UC Santa Cruz

UC Santa Cruz Previously Published Works

Title

Highly accurate long-read HiFi sequencing data for five complex genomes.

Permalink

https://escholarship.org/uc/item/26s4r638

Journal

Scientific data, 7(1)

ISSN

2052-4463

Authors

Hon, Ting
Mars, Kristin
Young, Greg
et al.

Publication Date

2020-11-01

DOI

10.1038/s41597-020-00743-4

Peer reviewed

SCIENTIFIC DATA

OPEN

DATA DESCRIPTOR

Highly accurate long-read HiFi sequencing data for five complex genomes

Ting Hon[1], Kristin Mars[1], Greg Young[1], Yu-Chih Tsai[1], Joseph W. Karalius[1], Jane M. Landolin[2], Nicholas Maurer[3], David Kudrna[4], Michael A. Hardigan[5], Cynthia C. Steiner[6], Steven J. Knapp[5], Doreen Ware[7,8], Beth Shapiro[3,9], Paul Peluso[1] & David R. Rank[1] ✉

The PacBio® HiFi sequencing method yields highly accurate long-read sequencing datasets with read lengths averaging 10–25 kb and accuracies greater than 99.5%. These accurate long reads can be used to improve results for complex applications such as single nucleotide and structural variant detection, genome assembly, assembly of difficult polyploid or highly repetitive genomes, and assembly of metagenomes. Currently, there is a need for sample data sets to both evaluate the benefits of these long accurate reads as well as for development of bioinformatic tools including genome assemblers, variant callers, and haplotyping algorithms. We present deep coverage HiFi datasets for five complex samples including the two inbred model genomes *Mus musculus* and *Zea mays*, as well as two complex genomes, octoploid *Fragaria × ananassa* and the diploid anuran *Rana muscosa*. Additionally, we release sequence data from a mock metagenome community. The datasets reported here can be used without restriction to develop new algorithms and explore complex genome structure and evolution. Data were generated on the PacBio Sequel II System.

Background & Summary

Until recently, DNA sequencing technologies produced either short highly accurate reads (up to 300 bases at 99% accuracy)[1,2] or less-accurate long reads (10–100 s of kb at 75–90% accuracy)[3,4]. Highly accurate short reads are appropriate for germline[5] and somatic[6] variant detection, exome sequencing[7], liquid biopsy[8], non-invasive prenatal testing[9], and counting applications such as transcript profiling[10] or single-cell analysis[11]. In contrast, error-prone long reads are more appropriate for *de novo* genome assembly[12–14], haplotype phasing[15], structural variant identification[16–18], full-length mRNA sequencing and mRNA isoform discovery[19].

To increase the utility of noisy long-read sequencing, several error correction methods have been devised to improve the accuracy of long reads by combining the data from either multiple independent long-read molecules or combining data from long- and short-read technologies[12,14]. These error-corrected reads can then be used for assembly or other downstream applications. In general, these error correction methods suffer from mis-mapping induced errors inherent to the multi-molecule approach[20] that hinder downstream applications.

A third sequencing data type leveraging multiple pass circular consensus sequencing of long (up to ~25 kb) individual molecules produces highly accurate long sequencing reads (HiFi reads)[21]. The HiFi sequencing protocol, data generation, and applications are described in Fig. 1. In the initial publication[21], 28-fold coverage of a human genome was sequenced with average read length of 13.5 kb and an average accuracy of 99.8%. The data

[1]Pacific Biosciences of California Inc., 1305 O'Brien Dr., Menlo Park, CA, 94025, USA. [2]Ravel Biotechnology Inc., 953 Indiana St., San Francisco, CA, 94107, USA. [3]Department of Ecology and Evolutionary Biology, University of California Santa Cruz, Santa Cruz, CA, 95064, USA. [4]Arizona Genomics Institute and School of Plant Sciences, University of Arizona, Tucson, AZ, 85721, USA. [5]Department of Plant Sciences, University of California, Davis, One Shields Ave, Davis, CA, 95616-8571, USA. [6]Conservation Genetics, Beckman Center for Conservation Research, San Diego Zoo Global, 15600 San Pasqual Valley Road, Escondido, CA, 92027, USA. [7]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 11724, USA. [8]USDA-ARS, Plant, Soil, and Nutrition Research Unit, Ithaca, NY, 14853, USA. [9]Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, CA, 95064, USA. ✉e-mail: drank@pacb.com
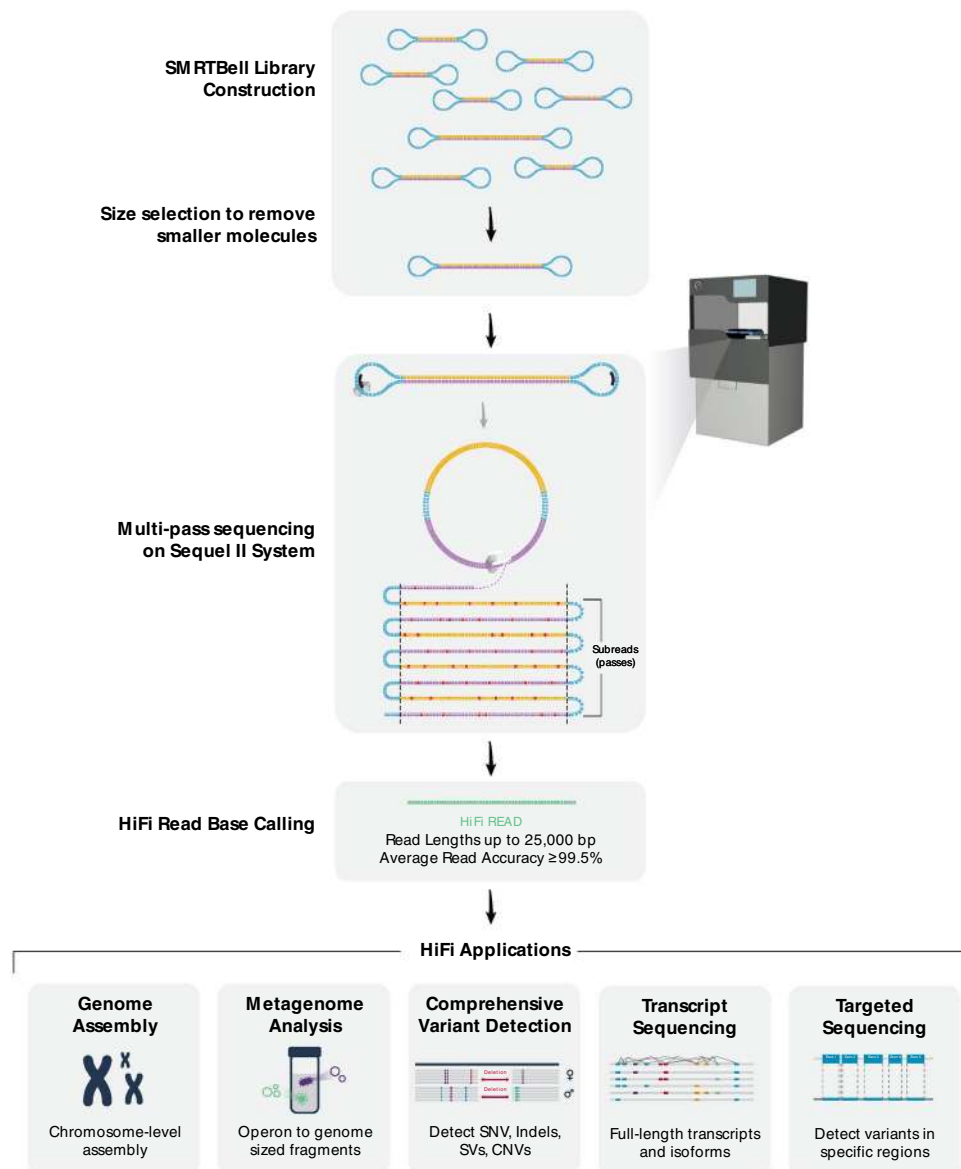
Fig. 1 Flowchart of HiFi sequence read generation and downstream applications.

has demonstrated superior assembly and haplotyping results for the human genome as measured by contiguity and accuracy when compared to traditional noisy long- or short-read methods. Additionally, single nucleotide variants were called at comparable precision and recall to Illumina® NovaSeq™ data. Since the initial publication, greatly improved assembly results have been observed in other human sequencing projects[22–25] using HiFi reads.

To encourage further application, software development, and interest in the HiFi sequencing data type, we report here the release of five deep coverage data sets spanning a set of complex genomes including *Mus musculus*, *Zea mays*, *Fragaria × ananassa* (Weston) Duchesne ex Rozier, *Rana muscosa*, and a standard metagenomic collection of 20 microbes formulated at staggered concentrations (ATCC® MSA-1003™). The data released in this study covers a wide breadth of highly complex plant, animal, and microbial organisms and will provide a useful sequence resource, driving the sequencing standards toward higher quality in the future[25].

Methods

Sample selection. Organisms sequenced in this study include *M. musculus, Z. mays, F. × ananassa*, and *R. muscosa*. The strain of each organism, source of the material, ploidy level, inbreeding status, reference genome sequence, and genome sizes are described in Tables 1 and 2. Additionally, we are releasing sequencing reads from a mock metagenomic sample (ATCC MSA-1003) consisting of 20 bacterial DNA samples at staggered concentrations ranging from 0.02% to 18% composition of the sample. The composition of the mock metagenomic sample as well as genome sizes of the individual bacterial species and their reference sequence accessions are listed in Supplementary Table 1.

Sample	Strain (Cultivar/ Cell line)	Sample Origin	Sequence Reference	SRA BioSample ID
M. musculus	C57BL/6 J	Jackson Labs	GRCm38.p6[60]	SAMN14691541[61]
Z. mays	B73	M. Hufford	Zm-B73-REFERENCE-NAM-5.0[62]	ERS3371164[63] SAMN14691542[64]
F. × ananassa	Royal Royce	S. Knapp	N/A	SAMN14691544[65]
R. muscosa	KB 21384; ISIS # 916035	San Diego Zoo Global	N/A	SAMN14691543[66]
Metagenome Std	MSA-1003	ATCC	See Supplementary Table 1	SAMN14691545[67]

Table 1. Sample description: strain names, origins, available reference sequences, and SRA BioSample IDs are detailed for each HiFi dataset.

....

Organism	Strain	Ploidy	Inbred	Haploid Genome Size (Mb)
M. musculus	C57BL/6 J	2n	Yes	2,700
Z. mays	B73	2n	Yes	2,200
F. × ananassa	Royal Royce	8n	No	800*
R. muscosa	KB 21384; ISIS # 916035	2n	No	9,000‡
Metagenome Std	MSA-1003	N/A	N/A	67

Table 2. Background genomic information for each sample: strain or sample ID, expected ploidy level, inbred status, and haploid genome size for each HiFi read dataset. *The estimate haploid genome size of *F. × ananassa* 'Royal Royce' is based on the size of the sequenced *F. × ananassa* 'Camarosa'[26]. ‡The haploid genome size of *R. muscosa* is estimated at 9 Gb based on the estimated genome sizes of 8,600 to 9,100 Mb for two closely related species (*R. aurora* and *R. cascadae*)[27] as well as the size estimate provided by our k-mer analysis.

....

Organism	HiFi library size (kb)	Sequel II Runs (number)	Bases > RQ20 (Gb)	Average RL (kb)	Reads (Millions)	Quality Value* (avg)	Data Record
M. musculus	15.9	2	66.5	16.4	4.1	31	SRR11606870[37]
Z. mays	15.0	2	48.1	15.6	3.1	30	SRR11606869[38]
F. × ananassa	23.0	1	29.7	21.7	1.4	28	SRR11606867[39]
R. muscosa	15.8	8	189.1	15.7	12.1	31	SRR11606868[40]
ATCC MSA-1003	14.1	2	59.1	10.5	5.6	35	SRR11606871[41]

Table 3. Library molecule sizes, sequencing metrics, and SRA accession numbers for each HiFi read dataset. *Predicted RQ values from the PacBio software are in Phred quality scale $= -10 \log10 (P)$ where P is the probability of error.

....

Excluding the metagenomic sample, the expected assembly sizes for the genomes sequenced in this study ranged from the 1,600 Mb for the outbred and octoploid *F. × ananassa*[26] to approximately 18,000 Mb for the outbred and diploid *R. muscosa* (estimate based on genome sizes of two related species *Rana aurora* and *Rana cascadae*)[27]. The individual genome sizes of the metagenomic sample range from 1.67 to 6.34 Mb, totaling 67 Mb of bacterial sequence (Supplementary Table 1).

Sequencing library preparation. Genomic DNA extraction methods and details of individual library preparations are described in the sample specific sections below. In general, if the starting genomic DNA sample was larger than 25 kb, the DNA was sheared to between 15 kb and 23 kb using the Megaruptor® 3 (Diagenode). HiFi sequencing libraries were prepared[28] using SMRTbell™ Express Template Prep Kit 2.0 and followed by immediate treatment with the Enzyme Clean Up Kit (PN: 101-843-100). The libraries were further size selected electrophoretically using either the SageELF or BluePippin Systems from SAGE Science. The appropriate fractions for sequencing runs were identified on the Femto Pulse System (Agilent). After pooling the desired size fractions, the final libraries were further cleaned up and concentrated using AMPure PB beads (Pacific Biosciences PN:100-265-900). Finally, all libraries were checked for concentration using Qubit™ 1X dsDNA HS Assay Kit (Thermo Fisher PN: Q33231) and final size distribution was confirmed on the Femto Pulse. All library sizes are described in Table 3.

***M. musculus* 'C57BL/6 J' sample acquisition, DNA extraction, and modifications to sequencing library preparation.** C57BL/6 J genomic DNA was obtained from The Jackson Laboratory (PN: GTC4560). The DNA arrived at an appropriate size for HiFi library preparation (~20 kb) and no shearing was required. Library preparation method, kit, and conditions were as described above. In order to tighten the size distribution of the SMRTbell library, the DNA was size fractionated using the SageELF following library preparation. The SMRTbell library was prepared with loading solution/Marker75 then loaded onto a 0.75% agarose 1kb-18 kb gel cassette (PN: ELD7510). Size fractionation was performed electrophoretically with a target size of 3,500 bp set for

elution well 12, which allowed for the collection of the appropriately sized library fractions (15–23 kb) in other elution wells of the SageELF device.

Z. mays 'B73' sample acquisition, DNA extraction, and modifications to sequencing library preparation. Leaf tissue for the B73 maize inbred was frozen and provided by Matthew Hufford at Iowa State University, Department of Ecology, Evolution, and Organismal Biology. Genomic DNA was isolated from the frozen leaf tissue at the University of Arizona Genomics Institute using methods previously described[29]. The high molecular weight DNA was sheared using the Megaruptor 3 targeting a size distribution between 15 and 20 kb. Library preparation method, kit and conditions were as described above. Library size selection was performed on the Sage BluePippin using the 0.75% Agarose dye-free Gel Cassette (PN: BLF7510) and the S1 Marker. To ensure suitable yields, the 3–10 kb Improved Recovery cassette definition was run for the size selection and high pass elution mode was chosen to target recovery of molecules greater than 15 kb.

F. × ananassa 'Royal Royce' sample acquisition, DNA extraction, and modifications to sequencing library preparation. The plant material was obtained from foundation stock of the culti-var 'Royal Royce' maintained by the UC Davis Strawberry Breeding Program. DNA was isolated as previously described[30]. The genomic DNA was larger than required for HiFi library production and was sheared using the Megaruptor 3 targeting a size distribution centered around 22 kb. Library preparation method, kit, and conditions were as described above. The SageELF was used for size selection, with similar conditions as described for *M. musculus* above, in order to generate a library with an appropriately sized distribution.

R. muscosa sample acquisition, DNA extraction, and modifications to sequencing library preparation. *R. muscosa*, the Mountain Yellow-legged Frog, is an endangered species endemic to California. To prevent sacrificing an individual, DNA was prepared from a fibroblast cell line (KB 21384; ISIS # 916035) originally derived from a 25-day old tadpole of undetermined sex. The cells were grown at room temperature in low O_2 from explants in alpha MEM with 1% NEAA. Approximately two million cells were harvested at passage 7 and frozen in a 1X solution of PBS buffer with 10% DMSO and 10% glycerol. Genomic DNA was isolated from these cells using Qiagen's MagAttract HMW DNA Kit (PN: 67563) following the manufacture's protocol. The resulting HMW gDNA was sheared to a target size of 22 kb on the MegaRuptor 3 prior to library preparation. Library preparation, kit and conditions were as described above. In order to tighten the size distribution, the SMRTbell library was size fractionated using SageELF System from Sage Science. The DNA was premixed with loading solution/Marker40 and loaded onto a 0.75% Agarose 10–40 kb Cassette (PN: ELD4010). Size fractiona-tion was performed electrophoretically with a target size of 7,000 bp set for elution well 12 in order to achieve the appropriate resolution in size separation. Fractions having the desired size distribution ranges were identified on the Femto Pulse to generate a final size selected library used in the Sequel II sequencing runs. An additional DNA damage repair step was performed using the SMRTbell Damage Repair Kit (PN:100-992-200) as this was found helpful to improve library performance in sequencing runs.

Mock metagenome sample acquisition, DNA extraction, and modifications to sequencing library preparation. ATCC offers a mock metagenomic community (MSA 1003) of 20 bacteria species rang-ing in composition from 0.02% to 18% of the sample. Isolated DNA from this sample arrived with genomic DNA having a broad distribution of sizes and was sheared using the MegaRuptor 3 to a uniform size of 13.7 kb. Library preparation method, kit and condition were described above. Rather than using electrophoretic size selection, the resulting library was size selected using AMPure PB beads (35% v/v) to remove all small fragments.

Sequencing and data processing. SMRTbell libraries were bound to the sequencing polymerase enzyme using the Sequel II Binding Kit 2.0 (PN:101-842-900) with the modification that the Sequencing Primer v2 (PN:101-847–900) was annealed to the template instead of the standard primer which comes with Sequel II Binding Kit 2.0. All incubations were performed per manufacturer's recommendations. Prior to sequenc-ing, unbound polymerase enzyme was removed using a modified AMPure PB bead method as previously described[21,31]. Shotgun genomic DNA sequence data was collected on the Pacific Biosciences Sequel II system using HiFi sequencing protocols[31] and Sequencing kit V2 (PN: 101-820-200). Sequence data collection was stand-ardized to 30 hours for this study to allow ample time for multiple pass sequencing around SMRTbell template molecules of 10–25 kb which yields high quality circular consensus sequencing (HiFi) results[21]. Raw base-called data was moved from the sequencing instrument and the imported into SMRTLink[32] to generate HiFi reads using the CCS algorithm (version 8.0.0.80529) which processed the raw data and generated the HiFi fastq files with the following settings: minimum pass 3, minimum predicted RQ 20.

K-mer analysis. Using Jellyfish[33] (v.2.2.10) a k-mer analysis was performed on each of the HiFi data sets individually using a k-mer size of 21. Counting was done using a two-pass method. First, a Bloom counter was created for each HiFi read dataset using the command described in Box 1.

Box 1 Running Jellyfish to create Bloom counter.

```
jellyfish bc -m 21 -s <Input Size> -t <nproc> -C -o
HiFiReadSetFilename.bc HiFiReadSet.fasta
```

where Input Size = 100G (*M. musculus, Z. mays, F. × ananassa* and *R. muscosa*) and 5G (ATCC MSA-1003).

After generating the Bloom counter, a frequency count of k-mers (size = 21) was run using the command shown in Box 2:

Box 2 Running Jellyfish to obtain a frequency count of k-mers.

```
jellyfish count -m 21 -s <Input Size> -t <nproc> -C --bc
HiFiReadSetFilename.bc HiFiReadSet.fasta
```

Where Input Size = 20G (*R. muscosa*), 3G (*M. musculus* and *Z. mays*), 2G (*F. × ananassa*) and 200M (ATCC MSA-1003).

Finally, a histogram of the k-mer frequency was generated for each dataset by using the command in Box 3.

Box 3 Generating k-mer histogram.

```
jellyfish histo HiFiReadSet_21mer counts.jf >
HiFiReadSet_21mer_Histogram.out
```

These outputs were then used to generate the additional summary analysis and determine genome sizes for each sample where applicable. Genome sizes were estimated from the ratio of total HiFi bases divided by the frequency mode from each k-mer distribution.

Mapping accuracies and read lengths. In the cases where references were available (*M. musculus*, *Z. mays*, and the concatenated genomes comprising the ATCC MSA-1003 sample), HiFi reads were mapped to the references using pbmm2 version 1.2.0 (https://github.com/PacificBiosciences/pbmm2) which is a customized wrapper for minimap2[34] using the command demonstrated in Box 4.

Box 4 Mapping HiFi reads to a reference with pbmm2.

```
pbmm2 align REF.fasta HiFiReadSet.fastq
HiFiReadSet.REF.sorted.bam --preset CCS --sort -j 48 -J 16
```

(where j + J = nproc=64)

To extract accuracy metrics from each bam file using Samtools[35] version 1.9, the command shown in Box 5 was used:

Box 5 Extracting accuracy metrics from bam file using Samtools.

```
samtools view HiFiReadSet.REF.sorted.bam | awk '{ mc="";
for(i=12;i<=NF;i++) { split($i,TAG,":"); if(TAG[1]=="mc") {
mc=TAG[3]; break; } } if(mc != "") { print $1 "\t" mc; } }' >
MappedConcordance.HiFiReadSet.Genome.out
```

Box 6 shows the command used to extract read length metrics from each bam file using Samtools,

---

Box 6 Extracting read length metrics from bam file using Samtools.

```
samtools view HiFiReadSet.REF.sorted.bam | head -n <input # of
HiFi Reads> | cut -f 10 | perl -ne 'chomp;print length($_).
"\n"' | sort | uniq -c > MappedRL.HiFiReadSet.Genome.out
```

Finally, coverage metrics were obtained from each bam files using the Samtools with the command listed in Box 7.

Box 7 Extracting coverage metrics using Samtools.

```
samtools depth -a HiFiReadSet.REF.sorted.bam >
HiFiReadSet.REF.sorted.Depth.out
```

Sample	K-mer based Genome Coverage (fold)	Reference Mapped Genome Coverage (fold)	Median Read Accuracy (percent)	Mean Read Accuracy (percent)
M. musculus	25	27	99.869	99.176
Z. mays	21	23	99.844	99.686
F. × ananassa	17/37/74/109	N/A#	N/A#	N/A#
R. muscosa	20	N/A#	N/A#	N/A#
ATCC MSA-1003	2–4000	1–8,000§	99.995	99.733

Table 4. Technical validation summary: k-mer based genome size estimates, average mapped HiFi read coverage for samples with references[59,61] genomes, and average mapped HiFi read accuracy for each dataset. #No published reference. §See Supplementary Table 1 for reference genome file names and locations.

Data Records

All sequencing data presented are available at the Sequencing Read Archive (SRA) under the SRA study accession SRP258341[36]. The HiFi sequencing data is stored as fastq files with one file for each Sequel II sequencing run. Information describing each data record is presented in Table 3 and described below.

SRR11606870[37] The *M. musculus* 'C57BL/6 J' data record is composed of two Sequel II runs (total of two SMRT Cell 8 M) containing 4.1 M sequencing reads and 66.5 Gb of sequence which corresponds to 25-fold coverage of the mouse genome. The average read length is 16.4 kb with an average PacBio predicted quality value (RQ) of 31.

SRR11606869[38] The *Z. mays* 'B73' data record is composed of two Sequel II runs (total of two SMRT Cell 8 M) containing 3.1 M sequencing reads and 48.1 Gb of sequence which corresponds to 22-fold coverage of the maize genome. The average read length is 15.6 kb with an average PacBio predicted quality value of 30.

SRR11606867[39] The *F. × ananassa* 'Royal Royce' data record is composed of one Sequel II run (total of one SMRT Cell 8 M) containing 1.4 M sequencing reads and 29.7 Gb of sequence of the octoploid Royal Royce genome. The average read length is 21.7 kb with an average RQ value of 28.

SRR11606868[40] The *R. muscosa* (cell line KB 21384; ISIS # 916035) data record is composed of 8 Sequel II runs (total of eight SMRT Cell 8 M) containing 12.1 M sequencing reads and 189.1 Gb bases of sequence which corresponds to approximately 20-fold coverage of the *R. mucosa* genome. The average read length is 15.7 kb with an average RQ of 31.

SRR11606871[41] is the data record for the ATTC MSA-1003 mock metagenome community which is composed of 20 bacterial organisms reported to be mixed at relative amounts differing by 900-fold from highest to lowest (Supplementary Table 1). The files in this sequence record span two Sequel II runs (total of two SMRT Cell 8 M) containing 5.6 M sequencing reads with 59.1 Gb of sequence which corresponds to between ~3 and ~5,000-fold coverage of the individual bacterial genomes. The average read length is 10.5 kb with an average RQ value of 35.

Additionally, the raw base-called subreads from which the HiFi consensus reads were derived have been made available as a resource for developers interested in improving circular consensus sequencing algorithms. These reads have been deposited to the SRA under the following data records **SRR12358174**[42] **and SRR12371718**[43] (*M. musculus*); **SRR12358173**[44] (*Z. mays*); **SRR12358171**[45] (*F. x ananassa*); **SRR12371721**[46], **SRR12371723**[47], **SRR12371724**[48], **SRR12371725**[49], **SRR12371726**[50], **SRR12371727**[51], **SRR12371722**[52], **SRR12358172**[53] (*R. muscosa*); and **SRR12371719**[54] and **SRR12358170**[55] (mock metagenome), and are further described in Supplementary Table 2.

Technical Validation

Two of the non-microbial organisms sequenced, *M. musculus*, and *Z. mays*, have high quality reference genomes available[56,57] allowing for detailed validation of the sequencing data. Additionally, reads from sequencing the mock metagenome sample were aligned to a concatenated file containing all microbial references listed in Supplementary Table 1 and used for validation. All reads were aligned to their corresponding references using pbmm2 resulting in over 98.9% of reads mapping to their respective genomes and 98.4% to 99.3% of the alignments being unique within the respective references (Supplementary Table 3). The mapped read lengths and read accuracies are reported in Tables 3 and 4 and distributions are presented in Fig. 2 with a breakdown of error types and their distribution described in Supplementary Figure 1. In agreement with previously published reports[21], the accuracy of the HiFi reads exceeds 99.5% with sequencing errors predominantly arising from indels (Supplementary Figure 1b–d). Median accuracies are 99.87%, 99.84%, and 99.99% for the mouse, maize and mock metagenome samples respectively with mean accuracies of 99.18% (mouse), 99.69% (maize) and 99.73% (mock metagenome). Sequencing read lengths (Table 3) ranged from 10.5 kb (mock metagenome) to 21.7 kb (*F. × ananassa*) and were dependent on the final size distributions of the sequencing libraries.

The data for all five organisms was used to generate k-mer plots using a k-mer size of 21 (Fig. 3a–e) to estimate the sequencing coverage and complexity of for each sample. K-mer based sequencing coverage was measured at 17 to 25-fold (Table 4) for each of the individual diploid genomes sequenced (*M. musculus, Z. mays, and R. muscosa*) and as expected produced a multimodal distribution for the octoploid *F. × ananassa* and a complex curve for the metagenome sample.

Additionally, the k-mer plots can be used to characterize genome complexity such as ploidy and/or genome duplications as evidenced by multimode distributions within the k-mer plots caused by inherent polymorphism within the respective genomes. As expected, the inbred mouse C57BL/6 J, shows a single k-mer distribution
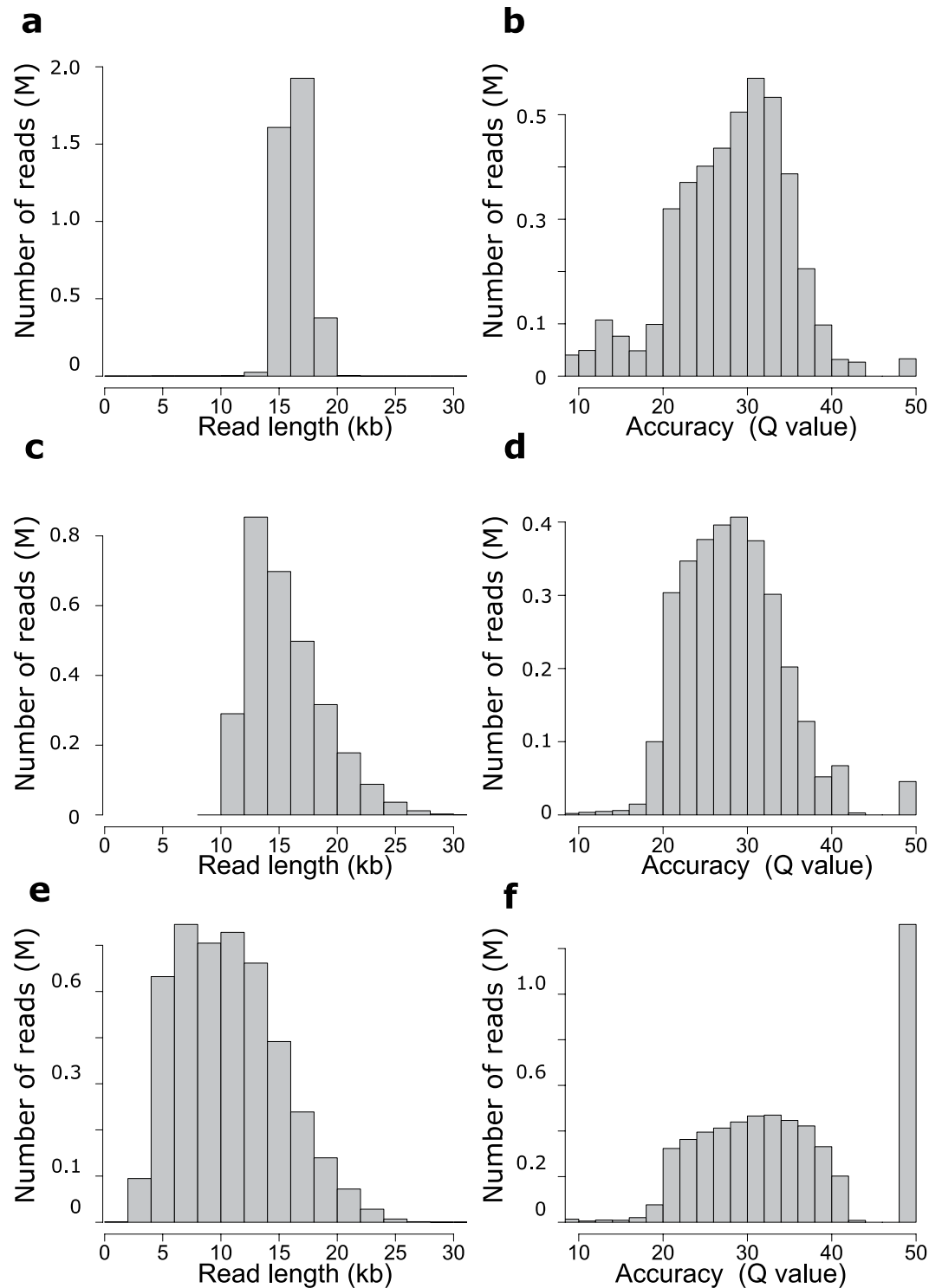
Fig. 2 Read length and quality distributions for the three sequenced samples with high quality finished sequence references. *M. musculus* read length (**a**) and accuracy (**b**), *Z. mays* read length (**c**) and accuracy (**d**), and Mock metagenome community ATTC MSA-1003 read length (**e**) and accuracy (**f**). All data is mapped to the genomic references (Table 1 and Supplementary Table 1) using minmap2. Accuracies are reported in Phred read quality space (Q value) $= -10 \times \log10(P)$ where P is the measured error rate.

consistent with the single haplotype present in the inbred animal. The inbred B73 maize shows a dominant k-mer coverage peak at 21-fold as one would expect, but also a minor peak at 42-fold which is consistent with an ancient duplication and polyploidization[58] of this inbred sample. The high heterozygocity and ploidy of *F. × ananassa* contribute to a complex and ill-defined k-mer spectrum which is consistent with previous observations[59]. Major
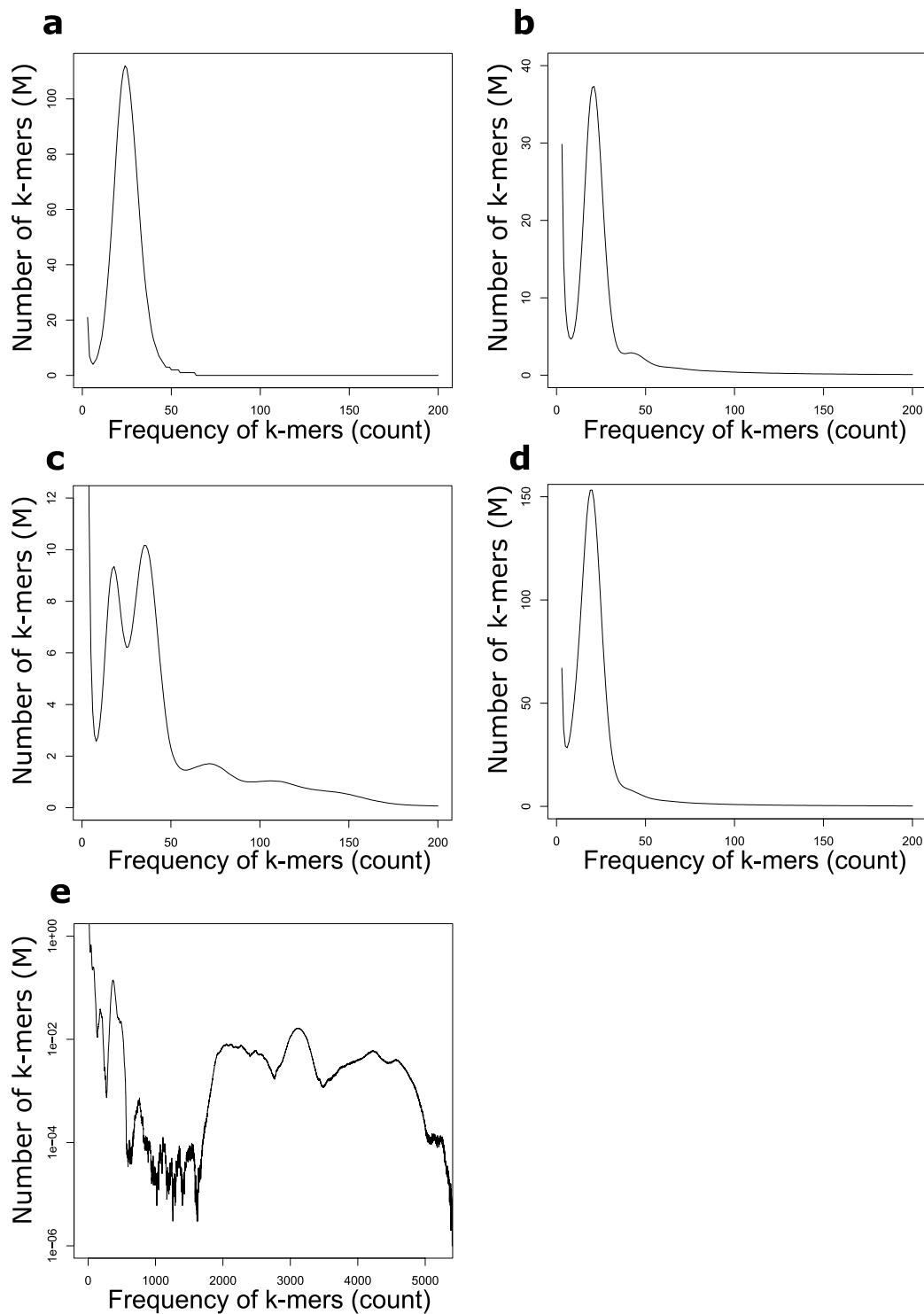
Fig. 3 K-mer (length 21) distribution for all HiFi reads for each sequencing dataset. (**a**) *M. musculus* (**b**) *Z. mays* (**c**) *F. × ananassa* (**d**) *R. muscosa* (**e**) Mock metagenome community ATTC MSA-1003.

k-mer frequency peaks at 17, 37, and minor peaks at 74 and 109-fold presumably represent the 8n, 4n, and less well defined 2n and 1n components of the genome, respectively. These k-mer peak identifications are consistent with an 8n genome size of 1.7 Gb (29.5 Gb of sequence/17-fold coverage) which agrees with previously published genome size[26] for octoploid strawberry.

The diploid *R. muscosa* sample demonstrates a more interesting case with respect to k-mer analysis as the frequency distribution shows one single haplotype at 20-fold coverage. The presence of a single k-mer peak in the genomic reads likely speaks to population bottlenecking which reduced the level of polymorphism in the genome resulting in collapse of the paternal and maternal haplotypes into one frequency peak for a k-mer size of

21. This is further supported by an apparent haploid genome size of 9,000 Mb (as calculated by the total number of sequenced bases / frequency mode of the k-mer histogram) which is equal to one half the size of the measured diploid genome sizes (~18 Gb) of two closely related *species* (*R. aurora* and *R. cascadae*)[27].

Alternatively, the genome coverage can be measured by mapping the HiFi reads to published references. The genome wide mapping-based coverages are reported in Table 4 and distributions are show in Supplementary Figures 2–4 and agree with the k-mer based estimates for those samples with known references. Minimal impact of GC composition is observed on HiFi sequencing coverage for the mouse and maize samples (Supplementary Figure 2 and 3). The read mapping method for genome coverage also produces coverage distributions and values for each member of the mock metagenome community sample (Supplementary Figure 4) and is consistent with the genomic complexity displayed in the k-mer plot (Fig. 3e), and agrees with the uneven representation of the abundance of each microbe in the mixture (Supplementary Figure 5).

Usage Notes

The data presented in this manuscript should provide ample DNA sequence for genome assembly, variant detection, evaluation of metagenome completeness and metagenome assembly for the samples covered. Additionally, the data should prove useful for bioinformaticians developing, improving, and validating assembly algorithms, developing haplotyping tools, and variant detection algorithms. High contiguity and high-quality genome assemblies should also be possible for the two unpublished genomes presented in this study (*F. × ananassa* 'Royal Royce', and the endangered anuran *R. muscosa)*. Recently, HiFi read based assemblies have reconstructed several centromeres of the human genome[25], and the HiFi data presented here will be useful for future updates of the reference genomes for both *Z. mays* 'B73' and *M. muscosa* 'C57BL/6 J' by adding previously unresolvable regions, possibly including some complete centromeres of these genomes. The data from the metagenome mock community should prove valuable for metagenome assembly algorithms, and other analytical tool development allowing for the assembly of complete bacterial genomes from metagenomic samples displaying high heterogeneity in individual bacterial species and relative abundance.

Code availability

Bioinformatic tools used for validation are all open source and feely available. We used jellyfish[33] version 2.2.10 to count k-mers, pbmm2 version 1.2.0 to map to a reference, and samtools[35] version 1.9 to summarize metrics. Sequencing accuracy breakdowns, error type determination and sequencing coverage measured across GC composition bins were determined as previously described[21].

References

1. Bentley, D. R. *et al*. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
2. Rothberg, J. M. *et al*. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**, 348–352 (2011).
3. Eid, J. *et al*. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science* **323**, 133–138 (2009).
4. Mikheyev, A. S. & Tin, M. M. Y. A first look at the Oxford Nanopore MinION sequencer. *Mol. Ecol. Resour.* **14**, 1097–1102 (2014).
5. Koboldt, D. C., Larson, D. E. & Wilson, R. K. Using VarScan 2 for Germline Variant Calling and Somatic Mutation Detection. *Curr. Protoc. Bioinforma.* **44**, 15.4.1–15.4.17 (2013).
6. Krøigård, A. B., Thomassen, M., Lænkholm, A.-V., Kruse, T. A. & Larsen, M. J. Evaluation of Nine Somatic Variant Callers for Detection of Somatic Mutations in Exome and Targeted Deep Sequencing Data. *PLoS ONE* **11**, (2016).
7. Ng, S. B. *et al*. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276 (2009).
8. Plagnol, V. *et al*. Analytical validation of a next generation sequencing liquid biopsy assay for high sensitivity broad molecular profiling. *PLoS ONE* **13**, (2018).
9. Chitty, L. S. *et al*. Non-invasive prenatal diagnosis of achondroplasia and thanatophoric dysplasia: next-generation sequencing allows for a safer, more accurate, and comprehensive approach. *Prenat. Diagn.* **35**, 656–662 (2015).
10. Cloonan, N. *et al*. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* **5**, 613–619 (2008).
11. Tang, F. *et al*. RNA-Seq analysis to capture the transcriptome landscape of a single cell. *Nat. Protoc.* **5**, 516–535 (2010).
12. Chin, C.-S. *et al*. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
13. Chin, C.-S. *et al*. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
14. Koren, S. *et al*. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
15. Patterson, M. *et al*. WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads. *J. Comput. Biol.* **22**, 498–509 (2015).
16. Sedlazeck, F. J. *et al*. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
17. Rausch, T. *et al*. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
18. Chen, X. *et al*. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).
19. Cartolano, M., Huettel, B., Hartwig, B., Reinhardt, R. & Schneeberger, K. cDNA Library Enrichment of Full Length Transcripts for SMRT Long Read Sequencing. *PLoS ONE* **11**, (2016).
20. Heydari, M., Miclotte, G., Demeester, P., Van de Peer, Y. & Fostier, J. Evaluation of the impact of Illumina error correction tools on de novo genome assembly. *BMC Bioinformatics* **18**, 374 (2017).
21. Wenger, A. M. *et al*. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
22. Porubsky, D. *et al*. A fully phased accurate assembly of an individual human genome. Preprint at https://doi.org/10.1101/855049 (2019).
23. Garg, S. *et al*. Efficient chromosome-scale haplotype-resolved assembly of human genomes. Preprint at https://doi.org/10.1101/810341 (2019).
24. Shumate, A. *et al*. Assembly and annotation of an Ashkenazi human reference genome. *Genome Biol.* **21**, 129 (2020).

25. Nurk, S. *et al*. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res*. gr.263566.120, https://doi.org/10.1101/gr.263566.120 (2020).
26. Edger, P. P. *et al*. Origin and evolution of the octoploid strawberry genome. *Nat. Genet*. **51**, 541–547 (2019).
27. Vinogradov, A. E. Genome size and GC-percent in vertebrates as determined by flow cytometry: The triangular relationship. *Cytometry* **31**, 100–109 (1998).
28. Travers, K. J., Chin, C.-S., Rank, D. R., Eid, J. S. & Turner, S. W. A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res*. **38**, e159–e159 (2010).
29. Doyle, J. J. & Doyle, J. L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* **19**, 11–15 (1987).
30. Li, Z., Parris, S. & Saski, C. A. A simple plant high-molecular-weight DNA extraction method suitable for single-molecule technologies. *Plant Methods* **16**, 38 (2020).
31. Procedure & Checklist - Preparing HiFi SMRTbell Libraries using SMRTbell Template Prep Kit 2.0, https://www.pacb.com/wp-content/uploads/Procedure-Checklist-Preparing-HiFi-SMRTbell-Libraries-using-SMRTbell-Express-Template-Prep-Kit-2.0.pdf (2020).
32. PacBio SMRT Link, https://www.pacb.com/support/software-downloads (2020).
33. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
34. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
35. Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
36. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP258341 (2020).
37. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR11606870 (2020).
38. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR11606869 (2020).
39. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR11606867 (2020).
40. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR11606868 (2020).
41. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR11606871 (2020).
42. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR12358174 (2020).
43. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR12371718 (2020).
44. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR12358173 (2020).
45. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR12358171 (2020).
46. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR12371721 (2020).
47. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR12371723 (2020).
48. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR12371724 (2020).
49. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR12371725 (2020).
50. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR12371726 (2020).
51. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR12371727 (2020).
52. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR12371722 (2020).
53. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR12358172 (2020).
54. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR12371719 (2020).
55. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR12358170 (2020).
56. Sarsani, V. K. *et al*. The Genome of C57BL/6J "Eve", the Mother of the Laboratory Mouse Genome Reference Strain. *G3 Genes Genomes Genet*. **9**, 1795–1805 (2019).
57. Jiao, Y. *et al*. Improved maize reference genome with single-molecule technologies. *Nature* **546**, 524–527 (2017).
58. Wei, F. *et al*. Physical and Genetic Structure of the Maize Genome Reflects Its Complex Evolutionary History. *PLoS Genet*. **3**, (2007).
59. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun*. **11**, 1432 (2020).
60. *NCBI Assembly* https://identifiers.org/ncbi/insdc.gca:GCF_000001635.26 (2017).
61. *NCBI BioSample* https://identifiers.org/ncbi/BioSample:SAMN14691541 (2020).
62. *NCBI Assembly* https://identifiers.org/ncbi/insdc.gca:GCA_902167145.1 (2020).
63. *NCBI BioSample* https://identifiers.org/ncbi/BioSample:SAMEA5569141 (2020).
64. *NCBI BioSample* https://identifiers.org/ncbi/BioSample:SAMN14691542 (2020).
65. *NCBI BioSample* https://identifiers.org/ncbi/BioSample:SAMN14691544 (2020).
66. *NCBI BioSample* https://identifiers.org/ncbi/BioSample:SAMN14691543 (2020).
67. *NCBI BioSample* https://identifiers.org/ncbi/BioSample:SAMN14691545 (2020).

Acknowledgements

Author contributions

T.H., K.M., G.Y. and Y.-C.T. library preparation, DNA sequencing and data quality control, and manuscript preparation. J.M.L. and J.W.K. collating sequencing data, posting to data repositories, and manuscript preparation. N.M., D.K., M.A.H. sample selection, sample preparation, and DNA isolation. C.C.S., S.J.K., D.W. and B.S. sample selection and manuscript preparation. P.S.P. experimental design, sequencing coordination, data submission, technical validation, bioinformatic analysis, and manuscript preparation. D.R.R. experimental design, technical evaluation, and manuscript preparation.

Competing interests

T.H., K.M., G.Y., Y-C. T., J.W.K., P.S.P. and D.R.R. are employees of Pacific Biosciences of California Inc. a company commercializing DNA sequencing technology. J.M.L. is an employee of Ravel Biotechnology Inc. a company commercializing disease detection from cell-free DNA. All other authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41597-020-00743-4.

Correspondence and requests for materials should be addressed to D.R.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.