15kb	2	pbccs	439410	65091	221210	0.6725	0.871	0.7590
		DeepConsensus	**462380**	**42121**	**128197**	**0.7888**	**0.9165**	**0.8478**
	3	pbccs	479075	25426	48570	0.9108	0.9496	0.9298
		DeepConsensus	**485407**	**19094**	**25203**	**0.9523**	**0.9622**	**0.9572**
HG004 15kb	2	pbccs	425671	84848	218034	0.6686	0.8338	0.7421
		DeepConsensus	**460158**	**50361**	**145983**	**0.7656**	**0.9014**	**0.8279**
	3	pbccs	478488	32031	56204	0.8981	0.9373	0.9173
		DeepConsensus	**488804**	**21715**	**32312**	**0.94**	**0.9575**	**0.9487**
HG006 15kb	2	pbccs	395841	32397	53935	0.8831	0.9243	0.9033
		DeepConsensus	**407094**	**21144**	**30961**	**0.9312**	**0.9506**	**0.9408**
	3	pbccs	409551	18687	22326	0.9497	0.9564	0.953
		DeepConsensus	**413053**	**15185**	**13441**	**0.9694**	**0.9645**	**0.9669**
HG007 15kb	2	pbccs	318117	112209	526259	0.3825	0.7392	0.5041
		DeepConsensus	**366759**	**63567**	**244628**	**0.6064**	**0.8523**	**0.7086**

Supplementary Table 9: Asmgene single-copy gene completeness analysis of hifiasm assemblies with pbccs and DeepConsensus reads.

Sample	SMRT Cells	Method	# Single copy ref	Single copy		False duplications		Complete (%)	Duplicated (%)
				Hap1	Hap2	Hap1	Hap2		
HG003 15kb	2	pbccs	35374	32045	31682	623	375	90.076	1.411
		DeepConsensus	35374	**32882**	**32864**	**301**	**203**	**92.93**	**0.712**
	3	pbccs	35374	33852	33668	180	141	95.437	0.454
		DeepConsensus	35374	**34201**	**34302**	**126**	**134**	**96.827**	**0.368**
HG004 15kb	2	pbccs	35374	30973	30779	1250	647	87.284	2.681
		DeepConsensus	35374	**32306**	**31951**	**279**	**202**	**90.825**	**0.68**
	3	pbccs	35374	33512	33183	225	142	94.271	0.519
		DeepConsensus	35374	**33897**	**33898**	**145**	**140**	**95.826**	**0.403**
HG006 15kb	2	pbccs	35374	33149	33145	**184**	**143**	93.704	**0.462**
		DeepConsensus	35374	**33893**	**33765**	186	**143**	**95.632**	0.465
	3	pbccs	35374	34134	34021	146	139	96.335	0.403
		DeepConsensus	35374	**34399**	**34341**	**141**	**122**	**97.162**	**0.372**
HG007 15kb	2	pbccs	35374	29284	29056	649	453	82.462	1.558
		DeepConsensus	35374	**30935**	**30925**	**309**	**226**	**87.437**	**0.756**

Supplementary Table 10: Asmgene multi-copy gene completeness analysis of hifiasm assemblies with pbccs and DeepConsensus reads.

Sample	SMRT Cells	Method		Multi copy		Complete multi copy (%)	Missing multi copy (%)
			# Multi copy ref	Hap1	Hap2		
HG003 15kb	2	pbccs	1253	968	936	75.978	24.022
		DeepConsensus	1253	**980**	**982**	**78.292**	**21.708**
	3	pbccs	1253	**1006**	983	**79.37**	**20.63**
		DeepConsensus	1253	988	**989**	78.891	21.109
HG004 15kb	2	pbccs	1253	932	**959**	75.459	24.541
		DeepConsensus	1253	**967**	942	**76.177**	**23.823**
	3	pbccs	1253	**1007**	974	**79.05**	**20.95**
		DeepConsensus	1253	993	**987**	79.01	20.99
HG006 15kb	2	pbccs	1253	964	**995**	78.172	21.828
		DeepConsensus	1253	**1019**	957	**78.851**	**21.149**
	3	pbccs	1253	1003	**995**	79.729	20.271
		DeepConsensus	1253	**1042**	972	**80.367**	**19.633**
HG007 15kb	2	pbccs	1253	886	851	69.314	30.686
		DeepConsensus	1253	**920**	**951**	**74.661**	**25.339**

Supplementary Table 11: QUAST reported assembly statistics of hifiasm assemblies for Maize and Mus Musculus. We were not able to successfully run Quast on DeepConsensus Maize reads with the available resources.

Sample	Read length	SMRT Cells	Method	N50 (Mb)	NG50 (Mb)	Assembly size (Gb)	Assembly completeness against reference
Maize	15kb	2	pbccs	45.5	45.5	2.19	98.689
			DeepConsensus	**57.6**	**57.6**	2.19	N/A
Mus Musculus	15kb	2	pbccs	23.4	22.43	2.61	95.666
			DeepConsensus	**43.33**	**41.31**	2.61	**95.728**

Supplementary Table 12: Asmgene single-copy gene completeness analysis of Maize and Mus Musculus hifiasm assemblies with pbccs and DeepConsensus reads.

Sample	Read length	SMRT Cells	Method	# Single copy ref	Single copy	False duplications	Complete (%)	Duplicated (%)
					Primary	Primary		
Maize	15kb	2	pbccs	30651	**30562**	**50**	**99.71**	**0.163**
			DeepConsensus	30651	**30562**	57	**99.71**	0.186
Mus Musculus	15kb	2	pbccs	20179	20066	55	99.44	0.273
			DeepConsensus	20179	**20084**	**53**	**99.529**	**0.263**

Supplementary Table 13: Asmgene multi-copy gene completeness analysis of Mus Musculus hifiasm assemblies with pbccs and DeepConsensus reads.

Sample	Read length	SMRT Cells	Method	# Multi copy ref	Multi copy	Complete multi copy (%)	Missing multi copy (%)
					Primary		
Mus Musculus	15kb	2	pbccs	1015	856	84.335	15.665
			DeepConsensus	1015	**863**	**85.025**	**14.975**

Supplementary Table 14: QUAST reported assembly statistics of hifiasm assemblies with PBCCS and DeepConsensus on ZMWs above Q20 for each method. DC-HiFi-reads sample includes all reads that pass the Q20 filter for DeepConsensus. PBCCS-HiFi-reads includes all reads that pass the Q20 filter for PBCCS.

Sample	SMRT Cells	Method	N50 (Mb)	NG50 (Mb)	Assembly size (Gb)	Assembly completeness against GRCh38 (%)
HG006 DC-HiFi-reads	2	pbccs	25.1	24.56	3.02	97.58
		DeepConsensus	31.68	31.54	3.02	97.5
	3	pbccs	44.4	42.93	3.04	97.69
		DeepConsensus	51.68	51.68	3.04	97.88
HG006 PBCCS-HiFi-reads	2	pbccs	18.89	18.55	3.01	97.37
		DeepConsensus	20.24	18.89	3.01	97.34
	3	pbccs	42.87	41.19	3.04	97.76
		DeepConsensus	38.96	38.86	3.04	97.84

Supplementary Table 15: YAK estimated phred-scale QV of hifiasm assemblies using ZMWs above Q20 for each method. DC-HiFi-reads sample includes all reads that pass the Q20 filter for DeepConsensus. PBCCS-HiFi-reads includes all reads that pass the Q20 filter for PBCCS.

Sample	SMRT Cells	Method	HAP1 QV	HAP2 QV
HG006 DC-HiFi-reads	2	pbccs	45.346	45.375
		DeepConsensus	46.637	46.503
	3	pbccs	46.399	46.212
		DeepConsensus	47.421	47.483
HG006 PBCCS-HiFi-reads	2	pbccs	45.778	45.806
		DeepConsensus	46.995	46.902
	3	pbccs	46.972	46.868
		DeepConsensus	47.786	47.727

Supplementary Table 16: Assembly-based small variant calling evaluation of hifiasm assemblies with ZMWs above Q20 for each method. DC-HiFi-reads sample includes all reads that pass the Q20 filter for DeepConsensus. PBCCS-HiFi-reads includes all reads that pass the Q20 filter for PBCCS.

Sample	SMRT Cells	Method	SNPs					
			True positives	False negatives	False positives	Precision	Recall	F1-score
HG006 DC-HiFi-reads	2	pbccs	3163142	106718	14358	0.9955	0.9674	0.9812
		DeepConsensus	3174650	95210	11817	0.9963	0.9709	0.9834
	3	pbccs	3209208	60652	10929	0.9966	0.9815	0.989
		DeepConsensus	3214917	54943	11265	0.9965	0.9832	0.9898
HG006 PBCCS-HiFi-reads	2	pbccs	3093947	175913	13177	0.9958	0.9462	0.9703
		DeepConsensus	3091516	178344	11265	0.9964	0.9455	0.9702
	3	pbccs	3198263	71597	10844	0.9966	0.9781	0.9873
		DeepConsensus	3196618	73242	8672	0.9973	0.9776	0.9874

Supplementary Table 17: Assembly-based small variant calling evaluation of hifiasm assemblies with ZMWs above Q20 for each method. DC-HiFi-reads sample includes all reads that pass the Q20 filter for DeepConsensus. PBCCS-HiFi-reads includes all reads that pass the Q20 filter for PBCCS.

Sample	SMRT Cells	Method	INDELs					
			True positives	False negatives	False positives	Precision	Recall	F1-score
HG006 DC-HiFi-reads	2	pbccs	404384	23854	48707	0.8952	0.9443	0.9191
		DeepConsensus	407094	21144	30961	0.9312	0.9506	0.9408
	3	pbccs	410918	17320	21145	0.9524	0.9596	0.956
		DeepConsensus	413053	15185	13441	0.9694	0.9645	0.9669
HG006 PBCCS-HiFi-reads	2	pbccs	395841	32397	53935	0.8831	0.9243	0.9033
		DeepConsensus	396632	31606	29926	0.9317	0.9262	0.929
	3	pbccs	409551	18687	22326	0.9497	0.9564	0.953
		DeepConsensus	411029	17209	12264	0.9718	0.9598	0.9658

Supplementary Table 18: Asmgene single-copy gene completeness analysis of hifiasm assemblies with ZMWs above Q20 for each method. DC-HiFi-reads sample includes all reads that pass the Q20 filter for DeepConsensus. PBCCS-HiFi-reads includes all reads that pass the Q20 filter for PBCCS.

Sample	SMRT Cells	Method	# Single copy ref	Single copy		False duplications		Complete (%)	Duplicated (%)
				Hap1	Hap2	Hap1	Hap2		
HG006 DC-HiFi-reads	2	pbccs	35374	33768	33777	208	136	95.473	0.486
		DeepConsensus	35374	33893	33765	186	143	95.632	0.465
	3	pbccs	35374	34164	34332	137	158	96.817	0.417
		DeepConsensus	35374	34399	34341	141	122	97.162	0.372
HG006 PBCCS-HiFi-reads	2	pbccs	35374	33149	33145	184	143	93.704	0.462
		DeepConsensus	35374	33111	33098	171	138	93.584	0.437
	3	pbccs	35374	34134	34021	146	139	96.335	0.403
		DeepConsensus	35374	33837	34052	145	136	95.959	0.397

Supplementary Table 19: Asmgene multi-copy gene completeness analysis of hifiasm assemblies with ZMWs above Q20 for each method. DC-HiFi-reads sample includes all reads that pass the Q20 filter for DeepConsensus. PBCCS-HiFi-reads includes all reads that pass the Q20 filter for PBCCS.

Sample	SMRT Cells	Method	# Multi copy ref	Multi copy		Complete multi copy (%)	Missing multi copy (%)
				Hap1	Hap2		
HG006 DC-HiFi-reads	2	pbccs	1253	967	1018	79.21	20.79
		DeepConsensus	1253	1019	957	78.851	21.149
	3	pbccs	1253	972	1039	80.247	19.753
		DeepConsensus	1253	1042	972	80.367	19.633
HG006 PBCCS-HiFi-reads	2	pbccs	1253	964	995	78.172	21.828
		DeepConsensus	1253	1041	951	79.489	20.511
	3	pbccs	1253	1003	995	79.729	20.271
		DeepConsensus	1253	982	1040	80.686	19.314

Supplementary Table 20: Whole genome variant calling results for DeepConsensus and pbccs reads.

Sample	SMRT Cells	Method	Type	Recall	Precision	F1 Score	False Negatives	False Positives
HG003 15kb	2	pbccs	INDEL	0.969544	0.977546	0.973528	15365	11674
			SNP	0.994683	0.998822	0.996748	17693	3907
		DeepConsensus + re-trained DV	INDEL	**0.976595**	**0.978256**	**0.977425**	**11808**	**11383**
			SNP	0.996891	**0.99904**	0.997964	10345	**3191**

		DeepConsensus + DeepVariant v1.2	INDEL	0.941865	0.937756	0.939806	29329	32591
			SNP	**0.996948**	0.999033	**0.997989**	**10155**	3214
	3	pbccs	INDEL	0.986755	**0.988421**	0.987587	6682	**6069**
			SNP	0.998211	0.999261	0.998736	5954	2457
		DeepConsensus + re-trained DV	INDEL	**0.988248**	0.988042	**0.988145**	**5929**	6278
			SNP	0.998487	**0.999382**	0.998934	5035	**2056**
		DeepConsensus + DeepVariant v1.2	INDEL	0.956454	0.954828	0.95564	21969	23606
			SNP	**0.99856**	0.999318	**0.998939**	**4791**	2268
HG004 15kb	2	pbccs	INDEL	0.966677	0.976135	0.971383	17012	12540
			SNP	0.992949	0.99839	0.995662	23596	5363
		DeepConsensus + re-trained DV	INDEL	**0.9746**	**0.977708**	**0.976152**	**12967**	**11797**
			SNP	**0.995903**	**0.998765**	**0.997332**	**13711**	**4123**
	3	pbccs	INDEL	0.98602	0.988074	0.987046	7137	6327
			SNP	0.99782	0.999242	0.99853	7297	2535
		DeepConsensus + re-trained DV	INDEL	**0.988143**	**0.988352**	**0.988248**	**6053**	**6190**
			SNP	**0.998234**	**0.999352**	**0.998793**	**5910**	**2168**
HG006 15kb	2	pbccs	INDEL	0.989999	0.991616	0.990807	4283	3705
			SNP	0.998028	0.999078	0.998553	6448	3014
		DeepConsensus + re-trained DV	INDEL	**0.992824**	**0.993181**	**0.993002**	**3073**	**3017**
			SNP	**0.998547**	**0.999332**	**0.99894**	**4750**	**2183**
	3	pbccs	INDEL	0.993529	0.993174	0.993351	2771	3024
			SNP	0.998865	0.999451	0.999158	3712	1795
		DeepConsensus + re-trained DV	INDEL	**0.996324**	**0.996303**	**0.996314**	**1574**	**1637**
			SNP	**0.998977**	**0.999538**	**0.999258**	**3346**	**1509**
HG007 15kb	2	pbccs	INDEL	0.929728	0.938265	0.933977	30240	27097
			SNP	0.987392	0.997745	0.992541	41412	7333
		DeepConsensus + re-trained DV	INDEL	**0.960992**	**0.96166**	**0.961326**	**16786**	**17006**
			SNP	**0.995266**	**0.998799**	**0.997029**	**15549**	**3933**

Supplementary Table 21: Error reduction in HG006 dipcall variant calling between PBCCS and DeepConsensus stratified by Genome in a Bottle region.

SMRT Cells	Stratification	Method	Indel FN	Indel FP	SNP FN	SNP FP	Total Errors	DC errors relative to PBCCS
2	<30% GC	pbccs	3848	5393	29676	5053	43970	54.50%
		DeepConsensus	**2498**	**3176**	**14054**	**4151**	**23879**	
	30-55 % GC	pbccs	12625	20182	99606	19053	151466	55.60%
		DeepConsensus	**8307**	**10612**	**51065**	**14272**	**84256**	
	>55 % GC	pbccs	3573	7171	33486	6415	50645	56.30%
		DeepConsensus	**2252**	**3330**	**18295**	**4629**	**28506**	
	All Homopolymers	pbccs	903	1225	4871	960	7959	56.70%
		DeepConsensus	**586**	**642**	**2548**	**737**	**4513**	
	AllTandemRepeatsand HomoPolymers	pbccs	1887	2744	13603	2352	20586	60.90%
		DeepConsensus	**1250**	**1589**	**7755**	**1946**	**12540**	
	Dinucledotide repeats >10bp	pbccs	199	211	837	182	1429	61.50%
		DeepConsensus	**144**	**152**	**441**	**142**	**879**	
	All	pbccs	20078	32966	164546	30928	248518	55.70%
		DeepConsensus	**13101**	**17149**	**84972**	**23261**	**138483**	
3	<30% GC	pbccs	2089	2856	7270	**2062**	14277	75.10%
		DeepConsensus	**1701**	**1717**	**5223**	2081	**10722**	
	30-55 % GC	pbccs	7483	11163	32805	7993	59444	74.40%
		DeepConsensus	**6000**	**5868**	**24503**	**7832**	**44203**	
	>55 % GC	pbccs	2448	4082	18467	3326	28323	69.10%
		DeepConsensus	**1886**	**1773**	**13254**	**2671**	**19584**	
	All Homopolymers	pbccs	605	835	1751	**419**	3610	70.00%
		DeepConsensus	**434**	**429**	**1237**	427	**2527**	
	AllTandemRepeatsand HomoPolymers	pbccs	1250	1781	6154	**1016**	10201	75.40%
		DeepConsensus	**934**	**964**	**4706**	1094	**7698**	
	Dinucledotide repeats >10bp	pbccs	129	148	297	**61**	635	79.40%
		DeepConsensus	**123**	**105**	**201**	75	**504**	
	All	pbccs	12116	18201	60996	13520	104833	73.10%
		DeepConsensus	**9645**	**9288**	**44791**	**12877**	**76601**	

Supplementary Table 22. Yield for 15kb and 24kb reads at Q20 used for assembly and variant calling experiments.

Sample	Read Length	SMRT Cells	Method	Yield
HG002	15 kb	2	DeepConsensus	88.98 Gb
	24 kb	2	DeepConsensus	**103.61 Gb**

Supplementary Table 23: QUAST reported assembly statistics of hifiasm assemblies with 15kb and 24kb reads.

Sample	Read length	SMRT Cells	Method	N50 (Mb)	NG50 (Mb)	Assembly size (Gb)	Assembly completeness against GRCh38
HG002	15kb	2	DeepConsensus	24.81	24.81	3.09	97.654
	24kb	2	DeepConsensus	**33.14**	**34.05**	3.13	**97.68**

Supplementary Table 24: YAK estimated phred-scale Q value of hifiasm assemblies with 15kb and 24kb reads.

Sample	Read length	SMRT Cells	Method	HAP1 Q	HAP2 Q
HG002	15kb	2	DeepConsensus	**51.879**	**51.534**
	24kb	2	DeepConsensus	50.901	50.614

Supplementary Table 25: Assembly-based small variant calling SNP evaluation of hifiasm assemblies with 15kb and 24kb reads.

Sample	Read length	SMRT Cells	Method	SNPs					
				True positives	False negatives	False positives	Precision	Recall	F1-score
HG002	15kb	2	DeepConsensus	**3318960**	**46167**	**7073**	**0.9979**	**0.9863**	**0.992**
	24kb	2	DeepConsensus	3312496	52631	7404	0.9978	0.9844	0.991

Supplementary Table 26: Assembly-based small variant calling INDEL evaluation of hifiasm assemblies with 15kb and 24kb reads.

Sample	Read length	SMRT Cells	Method	INDELs					
				True positives	False negatives	False positives	Precision	Recall	F1-score
HG002	15kb	2	DeepConsensus	507578	17891	19463	0.9643	0.966	0.9651
	24kb	2	DeepConsensus	**511634**	**13835**	**19131**	**0.9652**	**0.9737**	**0.9694**

Supplementary Table 27: Asmgene single-copy gene completeness analysis of hifiasm assemblies with 15kb and 24kb reads.

Sample	Read length	SMRT Cells	Method	# Single copy ref	Single copy		False duplications		Complete (%)	Duplicated (%)
					Hap1	Hap2	Hap1	Hap2		
HG002	15kb	2	DeepConsensus	35374	**34553**	**34353**	**153**	**114**	**97.396**	**0.377**
	24kb	2	DeepConsensus	35374	34517	34269	191	189	97.227	0.537

Supplementary Table 28: Asmgene multi-copy gene completeness analysis of hifiasm assemblies with 15kb and 24kb reads.

Sample	Read length	SMRT Cells	Method	# Multi copy ref	Multi copy		Complete multi copy (%)	Missing multi copy (%)
					Hap1	Hap2		
HG002	15kb	2	DeepConsensus	1253	1024	904	76.935	23.065
	24kb	2	DeepConsensus	1253	**1043**	**985**	**80.926**	**19.074**

Supplementary Table 29: Chr20 variant calling results for DeepConsensus reads.

Sample	SMRT Cells	Method	Type	Recall	Precision	F1 Score	False Negatives	False Positives
HG002 15kb	2	DeepConsensus	INDEL	0.989606	0.987434	0.988518	117	147
			SNP	0.998822	**0.999748**	0.999285	84	**18**
HG002 24kb	2	DeepConsensus	INDEL	**0.989961**	**0.989044**	**0.989502**	**113**	**128**
			SNP	**0.999257**	0.999734	**0.999495**	**53**	19