bioRxiv preprint doi: https://doi.org/10.1101/008037; this version posted August 15, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Long-read, whole-genome shotgun sequence data for five model organisms

Kristi E. Kim[1], Paul Peluso[1], Primo Babayan[1], P. Jane Yeadon[2], Charles Yu[3], William W. Fisher[3], Chen-Shan Chin[1], Nicole Rapicavoli[1], David R. Rank[1], Joachim Li[4], David E. A.Catcheside[2], Susan E. Celniker [3], Adam M. Phillippy[5], Casey M. Bergman[6], Jane M. Landolin[1]

[1] Pacific Biosciences of California Inc., 1380 Willow Road, Menlo Park, CA

[2] Flinders University, School of Biological Sciences, PO Box 2100, Adelaide, SA 5001, Australia

[3] Berkeley Drosophila Genome Center, Lawrence Berkeley National Laboratory, Berkeley, CA

[4] Department of Microbiology and Immunology, University of California, San Francisco, San Francisco, CA

[5] National Biodefense Analysis and Countermeasures Center, 110 Thomas Johnson Drive, Frederick, MD 21702, USA

[6] Faculty of Life Sciences, University of Manchester, Oxford Road, Manchester, UK M13 9PT

Correspondence to:

Jane M. Landolin, Ph.D.
Pacific Biosciences
1380 Willow Road
Menlo Park, CA 94025
E-mail: jlandolin@pacificbiosciences.com

Keywords: genome sequence, open data, model organism, PacBio, single-molecule real-time sequencing, E. coli, S. cerevisiae, Neurospora, Arabidopsis, Drosophila

bioRxiv preprint doi: https://doi.org/10.1101/008037; this version posted August 15, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Abstract

Single molecule, real-time (SMRT) sequencing from Pacific Biosciences is increasingly used in many areas of biological research including *de novo* genome assembly, structural-variant identification, haplotype phasing, mRNA isoform discovery, and base-modification analyses. High-quality, public datasets of SMRT sequences can spur development of analytic tools that can accommodate unique characterisitcs of SMRT data (long read lengths, lack of GC or amplification bias, and a random error profile leading to high consensus accuracy). In this paper, we describe eight high-coverage SMRT sequence datasets from five organisms (*Escherichia coli*, *Saccharomyces cerevisiae*, *Neurospora crassa*, *Arabidopsis thaliana*, and *Drosophila melanogaster)* that have been publicly released to the general scientific community (NCBI Sequence Read Archive ID SRP040522). Data were generated using two sequencing chemistries (P4-C2 and P5-C3) on the PacBio RS II instrument. The datasets reported here can be used without restriction by the research community to generate whole-genome assemblies, test new algorithms, investigate genome structure and evolution, and identify base modifications in some of the most widely-studied model systems in biological research.

bioRxiv preprint doi: https://doi.org/10.1101/008037; this version posted August 15, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Background and Summary

Single-molecule, real-time (SMRT®) DNA sequencing occurs by optically detecting a fluorescent signal when a nucleotide is being incorporated by a DNA polymerase [1-4]. This relatively new technology enables detection of DNA sequences that have unique characteristics, such as long read lengths, lack of CG bias, and random error profiles, and can yield highly accurate consensus sequences [5]. Kinetic information such as pulse width and interpulse duration are also recorded and can be used to detect base modifications [6-8].

Since its introduction, investigators have published on a range of applications using SMRT sequencing. For example, the developers of GATK (Genome Analysis Toolkit) demonstrated that single nucleotide polymorphisms (SNPs) could be detected using SMRT sequences [9, 10] due to their lack of context-specific bias and systematic error [5, 10]. Likewise, the developers of PBcR (PacBio error correction) [11, 12] showed that complete bacterial genome assemblies using SMRT sequence data had greater than Q60 base quality [12]. PBcR was later incorporated as the "pre-assembly" step in the HGAP (hierarchical genome assembly process) system [13], followed by consensus polishing using the Quiver algorithm [13] to produce a complete assembly pipeline for SMRT sequence data. In addition, other third-party tools now support long reads for various applications such as mapping [14, 15], scaffolding [16], structural-variation discovery [17], and genome assembly [11, 18]. Other applications such as 16S rRNA sequencing [19], characterization of entire transcriptomes [20, 21], genome-editing studies [22], base-modification studies [7, 8, 23-25], and validation of CRISPR targets [26] have also been published.

To encourage interest in further applications and tool development for SMRT sequence data, we report here the release of whole-genome shotgun-sequence datasets from five model organisms (*E. coli*, *S. cerevisiae*, *N. crassa*, *A. thaliana*, and *D. melanogaster*). These organisms have among the most complete and well-annotated reference genome sequences, due to continual refinement by dedicated teams of scientists. Despite continued improvement of these genome sequences with new technologies, few are completely finished with fully contiguous assemblies of all chromosomes. The gaps remaining arise from complex structures such as transposable elements, repeats, segmental duplications, or other dynamic regions of the genome that cannot be easily assembled. Structural differences in these regions can account for variability in millions of nucleotides within every genome, and mounting evidence suggest that such mutations are important for human diversity and disease susceptibility in many complex traits including autism and schizophrenia [27-29]. SMRT sequencing data can therefore play an important role in the completion of these and other reference genomes, providing a platform for new insights into genome biology.

bioRxiv preprint doi: https://doi.org/10.1101/008037; this version posted August 15, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Methods

We generated eight whole-genome shotgun-sequence datasets from five model organisms using the P4C2 or P5C3 polymerase and chemistry combinations, totaling nearly 1000 gigabytes (GB) of raw data (See Data Records section). Genomic DNA was either purchased from commercial sources or generously provided by collaborators.

DNA from the reference K12 strain of *E. coli* was purchased from Lofstrand Labs Limited (K12 MG1655 *E. coli*, cat# L3-4001SP2). DNA from the reference OR74A strain of *N. crassa* was purchased from the Fungal Genetics Stock Center (FGSC). A standard Ler-0 strain of *A. thaliana* plants was grown from seeds purchased from Lehle seeds (WT-04-19-02) and DNA was extracted at Pacific Biosciences. The protocol is available on Sample Net [30] and summarized in the organism-specific methods section of this paper. DNA from the 9464 strain of *S. cerevisiae* was provided by J. Li at University of California San Francisco. The 9464 strain is a daughter of the reference WG303 strain. DNA from the T1 strain of *N. crassa* was obtained from D. Catcheside at Flinders University who has an interest in polymorphic genes regulating recombination. The T1 strain is an A mating type strain which, like OR74A, was derived from a cross between the Em a 5297 and Em A 5256 strains . DNA from the ISO1 strain [31] of *D. melanogaster* was obtained from S. Celniker at Lawrence Berkeley National Laboratory. This is the reference strain of *D. melanogaster* that was originally chosen to be the first large genome to be sequenced and assembled using a whole-genome shotgun approach [32]. It continues to serve as the reference strain in subsequent releases and numerous annotations of the *D. melanogaster* genome.

DNA extraction methods were species-specific and optimized for each organism (See organism-specific methods below). In general, the steps are: (1) remove debris and particulate material, (2) lyse cells, (3) remove membrane lipids, proteins and RNA, (4) DNA purification.

SMRTbell™ libraries for sequencing [9] were prepared using either 10 kb [33, 34] or 20 kb [35] preparation protocols to optimize for the most high-quality and longest reads. The main steps for library preparation are: (1) Shearing (2) DNA damage repair, (3) blunt end-ligation with hairpin adapters supplied in the DNA Template Prep Kit 2.0 (Pacific Biosciences), (4) size selection, and (5) binding to polymerase using the DNA Sequencing Kit 3.0 (Pacific Biosciences).

bioRxiv preprint doi: https://doi.org/10.1101/008037; this version posted August 15, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Table 1: Summary of DNA Samples. The NCBI sample ID associated with each dataset is provided. DNA was extracted in a species-specific manner, yielding genomic DNA of various sizes. All DNA was size selected using the Blue Pippin system (Sage Sciences), and select samples were sheared with g-TUBEs (Covaris).

Dataset Name	Sample ID	DNA extraction	gDNA size (kb)	Shearing	Size selection
E. coli MG1655 P4C2	SAMN02951645	ammonium acetate or SDS, proteinase K, phenol-chloroform	10	none	Blue Pippin (7kb)
E. coli MG1655 P5C3	SAMN02743420	ammonium acetate or SDS, proteinase K, phenol-chloroform	10	none	Blue Pippin (7kb)
S. cerevisiae 9464 P4C2	SAMN02731377	contact J. Li at UCSF	>40	g-TUBE	Blue Pippin (17kb)
N. crassa OR74A P4C2	SAMN02724975	BashingBeads, Zymo Research kit	6	none	Blue Pippin (4kb)
N. crassa T1 P4C2	SAMN02724976	SDS, proteinase K, phenol-chloroform, RNAase, isopropanol	15	none	Blue Pippin (7kb)
A. thaliana Ler-0 P5C3	SAMN02724977	CTAB, chloroform:isoamyl, isopropanol precip.	>40	g-TUBE	Blue Pippin (15kb)
A. thaliana Ler-0 P4C2	SAMN02731378	CTAB, chloroform:isoamyl, isopropanol precip.	>40	g-TUBE	Blue Pippin (7kb)
D. melanogaster ISO1 P5C3	SAMN02614627	SDS, phenol-chloroform, CsCl banding, ethanol precip.	>40	g-TUBE	Blue Pippin (17kb)

E. coli collection, DNA Extraction, and SMRTbell Library Preparation

Both P4C2 and P5C3 samples were prepared in the same way. *E. coli* K12 genomic DNA was ordered and purified by Lofstrand Labs Limited (K12MG1655 *E. coli*, cat# L3-4001SP2). Field Inversion Gel Electrophoresis (FIGE) was run to ensure presence of high-molecular-weight gDNA. Ten micrograms of gDNA was sheared using g-TUBE devices (Covaris, Inc) spun at 5500 rpm for 1 minute. Three microliters of elution buffer (EB) was added to rinse the upper chamber, spun at 6000 rpm, and spun again at 5500 rpm after inverting the g-TUBE device. SMRTbell libraries were created using the Procedure & Checklist – 20 kb Template Preparation using BluePippin™ Size Selection protocol[35]. Briefly, the library was run on a BluePippin system (Sage Science, Inc., Beverly, MA, USA) to select for SMRTbell templates greater than 10 kb. The resulting average insert size was 17 kb based on 2100 Bioanalyzer instrument (Agilent Technologies Genomics, Santa Clara, CA., USA). Sequencing primers were annealed to the hairpins of the SMRTbell templates followed by binding with the P5 sequencing polymerase and MagBeads (Pacific Biosciences, Menlo Park, CA, USA). One SMRT Cell was run on the PacBio® RS II system with an on-plate concentration of 150 pM using P5-C3 chemistry and a 180-minute data-collection mode.

S. cerevisiae collection, DNA Extraction, and SMRTbell Library Preparation

Please contact J. Li at University of California, San Francisco to obtain the protocol.

A. thaliana collection, DNA Extraction, and SMRTbell Library Preparation

Plants were grown from seeds provided by Lehle seeds (WT-04-19-02). Shoots and leaves were harvested at three weeks and ground in liquid nitrogen using a mortar and pestle. The complete protocol is described in the document "Preparing *Arabidopsis* Genomic DNA for Size-Selected ~20 kb SMRTbell™ Libraries" [36]. This protocol can be used to prepare purified *Arabidopsis* genomic DNA for size-selected SMRTbell templates with average insert sizes of 10 to 20 kb. We recommend

bioRxiv preprint doi: https://doi.org/10.1101/008037; this version posted August 15, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

starting with 20-40 grams of three-week-old *Arabidopsis* whole plants, which can generate >100 µg of purified genomic DNA. SMRTbell libraries were created using the document "Procedure & Checklist – 20 kb Template Preparation using BluePippin™ Size Selection protocol" [35]. Eighty-five SMRT Cells were run on the PacBio RS II system using P4-C2 chemistry and a 180-minute data-collection mode. Forty-six SMRT Cells were run on the PacBio RS II system using P5-C3 chemistry and a 180-minute data-collection mode.

N. crassa OR74A, collection, DNA Extraction, and SMRTbell Library Preparation

The T1 strain of N. crassa, is an A mating type strain derived by DG Catcheside from a cross between the Em a 5297 and Em A 5256 strains he obtained from Stirling Emerson in 1955. The fungus was grown in shake culture for 72 hr at 25°C in 500 ml Vogel's [37] minimal medium containing 2% sucrose. Mycelium was harvested by filtration, ground in liquid nitrogen, resuspended in 10 ml of a buffer containing 0.15 M NaCl, 0.1 M EDTA, 2% SDS at pH 9.5, and incubated overnight at 37°C with 1 mg protease K. Debris was precipitated by centrifugation and 10 ml distilled water was added to the supernatant, which was extracted once with an equal volume of water saturated phenol and once with chloroform. Nucleic acids were precipitated from the aqueous phase with 0.6 volumes of isopropanol. Following centrifugation, the pellet was dried and dissolved in 1 ml TE buffer (TRIS 10 mM, 1 mM EDTA pH 8.0). RNA and protein were digested by overnight incubation at 37°C with RNAase (50 µg) followed by addition of protease K (50 µg) and further incubation for 2 hr. The digest was extracted once with water-saturated phenol and once with chloroform. DNA was collected by precipitation with 0.6 volumes of isopropanol and, following centrifugation, the pellet was dried, dissolved in 500 µl TE buffer and stored at 4°C. Field Inversion Gel Electrophoresis (FIGE) was run to ensure presence of high-molecular-weight gDNA. The genomic DNA was approximately 25 kb and was not sheared. SMRTbell libraries were created using the document "Procedure and Checklist – 10 kb Template Preparation and Sequencing (with Low-Input DNA)"[33] . Two SMRT Cells were run on the PacBio RS II system using P4C2 chemistry and a 180-minute data collection mode.

N. crassa T1 collection, DNA Extraction, and SMRTbell Library Preparation

The T1 strain of *N. crassa*, is an A mating type strain derived by DG Catcheside from a cross between the Em a 5297 and Em A 5256 strains he obtained from Stirling Emerson in 1955. The fungus was grown in shake culture for 72 hr at 25°C in 500 ml Vogel's N [37] minimal medium containing 2% sucrose. Mycelium was harvested by filtration, ground in liquid nitrogen, resuspended in 10 ml of a buffer containing 0.15 M NaCl, 0.1 M EDTA, 2% SDS at pH 9.5, and incubated overnight at 37°C with 1 mg protease K. Debris was precipitated by centrifugation and 10 ml distilled water was added to the supernatant, which was extracted once with an equal volume of water-saturated phenol and once with chloroform. Nucleic acids were precipitated from the aqueous phase with 0.6 volumes of isopropanol. Following centrifugation, the pellet was dried and dissolved in 1 ml TE buffer (TRIS 10 mM, 1 mM EDTA pH 8.0). RNA and protein were digested by overnight incubation at 37°C with RNAase (50 µg) followed by addition of protease K (50 µg) and further incubation for 2 hr. The digest was extracted once with water saturated phenol and once with chloroform. DNA was collected by precipitation with 0.6 volumes of isopropanol and, following centrifugation, the pellet was dried, dissolved in 500 µl TE buffer and stored at 4°C. *Field Inversion Gel Electrophoresis (FIGE) was run to ensure presence of high-molecular-weight gDNA. The genomic DNA was approximately 25 kb and* was not sheared. SMRTbell libraries were created using the document "Procedure and Checklist – 10 kb Template Preparation and Sequencing (with Low-Input DNA)" [33]. Eighteen SMRT Cells were run on the PacBio RS II system using P4-C2 chemistry and a 180-minute data-collection mode.

D. melanogaster collection, DNA Extraction, and SMRTbell Library Preparation

A total of 1.2 g of adult male ISO1 flies corresponding to 1950 animals were collected, starved for 90-120 min and frozen. The flies ranged in age from 0-7 days based on four collections (1) 0-2 days old, 500 males, 0.33 g; (2) 0-4 days old, 500 males, 0.29 g; (3) 0-7 days old, 500 males, 0.29 g; (4) 0-2 days old, 450 males, 0.29 g. Flies were ground in liquid nitrogen to a fine powder and genomic DNA was purified by phenol-chloroform extraction and CsCl banding in the ultracentrifuge. Briefly, the pulverized fly extract was gently re-suspended in 5 ml of HB buffer (7 M Urea, 2% SDS, 50 mM Tris pH7.5, 10 mM EDTA and 0.35 M NaCl) and 5 ml of 1:1 phenol/chloroform. The mixture was shaken slowly for 30 minutes and then centrifuged at 18K rpm for 10 min at 20°C. The aqueous phase was re-extracted twice as above and then precipitated by adding two volumes of ethanol and centrifuging at 18K rpm for 10 min at 20°C. The pellet was re-suspended in 3 ml of TE (10 mM Tris 1 mM EDTA pH 8.0) by gentle

bioRxiv preprint doi: https://doi.org/10.1101/008037; this version posted August 15, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

inversion. To the re-suspended DNA, 3 g CsCl and 0.3 ml of 10 mg/ml ethidium bromide (EtBr) were added and the mixture centrifuged at 45K rpm for 16 hrs at 15°C. The DNA band was collected and the EtBr removed by extraction with CsCl-saturated butanol. The DNA was diluted three-fold with TE, 1/10 vol, 5 M NaCl was added and the DNA precipitated with two volumes of ethanol. After centrifugation, the pellet was washed in 70% ethanol. The DNA was resuspended in 100 µl TE at a concentration of 1.4 µg/µl and quantified using a Nanodrop instrument. This protocol routinely yields at least 10 ng DNA per mg of flies with an estimated DNA size >100 kb.

Genomic DNA was sheared, using a g-TUBE device (Covaris),at 4800 RPM, 150 ng/µl and purified using 0.45x volume ratio of AMPure PB beads. SMRTbell libraries were created using the Procedure & Checklist – 20 kb Template Preparation using BluePippin™ Size Selection [35]. Libraries were ligated with excess adapters and an overnight incubation was performed to increase the yield of ligated fragments larger than 20 kb. Smaller fragments and adapter dimers were then removed by >15 kb size selection using the BluePippin DNA size selection system by Sage Science. Forty-two SMRT Cells were run on the PacBio RS II system. The first run was composed of four SMRT Cells, loaded at 75 pM, 150 pM, 300 pM, and 400 pM in order to determine the optimal loading concentration of the sample. The remaining 38 SMRT Cells were loaded at 400 pM.

Data Records

After DNA extraction, libraries were generated and sequenced at Pacific Biosciences of California, uploaded to Amazon Web Services' Simple Storage Service (S3), and then submitted to the Sequence Read Archive at NCBI under Project ID SRP040522. The corresponding accession numbers and file sizes are listed in Table 1. More detailed information including md5 checksums and links to download the original data from AWS S3 are provided in Supplementary Table S1.

Table 2: Summary of Datasets. Eight datasets from five organisms are described in this paper. Data can be accessed from the Sequence Read Archive (SRA) using the accession numbers provided.

Organism	Strain	Origin	Polymerase & Chemistry Library kits	SRA Accession	Size (GB)
E. coli	MG1655	Lofstrand Labs	P4C2	SRX669475	6.0
E. coli	MG1655	Lofstrand Labs	P5C3	SRX533603	3.8
S. cerevisiae	9464	J. Li	P4C2	SRX533604	38
N. crassa	OR74A	FGSC	P4C2	SRX533605	29
N. crassa	T1	D. Catcheside	P4C2	SRX533606	143
A thaliana	Ler-0	Lehle Seeds	P4C2	SRX533608	263
A. thaliana	Ler-0	Lehle Seeds	P5C3	SRX533607	252
D. melanogaster	ISO1	S. Celniker	P5C3	SRX499318	187

Raw data was transferred from the instrument to a storage location and organized first by the run name, and then by the SMRT Cell directory. Each run contained one or more SMRT Cells. Each SMRT Cell produced a metadata.xml file that recorded the run conditions and barcodes of sequencing kits, three bax.h5 files that contained base call and quality information of actual sequenced data, and one bas.h5

bioRxiv preprint doi: https://doi.org/10.1101/008037; this version posted August 15, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

file that acted as a pointer to consolidate the three bax.h5 files. The "h5" suffix denotes that these are Hierarchical Data format 5 (HDF5) files. The specific contents and structure of a PacBio bax.h5 file is described in more detail in online documentation [38].

Recall the "SMRT bell" structure that underwent sequencing was created by the library preparation process [9]. Sequenced SMRT Bells corresponded to raw reads that may pass around the same base multiple times. A raw read could therefore have a structure that is composed of left adapter→ DNA insert → right adapter → reverse complement of DNA insert → left adapter → DNA insert → and so on. This raw read is typically processed downstream to remove adapters and create subreads composed of the DNA sequence of interest to the investigator. Typical filtering conditions for high-quality SMRT sequence data are read score > 0.8, read length > 100, subread length > 500. In addition, the ends of reads are trimmed if they are outside of high-quality (HQ) regions, and adapter sequences between subreads are removed.

The post-filter statistics of each dataset are listed in Table 3. While raw read lengths reflect the true sequencing capacity of the instrument; only subreads are summarized in Table 3 because it is used in downstream analysis algorithms such as *de novo* assemblers. Multiple subreads can be contained within one raw read, and subreads exclude adapters and low quality sequence. N50 is a statistic used to describe the length distribution of a collection of reads, contigs, or scaffolds, and is defined as the length where 50% of all bases are contained in sequences longer than that length. The N50 filtered subread lengths ranged from 7.6 kb to 10.5 kb for datasets generated with P4-C2 chemistry and ranged from 12.2 kb to 14.2 kb for datasets generated with P5-C3 chemistry. With the exception of *N. crassa* OR74A, all datasets were sequenced to high-coverage (>68X) and sufficient for *de novo* genome assembly applications.

bioRxiv preprint doi: https://doi.org/10.1101/008037; this version posted August 15, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Table 3: Summary statistics of filtered data. Results shown for each dataset are based on output of SMRT Portal analysis using the default filtering parameters (see text for details). Fold coverage is calculated relative to the estimated genome size.

Dataset Name	Number of filtered subreads	N50 filtered subread length (nt)	Maximum filtered subread length (nt)	Total filtered subread (nt)	Estimated genome size (Mb)	Fold coverage
E. coli MG1655 P4C2	61,019	7,586	22,609	331,516,965	5	66X
E. coli MG1655 P5C3	43,063	12,041	28,647	373,874,428	5	75X
S. cerevisiae 9464 P4C2	269,145	8,821	30,164	1,597,871,118	12	133X
N. crassa OR74A P4C2	175,926	7,617	30,845	981,884,113	40	25X
N. crassa T1 P4C2	210,480	10,462	36,227	11,497,185,440	40	287X
A. thaliana Ler-0 P4C2	1,338,320	8,769	41,753	8,129,670,483	120	68X
A. thaliana Ler-0 P5C3	2,067,212	12,188	47,445	17,714,447,516	120	148X
D. melnogaster ISO1 P5C3	1,561,929	14,214	44,766	15,194,174,294	160	95X

Technical Validation

DNA and Sample preparation

To assess the quality of genomic DNA received, we used Qbit (Life Technologies) and Nanodrop (Thermo Scientific) to measure the concentration of genomic DNA. Ideal samples had similar concentration estimates on both platforms, with $A_{230/260/230}$ ratios close to 1:1.8:1, corresponding to what is expected of pure DNA. All samples presented here passed this screening criterion.

Next we assessed the size of the genomic DNA received. For genomic DNA where the size range was less than 17kb, we used the Bioanalyzer 21000 (Agilent) to determine the actual size distribution. For genomic DNA were the size range was greater than 17kb, we opted for pulse field gel electrophoresis to better estimate the larger size distributions. The sizes of the genomic DNA for each sample are listed in Table 1.

To ensure that the library insert sizes were in the optimal size range, we sheared genomic DNA using gTubes if the apparent size was greater than 40 kb. Alternatively, if the size was less than 40kb, then the DNA was not sheared and carried straight through to library preparation. Extremely small fragments (<100bp) and adapter dimers are eliminated by Ampure Beads. Adapter Dimer (0-10bp) and small inserts (11-100bp) represented less than 0.01% of all the reads sequenced in all datasets. We additionally use the Blue Pippin (Sage Science) to select ensure that the libraries had a physical size of 10kb or greater. The size cutoffs used for each sample are listed in Table 1.

bioRxiv preprint doi: https://doi.org/10.1101/008037; this version posted August 15, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Analysis and Quality Filtering

To assess the quality of the libraries sequenced, we examined the percent of bases filtered by a standard QC procedure. Filtering conditions for high-quality SMRT sequence data are read score > 0.8, read length > 100, subread length > 500. In addition, the ends of reads are trimmed if they are outside of high-quality (HQ) regions, and adapter sequences between subreads are removed. All samples retained 71-97% of the bases after filtering.

To ensure that the sequences matched the model organism of interest, we examined the percent of post-filter bases that were mapped to the closest reference genome available. All samples had a mapping rate of 81-95%, with the exception of the Neurospora T1 sample that had a mapping rate of 62%. This sample may have some damaged DNA as it had been stored in a freezer for over 20 years. Nonetheless, preliminary unpublished results show that the sequence from the Neurosporta T1 sample can be successfully assembled into a genome that is more contiguous than the existing reference genome for Neurospora [39].

Usage Notes

The datasets described in this paper were first released on DevNet [40], the PacBio Software Developer Community Network website, with brief descriptions on the PacBio blog. DevNet typically hosts open-source software; SampleNet [30], the PacBio Sample Preparation Community Network website, typically hosts protocols for DNA extraction and library preparation. These websites provide valuable data and documentation about the technology, but are not considered a part of the traditional academic record. This paper in Nature Scientific Data provides an opportunity to describe the methodology and characteristics of the eight datasets in more detail, creates a citable entity for the scientific community, and allows the data to be continually hosted and maintained by the Sequence Read Archive.

DNA sequencing instruments and chemistries change rapidly, and PacBio SMRT sequencing is no exception. The datasets presented here are from P4-C2 and P5-C3 polymerase-chemistry combinations, spanning release dates from late-2013 to early-2014. These datasets represent some of the longest read lengths to date for these chemistries, and can be used to benchmark and develop new algorithms and the state of the art as the technology evolves.

Acknowledgements

The contributions of AMP were funded under Agreement No. HSHQDC-07-C-00020 awarded by the Department of Homeland Security Science and Technology Directorate (DHS/S&T) for the management and operation of the National Biodefense Analysis and Countermeasures Center (NBACC), a Federally Funded Research and Development Center. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security. In no event shall the DHS,

bioRxiv preprint doi: https://doi.org/10.1101/008037; this version posted August 15, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

NBACC, or Battelle National Biodefense Institute (BNBI) have any responsibility or liability for any use, misuse, inability to use, or reliance upon the information contained herein. The Department of Homeland Security does not endorse any products or commercial services mentioned in this publication. CMB was supported by Human Frontier Science Program Young Investigator grant RGY0093/2012.

We thank J. Korlach and E. Hauw for assistance in manuscript preparation, R. Stainer for *Neurospora* T1 sample preparation, and J. Trow for assistance with data submission.

Author contributions

KEK prepared libraries, sequenced, and analyzed data for the *N. crassa* OR74A, *N. crassa* T1, and *D. melanogaster* samples. PP and DRR grew plants from seed, prepared libraries, and sequenced the *A. thaliana* P4C2 and P5C3 datasets. PB prepared libraries and sequenced the *E. coli* datasets. PJY and DE provided DNA for *N. crassa* T1. CY and SEC extracted DNA, and WWF collected male flies for the *D. melanogaster* dataset. NAR and JL extracted DNA and prepared libraries for and PP sequenced the *S. cerevisae* 9464 sample. JML deposited data to the SRA. CSC, AP, CMB and JML analyzed the data and prepared the manuscript. CMB and JML coordinated the project.

Competing Financial Interests

The authors declare competing financial interests. KEK, PP, PB, CSC, NAR, DRR, and JML are employees of Pacific Biosciences of California, Inc., a company commercializing DNA sequencing technologies.

bioRxiv preprint doi: https://doi.org/10.1101/008037; this version posted August 15, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

References

1. Eid, J., et al., *Real-time DNA sequencing from single polymerase molecules.* Science, 2009. **323**(5910): p. 133-8.
2. Korlach, J., et al., *Real-time DNA sequencing from single polymerase molecules.* Methods Enzymol. **472**: p. 431-55.
3. Levene, M.J., et al., *Zero-mode waveguides for single-molecule analysis at high concentrations.* Science, 2003. **299**(5607): p. 682-6.
4. Lundquist, P.M., et al., *Parallel confocal detection of single molecules in real time.* Opt Lett, 2008. **33**(9): p. 1026-8.
5. Roberts, R.J., M.O. Carneiro, and M.C. Schatz, *The advantages of SMRT sequencing.* Genome Biol. **14**(6): p. 405.
6. Clark, T.A., et al., *Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing.* Nucleic Acids Res. **40**(4): p. e29.
7. Song, C.X., et al., *Sensitive and specific single-molecule sequencing of 5-hydroxymethylcytosine.* Nat Methods. **9**(1): p. 75-7.
8. Fang, G., et al., *Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing.* Nat Biotechnol. **30**(12): p. 1232-9.
9. Travers, K.J., et al., *A flexible and efficient template format for circular consensus sequencing and SNP detection.* Nucleic Acids Res. **38**(15): p. e159.
10. Carneiro, M.O., et al., *Pacific biosciences sequencing technology for genotyping and variation discovery in human data.* BMC Genomics. **13**: p. 375.
11. Koren, S., et al., *Hybrid error correction and de novo assembly of single-molecule sequencing reads.* Nat Biotechnol. **30**(7): p. 693-700.
12. Koren, S., et al., *Reducing assembly complexity of microbial genomes with single-molecule sequencing.* Genome Biol. **14**(9): p. R101.
13. Chin, C.S., et al., *Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data.* Nat Methods. **10**(6): p. 563-9.
14. Li, H. and R. Durbin, *Fast and accurate long-read alignment with Burrows-Wheeler transform.* Bioinformatics. **26**(5): p. 589-95.
15. Chaisson, M.J. and G. Tesler, *Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory.* BMC Bioinformatics. **13**: p. 238.
16. English, A.C., et al., *Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology.* PLoS One. **7**(11): p. e47768.
17. English, A.C., W.J.D. Salerno, and J.G.D. Reid, *PBHoney: Identifying Genomic Variants via Long-Read Discordance and Interrupted Mapping.* BMC Bioinformatics. **15**(1): p. 180.
18. Bankevich, A., et al., *SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing.* J Comput Biol. **19**(5): p. 455-77.
19. Mosher, J.J., et al., *Improved performance of the PacBio SMRT technology for 16S rDNA sequencing.* J Microbiol Methods. **104C**: p. 59-60.
20. Tilgner, H., et al., *Defining a personal, allele-specific, and single-molecule long-read transcriptome.* Proc Natl Acad Sci U S A. **111**(27): p. 9869-74.
21. Thomas, S., et al., *Long-read sequencing of chicken transcripts and identification of new transcript isoforms.* PLoS One. **9**(4): p. e94650.
22. Voit, R.A., et al., *Nuclease-mediated gene editing by homologous recombination of the human globin locus.* Nucleic Acids Res. **42**(2): p. 1365-78.

bioRxiv preprint doi: https://doi.org/10.1101/008037; this version posted August 15, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

23. Bendall, M.L., et al., *Exploring the roles of DNA methylation in the metal-reducing bacterium Shewanella oneidensis MR-1.* J Bacteriol. **195**(21): p. 4966-74.
24. Flusberg, B.A., et al., *Direct detection of DNA methylation during single-molecule, real-time sequencing.* Nat Methods. **7**(6): p. 461-5.
25. Kozdon, J.B., et al., *Global methylation state at base-pair resolution of the Caulobacter genome throughout the cell cycle.* Proc Natl Acad Sci U S A. **110**(48): p. E4658-67.
26. Brown, S.D., et al., *Comparison of single-molecule sequencing and hybrid approaches for finishing the genome of Clostridium autoethanogenum and analysis of CRISPR systems in industrial relevant Clostridia.* Biotechnol Biofuels. **7**: p. 40.
27. Itsara, A., et al., *Population analysis of large copy number variants and hotspots of human genetic disease.* Am J Hum Genet, 2009. **84**(2): p. 148-61.
28. Feuk, L., A.R. Carson, and S.W. Scherer, *Structural variation in the human genome.* Nat Rev Genet, 2006. **7**(2): p. 85-97.
29. Stankiewicz, P. and J.R. Lupski, *Structural variation in the human genome and its role in disease.* Annu Rev Med. **61**: p. 437-55.
30. Biosciences, P. *Pacific Biosciences Sample Preparation Community Network*. 2014 [cited; Available from: http://www.smrtcommunity.com/SampleNet.
31. Brizuela, B.J., et al., *Genetic analysis of the brahma gene of Drosophila melanogaster and polytene chromosome subdivisions 72AB.* Genetics, 1994. **137**(3): p. 803-13.
32. Celniker, S.E., et al., *Finishing a whole-genome shotgun: release 3 of the Drosophila melanogaster euchromatic genome sequence.* Genome Biol, 2002. **3**(12): p. RESEARCH0079.
33. Biosciences, P. *Procedure & Checklist - 10 kb Template Preparation and Sequencing (with Low-Input DNA)*. SampleNet 2014 [cited PN 100-152-400-05; Available from: https://na5.salesforce.com/sfc/p/#70000000IVif/a/70000000PVYH/qX1CL1upbnO0rvoeVbk6ZtPPmY4018nY1JzHJKaMYe0=.
34. Biosciences, P. *Procedure & Checklist - Greater Than 10 kb Template Preparation Using AMPure PB Beads*. SampleNet 2014 [cited PN 100-286-100-02; Available from: https://na5.salesforce.com/sfc/p/#70000000IVif/a/70000000PYNC/heYx8OfGiFWX1PwhotTAfUjROSOwZaRMP4FJUXJD6tc=.
35. Biosciences, P. *Procedure & Checklist - 20 kb Template Preparation Using BluePippinTM Size Selection System*. SampleNet 2014 [cited PN 100-286-000-03; Available from: https://na5.salesforce.com/sfc/p/70000000IVif/a/70000000PYNR/UM0ZNjFScqg8WtjFaR2f4YsQTbBVyXIRCjCu9kxLpLM=.
36. Biosciences, P. *Preparing Arabidopsis Genomic DNA for Size-Selected ~20 kb SMRTbell™ Libraries*. 2014 [cited; Available from: http://www.smrtcommunity.com/servlet/servlet.FileDownload?file=00P7000000KMpFEEA1.
37. Vogel, H., *Distrbution of lysine pathways among fungi: Evolutionary implications.* Am Naturalist, 1964. **98**(903): p. 435-446.
38. Biosciences, P. *Pacific Biosciences .bas.h5 file reference guide*. [cited; Available from: http://files.pacb.com/software/instrument/2.0.0/bas.h5%20Reference%20Guide.pdf.
39. P.J. Yeadon, K.E.K., Elizabeth Tseng, Susana Wang, Joan Wilson, David Catcheside, Jane Landolin, *Integrative Biology of a Fungus: User PacBio SMRT Sequencing to interrogate the genome, epigenome, and transcriptome of Neurospora crassa.* http://figshare.com/articles/ENCODE_like_study_using_PacBio_sequencing/928630, FigShare, 2013.
40. Biosciences, P., *Pacific Biosciences Software Developer Community Network*. 2014.

Filename	md5sum	File size (bytes)	Size	S3 location
ecoliK12_tar.gz	07d8f9bcca61876d5d8a5360aa5cd823	6354430732	6G	http://files.pacb.com/datasets/secondary-analysis/ecoli-k12-P4C2-20KSS/ecoliK12.tar.gz
ecoli_P5C3.tgz	e6cd7f18622e4818bbb68f6b8be55a5a	3827754122	3.6G	https://s3.amazonaws.com/datasets.pacb.com/2014/ecoli_P5C3/raw/ecoli_P5C3.tgz
Yeast_9464.tgz	de893b28b3ce0f06a11edfcbe2f61e44	37795558610	35G	https://s3.amazonaws.com/datasets.pacb.com/2013/Yeast/Yeast_9464.tgz
OR74A_rawdata.tgz	d34bb5dd471aa656803567f255be1e8e	29026750071	27G	https://s3.amazonaws.com/datasets.pacb.com/2014/Neurospora/OR74A/raw/OR74A_rawdata.tgz
28SEPT2013_Neuro_371.tgz	5c957a3e9b3dccf108c1d0e6fbdf6522	23366606231	22G	https://s3.amazonaws.com/datasets.pacb.com/2014/Neurospora/T1/raw/28SEPT2013_Neuro_371.tgz
29SEPT2013_Neuro_T1_set1.tgz	8549bc9d314b02b267b26e0375106df1	34598987340	33G	https://s3.amazonaws.com/datasets.pacb.com/2014/Neurospora/T1/raw/29SEPT2013_Nero_T1_set1.tgz
29SEPT2013_Neuro_T1_set2.tgz	d2230963b066f2279c6069fbe7745012	29379712258	28G	https://s3.amazonaws.com/datasets.pacb.com/2014/Neurospora/T1/raw/29SEPT2013_Nero_T1_set2.tgz
29SEPT2013_Neuro_T1_set3.tgz	3080dfe26846f6febc89df1a9f15a715	24413958417	23G	https://s3.amazonaws.com/datasets.pacb.com/2014/Neurospora/T1/raw/29SEPT2013_Nero_T1_set3.tgz
29SEPT2013_Neuro_T1_set4.tgz	6a178ddd45d3cbe4cc37c3dccfaabf79	31549546018	30G	https://s3.amazonaws.com/datasets.pacb.com/2014/Neurospora/T1/raw/29SEPT2013_Nero_T1_set4.tgz
Arabidopsis0_P5C3.tgz	6b867d48b827c684cdab844b64639252	56433467879	53G	https://s3.amazonaws.com/datasets.pacb.com/2014/Arabidopsis/raw/Arabidopsis0_P5C3.tgz
Arabidopsis1_P5C3.tgz	38c8ad4d89cf9c0f7e47f0851b184021	82859079680	82G	https://s3.amazonaws.com/datasets.pacb.com/2014/Arabidopsis/raw/Arabidopsis1_P5C3.tgz
Arabidopsis2_P5C3.tgz	f129bf9497670da4a552ce44705ef458	50116668079	47G	https://s3.amazonaws.com/datasets.pacb.com/2014/Arabidopsis/raw/Arabidopsis2_P5C3.tgz
Arabidopsis3_P5C3.tgz	55cdc7011c9d90e7d67cf58d44ee4e1a	40763469988	38G	https://s3.amazonaws.com/datasets.pacb.com/2014/Arabidopsis/raw/Arabidopsis3_P5C3.tgz
Arabidopsis4_P5C3.tgz	0a2764c62f89ad1e67f663bd4e132177	21868137300	21G	https://s3.amazonaws.com/datasets.pacb.com/2014/Arabidopsis/raw/Arabidopsis4_P5C3.tgz
Arabidopsis0_P4C2.tgz	ba0792cd81343e630b3235e00ed92772	25768905447	24G	https://s3.amazonaws.com/datasets.pacb.com/2013/Arabidopsis-Ler0/raw/Arabidopsis0_P4C2.tgz
Arabidopsis1_P4C2.tgz	814f64f863dbee7d0ab89c229c0197e3	27751138029	26G	https://s3.amazonaws.com/datasets.pacb.com/2013/Arabidopsis-Ler0/raw/Arabidopsis1_P4C2.tgz
Arabidopsis2_P4C2.tgz	5f7c44faaee8b746a0439edf7f7d35f6	35773911210	34G	https://s3.amazonaws.com/datasets.pacb.com/2013/Arabidopsis-Ler0/raw/Arabidopsis2_P4C2.tgz
Arabidopsis3_P4C2.tgz	81f7bf760f7a36c81958a3ce67df7ef6	28015147756	27G	https://s3.amazonaws.com/datasets.pacb.com/2013/Arabidopsis-Ler0/raw/Arabidopsis3_P4C2.tgz
Arabidopsis4_P4C2.tgz	c94598000d9467ca7c45835296bfffdd	27926444439	27G	https://s3.amazonaws.com/datasets.pacb.com/2013/Arabidopsis-Ler0/raw/Arabidopsis4_P4C2.tgz
Arabidopsis5_P4C2.tgz	3d86ac3875ae07f19cb862fd959af535	32736617220	31G	https://s3.amazonaws.com/datasets.pacb.com/2013/Arabidopsis-Ler0/raw/Arabidopsis5_P4C2.tgz
Arabidopsis6_P4C2.tgz	fefb945217c6fcd3de62270d8449639a	35652831091	34G	https://s3.amazonaws.com/datasets.pacb.com/2013/Arabidopsis-Ler0/raw/Arabidopsis6_P4C2.tgz
Arabidopsis7_P4C2.tgz	e6bb14a9cdff49ab3d6daacbd355fa2d	29285289766	28G	https://s3.amazonaws.com/datasets.pacb.com/2013/Arabidopsis-Ler0/raw/Arabidopsis7_P4C2.tgz
Arabidopsis8_P4C2.tgz	ffc0c7ff9e118f270dcc84d109aba46f	20047936437	19G	https://s3.amazonaws.com/datasets.pacb.com/2013/Arabidopsis-Ler0/raw/Arabidopsis8_P4C2.tgz
Dro1_24NOV2013_398	00a51e3e91a7e1124ed6e159f35183bf	14456850906	14G	https://s3.amazonaws.com/datasets.pacb.com/2014/Drosophila/raw/Dro1_24NOV2013_398.tgz
Dro2_25NOV2013_399	473ddb95c959da8508382b7684cb743a	29367287985	27G	https://s3.amazonaws.com/datasets.pacb.com/2014/Drosophila/raw/Dro2_25NOV2013_399.tgz
Dro3_26NOV2013_400	fe0f04dba635f32b475f8c9f2eb46ab4	47083216413	44G	https://s3.amazonaws.com/datasets.pacb.com/2014/Drosophila/raw/Dro3_26NOV2013_400.tgz
Dro4_28NOV2013_401	d9510971c222b70235834aceab5cecfd	42208205056	39G	https://s3.amazonaws.com/datasets.pacb.com/2014/Drosophila/raw/Dro4_28NOV2013_401.tgz
Dro5_29NOV2013_402	7fe82f4448ef6e05afe946a82938ab5d	28078874458	26G	https://s3.amazonaws.com/datasets.pacb.com/2014/Drosophila/raw/Dro5_29NOV2013_402.tgz
Dro6_1DEC2013_403	b412d5dcc9c66155d0374dbf4806a931	26044653729	24G	https://s3.amazonaws.com/datasets.pacb.com/2014/Drosophila/raw/Dro6_1DEC2013_403.tgz

bioRxiv preprint doi: https://doi.org/10.1101/008037; this version posted August 15, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.